Submitted by-Aditya Gautam

Roll No-12

M.sc. 3rd semester

Date of Assignment-21/10/2020

Date of Submission-31/10/2020

Experiment No-03

Topic- FITTING OF MULTIPLE LINEAR REGRESSION MODEL.

Problem- The following arrangement of the sum of squares and sum of product corrected for mean, each being based on 28 observations.

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array}\begin{pmatrix} x_1 & x_2 & x_3 & y \\ 10.458 & 0.238 & 7.032 & 1.379 \\ & 1.316 & -0.268 & 2.198 \\ & & 16.047 & 0.368 \\ & & & 4.244 \end{pmatrix}$$

1) Fit a regression model of the form
   $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
2) Test for the significance of the multiple correlation coefficient of Y on (X_1, X_2, X_3). Also construct the 95% confidence interval for each of the regression coefficient.

Theory-

(1) Suppose we are to fit a regression model of the form-

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots + \beta_p X_p + \varepsilon$ where Y is the dependent variables, X_i' s are independent variables (i=1,2,3, ………, p) and β_i 's are the regression coefficient (i=1,2,3,…………., p). Let us further suppose that there are 'n' observations under each of the 'p-variable' and Y. Therefore, for the i^{th} observation , we have-

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots\cdots + \beta_p X_{pi} + \varepsilon_i \quad ; i = 1,2,3,,,,,,,,n \qquad ....(1)$$

Here it is assume that-

$$E(\varepsilon_i) = 0$$

$$E(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 \ ; & i \neq j \\ \sigma^2 \ ; & i = j \end{cases} \qquad i, j = 1,2,3,......., n$$

Also, ε_i ~ Normal distribution with mean 0 and variance σ^2

Taking sum over all the 'n' observation on equation (1) and dividing by 'n' we get—

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{n\beta_0}{n} + \beta_1 \frac{1}{n}\sum_{i=1}^{n} X_{1i} + \beta_2 \frac{1}{n}\sum_{i=1}^{n} X_{2i} + \cdots\cdots + \beta_p \frac{1}{n}\sum_{i=1}^{n} X_{pi} + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$$

$$\Rightarrow \bar{Y} = \beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \cdots\cdots + \beta_p \bar{X}_p + \bar{\varepsilon} \qquad \text{---------------------(2)}$$

Subtracting eqn (2) from each of the 'n' equation in (1), we get-

$$\Rightarrow Y_i - \bar{Y} = (\beta_0 - \beta_0) + \beta_1 (X_{1i} - \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2) + \cdots\cdots + \beta_p (X_{pi} - \bar{X}_p) + (\varepsilon_i - \bar{\varepsilon})$$

$$\Rightarrow Y_i = \beta_1 (X_{1i} - \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2) + \cdots\cdots + \beta_p (X_{pi} - \bar{X}_p) + (\varepsilon_i - \bar{\varepsilon})$$

$$\Rightarrow y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots\cdots + \beta_p x_{pi} + \varepsilon_i \qquad ; i = 1,2,3,.......n \qquad \text{----------------(3)}$$

Where x_{ki} denote deviation of the i^{th} observation under k^{th} variable from its mean,

k=1,2,3,………………p and i=1,2,3,……………..n

The 'n' equations in no (3) may be written in the matrix form as-

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

Where,

$$\underset{\sim}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}, \quad \underset{\sim}{X} = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{p1} \\ x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{pn} \end{pmatrix}_{n\times p}, \quad \underset{\sim}{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{p\times 1}, \quad \underset{\sim}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix}_{p\times 1}$$

The least square estimates of $\hat{\beta}$ is given by-

$$\hat{\beta} = (X'X)^{-1} X'\underset{\sim}{Y} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}_{p\times 1}$$

Where,

$$X'X = \begin{pmatrix} \sum_{i=1}^{n} x_{1i}{}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{pi} \\ \sum_{i=1}^{n} x_{2i}x_{1i} & \sum_{i=1}^{n} x_{2i}{}^2 & \cdots & \sum_{i=1}^{n} x_{2i}x_{pi} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{pi}x_{1i} & \sum_{i=1}^{n} x_{pi}x_{pi} & \cdots & \sum_{i=1}^{n} x_{pi}{}^2 \end{pmatrix}, \text{ and } \quad \underset{\sim}{X}'\underset{\sim}{Y} = \begin{pmatrix} \sum_i x_{1i}y_i \\ \sum_i x_{2i}y_i \\ \vdots \\ \sum_i x_{pi}y_i \end{pmatrix}$$

The required multiple linear regression model to be fitted is -

$$Y = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots\cdots + \hat{\beta}_p X_p$$

(2) Here the null hypothesis to be tested is -

H_0: The multiple correlation coefficient of Y on (X_1, X_2, X_3) is zero.

Under H_0 the test statistic is-

$$F = \frac{R^2}{1-R^2}\left(\frac{n-p-1}{p}\right) \sim F(p, n-p-1)$$

Where, R is the multiple correlation coefficient of a variable with the other ‘p’ variables in a random sample of size ‘n’ from a (p+1) variate distribution.

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n} y_i^2 - \hat{\beta}'(X'X)\hat{\beta}}{\sum_{i=1}^{n} y_i^2} \times \frac{(n-1)}{(n-p)} \right]$$

The calculated value is compared with the tabulated value and conclusion are drawn accordingly.

Confidence Interval Contraction-

Suppose in a regression model $\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$, X has full rank ‘p’ and $\underset{\sim}{\varepsilon}$ is distributed

as $N_w(0, \sigma^2 I_n)$. The simultaneous 100(1-α)% confidence interval for β_i is given by -

$$\hat{\beta}_i \pm \left(\sqrt{Va\hat{r}(\hat{\beta}_i)} \times \sqrt{p \times F_{p,n-p}(\alpha)} \right)$$

Where $Va\hat{r}(\hat{\beta}_i)$ is the diagonal element of $\left\lfloor s^2 \left(X'X \right)^{-1} \right\rfloor$ corresponding to β_i and

$s^2 = \dfrac{\underset{\sim}{\hat{\varepsilon}}'\,\underset{\sim}{\hat{\varepsilon}}}{n-p}$ and $F_{p,n-p}(\alpha)$ is the upper 100(1-α)th percentile of the F distribution with p and (n-p) d.f.

The program for finding the solution to the given problem is-

$$X'X = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{pmatrix} x_1 & x_2 & x_3 \\ 10.458 & 0.238 & 7.032 \\ 0.238 & 0.316 & -0.268 \\ 7.03 & -0.268 & 16.047 \end{pmatrix} \quad , \quad \text{and} \quad X'Y = \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \begin{pmatrix} 1.379 \\ 2.198 \\ 0.368 \end{pmatrix}$$

R Program-

```
XX=array(c(10.458,0.238,7.032,0.238,1.316,-0.268,7.032,-0.268,16.047),dim=c(3,3))

XX

XY=array(c(1.379,2.198,0.368),dim=c(3,1))

XY

XX_inv=solve(XX)

B=(XX_inv)%*%(XY)

B

den=4.244

num=den-(t(B)%*%(XX)%*%B)

n=28

p=3

R_sqr=1-((num/den)*((n-1)/(n-p)))

R_sqr

F_cal=( R_sqr/(1- R_sqr))*((n-p-1)/p)

F_cal

F_tab=qf(1-0.05,3,24)

F_tab

EE=num

S_sqr=EE/(n-p)

V=S_sqr[1,1]*(XX_inv)

V

V_B=array(c(sqrt(V[1,1]), sqrt(V[2,2]), sqrt(V[3,3])),dim=c(3,1))
```

```
T2=sqrt(p*qf(0.05,3,25))

LCL=B-(V_B*T2)

LCL

UCL= B+(V_B*T2)

UCL

#We write the following program to get the pair (LCL, UCL) for Beta1, Beta2 and Beta3

CONF_INTRVL=mat.or.vec(3,2)

for (i in 1:3){

CONF_INTRVL[i,1]=c(LCL[i])

CONF_INTRVL[i,2]=c(UCL[i])}

CONF_INTRVL
```

#Result from the R-Programming-

```
#F_cal=57.7116

#F_tab=3.008787

#THE VALUE OF REGRESSION COEFFICIENTS ARE-

#Beta1=0.08521572

#Beta2=1.65750420

#Beta3=0.01327190

#AND CONFIDENCE INTERVAL FOR Beta1, Beta2 AND Beta3 IS GIVEN BY-

#  0.05499460      0.11543684

#  1.58583258      1.72917582

#  -0.01111644      0.03766024
```

Result and Conclusion-

(i) The required multiple linear regression model to be fitted is -

$$Y = 0.08521572X_1 + 1.65750420X_2 + 0.01327190X_3$$

(ii) Since the calculated value of F (i.e. cal F=57.7116) is greater than the tabulated value of F (i.e. tab F=3.008787). Therefore we reject our null hypothesis & we conclude that the multiple correlation coefficient of Y on (X_1, X_2, X_3) is not equal to zero.

The 95% confidence interval for β_i are given below-
For β_1 is = (0.05499460, 0.11543684)
For β_2 is = (1.58583258, 1.72917582)
For β_3 is = (-0.01111644, 0.03766024)