**MODERN COLLEGE OF ARTS,SCI. & COMM. PUNE-05.**

**DEPARTMENT OF STATISTICS.**

**M.Sc.( I ) Sem II**

**ST- 28**

**EXPT.NO. 4.**

**TITLE :Multiple regression (selection of variable)**

1)Consider the following solar thermal energy data given below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Y | x1 | x2 | x3 | x4 | x5 |
| 271.8 | 783.35 | 33.53 | 40.55 | 16.66 | 13.2 |
| 264 | 748.45 | 36.5 | 36.19 | 16.46 | 14.11 |
| 238.8 | 684.45 | 34.66 | 37.31 | 17.66 | 15.68 |
| 230.7 | 827.8 | 33.13 | 32.52 | 17.5 | 10.53 |
| 251.6 | 860.45 | 35.75 | 33.71 | 16.4 | 11 |
| 257.9 | 875.15 | 34.46 | 34.14 | 16.28 | 11.31 |
| 263.9 | 909.45 | 34.6 | 34.85 | 16.06 | 11.96 |
| 266.5 | 905.55 | 35.38 | 35.89 | 15.93 | 12.58 |
| 229.1 | 756 | 35.85 | 33.53 | 16.6 | 10.66 |
| 239.3 | 769.35 | 35.68 | 33.79 | 16.41 | 10.85 |
| 258 | 793.5 | 35.35 | 34.72 | 16.17 | 11.41 |
| 257.6 | 801.65 | 35.04 | 35.22 | 15.92 | 11.91 |
| 267.3 | 819.65 | 34.07 | 36.5 | 16.04 | 12.85 |
| 267 | 808.55 | 32.2 | 37.6 | 16.19 | 13.58 |
| 259.6 | 774.95 | 34.32 | 37.89 | 16.62 | 14.21 |
| 240.4 | 711.85 | 31.08 | 37.71 | 17.37 | 15.56 |
| 227.2 | 694.85 | 35.73 | 37 | 18.12 | 15.83 |
| 196 | 638.1 | 34.11 | 36.76 | 18.53 | 16.41 |
| 278.7 | 774.55 | 34.79 | 34.62 | 15.54 | 13.1 |
| 272.3 | 757.9 | 35.77 | 35.4 | 15.7 | 13.63 |
| 267.4 | 753.35 | 36.44 | 35.96 | 16.45 | 14.51 |
| 254.5 | 704.7 | 37.82 | 36.26 | 17.62 | 15.38 |
| 224.7 | 666.8 | 35.07 | 36.34 | 18.12 | 16.1 |
| 181.5 | 568.55 | 35.26 | 35.9 | 19.05 | 16.73 |
| 227.5 | 653.1 | 35.56 | 31.84 | 16.51 | 10.58 |
| 253.6 | 704.05 | 35.73 | 33.16 | 16.02 | 11.28 |
| 263 | 709.6 | 36.46 | 33.83 | 15.89 | 11.91 |
| 265.8 | 726.9 | 36.26 | 34.89 | 15.83 | 12.65 |
| 263.8 | 697.15 | 37.2 | 36.27 | 16.71 | 14.06 |

a) Use forward selection method algorithm to select best subset regression model.

b) Use backward elimination algorithm to select best subset regression model.

c) Use step wise regression to select best subset regression model.

d) Compare R^2, Cp and MSRes for each model and write the best subset regression model.

2) Hald [1952] presents data concerning the heat evolved in calories per gram of cement (y) as a function of the amount of each of four ingredients in the mix: tricalcium aluminate (x1), tricalcium silicate(x2), tetra calcium alumino ferrite (x3) and dicalcium silicate (x4).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Observation  i | yi | xi1 | xi2 | xi3 | xi4 |
| 1 | 78.5 | 7 | 26 | 6 | 60 |
| 2 | 74.3 | 1 | 29 | 15 | 52 |
| 3 | 104.3 | 11 | 56 | 8 | 20 |
| 4 | 87.6 | 11 | 31 | 8 | 47 |
| 5 | 95.9 | 7 | 52 | 6 | 33 |
| 6 | 109.2 | 11 | 55 | 9 | 22 |
| 7 | 102.7 | 3 | 71 | 17 | 6 |
| 8 | 72.5 | 1 | 31 | 22 | 44 |
| 9 | 93.1 | 2 | 54 | 18 | 22 |
| 10 | 115.9 | 21 | 47 | 4 | 26 |
| 11 | 83.8 | 1 | 40 | 23 | 34 |
| 12 | 113.3 | 11 | 66 | 9 | 12 |
| 13 | 109.4 | 10 | 68 | 8 | 12 |

a) Use forward selection method algorithm to select best subset regression model.

b) Use backward elimination algorithm to select best subset regression model.

c) Use step wise regression to select best subset regression model.

d) Compare R^2, Cp and MSRes for each model and write the best subset regression model.

ALGORITHM

1. Forward selection method:-

step 1:- this method being with a regression model consisting of the intercept form only.

step 2:- we have a set of the p regressors the first regressor selected in the one that has the largest simple correlation with the response variable y. suppose that X1 has the largest correlation with the λ1 (largest value of a F-statistic) .

when we add X1 in the model

if F-statistic >Fin where ,Fin=F1,n-p-1,α

step 3:- Now the next variable can be added by using partial F-statistics

partial F- statistics for regressor xj given x1 is

Fxj |x1 = SSR(xj|x1) / MSres(x1,xj) j=2,3,……,p

= SSR(xj,x1) – SSR(x1) / MSres(x1,xj)

Find the largest partial F-statistics

Suppose partial F-statistics corresponding to x2 largest

Step 4:- when we add x2

If partial F-statistics X2  > Fin where ,Fin=F1,n-p-1,α

Step 5:- Now x1 and x2 are in the model then we add the next variable by calculating partial F-statistics for remaining variables

Fxj |x1,x2  = SSR(xj|x1,x2) / MSres(x1,x2,xj) j=2,3,……,p

= SSR(xj,x1,x2) – SSR(x1,x2) / MSres(x1,x2,xj)

Find the largest partial F-statistics

Suppose partial F-statistics corresponding to x3 largest

Step 6 :- we terminate this procedure if partial F-statistics < Fin

1. Backward elimination method:-

Step 1:- Work in the opposite direction of a forward selection method

step 2 :- Begins with the regression model that includes all the regressors

step 3:- the lowest partial F-statistic value Flow is compared with a preselected

F value say Fout

step 4:- Fout = F1,n-p-1,α where p is no. of regressor

step 5:- if Flow < Fout the regressor is a removed from the model

step 6:- now fit the regression model with the remaining regressor

step 7:- partial F- statistics are calculated and the procedure is repeated

Step 8:- the method terminates when the smallest partial F value is greater than the preselected cut-off value Fout

1. Stepwise regression method

step 1:-it is a combination of a forward selection and backward elimination method

step 2:- begins with the no regressor

step 3:- enter the variable using the forward selection method (for example X1)

step 4 :- now enter the next variable using the forward selection method (for example X2)

step 5:- now we check whether the regression X1 included in the step 3 is to be removed or not

step 6:- we calculate partial F-statistic

F1|2  = fit a regression model with X1 and X2

Fout = F1,n-p-1,α where p is no. of regressor

step 7:- Since F1|4 >Fout regressor X1 is remain in the model ( note that there are we have used a backward elimination method)

step 8:- we use a forward selection method to check whether X3 or X4 can be added in the model

partial F- statistic are

F3|1,2 = fit a regression model on X1 X2 X3

F4|1,2 = fit a regression model on X1 X2 X4

Fin=F1,n-p-1,α

Step 9:- if the largest partial F-statistic value is less than Fin  this implies that procedure is terminates here.

**SOLUTION**

Q1)

1. **Forward selection method :-**

**Correlation: Y, x1, x2, x3, x4, x5**

Y x1 x2 x3 x4

x1 0.628

x2 0.102 -0.204

x3 0.112 -0.107 -0.329

x4 **-0.849** -0.634 -0.117 0.287

x5 -0.351 -0.584 -0.065 0.697 0.685

**Regression Analysis: Y versus x1, x2, x3, x4, x5**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 5 13195.5 89.88% 13195.5 2639.11 40.84 0.000

x1 1 5783.8 39.39% 350.6 350.56 5.42 0.029

x2 1 811.7 5.53% 270.1 270.13 4.18 0.053

x3 1 1181.5 8.05% 437.2 437.15 6.76 0.016

x4 1 5303.0 36.12% 4656.6 4656.56 **72.05** 0.000

x5 1 115.5 0.79% 115.5 115.50 1.79 0.194

Error 23 1486.4 10.12% 1486.4 64.63

**Regression Analysis: Y versus x4**

The regression equation is

Y = 607.1 - 21.40 x4

S = 12.3277 R-Sq = 72.1% R-Sq(adj) = 71.0%

Analysis of Variance

Source DF SS MS F P

Regression 1 10578.7 10578.7 0.000

Error 27 4103.2 152.0

Total 28 14681.9

F distribution with 1 DF in numerator and 27 DF in denominator

P( X ≤ x ) x

0.95 4.21001

Decision:-

Here, Fcal>Fin i.e**.** 69.61**>** 4.21001

The 1 st entering variable is X4.

Therefore , regression equation is

Y = 607.1 - 21.40 x4

**Next entering variable is as follows :-**

Calculating partial F statistics:-

**Regression Analysis: Y versus x4, x1**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 10774.5 73.39% 10774.5 5387.2 35.85 0.000

x4 1 10578.7 72.05% 4990.7 4990.7 33.21 0.000

x1 1 195.8 1.33% 195.8 195.8 **1.30** 0.264

Error 26 3907.4 26.61% 3907.4 150.3

Total 28 14681.9 100.00%

**Regression Analysis: Y versus x4, x2**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 10578.8 72.05% 10578.8 5289.4 33.52 0.000

x4 1 10578.7 72.05% 10425.0 10425.0 66.06 0.000

**x2** 1 0.1 0.00% 0.1 0.1 **0.00** 0.979

Error 26 4103.1 27.95% 4103.1 157.8

Total 28 14681.9 100.00%

**Regression Analysis: Y versus x4, x3**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 12608 85.87% 12608 6303.8 79.01 0.000

x4 1 10579 72.05% 12423 12423.1 155.71 0.000

**x3**  1 2029 13.82% 2029 2028.9 25.43 0.000

Error 26 2074 14.13% 2074 79.8

Total 28 14682 100.00%

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 12042 82.02% 12042 6021.1 59.31 0.000

x4 1 10579 72.05% 10232 10232.1 100.78 0.000

**x5** 1 1464 9.97% 1464 1463.5 **14.42** 0.001

Error 26 2640 17.98% 2640 101.5

Total 28 14682 100.00%

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 26 DF in denominator

P( X ≤ x ) x

0.95 4.22520

The largest partial F-statistic Y on X4,X3 is 25.43

Decision:-

Here, Fcal>Fin

25.43 > 4.22520

**Therefore next entering variable is X3. So, the regression model is Y on X4 ,X3 is**

**Regression Analysis: Y versus x4, x3**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 12608 85.87% 12608 6303.8 79.01 0.000

x4 1 10579 72.05% 12423 12423.1 155.71 0.000

x3 1 2029 13.82% 2029 2028.9 25.43 0.000

Error 26 2074 14.13% 2074 79.8

Total 28 14682 100.00%

Regression Equation

**Y = 483.7 - 24.22 x4 + 4.796 x3**

**Next entering variable is,**

**Regression Analysis: Y versus x4, x3, x1**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 3 12696.8 86.48% 12696.8 4232.26 53.30 0.000

x4 1 10578.7 72.05% 6436.9 6436.92 81.06 0.000

x3 1 2028.9 13.82% 1922.3 1922.28 24.21 0.000

x1 1 89.2 0.61% 89.2 89.17 **1.12** 0.299

Error 25 1985.2 13.52% 1985.2 79.41

**Regression Analysis: Y versus x4, x3, x2**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 3 12833.9 87.41% 12833.9 4278.0 57.87 0.000

x4 1 10578.7 72.05% 12330.6 12330.6 166.80 0.000

x3 1 2028.9 13.82% 2255.1 2255.1 30.51 0.000

x2 1 226.3 1.54% 226.3 226.3 **3.06** 0.092

Error 25 1848.1 12.59% 1848.1 73.9

Total 28 14681.9 100.00%

**Regression Analysis: Y versus x4, x3, x5**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 3 12681.7 86.38% 12681.7 4227.23 52.83 0.000

x4 1 10578.7 72.05% 7244.3 7244.30 90.54 0.000

x3 1 2028.9 13.82% 639.5 639.49 7.99 0.009

x5 1 74.1 0.50% 74.1 74.10 **0.93** 0.345

Error 25 2000.2 13.62% 2000.2 80.01

Total 28 14681.9 100.00%

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 25 DF in denominator

P( X ≤ x ) x

0.95 **4.24170**

The largest partial F-statistic Y on X4,X3,X2

is **3.06**

decision:-

Here, Fcal< Fin

**3.06** < **4.24170**

**The largest partial F-statistic is less than Fin**

**Therefore , we terminated this procedure here and**

**Final regression model is :-**

**Regression Analysis: Y versus x4, x3**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 12608 85.87% 12608 6303.8 79.01 0.000

x4 1 10579 72.05% 12423 12423.1 155.71 0.000

x3 1 2029 13.82% 2029 2028.9 25.43 0.000

Error 26 2074 14.13% 2074 79.8

Total 28 14682 100.00%

Regression Equation

**Y = 483.7 - 24.22 x4 + 4.796 x3**

**b)**

**backward elimination method:-**

**Regression Analysis: Y versus x1, x2, x3, x4, x5**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 5 13195.5 89.88% 13195.5 2639.11 40.84 0.000

x4 1 10578.7 72.05% 4656.6 4656.56 72.05 0.000

x3 1 2028.9 13.82% 437.2 437.15 6.76 0.016

**x5** 1 74.1 0.50% 115.5 115.50 **1.79** 0.194

x1 1 243.7 1.66% 350.6 350.56 5.42 0.029

x2 1 270.1 1.84% 270.1 270.13 4.18 0.053

Error 23 1486.4 10.12% 1486.4 64.63

Total 28 14681.9 100.00%

Regression Equation

Y = 325.4 - 22.95 x4 + 3.80 x3 + 2.42 x5 + 0.0675 x1 + 2.55 x2

Fout=F1,n-p-1 , Flow= **1.79**

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 23 DF in denominator

P( X ≤ x ) x

0.95 4.27934

Decision:-

here, Fout > Flow  i.e. 4.27934**> 1.79**

Therefore 1 st leaving variable is X5.

Partial F statistic are :-

**Regression Analysis: Y versus x1, x2, x3, x4**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 4 13080.0 89.09% 13080.0 3270.01 48.99 0.000

x4 1 10578.7 72.05% 5303.0 5303.03 79.45 0.000

x3 1 2028.9 13.82% 2271.9 2271.88 34.04 0.000

**x1** 1 89.2 0.61% 246.2 246.16 **3.69** 0.067

x2 1 383.3 2.61% 383.3 383.26 5.74 0.025

Error 24 1601.9 10.91% 1601.9 66.75

Total 28 14681.9 100.00%

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 24 DF in denominator

P( X ≤ x ) x

0.95 **4.25968**

Decision:-

here, Fout > Flow  i. **4.25968 > 3.69**

Therefore 2 nd leaving variable is X1.

Regression model is :-

**Regression Analysis: Y versus x2, x3, x4**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 3 12833.9 87.41% 12833.9 4278.0 57.87 0.000

x4 1 10578.7 72.05% 12330.6 12330.6 166.80 0.000

x3 1 2028.9 13.82% 2255.1 2255.1 30.51 0.000

**x2** 1 226.3 1.54% 226.3 226.3 **3.06** 0.092

Error 25 1848.1 12.59% 1848.1 73.9

Total 28 14681.9 100.00%

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 25 DF in denominator

P( X ≤ x ) x

0.95 4.24170

Decision:-

here, Fout > Flow  i.e. 4.24170**> 3.06**

Therefore 3rd leaving variable is X2.

Regression model is :-

**Regression Analysis: Y versus x4, x3**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 12608 85.87% 12608 6303.8 79.01 0.000

x4 1 10579 72.05% 12423 12423.1 155.71 0.000

**x3** 1 2029 13.82% 2029 2028.9 **25.43** 0.000

Error 26 2074 14.13% 2074 79.8

Total 28 14682 100.00%

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 26 DF in denominator

P( X ≤ x ) x

0.95 4.22520

Decision:-

here, Fout < Flow  i.e. 4.22520 **<** **25.43**

**Here, the backward elimination method is terminated because smallest partial F statistic is greater than preselected cut of value .**

**Final regression model is :-**

**Regression Analysis: Y versus x4, x3**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 12608 85.87% 12608 6303.8 79.01 0.000

x4 1 10579 72.05% 12423 12423.1 155.71 0.000

x3 1 2029 13.82% 2029 2028.9 25.43 0.000

Error 26 2074 14.13% 2074 79.8

Total 28 14682 100.00%

Regression Equation

**Y = 483.7 - 24.22 x4 + 4.796 x3**

**c)**

**stepwise regression method:-**

**Correlation: Y, x1, x2, x3, x4, x5**

Y x1 x2 x3 x4

x1 0.628

x2 0.102 -0.204

x3 0.112 -0.107 -0.329

x4 **-0.849** -0.634 -0.117 0.287

x5 -0.351 -0.584 -0.065 0.697 0.685

**Regression Analysis: Y versus x1, x2, x3, x4, x5**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 5 13195.5 89.88% 13195.5 2639.11 40.84 0.000

x1 1 5783.8 39.39% 350.6 350.56 5.42 0.029

x2 1 811.7 5.53% 270.1 270.13 4.18 0.053

x3 1 1181.5 8.05% 437.2 437.15 6.76 0.016

**x4** 1 5303.0 36.12% 4656.6 4656.56 **72.05** 0.000

x5 1 115.5 0.79% 115.5 115.50 1.79 0.194

Error 23 1486.4 10.12% 1486.4 64.63

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 23 DF in denominator

P( X ≤ x ) x

0.95 **4.27934**

Decision:-

Here, Fcal>Fvalue i.e. **72.05** > **4.27934**

The 1 st entering variable is X4.

Therefore , regression equation is

**Regression Analysis: Y versus x4**

The regression equation is

Y = 607.1 - 21.40 x4

S = 12.3277 R-Sq = 72.1% R-Sq(adj) = 71.0%

Analysis of Variance

Source DF SS MS F P

Regression 1 10578.7 10578.7 69.61 0.000

Error 27 4103.2 152.0

Total 28 14681.9

**Next entering variable is as follows :-**

Calculating partial F statistics:-

**Regression Analysis: Y versus x4, x1**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 10774.5 73.39% 10774.5 5387.2 35.85 0.000

x4 1 10578.7 72.05% 4990.7 4990.7 33.21 0.000

**x1** 1 195.8 1.33% 195.8 195.8 **1.30** 0.264

Error 26 3907.4 26.61% 3907.4 150.3

Total 28 14681.9 100.00%

**Regression Analysis: Y versus x4, x2**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 10578.8 72.05% 10578.8 5289.4 33.52 0.000

x4 1 10578.7 72.05% 10425.0 10425.0 66.06 0.000

**x2** 1 0.1 0.00% 0.1 0.1 **0.00** 0.979

Error 26 4103.1 27.95% 4103.1 157.8

Total 28 14681.9 100.00%

**Regression Analysis: Y versus x4, x3**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 12608 85.87% 12608 6303.8 79.01 0.000

x4 1 10579 72.05% 12423 12423.1 155.71 0.000

**x3** 1 2029 13.82% 2029 2028.9 25.43 0.000

Error 26 2074 14.13% 2074 79.8

Total 28 14682 100.00%

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 12042 82.02% 12042 6021.1 59.31 0.000

x4 1 10579 72.05% 10232 10232.1 100.78 0.000

**x5** 1 1464 9.97% 1464 1463.5 **14.42** 0.001

Error 26 2640 17.98% 2640 101.5

Total 28 14682 100.00%

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 26 DF in denominator

P( X ≤ x ) x

0.95 4.22520

Decision :-

The largest partial F-statistic Y on X4,X3 is 25.43

25.43 > 4.22520

**Therefore next entering variable is X3. So, next regression model is Y on X4 ,X3 is**

**Regression Analysis: Y versus x4, x3**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 12608 85.87% 12608 6303.8 79.01 0.000

x4 1 10579 72.05% 12423 12423.1 155.71 0.000

**x3** 1 2029 13.82% 2029 2028.9 **25.43** 0.000

Error 26 2074 14.13% 2074 79.8

Total 28 14682 100.00%

Regression Equation

**Y = 483.7 - 24.22 x4 + 4.796 x3**

Here , first we have to check leaving variable :-

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 26 DF in denominator

P( X ≤ x ) x

0.95 4.22520

Decision:-

here, Fout < Flow  i.e. 4.22520**< 25.43**

Therefore, variable X3 is in the model.

**Selecting of next entering variable:-**

**Regression Analysis: Y versus x4, x3, x1**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 3 12696.8 86.48% 12696.8 4232.26 53.30 0.000

x4 1 10578.7 72.05% 6436.9 6436.92 81.06 0.000

x3 1 2028.9 13.82% 1922.3 1922.28 24.21 0.000

**x1** 1 89.2 0.61% 89.2 89.17 **1.12** 0.299

Error 25 1985.2 13.52% 1985.2 79.41

**Regression Analysis: Y versus x4, x3, x2**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 3 12833.9 87.41% 12833.9 4278.0 57.87 0.000

x4 1 10578.7 72.05% 12330.6 12330.6 166.80 0.000

x3 1 2028.9 13.82% 2255.1 2255.1 30.51 0.000

**x2** 1 226.3 1.54% 226.3 226.3 **3.06** 0.092

Error 25 1848.1 12.59% 1848.1 73.9

Total 28 14681.9 100.00%

**Regression Analysis: Y versus x4, x3, x5**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 3 12681.7 86.38% 12681.7 4227.23 52.83 0.000

x4 1 10578.7 72.05% 7244.3 7244.30 90.54 0.000

x3 1 2028.9 13.82% 639.5 639.49 7.99 0.009

**x5** 1 74.1 0.50% 74.1 74.10 **0.93** 0.345

Error 25 2000.2 13.62% 2000.2 80.01

Total 28 14681.9 100.00%

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 25 DF in denominator

P( X ≤ x ) x

0.95 4.24170

Decision:-

The largest partial F-statistic Y on X4,X3,X2

is **3.06**

Here, Fcal< Fin

**3.06 <** 4.24170

**The largest partial F-statistic is less than Fin**

**Therefore , we terminated this procedure here and**

**Final regression model is :-**

**Regression Analysis: Y versus x4, x3**

Analysis of Variance

Source DF Seq SS Contribution Adj SS Adj MS F-Value P-Value

Regression 2 12608 85.87% 12608 6303.8 79.01 0.000

x4 1 10579 72.05% 12423 12423.1 155.71 0.000

x3 1 2029 13.82% 2029 2028.9 25.43 0.000

Error 26 2074 14.13% 2074 79.8

Total 28 14682 100.00%

Model Summary

S R-sq R-sq(adj) PRESS R-sq(pred)

8.93207 85.87% 84.78% 2736.51 81.36%

Coefficients

Term Coef SE Coef 95% CI T-Value P-Value VIF

Constant 483.7 39.6 ( 402.3, 565.0) 12.22 0.000

x4 -24.22 1.94 (-28.20, -20.23) -12.48 0.000 1.09

x3 4.796 0.951 ( 2.841, 6.751) 5.04 0.000 1.09

**Regression Equation**

**Y = 483.7 - 24.22 x4 + 4.796 x3**

Fits and Diagnostics for Unusual Observations

Obs Y Fit SE Fit 95% CI Resid Std Resid Del Resid HI

1 271.80 274.74 5.08 (264.30, 285.17) -2.94 -0.40 -0.39 0.323182

22 254.50 230.91 2.39 (226.01, 235.82) 23.59 2.74 3.19 0.071388

Obs Cook’s D DFITS

1 0.03 -0.271688 X

22 0.19 0.883376 R

R Large residual

X Unusual X

**d)**

**Best Subsets Regression: Y versus x1, x2, x3, x4, x5**

Response is Y

R-Sq R-Sq Mallows x x x x x

Vars R-Sq (adj) PRESS (pred) Cp S 1 2 3 4 5

1 72.1 71.0 4855.9 66.9 38.5 12.328 X

1 39.4 37.1 10822.6 26.3 112.7 18.154 X

2 85.9 84.8 2736.5 81.4 9.1 8.9321 X X

2 82.0 80.6 3786.4 74.2 17.8 10.076 X X

3 87.4 85.9 3089.7 79.0 7.6 8.5978 X X X

3 86.5 84.9 2725.9 81.4 9.7 8.9110 X X X

4 89.1 87.3 2847.2 80.6 5.8 8.1698 X X X X

4 88.0 86.0 3045.7 79.3 8.2 8.5550 X X X X

5 89.9 87.7 3109.9 78.8 6.0 8.0390 X X X X X

In the above table we observed that 4 variable regression model(subset) have low mallow cp statistic & Msres and high Rsq(adj) therefore, 4 variable regression model (subset)t is best(adequate) among all subsets.

**Q2)**

**a)**

**Forward selection method**

**Correlation: yi, xi1, xi2, xi3, xi4**

yi xi1 xi2 xi3

xi1 0.731

xi2 0.816 0.229

xi3 -0.535 -0.824 -0.139

**xi4 -0.821** -0.245 -0.973 0.030

**Here, maximum correlation occurs in y and x4i .therefore 1st entering variable is x4i.**

**Since , regression equation is**

**Regression Analysis: yi versus xi4**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 1 1831.9 1831.90 22.80 0.001

**xi4** 1 1831.9 1831.90 **22.80** 0.001

Error 11 883.9 80.35

Lack-of-Fit 9 746.7 82.96 1.21 0.532

Pure Error 2 137.2 68.61

Total 12 2715.8

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 11 DF in denominator

P( X ≤ x ) x

0.95 4.8443

Decision:-

Here, Fcal>Fin i.e. **22.80** > 4.8443

The 1 st entering variable is xi4..

Therefore , regression equation is

**Regression Analysis: yi versus xi4**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 1 1831.9 1831.90 22.80 0.001

xi4 1 1831.9 1831.90 22.80 0.001

Error 11 883.9 80.35

Lack-of-Fit 9 746.7 82.96 1.21 0.532

Pure Error 2 137.2 68.61

Total 12 2715.8

**egression Equation**

**yi = 117.57 - 0.738 xi4**

**Next entering variable is as follows :-**

Calculating partial F statistics:-

**Regression Analysis: yi versus xi4, xi1**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2641.00 1320.50 176.63 0.000

xi4 1 1190.92 1190.92 159.30 0.000

**xi1** 1 809.10 809.10 **108.22** 0.000

Error 10 74.76 7.48

Total 12 2715.76

**Regression Analysis: yi versus xi4, xi2**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 1846.88 923.44 10.63 0.003

xi4 1 37.46 37.46 0.43 0.526

**xi2** 1 14.99 14.99 **0.17** 0.687

Error 10 868.88 86.89

Total 12 2715.76

**Regression Analysis: yi versus xi4, xi3**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2540.0 1270.01 72.27 0.000

xi4 1 1763.7 1763.66 100.36 0.000

**xi3** 1 708.1 708.13 **40.29** 0.000

Error 10 175.7 17.57

Total 12 2715.8

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 10 DF in denominator

P( X ≤ x ) x

0.95 4.96460

The largest partial F-statistic Y on x2i,x1i is **108.22**

decision:-

Here, Fcal>Fin

**108.22** > 4.96460

**Therefore 2nd entering variable is xi1.**

**Regression Analysis: yi versus xi4, xi1**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2641.00 1320.50 176.63 0.000

xi4 1 1190.92 1190.92 159.30 0.000

**xi1** 1 809.10 809.10 **108.22** 0.000

Error 10 74.76 7.48

Total 12 2715.76

**Regression Equation**

**yi = 103.10 - 0.6140 xi4 + 1.440 xi1**

next entering variableis:-

partial F-statistics are:-

**Regression Analysis: yi versus xi4, xi1, xi2**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 3 2667.79 889.263 166.83 0.000

xi4 1 9.93 9.932 1.86 0.205

xi1 1 820.91 820.907 154.01 0.000

**xi2** 1 26.79 26.789 **5.03** 0.052

Error 9 47.97 5.330

Total 12 2715.76

**Regression Analysis: yi versus xi4, xi1, xi3**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 3 2664.93 888.31 157.27 0.000

xi4 1 1176.24 1176.24 208.24 0.000

xi1 1 124.90 124.90 22.11 0.001

**xi3** 1 23.93 23.93 **4.24** 0.070

Error 9 50.84 5.65

Total 12 2715.76

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 9 DF in denominator

P( X ≤ x ) x

0.95 5.11736

The largest partial F-statistic Y on x2i,x1i,x4i is  **5.03**

Decision:-

Here, Fcal>Fin

**5.03** < 5.11736

**The largest partial F-statistic is less than Fin**

**Therefore , we terminated this procedure here and**

**Final regression model is :-**

**Regression Analysis: yi versus xi4, xi1**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2641.00 1320.50 176.63 0.000

xi4 1 1190.92 1190.92 159.30 0.000

**xi1** 1 809.10 809.10 **108.22** 0.000

Error 10 74.76 7.48

Total 12 2715.76

**Regression Equation**

**yi = 103.10 - 0.6140 xi4 + 1.440 xi1**

**b)**

**Backward elimination method is,**

**Regression Analysis: yi versus xi1, xi2, xi3, xi4**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 4 2667.90 666.975 111.48 0.000

xi2 1 2.97 2.972 0.50 0.501

xi1 1 25.95 25.951 4.34 0.071

xi4 1 0.25 0.247 0.04 0.844

**xi3** 1 0.11 0.109 **0.02** 0.896

Error 8 47.86 5.983

Total 12 2715.76

Model Summary

S R-sq R-sq(adj) R-sq(pred)

2.44601 98.24% 97.36% 95.94%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 62.4 70.1 0.89 0.399

xi2 0.510 0.724 0.70 0.501 254.42

xi1 1.551 0.745 2.08 0.071 38.50

xi4 -0.144 0.709 -0.20 0.844 282.51

xi3 0.102 0.755 0.14 0.896 46.87

Regression Equation

yi = 62.4 + 0.510 xi2 + 1.551 xi1 - 0.144 xi4 + 0.102 xi3

Fout=F1,n-p-1 , lowest partial F statistic is :Flow= **0.02**

**Inverse Cumulative Distribution Function**

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 8 DF in denominator

P( X ≤ x ) x

0.95 5.31766

Decision:-

here, Fout > Flow  i.e. 5.31766 **> 0.02**

Therefore 1 st leaving variable is xi3.

Regression model is :-

**Regression Analysis: yi versus xi1, xi2, xi4**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 3 2667.79 889.263 166.83 0.000

xi2 1 26.79 26.789 5.03 0.052

xi1 1 820.91 820.907 154.01 0.000

xi4 1 9.93 9.932 1.86 0.205

Error 9 47.97 5.330

Total 12 2715.76

Regression Equation

**yi = 71.6 + 0.416 xi2 + 1.452 xi1 - 0.237 xi4**

**Partial F statistic are :-**

**Regression Analysis: yi versus xi1, xi2, xi4**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 3 2667.79 889.263 166.83 0.000

xi2 1 26.79 26.789 5.03 0.052

xi1 1 820.91 820.907 154.01 0.000

**xi4** 1 9.93 9.932 **1.86** 0.205

Error 9 47.97 5.330

Total 12 2715.76

Fout=F1,n-p-1 , lowest partial F statistic is :Flow= **0.02**

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 9 DF in denominator

P( X ≤ x ) x

0.95 5.11736

Decision:-

here, Fout > Flow  i.e. 5.11736**> 1.86**

Therefore 2nd leaving variable is xi4.

Regression model is:-

**Regression Analysis: yi versus xi1, xi2**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2657.86 1328.93 229.50 0.000

xi2 1 1207.78 1207.78 208.58 0.000

xi1 1 848.43 848.43 146.52 0.000

Error 10 57.90 5.79

Total 12 2715.76

Model Summary

S R-sq R-sq(adj) R-sq(pred)

2.40634 97.87% 97.44% 96.54%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 52.58 2.29 23.00 0.000

xi2 0.6623 0.0459 14.44 0.000 1.06

xi1 1.468 0.121 12.10 0.000 1.06

**Regression Equation**

**yi = 52.58 + 0.6623 xi2 + 1.468 xi1**

**Partial f-statistic are:-**

**Regression Analysis: yi versus xi1, xi2**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2657.86 1328.93 229.50 0.000

xi2 1 1207.78 1207.78 208.58 0.000

**xi1** 1 848.43 848.43 **146.52** 0.000

Error 10 57.90 5.79

Total 12 2715.76

Fout=F1,n-p-1 , lowest partial F statistic is :Flow= **146.52**

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 10 DF in denominator

P( X ≤ x ) x

0.95 4.96460

Decision:-

here, Fout > Flow  i.e. 4.96460 < **146.52**

**Here, the backward elimination method is terminated because smallest partial F statistic is greater than preselected cut of value .**

**Final regression model is :-**

**Regression Analysis: yi versus xi1, xi2**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2657.86 1328.93 229.50 0.000

xi2 1 1207.78 1207.78 208.58 0.000

xi1 1 848.43 848.43 146.52 0.000

Error 10 57.90 5.79

Total 12 2715.76

Model Summary

S R-sq R-sq(adj) R-sq(pred)

2.40634 97.87% 97.44% 96.54%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 52.58 2.29 23.00 0.000

xi2 0.6623 0.0459 14.44 0.000 1.06

xi1 1.468 0.121 12.10 0.000 1.06

**Regression Equation**

**yi = 52.58 + 0.6623 xi2 + 1.468 xi1**

**c)**

**Stepwise selection method :-**

**Correlation: yi, xi1, xi2, xi3, xi4**

yi xi1 xi2 xi3

xi1 0.731

xi2 0.816 0.229

xi3 -0.535 -0.824 -0.139

**xi4 -0.821** -0.245 -0.973 0.030

**Here, maximum correlation occurs in y and x4i .therefore 1st entering variable is x4i.**

**Since , regression equation is**

**Regression Analysis: yi versus xi4**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 1 1831.9 1831.90 22.80 0.001

**xi4** 1 1831.9 1831.90 **22.80** 0.001

Error 11 883.9 80.35

Lack-of-Fit 9 746.7 82.96 1.21 0.532

Pure Error 2 137.2 68.61

Total 12 2715.8

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 11 DF in denominator

P( X ≤ x ) x

0.95 4.84434

Decision:-

Here, Fcal>Fin i.e. **22.80** > 4.84434

The 1 st entering variable is xi4..

Therefore , regression equation is

**Regression Analysis: yi versus xi4**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 1 1831.9 1831.90 22.80 0.001

xi4 1 1831.9 1831.90 22.80 0.001

Error 11 883.9 80.35

Lack-of-Fit 9 746.7 82.96 1.21 0.532

Pure Error 2 137.2 68.61

Total 12 2715.8

**egression Equation**

**yi = 117.57 - 0.738 xi4**

**Next entering variable is as follows :-**

Calculating partial F statistics:-

**Regression Analysis: yi versus xi4, xi1**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2641.00 1320.50 176.63 0.000

xi4 1 1190.92 1190.92 159.30 0.000

**xi1** 1 809.10 809.10 **108.22** 0.000

Error 10 74.76 7.48

Total 12 2715.76

**Regression Analysis: yi versus xi4, xi2**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 1846.88 923.44 10.63 0.003

xi4 1 37.46 37.46 0.43 0.526

**xi2** 1 14.99 14.99 **0.17** 0.687

Error 10 868.88 86.89

Total 12 2715.76

**Regression Analysis: yi versus xi4, xi3**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2540.0 1270.01 72.27 0.000

xi4 1 1763.7 1763.66 100.36 0.000

**xi3** 1 708.1 708.13 **40.29** 0.000

Error 10 175.7 17.57

Total 12 2715.8

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 10 DF in denominator

P( X ≤ x ) x

0.95 **4.96460**

The largest partial F-statistic Y on x2i,x1i is **108.22**

decision:-

Here, Fcal>Fin

**108.22** > **4.96460**

**Therefore 2nd entering variable is xi1.**

**Regression Analysis: yi versus xi4, xi1**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2641.00 1320.50 176.63 0.000

xi4 1 1190.92 1190.92 159.30 0.000

**xi1** 1 809.10 809.10 **108.22** 0.000

Error 10 74.76 7.48

Total 12 2715.76

**Regression Equation**

**yi = 103.10 - 0.6140 xi4 + 1.440 xi1**

in stepwise selection method 1st check 1 st entering variable is leaving the model or not .therefore,

Fout=F1,n-p-1 , lowest partial F statistic is :Flow= **108.22**

**Inverse Cumulative Distribution Function**

F distribution with 1 DF in numerator and 10 DF in denominator

P( X ≤ x ) x

0.95 **4.96460**

Decision:-

here, Fout < Flow  i.e **4.96460**< **108.22**

**Here, by using backward elimination method is terminated because smallest partial F statistic is greater than preselected cut of value .**

**Final regression model is :-**

**Regression Analysis: yi versus xi4, xi1**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2641.00 1320.50 176.63 0.000

xi4 1 1190.92 1190.92 159.30 0.000

xi1 1 809.10 809.10 108.22 0.000

Error 10 74.76 7.48

Total 12 2715.76

Model Summary

S R-sq R-sq(adj) R-sq(pred)

2.73427 97.25% 96.70% 95.54%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 103.10 2.12 48.54 0.000

xi4 -0.6140 0.0486 -12.62 0.000 1.06

xi1 1.440 0.138 10.40 0.000 1.06

**Regression Equation**

**yi = 103.10 - 0.6140 xi4 + 1.440 xi1**

Fits and Diagnostics for Unusual Observations

Obs yi Fit Resid Std Resid

8 72.50 77.52 -5.02 -2.06 R

R Large residual

**d)**

**Best Subsets Regression: yi versus xi1, xi2, xi3, xi4**

Response is yi

x x x x

R-Sq R-Sq Mallows i i i i

Vars R-Sq (adj) PRESS (pred) Cp S 1 2 3 4

1 67.5 64.5 1194.2 56.0 138.7 8.9639 X

1 66.6 63.6 1202.1 55.7 142.5 9.0771 X

2 97.9 97.4 93.9 96.5 2.7 2.4063 X X

2 97.2 96.7 121.2 95.5 5.5 2.7343 X X

3 98.2 97.6 85.4 96.9 3.0 2.3087 X X X

3 98.2 97.6 90.0 96.7 3.0 2.3121 X X X

4 98.2 97.4 110.3 95.9 5.0 2.4460 X X X X

In the above table we observed that 3 variable regression model(subset) have low mallow cp statistic & Msres and high Rsq(adj). Therefore, 3 variable regression model (subset)t is best(adequate) among all subsets.