MA-4710 Regression Analysis Final Project

Michigan Tech

Final Project -Multilinear Regression Analysis

BY GROUP 18: **Instructor**

Bhavani Chalamalla **Prof. Byung-Jun-Kim**

Haranadh Reddy Ravi

ABSTRACT

This study delves into the comprehensive analysis of hospitals participating in the Study on the Efficacy of Nosocomial Infection Control (SENIC) project, leveraging a dataset comprising 113 samples and 11 variables. The dataset encapsulates crucial aspects of hospital characteristics, from the average length of stay and patient age to infection risk and the presence of medical schools. Employing statistical methodologies, this investigation aims to unravel patterns, relationships, and insights within the dataset, shedding light on the factors contributing to the efficacy of infection control measures in healthcare institutions.

INTRODUCTION

In the context of healthcare, effective infection control is paramount to ensuring the well-being of patients and the overall quality of medical services. This study delves into the intricacies of hospital operations by analyzing the Study on the Efficacy of Nosocomial Infection Control (SENIC) dataset, a compilation of 113 samples featuring 11 variables. These variables encompass crucial aspects of hospital characteristics, ranging from the average length of patient stays and age demographics to infection risk estimations and institutional practices such as routine culturing and X-ray ratios. Hospitals' structural components are also explored, including bed capacity, association with medical schools, and geographic regions. By employing rigorous statistical methodologies, this analysis aims to unravel patterns and relationships within the dataset, shedding light on the nuanced factors that contribute to the efficacy of infection control measures in healthcare institutions. The findings are poised to inform and enhance strategies for mitigating hospital-acquired infections and improving the overall resilience of healthcare systems.

THE GOAL OF THE PROJECT:

The goal of the analysis is to determine which predictor variables in this dataset can help to better understand and predict the length of stay of the patient in the US hospital by building the multiple linear regression model.

Description of the variables in the dataset (in ascending order):

• Length of Stay (Y): The average length of stay of all patients in the hospital (in days).

• Age (X1): The average age of patients (in years).

• Infection Risk (X2): The average estimated probability of acquiring infection in a hospital (in percent).

• Routine Culturing Ratio (X3): The ratio of the number of cultures performed to the number of patients without signs or symptoms of hospital-acquired infection, times 100.

• Routine Chest X-ray Ratio (X4)**:** The ratio of the number of X-rays performed to the number of patients without signs or symptoms of pneumonia, times 100.

• Number of Beds (X5): The average number of beds in the hospital during the study period.

• Medical School (X6): Indicator of whether the hospital is associated with a medical school (1 = Yes, 2 = No).

• Region (X7): Indicator of the geographic region for the hospital (1 = NE, 2 = NC, 3 = S, 4 = W).

• Average daily Census (X8): The average number of patients per day in the hospital during the study period.

• Number of nurses (X9): The average number of full-time equivalent registered and licensed practical nurses during the study period (number of full-time plus one-half the number of part-time).

• Available facilities and services (X10): A percent of 35 potential facilities and services are provided by the hospital.

Numerical Variable: "X1" "X2" "X3" "X4" "X5" "X8" "X9" "X10"

Categorical Variable: **“**X6”, “X7”

Response Variable: “Y”

EXPLORATORY DATA ANALYSIS:

Histograms of Y and Xs:

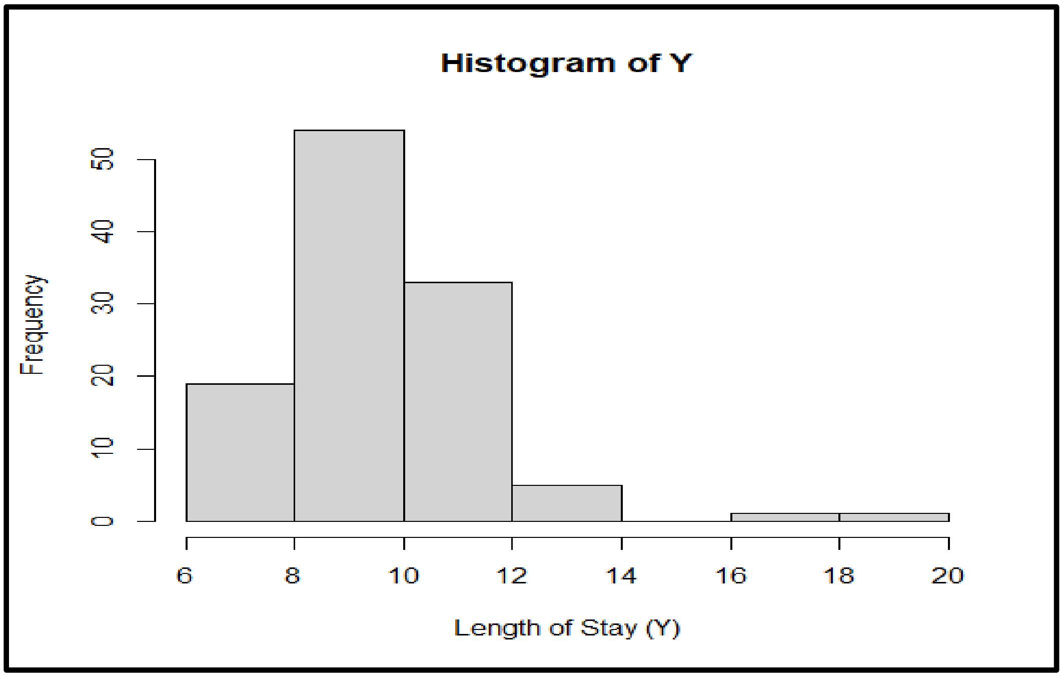
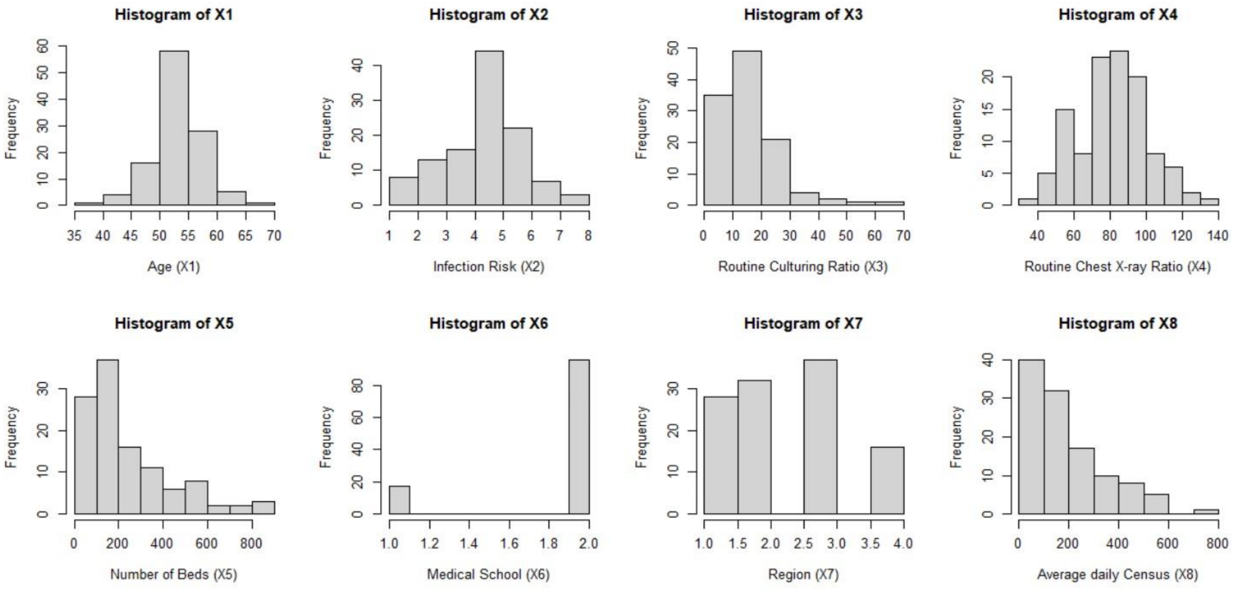


Figure 1. Histogram of the response variable; Length of Stay (Y).

The histogram for our response variable (i.e., Length of Stay), which is displayed in above figure 1 has the right-skewed distribution.



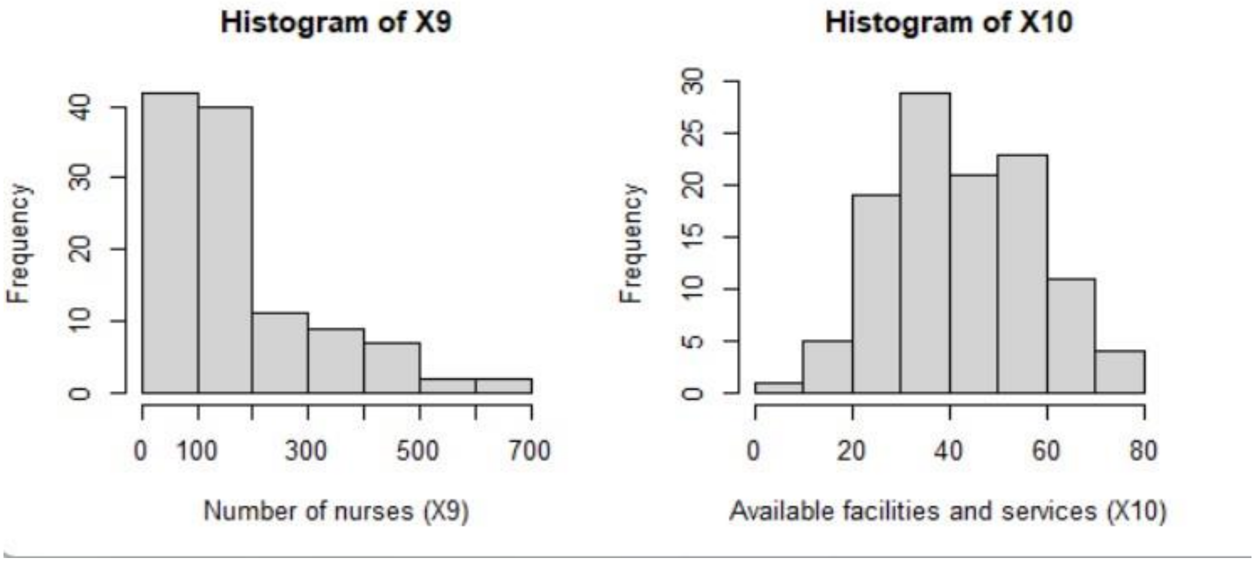
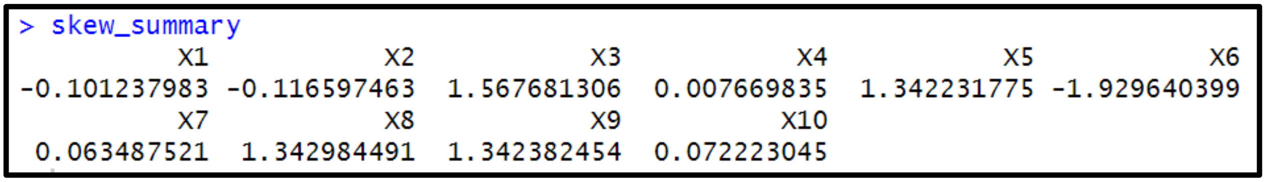


Figure 2. Histograms of the predictor variables (Xs).

.



```
> skew_summary
         X1           X2           X3           X4           X5           X6 
-0.101237983 -0.116597463  1.567681306  0.007669835  1.342231775 -1.929640399 
         X7           X8           X9          X10 
 0.063487521  1.342984491  1.342382454  0.072223045 
```

Figure 3. Skewness of the predictor variables (Xs).

The histogram displayed in Figures 2 & 3 reveals that our predictor variable, Routine Culturing Ratio (X3) **with skewness of 1.567681306**, Number of Beds (X5) **with skewness of 1.342231775**, Average Daily Census (X8) **with skewness 1.342984491**, and Number of Nurses (X9) **with skewness 1.342382454** have a **right-skewed distribution**. On the other hand, Age (X1) **with skewness -0.101237983**, Infection Risk (X2) **with skewness -0.116597463**, Routine Chest X-ray Ratio (X4) **with skewness 0.007669835,** and Available Facilities and Services (X10) **with skewness 0.072223045,** have a **normal distribution**. In addition to these variables, we have two categorical variables, Medical School (X6) and Region (X7).

Boxplots of Y and Xs:

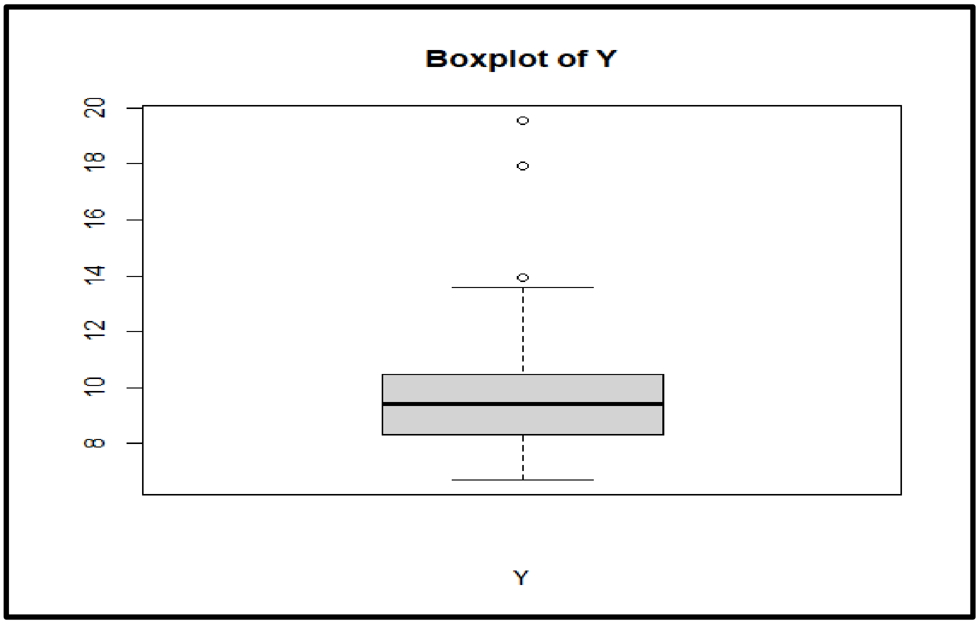


Figure 4. Boxplot of the response variable; Length of Stay (Y).

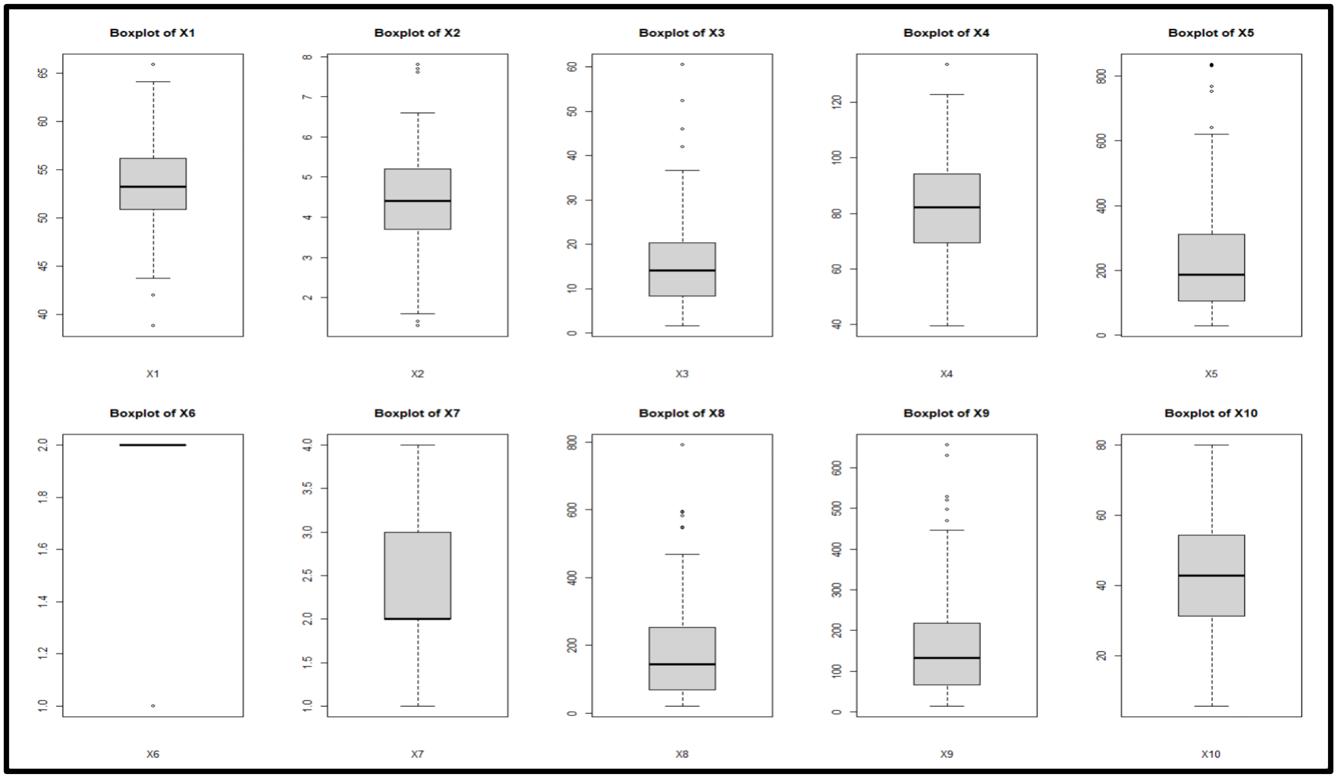


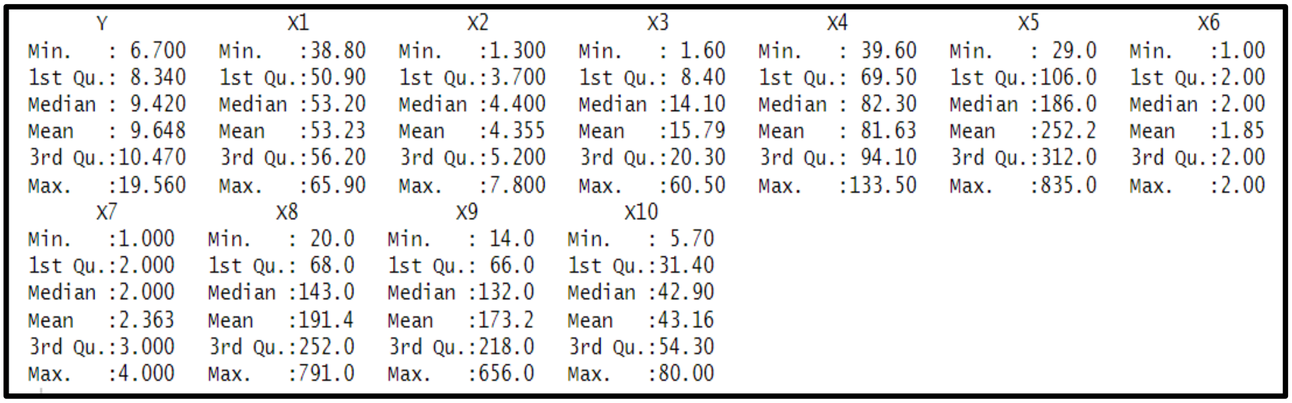
Figure 5. Boxplots of the predictor variables (Xs).

According to Figures 4 & 5, the boxplot of the response variable is not centered since one side of the median is larger than the other, and it contains a few outliers. This suggests that the response variable is not normally distributed.

Figure 4 shows that the boxplots of Routine Culturing Ratio (X3), Number of Beds (X5), Average daily Census (X8), and Number of Nurses (X9) are right skewed with multiple outliers. This indicates that these predictor variables are also not normally distributed. On the other hand, the boxplot for Age (X1), Infection Risk (X2), Routine Chest X-ray Ratio (X4), and Available facilities and services (X10) are centered around the median. Therefore, we can say that they are normally distributed.

We also have two categorical variables, Medical School (X6) and Region (X7). Since they cannot be plotted on a box plot, we cannot make any assumptions about their distribution.

Summary Statistics:-



```
       Y                X1              X2              X3               X4               X5               X6
 Min.   : 6.700   Min.   :38.80   Min.   :1.300   Min.   : 1.60   Min.   : 39.60   Min.   : 29.0   Min.   :1.00
 1st Qu.: 8.340   1st Qu.:50.90   1st Qu.:3.700   1st Qu.: 8.40   1st Qu.: 69.50   1st Qu.:106.0   1st Qu.:2.00
 Median : 9.420   Median :53.20   Median :4.400   Median :14.10   Median : 82.30   Median :186.0   Median :2.00
 Mean   : 9.648   Mean   :53.23   Mean   :4.355   Mean   :15.79   Mean   : 81.63   Mean   :252.2   Mean   :1.85
 3rd Qu.:10.470   3rd Qu.:56.20   3rd Qu.:5.200   3rd Qu.:20.30   3rd Qu.: 94.10   3rd Qu.:312.0   3rd Qu.:2.00
 Max.   :19.560   Max.   :65.90   Max.   :7.800   Max.   :60.50   Max.   :133.50   Max.   :835.0   Max.   :2.00
       X7              X8              X9              X10
 Min.   :1.000   Min.   : 20.0   Min.   : 14.0   Min.   : 5.70
 1st Qu.:2.000   1st Qu.: 68.0   1st Qu.: 66.0   1st Qu.:31.40
 Median :2.000   Median :143.0   Median :132.0   Median :42.90
 Mean   :2.363   Mean   :191.4   Mean   :173.2   Mean   :43.16
 3rd Qu.:3.000   3rd Qu.:252.0   3rd Qu.:218.0   3rd Qu.:54.30
 Max.   :4.000   Max.   :791.0   Max.   :656.0   Max.   :80.00
```

Figure 6. Summary Statistics of the data

It appears that the predictor variable Age (X1) does not provide much information about children or elderly people, as the minimum and maximum ages are between 38 and 66. Additionally, the Average Census (X8) has a wide range of values, from 20.0 to 791.0. The difference between the median and mean suggests that there may be outliers in X8. Furthermore, the Number of Beds (X5), Number of Nurses (X9), and Available Facilities and Services (X10) also have a wide range of values, indicating the potential presence of outliers.

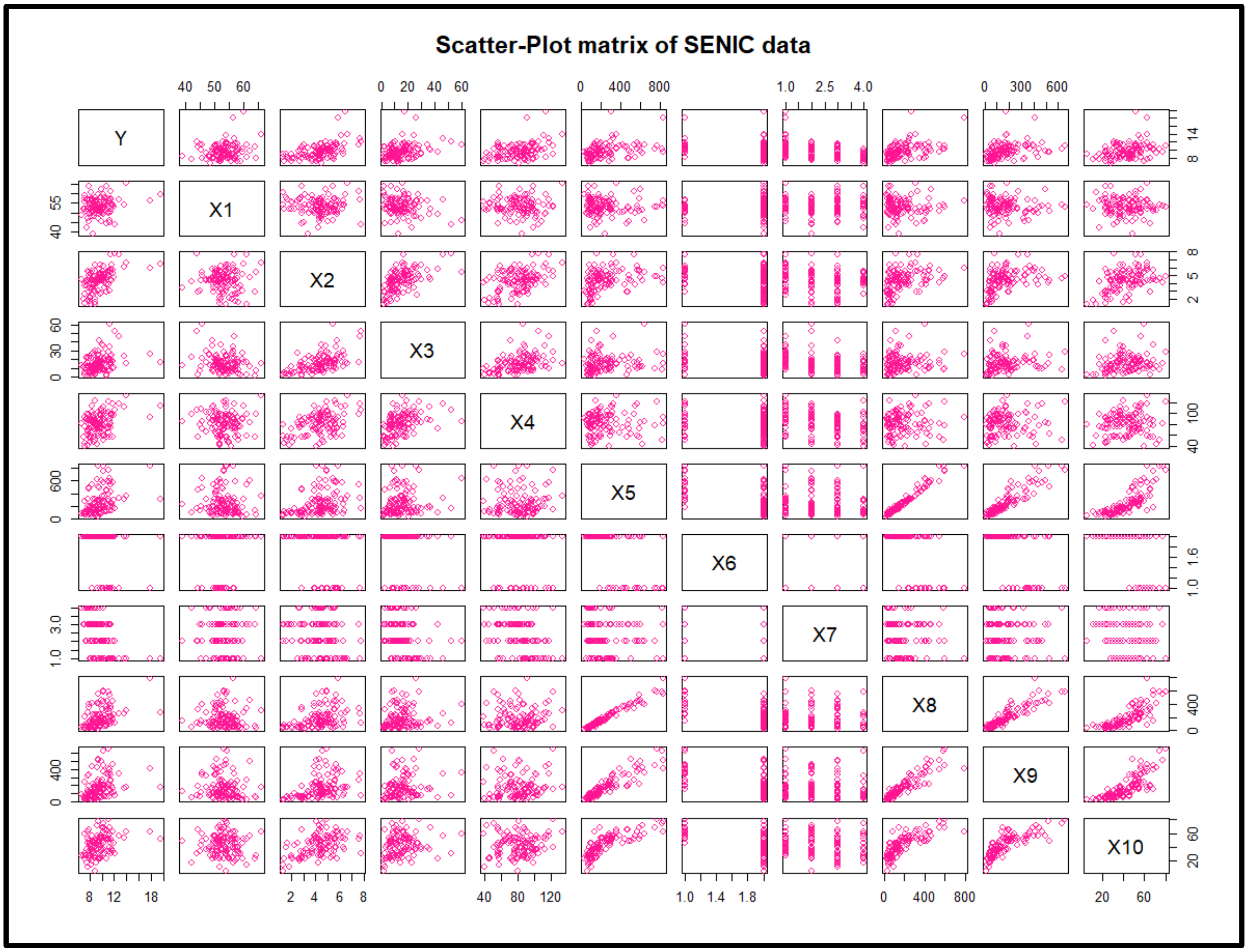


Figure 7. Scatter Plot Matrix

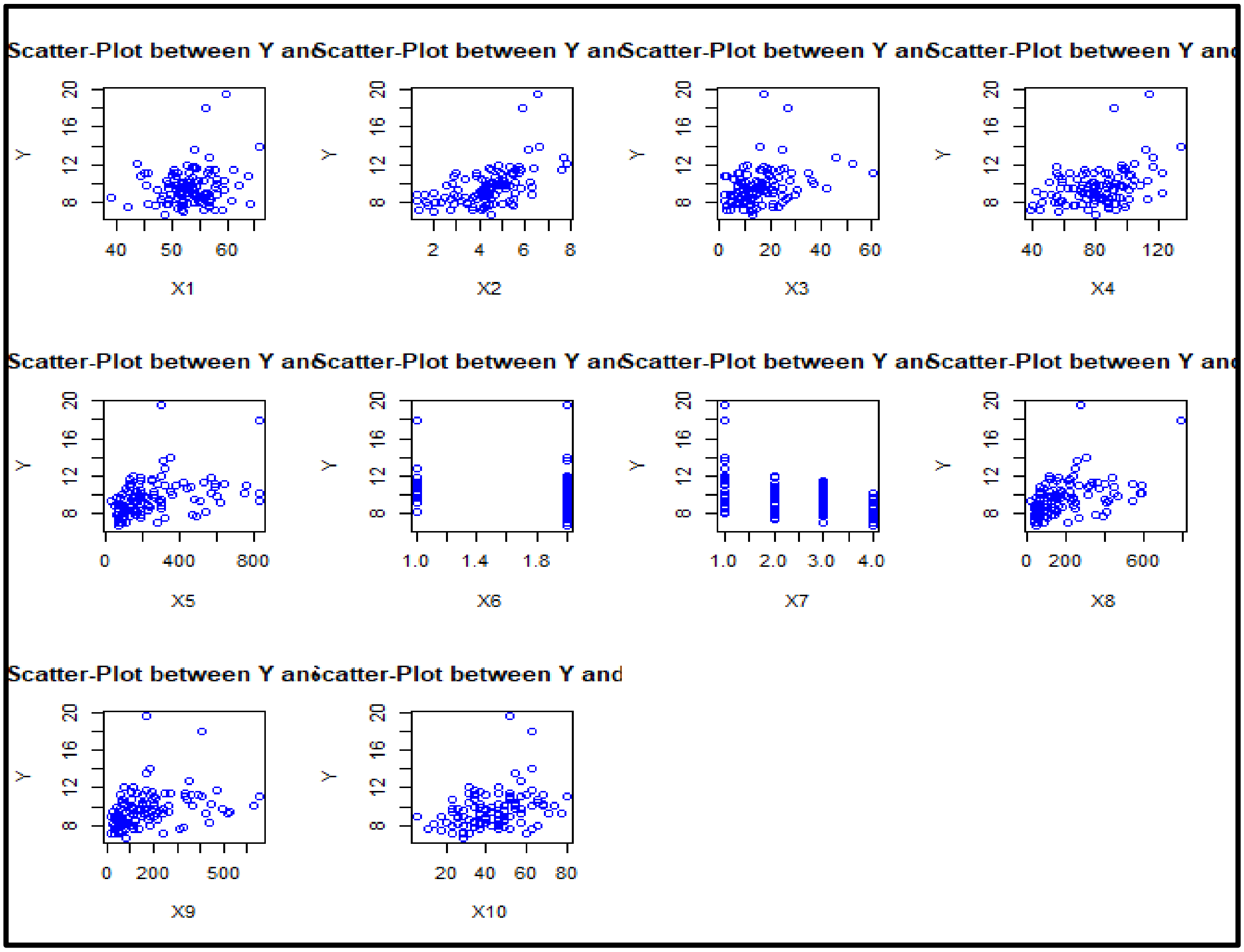


Figure 8. Scatter Plot Matrix of the response variable against the predictor variables.

We can conclude from the Scatter plot (Figures 7&8) that all predictor variables have a kind of linear relationship with the response variable, except for the two categorical variables, Medical School (X6) and Region (X7).

Added variable plots:

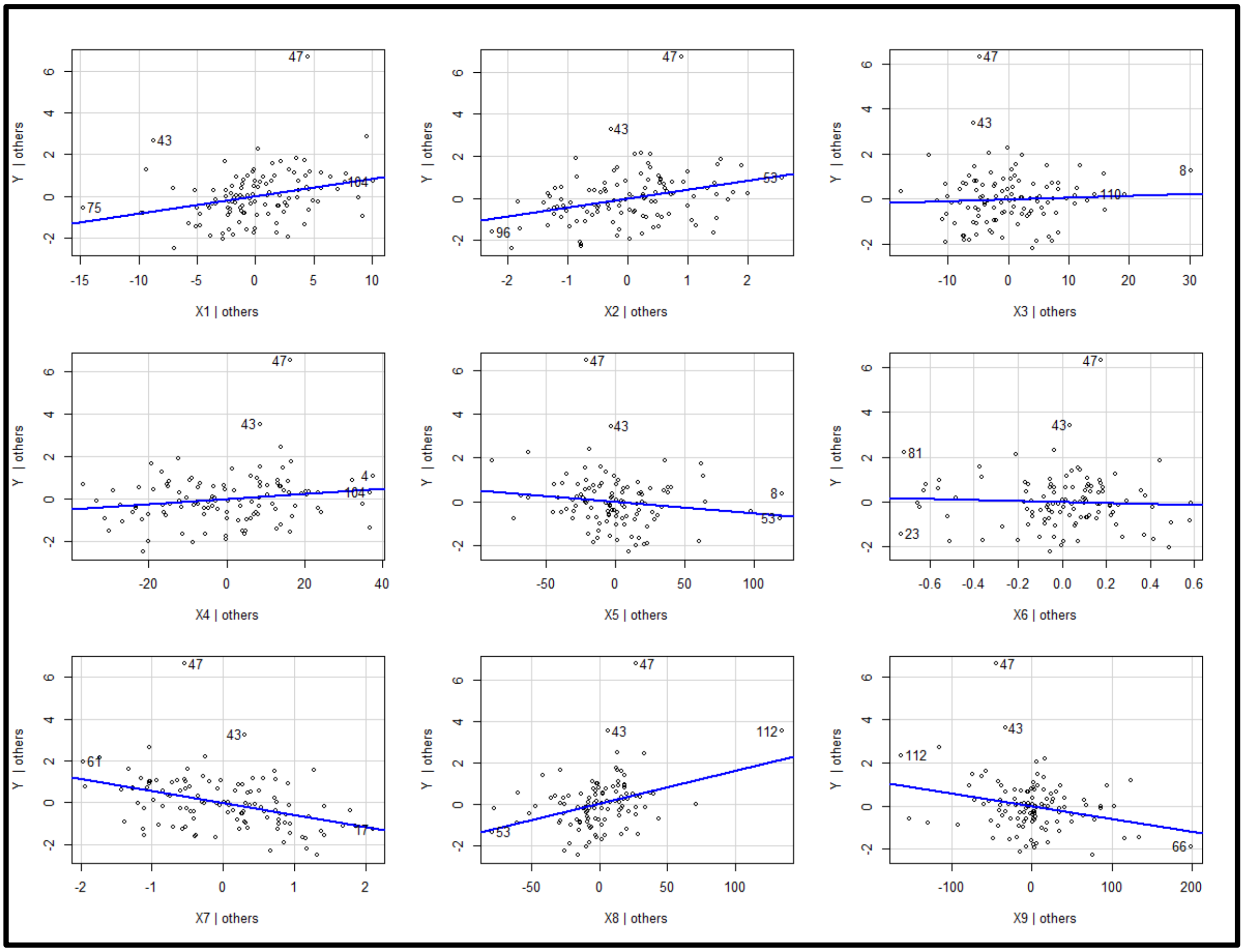
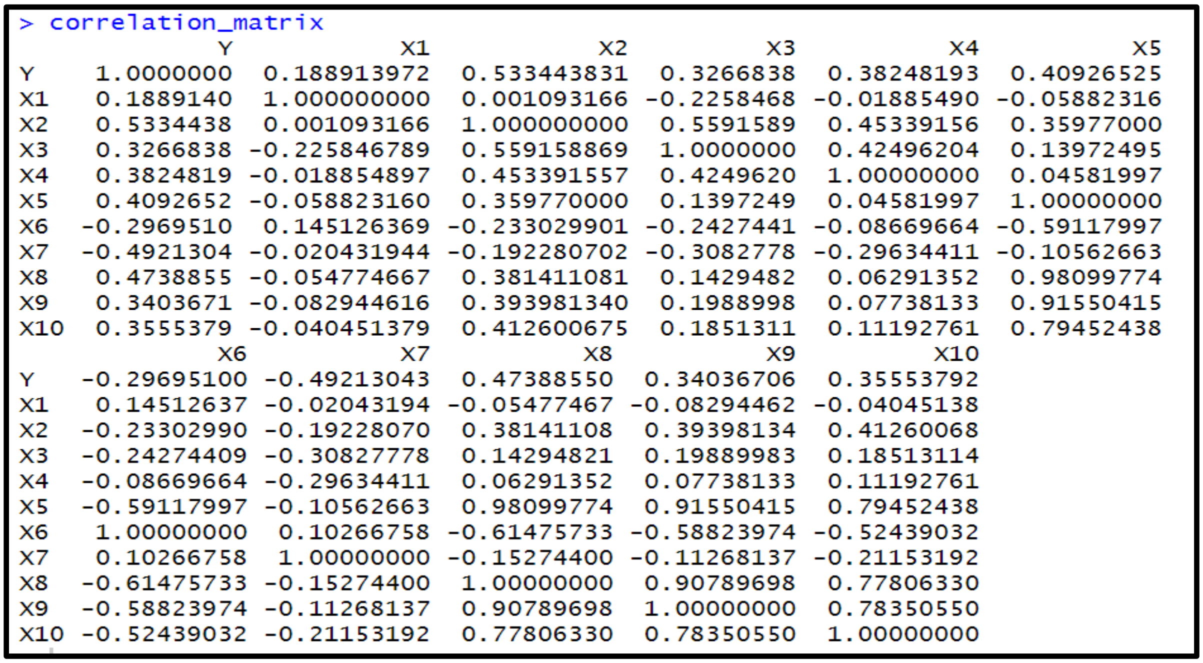


Figure 9. Added-Variable Plots of the response variable against the predictor variables.

The relationship regression lines between the predictor and response variables are provided by this, based on the Added-Variable Plots (Figure 9). Regression analysis reveals that the variables with the highest impact on the response variable are Infection Risk (X2), Average Daily Census (X8), and Routine Culturing Ratio (X3), Medical School (X6). The variables with the lowest impact are others.

Correlation Matrix:



```
> correlation_matrix
            Y           X1          X2         X3          X4          X5
Y    1.0000000  0.188913972  0.533443831  0.3266838  0.38248193  0.40926525
X1   0.1889140  1.000000000  0.001093166 -0.2258468 -0.01885490 -0.05882316
X2   0.5334438  0.001093166  1.000000000  0.5591589  0.45339156  0.35977000
X3   0.3266838 -0.225846789  0.559158869  1.0000000  0.42496204  0.13972495
X4   0.3824819 -0.018854897  0.453391557  0.4249620  1.00000000  0.04581997
X5   0.4092652 -0.058823160  0.359770000  0.1397249  0.04581997  1.00000000
X6  -0.2969510  0.145126369 -0.233029901 -0.2427441 -0.08669664 -0.59117997
X7  -0.4921304 -0.020431944 -0.192280702 -0.3082778 -0.29634411 -0.10562663
X8   0.4738855 -0.054774667  0.381411081  0.1429482  0.06291352  0.98099774
X9   0.3403671 -0.082944616  0.393981340  0.1988998  0.07738133  0.91550415
X10  0.3555379 -0.040451379  0.412600675  0.1851311  0.11192761  0.79452438
            X6          X7          X8          X9         X10
Y   -0.29695100 -0.49213043  0.47388550  0.34036706  0.35553792
X1   0.14512637 -0.02043194 -0.05477467 -0.08294462 -0.04045138
X2  -0.23302990 -0.19228070  0.38141108  0.39398134  0.41260068
X3  -0.24274409 -0.30827778  0.14294821  0.19889983  0.18513114
X4  -0.08669664 -0.29634411  0.06291352  0.07738133  0.11192761
X5  -0.59117997 -0.10562663  0.98099774  0.91550415  0.79452438
X6   1.00000000  0.10266758 -0.61475733 -0.58823974 -0.52439032
X7   0.10266758  1.00000000 -0.15274400 -0.11268137 -0.21153192
X8  -0.61475733 -0.15274400  1.00000000  0.90789698  0.77806330
X9  -0.58823974 -0.11268137  0.90789698  1.00000000  0.78350550
X10 -0.52439032 -0.21153192  0.77806330  0.78350550  1.00000000
```

Figure 10. Correlation matrix of the response variable against the predictor variables.0.9

Correlation plot:

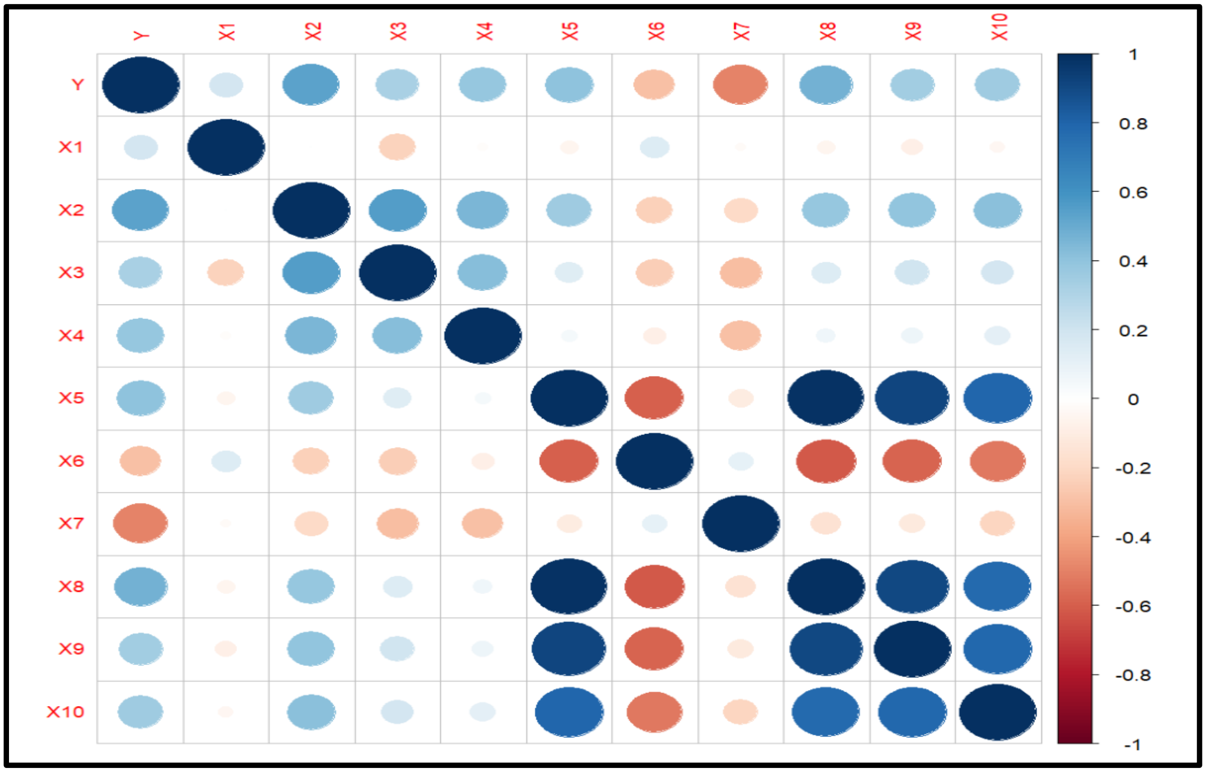


Figure 11. Correlation plot of the response variable against the predictor variables.

The number of beds (X5) has a strong correlation with X8(0.98099774) , X9(0.91550415), and X10(0.79452438), as shown by the correlation matrix and plot (Figures 10&11). A strong correlation has also been observed between X8, the average daily Census, and X9(0.90789698) and X10(0.77806330). In addition, X9—the number of nurses—correlates strongly with X10(0.78350550).

2. MODEL/METHODS

2.1 Model Fitting

Given our dataset with a response variable Y (Length of stay) and predictor variables □1 to □10, a potential multiple linear regression model could be formulated as follows:

□ = □0 + □1□1 + □2□2 + □3□3 + □4□4 + □5□5 + □6□6 + □7□7 + □8□8 + □9□9 + □10□10 + □

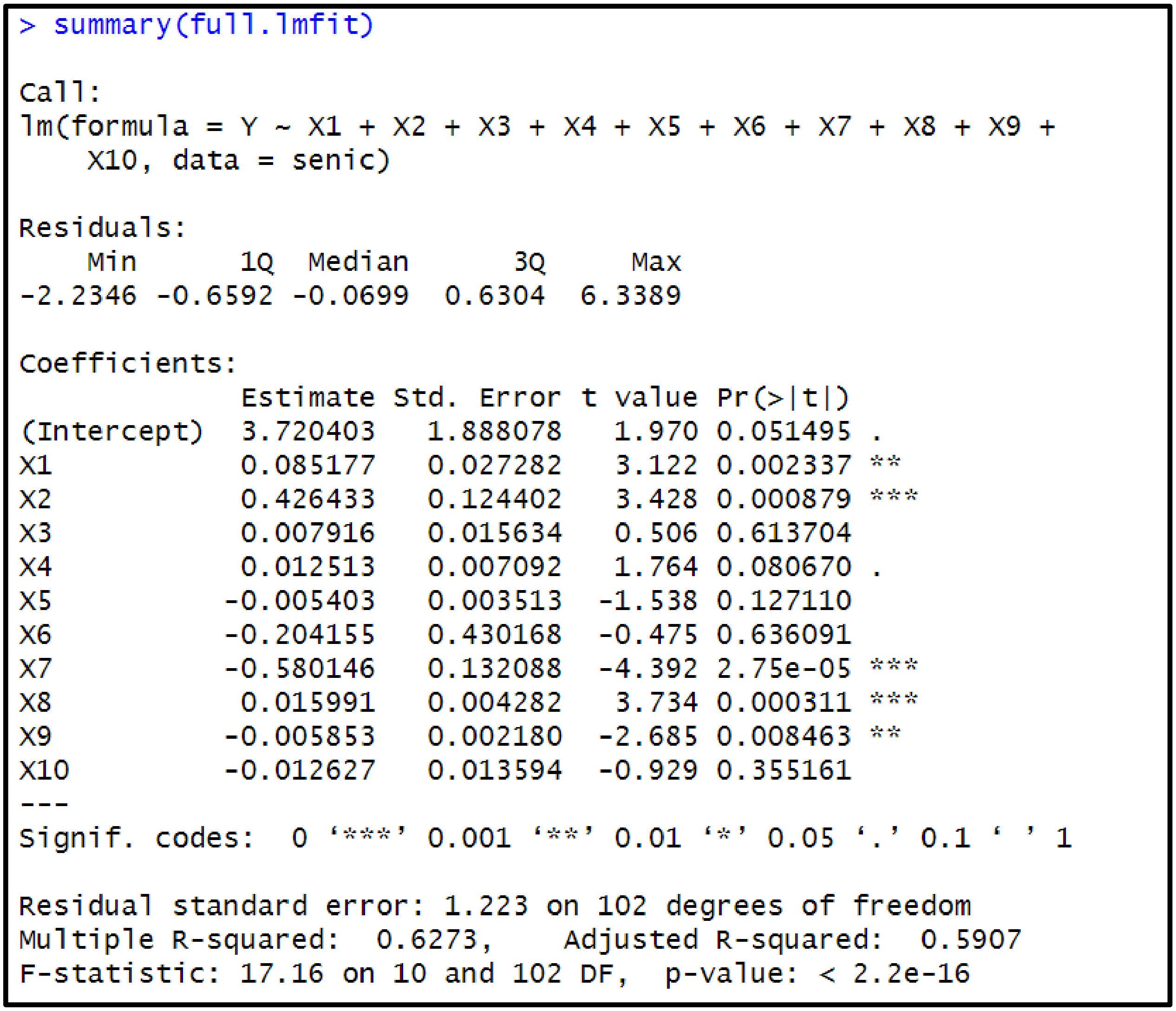
Where:

□□ is the y-intercept of the regression line.

□□ □□ □**10** are the coefficients for each predictor variable, representing the change in the response variable for a one-unit change in the predictor, all else being equal.

ϵ is the error term, representing the residual effect unexplained by the predictors.

We will now fit the full model including all predictor variables using the ‘lm’ function in R. This model will serve as a baseline for comparison.



```
> summary(full.lmfit)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + 
    X10, data = senic)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.2346 -0.6592 -0.0699  0.6304  6.3389 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.720403   1.888078   1.970 0.051495 .  
X1           0.085177   0.027282   3.122 0.002337 ** 
X2           0.426433   0.124402   3.428 0.000879 ***
X3           0.007916   0.015634   0.506 0.613704    
X4           0.012513   0.007092   1.764 0.080670 .  
X5          -0.005403   0.003513  -1.538 0.127110    
X6          -0.204155   0.430168  -0.475 0.636091    
X7          -0.580146   0.132088  -4.392 2.75e-05 ***
X8           0.015991   0.004282   3.734 0.000311 ***
X9          -0.005853   0.002180  -2.685 0.008463 ** 
X10         -0.012627   0.013594  -0.929 0.355161    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.223 on 102 degrees of freedom
Multiple R-squared:  0.6273,	Adjusted R-squared:  0.5907 
F-statistic: 17.16 on 10 and 102 DF,  p-value: < 2.2e-16
```

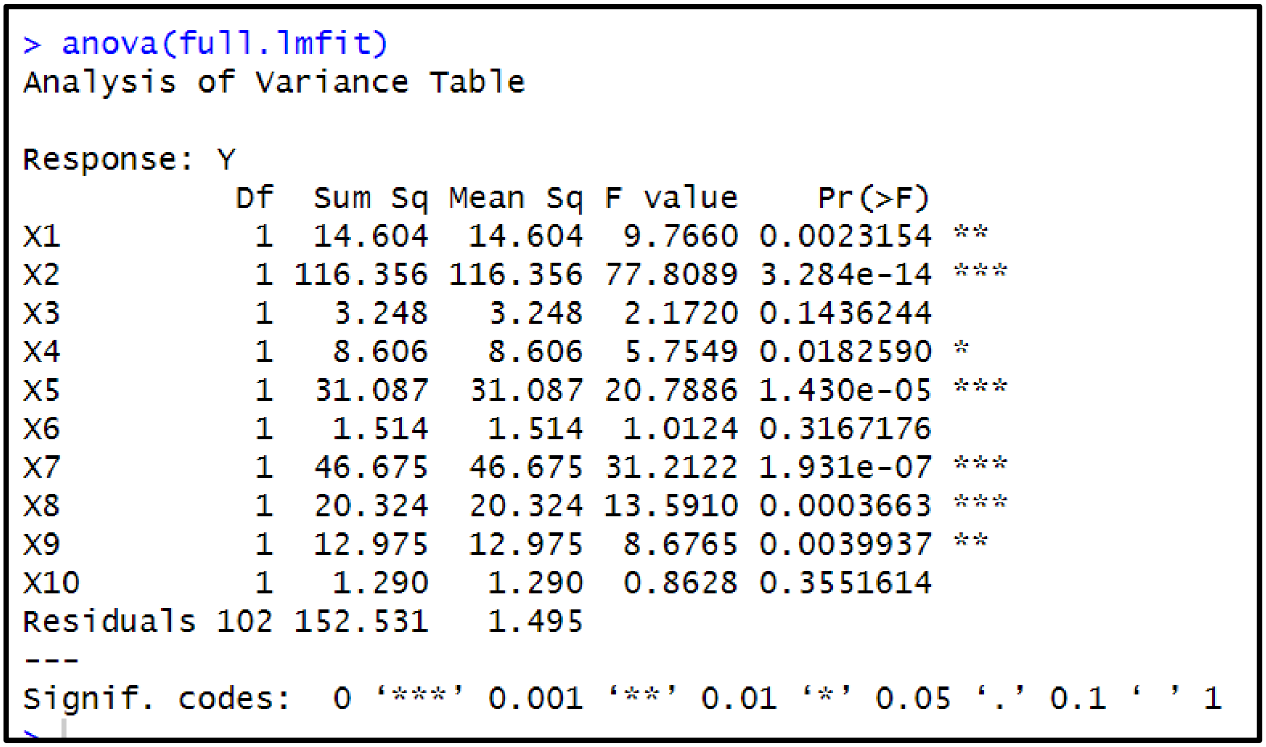
Figure 12: Full Model

The linear regression model was applied to predict the response variable (Y) based on ten predictor variables (X1 to X10). The linear regression model was fitted with response variable Y and predictors X1 through X10. Among the predictors, X2, X7, and X8 showed statistically significant positive effects on Y, while X5 and X9 had significant negative effects. However, variables X3, X6, and X10 were not statistically significant ($p > 0.05$) and can be removed from the model. The adjusted R-squared was 0.5907, indicating that the model explained approximately 59.07% of the variability in Y. The p-value of X5 was very close to 0.1 so I had to investigate it further.

So now the regression equation from the above data is:

$$Y = 3.720 + 0.085X_1 + 0.426X_2 + 0.0079X_3 + 0.0125X_4 - 0.0054X_5 - 0.204X_6 - 0.580X_7 + 0.016X_8 - 0.0059X_9 - 0.0126X_{10}$$

Now we will test the significance of the model through the ANOVA test.



```
> anova(full.lmfit)
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)    
X1          1  14.604  14.604  9.7660 0.0023154 ** 
X2          1 116.356 116.356 77.8089 3.284e-14 ***
X3          1   3.248   3.248  2.1720 0.1436244    
X4          1   8.606   8.606  5.7549 0.0182590 *  
X5          1  31.087  31.087 20.7886 1.430e-05 ***
X6          1   1.514   1.514  1.0124 0.3167176    
X7          1  46.675  46.675 31.2122 1.931e-07 ***
X8          1  20.324  20.324 13.5910 0.0003663 ***
X9          1  12.975  12.975  8.6765 0.0039937 ** 
X10         1   1.290   1.290  0.8628 0.3551614    
Residuals 102 152.531   1.495                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Fig 13: ANNOVA MODEL

The ANOVA table indicates a highly significant F-statistic of 17.16 (p < 2.2e-16), confirming the model's overall explanatory power. Significant predictors include X1, X2, X4, X5, X7, X8, and X9, with positive impacts on Y. Notably, X2 has the highest impact (F = 77.81, p < 3.284e-14), while X5 has a negative effect. In contrast, X3, X6, and X10 do not significantly contribute to Y. The residual mean square is 1.495, representing unexplained variability. Model refinement is suggested, focusing on removing non-significant predictors and validating assumptions like normality and homoscedasticity.

We can see that the model is significant, but some of the individual predictors are not significant. Running the best subset and stepwise regression on this full model results in the following. First, we will look at the best subset for each number of predictor variables selected based on the highest adjusted R2.

3.2 Model Selection/ Step-wise Regression

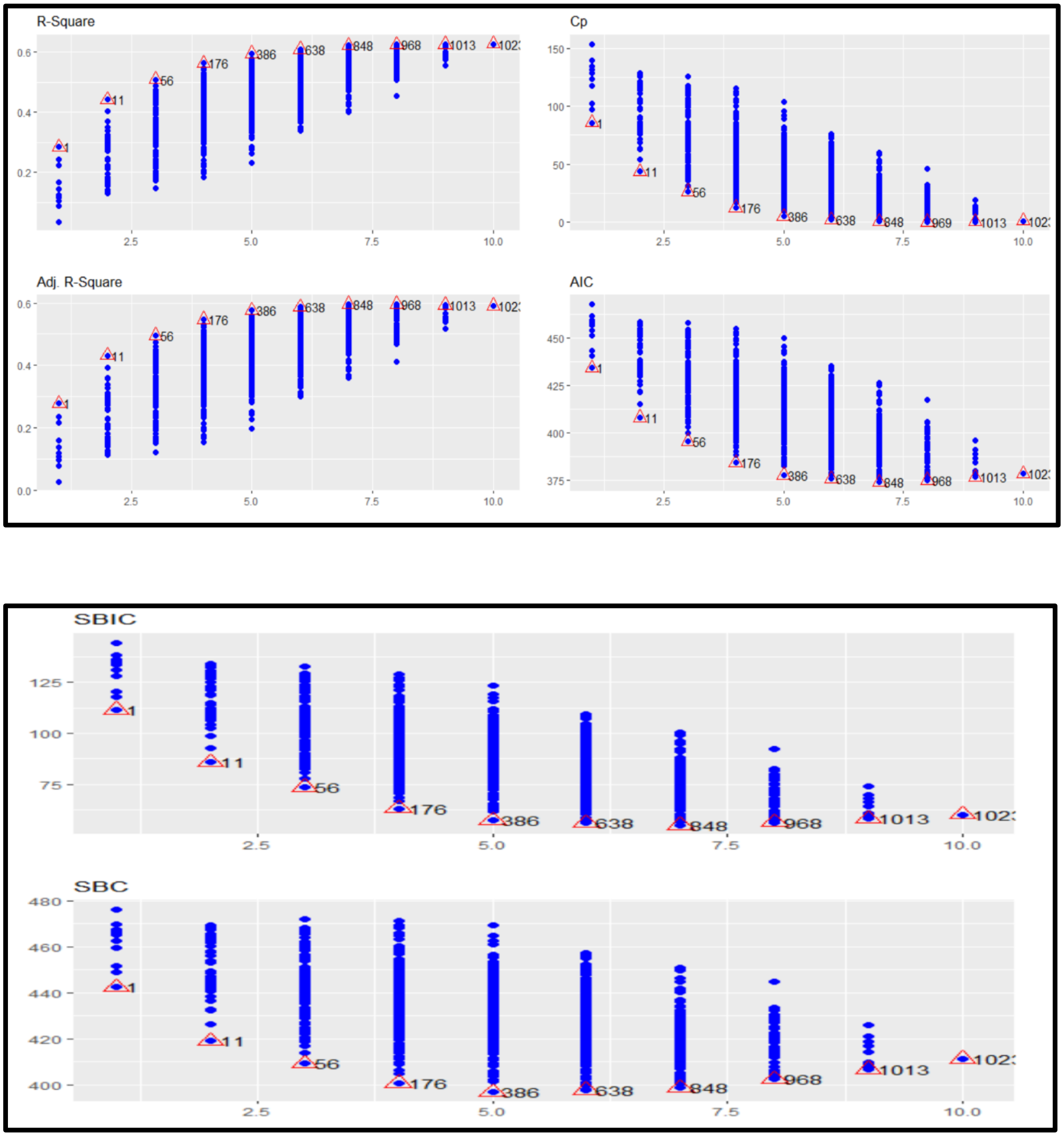
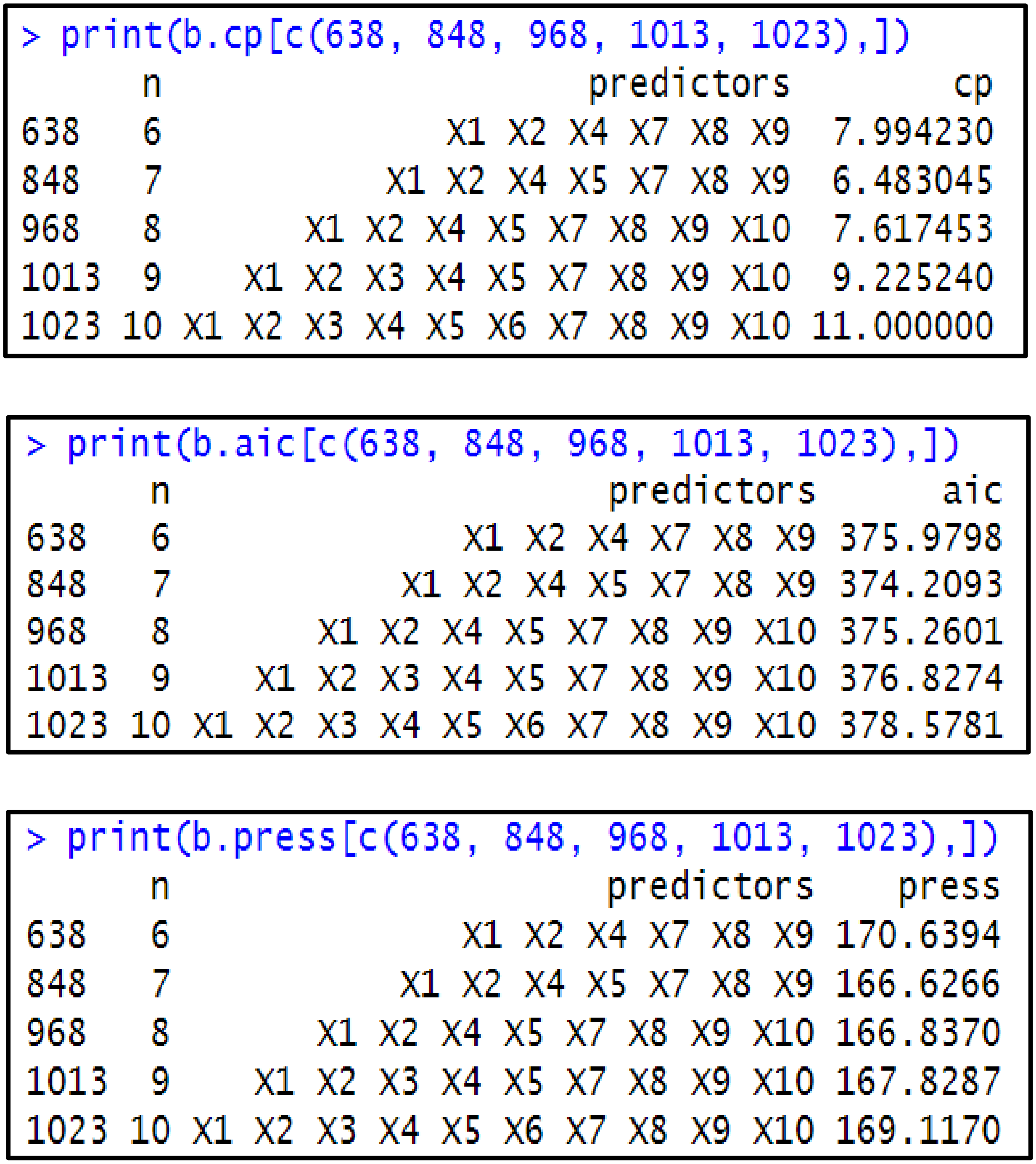


Figure 14: R square, Adjusted □^□, Mallows' Cp, AIC, BIC, SBIC, SBC

The goal of these criteria is to find a model that has the best trade-off between explaining the data and not becoming overly complex. Overly complex models may fit the current data well but can fail to generalize to new data. These criteria help to identify a model that is expected to have the best predictive performance on data. Comparing the above plot we can come up with the following subset of the full model. We can see that all our procedures agree on a model.



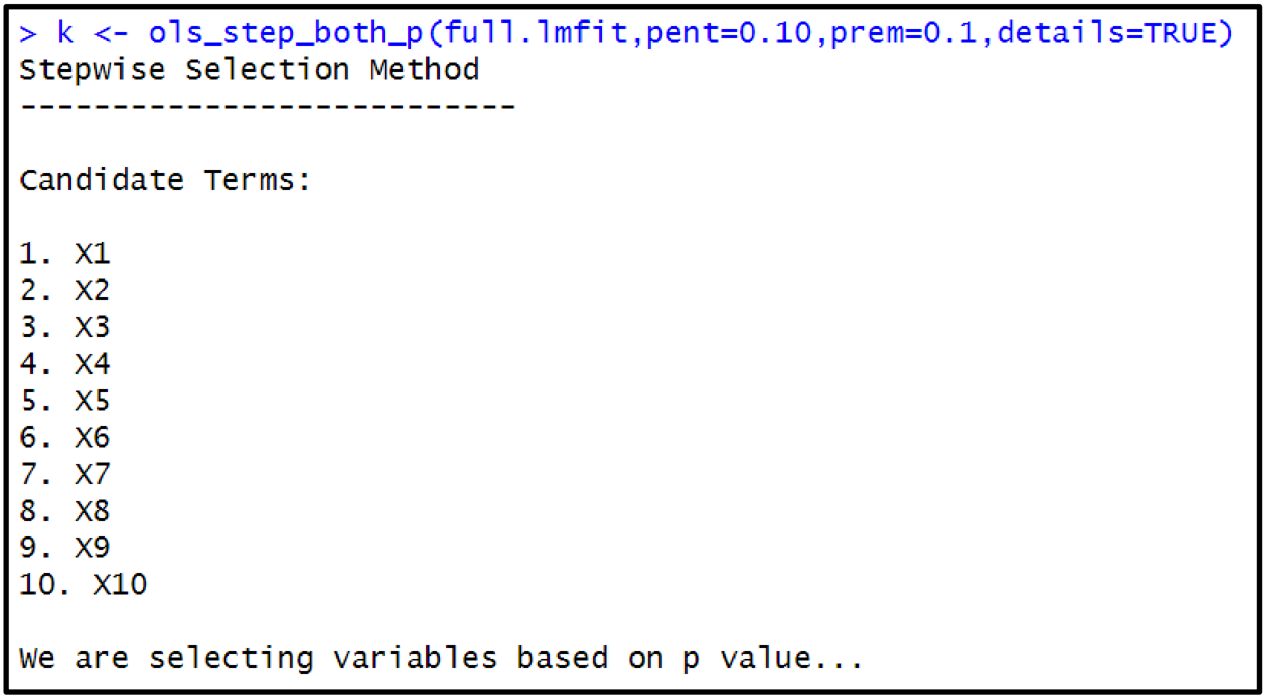
```
> print(b.cp[c(638, 848, 968, 1013, 1023),])
      n                          predictors        cp
638   6                   X1 X2 X4 X7 X8 X9  7.994230
848   7                X1 X2 X4 X5 X7 X8 X9  6.483045
968   8            X1 X2 X4 X5 X7 X8 X9 X10  7.617453
1013  9         X1 X2 X3 X4 X5 X7 X8 X9 X10  9.225240
1023 10 X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 11.000000
```

```
> print(b.aic[c(638, 848, 968, 1013, 1023),])
      n                          predictors      aic
638   6                   X1 X2 X4 X7 X8 X9 375.9798
848   7                X1 X2 X4 X5 X7 X8 X9 374.2093
968   8            X1 X2 X4 X5 X7 X8 X9 X10 375.2601
1013  9         X1 X2 X3 X4 X5 X7 X8 X9 X10 376.8274
1023 10 X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 378.5781
```

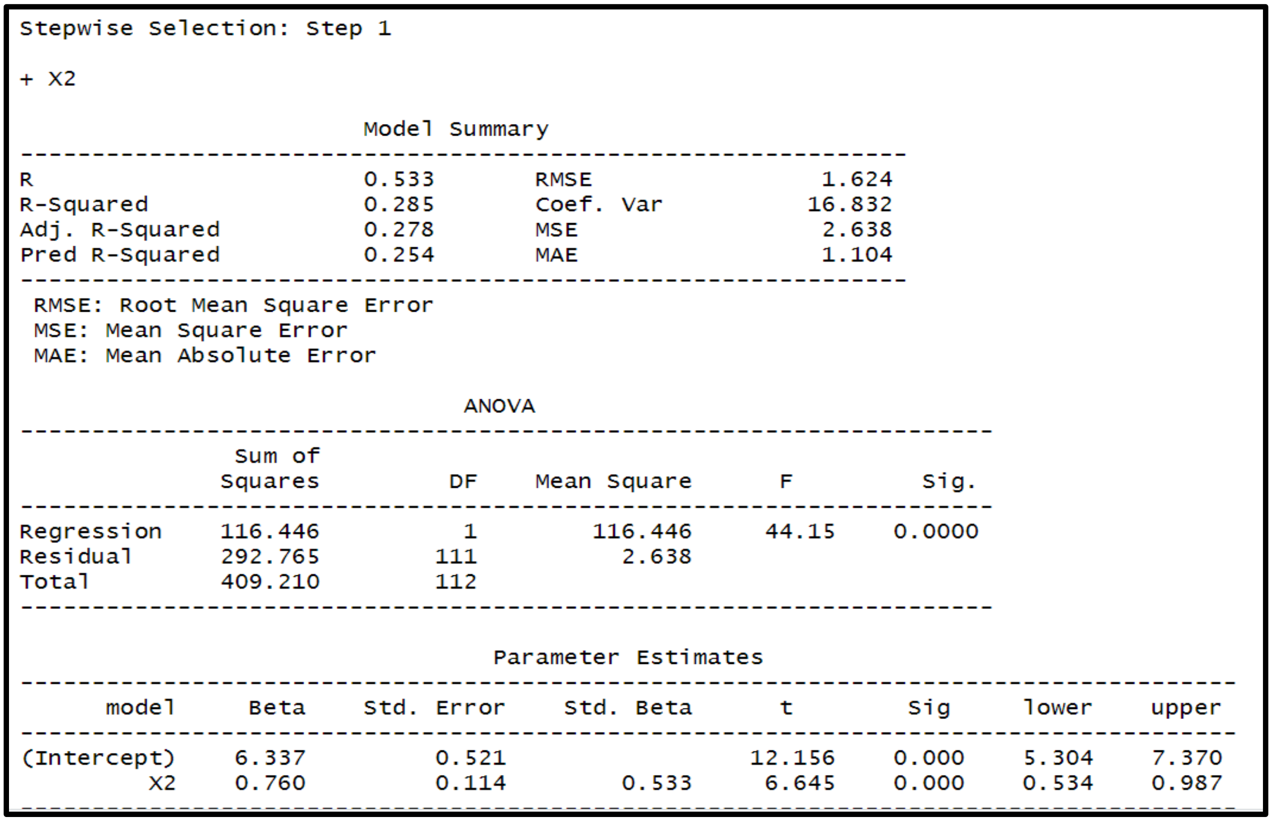
```
> print(b.press[c(638, 848, 968, 1013, 1023),])
      n                          predictors    press
638   6                   X1 X2 X4 X7 X8 X9 170.6394
848   7                X1 X2 X4 X5 X7 X8 X9 166.6266
968   8            X1 X2 X4 X5 X7 X8 X9 X10 166.8370
1013  9         X1 X2 X3 X4 X5 X7 X8 X9 X10 167.8287
1023 10 X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 169.1170
```

Figure 15: The subset of the full model.

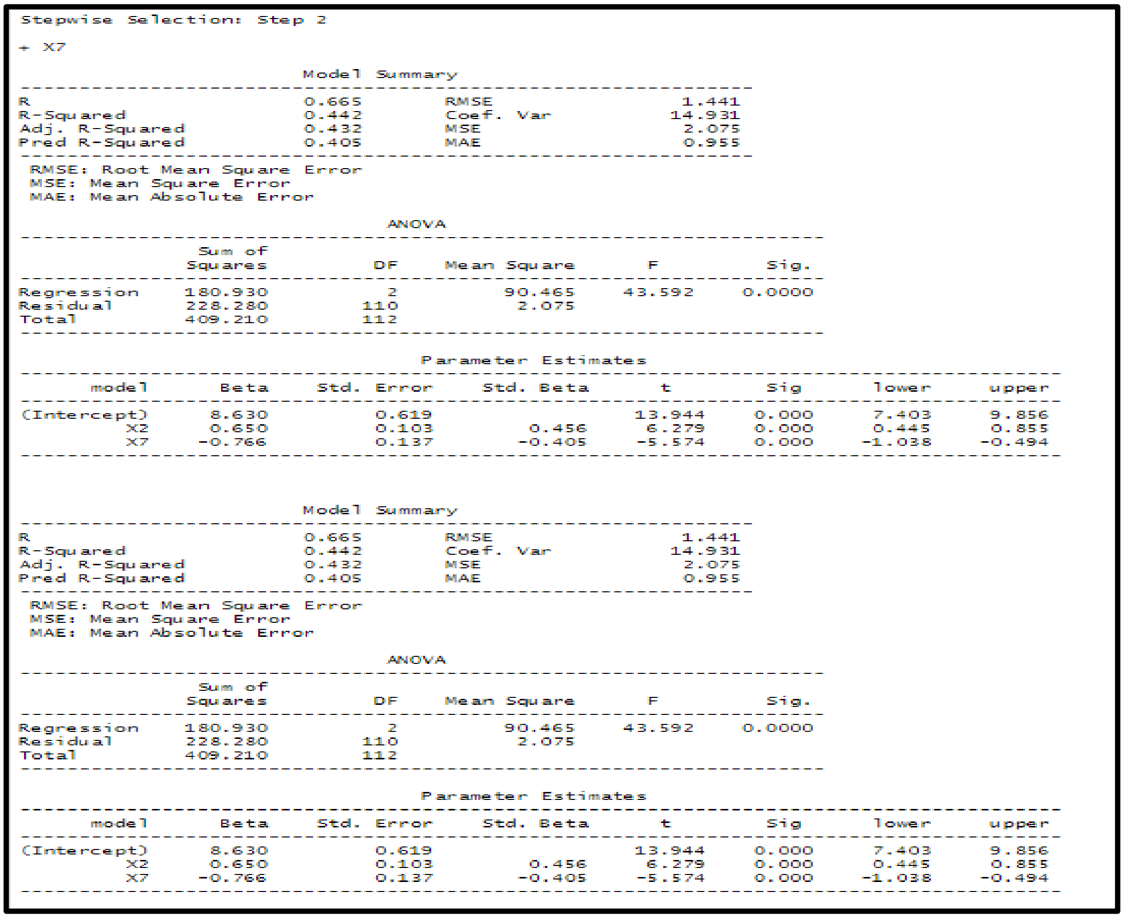
Stepwise Model Selection:



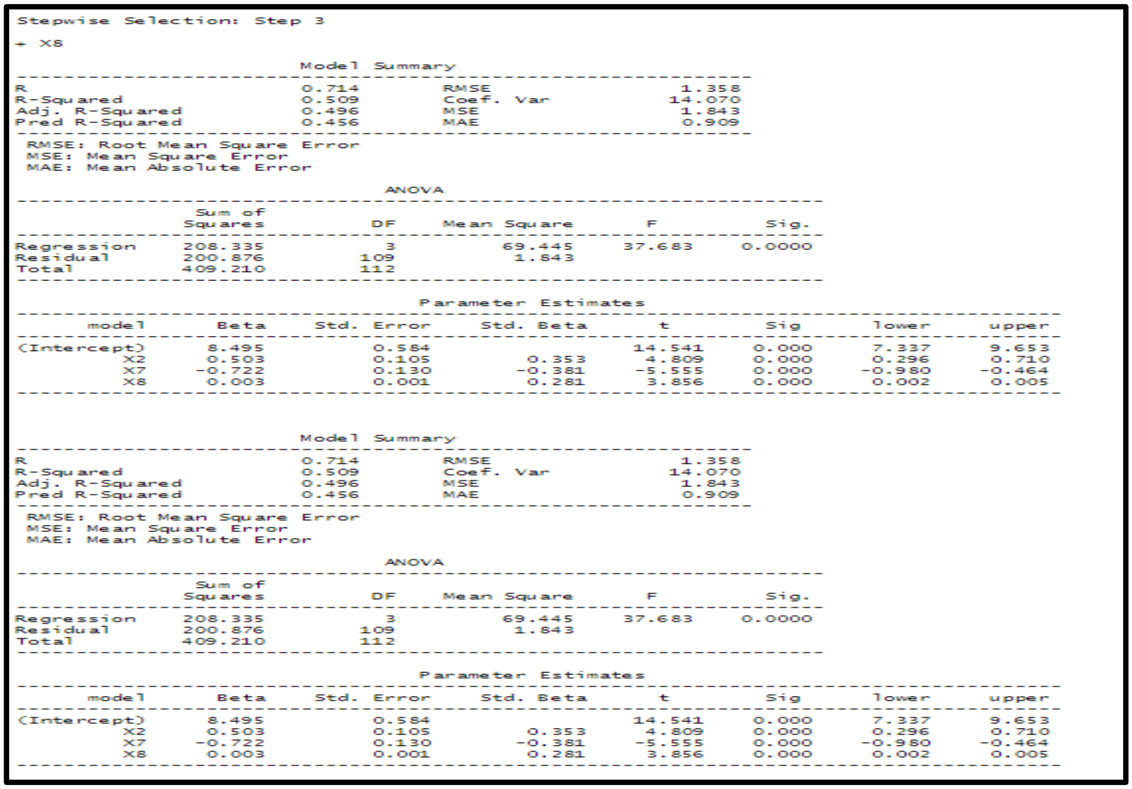
```
> k <- ols_step_both_p(full.lmfit,pent=0.10,prem=0.1,details=TRUE)
Stepwise Selection Method
---------------------------

Candidate Terms:

1. X1
2. X2
3. X3
4. X4
5. X5
6. X6
7. X7
8. X8
9. X9
10. X10

We are selecting variables based on p value...
```



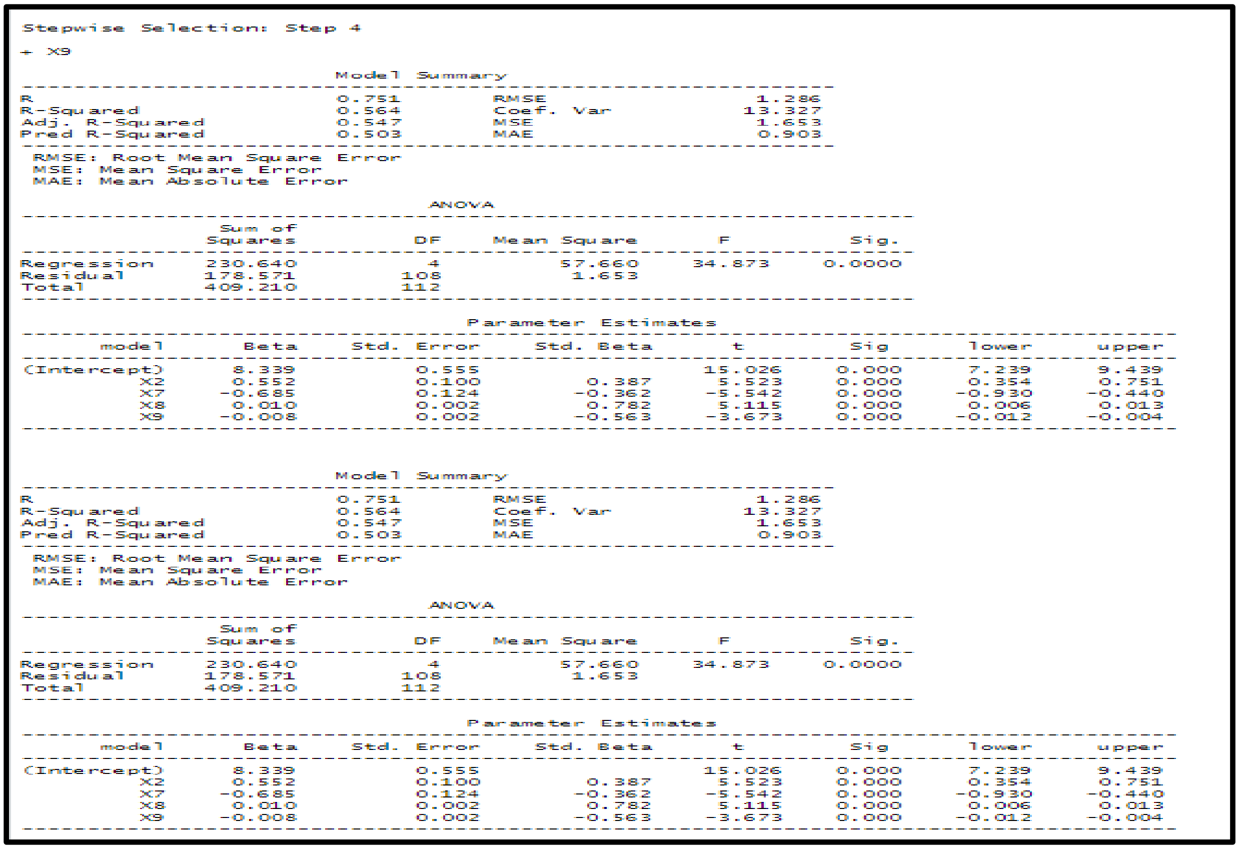
```
Stepwise Selection: Step 1

+ X2

                        Model Summary
--------------------------------------------------------------
R                        0.533       RMSE               1.624
R-Squared                0.285       Coef. Var         16.832
Adj. R-Squared           0.278       MSE                2.638
Pred R-Squared           0.254       MAE                1.104
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                               ANOVA
--------------------------------------------------------------------
                Sum of
               Squares         DF     Mean Square      F         Sig.
--------------------------------------------------------------------
Regression     116.446          1         116.446    44.15     0.0000
Residual       292.765        111           2.638
Total          409.210        112
--------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------------
      model       Beta     Std. Error     Std. Beta       t         Sig      lower    upper
-------------------------------------------------------------------------------------
(Intercept)      6.337          0.521                  12.156      0.000     5.304    7.370
         X2      0.760          0.114         0.533     6.645      0.000     0.534    0.987
-------------------------------------------------------------------------------------
```



```
Stepwise Selection: Step 2

+ X7

                          Model Summary
--------------------------------------------------------------
R                        0.665       RMSE                1.441
R-Squared                0.442       Coef. Var          14.931
Adj. R-Squared           0.432       MSE                 2.075
Pred R-Squared           0.405       MAE                 0.955
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                  ANOVA
--------------------------------------------------------------------
              Sum of
             Squares        DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression   180.930         2         90.465    43.592    0.0000
Residual     228.280       110          2.075
Total        409.210       112
--------------------------------------------------------------------

                                Parameter Estimates
------------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta        t        Sig      lower     upper
------------------------------------------------------------------------------------------
(Intercept)      8.630         0.619                 13.944     0.000     7.403     9.856
         X2      0.650         0.103        0.456     6.279     0.000     0.445     0.855
         X7     -0.766         0.137       -0.405    -5.574     0.000    -1.038    -0.494
------------------------------------------------------------------------------------------


                          Model Summary
--------------------------------------------------------------
R                        0.665       RMSE                1.441
R-Squared                0.442       Coef. Var          14.931
Adj. R-Squared           0.432       MSE                 2.075
Pred R-Squared           0.405       MAE                 0.955
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                  ANOVA
--------------------------------------------------------------------
              Sum of
             Squares        DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression   180.930         2         90.465    43.592    0.0000
Residual     228.280       110          2.075
Total        409.210       112
--------------------------------------------------------------------

                                Parameter Estimates
------------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta        t        Sig      lower     upper
------------------------------------------------------------------------------------------
(Intercept)      8.630         0.619                 13.944     0.000     7.403     9.856
         X2      0.650         0.103        0.456     6.279     0.000     0.445     0.855
         X7     -0.766         0.137       -0.405    -5.574     0.000    -1.038    -0.494
------------------------------------------------------------------------------------------
```

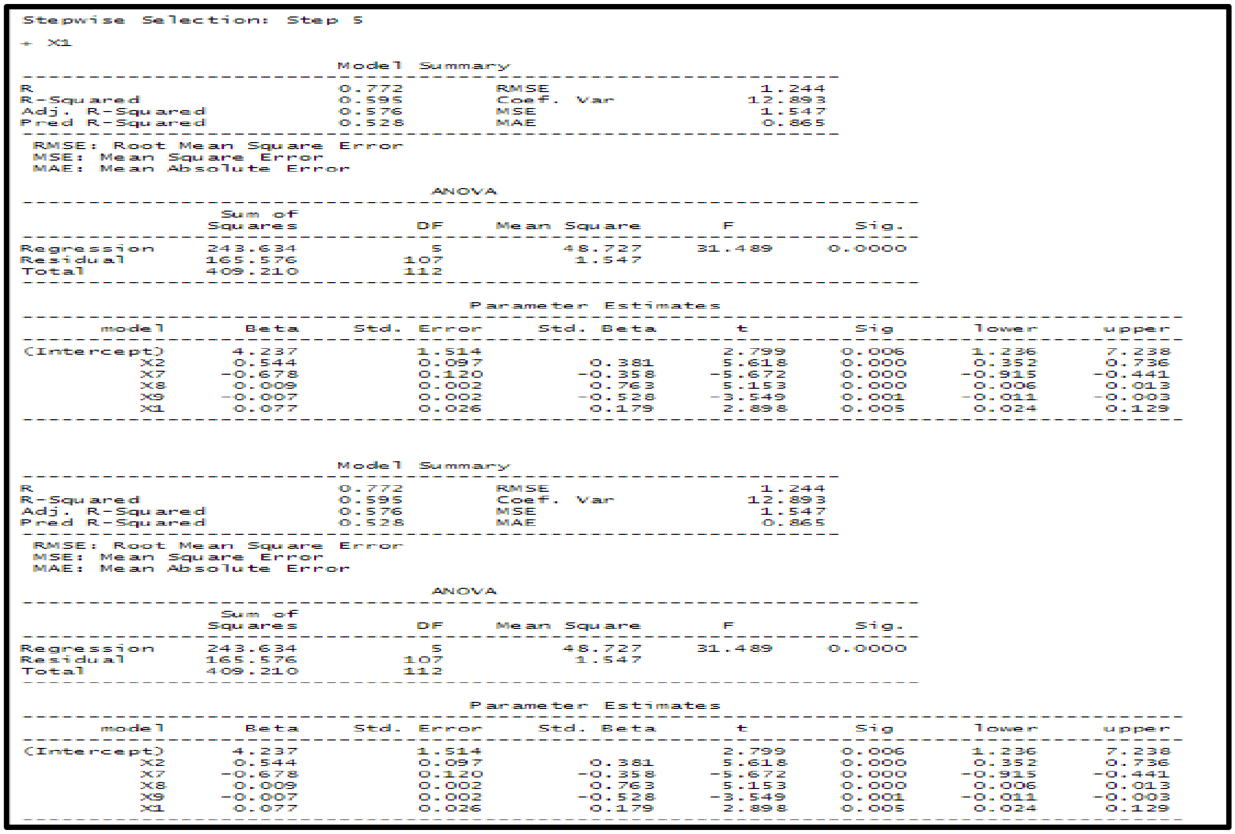


```
Stepwise Selection: Step 3

+ X8

                          Model Summary
--------------------------------------------------------------
R                        0.714       RMSE                1.358
R-Squared                0.509       Coef. Var          14.070
Adj. R-Squared           0.496       MSE                 1.843
Pred R-Squared           0.456       MAE                 0.909
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                  ANOVA
--------------------------------------------------------------------
              Sum of
             Squares        DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression   208.335         3         69.445    37.683    0.0000
Residual     200.876       109          1.843
Total        409.210       112
--------------------------------------------------------------------

                                Parameter Estimates
------------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta        t        Sig      lower     upper
------------------------------------------------------------------------------------------
(Intercept)      8.495         0.584                 14.541     0.000     7.337     9.653
         X2      0.503         0.105        0.353     4.809     0.000     0.296     0.710
         X7     -0.722         0.130       -0.381    -5.555     0.000    -0.980    -0.464
         X8      0.003         0.001        0.281     3.856     0.000     0.002     0.005
------------------------------------------------------------------------------------------


                          Model Summary
--------------------------------------------------------------
R                        0.714       RMSE                1.358
R-Squared                0.509       Coef. Var          14.070
Adj. R-Squared           0.496       MSE                 1.843
Pred R-Squared           0.456       MAE                 0.909
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                  ANOVA
--------------------------------------------------------------------
              Sum of
             Squares        DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression   208.335         3         69.445    37.683    0.0000
Residual     200.876       109          1.843
Total        409.210       112
--------------------------------------------------------------------

                                Parameter Estimates
------------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta        t        Sig      lower     upper
------------------------------------------------------------------------------------------
(Intercept)      8.495         0.584                 14.541     0.000     7.337     9.653
         X2      0.503         0.105        0.353     4.809     0.000     0.296     0.710
         X7     -0.722         0.130       -0.381    -5.555     0.000    -0.980    -0.464
         X8      0.003         0.001        0.281     3.856     0.000     0.002     0.005
------------------------------------------------------------------------------------------
```



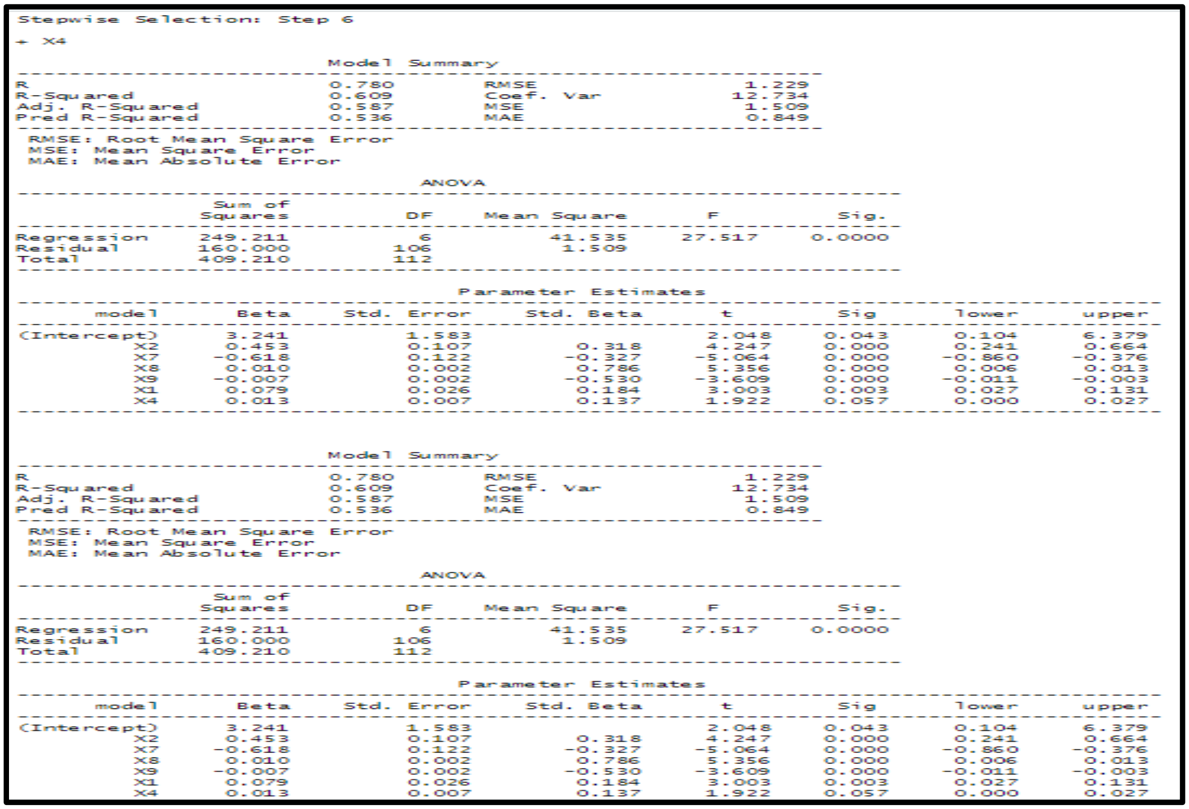
```
Stepwise Selection: Step 4

+ X9

                              Model Summary
--------------------------------------------------------------
R                             0.751       RMSE             1.286
R-Squared                     0.564       Coef. Var       13.327
Adj. R-Squared                0.547       MSE              1.653
Pred R-Squared                0.503       MAE              0.903
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                  ANOVA
--------------------------------------------------------------------
              Sum of
             Squares         DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression   230.640          4         57.660    34.873     0.0000
Residual     178.571        108          1.653
Total        409.210        112
--------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta        t        Sig      lower     upper
-------------------------------------------------------------------------------------------
(Intercept)      8.339         0.555                 15.026     0.000      7.239     9.439
         X2      0.552         0.100        0.387     5.523     0.000      0.354     0.751
         X7     -0.685         0.124       -0.362    -5.542     0.000     -0.930    -0.440
         X8      0.010         0.002        0.782     5.115     0.000      0.006     0.013
         X9     -0.008         0.002       -0.563    -3.673     0.000     -0.012    -0.004
-------------------------------------------------------------------------------------------
```

```
                              Model Summary
--------------------------------------------------------------
R                             0.751       RMSE             1.286
R-Squared                     0.564       Coef. Var       13.327
Adj. R-Squared                0.547       MSE              1.653
Pred R-Squared                0.503       MAE              0.903
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                  ANOVA
--------------------------------------------------------------------
              Sum of
             Squares         DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression   230.640          4         57.660    34.873     0.0000
Residual     178.571        108          1.653
Total        409.210        112
--------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta        t        Sig      lower     upper
-------------------------------------------------------------------------------------------
(Intercept)      8.339         0.555                 15.026     0.000      7.239     9.439
         X2      0.552         0.100        0.387     5.523     0.000      0.354     0.751
         X7     -0.685         0.124       -0.362    -5.542     0.000     -0.930    -0.440
         X8      0.010         0.002        0.782     5.115     0.000      0.006     0.013
         X9     -0.008         0.002       -0.563    -3.673     0.000     -0.012    -0.004
-------------------------------------------------------------------------------------------
```



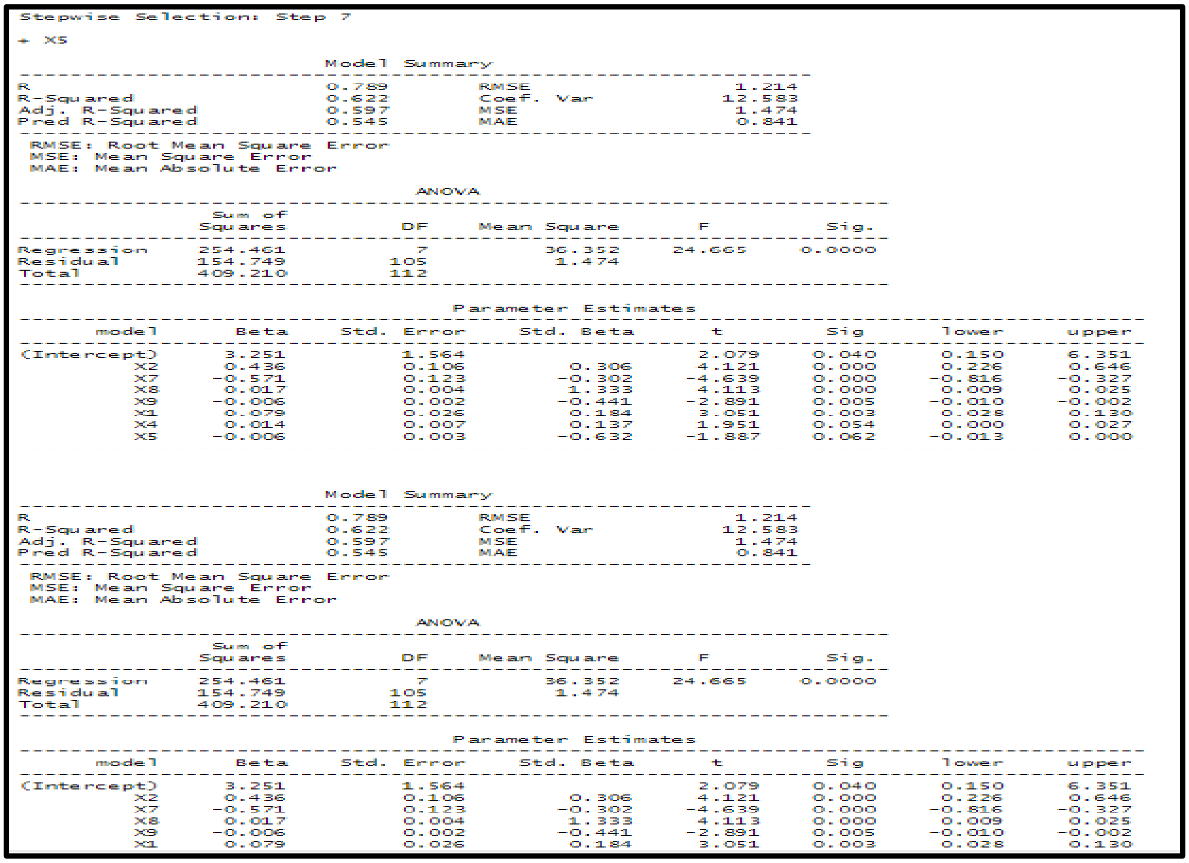
```
Stepwise Selection: Step 5

+ X1

                              Model Summary
--------------------------------------------------------------
R                             0.772       RMSE             1.244
R-Squared                     0.595       Coef. Var       12.893
Adj. R-Squared                0.576       MSE              1.547
Pred R-Squared                0.528       MAE              0.865
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                  ANOVA
--------------------------------------------------------------------
              Sum of
             Squares         DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression   243.634          5         48.727    31.489     0.0000
Residual     165.576        107          1.547
Total        409.210        112
--------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta        t        Sig      lower     upper
-------------------------------------------------------------------------------------------
(Intercept)      4.237         1.514                  2.799     0.006      1.236     7.238
         X2      0.544         0.097        0.381     5.618     0.000      0.352     0.736
         X7     -0.678         0.120       -0.358    -5.672     0.000     -0.915    -0.441
         X8      0.009         0.002        0.763     5.153     0.000      0.006     0.013
         X9     -0.007         0.002       -0.528    -3.549     0.001     -0.011    -0.003
         X1      0.077         0.026        0.179     2.898     0.005      0.024     0.129
-------------------------------------------------------------------------------------------
```

```
                              Model Summary
--------------------------------------------------------------
R                             0.772       RMSE             1.244
R-Squared                     0.595       Coef. Var       12.893
Adj. R-Squared                0.576       MSE              1.547
Pred R-Squared                0.528       MAE              0.865
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                  ANOVA
--------------------------------------------------------------------
              Sum of
             Squares         DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression   243.634          5         48.727    31.489     0.0000
Residual     165.576        107          1.547
Total        409.210        112
--------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta        t        Sig      lower     upper
-------------------------------------------------------------------------------------------
(Intercept)      4.237         1.514                  2.799     0.006      1.236     7.238
         X2      0.544         0.097        0.381     5.618     0.000      0.352     0.736
         X7     -0.678         0.120       -0.358    -5.672     0.000     -0.915    -0.441
         X8      0.009         0.002        0.763     5.153     0.000      0.006     0.013
         X9     -0.007         0.002       -0.528    -3.549     0.001     -0.011    -0.003
         X1      0.077         0.026        0.179     2.898     0.005      0.024     0.129
-------------------------------------------------------------------------------------------
```



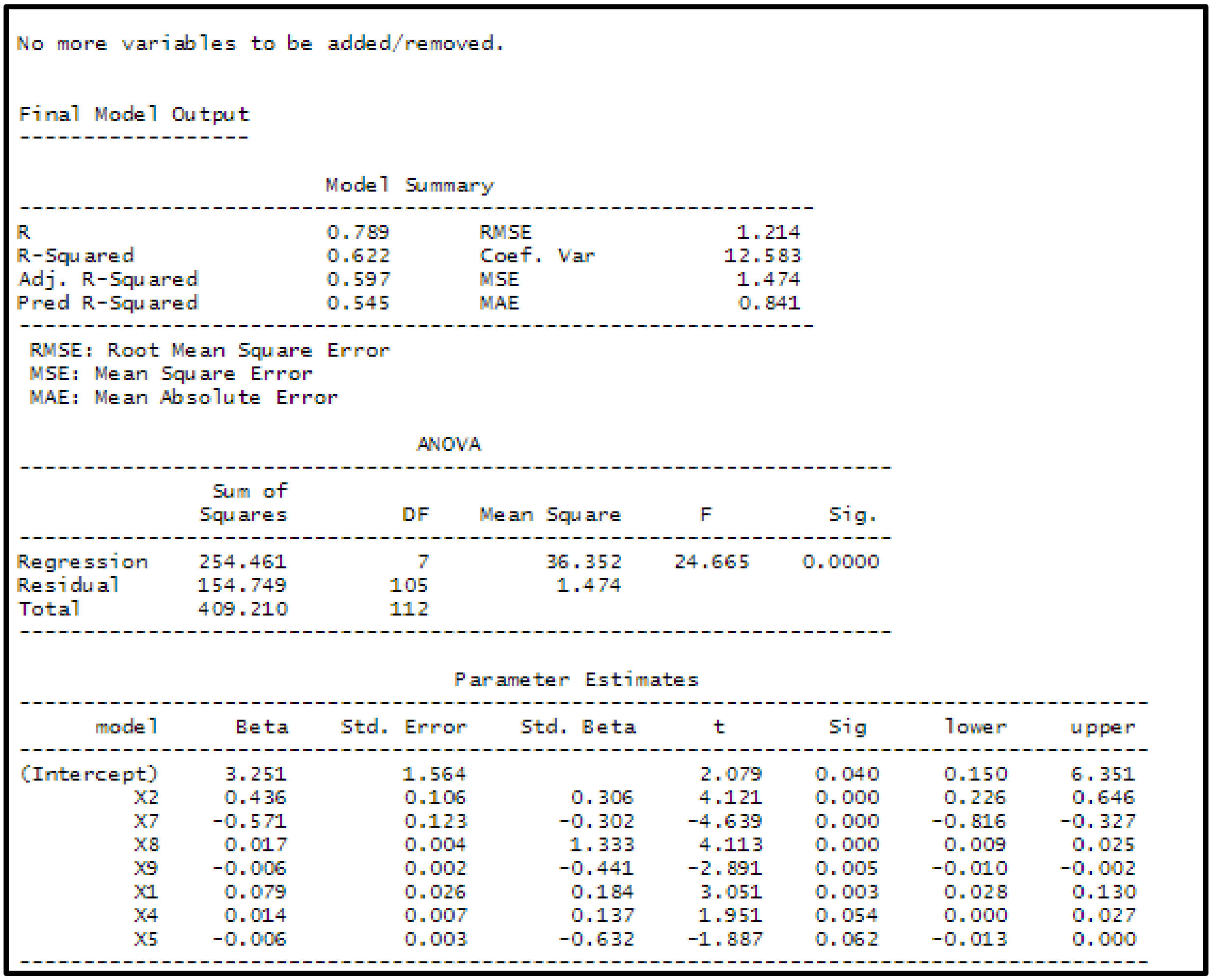
```
Stepwise Selection: Step 6

+ X4

                              Model Summary
---------------------------------------------------------------
R                             0.780       RMSE            1.229
R-Squared                     0.609       Coef. Var      12.734
Adj. R-Squared                0.587       MSE             1.509
Pred R-Squared                0.536       MAE             0.849
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                   ANOVA
---------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square      F        Sig.
---------------------------------------------------------------------
Regression     249.211          6         41.535    27.517    0.0000
Residual       160.000        106          1.509
Total          409.210        112
---------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------------------
      model         Beta    Std. Error    Std. Beta       t        Sig      lower     upper
-------------------------------------------------------------------------------------------
(Intercept)        3.241         1.583                  2.048    0.043     0.104     6.379
         X2        0.453         0.107        0.318     4.247    0.000     0.241     0.664
         X7       -0.618         0.122       -0.327    -5.064    0.000    -0.860    -0.376
         X8        0.010         0.002        0.786     5.356    0.000     0.006     0.013
         X9       -0.007         0.002       -0.530    -3.609    0.000    -0.011    -0.003
         X1        0.079         0.026        0.184     3.003    0.003     0.027     0.131
         X4        0.013         0.007        0.137     1.922    0.057     0.000     0.027
-------------------------------------------------------------------------------------------
```

```
                              Model Summary
---------------------------------------------------------------
R                             0.780       RMSE            1.229
R-Squared                     0.609       Coef. Var      12.734
Adj. R-Squared                0.587       MSE             1.509
Pred R-Squared                0.536       MAE             0.849
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                   ANOVA
---------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square      F        Sig.
---------------------------------------------------------------------
Regression     249.211          6         41.535    27.517    0.0000
Residual       160.000        106          1.509
Total          409.210        112
---------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------------------
      model         Beta    Std. Error    Std. Beta       t        Sig      lower     upper
-------------------------------------------------------------------------------------------
(Intercept)        3.241         1.583                  2.048    0.043     0.104     6.379
         X2        0.453         0.107        0.318     4.247    0.000     0.241     0.664
         X7       -0.618         0.122       -0.327    -5.064    0.000    -0.860    -0.376
         X8        0.010         0.002        0.786     5.356    0.000     0.006     0.013
         X9       -0.007         0.002       -0.530    -3.609    0.000    -0.011    -0.003
         X1        0.079         0.026        0.184     3.003    0.003     0.027     0.131
         X4        0.013         0.007        0.137     1.922    0.057     0.000     0.027
```

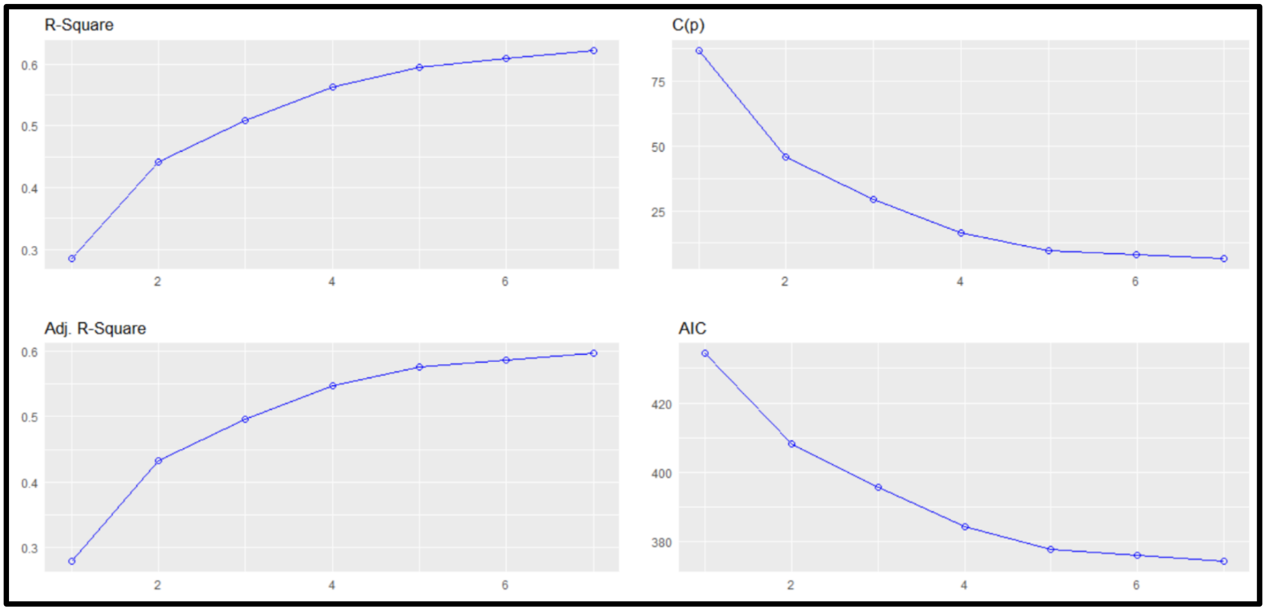


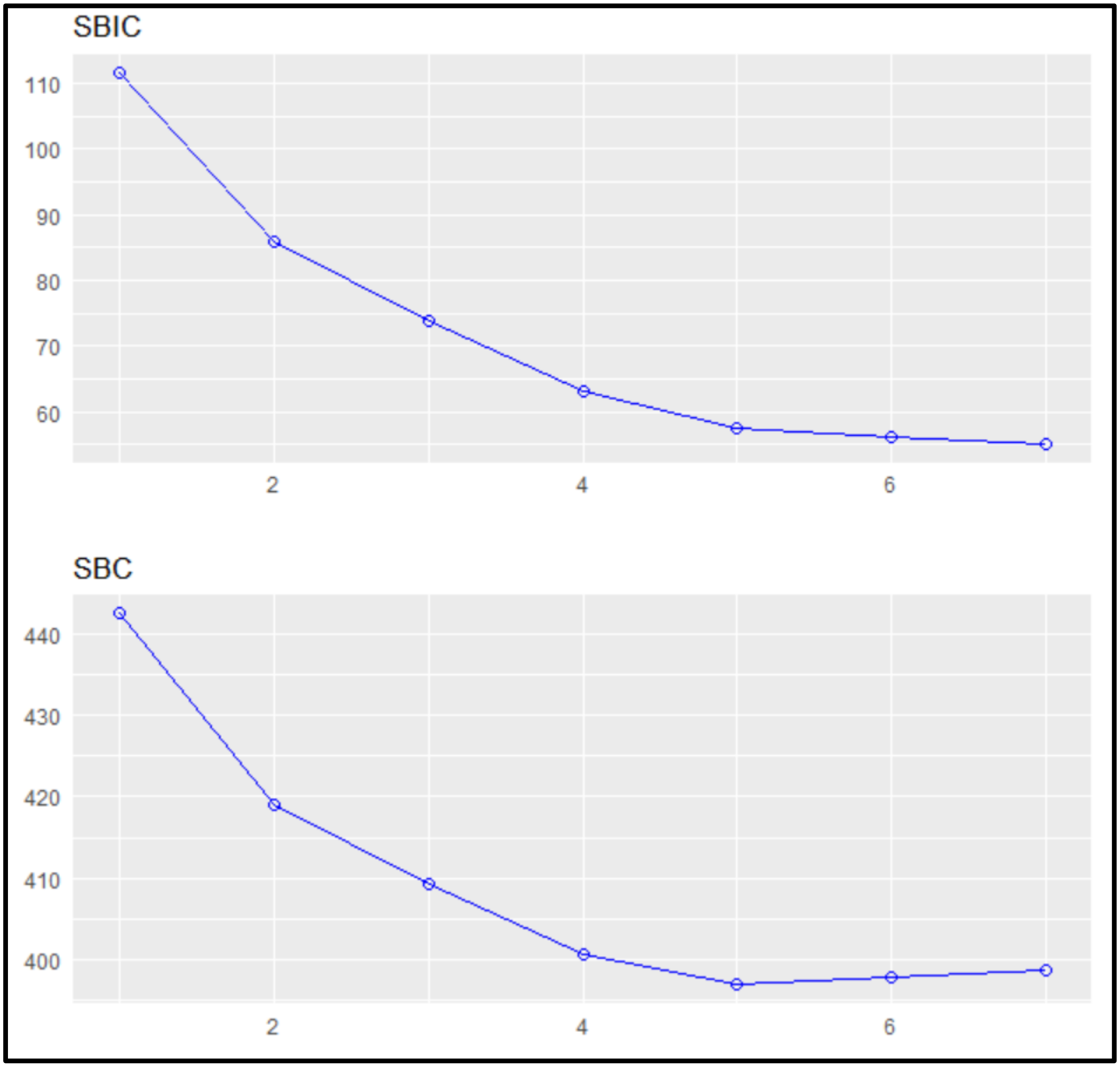
```
Stepwise Selection: Step 7

+ X5

                              Model Summary
---------------------------------------------------------------
R                             0.789       RMSE            1.214
R-Squared                     0.622       Coef. Var      12.583
Adj. R-Squared                0.597       MSE             1.474
Pred R-Squared                0.545       MAE             0.841
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                   ANOVA
---------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square      F        Sig.
---------------------------------------------------------------------
Regression     254.461          7         36.352    24.665    0.0000
Residual       154.749        105          1.474
Total          409.210        112
---------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------------------
      model         Beta    Std. Error    Std. Beta       t        Sig      lower     upper
-------------------------------------------------------------------------------------------
(Intercept)        3.251         1.564                  2.079    0.040     0.150     6.351
         X2        0.436         0.106        0.306     4.121    0.000     0.226     0.646
         X7       -0.571         0.123       -0.302    -4.639    0.000    -0.816    -0.327
         X8        0.017         0.004        1.333     4.113    0.000     0.009     0.025
         X9       -0.006         0.002       -0.441    -2.891    0.005    -0.010    -0.002
         X1        0.079         0.026        0.184     3.051    0.003     0.028     0.130
         X4        0.014         0.007        0.137     1.951    0.054     0.000     0.027
         X5       -0.006         0.003       -0.632    -1.887    0.062    -0.013     0.000
-------------------------------------------------------------------------------------------
```

```
                              Model Summary
---------------------------------------------------------------
R                             0.789       RMSE            1.214
R-Squared                     0.622       Coef. Var      12.583
Adj. R-Squared                0.597       MSE             1.474
Pred R-Squared                0.545       MAE             0.841
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                   ANOVA
---------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square      F        Sig.
---------------------------------------------------------------------
Regression     254.461          7         36.352    24.665    0.0000
Residual       154.749        105          1.474
Total          409.210        112
---------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------------------
      model         Beta    Std. Error    Std. Beta       t        Sig      lower     upper
-------------------------------------------------------------------------------------------
(Intercept)        3.251         1.564                  2.079    0.040     0.150     6.351
         X2        0.436         0.106        0.306     4.121    0.000     0.226     0.646
         X7       -0.571         0.123       -0.302    -4.639    0.000    -0.816    -0.327
         X8        0.017         0.004        1.333     4.113    0.000     0.009     0.025
         X9       -0.006         0.002       -0.441    -2.891    0.005    -0.010    -0.002
         X1        0.079         0.026        0.184     3.051    0.003     0.028     0.130
```

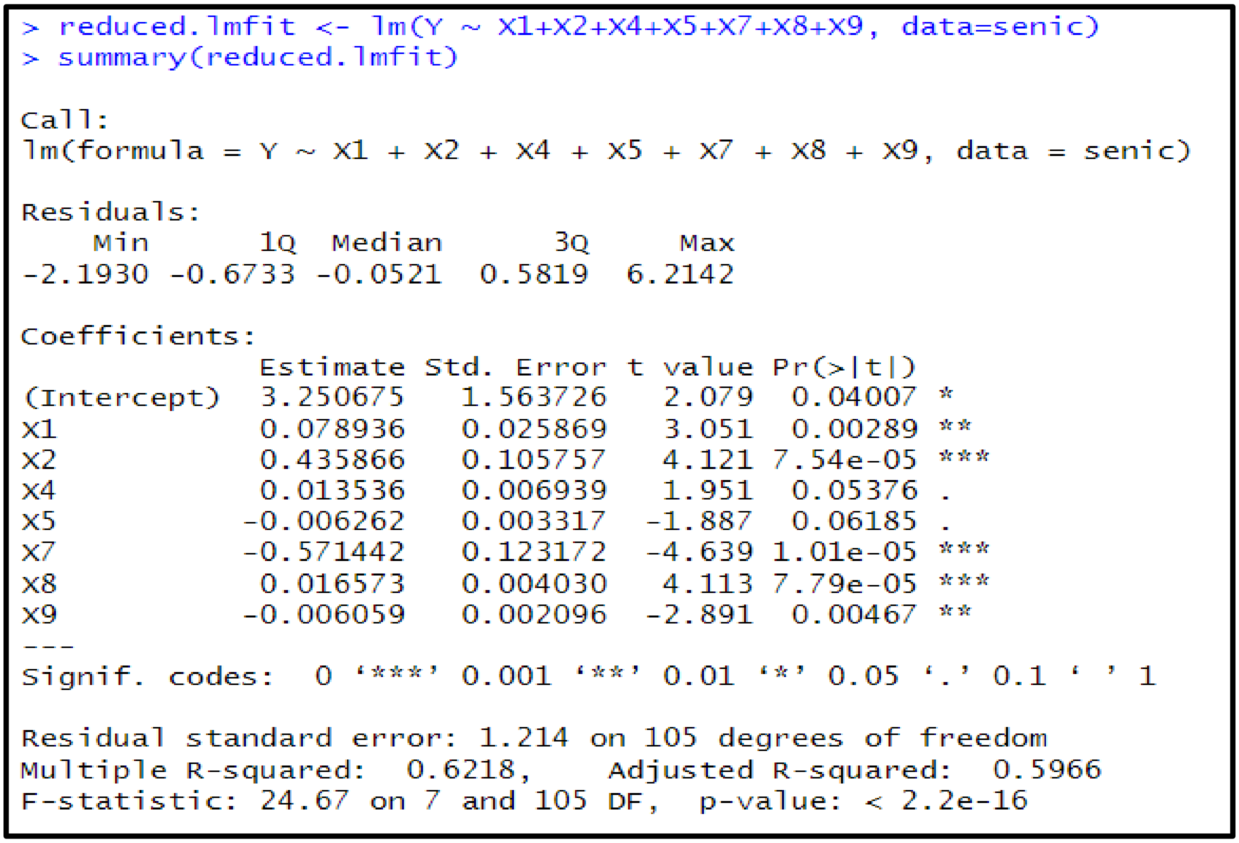


```
No more variables to be added/removed.


Final Model Output
------------------

                            Model Summary
-----------------------------------------------------------------
R                        0.789       RMSE                1.214
R-Squared                0.622       Coef. Var          12.583
Adj. R-Squared           0.597       MSE                 1.474
Pred R-Squared           0.545       MAE                 0.841
-----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                   ANOVA
-----------------------------------------------------------------------
              Sum of
             Squares        DF    Mean Square      F          Sig.
-----------------------------------------------------------------------
Regression   254.461         7         36.352    24.665    0.0000
Residual     154.749       105          1.474
Total        409.210       112
-----------------------------------------------------------------------

                                      Parameter Estimates
-------------------------------------------------------------------------------------------
      model        Beta    Std. Error    Std. Beta      t        Sig      lower     upper
-------------------------------------------------------------------------------------------
(Intercept)       3.251         1.564                 2.079    0.040      0.150     6.351
         X2       0.436         0.106        0.306    4.121    0.000      0.226     0.646
         X7      -0.571         0.123       -0.302   -4.639    0.000     -0.816    -0.327
         X8       0.017         0.004        1.333    4.113    0.000      0.009     0.025
         X9      -0.006         0.002       -0.441   -2.891    0.005     -0.010    -0.002
         X1       0.079         0.026        0.184    3.051    0.003      0.028     0.130
         X4       0.014         0.007        0.137    1.951    0.054      0.000     0.027
         X5      -0.006         0.003       -0.632   -1.887    0.062     -0.013     0.000
-------------------------------------------------------------------------------------------
```




SBIC
110
100
90
80
70
60
2
4
6
SBC
440
430
420
410
400
2
4
6

MODEL EVALUATION

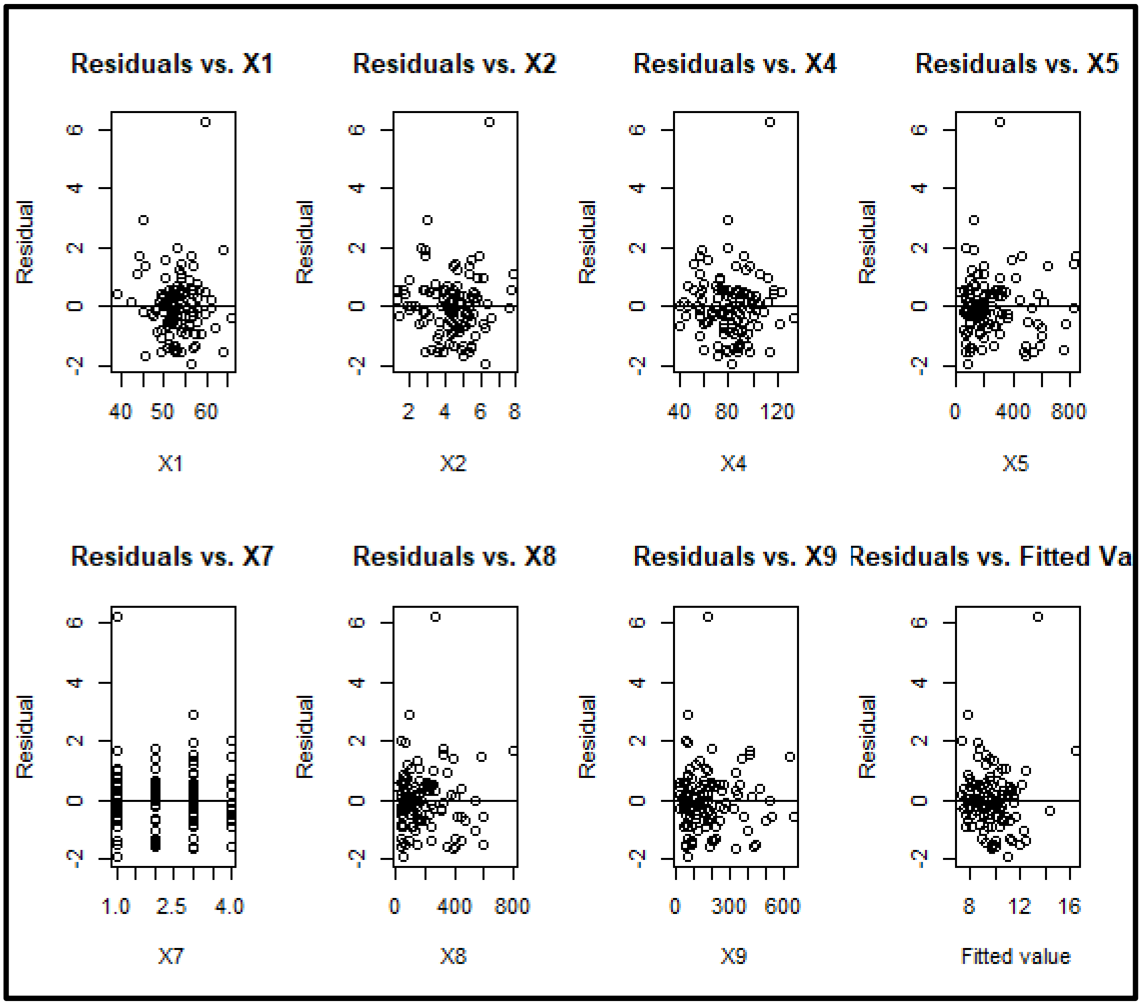


```
> reduced.lmfit <- lm(Y ~ X1+X2+X4+X5+X7+X8+X9, data=senic)
> summary(reduced.lmfit)

Call:
lm(formula = Y ~ X1 + X2 + X4 + X5 + X7 + X8 + X9, data = senic)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1930 -0.6733 -0.0521  0.5819  6.2142 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.250675   1.563726   2.079  0.04007 *  
X1           0.078936   0.025869   3.051  0.00289 ** 
X2           0.435866   0.105757   4.121 7.54e-05 ***
X4           0.013536   0.006939   1.951  0.05376 .  
X5          -0.006262   0.003317  -1.887  0.06185 .  
X7          -0.571442   0.123172  -4.639 1.01e-05 ***
X8           0.016573   0.004030   4.113 7.79e-05 ***
X9          -0.006059   0.002096  -2.891  0.00467 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.214 on 105 degrees of freedom
Multiple R-squared:  0.6218,	Adjusted R-squared:  0.5966 
F-statistic: 24.67 on 7 and 105 DF,  p-value: < 2.2e-16
```

In this reduced linear regression model, we assess the impact of predictors on the response variable Y. The intercept of 3.2507 is statistically significant (p = 0.04007), suggesting that when all predictors are zero, the estimated mean response is 3.2507. Notably, X2 has a substantial positive effect (estimate = 0.4359, p < 7.54e-05), indicating that a one-unit increase in X2 is associated with an increase in Y. Conversely, X7 has a negative impact (estimate = -0.5714, p < 1.01e-05), suggesting that higher values of X7 correspond to lower values of Y. The model overall is significant (F-statistic = 24.67, p < 2.2e-16), explaining 62.18% of the variance in Y. However, attention should be given to predictors X4 and X5, which have marginal significance, and further model refinement is recommended. The residuals exhibit a standard error of 1.214, indicating the unexplained variability in the model, and the adjusted R-squared is 0.5966.

3.3 Model diagnostics/ Regression Diagnostics:

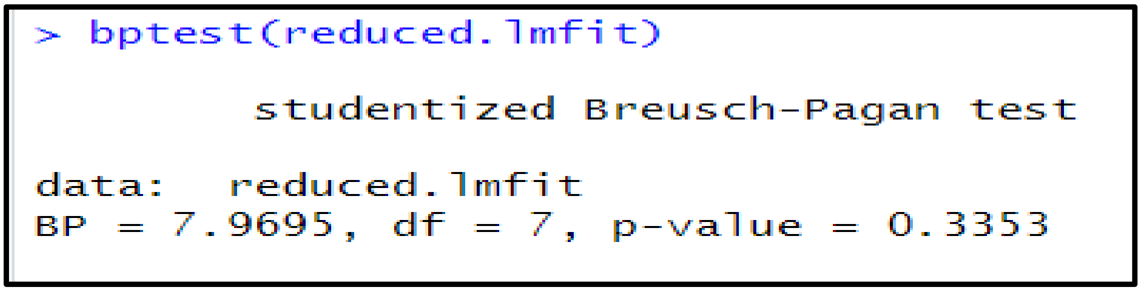
Linearity



Residual plot of Final Model

Looking at our residual vs. fitted value plots, we can see that our model meets the linearity assumption as the residuals are randomly distributed around the fitted values. Similarly, the jackknifed residual vs. predictor value plots also indicates that the linearity assumption is met as the residuals are randomly distributed around the predictor variables. That means our model is a good fit for our data.

The constancy of error Variance

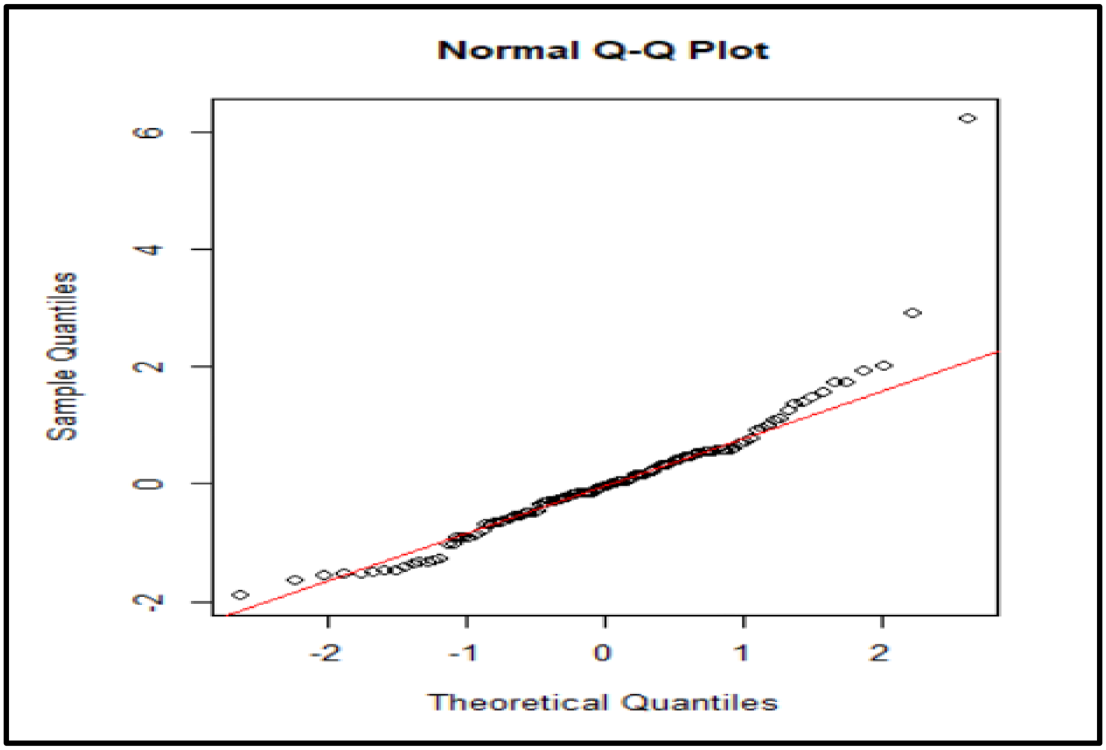


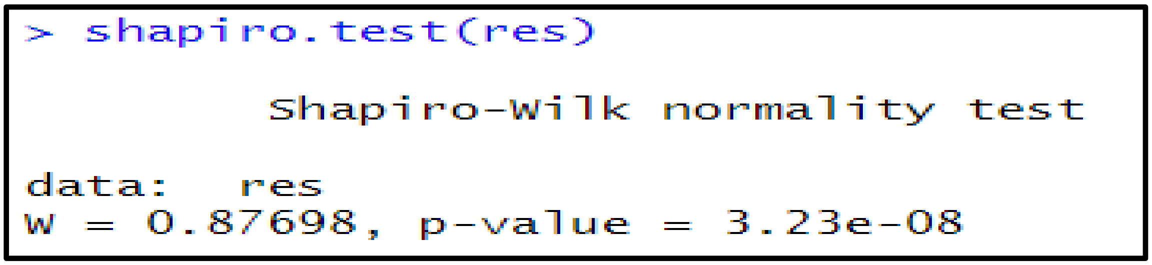
```
> bptest(reduced.lmfit)

        studentized Breusch-Pagan test

data:  reduced.lmfit
BP = 7.9695, df = 7, p-value = 0.3353
```

In the Breusch-Pagan test, the null hypothesis states that there is constant error variance and the alternative hypothesis states that there is not constant variance. The decision rule is that if the p-value is less than the significance level of 0.05, we will reject the null hypothesis and conclude that the error variance is not constant. If the p-value is greater than the significance level of 0.05, we will fail to reject the null hypothesis and conclude that the error variance is constant.

We calculated a test statistic of 7.9695 and a p-value of 0.3353 so we failed to reject the null hypothesis and conclude that the error terms are constant.

Normality



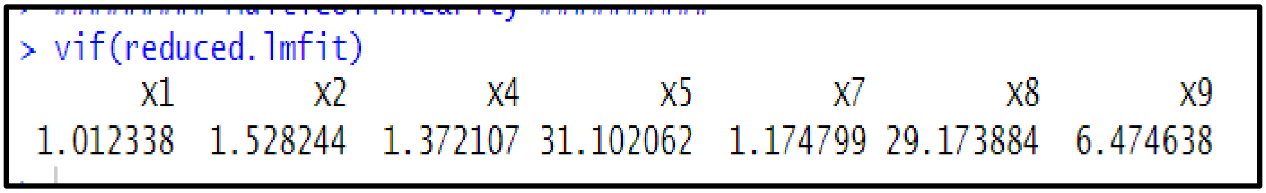


```
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.87698, p-value = 3.23e-08
```

The normal probability plot indicates that the data nearly forms a normal distribution because the data points mostly align with only some slight deviation near the lowest and highest values. In the Shapiro-Wilk test, the null hypothesis states that the error terms are normally distributed while the alternative hypothesis states that the error terms are not normally distributed. Our decision rule states that if the test statistic is small and the p-value is less than the significance level (alpha = 0.05), then we must reject the null hypothesis. If the test statistic is large and the p-value is greater than the significance level, we must fail to reject the null hypothesis.

We calculated a test statistic of 0.87698 and a p-value of 3.23e-08. Thus, we can reject the null hypothesis and conclude that the error terms are not normally distributed.

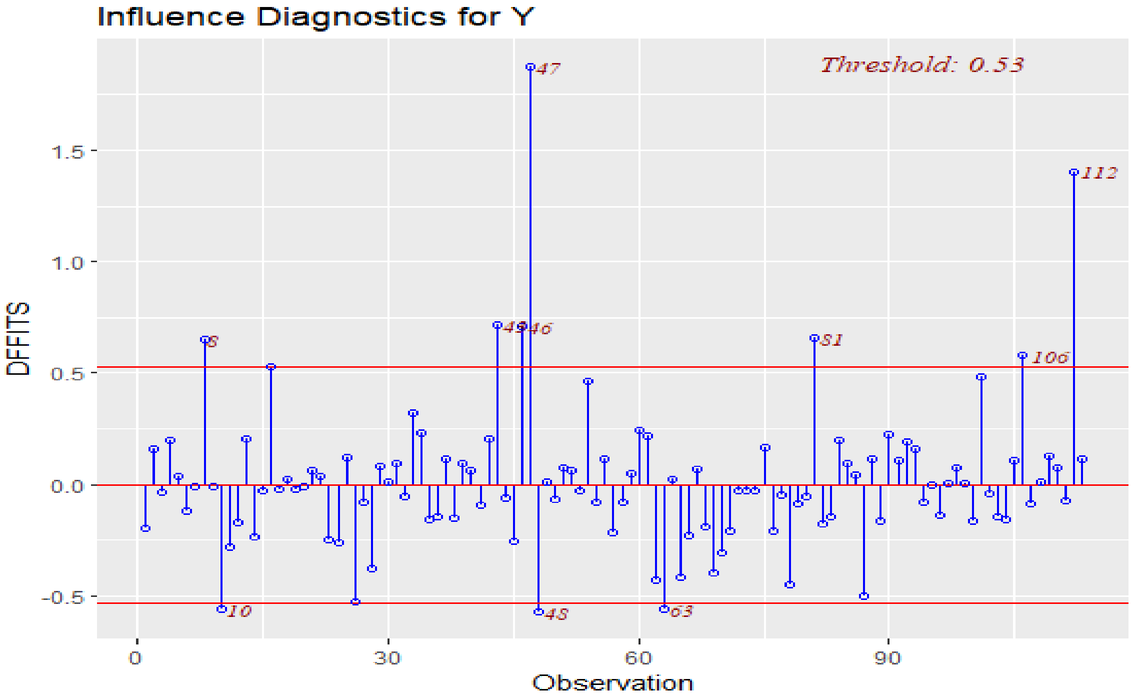
Multicollinearity

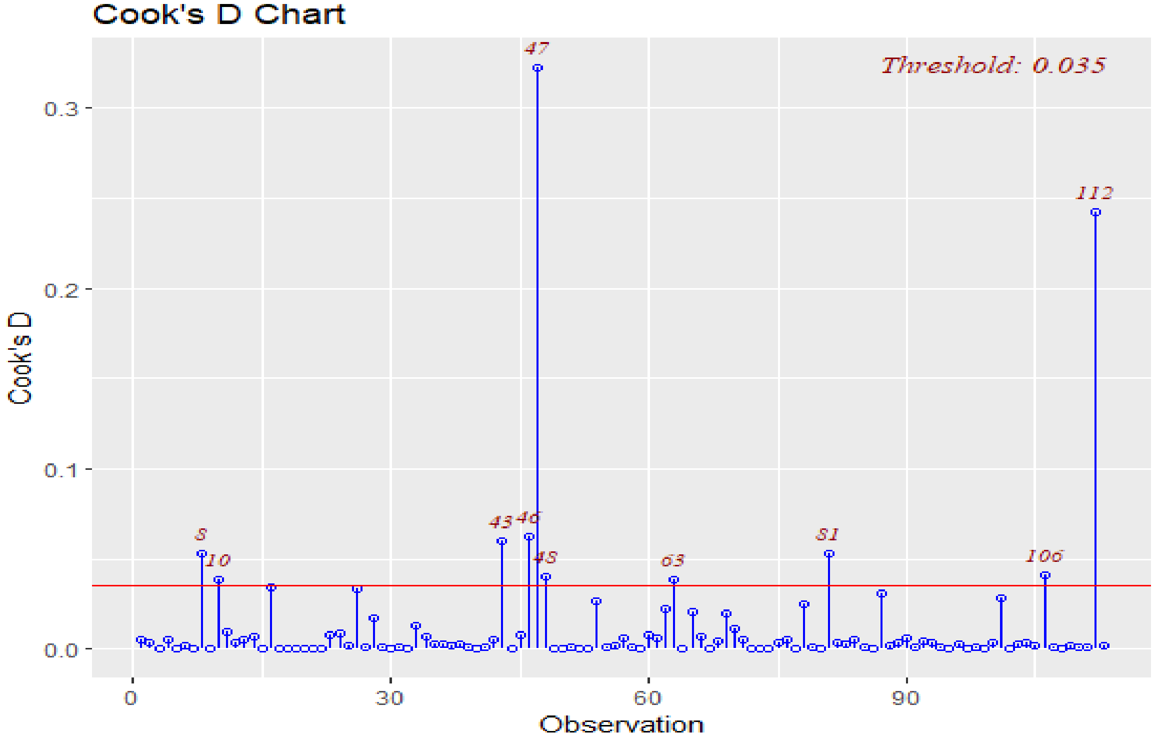


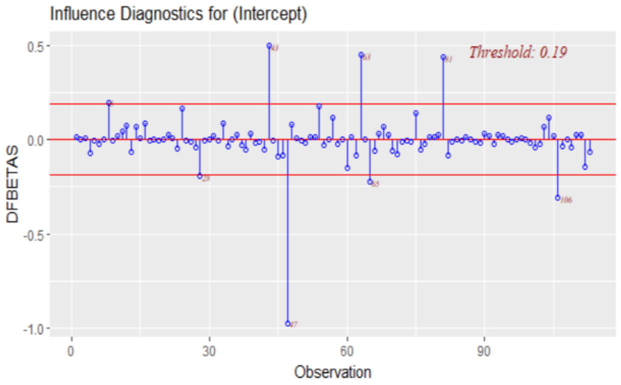
```
> vif(reduced.lmfit)
       X1        X2        X4        X5        X7        X8        X9 
 1.012338  1.528244  1.372107 31.102062  1.174799 29.173884  6.474638 
```

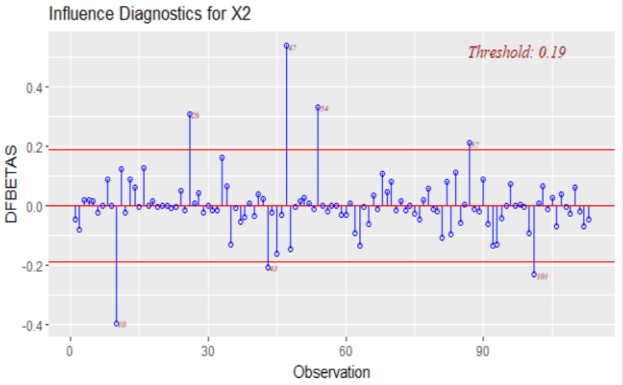
The values obtained from the variance inflation factor analysis indicate that multicollinearity is a major problem in our final model. This is because the VIF values for X5 and X8 are more than 10.

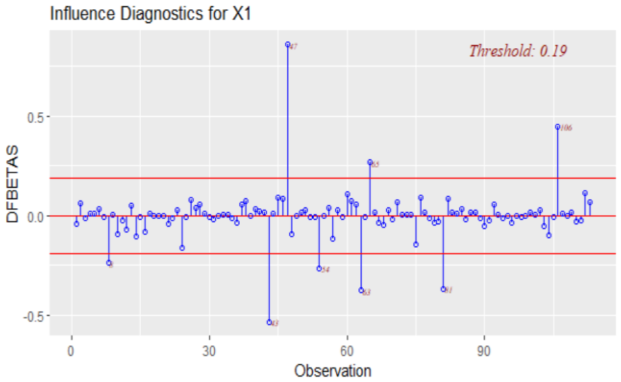
Influential Plots

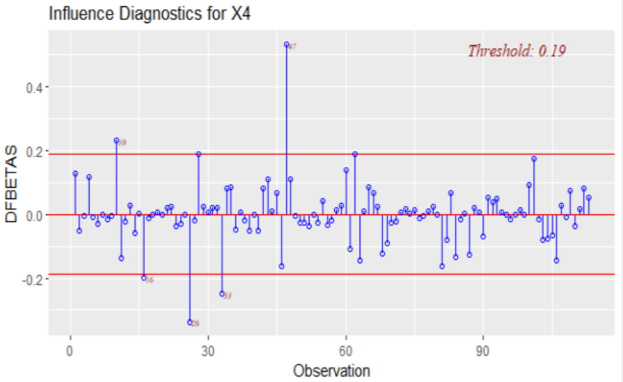


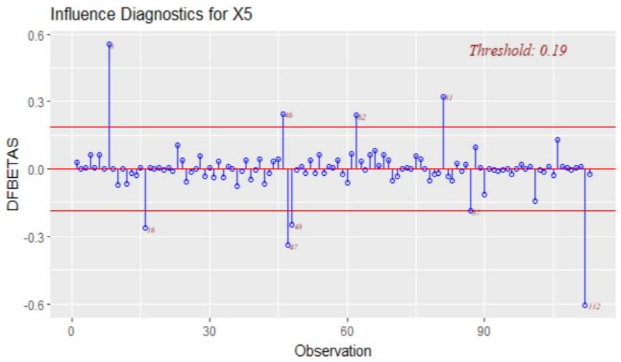


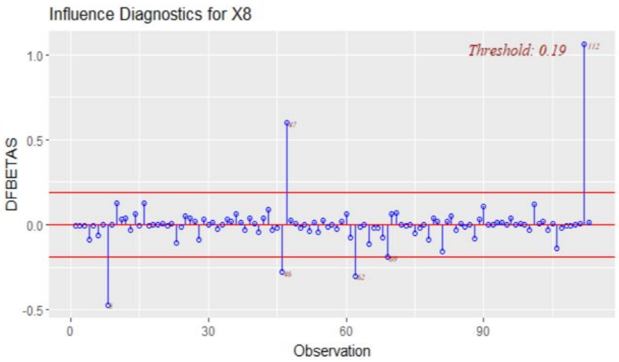

Influence Diagnostics for (Intercept)
Threshold: 0.19
DFBETAS
0.5
0.0
-0.5
-1.0
0
30
60
90
Observation

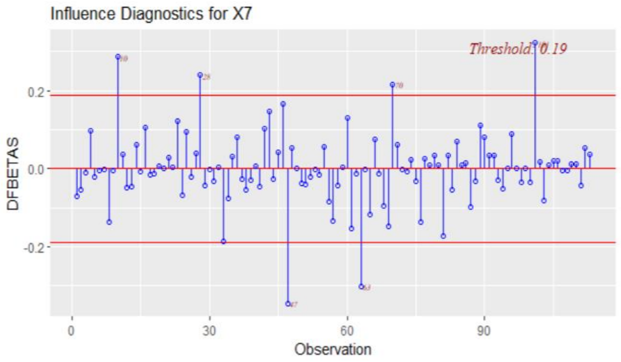

Influence Diagnostics for X2
Threshold: 0.19
DFBETAS
0.4
0.2
0.0
-0.2
-0.4
0
30
60
90
Observation

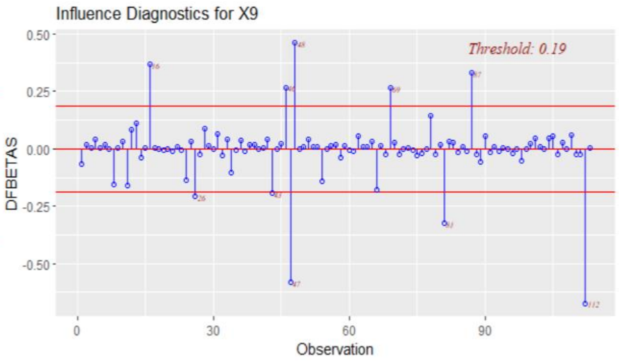

Influence Diagnostics for X1
Threshold: 0.19
DFBETAS
0.5
0.0
-0.5
0
30
60
90
Observation


Influence Diagnostics for X4
Threshold: 0.19
DFBETAS
0.4
0.2
0.0
-0.2
0
30
60
90
Observation


Influence Diagnostics for X5
Threshold: 0.19
DFBETAS
0.6
0.3
0.0
-0.3
-0.6
0
30
60
90
Observation

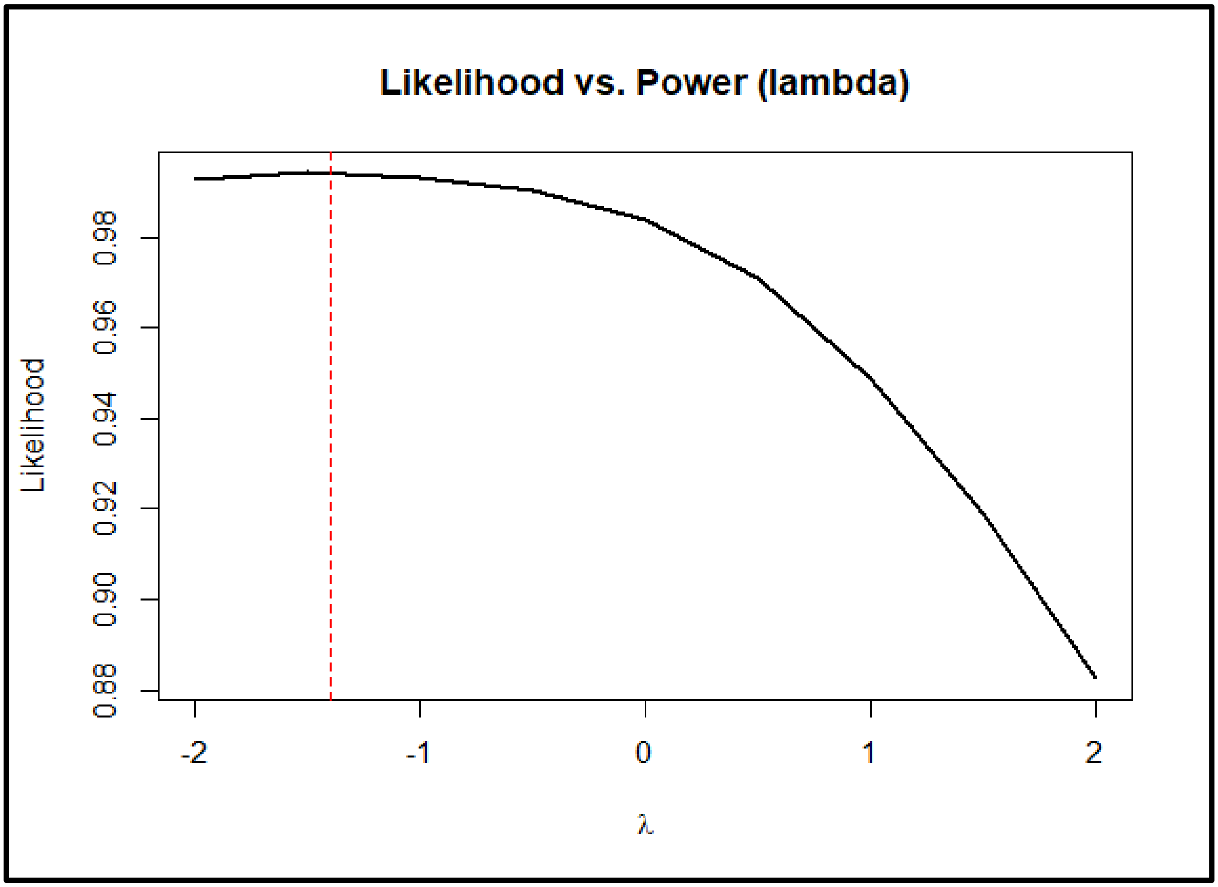

Influence Diagnostics for X8
Threshold: 0.19
DFBETAS
1.0
0.5
0.0
-0.5
0
30
60
90
Observation

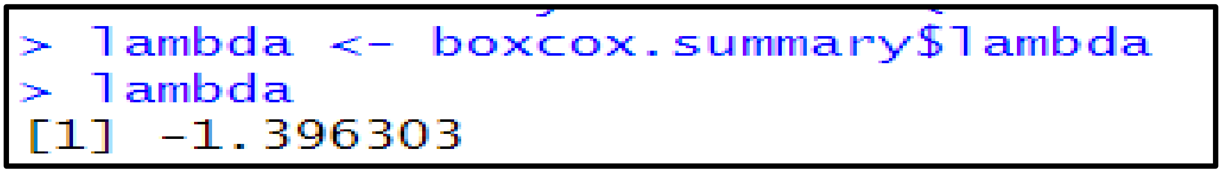

Influence Diagnostics for X7
Threshold: 0.19
DFBETAS
0.2
0.0
-0.2
0
30
60
90
Observation


Influence Diagnostics for X9
Threshold: 0.19
DFBETAS
0.50
0.25
0.00
-0.25
-0.50
0
30
60
90
Observation

From the graphs, we can see that there are many influential observations. The Cook's D plot shows that there are many outliers outside of the threshold of 0.035. Some of these major outliers include 8, 10, 43, 46, 47, 48, 63, 81, 112, etc. The DFFITS plot also indicates that there are many outliers both above and below the threshold of 0.53, including 8, 47, 81, 106, 112, etc.

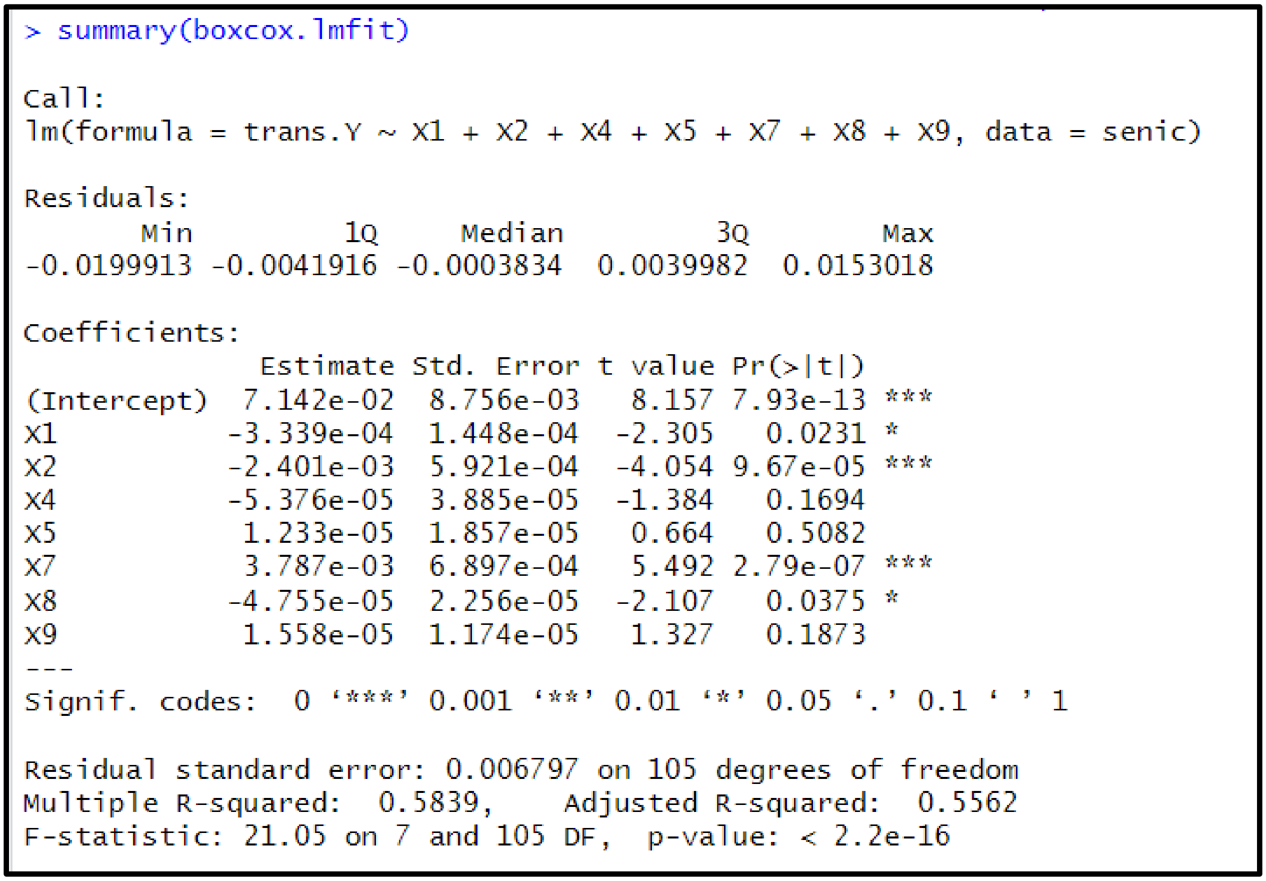
Remedial Actions/Transformations:: Based on our assumption checking, we found that our model satisfied the linearity and the homoscedasticity assumptions. Therefore we don't have to use the method of Weighted Least Squares (WLS). However, a box-cox transformation is required to make the model normal. The box-cox transformation will not correct the multicollinearity discovered earlier. Using the box-cox function in R, the lambda value needed can be determined.





```
> lambda <- boxcox.summary$lambda
> lambda
[1] -1.396303
```

With a lambda value of -1.396303, the response variable can be transformed and a new model created.

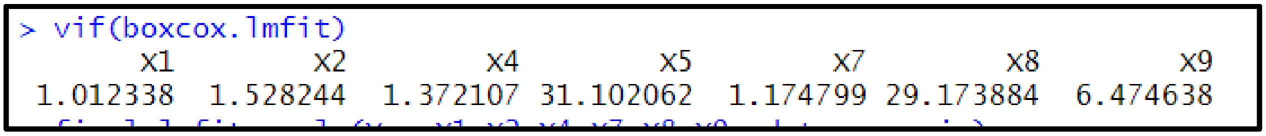


```
> summary(boxcox.lmfit)

Call:
lm(formula = trans.Y ~ X1 + X2 + X4 + X5 + X7 + X8 + X9, data = senic)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0199913 -0.0041916 -0.0003834  0.0039982  0.0153018 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.142e-02  8.756e-03   8.157 7.93e-13 ***
X1          -3.339e-04  1.448e-04  -2.305   0.0231 *  
X2          -2.401e-03  5.921e-04  -4.054 9.67e-05 ***
X4          -5.376e-05  3.885e-05  -1.384   0.1694    
X5           1.233e-05  1.857e-05   0.664   0.5082    
X7           3.787e-03  6.897e-04   5.492 2.79e-07 ***
X8          -4.755e-05  2.256e-05  -2.107   0.0375 *  
X9           1.558e-05  1.174e-05   1.327   0.1873    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.006797 on 105 degrees of freedom
Multiple R-squared:  0.5839,	Adjusted R-squared:  0.5562 
F-statistic: 21.05 on 7 and 105 DF,  p-value: < 2.2e-16
```

Fig Summary of Boxcox transformed model

To be sure that the model was truly appropriate for the data, the assumptions for linearity, normality, homoscedasticity, outlier/influential points, and multicollinearity had to be checked again with the transformed values.

Multicollinearity



```
> vif(boxcox.lmfit)
       X1        X2        X4        X5        X7        X8        X9 
 1.012338  1.528244  1.372107 31.102062  1.174799 29.173884  6.474638 
```

Fig VIF

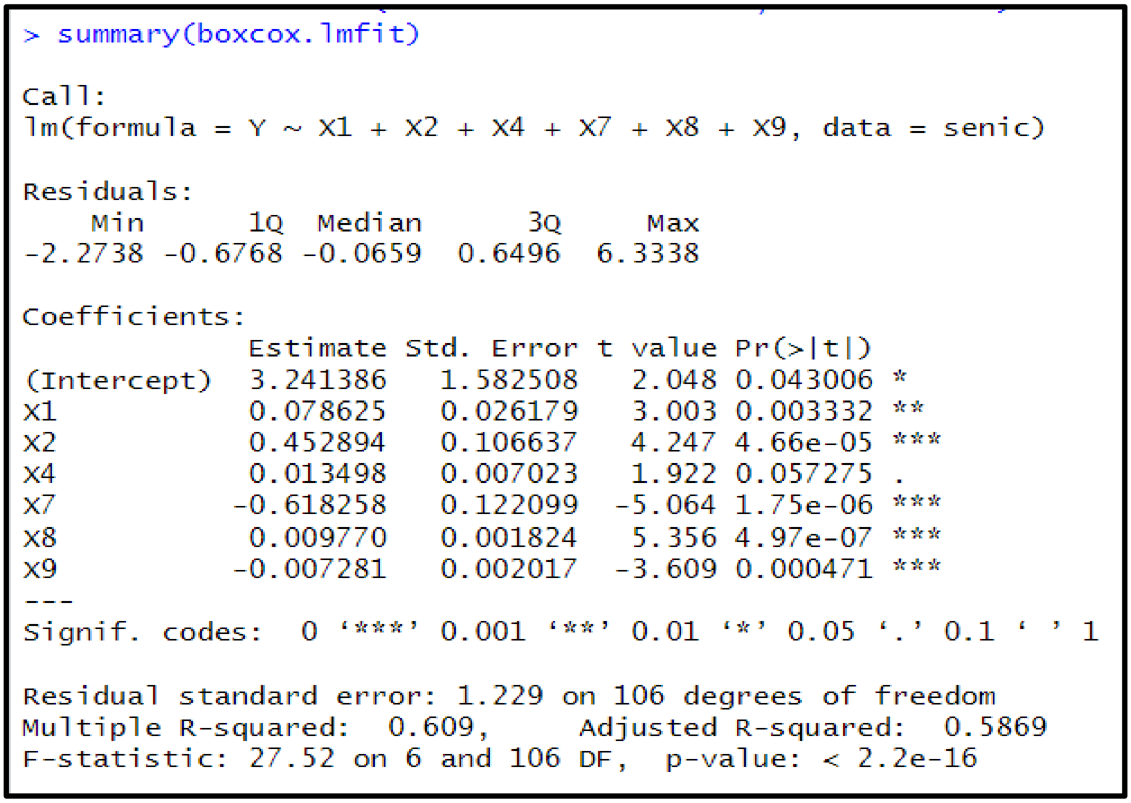
53The values obtained from the variance inflation factor analysis indicate that multicollinearity is a major problem in our final model. This is because the VIF values for X5 and X8 are more than 10. The variance inflation factors did not change between the untransformed and transformed models.

Did the transformation method work?

Our transformed regression model was able to solve the problem of normality. However, our model still has multicollinearity. So we will remove one of both (X5 and X8) as they are highly correlated variables.

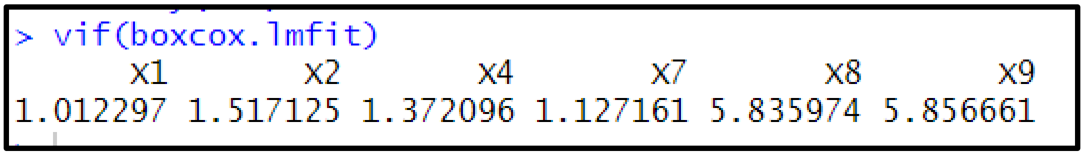
Final Model

We tried removing one of X5 And X8. Removing X5 gave best model.



```
> summary(boxcox.lmfit)

Call:
lm(formula = Y ~ X1 + X2 + X4 + X7 + X8 + X9, data = senic)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.2738 -0.6768 -0.0659  0.6496  6.3338 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.241386   1.582508   2.048 0.043006 *  
X1           0.078625   0.026179   3.003 0.003332 ** 
X2           0.452894   0.106637   4.247 4.66e-05 ***
X4           0.013498   0.007023   1.922 0.057275 .  
X7          -0.618258   0.122099  -5.064 1.75e-06 ***
X8           0.009770   0.001824   5.356 4.97e-07 ***
X9          -0.007281   0.002017  -3.609 0.000471 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.229 on 106 degrees of freedom
Multiple R-squared:  0.609,	Adjusted R-squared:  0.5869 
F-statistic: 27.52 on 6 and 106 DF,  p-value: < 2.2e-16
```

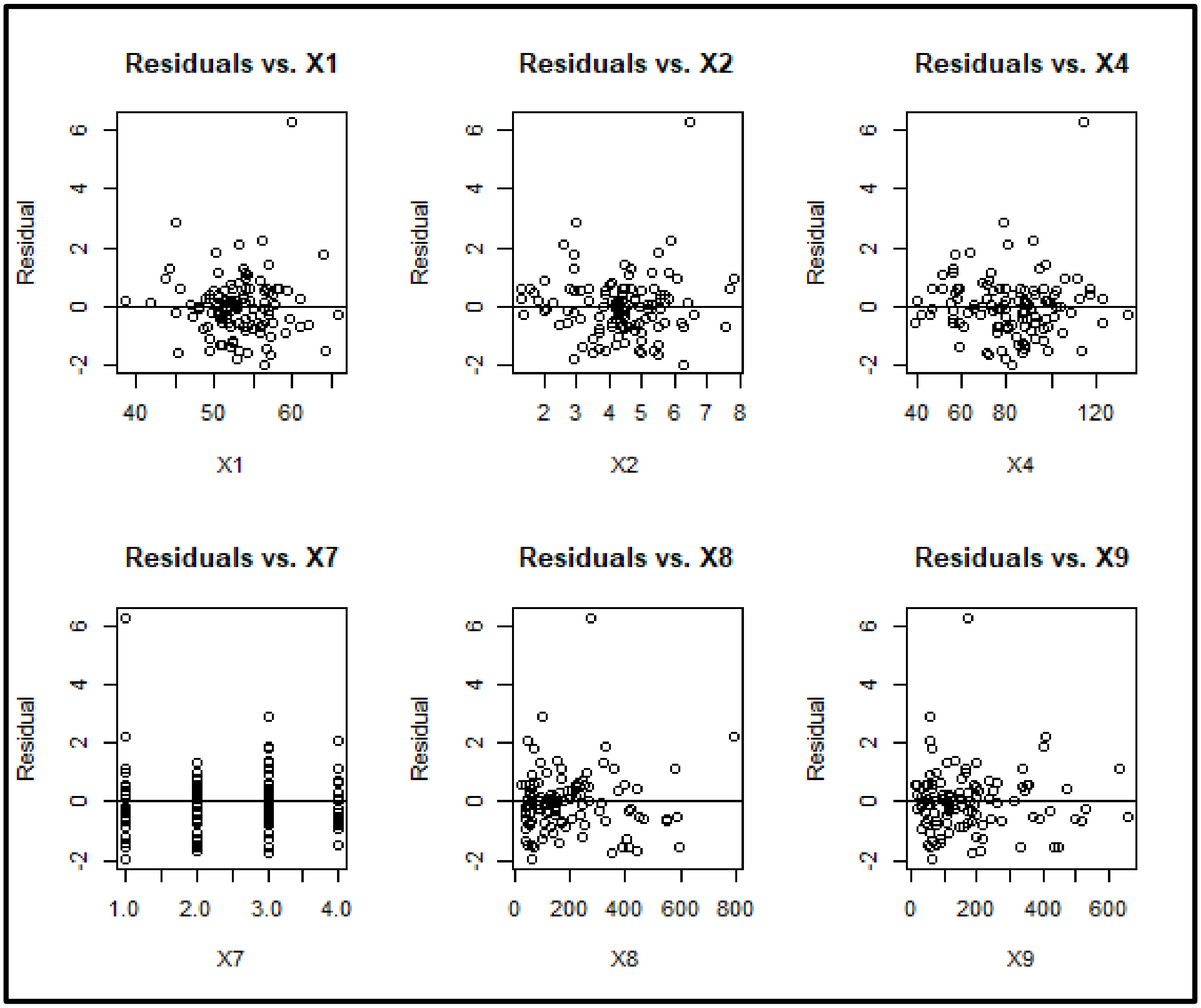
Multicollinearity

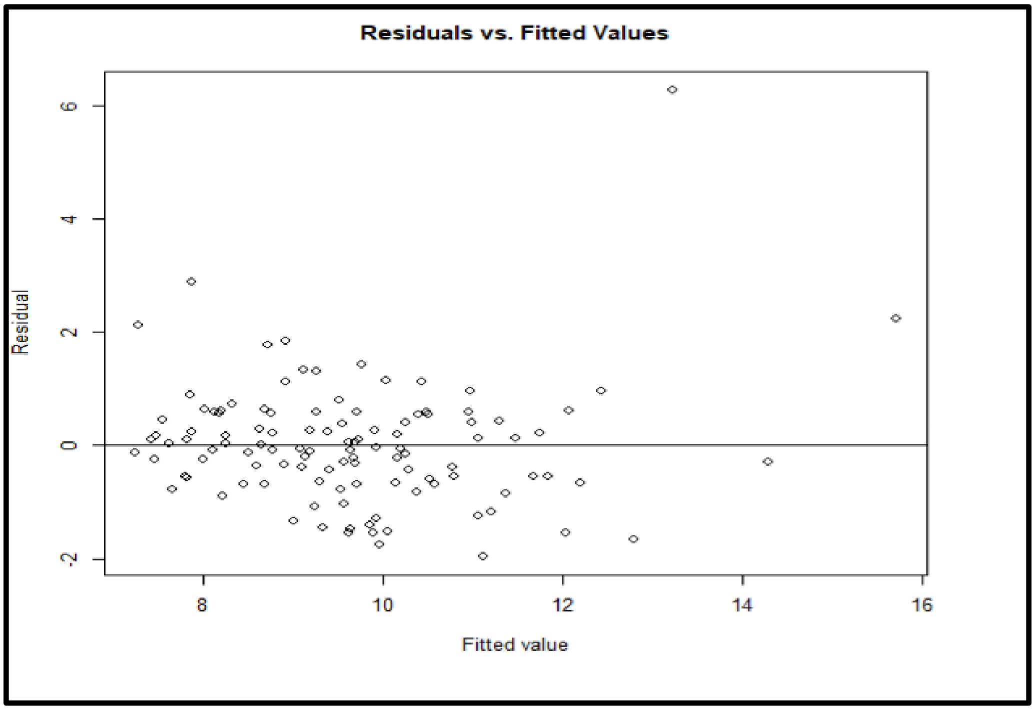


```
> vif(boxcox.lmfit)
      X1       X2       X4       X7       X8       X9 
1.012297 1.517125 1.372096 1.127161 5.835974 5.856661 
```

After removing the highly correlated variable X5, the values obtained from the variance inflation factor analysis indicate that multicollinearity is removed from our final model.

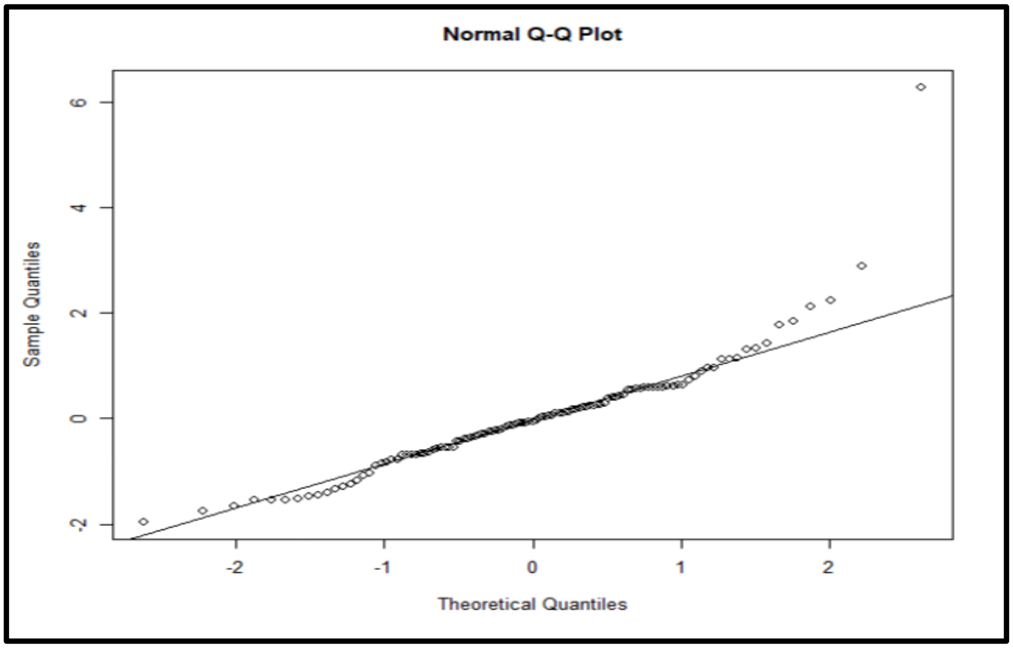
Linearity

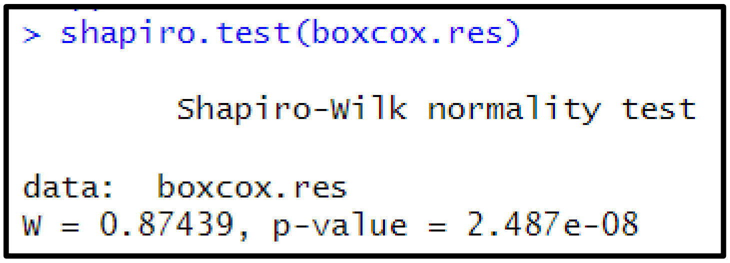

Residuals vs. X1
Residuals vs. X2
Residuals vs. X4
Residuals vs. X7
Residuals vs. X8
Residuals vs. X9
Residual
-2
0
2
4
6
40
50
60
X1
2
3
4
5
6
7
8
X2
40
60
80
120
X4
1.0
2.0
3.0
4.0
X7
0
200
400
600
800
X8
0
200
400
600
X9



To assess the linearity of our chosen transformed model we analyzed the scatter plots between the jackknifed residual and fitted values, we can see that the model meets the linearity assumption as the residuals are randomly distributed around the fitted values. Similarly, the jackknifed residual vs. predictor value plots for the transformed model also indicate that the linearity assumption is met as the residuals are randomly distributed around the predictor variables.

Normality





```
> shapiro.test(boxcox.res)

        Shapiro-Wilk normality test

data:  boxcox.res
W = 0.87439, p-value = 2.487e-08
```

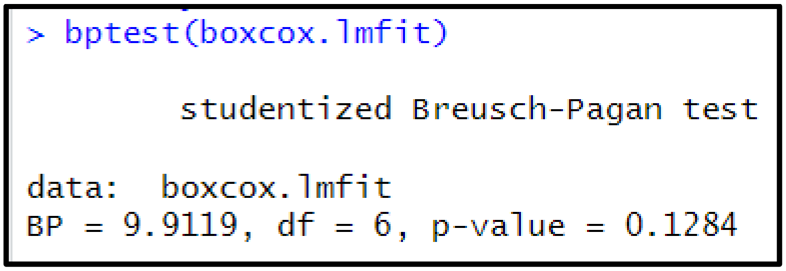
The normal probability plot of the transformed model indicates that the data nearly forms a normal distribution because the data points mostly align with only some slight deviation near the lowest and highest values.

In the Shapiro-Wilk test, the null hypothesis states that the error terms are normally distributed while the alternative hypothesis states that the error terms are not normally distributed. Our decision rule states that if the test statistic is small and the p-value is less than the significance level (alpha = 0.05), then we must reject the null hypothesis. If the test statistic is large and the p-value is greater than the significance level, we must fail to reject the null hypothesis.

For our transformed model:

We calculated a test statistic of 0.87439 and a p-value of 2.487e-08 Thus, we fail to reject the null hypothesis and conclude that the error terms are normally distributed.

Homoscedasticity Assumption



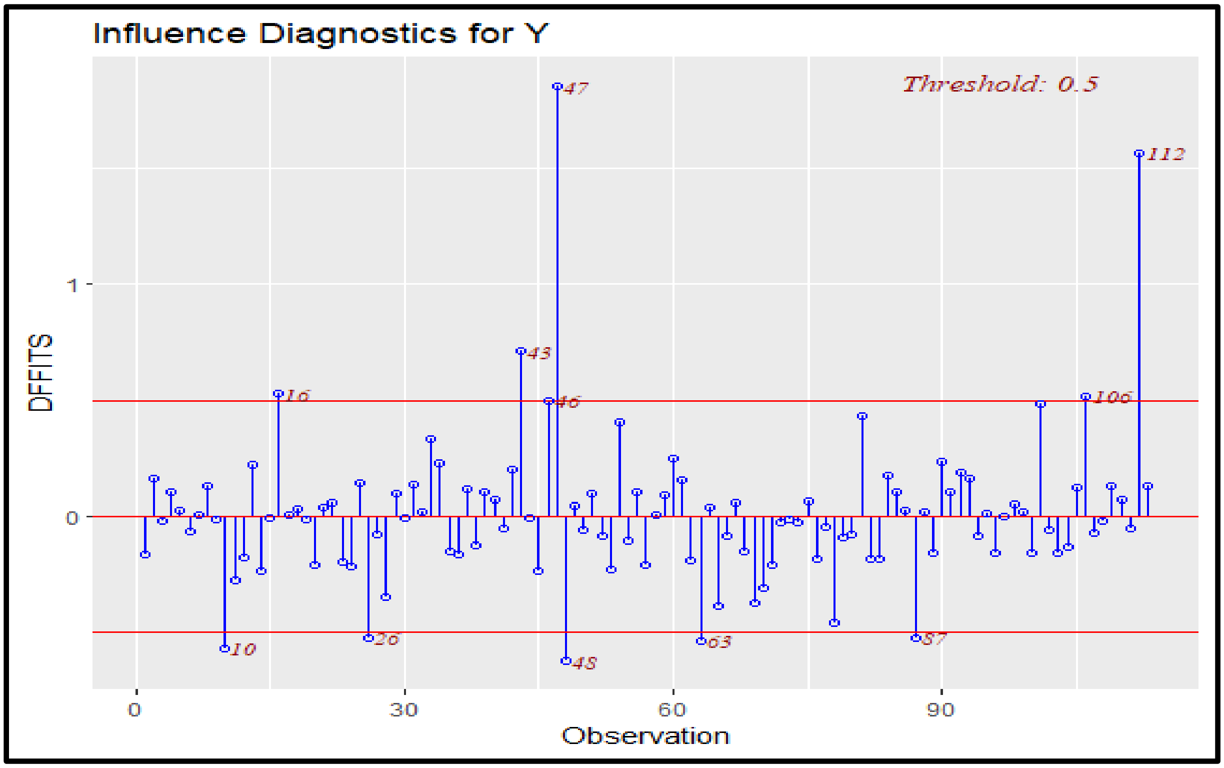
```
> bptest(boxcox.lmfit)

        studentized Breusch-Pagan test

data:  boxcox.lmfit
BP = 9.9119, df = 6, p-value = 0.1284
```

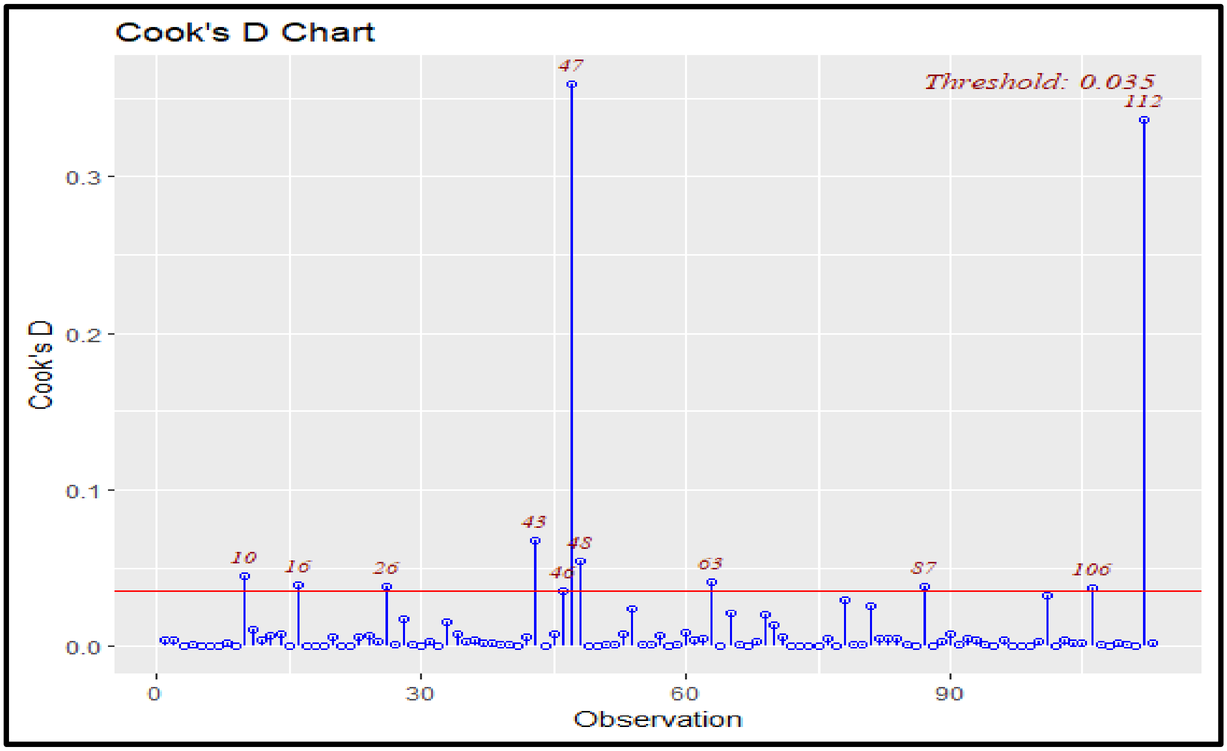
In the Breusch-Pagan test, the null hypothesis states that there is constant error variance and the alternative hypothesis states that there is no constant variance. The decision rule is that if the p-value is less than the significance level of 0.05, we will reject the null hypothesis and conclude that the error variance is not constant. If the p-value is greater than the significance level of 0.05, we will fail to reject the null hypothesis and conclude that the error variance is constant.

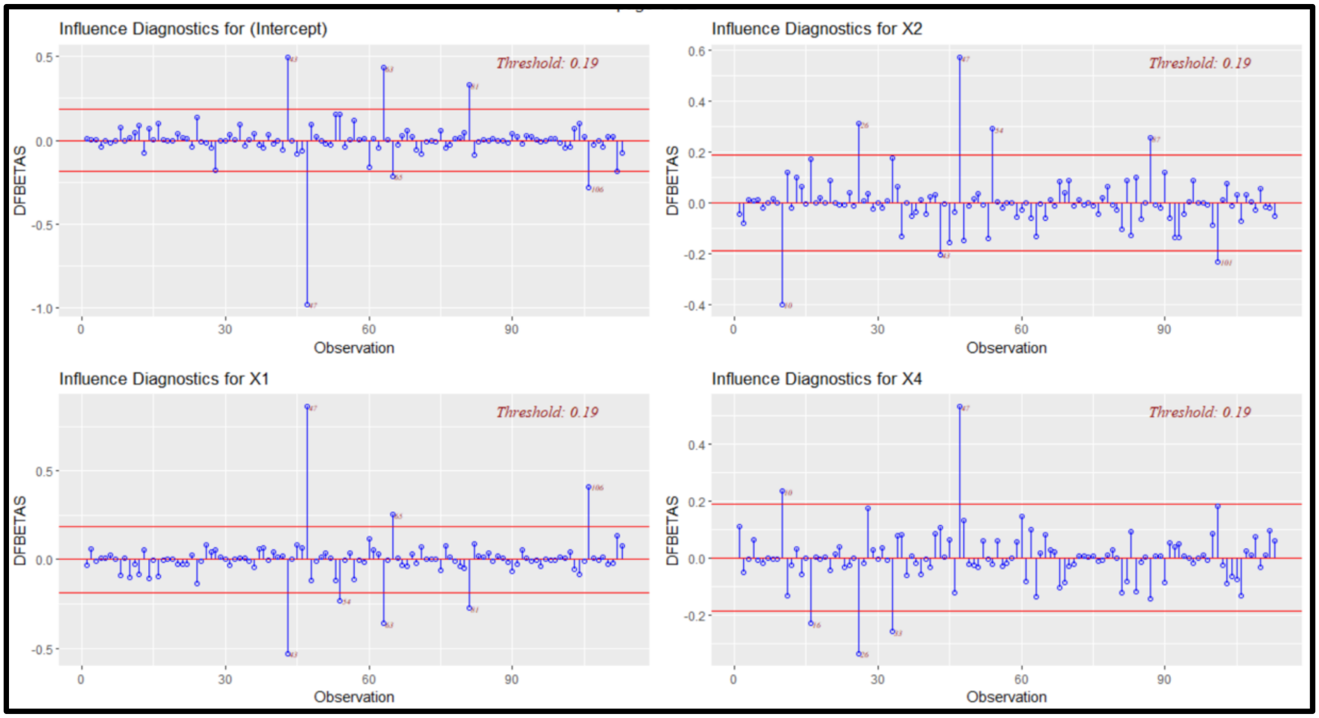
For our transformed model:

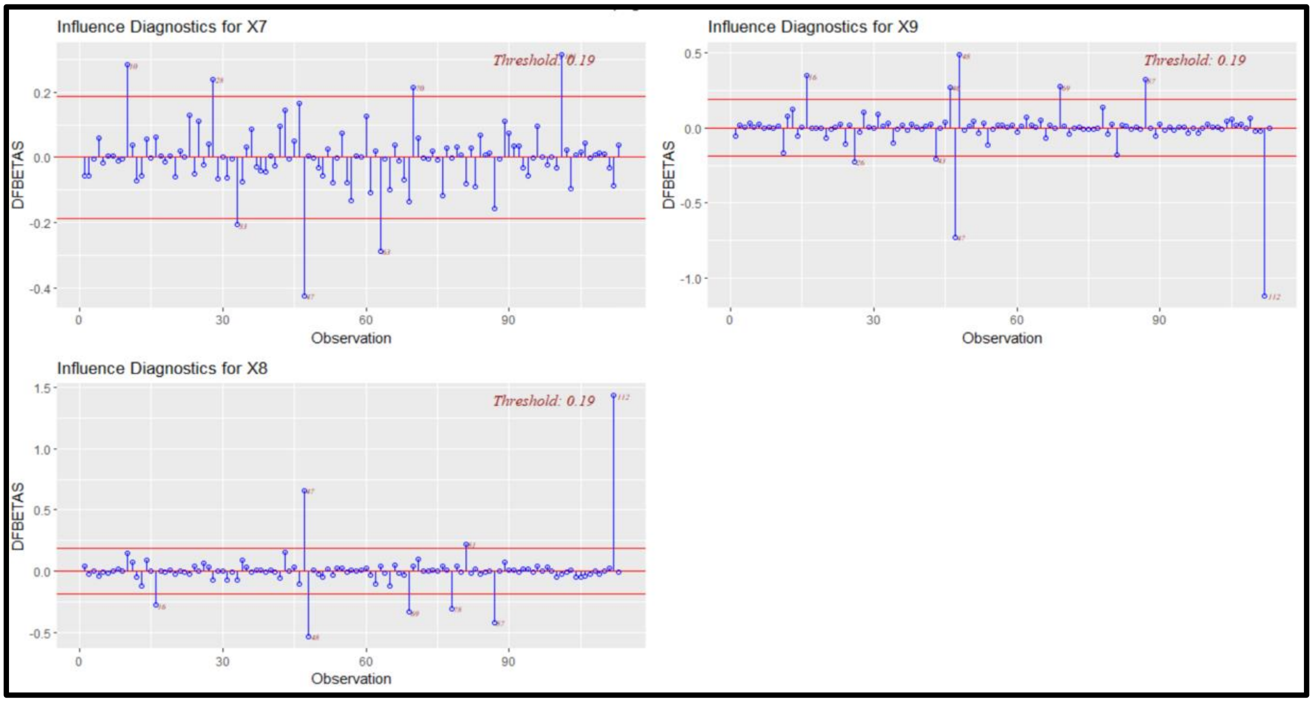
We calculated a test statistic of 9.9119 and a p-value of 0.1284 so we fail to reject the null hypothesis and conclude that the error terms are constant.

Influential Observations:



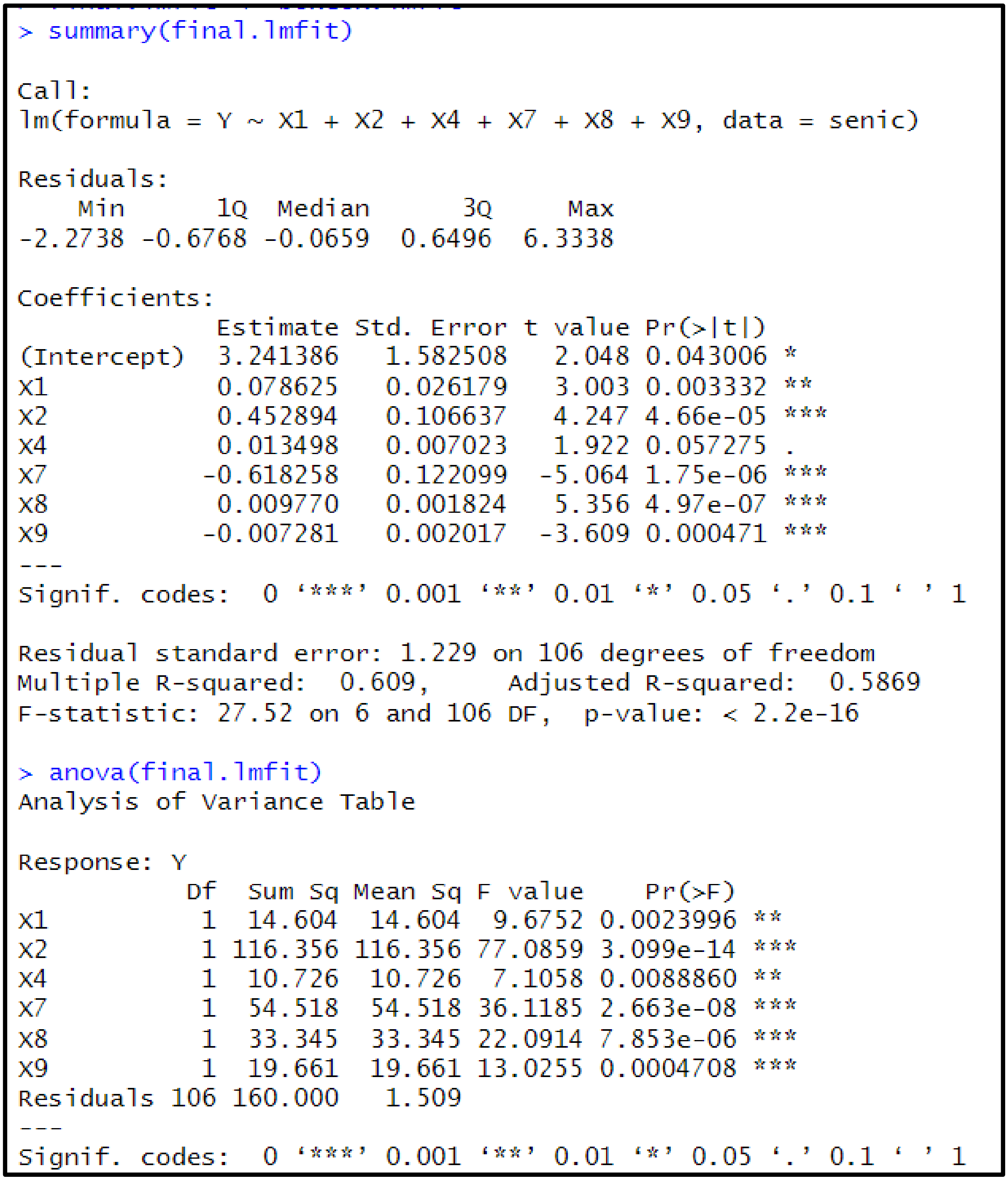



Influence Diagnostics for (Intercept)
DFBETAS
0.5
0.0
-0.5
-1.0
Threshold: 0.19
43
63
81
65
106
47
0
30
60
90
Observation
Influence Diagnostics for X2
DFBETAS
0.6
0.4
0.2
0.0
-0.2
-0.4
Threshold: 0.19
47
26
54
87
43
101
10
0
30
60
90
Observation
Influence Diagnostics for X1
DFBETAS
0.5
0.0
-0.5
Threshold: 0.19
47
106
65
54
81
63
43
0
30
60
90
Observation
Influence Diagnostics for X4
DFBETAS
0.4
0.2
0.0
-0.2
Threshold: 0.19
47
10
16
33
26
0
30
60
90
Observation

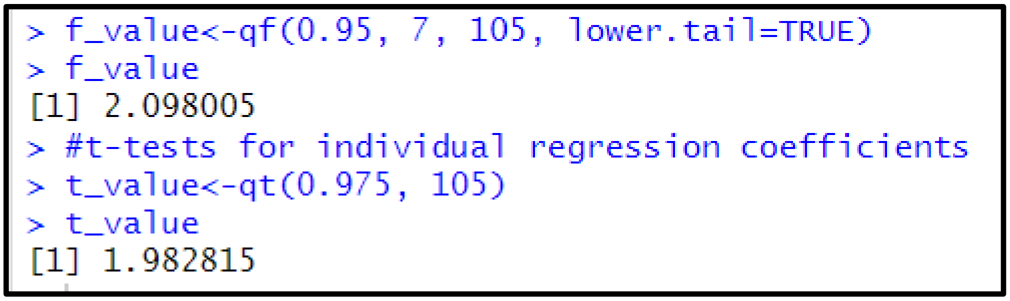

Influence Diagnostics for X7
DFBETAS
0.2
0.0
-0.2
-0.4
Threshold: 0.19
10
28
70
33
63
47
0
30
60
90
Observation
Influence Diagnostics for X9
DFBETAS
0.5
0.0
-0.5
-1.0
Threshold: 0.19
48
16
69
87
26
43
47
112
0
30
60
90
Observation
Influence Diagnostics for X8
DFBETAS
1.5
1.0
0.5
0.0
-0.5
Threshold: 0.19
112
47
81
16
69
78
87
48
0
30
60
90
Observation

From the graphs, we can see that there are many influential observations. The Cook's D plot shows that there are many outliers outside of the threshold of 0.035. Some of these major outliers include 10, 26, 43, 76, 81, 101, 106. The DFFITS plot also indicates that there are many outliers both above and below the threshold of 0.53, including 26, 43, 46, 76, 81,101, 106, etc.

Result



```
> summary(final.lmfit)

Call:
lm(formula = Y ~ X1 + X2 + X4 + X7 + X8 + X9, data = senic)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.2738 -0.6768 -0.0659  0.6496  6.3338 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.241386   1.582508   2.048 0.043006 *  
X1           0.078625   0.026179   3.003 0.003332 ** 
X2           0.452894   0.106637   4.247 4.66e-05 ***
X4           0.013498   0.007023   1.922 0.057275 .  
X7          -0.618258   0.122099  -5.064 1.75e-06 ***
X8           0.009770   0.001824   5.356 4.97e-07 ***
X9          -0.007281   0.002017  -3.609 0.000471 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.229 on 106 degrees of freedom
Multiple R-squared:  0.609,	Adjusted R-squared:  0.5869 
F-statistic: 27.52 on 6 and 106 DF,  p-value: < 2.2e-16

> anova(final.lmfit)
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)    
X1          1  14.604  14.604  9.6752 0.0023996 ** 
X2          1 116.356 116.356 77.0859 3.099e-14 ***
X4          1  10.726  10.726  7.1058 0.0088860 ** 
X7          1  54.518  54.518 36.1185 2.663e-08 ***
X8          1  33.345  33.345 22.0914 7.853e-06 ***
X9          1  19.661  19.661 13.0255 0.0004708 ***
Residuals 106 160.000   1.509                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
> f_value<-qf(0.95, 7, 105, lower.tail=TRUE)
> f_value
[1] 2.098005
> #t-tests for individual regression coefficients
> t_value<-qt(0.975, 105)
> t_value
[1] 1.982815
```

Final model:

Y = 3.241386 +0.078625 X1 +0.452894 X2 +0.013498 X4 -0.618258 X7 +0.009770 X8 - 0.007281X9.

F-test: (alpha = 0.05)

Hypothesis:

H0: All the regression coefficients = 0 vs Ha: not H0

F*= 27.52

f-value(0.95, 7, 105)= 2.098005

At a significance level of 0.05, F*>f-value, so we reject the null hypothesis and conclude that a model with the set of the seven predictor variables is a better fit than an only-intercept model.

Also, p-value= 2.2e-16 which is less than 0.05, so we reject the null hypothesis and conclude that a model with the set of the seven predictor variables is a better fit than an only-intercept model.

Adjusted R2:

The adjusted R2 value is 0.5869, which means that about 58% (more than half) of the variation in the length of the stay in hospitals can be described by our model.

Significance of individual predictors:

T-test: (alpha = 0.05)

t-value(0.975, 105) = 1.982815

We can see that the absolute value of the t-statistic for variables X4, and X9 is less than the t-value at a significance level of 0.05. Hence, we cannot conclude that they are significant in the response variable. Also,

for X1, p-value= 0.0231<0.05 so we conclude that it is significant on the response variable.

For X2, p-value= 9.67e-05<0.05 so we conclude that it is significant on the response variable.

For X4, p-value= 0.1694>0.05 so we cannot conclude that it is significant on the response variable.

For X7, p-value= 2.79e-07<0.05 so we conclude that it is significant on the response variable.

For X8, p-value= 0.0375<0.05 so we conclude that it is significant on the response variable.

For X9, p-value= 0.1873>0.05 so we cannot conclude that it is significant on the response variable.

Interpretation of Coefficients:

For unit increase in X1, the mean of probability distribution of Y changes by -3.339e-04

when X2, X4, X7, X8, X9 are held constant.

For unit increase in X2, the mean of probability distribution of Y changes by -2.401e-03

when X1, X4, X7, X8, X9 are held constant.

For unit increase in X4, the mean of the probability distribution of Y changes by -5.376e-05

when X1, X2, X7, X8, and X9 are held constant.

For unit increase in X7, the mean of probability distribution of Y changes by 3.787e-03

when X1, X2, X4,, X8, X9 are held constant.

For unit increase in X8, the mean of probability distribution of Y changes by -4.755e-05 when X1, X2, X4, X7, X9 are held constant.

For unit increase in X9, the mean of the probability distribution of Y changes by 1.558e-05 when X1, X2, X4, X7, and X8 are held constant.

Conclusion:

I initially studied our data using visualizations like histograms, boxplots, scatterplots, correlation plots, and added-variance plots before developing a linear regression model. These plots helped me comprehend the skewness of our data and the relationship between the predictors and the response.

Following that, I fitted numerous regression models with various predictor variables to determine the relevance of certain variables based on their p-values. Then I ran model selection to see which variables the stepwise regression function in R recommended we maintain for our linear model. I kept seven predictor variables for the final model based on the AIC, BIC, Adj. R2, and Mallow's CP values.

I next examined the assumptions of this chosen model and discovered that it was not normal. As a result, I conducted a Box-Cox transformation on the altered model and questioned the assumptions. Overall, I found that the modified model was normal. According to the hypothesis test, all seven predictors are significant for the model.

With an adjusted R2 of 0.5869, I picked this changed model as my final model. Based on their age, infection risk, area, the routine x-ray, average census, and the number of nurses and beds, this model may be used to forecast the duration of stay of patients in hospitals.

APPENDIX

```
############################## Read a data into R ##############################
senic <- read.csv("C:/Users/ravih/downloads/SENIC.csv", header=TRUE)
head(senic)

# Check for missing values in the entire dataset
missing_values <- sum(is.na(senic))

# Display the number of missing values
cat("Number of missing values in the dataset:", missing_values, "\n")
### No missing values



##### Analysing response variable
par(mfrow= c(1,1))
hist(senic$Y)
boxplot(senic$Y)



########################### 1. Introduction ###################################
###################### 1.1 Exploratory Data Analysis. ##########################
## 1. Histograms of Y and Xs
library(dplyr)



par(mfrow= c(3,4))
for (col in c(names(senic))){
  senic %>% pull(col) %>% hist(main= col)
}


library(e1071)
skew_summary <- sapply(senic, function(x) skewness(x))
skew_summary


## 2. Boxplots of Y and Xs

par(mfrow= c(3,4))
for (col in c(names(senic))){
  senic %>% pull(col) %>% boxplot(main= col)
}


## 3. Summary Statistics
summary(senic)



## 4. Scatter Plot Matrix
pairs(senic, col= "#FF1493E2", main = "Scatter-Plot matrix of SENIC data")

par(mfrow=c(3,4))
plot(Y~X1, senic,col="blue", main="Scatter-Plot between Y and X1")
plot(Y~X2, senic,col="blue", main="Scatter-Plot between Y and X2")
plot(Y~X3, senic,col="blue", main="Scatter-Plot between Y and X3")
plot(Y~X4, senic,col="blue", main="Scatter-Plot between Y and X4")
plot(Y~X5, senic,col="blue", main="Scatter-Plot between Y and X5")
plot(Y~X6, senic,col="blue", main="Scatter-Plot between Y and X6")
plot(Y~X7, senic,col="blue", main="Scatter-Plot between Y and X7")
plot(Y~X8, senic,col="blue", main="Scatter-Plot between Y and X8")
plot(Y~X9, senic,col="blue", main="Scatter-Plot between Y and X9")
plot(Y~X10, senic,col="blue", main="Scatter-Plot between Y and X10")


## 5. Added-Variable Plots
library(car)
dev. off()
senic.lmfit <- lm(Y ~ X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data = senic)
avPlots(senic.lmfit)


## 6. correlation matrix
library(caret)
library(corrplot)

#dev.new()
correlation_matrix <- cor(senic)
correlation_matrix
# Create a correlation plot
corrplot(correlation_matrix, method = "circle", diag = TRUE, tl.cex = 0.8)
```

```
###########################2. Model/Methods ###################################
## Fit a regression model with all of the predictors
full.lmfit <- lm(Y ~ X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data = senic)
summary(full.lmfit)

anova(full.lmfit)



## Model Selection (Stepwise Regression)
# Install packages for the model selection
# install.packages("leaps")
# install.packages("HH")
# install.packages("StepReg")
# Load HH, leaps, and StepReg packages
library(leaps)
library(HH)
library(StepReg)



#### Stepwise Regression
par(mfrow=c(3,2))
library(olsrr)
b<- ols_step_all_possible(full.lmfit )
plot(b)



b.adjr = data.frame(n=b$n,predictors=b$predictors,adjr=b$adjr)
#print(b.adjr)
```

```
print(b.adjr[c(638, 848, 968, 1013, 1023),])



b.cp = data.frame(n=b$n,predictors=b$predictors,cp=b$cp)
#print(b.cp)
print(b.cp[c(638, 848, 968, 1013, 1023),])




b.aic = data.frame(n=b$n,predictors=b$predictors,aic=b$aic)
#print(b.aic)
print(b.aic[c(638, 848, 968, 1013, 1023),])



b.press = data.frame(n=b$n,predictors=b$predictors,press=b$msep)
#print(b.press)
print(b.press[c(638, 848, 968, 1013, 1023),])



k <- ols_step_both_p(full.lmfit,pent=0.10,prem=0.1,details=TRUE)
plot(k)



# #### Checking for correlated variables
# senic_data <- cbind(senic$Y, x1, x2, x4, x5, x7, x8, x9)
# senic_data <- as.data.frame(senic_data)
#
# # Create a correlation plot
# dev.new()
# corrplot(cor_matrix, method = "circle", diag = TRUE, tl.cex = 0.8)

library(dplyr)
Cols<- c("X3", "X6", "X10")
senic<-senic[, -which(names(senic) %in% Cols)]

cor_matrix <- cor(senic)
cor_matrix
corrplot(cor_matrix, method = "circle", diag = TRUE, tl.cex = 0.8)



# # As we can see, lots of variables are highly correlated.
# #Therefore Standardization needed for our variables.
#
# x1 <- (senic$X1 -mean(senic$X1))/sd(senic$X1)
# x2 <- (senic$X2 -mean(senic$X2))/sd(senic$X2)
# x3 <- (senic$X3 -mean(senic$X3))/sd(senic$X3)
# x4 <- (senic$X4 -mean(senic$X4))/sd(senic$X4)
# x5 <- (senic$X5 -mean(senic$X5))/sd(senic$X5)
# x6 <- (senic$X6 -mean(senic$X6))/sd(senic$X6)
# x7 <- (senic$X7 -mean(senic$X7))/sd(senic$X7)
# x8 <- (senic$X8 -mean(senic$X8))/sd(senic$X8)
# x9 <- (senic$X9 -mean(senic$X9))/sd(senic$X9)

# #Now very few of them are highly correlated to each other compare to without standardization.
# senic.itact.Std <- cbind(senic$Y,x1,x2,x4,x5,x7,x8,x9)
# senic.itact.Std <- as.data.frame(senic.itact.Std) # Converting to data Frame.
# head(senic.itact.Std)
# colnames(senic.itact.Std)[1] <- "Y"


################ Fit a reduced regression model
# reduced.lmfit <- lm(Y ~ x1+x2+x4+x5+x7+x8+x9, data=senic.itact.Std)
# summary(reduced.lmfit)

reduced.lmfit <- lm(Y ~ X1+X2+X4+X5+X7+X8+X9, data=senic)
summary(reduced.lmfit)



########3 Regression Diagnostics
res <- rstudent(reduced.lmfit)
fitted.y <- fitted(reduced.lmfit)



######## Residual Plots ##########
par(mfrow=c(2,4))
plot(res ~ senic$X1, xlab="X1", ylab="Residual", main="Residuals vs. X1")
abline(h=0)
plot(res ~ senic$X2, xlab="X2", ylab="Residual", main="Residuals vs. X2")
abline(h=0)
plot(res ~ senic$X4, xlab="X4", ylab="Residual", main="Residuals vs. X4")
abline(h=0)
plot(res ~ senic$X5, xlab="X5", ylab="Residual", main="Residuals vs. X5")
abline(h=0)
plot(res ~ senic$X7, xlab="X7", ylab="Residual", main="Residuals vs. X7")
```

```
abline(h=0)
plot(res ~ senic$X8, xlab="X8", ylab="Residual", main="Residuals vs. X8")
abline(h=0)
plot(res ~ senic$X9, xlab="X9", ylab="Residual", main="Residuals vs. X9")
abline(h=0)
plot(res ~ fitted.y, xlab="Fitted value", ylab="Residual", main="Residuals vs. Fitted Values")
abline(h=0)

######### Normality ###########
qqnorm(res);
qqline(res, col= "red")
shapiro.test(res)



######### Constancy of Error Variances #########
library(lmtest)
bptest(reduced.lmfit)

######### Multicollinearity ##########
vif(reduced.lmfit)

########## performing transformations as we have high multicollinearity
install.packages("EnvStats")
library(EnvStats)

boxcox.summary <- boxcox(reduced.lmfit, optimize=TRUE)
lambda <- boxcox.summary$lambda
lambda
```

```
trans.Y <- senic$Y^lambda

senic <- cbind(senic,trans.Y)
senic


######### Re-fitting a model using the transformed response variable. ##########
boxcox.lmfit <- lm(trans.Y ~ X1 + X2 + X4 + X5+ X7+ X8 +X9, data=senic)
summary(boxcox.lmfit)

boxcox.res <- rstudent(boxcox.lmfit)

boxcox.fitted.y <- fitted(boxcox.lmfit)


############ Checking if transformation decreased the multicollinearity problem
library(car)
vif(boxcox.lmfit)


####### Transformation didn't decrease the multi collinearity problem
## So removing highly correlated variable X5 and fitting the model again

boxcox.lmfit <- lm(Y ~ X1+X2+X4+X7+X8+X9, data = senic)
summary(boxcox.lmfit)
```

```
############ Now check the multicollinearity
library(car)
vif(boxcox.lmfit)

#### Stepwise Regression
par(mfrow=c(3,3))
library(olsrr)
b<- ols_step_all_possible(full.lmfit )
plot(b)


b.adjr = data.frame(n=b$n,predictors=b$predictors,adjr=b$adjr)
print(b.adjr)
print(b.adjr[c(256, 382, 466, 502, 511),])


b.cp = data.frame(n=b$n,predictors=b$predictors,cp=b$cp)
print(b.cp)
print(b.cp[c(256, 382, 466, 502, 511),])
b.aic = data.frame(n=b$n,predictors=b$predictors,aic=b$aic)
print(b.aic)
print(b.aic[c(256, 382, 466, 502, 511),])


b.press = data.frame(n=b$n,predictors=b$predictors,press=b$msep)
print(b.press)
print(b.press[c(256, 382, 466, 502, 511),])
```

```
k <- ols_step_both_p(full.lmfit,pent=0.10,prem=0.1,details=TRUE)
plot(k)


####### Fitting the reduced model
reduced.lmfit <- lm(Y ~ x1+x2+x4+x7+x8+x9, data=senic)
summary(reduced.lmfit)


########## Checking if removing x5 variable removed multi collinearity
vif(reduced.lmfit)
###### yes it did

################  MOdel Diagnostics
######## Residual Plots ##########

final.lmfit <- boxcox.lmfit
summary(final.lmfit)
anova(final.lmfit)


boxcox.res <- rstudent(boxcox.lmfit)

boxcox.fitted.y <- fitted(boxcox.lmfit)

par(mfrow=c(2,3))
plot(boxcox.res ~ senic$X1, xlab="X1", ylab="Residual", main="Residuals vs. X1")
```

```
abline(h=0)

plot(boxcox.res ~ senic$X2, xlab="X2", ylab="Residual", main="Residuals vs. X2")

abline(h=0)

plot(boxcox.res ~ senic$X4, xlab="X4", ylab="Residual", main="Residuals vs. X4")

abline(h=0)

plot(boxcox.res ~ senic$X7, xlab="X7", ylab="Residual", main="Residuals vs. X7")

abline(h=0)

plot(boxcox.res ~ senic$X8, xlab="X8", ylab="Residual", main="Residuals vs. X8")

abline(h=0)

plot(boxcox.res ~ senic$X9, xlab="X9", ylab="Residual", main="Residuals vs. X9")

abline(h=0)

plot(boxcox.res ~ boxcox.fitted.y, xlab="Fitted value", ylab="Residual", main="Residuals vs. Fitted Values")

abline(h=0)



######### Normality ###########

qqnorm(boxcox.res);

qqline(boxcox.res)

shapiro.test(boxcox.res)


######### Constancy of Error Variances #########

library(lmtest)

bptest(boxcox.lmfit)


# 1. DFFITS

ols_plot_dffits(boxcox.lmfit)
```

```
# 2. Cook's D
ols_plot_cooksd_chart(boxcox.lmfit)


# 3. DFBETAS
ols_plot_dfbetas(boxcox.lmfit)


######### Multicollinearity ##########
library(car)
vif(boxcox.lmfit)
```