01 December 2017
11:28

SEQUENTIAL SELECTION

It has 3 Methods:

1. Forward Selection
2. Backward Selection
3. Stepwise Selection

Forward Selection :

Let's assume 4 variables $y = x_1, x_2, x_3, x_4$. It adds one variable and its calculates the Extra Sum of Squares and looks at the 'F' value.

Step1:- No Regressor in the Model

Step2:- All possible models with one regressor are considered and F_{calc} for each regressor is computed. The regressor having the highest F_{calc} is added to the model provided.

$$F_{calc} > F_{Tab}(1, \text{Err df})$$

Step3 :-

partial F-statistic are computed for all of the remaining regressors in the presence of previously selected regressors. The one which is yeilding the highest F_{calc} is added to the model i.e. $F_{calc} > F_{\alpha}(1, \text{Err df})$

Step4 :-

Forward Selection terminates when the highest F_{calc} at a particular stage does not exceed $F_{Tab}(1, \text{Errordf})$ (or) when the last regressor is added to the model.

Let us work on example of "TEST-Hald Cement data.csv"

Step1:- No Regressor in the model.

Step2:- Note down the individual values when it is added independently

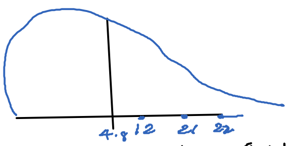
Y = Bo + B1x1 +	F calc (X1)	=	12.602
Y = Bo + B1x2 +	F calc (X2)	=	21.961
Y = Bo + B1x3 +	F calc (X3)	=	4.4034
Y = Bo + B1x4 +	F calc (X4)	=	22.799

Now, the question is which variable I need to add first in model equation ?

So, set the Hypothesis test i.e, $H_0 : \beta_i = 0$
$H_1 : \beta_i \neq 0$

and Draw the picture and check which value is very far to the F_{Tab} value. i.e, Higher value

So, $F_{Tab}(1, 11) = 4.8$



So, $F_{cal(x_4)}$ is very far to F_{Tab} value, so you need to add x_4 first to model first $\Rightarrow$ $Y = \beta_0 + \beta_4 x_4$

Step3 :- Now, I need to check in presence of x_4 How my other variables are going to influence on model.

So, I need to calculate the $F_{calc}(x_1/x_4)$, $F_{cal}(x_2/x_4)$, $F_{cal}(x_3/x_4)$

Case1:-

$$F_{cal}(x_1/x_4) = \frac{SS_{Reg}(x_1, x_4) - SS_{Reg}(x_4)}{MS_{Res}(x_1, x_4)}$$

$$= \frac{2641 - 1831.9}{7.48} = 108.16$$

Case 2:-

$$F_{cal}(x_2/x_4) = \frac{SS_{Reg}(x_2, x_4) - SS_{Reg}(x_4)}{MS_{Res}(x_2, x_4)}$$

$$= \frac{1846.89 - 1831.9}{86.89} = 0.172$$

d.f

Case 3:-

86.89

Case 3:-

$$F_{cal}(x_3 / x_4) = \frac{SS_{Reg}(x_3, x_4) - SS_{Reg}(x_4)}{MS_{Res}(x_3, x_4)}$$

d.f

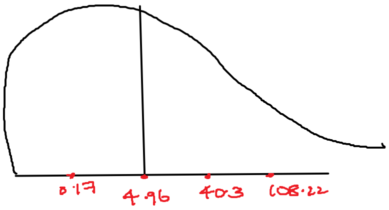
$(2-1) = 1$

$(n-p) = 10$

$$= \frac{2540.2 - 1831.9}{17.57} = 40.3$$

So, Here calculate the F_{Tab} at 5% $\Rightarrow F(1,10) = 4.96$

Let us Draw the picture and see which one's highest



So, the conclusion is among all the data points 108.22 is more highest so x_1 is going to be added in the existing model

i.e, $Y = \beta_0 + \beta_1 x_1 + \beta_4 x_4$

Step 4:- Now, In presence of x_1, x_4 what other variables are behaving we need to calculate.

Case 1:- $F_{cal}(x_2 / x_1, x_4)$

$$= \frac{SS_{Reg}(x_2, x_1, x_4) - SS_{Reg}(x_1, x_4)}{MS_{Res}(x_2, x_1, x_4)}$$

$$= \frac{2667.79 - 2641}{5.33} = 5.026$$

Case 2:- $F_{cal}(x_3 / x_1, x_4)$

$$= \frac{SS_{Reg}(x_3, x_1, x_4) - SS_{Reg}(x_1, x_4)}{MS_{Res}(x_3, x_1, x_4)}$$

$$= \frac{2664.93 - 2641}{5.65} = 4.23$$

So, Let us calculate $F_{Tab}(1,9) = 5.11$

So, Both values are coming under accepted region and < F_{Tab} value.

So, both values are rejected

So, ∴ forward selection method terminates at this stage.

∴ The Final model is $\boxed{\hat{Y} = 103.09 + 1.439 x_1 + (-0.613) x_4}$