Problem 5 - Model selection

Data Splitting and R code

→ According to the problem , the data is split into into training and validation data sets in 60% and 40% proportions.
→ The validation set has been used to build the model.
→ The following below steps and analysis have been performed using R and some required libraries. The R code has been attached as a file to this folder.

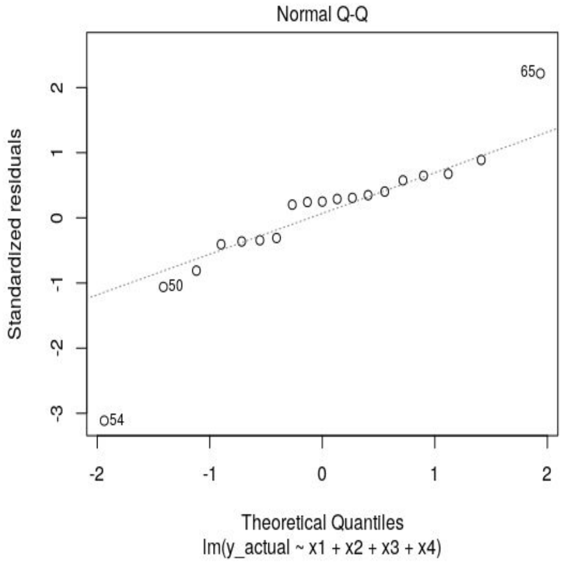
Outlier Analysis

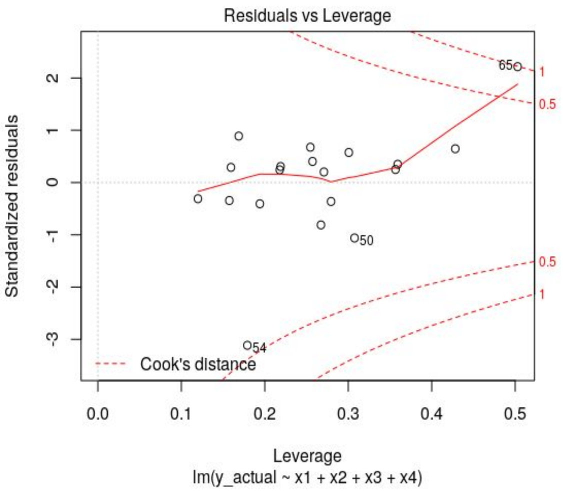
→ For outlier analysis the points which exceed the threshold of Leverage,COOK'S Distance DFBETA , DFFITS and covariance ratio are removed.

→ **Threshold values:** If p = no.of parameters and n = no.of data points

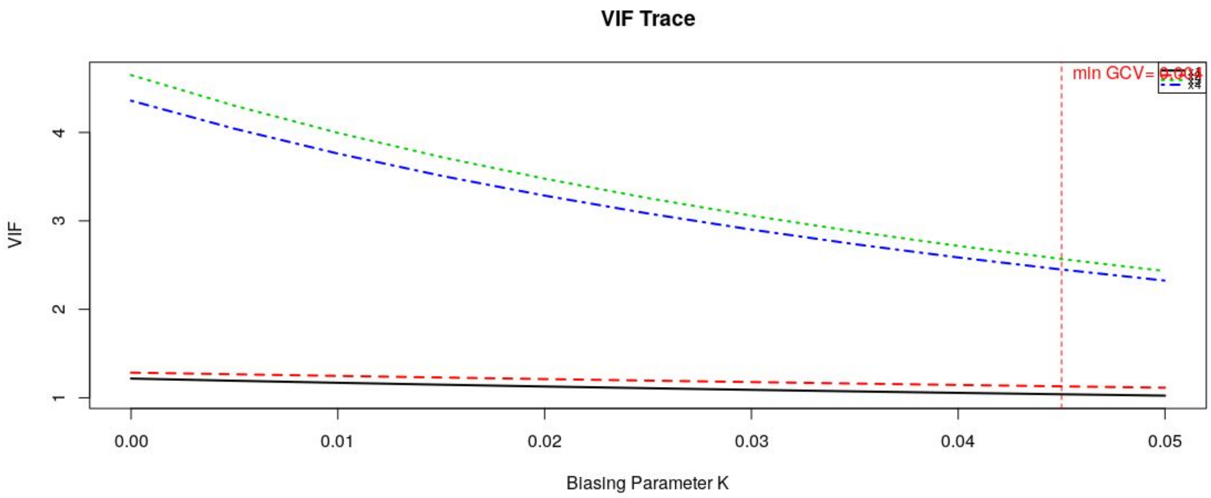
1. Leverage : should be less than **" 2 * (p+1)/n"**
2. Cook'S Distance: should be less than **1**
3. DFBETA : should be less than **2 / sqrt(n**)
4. DFFITS : should be less than **(2 * sqrt(p+1)) / (n-p-1**)

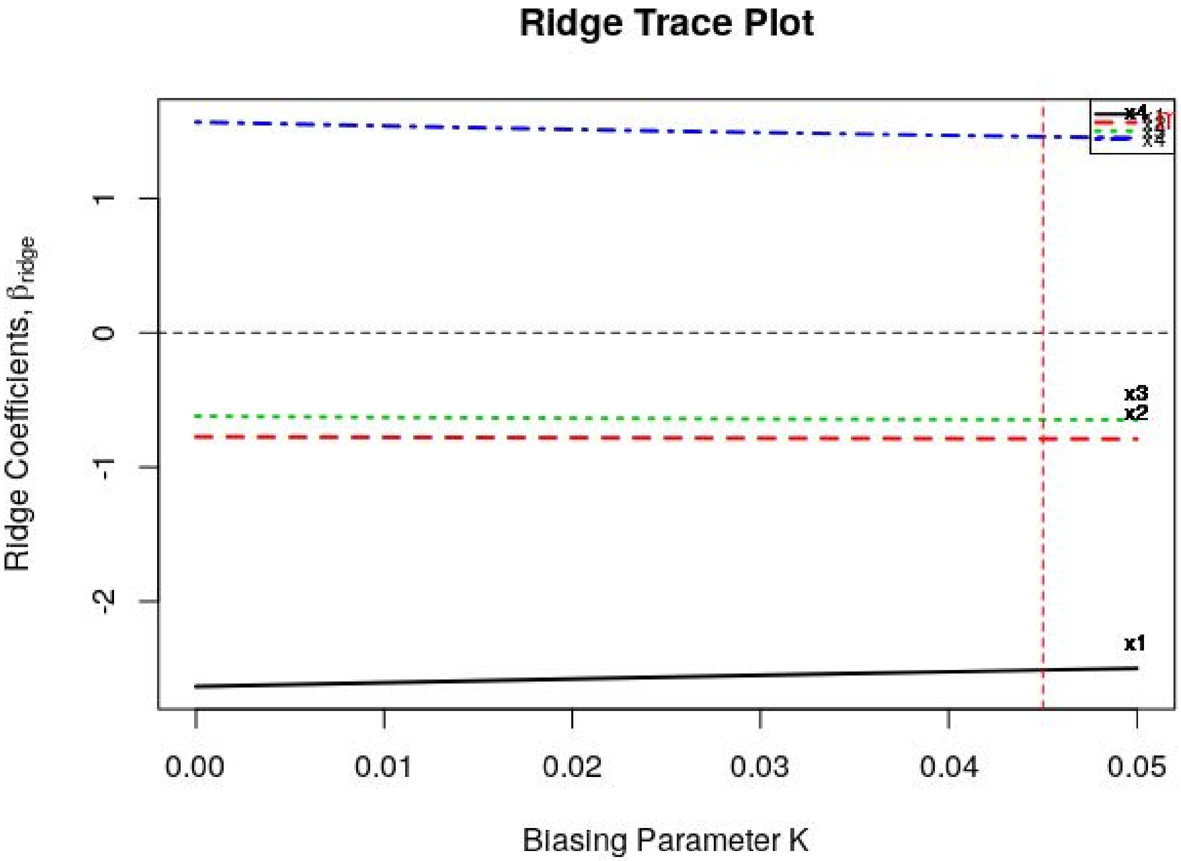
→ Plots and figures



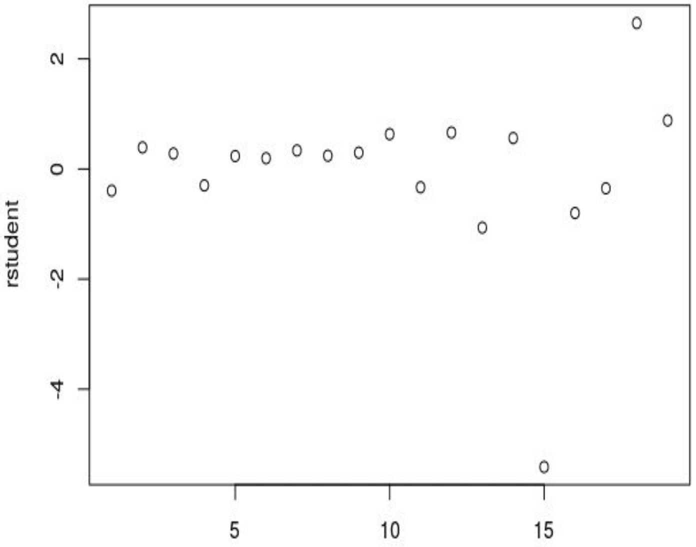


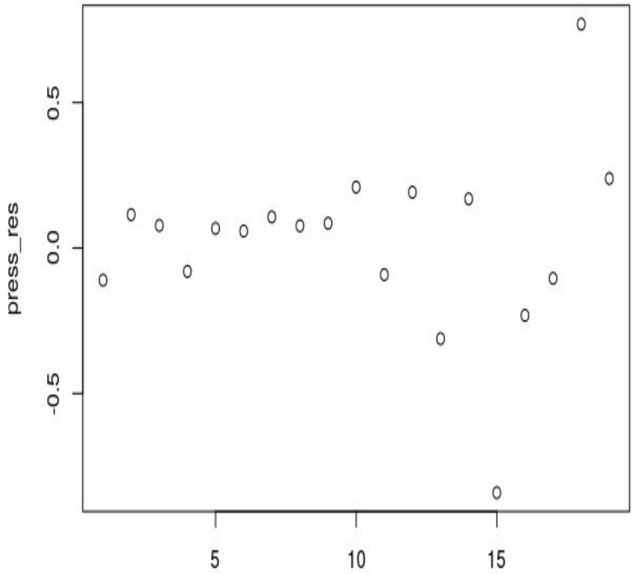
VIF and Ridge trace plot

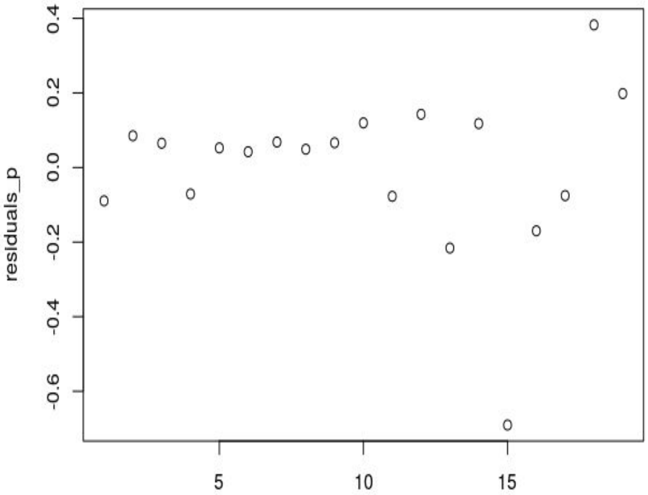


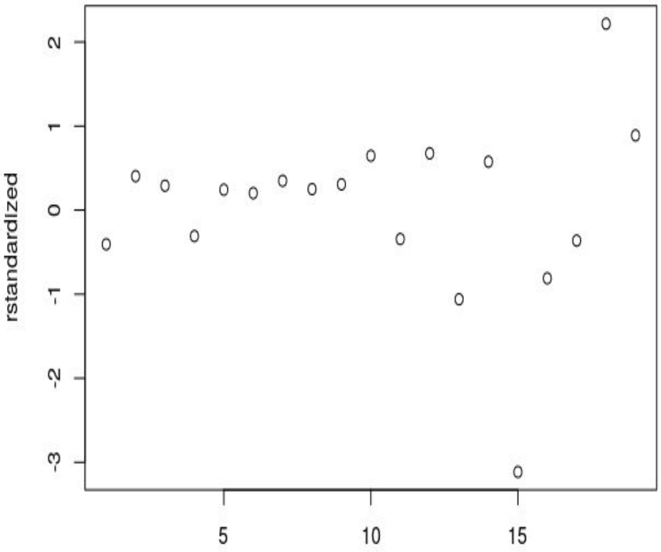


Residual analysis

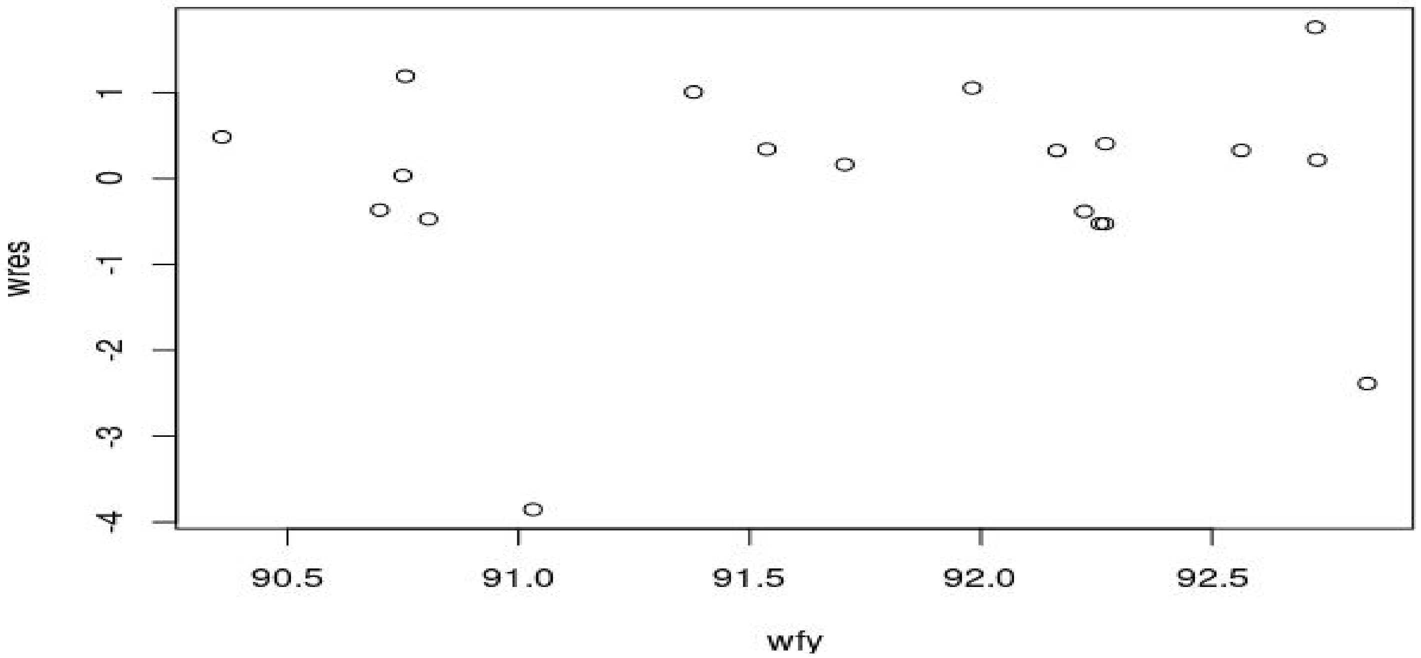




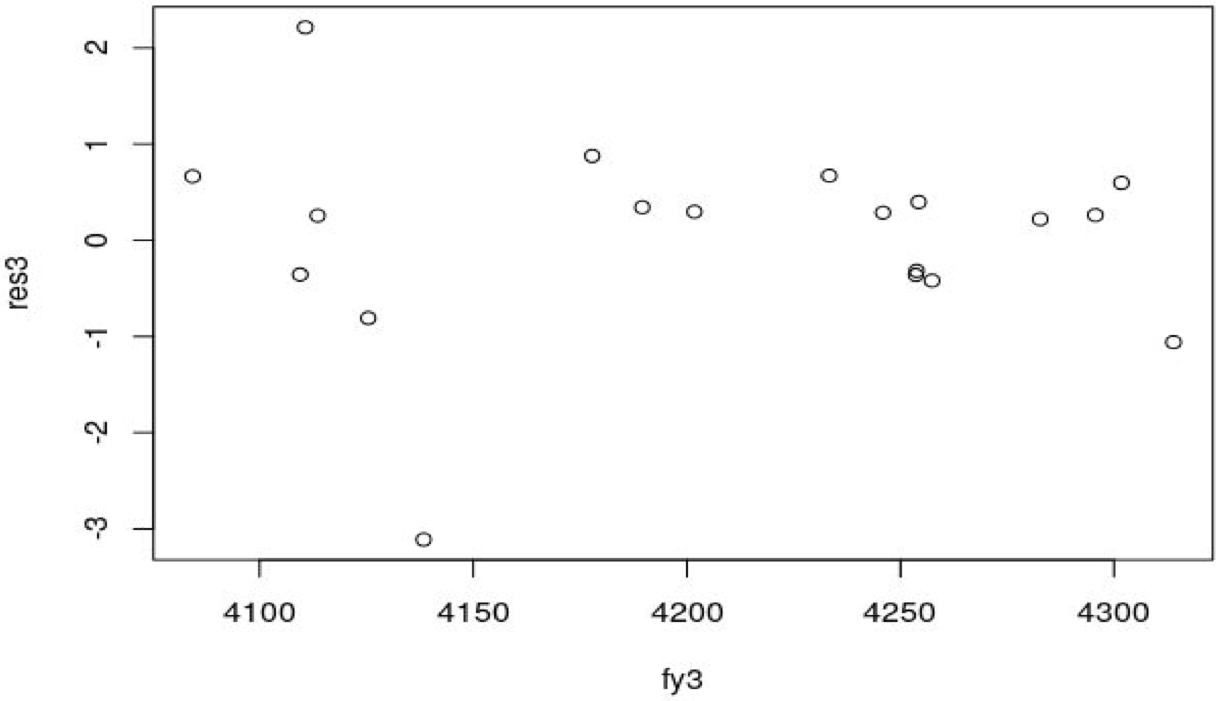




Weighted least square methods



Box-Cox Transformation



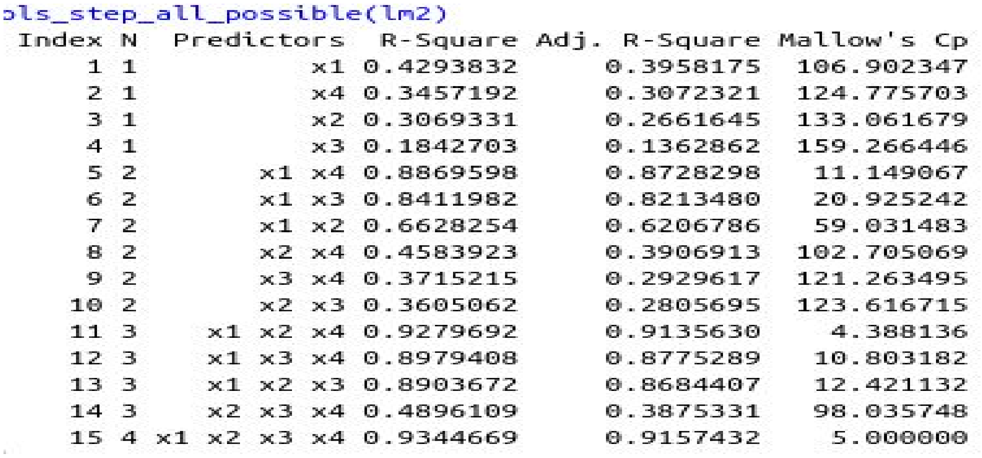
Step - wise model selection

→ The following model selection procedures - Backward elimination, Forward selection, and stepwise are performed on the data after removal of outliers.

→ Based on the values of R^2 , R^2_{adj} , MS_{Res} , AIC, BIC, and Mallow Cp statistics I have chosen the following two models out of many possible models outputted by the stepwise model selection method.

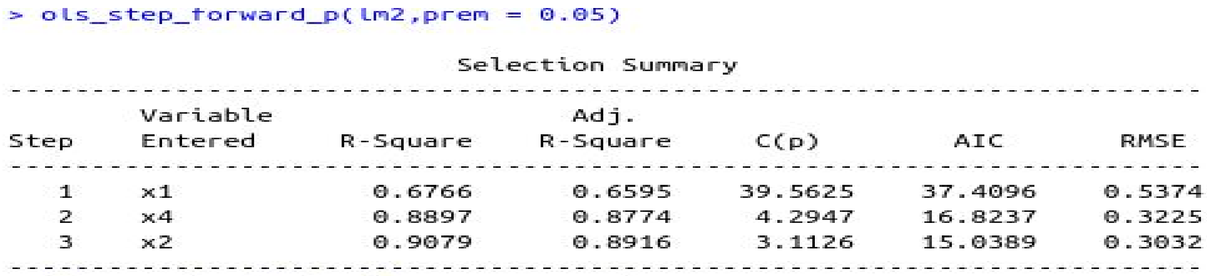
→ The selection of models is such that our MSE and above statistics values are less than other possible models.

→ Possible Models



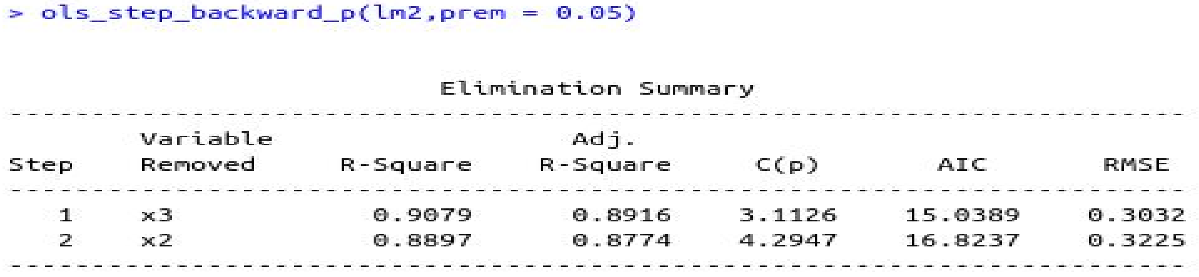
```
ols_step_all_possible(lm2)
 Index N   Predictors  R-Square Adj. R-Square Mallow's Cp
     1 1           x1 0.4293832     0.3958175  106.902347
     2 1           x4 0.3457192     0.3072321  124.775703
     3 1           x2 0.3069331     0.2661645  133.061679
     4 1           x3 0.1842703     0.1362862  159.266446
     5 2        x1 x4 0.8869598     0.8728298   11.149067
     6 2        x1 x3 0.8411982     0.8213480   20.925242
     7 2        x1 x2 0.6628254     0.6206786   59.031483
     8 2        x2 x4 0.4583923     0.3906913  102.705069
     9 2        x3 x4 0.3715215     0.2929617  121.263495
    10 2        x2 x3 0.3605062     0.2805695  123.616715
    11 3     x1 x2 x4 0.9279692     0.9135630    4.388136
    12 3     x1 x3 x4 0.8979408     0.8775289   10.803182
    13 3     x1 x2 x3 0.8903672     0.8684407   12.421132
    14 3     x2 x3 x4 0.4896109     0.3875331   98.035748
    15 4  x1 x2 x3 x4 0.9344669     0.9157432    5.000000
```

→ Backward elimination and Forward selection



```
> ols_step_forward_p(lm2,prem = 0.05)

                              Selection Summary
-----------------------------------------------------------------------
        Variable                   Adj.
Step    Entered      R-Square    R-Square     C(p)       AIC        RMSE
-----------------------------------------------------------------------
   1    x1             0.6766      0.6595    39.5625    37.4096    0.5374
   2    x4             0.8897      0.8774     4.2947    16.8237    0.3225
   3    x2             0.9079      0.8916     3.1126    15.0389    0.3032
-----------------------------------------------------------------------
```

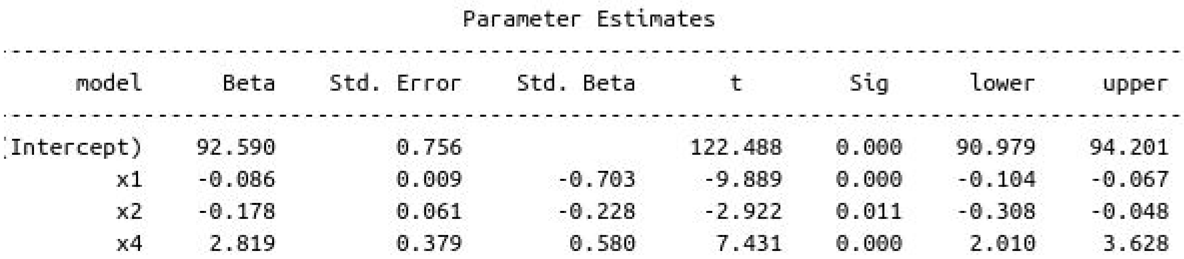


```
> ols_step_backward_p(lm2,prem = 0.05)


                              Elimination Summary
----------------------------------------------------------------------
        Variable                   Adj.
Step    Removed      R-Square    R-Square     C(p)      AIC       RMSE
----------------------------------------------------------------------
   1    x3             0.9079      0.8916    3.1126    15.0389    0.3032
   2    x2             0.8897      0.8774    4.2947    16.8237    0.3225
----------------------------------------------------------------------
```

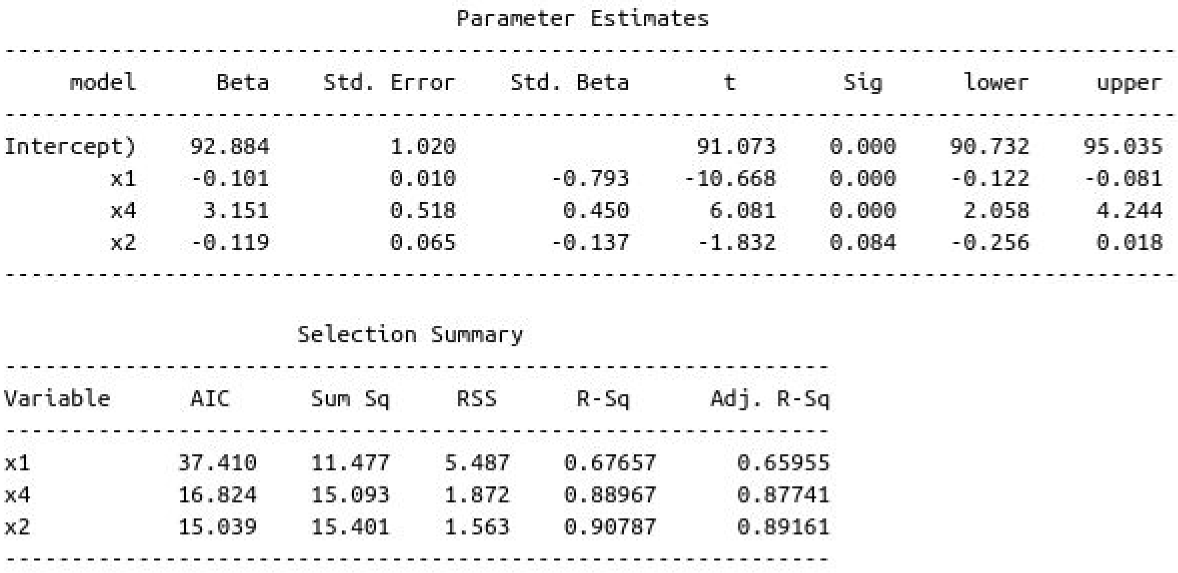
→ **Selected Models**

1. Model-1



```
                              Parameter Estimates
--------------------------------------------------------------------------------
      model        Beta    Std. Error    Std. Beta        t        Sig      lower     upper
--------------------------------------------------------------------------------
(Intercept)      92.590         0.756                 122.488    0.000     90.979    94.201
         x1      -0.086         0.009       -0.703     -9.889    0.000     -0.104    -0.067
         x2      -0.178         0.061       -0.228     -2.922    0.011     -0.308    -0.048
         x4       2.819         0.379        0.580      7.431    0.000      2.010     3.628
```

2. Model - 2



```
                              Parameter Estimates
--------------------------------------------------------------------------------
      model        Beta    Std. Error    Std. Beta        t        Sig      lower     upper
--------------------------------------------------------------------------------
Intercept)       92.884         1.020                  91.073    0.000     90.732    95.035
         x1      -0.101         0.010       -0.793    -10.668    0.000     -0.122    -0.081
         x4       3.151         0.518        0.450      6.081    0.000      2.058     4.244
         x2      -0.119         0.065       -0.137     -1.832    0.084     -0.256     0.018
--------------------------------------------------------------------------------

                     Selection Summary
--------------------------------------------------------
Variable      AIC       Sum Sq     RSS      R-Sq      Adj. R-Sq
--------------------------------------------------------
x1          37.410     11.477    5.487    0.67657      0.65955
x4          16.824     15.093    1.872    0.88967      0.87741
x2          15.039     15.401    1.563    0.90787      0.89161
--------------------------------------------------------
```