STEPWISE SELECTION

It is a combination of Backward Elimination and Forward Selection.

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 Fcalc is added to the model provided. But, no need to check the condition Fcalc > Ftab for 1st Regressor.

Step3: Partial F statistic are computed for all of the remaining regressors in the presence of previously selected regressor.
The one which is yielding the highest Fcalc is added to the model if Fcal > Fthreshold (5).

Step4: All the variables in the model are evaluated with partial F test to see if each one is still significant. At this step, any regressor that is no longer significant is dropped from the model.

Step 5: The Stepwise selection terminates when no other regressor yields a partial F greater than the threshold values of all the regressor in the model remains significant.

Let us work on the "Best-Hald Cement data.csv" and see what results will share us.

Step1: Since there is no regressor in the model, First we need to start with Forward and see the individual Fcalc values.

F-calc(X1) = 12.602
F-calc(X2) = 21.961
F-calc(X3) = 4.4034
F-calc(X4) = 22.799

Highest value is X4. Therefore X4 is added to the model.

Step2: In the presence of X4 we need to check X1, X2, X3.

Fcalc(X1/X4) = 108.22
Fcalc(X2/X4) = 0.1725
Fcalc(X3/X4) = 40.295

Therefore Highest partial Fcalc(X1/X4)= 108.22 is > Threshold (5). So, X1 is added to the model.

Step3: Fcalc(X1/X4) is added to the model, Now I need to check How X4 is behaving in presence of X1 i.e Fcalc(X4/X1)

$$Fcalc(X4/X1) = \frac{SS\ Reg(X4,X1) - SS\ Reg(X4,X1)}{MS\ Res\ (X4,X1)} = \frac{2641 - 1450.1}{7.48} = 159.22$$

We already known Fcalc(X1/X4)= 108.22 and now Fcalc(X4/X1)= 159.22 and Both the values are greater than 5. So both variables to be present in the model.

Step4: Now, we have X4,X1 in the model. We need to check how X2,X3 will behave in presence of X4,X1.

Case1:

$$Fcalc(X2/X4,X1) = \frac{SS\ Reg(X4,X1,X2) - SS\ Reg(X4,X1)}{MS\ Res\ (X4,X1,X2)} = \frac{2667.79 - 2641}{5.33} = 5.026$$

Case2:

$$Fcalc(X3/X4,X1) = \frac{SS\ Reg(X4,X1,X3) - SS\ Reg(X4,X1)}{MS\ Res\ (X4,X1,X3)} = \frac{2664.93 - 2641}{5.65} = 4.235$$

F-calc(X2/X4,X1) > 5 and F-calc(X3/X4,X1) < 5.
Hence X2 is added along with X4, X1 in the model.

Step 5:
From previous step, we understood that F-calc(X2/X4,X1) can be added in the model. But now we need to check back how F-calc(X1/X4,X2) and F-calc(X4/X1,X2) is behaving in presence of X2.

Case1:

$$F\text{-}calc(X1/X4,X2) = \frac{SS\ Reg(X4,X2,X1) - SS\ Reg(X4,X2)}{MS\ Res\ (X4,X2,X1)} = \frac{2667.8 - 1846.89}{5.33} = 154.01$$

Case2:

$$F\text{-}calc(X4/X1,X2) = \frac{SS\ Reg(X1,X2,X4) - SS\ Reg(X1,X2)}{MS\ Res\ (X1,X2,X4)} = \frac{2667.79 - 2657.9}{5.33} = 1.855$$

So, from above F-calc(X1/X4,X2) > 5 and F-calc(X4/X1,X2) < 5.
Hence X4 is dropped and X1 is added along with X2 in the model.
Therefore the model is now as Y=Bo + B1X1 + B2X2.

Step 6:
Now, we need to check in the presence of X1 and X2 how X3 and X4 are behaving in the model.

Case1:

$$F\text{-calc}(X3/X1,X2) = \frac{SS\ Reg(X1,X2,X3) - SS\ Reg(X1,X2)}{MS\ Res\ (X1,X2,X3)} = \frac{2667.65 - 2657.9}{5.35} = 1.82$$

Case2:

$$F\text{-calc}(X4/X1,X2) = \frac{SS\ Reg(X1,X2,X4) - SS\ Reg(X1,X2)}{MS\ Res\ (X1,X2,X4)} = \frac{2667.79 - 2657.9}{5.33} = 1.85$$

So, from above F-calc(X3/X1,X2) < 5 and F-calc(X4/X1,X2) < 5.
Hence X3,X4 are dropped from the model.
Therefore the model remains as Y=Bo + B1X1 + B2X2.

Y-hat = 52.57 + 1.468(X1) + 0.662(X2)

NOTE:
Different Procedures/Algorithms are provided different results.
Backward: X1, X2.
Forward: X1,X4.
Stepwise: X1,X2.

1. From above, two procedures are given same kind of results. So, we can go by that method.
2. Apart from other two the other procedures if the third one is much cheaper compare with other two same procedures. You can go head.
3. If all the three procedures are giving different kind of results. Then there is problem existed in the variable selections called Multicollinearity.