# Regression Analysis Final Project

**(MA 4710)**

**Udit Narayan Joshi**
**Michigan Technological University**
**December 2021**

# Table of Contents

# 1. Introduction

The primary objective of the Study on the Efficacy of Nosocomial Infection Control (SENIC Project) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial hospital-acquired) infection in United States hospitals. This data set consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed. The data presented here are for the 1975-76 study period.
The dataset contains 113 samples (rows) and 11 variables.

## *1.1 Description of the variables in the dataset (in ascending order):*

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

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

• **Infection Risk (X2)**
The average estimated probability of acquiring infection in 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 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)**
Percent of 35 potential facilities and services that are provided by the hospital.

## 1.2 *The Goal of Data Analysis:*

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.

The significance value (i.e., alpha) will be used as 0.10 throughout this analysis.

## *1.3 Exploratory Data Analysis:*

### *1.3.1 Histograms of Y and Xs:*

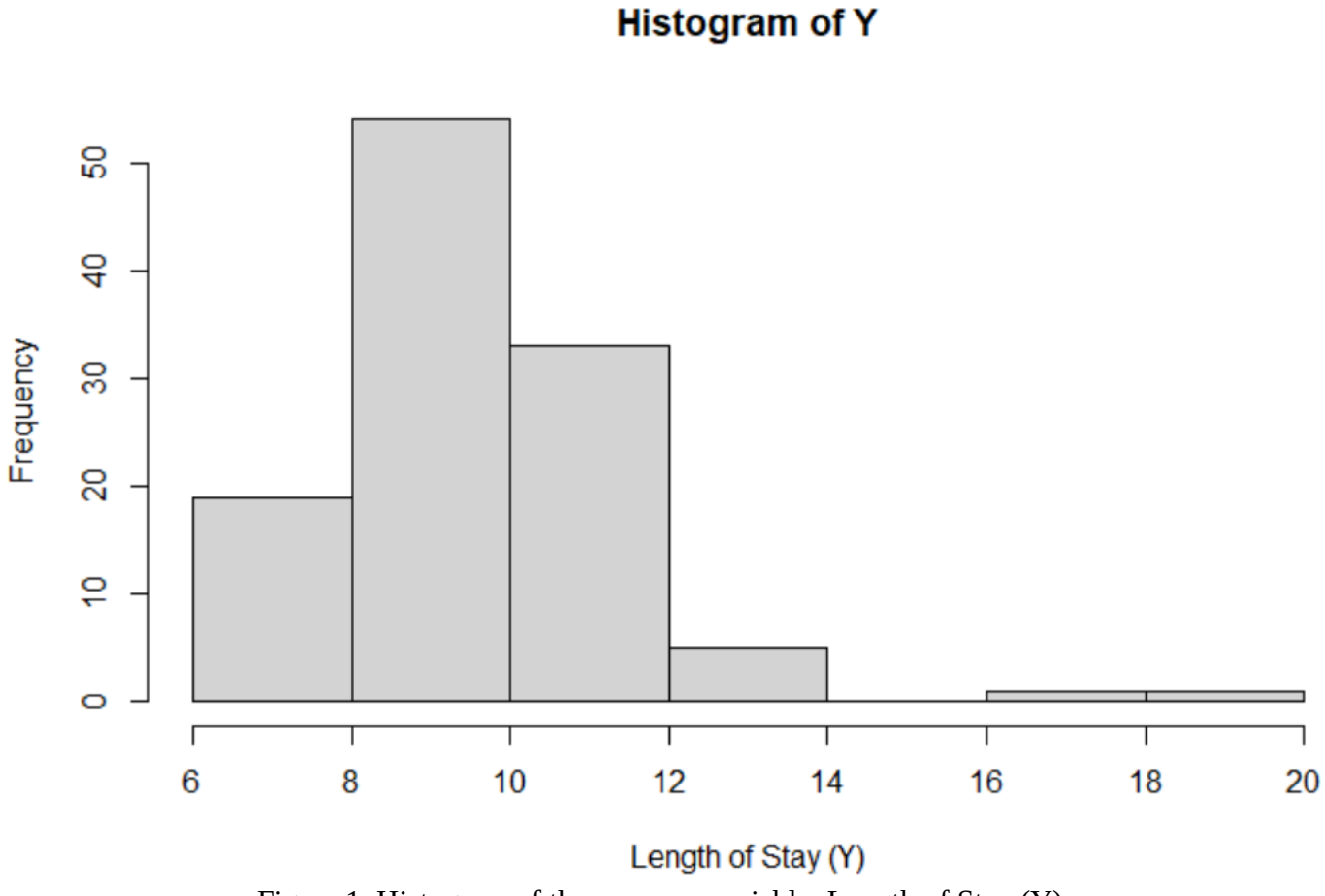



Figure 1. Histogram of the response variable; Length of Stay (Y).

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

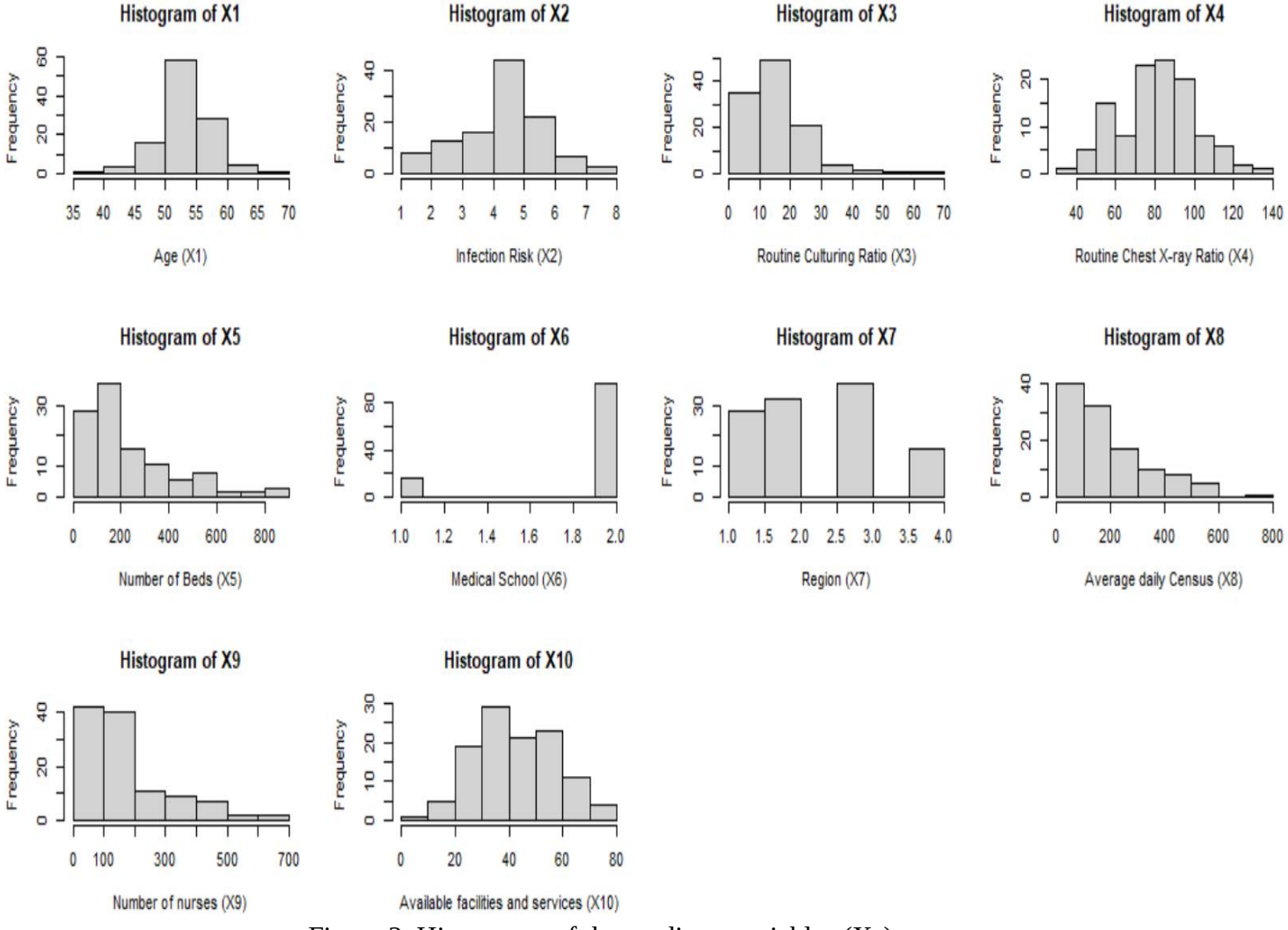



Figure 2. Histograms of the predictor variables (Xs).

.

The histogram for our predictor variable, which is displayed in the above figure 2 is having the right-skewed distribution for the Routine Culturing Ratio (X3), Number of Beds (X5), Average daily Census (X8), Number of nurses (X9). And the other columns such as Age (X1), Infection Risk (X2), Routine Chest X-ray Ratio (X4), and Available facilities and services (X10) are normally distributed.
Apart from these variables, we have two categorical variables as Medical School (X6) and Region (X7)

### 1.3.2 *Boxplots of Y and Xs:*

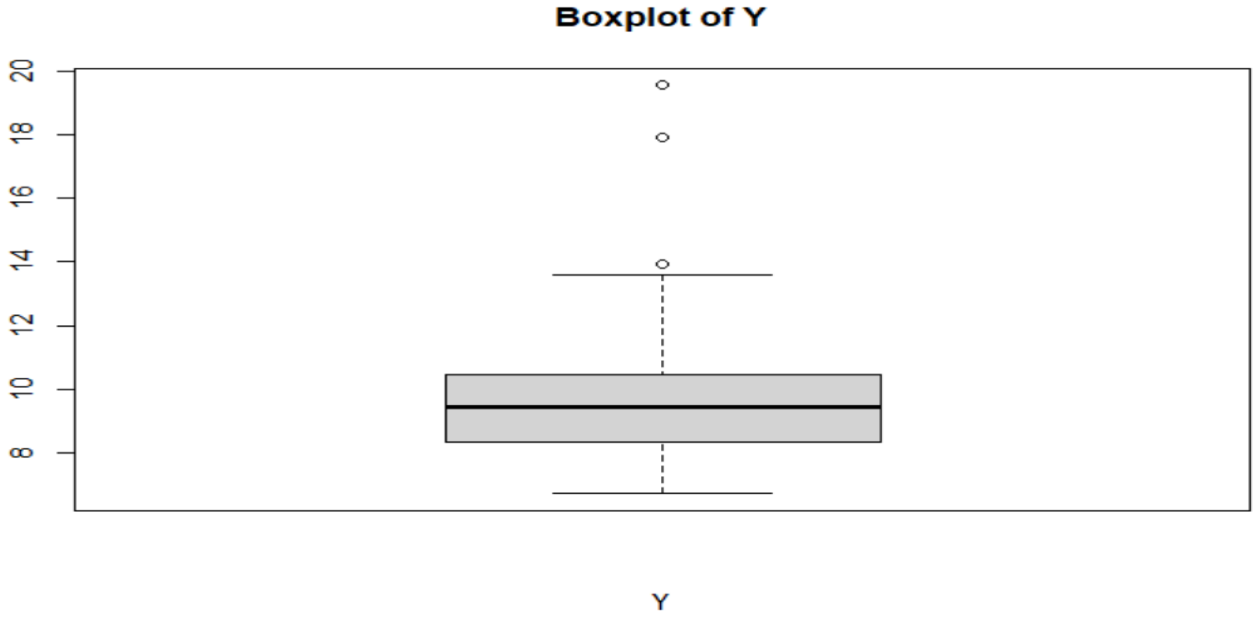



Figure 3. Boxplot of the response variable; Length of Stay (Y).

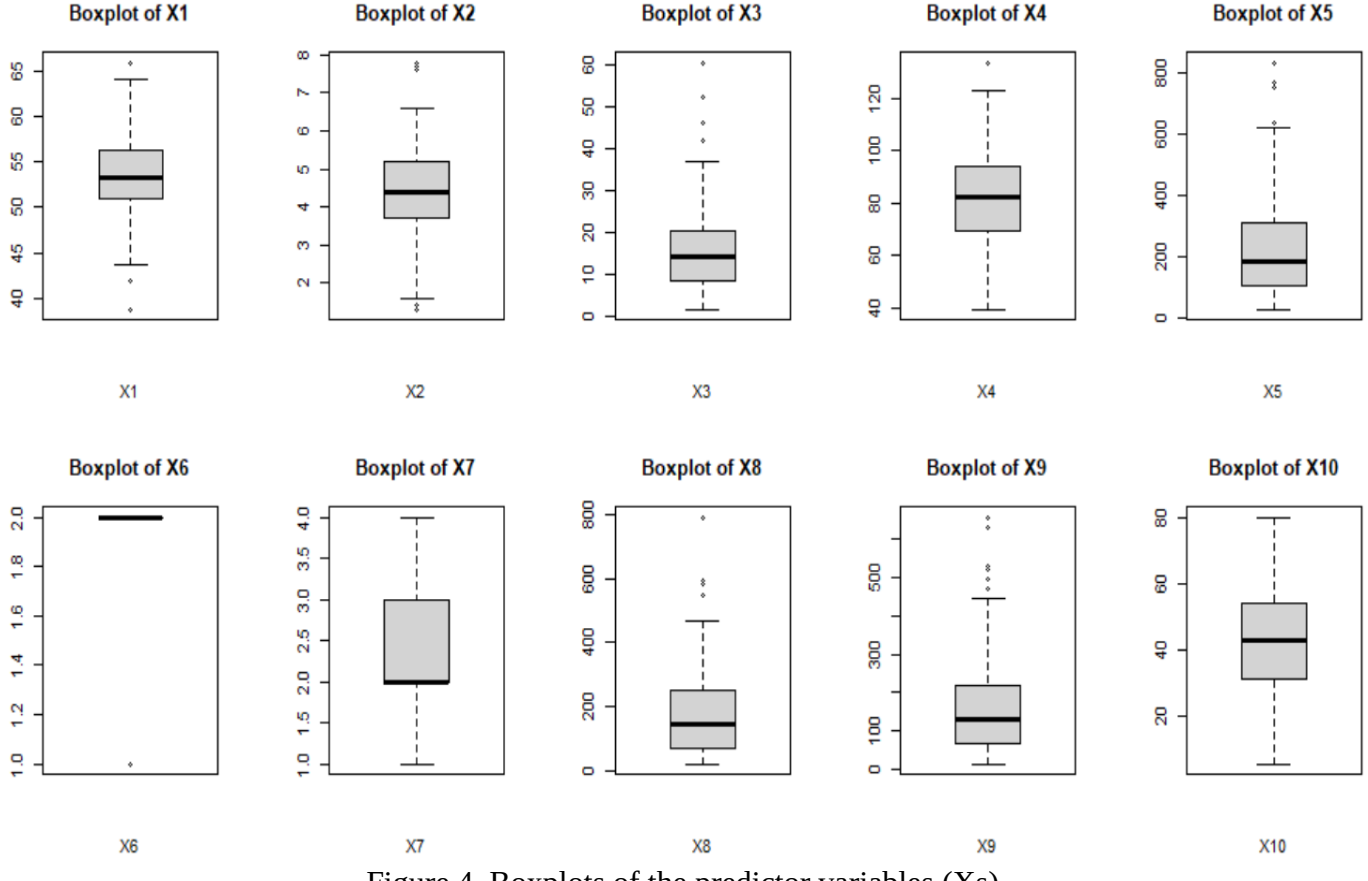



Figure 4. Boxplots of the predictor variables (Xs).

From Figure 3, we can say that the boxplot of the response variable is not centered because one side of the median is larger than the other side and it contains a few outliers also. This indicates that the response variable is not normally distributed.

From Figure 4, 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 not normally distributed.
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 so we can say that they all are normally distributed. And we have two categorical variables as Medical School (X6) and Region (X7) so we can't say anything about it.

### *1.3.3 Summary Statistics:*

| Y | X1 | X2 | X3 | X4 | X5 |
|---|---|---|---|---|---|
| Min. : 6.700 | Min. :38.80 | Min. :1.300 | Min. : 1.60 | Min. : 39.60 | Min. : 29.0 |
| 1st Qu.: 8.340 | 1st Qu.:50.90 | 1st Qu.:3.700 | 1st Qu.: 8.40 | 1st Qu.: 69.50 | 1st Qu.:106.0 |
| Median : 9.420 | Median :53.20 | Median :4.400 | Median :14.10 | Median : 82.30 | Median :186.0 |
| Mean : 9.648 | Mean :53.23 | Mean :4.355 | Mean :15.79 | Mean : 81.63 | Mean :252.2 |
| 3rd Qu.:10.470 | 3rd Qu.:56.20 | 3rd Qu.:5.200 | 3rd Qu.:20.30 | 3rd Qu.: 94.10 | 3rd Qu.:312.0 |
| Max. :19.560 | Max. :65.90 | Max. :7.800 | Max. :60.50 | Max. :133.50 | Max. :835.0 |

| X6 | X7 | X8 | X9 | X10 |
|---|---|---|---|---|
| Min. :1.00 | Min. :1.000 | Min. : 20.0 | Min. : 14.0 | Min. : 5.70 |
| 1st Qu.:2.00 | 1st Qu.:2.000 | 1st Qu.: 68.0 | 1st Qu.: 66.0 | 1st Qu.:31.40 |
| Median :2.00 | Median :2.000 | Median :143.0 | Median :132.0 | Median :42.90 |
| Mean :1.85 | Mean :2.363 | Mean :191.4 | Mean :173.2 | Mean :43.16 |
| 3rd Qu.:2.00 | 3rd Qu.:3.000 | 3rd Qu.:252.0 | 3rd Qu.:218.0 | 3rd Qu.:54.30 |
| Max. :2.00 | Max. :4.000 | Max. :791.0 | Max. :656.0 | Max. :80.00 |

Based on the Summary statistics, we can say that the predictor variable Age (X1) indicates no information about the children or very elderly persons are included like the minimum age is 38 and maximum age is 66 only. We can also see that the Average Census(X8) takes a wide range of values from 20.0 to 791.0 - the difference in the median and mean indicates that there may be some outliers in X8. And in the same way Number of Beds (X5), Number of nurses (X9), and Available facilities and services (X10) have a very wide range so there will be some potential outlier that exists.

## *1.3.4 Scatter Plot Matrix:*

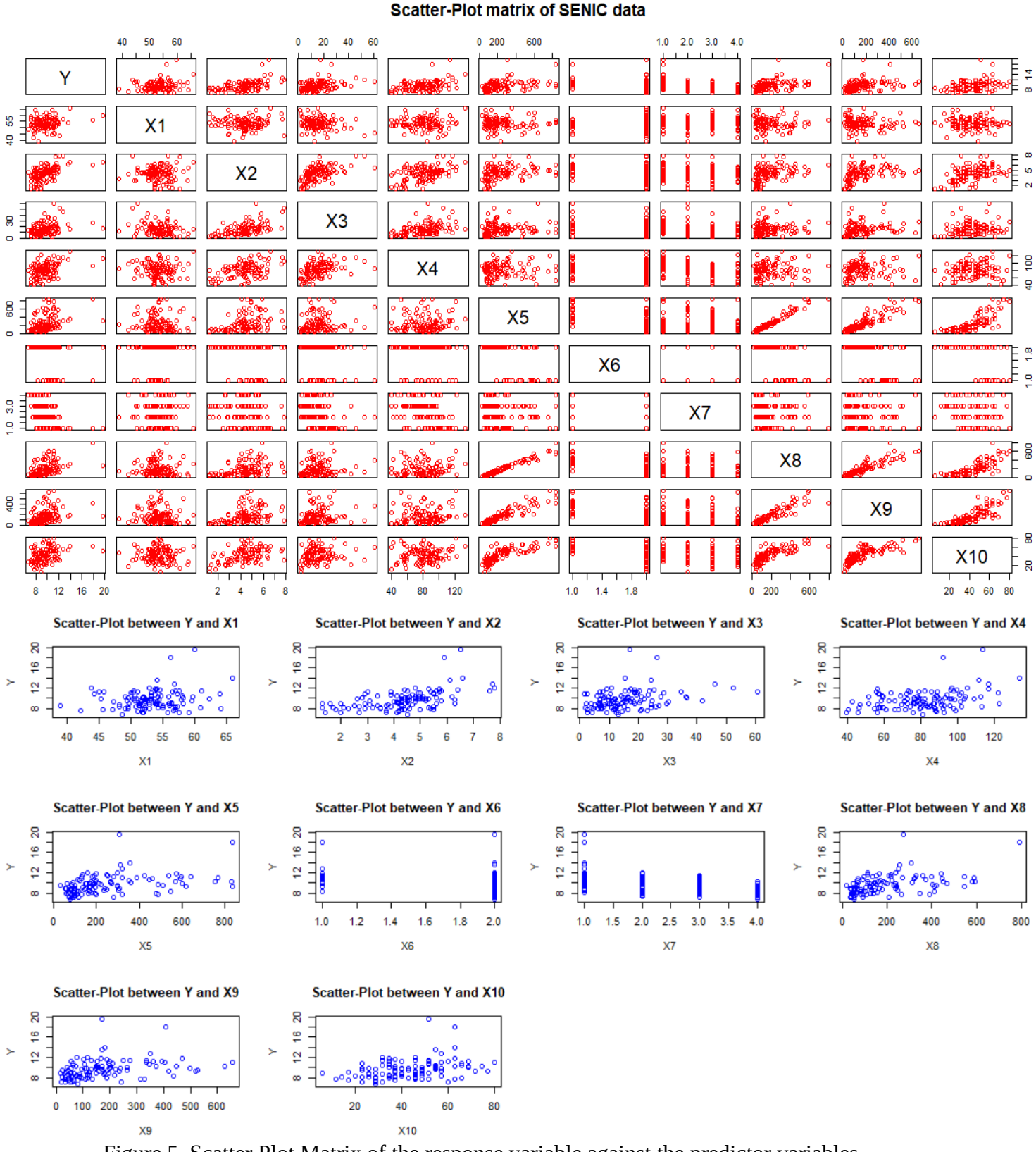



Figure 5. Scatter Plot Matrix of the response variable against the predictor variables.

Based on the Scatter plot (Figure 5), we can say that the all-predictor variables are having a kind of linear relationship with the response variable apart from the two categorical variables i.e., Medical School (X6) and Region (X7).

### *1.3.5 Added-Variable Plots:*

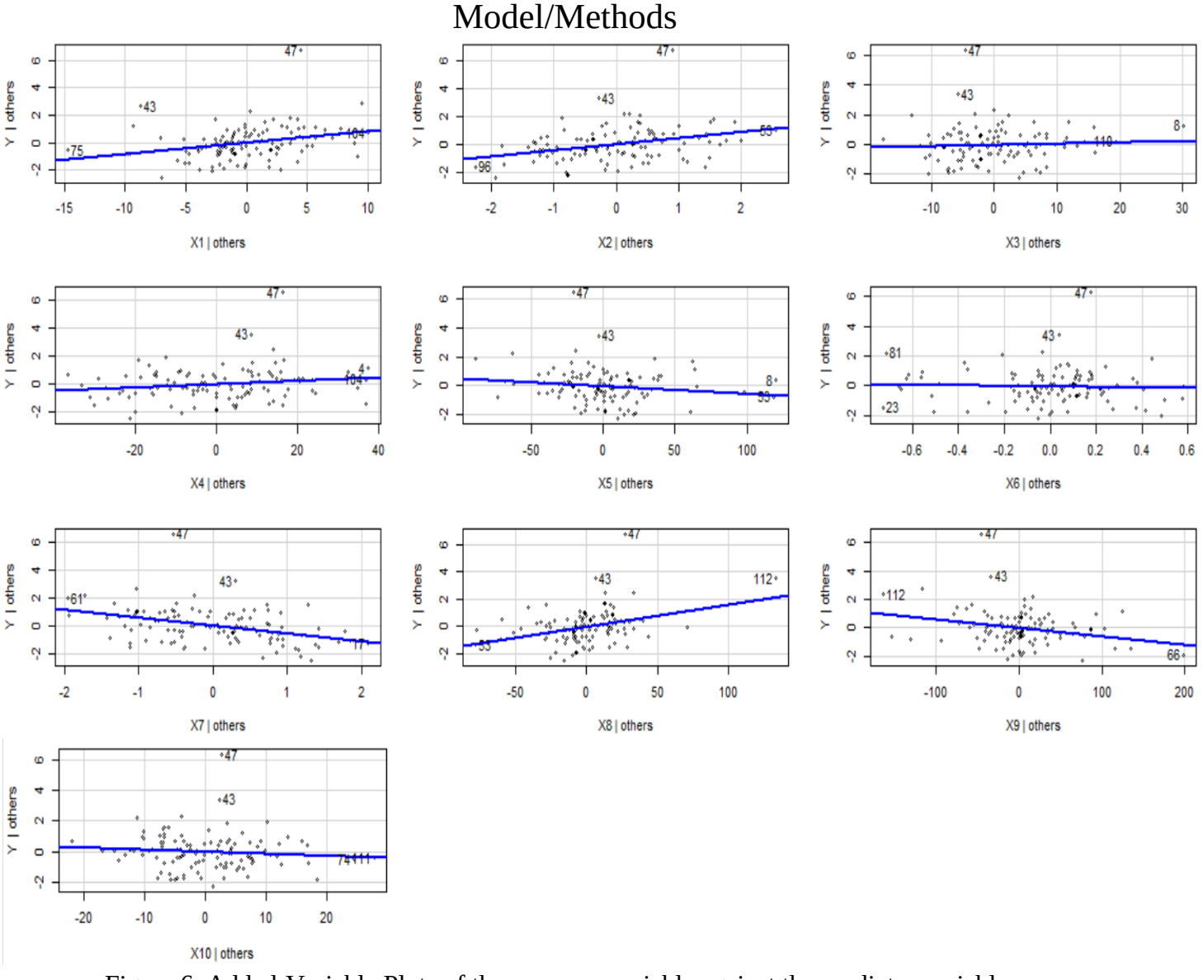



Figure 6. Added-Variable Plots of the response variable against the predictor variables.

Based on the Added-Variable Plots (Figure 6), it provides the regression lines for the relationships between the response variable and the predictor variables. By seeing the regression line we can say that the Infection Risk (X2) and Average daily Census (X8) is having the high effect on the response variable and Routine Culturing Ratio (X3) and Medical School (X6) is having the least effect on the response variable.

### *1.3.6 Correlation Matrix:*

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

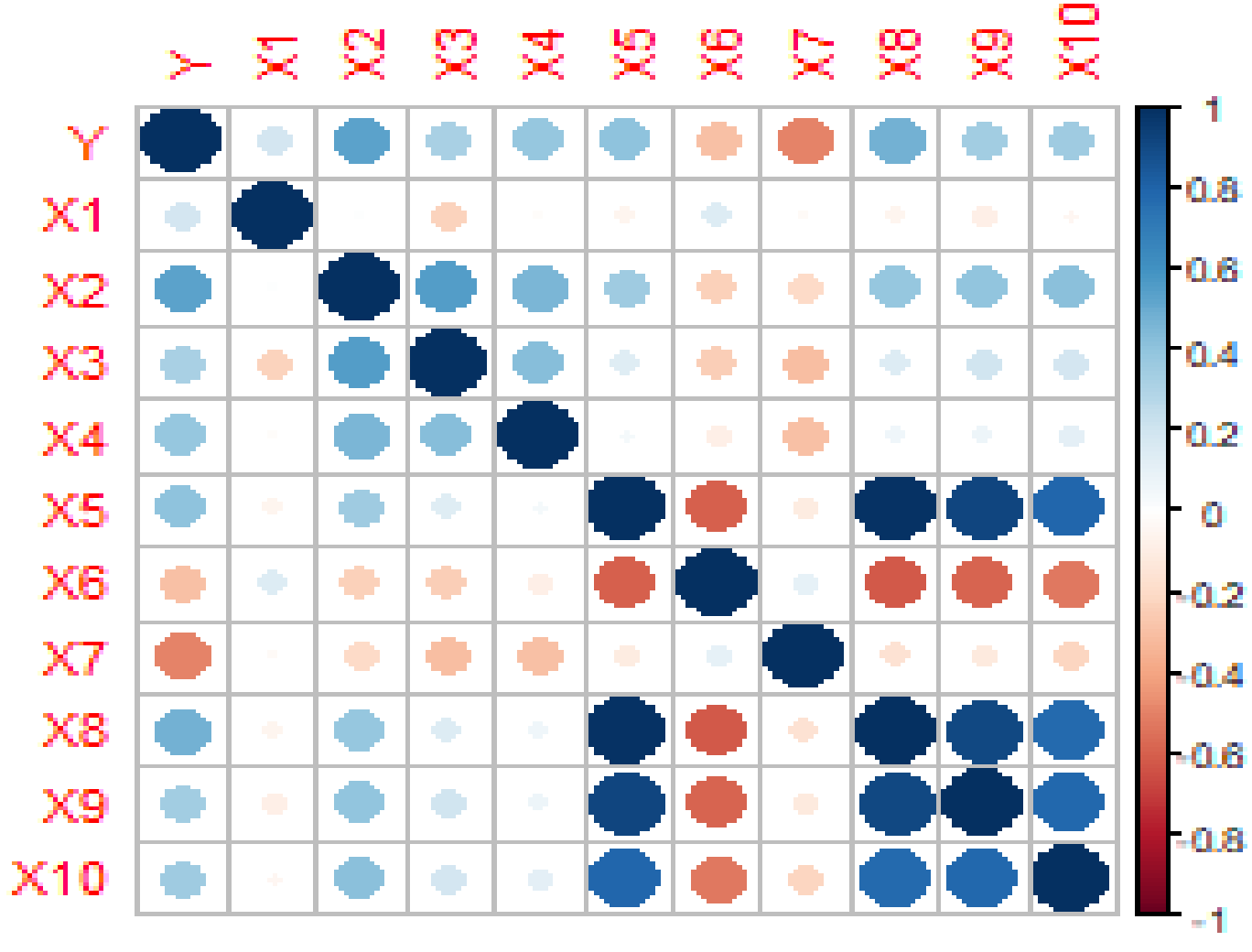



Figure 7. Correlation matrix of the response variable against the predictor variables.

Based on the Correlation matrix (Figure 7), we can see that Number of Beds (X5) is having a high correlation with X8, X9, and X10.
And Average daily Census (X8) is also having a high correlation with X9 and X10.
Along with that Number of nurses (X9) is having a high correlation with X10 also.

# 2. Model/Methods

## *2.1 Creating Interaction Terms and Checking the Correlation:*

First, we fit our model with all the linear terms, then through the Stepwise Regression, we will try to find the best predictor variables (i.e., Significant variables) for our model.

```
$variate
[1] "intercept" "X2"        "X7"        "X8"        "X9"        "X1"        "X4"        "X5"
```

So, we will get the above predictor variables through the model selection (like Adj R2, Cp, AIC, & BIC).

Now we will create the interaction variables for all the above-selected predictor variables only.

After combining all the predictor variables and interaction terms in the data, we check the correlation of the created interaction terms with the other variables.

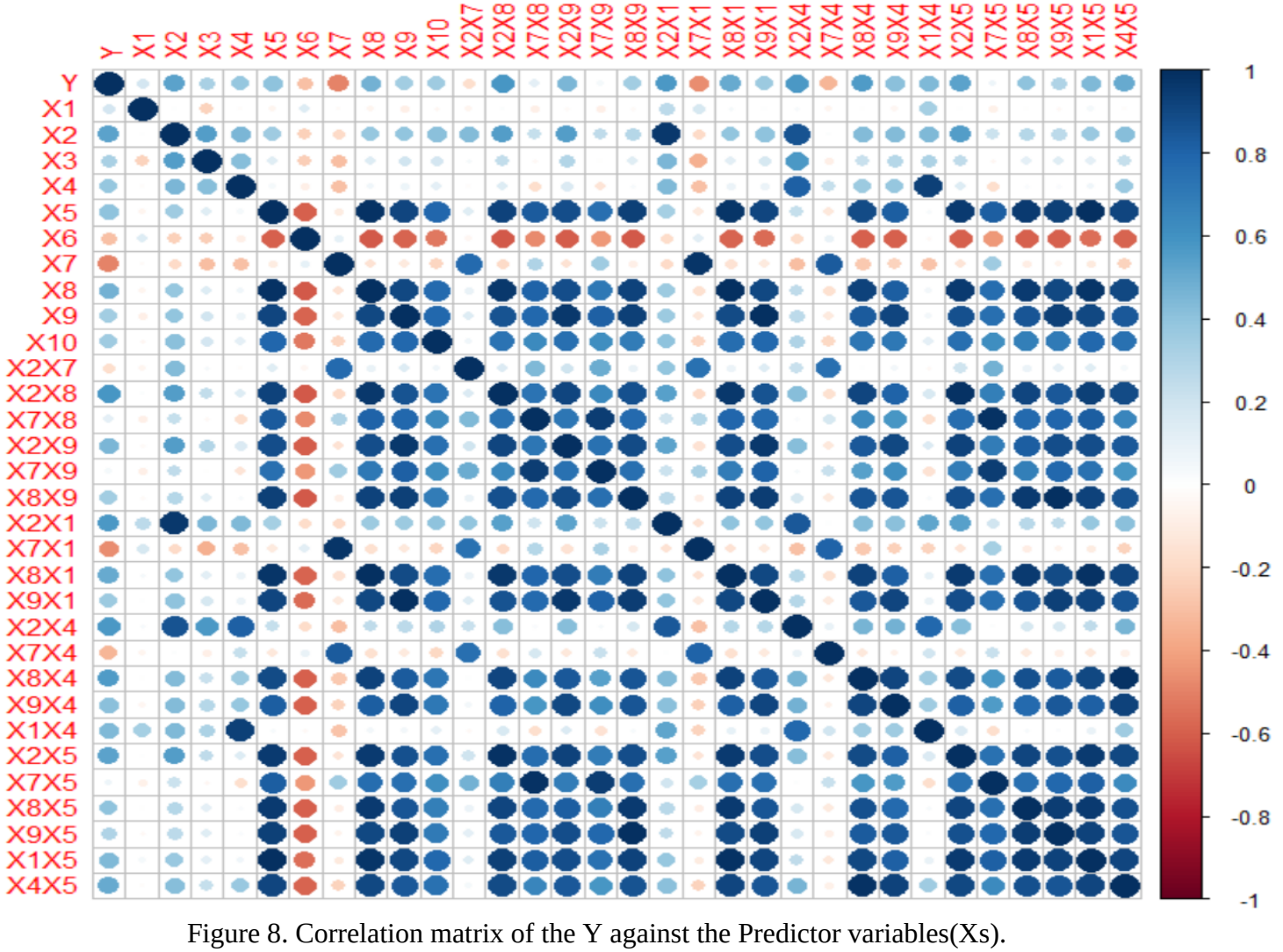



Figure 8. Correlation matrix of the Y against the Predictor variables(Xs).

So as we can see in the above correlation plot majority of variables are highly correlated because of the introduction of interaction variables. Therefore we have to standardize our data.

So, **standardization or centralization** is needed for our predictor variables.

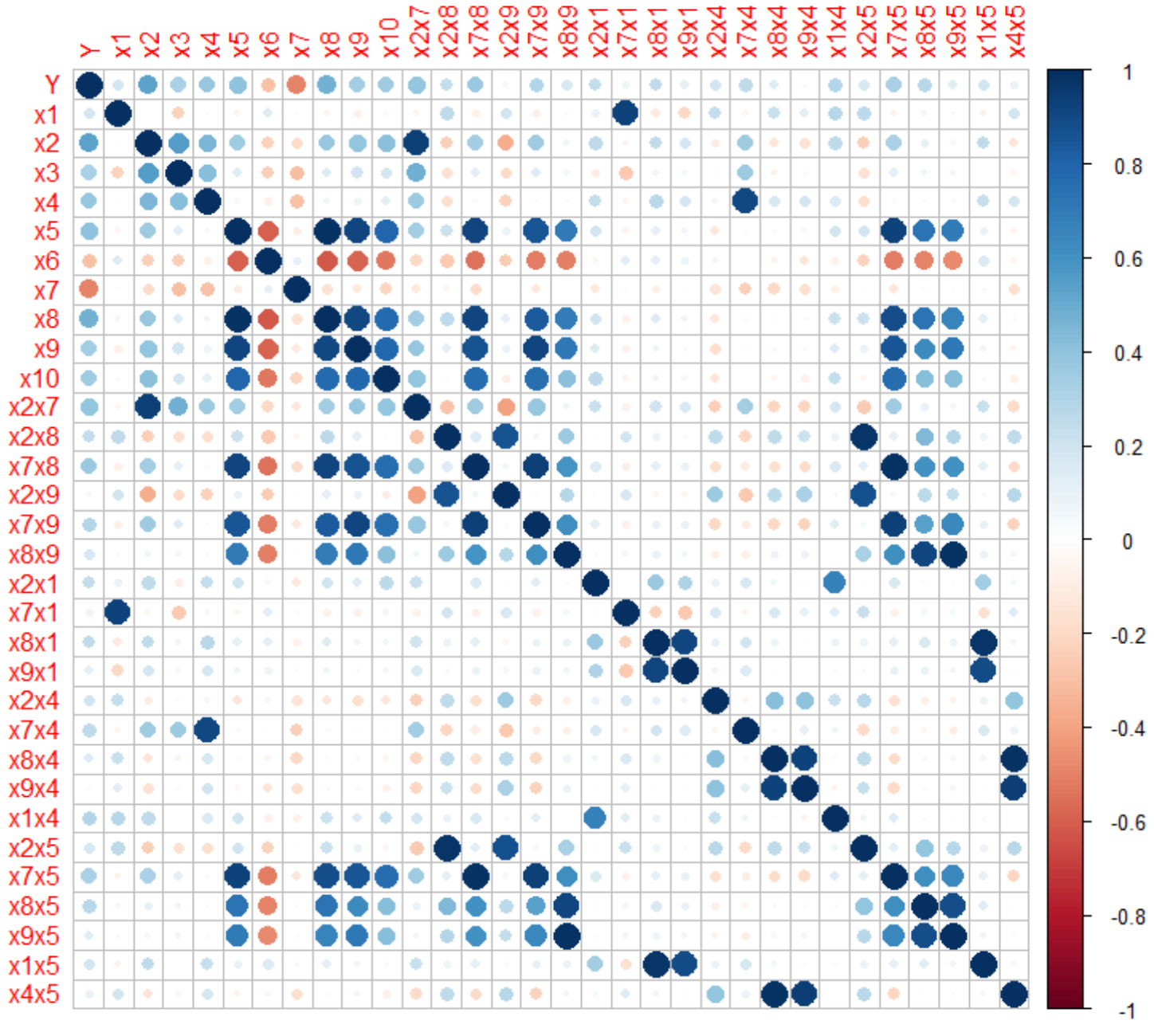



Figure 9. Correlation matrix of the Y against the Predictor variables(Xs) after standardization.

Now the data become standardized, and we can see only a few of them are correlated compared to the previous correlation plot (i.e., without standardization).

Now we will convert our data into a data frame for further analysis.

## *2.2 Building Model Based on Significance Test:*

Once we create the data frame then we will try to build the full model including all significant predictor variables and interaction variables.
And create a reduced model which does not contain any interactiom terms (i.e., only linear model).
Now we will test the significance of the interaction terms in the model through the ANOVA test.s

```
> ################ Test for significance of the interactive terms
> anova(reduced.lmfit, full.lmfit)
Analysis of Variance Table

Model 1: Y ~ x1 + x2 + x4 + x5 + x7 + x8 + x9
Model 2: Y ~ x1 + x2 + x4 + x5 + x7 + x8 + x9 + x2x7 + x2x8 + x7x8 + x2x9 +
    x7x9 + x8x9 + x2x1 + x7x1 + x8x1 + x9x1 + x2x4 + x7x4 + x8x4 +
    x9x4 + x1x4 + x2x5 + x7x5 + x8x5 + x9x5 + x1x5 + x4x5
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    105 154.75
2     84 101.75 21    53.004 2.0838 0.009851 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As we can see the p-value (0.009851) < 0.10, rejecting the null hypothesis. So, we can say that the model with interaction terms gives a better fit compared to the only linear terms.

Therefore, we will go with the full model which includes the interaction terms.

## *2.3 Model Selection:*

Based on the previous full model, we will perform the model selection through the stepwise regression function in order to select the best variables based on the adjusted R^2, the CP, the AIC, and the BIC.

```
> stepwise(data=senic.itact.Std,y="Y",select="adjRsq")
$process
   Step EffectEntered EffectRemoved EffectNumber    Select
1     0     intercept                          1 0.0000000
2     1            x2                          2 0.2781169
3     2            x7                          3 0.4320022
4     3          x2x8                          4 0.5495790
5     4          x2x5                          5 0.5866707
6     5          x2x7                          6 0.6035631
7     6            x8                          7 0.6144388
8     7            x9                          8 0.6270071
9     8            x1                          9 0.6350217
10    9            x4                         10 0.6435686
11   10          x2x4                         11 0.6483339
12   11          x8x5                         12 0.6518704
13   12          x2x9                         13 0.6539433

$variate
 [1] "intercept" "x2"        "x7"        "x2x8"      "x2x5"      "x2x7"      "x8"        "x9"
 [9] "x1"        "x4"        "x2x4"      "x8x5"      "x2x9"
```

Figure 10. Variable selection based on Adjusted R2.

```
> stepwise(data=senic.itact.Std,y="Y",select="CP")
$process
   Step EffectEntered EffectRemoved EffectNumber    Select
1     0     intercept                         1 226.83985
2     1            x2                         2 132.70336
3     2            x7                         3  81.46564
4     3          x2x8                         4  43.09418
5     4          x2x5                         5  31.65198
6     5          x2x7                         6  26.95306
7     6            x8                         7  24.27983
8     7            x9                         8  21.13612
9     8            x1                         9  19.49677
10    9            x4                        10  17.74039
11   10          x2x4                        11  17.19906
12   11                        x2x7           10  16.71587
13   12          x8x5                        11  16.22558

$variate
 [1] "intercept" "x2"        "x7"        "x2x8"      "x2x5"      "x8"        "x9"        "x1"
 [9] "x4"        "x2x4"      "x8x5"
```

Figure 11. Variable selection based on Cp.

```
> stepwise(data=senic.itact.Std,y="Y",select="AIC")
$process
   Step EffectEntered EffectRemoved EffectNumber   Select
1     0     intercept                         1 262.4131
2     1            x2                         2 226.5738
3     2            x7                         3 200.4599
4     3          x2x8                         4 175.2193
5     4          x2x5                         5 166.4668
6     5          x2x7                         6 162.7004
7     6            x8                         7 160.4961
8     7            x9                         8 157.6801
9     8            x1                         9 156.1442
10    9            x4                        10 154.3748
11   10          x2x4                        11 153.7514
12   11                        x2x7           10 153.3245
13   12          x8x5                        11 152.7301

$variate
 [1] "intercept" "x2"        "x7"        "x2x8"      "x2x5"      "x8"        "x9"        "x1"
 [9] "x4"        "x2x4"      "x8x5"
```

Figure 12. Variable selection based on AIC.

```
> stepwise(data=senic.itact.Std,y="Y",select="BIC")
$process
   Step EffectEntered EffectRemoved EffectNumber    Select
1     0     intercept                         1 147.19622
2     1            x2                         2 110.87681
3     2            x7                         3  84.73674
4     3          x2x8                         4  60.21125
5     4          x2x5                         5  51.80713
6     5          x2x7                         6  48.27076
7     6            x8                         7  46.31478
8     7            x9                         8  43.98071
9     8            x1                         9  42.90859
10    9            x4                        10  41.78204
11   10          x2x4                        11  41.72362
12   11                        x2x7           10  40.92136
13   12          x8x5                        11  40.90907

$variate
 [1] "intercept" "x2"        "x7"        "x2x8"      "x2x5"      "x8"        "x9"        "x1"
 [9] "x4"        "x2x4"      "x8x5"
```

Figure 13. Variable selection based on BIC.

We found that the best variables are:

```
"x2"    "x7"    "x2x8"    "x2x5"    "x8"    "x9"    "x1"    "x4"    "x2x4"    "x8x5"
```

## *2.4 Selecting Best Model:*

Now we are fitting our model based on the best-selected variables by the stepwise regression and checking the significance of the model through the Summary function.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4996     0.1513  62.782  < 2e-16 ***
x2            0.7531     0.1422   5.295 6.86e-07 ***
x7           -0.4640     0.1191  -3.894 0.000176 ***
x2x8          2.2452     0.6080   3.693 0.000358 ***
x2x5         -1.8005     0.5823  -3.092 0.002564 **
x8            1.1435     0.3235   3.535 0.000615 ***
x9           -0.5924     0.2764  -2.143 0.034464 *
x1            0.2298     0.1147   2.004 0.047711 *
x4            0.2838     0.1260   2.251 0.026508 *
x2x4          0.2137     0.1019   2.098 0.038384 *
x8x5         -0.1584     0.1029  -1.539 0.126869
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.128 on 102 degrees of freedom
Multiple R-squared:  0.6826,    Adjusted R-squared:  0.6515
F-statistic: 21.94 on 10 and 102 DF,  p-value: < 2.2e-16
```

As we can see that Interaction term( x8x5) is not significant to the model, based on the p-value(>0.10) So, we will drop this term. And rebuild our model and check again.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.3698     0.1265  74.082  < 2e-16 ***
x2            0.7870     0.1414   5.564 2.10e-07 ***
x7           -0.4953     0.1182  -4.191 5.87e-05 ***
x2x8          2.0449     0.5978   3.421 0.000897 ***
x2x5         -1.6758     0.5805  -2.887 0.004741 **
x8            0.9450     0.2986   3.165 0.002042 **
x9           -0.5918     0.2782  -2.127 0.035823 *
x1            0.2134     0.1149   1.857 0.066160 .
x4            0.2648     0.1263   2.097 0.038413 *
x2x4          0.2293     0.1020   2.247 0.026761 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.136 on 103 degrees of freedom
Multiple R-squared:  0.6752,    Adjusted R-squared:  0.6469
F-statistic:  23.8 on 9 and 103 DF,  p-value: < 2.2e-16
```

Now our model looks fine in terms of significance. So we will consider it as the full model.

Now we will test the significance of the full model against the reduced model through the ANOVA test. The full model contains all the above-selected variables and the reduced model contains only linear terms.

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

Model 1: Y ~ x2 + x7 + x8 + x9 + x1 + x4
Model 2: Y ~ x2 + x7 + x2x8 + x2x5 + x8 + x9 + x1 + x4 + x2x4
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    106 160.00
2    103 132.89  3    27.106 7.0028 0.0002474 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As we can see the p-value (0.0002474) < 0.10, rejecting the null hypothesis. So, we can say that the model with interaction terms gives a better fit compared to the only linear terms.

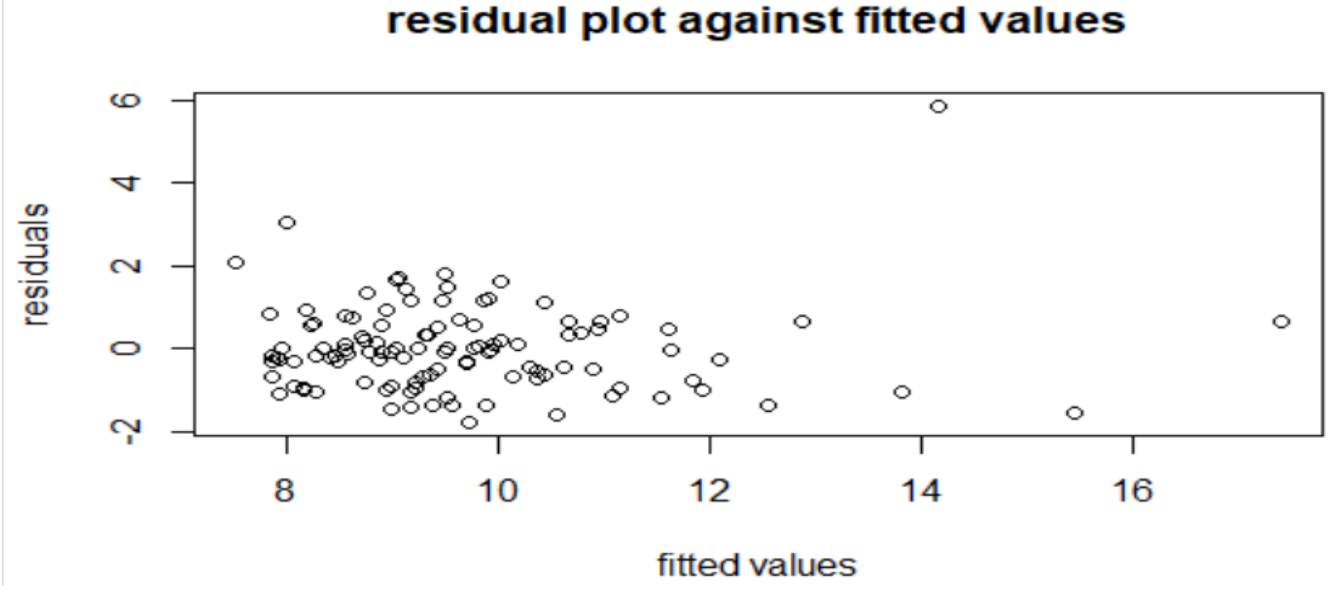
Therefore, we will go with the full model which includes the interaction terms.

So, we will consider this model as our final model. And in the upcoming steps, we will check the assumptions for this selected model.

Our Estimated regression function:
y^ = 9.4996 + 0.7531 x2 - 0.4597 x7 + 2.2452 x2x8 - 1.8005 x2x5 + 1.1435 x8 - 0.5924 x9 + 0.2298 x1 + 0.2838 x4 + 0.2137 x2x4.

## *2.5 Model Assumptions Checking:*

### *2.5.1 Linearity Assumptions Checking:*



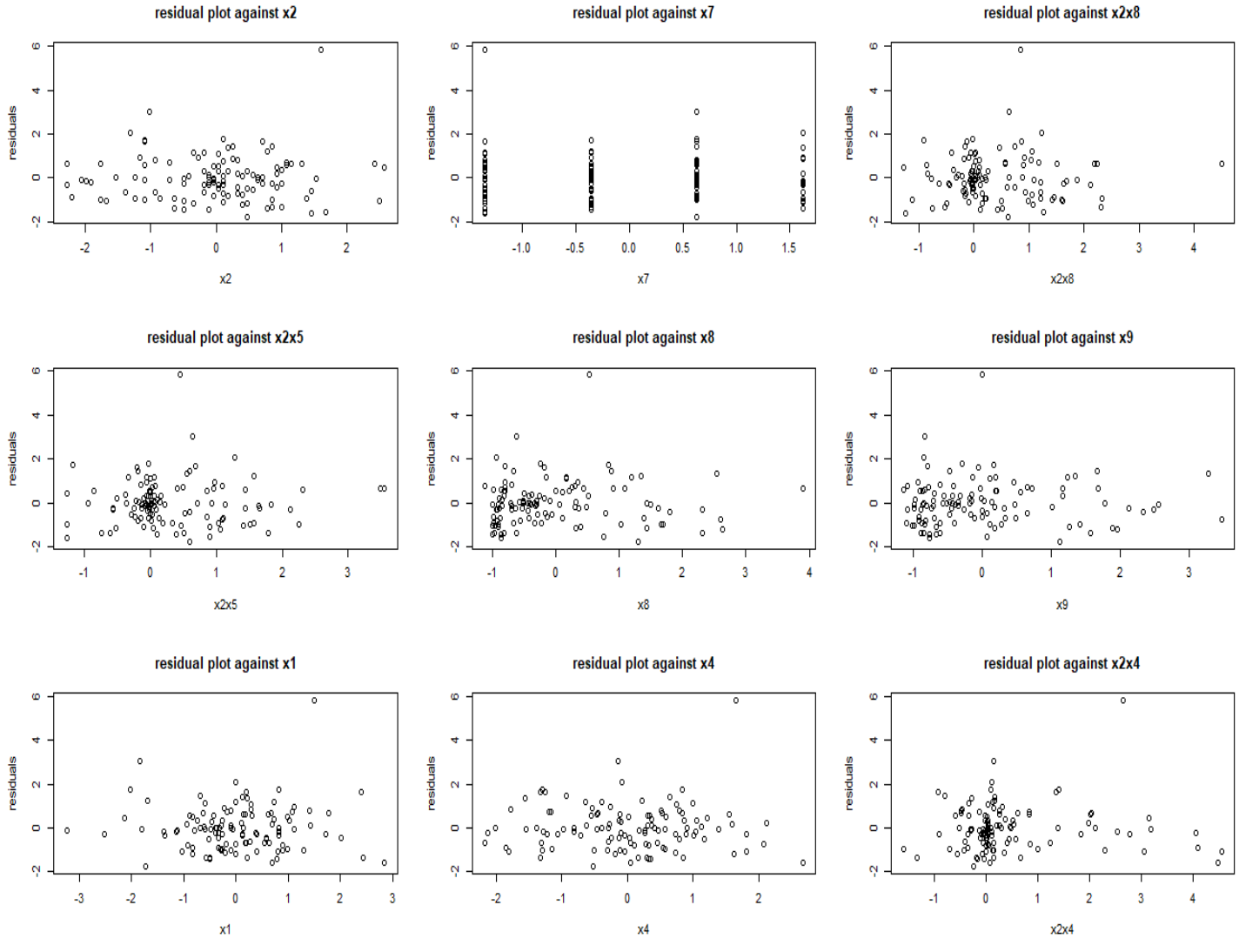



Figure 14. Residuals plot against fitted values and predictor variables.

From the above residual plot against the fitted value, we can see that residuals are looking a little decreasing with the increasing level of the fitted value. Because of the small sample size, we can't say that the linearity assumption is not satisfied. We have to check other plots or perform a lack of fit test, to come to a conclusion.

But from the residual plot against the predictor variable (Xs), data points are randomly scattered within a certain range indicates no serious departure from linearity.

So, we can say that residuals are linearly distributed.

### *2.5.2 Constant Variance Assumption Checking:*

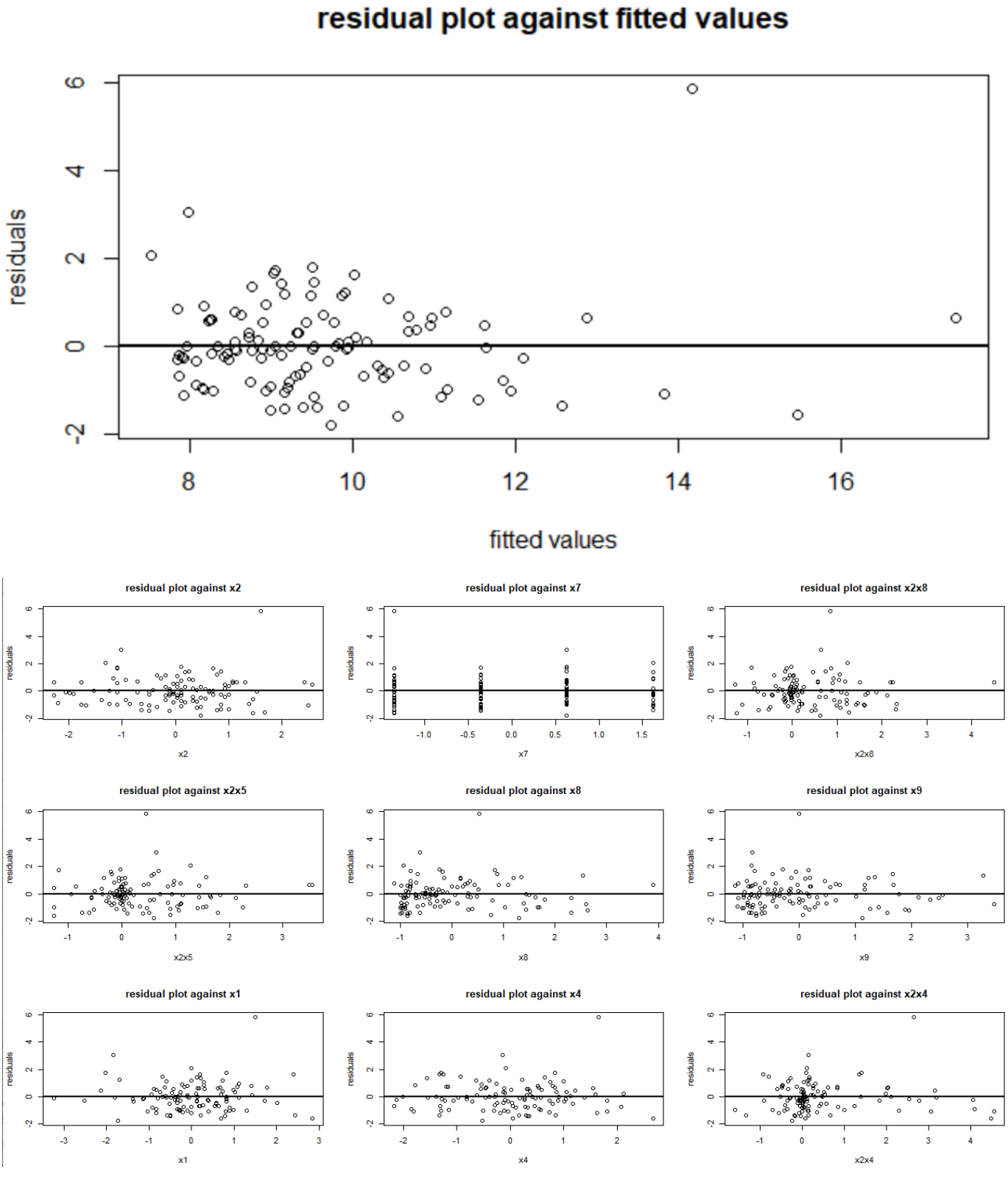



Figure 15. Residuals plot against fitted values and predictor variables.

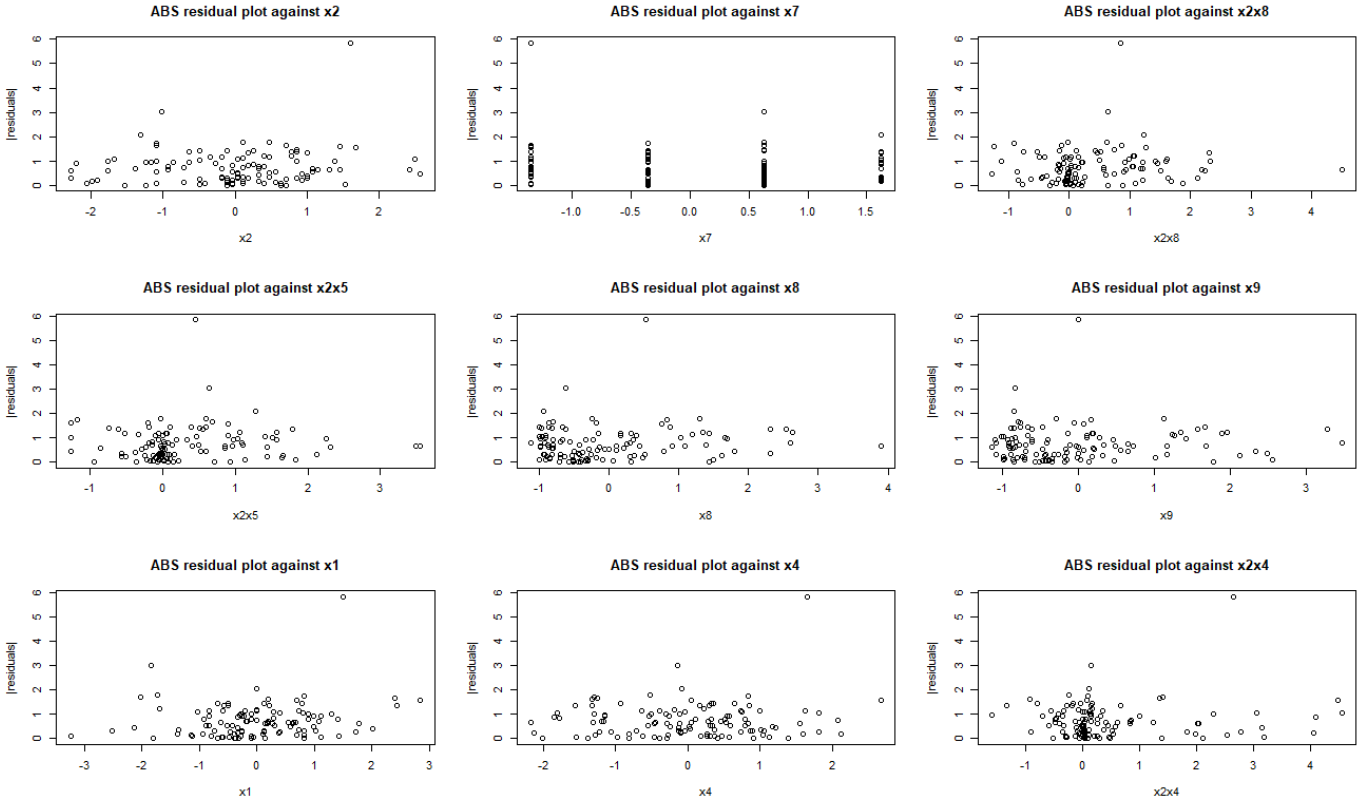



Figure 16. Absolute Residuals plot against the predictor variables.

From the above residual plot against the fitted value, we can see that residuals are decreasing with the increasing level of the fitted value. So, the variance is not looking constant.

But from the residual plot against the predictor variable (Xs) and absolute residual plot against the predictor variable (Xs), data points are randomly scattered within a certain range indicates variance is constant. Because of the small sample size, it is hard to say whether the variance is constant or not. So, we will check with the Breusch-Pagan (BP) test.

```
        studentized Breusch-Pagan test

data:  reduced.lmfit
BP = 10.086, df = 9, p-value = 0.3435
```

Figure 17. Breusch-Pagan Test output.

Based on the BP test, the p-value (0.3435) is greater than the significance level (i.e., 0.10), we will fail to reject the null hypothesis and conclude that error variance is constant.

### *2.5.3 Independence Assumption Checking:*

As we don't know whether our dataset is time series related or not. So, we will not be able to check the independence assumption.

### *2.5.4 Normality Assumption Checking:*

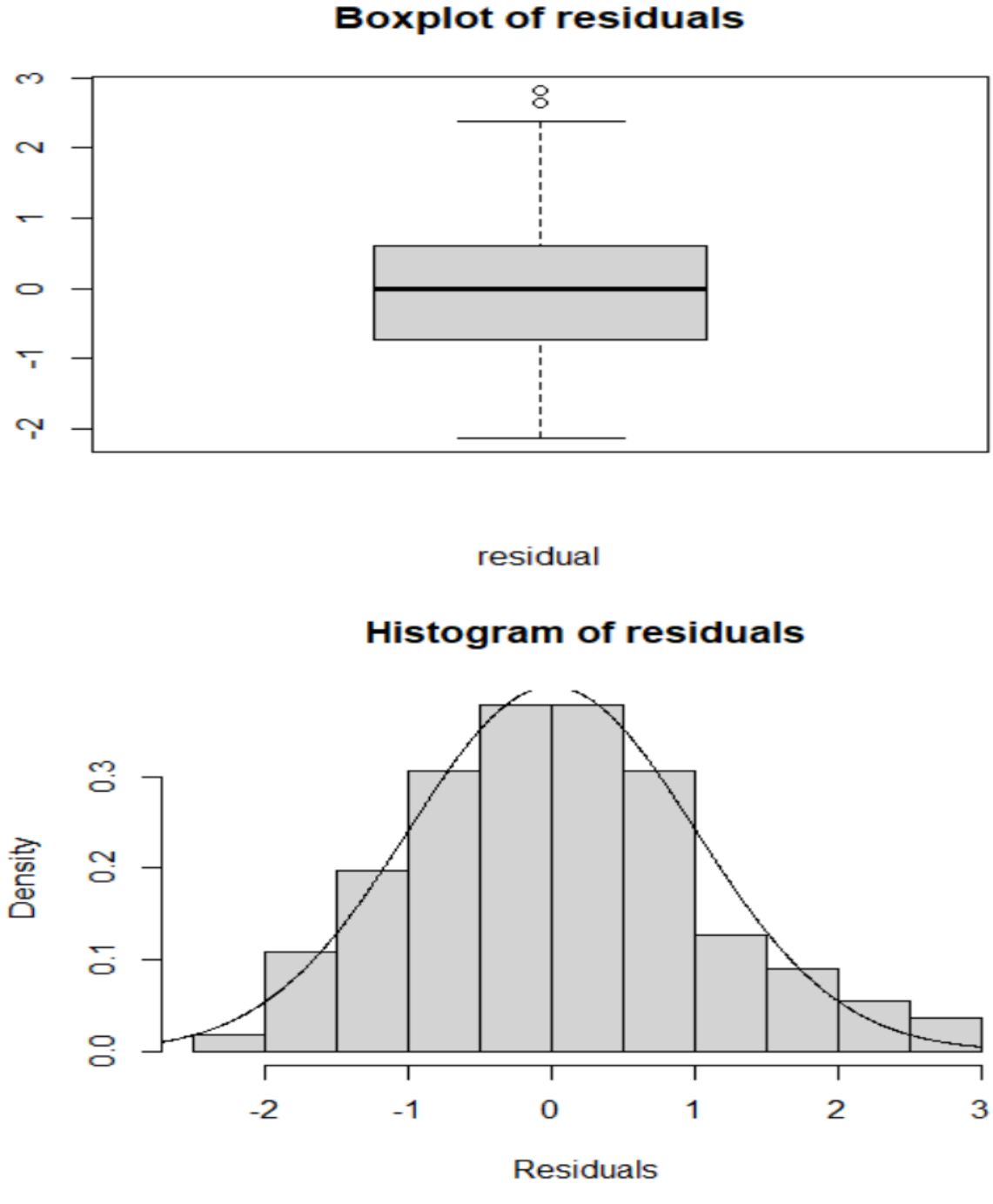



Figure 17. Boxplot and Histogram of Residuals.

From the Boxplot of residual, we can say that it looks symmetrical as it is distributed equally on each side of the medium other than a few outliers. But we have to confirm with other plots.

From the Histogram of residuals, it looks normally distributed but has a heavy tail on the right side.

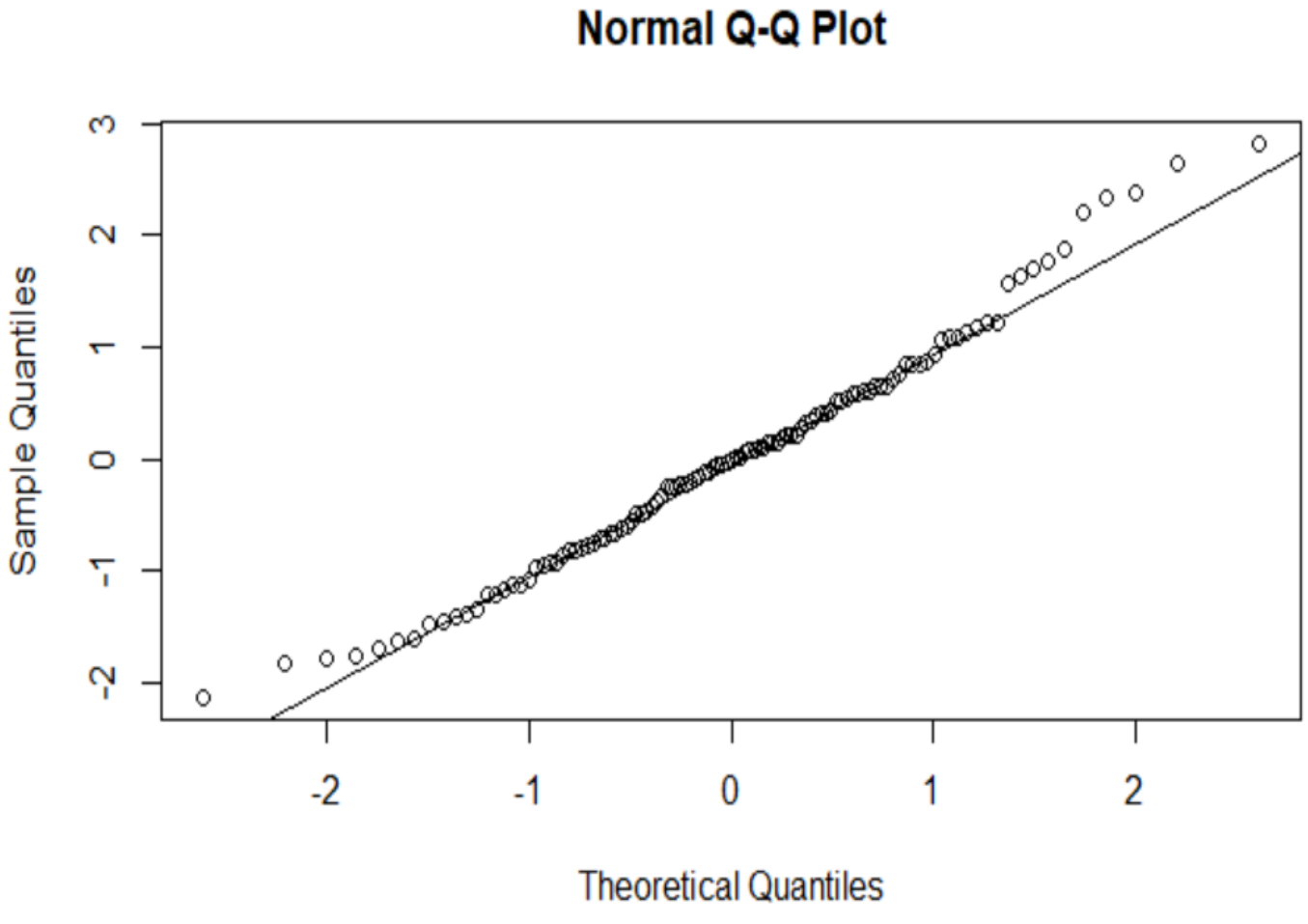



Figure 18. Normal Q-Q Plot of Residuals.

Based on the Normal Q-Q plot, there is no severe departure from the normality, but it is having heavy tails so we will check with the Shapiro-Wilk test for confirmation.

```
        Shapiro-Wilk normality test

data:  data.res
W = 0.98417, p-value = 0.2141
```

Figure 19. Shapiro-Wilk test output.

Based on the Shapiro-Wilk test, the p-value is greater than the significance value(i.e.,0.10), failing to reject the null hypothesis and we can say that our data is normally distributed.

### *2.5.5 Outlier Checking:*

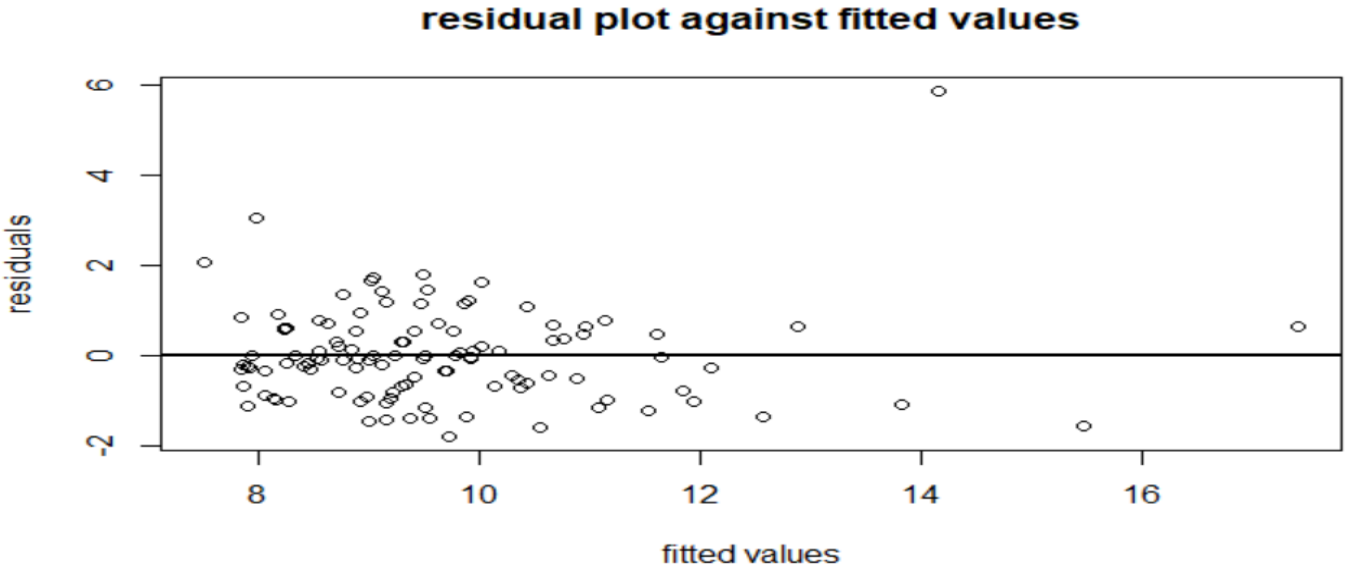



Figure 19. Residuals plot against the fitted values.

```
> which(abs(data.res) >3)
43 47
43 47
```

From the residual plot against the fitted value, we can see two observations (i.e., 43 & 47 index, calculated from R-code) is having a value greater than 3. So, it is a potential outlier.

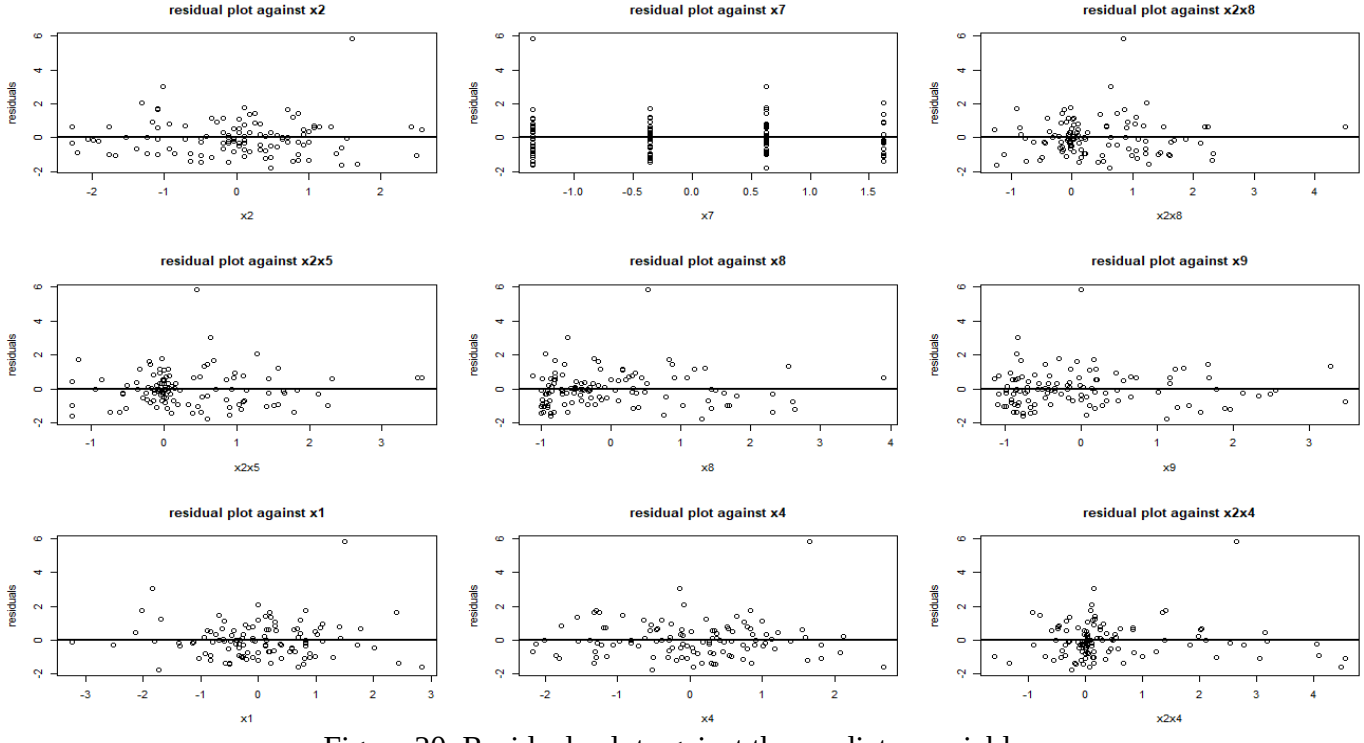



Figure 20. Residuals plot against the predictor variables.

From the residual plot against the predictor variables, we can see few observations are having a value greater than 3. So, those are the potential outlier.


Figure 21. Boxplot of the Residuals.

Boxplot of the residuals also shows some potential outlier far away from the medium values.

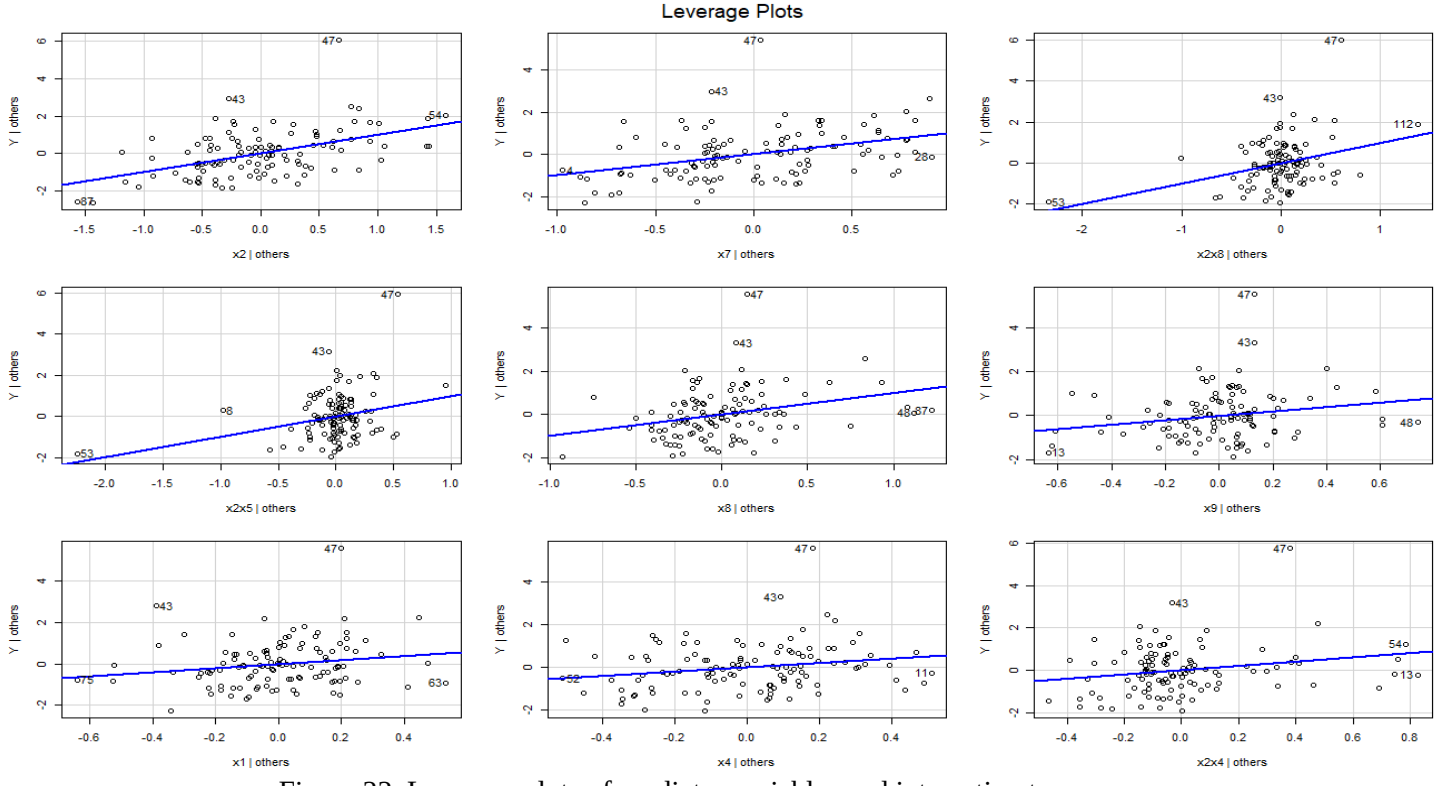



Figure 22. Leverage plots of predictor variables and interaction terms.

From the above leverage Plots, we can see that there are many influential observations. Some of the influential points are 43, 47,53,112, etc.

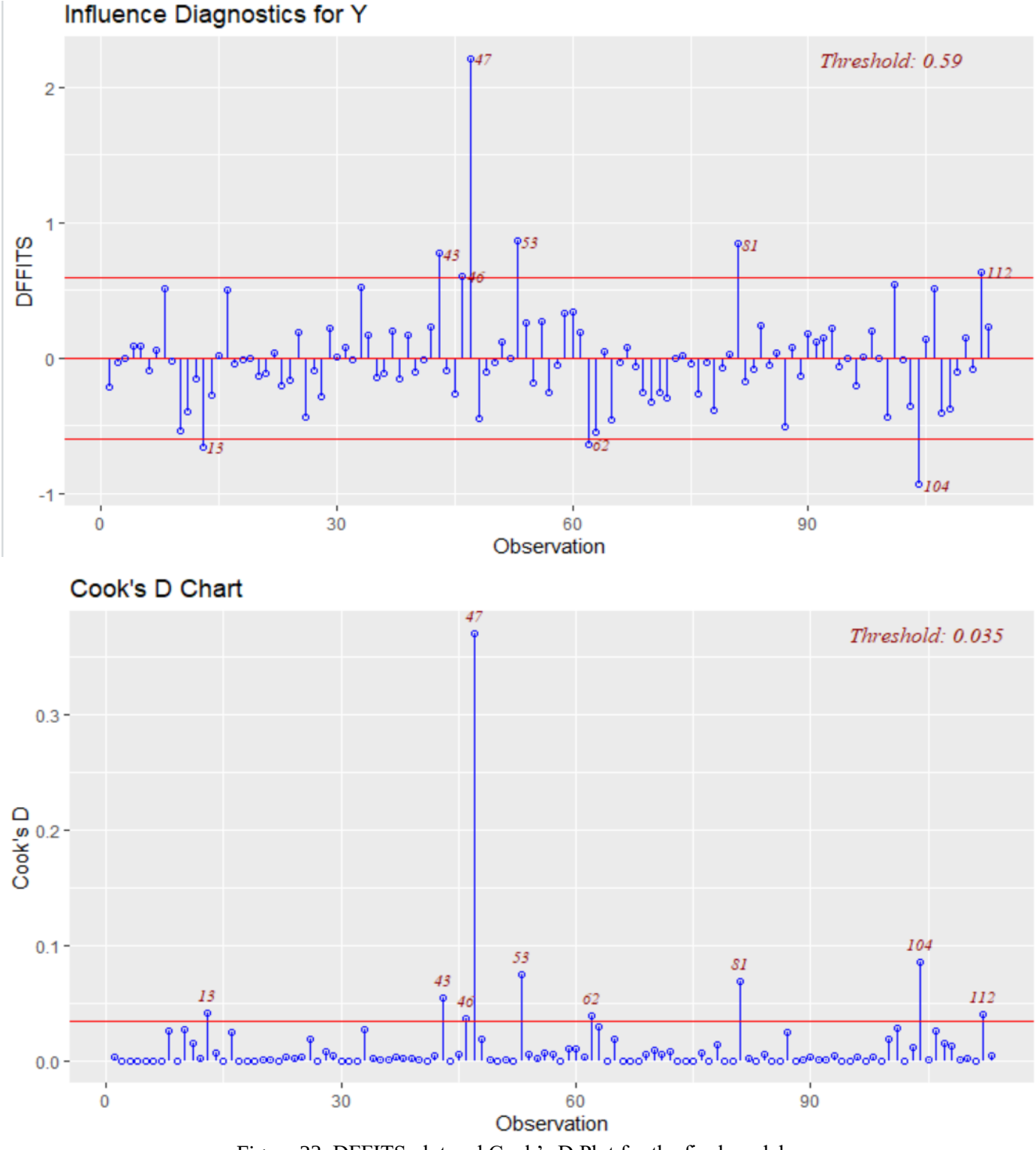
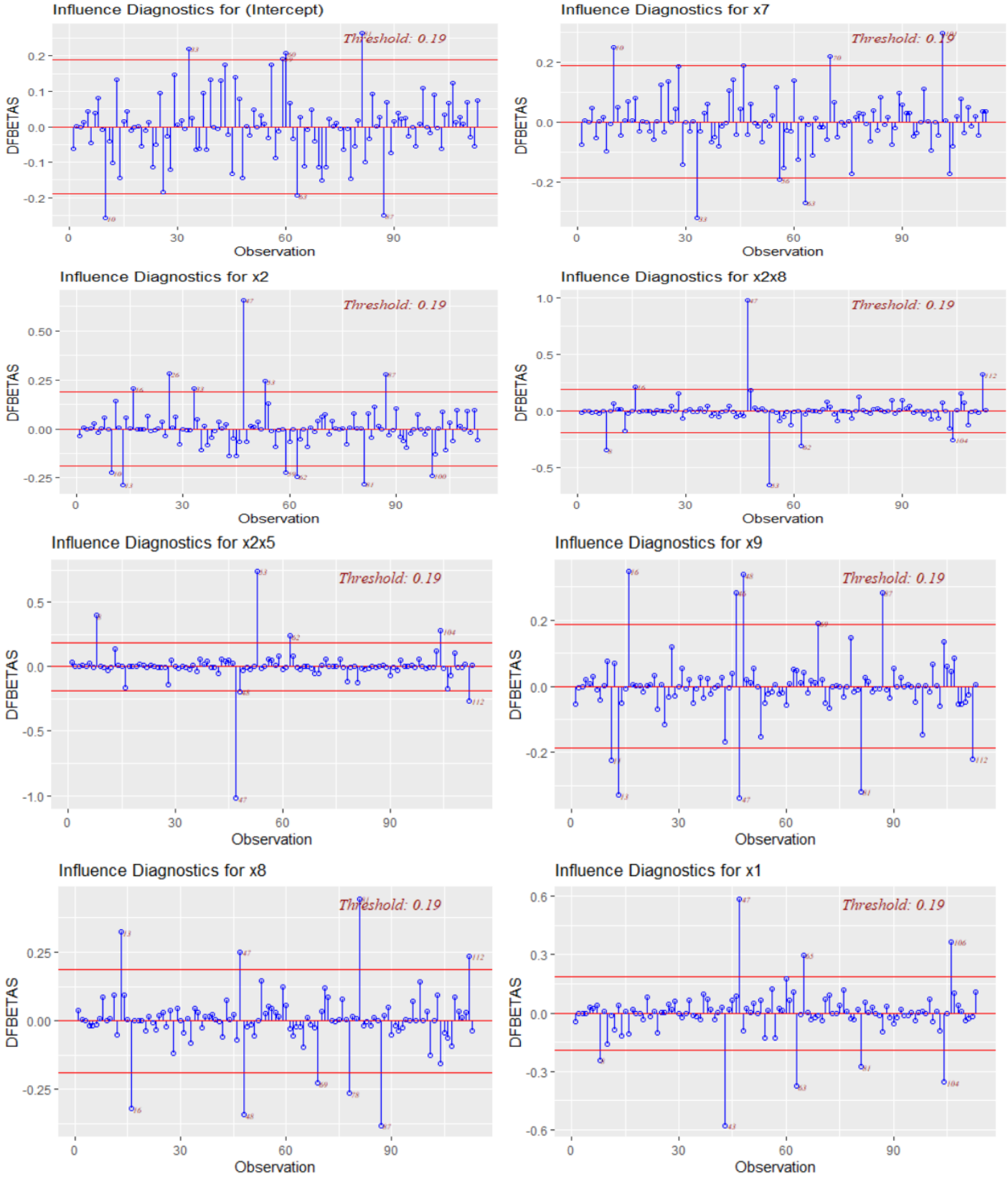



Figure 23. DFFITS plot and Cook's D Plot for the final model.


Influence Diagnostics for (Intercept)
Threshold: 0.19
DFBETAS
Observation
Influence Diagnostics for x7
Threshold: 0.19
DFBETAS
Observation
Influence Diagnostics for x2
Threshold: 0.19
DFBETAS
Observation
Influence Diagnostics for x2x8
Threshold: 0.19
DFBETAS
Observation
Influence Diagnostics for x2x5
Threshold: 0.19
DFBETAS
Observation
Influence Diagnostics for x9
Threshold: 0.19
DFBETAS
Observation
Influence Diagnostics for x8
Threshold: 0.19
DFBETAS
Observation
Influence Diagnostics for x1
Threshold: 0.19
DFBETAS
Observation

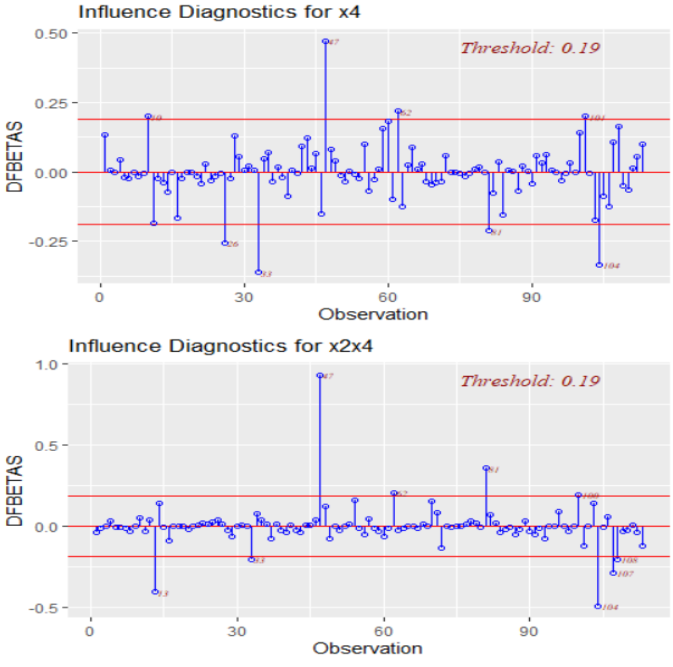



Figure 24. DFBETAS plot for the final model

Based on the DFFITS, Cook's D, and DFBETAS tests, there are many influential data points based on the threshold levels. Now we have to find the most far away influential point and will remove them in remedial part.

## *2.5.6 Multicollinearity Checking:*

```
Call:
lm(formula = Y ~ x2 + x7 + x2x8 + x2x5 + x8 + x9 + x1 + x4 +
    x2x4, data = senic.itact.Std)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9440 -0.7693 -0.0814  0.6083  5.4019

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.3698     0.1265  74.082  < 2e-16 ***
x2            0.7870     0.1414   5.564 2.10e-07 ***
x7           -0.4953     0.1182  -4.191 5.87e-05 ***
x2x8          2.0449     0.5978   3.421 0.000897 ***
x2x5         -1.6758     0.5805  -2.887 0.004741 **
x8            0.9450     0.2986   3.165 0.002042 **
x9           -0.5918     0.2782  -2.127 0.035823 *
x1            0.2134     0.1149   1.857 0.066160 .
x4            0.2648     0.1263   2.097 0.038413 *
x2x4          0.2293     0.1020   2.247 0.026761 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.136 on 103 degrees of freedom
Multiple R-squared:  0.6752,    Adjusted R-squared:  0.6469
F-statistic:  23.8 on 9 and 103 DF,  p-value: < 2.2e-16
```

As we can see in our summary table, the overall model is significant, and the individual predictor variable is also significant based on the p-value and value of standard error also very less which means there is no major concern about the multicollinearity existence.

Now we will confirm this by Variance Inflation Factor (VIF) analysis.

```
> vif(reduced.lmfit)
       x2        x7      x2x8      x2x5        x8        x9        x1        x4      x2x4
 1.736585  1.212257 23.384644 21.599480  7.741143  6.720237  1.146364  1.383953  1.215174
```

So Max VIF(x2x8) = 23.384644 >=10. Yes, serious multicollinearity problems exist.

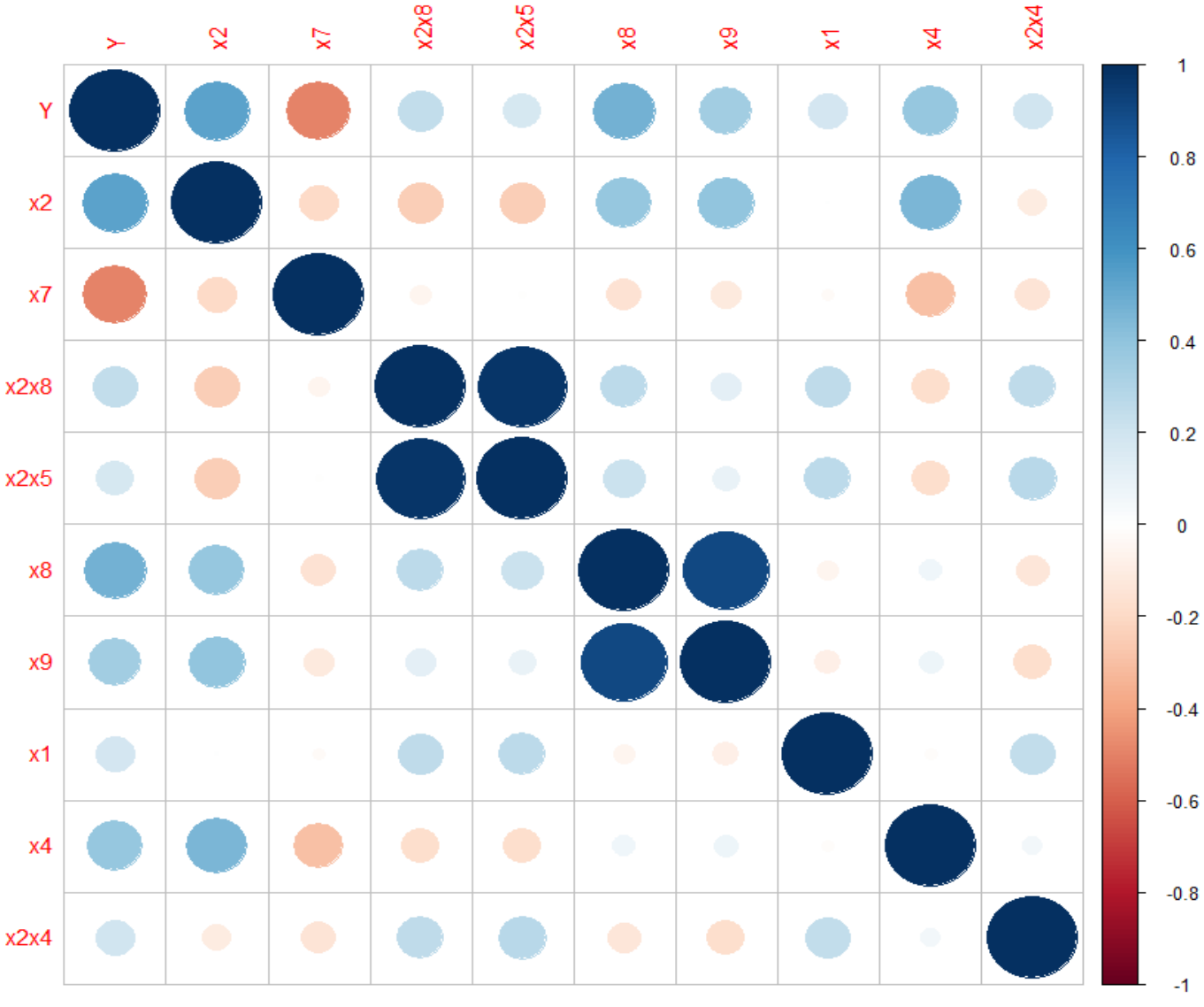



Figure 25. Correlation matrix of the Y against the Predictor variables(Xs).

## *2.6 Remedial Measures:*

Since our model satisfied all the mandatory assumptions such as Linearity, Constant error variance, and Normal distribution. So, no need to apply any transformation methods such as Transformation, Box-Cox Transformation, or Weighted Least Square (WLS).

But we have to drop some variables and observations because of multicollinearity and Outlier concerns.

### *2.6.1 Multicollinearity remedial measure:*

Based on Figure25 (Correlation Matrix), we can see that interaction terms x2x8 and x2x5 are highly correlated to each other so we will drop one of them from the model.

So based on the summary table, we will see that interaction term x2x5 is having a negative estimation value (i.e., -1.6758), So we will first try to drop it and see the outcome by re-evaluating the VIF.

```
> vif(reduced.lmfit)
       x2        x7      x2x8        x8        x9        x1        x4      x2x4
1.727112 1.155779 1.613621 7.144004 6.325394 1.146318 1.383122 1.203495
```

After deletion of the one highly correlated interaction term. We can see that now all VIF values for each variable are less than 10.

So, we can say that all the considered variables are independent of each other.

### *2.6.2 Delete Outliers:*

For the deletion of the Outlier, we have to first find the far-away potential outlier and try to remove them.

So based on the Rule of thumb:

```
> which(abs(data.res) >3)
43 47
43 47
```

We can see that 43rd and 47th index observation is considered as an Outlier.
So, we have to remove them from our data and check it again.

```
> which(abs(data.res) >3)
named integer(0)
```

After deletion of 43rd and 47th index observation. We can see that now there is no observation that comes under the rule of thumb.

After applying remedial measures, we will check the model summary table to check whether our overall model is still significant or not.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.37177    0.10685  87.712  < 2e-16 ***
x2           0.69817    0.12053   5.793 7.74e-08 ***
x7          -0.55693    0.09826  -5.668 1.35e-07 ***
x2x8         0.35522    0.13368   2.657  0.00914 **
x8           1.02293    0.24479   4.179 6.20e-05 ***
x9          -0.59834    0.23077  -2.593  0.01092 *
x1           0.20982    0.09987   2.101  0.03811 *
x4           0.19032    0.10761   1.769  0.07994 .
x2x4         0.12552    0.08713   1.441  0.15276
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9649 on 102 degrees of freedom
Multiple R-squared:  0.6911,    Adjusted R-squared:  0.6668
F-statistic: 28.52 on 8 and 102 DF,  p-value: < 2.2e-16
```

As we can see that our overall model is significant but the Interaction term( x2x4) becomes insignificant to the model, based on the p-value(>0.10) So, we will drop this term. And rebuild our model and check again.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.40958    0.10411  90.379  < 2e-16 ***
x2           0.69323    0.12110   5.724 1.03e-07 ***
x7          -0.57588    0.09788  -5.883 5.04e-08 ***
x2x8         0.40010    0.13067   3.062  0.00281 **
x8           0.99337    0.24520   4.051 9.89e-05 ***
x9          -0.60021    0.23197  -2.587  0.01106 *
x1           0.23081    0.09932   2.324  0.02210 *
x4           0.20108    0.10791   1.863  0.06525 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9699 on 103 degrees of freedom
Multiple R-squared:  0.6848,    Adjusted R-squared:  0.6634
F-statistic: 31.97 on 7 and 103 DF,  p-value: < 2.2e-16
```

Now our model looks fine in terms of significance. And there is not much reduction in the Adjusted R2 after the removal of the Interaction term( x2x4), So we will consider this as the final model.

Now we will check all the mandatory model assumptions again.

## *2.7 Model Assumptions Checking after Remedial Measures:*

### *2.7.1 Linearity Assumptions Checking:*

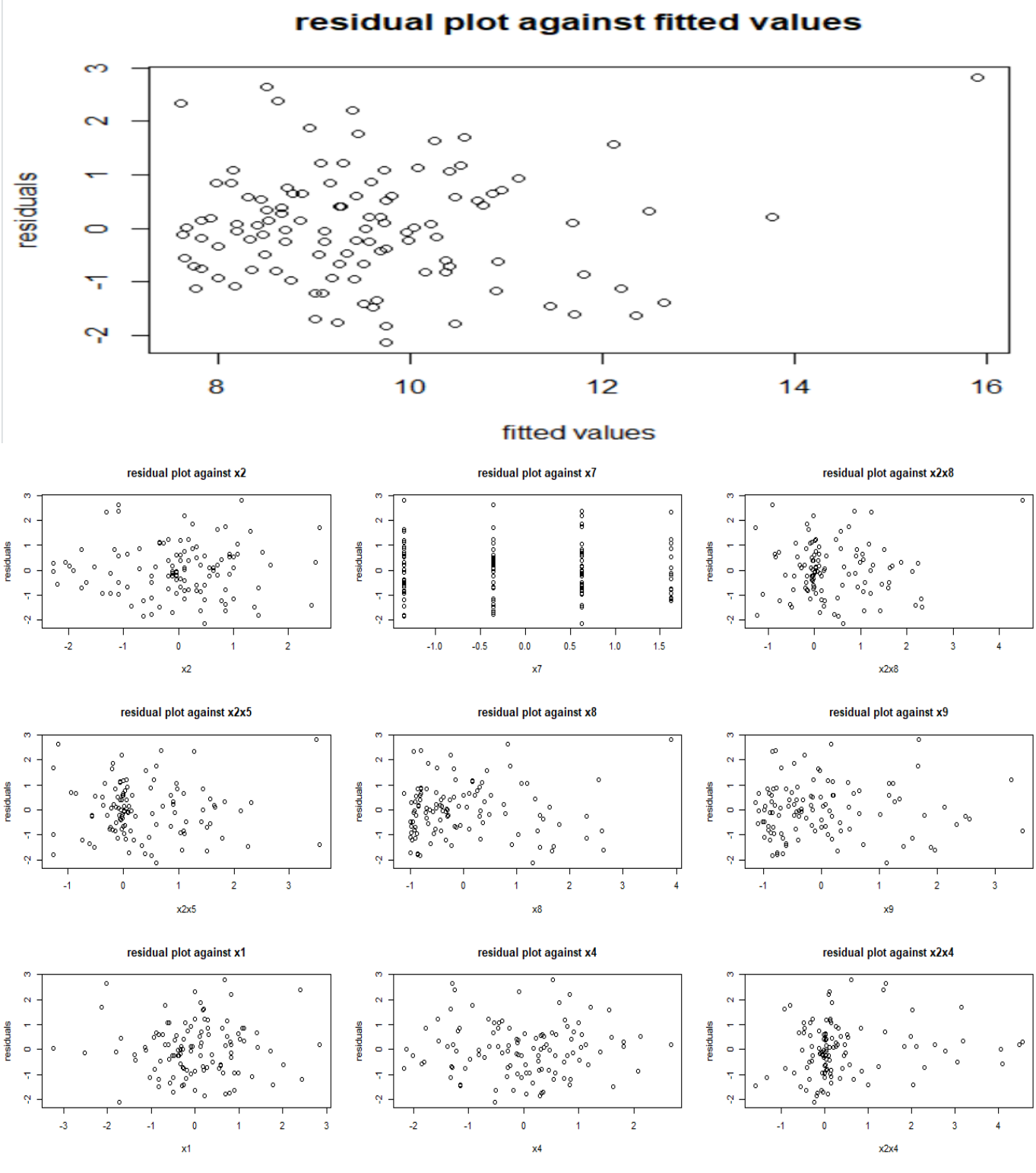



Figure 26. Residuals plot against fitted values and predictor variables.

From the above residual plot against the fitted value, we can see that data points are randomly scattered within a certain range indicates no serious departure from linearity.

But from the residual plot against the predictor variable (Xs), data points are randomly scattered within a certain range indicates no serious departure from linearity.

So, we can say that residuals are having a linear pattern.

### *2.5.2 Constant Variance Assumption Checking:*

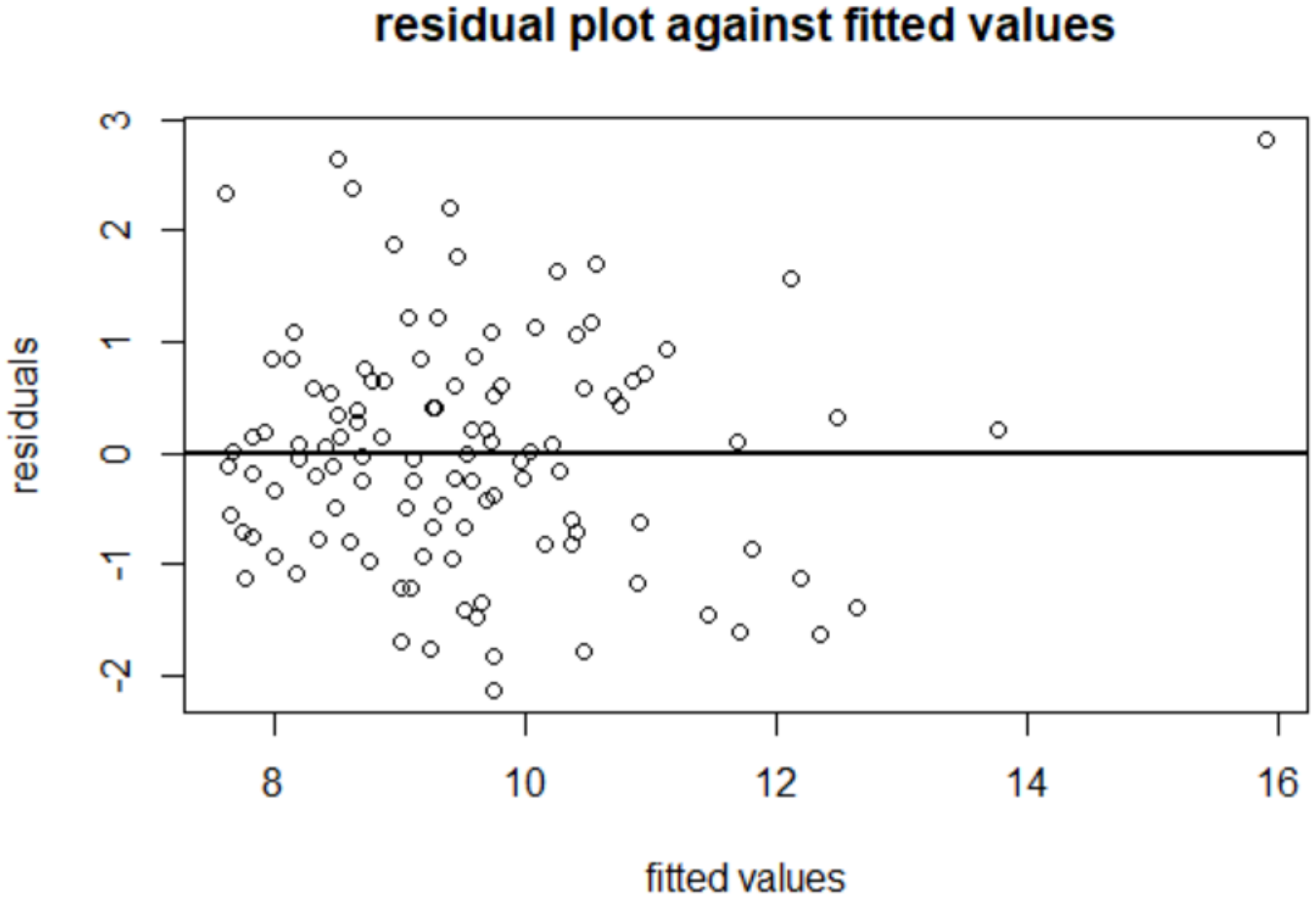



Figure 27. Residuals plot against the fitted values.

From the above residual plot against the fitted value, we can see that data points are randomly scattered within a certain range indicates no serious departure from constant variance.

Because of the small sample size, it is hard to say based on plots. So, we will check with the Breusch-Pagan (BP) test.

```
        studentized Breusch-Pagan test

data:  model
BP = 11.128, df = 7, p-value = 0.1331
```

Figure 28. Breusch-Pagan Test output.

Based on the BP test, the p-value (0.1331) is greater than the significance level (i.e., 0.10), we will fail to reject the null hypothesis and conclude that error variance is constant.

### *2.5.3 Independence Assumption Checking:*

As we don't know whether our dataset is time series related or not. So, we will not be able to check the independence assumption.

### *2.5.4 Normality Assumption Checking:*

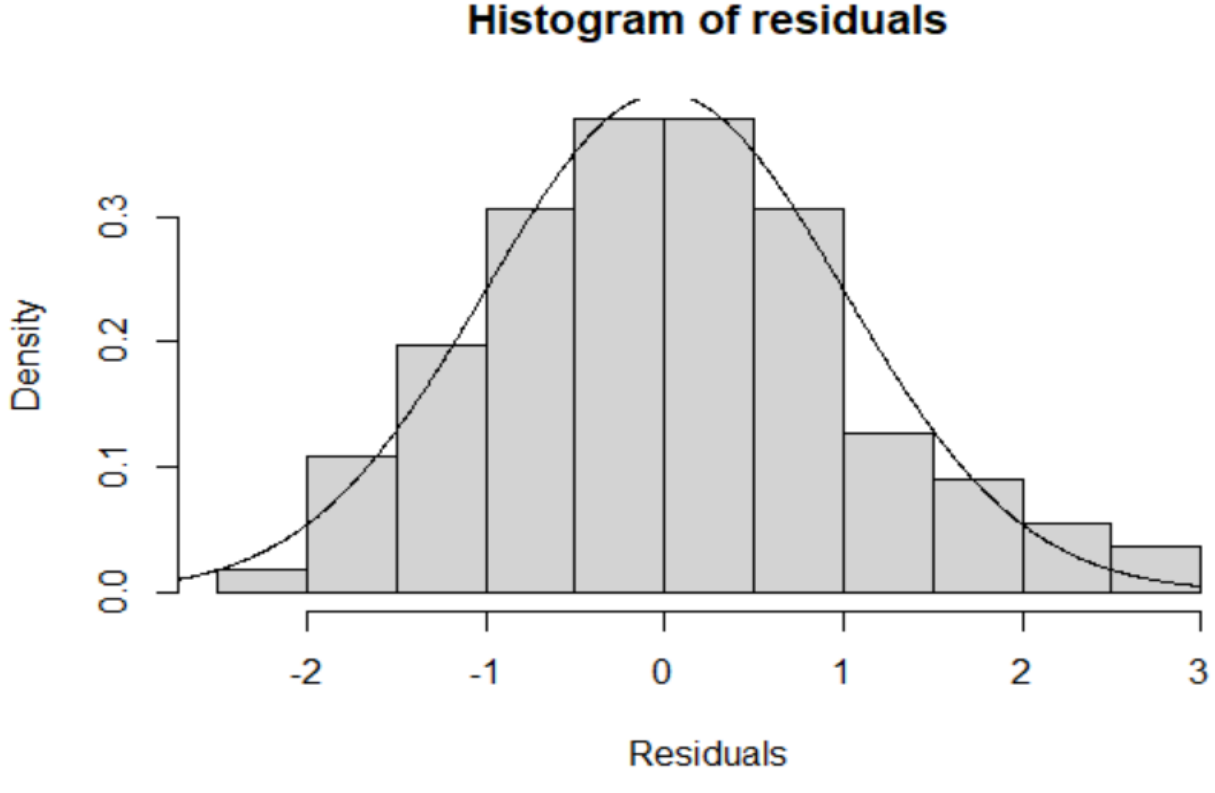



Figure 29. Histogram of Residuals.

From the Histogram of residuals, it looks normally distributed but has a heavy tail on the right side.

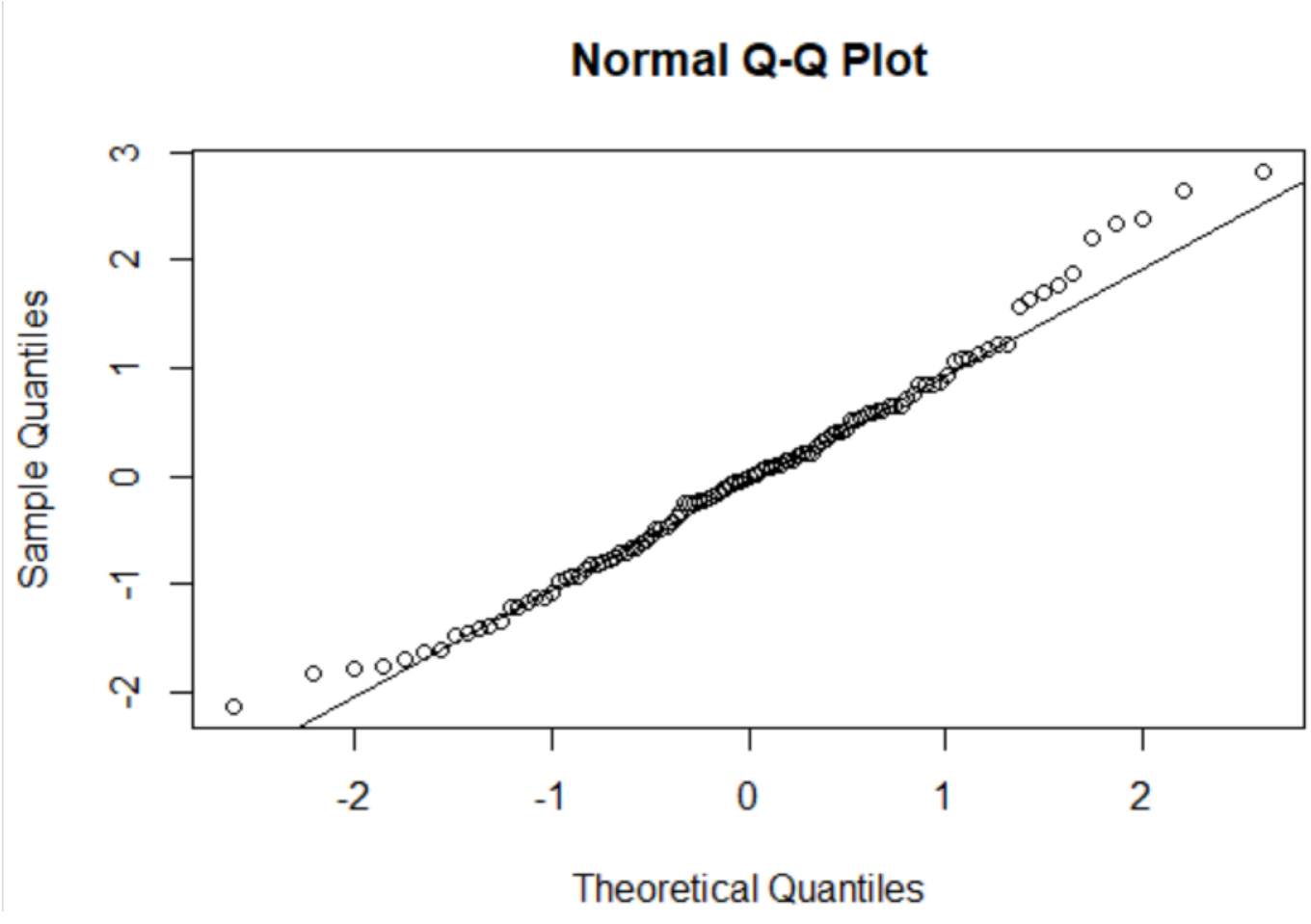



Figure 30. Normal Q-Q Plot of Residuals.

Based on the Normal Q-Q plot, there is no severe departure from the normality, but it is having heavy tails so we will check with the Shapiro-Wilk test for confirmation.

```
        Shapiro-Wilk normality test

data:  data.res
W = 0.98417, p-value = 0.2141
```

Figure 31. Shapiro-Wilk test output.

Based on the Shapiro-Wilk test, the p-value(0.2141) is greater than the significance value(i.e.,0.10), failing to reject the null hypothesis and we can say that our data is normally distributed.

### *2.5.5 Outlier Checking:*

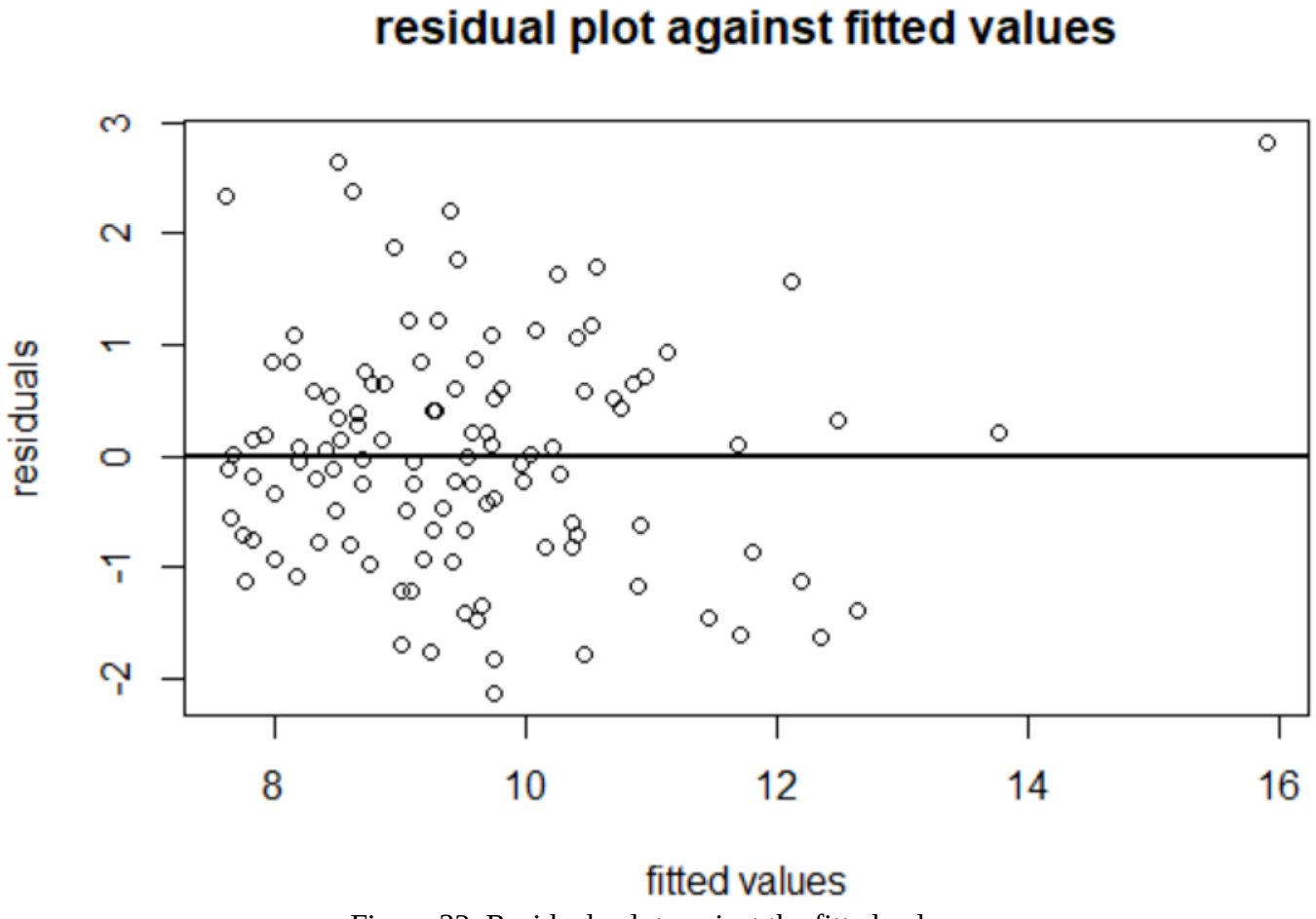



Figure 32. Residuals plot against the fitted values.

```
> which(abs(data.res) >3)
named integer(0)
```

From the residual plot against the fitted value, we can see there is no outlier.

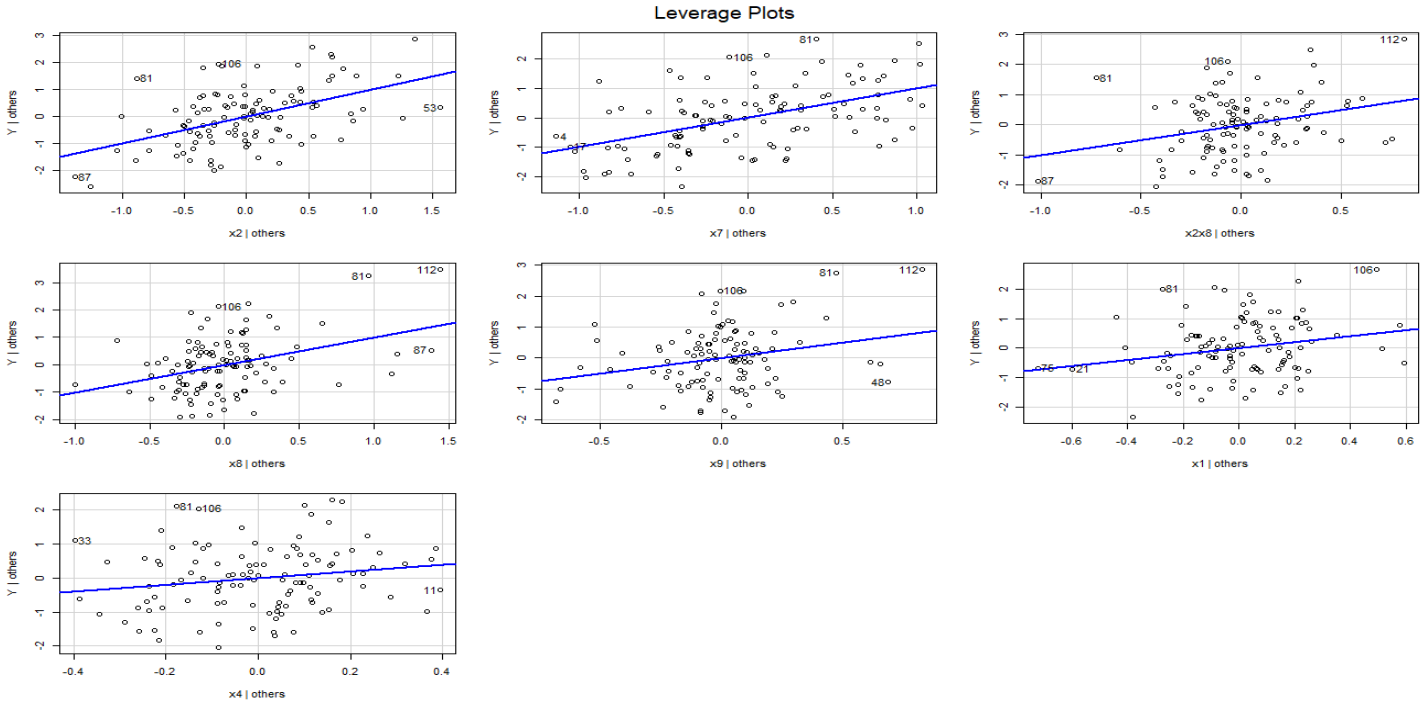
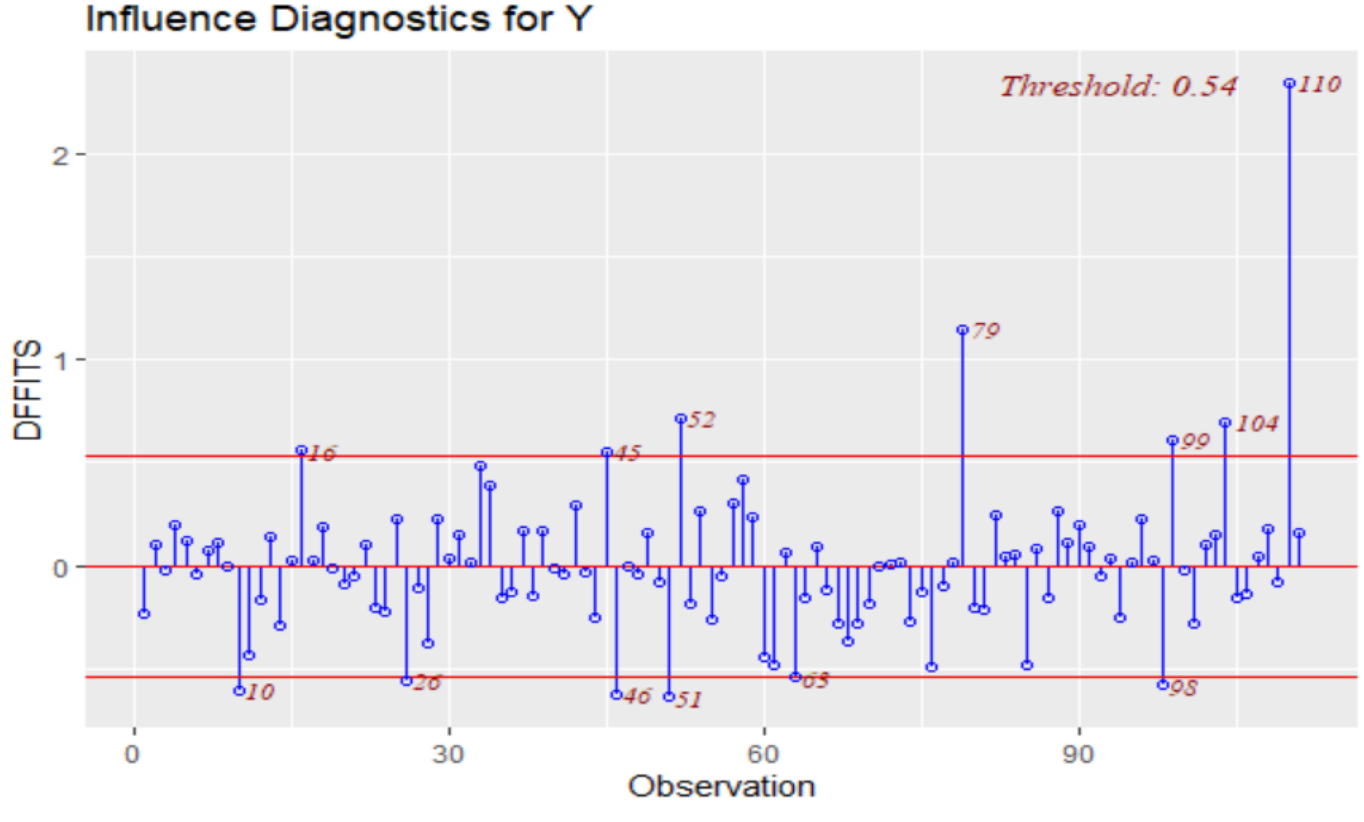



Figure 33. Leverage plots of predictor variables and interaction terms.

From the above leverage Plots, we can see that there are still some influential points present



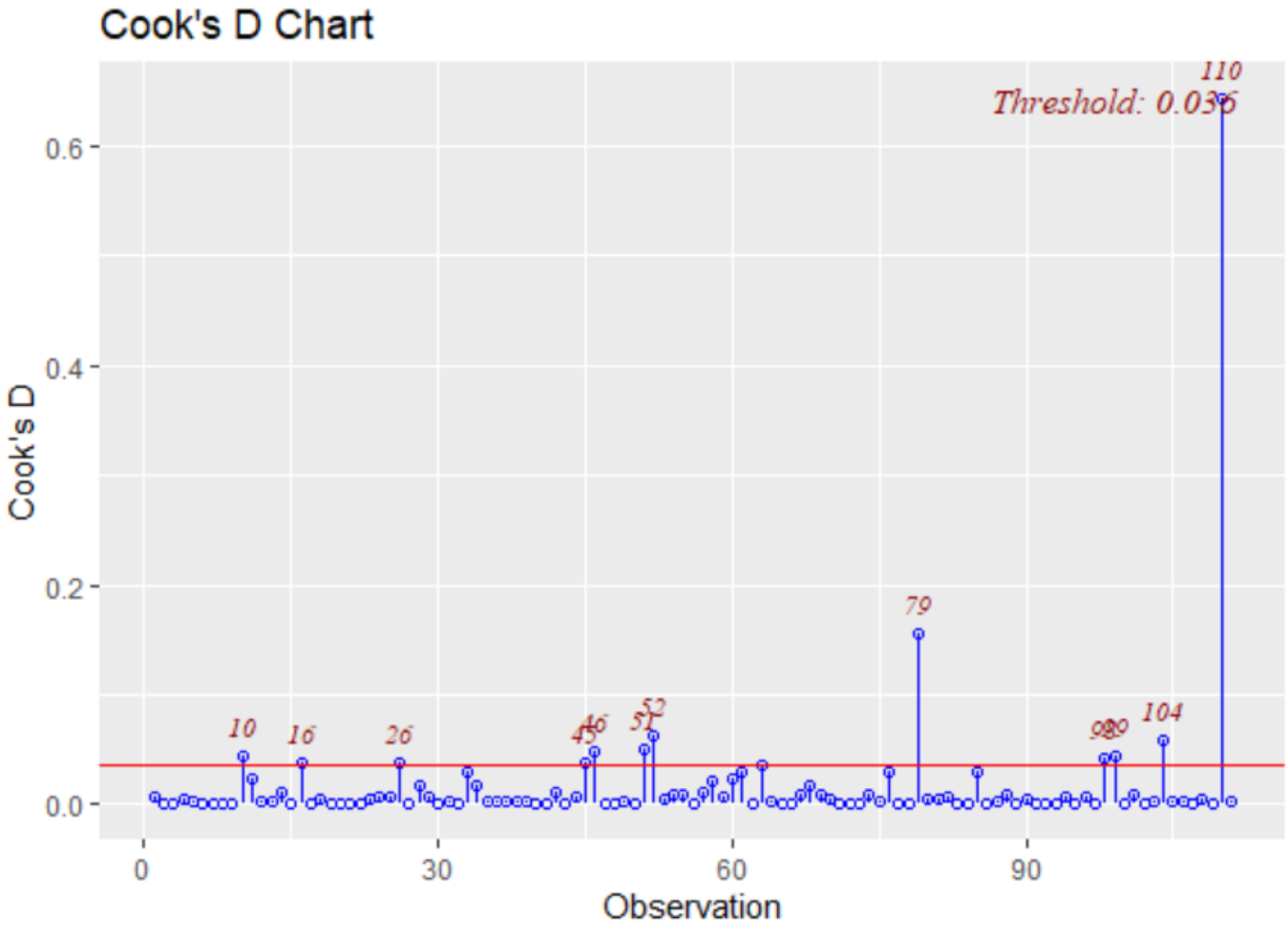



Figure 34. DFFITS plot and Cook's D Plot for the final model.

Based on the DFFITS, and Cook's D, tests, there are still few influential data points based on the threshold levels. But we can't remove all the potential outliers otherwise we will not able to give a good prediction.
Based on one scenario we have to remove and check the assumption if it is not affecting then its fine.
.

### *2.5.6 Multicollinearity Checking:*

By Variance Inflation Factor (VIF) analysis.

```
> vif(model)
      x2       x7     x2x8       x8       x9       x1       x4
1.690313 1.118002 1.527197 7.115137 6.365684 1.115340 1.351851
```

Based on the VIF value, there are no serious multicollinearity problems exist.

# 3. Result

The final model which I selected best fits our data is shown below in the summary output (Figure 35). This model contains the predictor variables such as Infection Risk (X2), Region (X7), Infection Risk (X2) * Average daily Census (X8), Average daily Census (X8), Number of nurses (X9), Age (X1), and Routine Chest X-ray Ratio (X4).

Our Estimated regression function:
y^ = 9.40958 + 0.69323 x2 - 0.57588 x7 + 0.40010 x2x8 + 0.99337 x8 - 0.60021 x9 + 0.23081 x1 + 0.20108 x4.

Based on the p-value (i.e., <2.2e-16), our model is having an overall significant relationship between the response variable and the predictors, and it provides a better fit than the intercept-only model.

The adjusted $R^2$ value is 0.6634, which means that about 66% of the variation in the "length of stay" data can be described by our model.

```
Call:
lm(formula = Y ~ x2 + x7 + x2x8 + x8 + x9 + x1 + x4, data = senic.itact.Std_ot)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9639 -0.6891 -0.0149  0.5787  2.2829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.40958    0.10411  90.379  < 2e-16 ***
x2           0.69323    0.12110   5.724 1.03e-07 ***
x7          -0.57588    0.09788  -5.883 5.04e-08 ***
x2x8         0.40010    0.13067   3.062  0.00281 **
x8           0.99337    0.24520   4.051 9.89e-05 ***
x9          -0.60021    0.23197  -2.587  0.01106 *
x1           0.23081    0.09932   2.324  0.02210 *
x4           0.20108    0.10791   1.863  0.06525 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9699 on 103 degrees of freedom
Multiple R-squared:  0.6848,    Adjusted R-squared:  0.6634
F-statistic: 31.97 on 7 and 103 DF,  p-value: < 2.2e-16
```

Figure 35. The summary output of the final model.

Figure 35 shows the t values for each estimated coefficient. We can see that each predictor variable is significant because in every case, the p-value is less than the significance level (i.e., 0.10).

**Interpretations of Estimated Coefficients:**

**β1**: When the Infection Risk (X2) increases by a unit and all the other variables are held constant, then the average Length of Stay (Y) will increase by 1 day.

**β2**: When the Region (X7) increases by a unit and all the other variables are held constant, then the average Length of Stay (Y) will decrease by 1 day.

**β3**: When the interaction term (Infection Risk (X2) * Average daily Census (X8)) increases by a unit and all the other variables are held constant, then the average Length of Stay (Y) will increase by 1 day.

**β4**: When the Average daily Census (X8) increases by a unit and all the other variables are held constant, then the average Length of Stay (Y) will increase by 1 day.

**β5**: When the Number of nurses (X9) increases by a unit and all the other variables are held constant, then the average Length of Stay (Y) will decrease by 1 day.

**β6**: When the Age (X1) increases by a unit and all the other variables are held constant, then the average Length of Stay (Y) will increase by 1 day.

**β7**: When the Routine Chest X-ray Ratio (X4) increases by a unit and all the other variables are held constant, then the average Length of Stay (Y) will increase by 1 day.

```
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
x2          1 91.964  91.964 97.7619 < 2.2e-16 ***
x7          1 53.684  53.684 57.0687 1.781e-11 ***
x2x8        1 40.467  40.467 43.0189 2.211e-09 ***
x8          1 10.114  10.114 10.7514  0.001422 **
x9          1  6.115   6.115  6.5002  0.012259 *
x1          1  4.876   4.876  5.1834  0.024872 *
x4          1  3.266   3.266  3.4724  0.065249 .
Residuals 103 96.891   0.941
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 36. Analysis of Variance Table (ANOVA) output of the final model.

Figure 36 shows the F value and P-value for each predictor variable. And in each predictor variable, the p-value is less than 0.10, so we can conclude that each predictor variables are significant and the relationship between the response variable and predictors variable is linear.

# 4. Conclusion

In Conclusion, I can say that based on the provided dataset I have created a good model as we can see it satisfied all the assumptions which are required for the linear model creation. And in our final model, all the predictor variables are significant to the response variable. Based on the adjusted $R^2$ value, about 66% of the variation in the "length of stay" data can be described by our model.

My finding from the regression analysis is that it's better to consider other forms of the model also for this dataset analysis because it's good if our model satisfied all required assumptions and give at least 80 to 85 percent accuracy then it makes sense to use it in the real-world application which is 66 percent in our case may be because of small sample size it is happening. But I find it's not a good idea to use it because of the average adjusted $R^2$.

To improve the model for the better fit, we can try to build the model by considering the polynomial regression, in this way maybe it will create a better model. Another thing we can try to remove some more outliers, as we see after applying remedial also there are still many outliers shown by the DFFIT, Cook D test. But due to the small sample size, we have to consider them one by one or if we get more data then we will be able to build a better model.

# 5. Appendix: R-Code

```
# MA4710 Final Project
# Created by Udit Joshi

############################### Read a data into R ##############################
senic <- read.csv("C:/Users/Uditnjoshi/Desktop/Fall_Sem/MA4710/Assignment/Project/SENIC.csv",
header=TRUE)
View(senic)
head(senic)

############################## 1. Introduction  ###################################

###################### 1.1 Exploratory Data Analysis. ##########################
## 1. Histograms of Y and Xs
hist(senic$Y,xlab="Length of Stay (Y)", main="Histogram of Y")

par(mfrow=c(3,4))
hist(senic$X1,xlab="Age (X1)", main="Histogram of X1")
hist(senic$X2,xlab="Infection Risk (X2)", main="Histogram of X2")
hist(senic$X3,xlab="Routine Culturing Ratio (X3)", main="Histogram of X3")
hist(senic$X4,xlab="Routine Chest X-ray Ratio (X4)", main="Histogram of X4")
hist(senic$X5,xlab="Number of Beds (X5)", main="Histogram of X5")
hist(senic$X6,xlab="Medical School (X6)", main="Histogram of X6")
hist(senic$X7,xlab="Region (X7)", main="Histogram of X7")
hist(senic$X8,xlab="Average daily Census (X8)", main="Histogram of X8")
hist(senic$X9,xlab="Number of nurses (X9)", main="Histogram of X9")
hist(senic$X10,xlab="Available facilities and services (X10)", main="Histogram of X10")


## 2. Boxplots of Y and Xs
boxplot(senic$Y, xlab="Y", main="Boxplot of Y")

par(mfrow=c(2,5))
boxplot(senic$X1, xlab="X1", main="Boxplot of X1")
boxplot(senic$X2, xlab="X2", main="Boxplot of X2")
boxplot(senic$X3, xlab="X3", main="Boxplot of X3")
boxplot(senic$X4, xlab="X4", main="Boxplot of X4")
boxplot(senic$X5, xlab="X5", main="Boxplot of X5")
boxplot(senic$X6, xlab="X6", main="Boxplot of X6")
boxplot(senic$X7, xlab="X7", main="Boxplot of X7")
boxplot(senic$X8, xlab="X8", main="Boxplot of X8")
boxplot(senic$X9, xlab="X9", main="Boxplot of X9")
boxplot(senic$X10, xlab="X10", main="Boxplot of X10")


## 3. Summary Statistics
summary(senic)
```

```
## 4. Scatter Plot Matrix
pairs(senic, col="red", 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)
senic.lmfit <- lm(Y ~ X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data = senic)
avPlots(senic.lmfit)


## 6. correlation matrix
cor(senic)

install.packages("corrplot")
library(corrplot)
corrplot(cor(senic))



############################## 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)

## 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)

par(mfrow=c(1,1))
#### Stepwise Regression
```

```
#### Adjusted R2
b = bestsubset(data=senic,y="Y",select="adjRsq",best=5)
print(b)
stepwise(data=senic,y="Y",select="adjRsq")
plot(b[,1:2])

#### Cp
b = bestsubset(data=senic,y="Y",select="CP",best=5)
print(b)
stepwise(data=senic,y="Y",select="CP")
plot(b[,1:2])

#### AIC
b = bestsubset(data=senic,y="Y",select="AIC",best=5)
print(b)
stepwise(data=senic,y="Y",select="AIC")
plot(b[,1:2])

#### BIC
b = bestsubset(data=senic,y="Y",select="BIC",best=5)
print(b)
stepwise(data=senic,y="Y",select="BIC")
plot(b[,1:2])

# Now try with the interaction term and include only those variable which is selected from Stepwise Regression.
#i.e., "X2", "X7", "X8", "X9", "X1", "X4", "X5"

# Model with interaction terms
X2X7 <- senic$X2*senic$X7 ;
X2X8 <- senic$X2*senic$X8 ; X7X8 <- senic$X7*senic$X8 ;
X2X9 <- senic$X2*senic$X9 ; X7X9 <- senic$X7*senic$X9 ; X8X9 <-senic$X8*senic$X9 ;
X2X1 <- senic$X2*senic$X1 ; X7X1 <- senic$X7*senic$X1 ; X8X1 <-senic$X8*senic$X1 ; X9X1 <-
senic$X9*senic$X1 ;
X2X4 <- senic$X2*senic$X4 ; X7X4 <- senic$X7*senic$X4 ; X8X4 <-senic$X8*senic$X4 ; X9X4 <-
senic$X9*senic$X4 ; X1X4 <-senic$X1*senic$X4 ;
X2X5 <- senic$X2*senic$X5 ; X7X5 <- senic$X7*senic$X5 ; X8X5 <-senic$X8*senic$X5 ; X9X5 <-
senic$X9*senic$X5 ; X1X5 <-senic$X1*senic$X5 ; X4X5 <-senic$X4*senic$X5 ;

senic.itact <-  cbind(senic,
                  X2X7 ,
                  X2X8 , X7X8 ,
                  X2X9 , X7X9 , X8X9 ,
                  X2X1 , X7X1 , X8X1 , X9X1 ,
                  X2X4 , X7X4 , X8X4 , X9X4 , X1X4 ,
                  X2X5 , X7X5 , X8X5 , X9X5 , X1X5 , X4X5)
head(senic.itact)
cor(senic.itact)
corrplot(cor(senic.itact))          #correlation matrix with interaction term

# 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$X9)
x10 <- (senic$X10 -mean(senic$X10))/sd(senic$X10)

x2x7 <- x2*x7 ;
x2x8 <- x2*x8 ; x7x8 <- x7*x8 ;
x2x9 <- x2*x9 ; x7x9 <- x7*x9 ; x8x9 <-x8*x9 ;
x2x1 <- x2*x1 ; x7x1 <- x7*x1 ; x8x1 <-x8*x1 ; x9x1 <-x9*x1 ;
x2x4 <- x2*x4 ; x7x4 <- x7*x4 ; x8x4 <-x8*x4 ; x9x4 <-x9*x4 ; x1x4 <-x1*x4 ;
x2x5 <- x2*x5 ; x7x5 <- x7*x5 ; x8x5 <-x8*x5 ; x9x5 <-x9*x5 ; x1x5 <-x1*x5 ; x4x5 <-x4*x5 ;

senic.itact.Std <- cbind(senic$Y,x1,x2,x4,x5,x7,x8,x9,
                         x2x7 ,
                         x2x8 , x7x8 ,
                         x2x9 , x7x9 , x8x9 ,
                         x2x1 , x7x1 , x8x1 , x9x1 ,
                         x2x4 , x7x4 , x8x4 , x9x4 , x1x4 ,
                         x2x5 , x7x5 , x8x5 , x9x5 , x1x5 , x4x5)
colnames(senic.itact.Std)[1] <- "Y"
head(senic.itact.Std)
cor(senic.itact.Std)
corrplot(cor(senic.itact.Std))        #correlation matrix after Standardization
#Now very few of them are highly correlated to each other compare to without standardization.

senic.itact.Std <- as.data.frame(senic.itact.Std)    # Converting to data Frame.
head(senic.itact.Std)


################ Fit a regression model with interaction terms
full.lmfit <- lm(Y ~ x1+x2+x4+x5+x7+x8+x9+
                 x2x7 +
                 x2x8 + x7x8 +
                 x2x9 + x7x9 + x8x9 +
                 x2x1 + x7x1 + x8x1 + x9x1 +
                 x2x4 + x7x4 + x8x4 + x9x4 + x1x4 +
                 x2x5 + x7x5 + x8x5 + x9x5 + x1x5 + x4x5, data = senic.itact.Std)
summary(full.lmfit)

################ Fit a regression model with no interaction terms
reduced.lmfit <- lm(Y ~ x1+x2+x4+x5+x7+x8+x9, data=senic.itact.Std)
summary(reduced.lmfit)

################ Test for significance of the interaction terms
anova(reduced.lmfit, full.lmfit)
```

```
## Again Model Selection (Stepwise Regression) after including interaction terms.
#### Stepwise Regression
#### Adjusted R2
stepwise(data=senic.itact.Std,y="Y",select="adjRsq")

#### Cp
stepwise(data=senic.itact.Std,y="Y",select="CP")

#### AIC
stepwise(data=senic.itact.Std,y="Y",select="AIC")

#### BIC
stepwise(data=senic.itact.Std,y="Y",select="BIC")


################ Fit a regression model with interaction terms
full.lmfit <- lm(Y ~ x2 + x7 + x2x8 + x2x5 + x8 + x9 + x1 + x4 + x2x4 + x8x5, data = senic.itact.Std)
summary(full.lmfit)
anova(full.lmfit)

#In the summary table we have seen one interaction variable is have higher p-value than 0.1 so we will drop this column
full.lmfit <- lm(Y ~ x2 + x7 + x2x8 + x2x5 + x8 + x9 + x1 + x4 + x2x4, data = senic.itact.Std)
summary(full.lmfit)

################ Fit a regression model with no interaction terms
reduced.lmfit <- lm(Y ~ x2 + x7 + x8 + x9 + x1 + x4, data=senic.itact.Std)
summary(reduced.lmfit)

################ Test for significance of the interaction terms
anova(reduced.lmfit, full.lmfit)

# Based on the anova test, we will go with the full model as our final model.
#### Final Model
reduced.lmfit <- lm(Y ~ x2 + x7 + x2x8 + x2x5 + x8 + x9 + x1 + x4 + x2x4, data=senic.itact.Std)
summary(reduced.lmfit)



######################### 3. Assumption Checking ################################
##### 3.1 Linearity assumption checking. #####
data.fitted <- fitted(reduced.lmfit)
data.res <- rstudent(reduced.lmfit)                #jackknifed residual

# Residual plot against fitted values.
plot(data.res ~ data.fitted, xlab="fitted values", ylab="residuals", main="residual plot against fitted values")

## Residual plot against predictor variables.
par(mfrow=c(3,3))
```

```
plot(data.res ~ senic.itact.Std$x2, xlab="x2", ylab="residuals", main="residual plot against x2")
plot(data.res ~ senic.itact.Std$x7, xlab="x7", ylab="residuals", main="residual plot against x7")
plot(data.res ~ senic.itact.Std$x2x8, xlab="x2x8", ylab="residuals", main="residual plot against x2x8")
plot(data.res ~ senic.itact.Std$x2x5, xlab="x2x5", ylab="residuals", main="residual plot against x2x5")
plot(data.res ~ senic.itact.Std$x8, xlab="x8", ylab="residuals", main="residual plot against x8")
plot(data.res ~ senic.itact.Std$x9, xlab="x9", ylab="residuals", main="residual plot against x9")
plot(data.res ~ senic.itact.Std$x1, xlab="x1", ylab="residuals", main="residual plot against x1")
plot(data.res ~ senic.itact.Std$x4, xlab="x4", ylab="residuals", main="residual plot against x4")
plot(data.res ~ senic.itact.Std$x2x4, xlab="x2x4", ylab="residuals", main="residual plot against x2x4")


##### 3.2 Constant Variance assumption checking.  #####
## Residual plot against fitted values.
plot(data.res ~ data.fitted, xlab="fitted values", ylab="residuals", main="residual plot against fitted values")
abline(h=0,lwd=2)

## Residual plot against predictor variables.
par(mfrow=c(3,3))
plot(data.res ~ senic.itact.Std$x2, xlab="x2", ylab="residuals", main="residual plot against x2")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x7, xlab="x7", ylab="residuals", main="residual plot against x7")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x2x8, xlab="x2x8", ylab="residuals", main="residual plot against x2x8")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x2x5, xlab="x2x5", ylab="residuals", main="residual plot against x2x5")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x8, xlab="x8", ylab="residuals", main="residual plot against x8")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x9, xlab="x9", ylab="residuals", main="residual plot against x9")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x1, xlab="x1", ylab="residuals", main="residual plot against x1")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x4, xlab="x4", ylab="residuals", main="residual plot against x4")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x2x4, xlab="x2x4", ylab="residuals", main="residual plot against x2x4")
abline(h=0,lwd=2)

## ABS Residual plot against predictor variables.
par(mfrow=c(3,3))
plot(abs(data.res) ~ senic.itact.Std$x2, xlab="x2", ylab="|residuals|", main="ABS residual plot against x2")
plot(abs(data.res) ~ senic.itact.Std$x7, xlab="x7", ylab="|residuals|", main="ABS residual plot against x7")
plot(abs(data.res) ~ senic.itact.Std$x2x8, xlab="x2x8", ylab="|residuals|", main="ABS residual plot against x2x8")
plot(abs(data.res) ~ senic.itact.Std$x2x5, xlab="x2x5", ylab="|residuals|", main="ABS residual plot against x2x5")
plot(abs(data.res) ~ senic.itact.Std$x8, xlab="x8", ylab="|residuals|", main="ABS residual plot against x8")
plot(abs(data.res) ~ senic.itact.Std$x9, xlab="x9", ylab="|residuals|", main="ABS residual plot against x9")
plot(abs(data.res) ~ senic.itact.Std$x1, xlab="x1", ylab="|residuals|", main="ABS residual plot against x1")
plot(abs(data.res) ~ senic.itact.Std$x4, xlab="x4", ylab="|residuals|", main="ABS residual plot against x4")
plot(abs(data.res) ~ senic.itact.Std$x2x4, xlab="x2x4", ylab="|residuals|", main="ABS residual plot against x2x4")

##### BP test.
library(lmtest)
```

```
bptest(reduced.lmfit)


##### 3.3 Independence assumption checking. #####
## Residual plot against time
## Data is not time series related. So, we can't check it.


##### 3.4 Normality assumption checking. #####
## Boxplot of residuals
boxplot(data.res, xlab="residual", main="Boxplot of residuals")

hist(data.res, main="Histogram of residuals", xlab="Residuals",probability=TRUE)
lines(seq(-3,3,length.out = 1000),dnorm(seq(-3,3,length.out = 1000)))

## QQ Plot
qqnorm(data.res); qqline(data.res)


##### Shapiro-Wilk test
shapiro.test(data.res)


##### 3.5 Outliers checking. #####
## Residual plot against fitted values
plot(data.res ~ data.fitted, xlab="fitted values", ylab="residuals", main="residual plot against fitted values")
abline(h=0,lwd=2)

which(abs(data.res) >3)

## Residual plot against predictor variables
par(mfrow=c(3,3))
plot(data.res ~ senic.itact.Std$x2, xlab="x2", ylab="residuals", main="residual plot against x2")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x7, xlab="x7", ylab="residuals", main="residual plot against x7")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x2x8, xlab="x2x8", ylab="residuals", main="residual plot against x2x8")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x2x5, xlab="x2x5", ylab="residuals", main="residual plot against x2x5")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x8, xlab="x8", ylab="residuals", main="residual plot against x8")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x9, xlab="x9", ylab="residuals", main="residual plot against x9")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x1, xlab="x1", ylab="residuals", main="residual plot against x1")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x4, xlab="x4", ylab="residuals", main="residual plot against x4")
abline(h=0,lwd=2)
plot(data.res ~ senic.itact.Std$x2x4, xlab="x2x4", ylab="residuals", main="residual plot against x2x4")
abline(h=0,lwd=2)
```

```
## Boxplot of residuals
boxplot(data.res, xlab="residual", main="Boxplot of residuals")

library(car)
leveragePlots(reduced.lmfit)


############ Detection methods of influential points
install.packages("olsrr")
library(olsrr)

## 1. DFFITS
ols_plot_dffits(reduced.lmfit)

## 2. Cook's D
ols_plot_cooksd_chart(reduced.lmfit)

## 3. DFBETAS
ols_plot_dfbetas(reduced.lmfit)


##### 3.6 Diagnose Multicollinearity. #####
summary(reduced.lmfit)
library(dplyr)
install.packages("corrplot")
library("corrplot")
corrplot(cor(select(senic.itact.Std,Y, x2 , x7 , x2x8 ,x2x5, x8 , x9 , x1 , x4, x2x4)))
anova(reduced.lmfit)

# Check Variance Inflation Factor(VIF).
library(car,carData)
vif(reduced.lmfit)
# Based on VIF we can say that interaction variables x2x8 and x2x5 are having
# high correlation so remove one of them from the model.

reduced.lmfit <- lm(Y ~ x2 + x7 + x2x8 + x8 + x9 + x1 + x4+ x2x4 , data=senic.itact.Std)
vif(reduced.lmfit)
summary(reduced.lmfit)



########################## 4. Remedial Measures ###################################
##### 4.1 Multicollinearity remedial measure. #####
# Based on VIF we can say that interaction variables x2x8 and x2x5 are having
# highly correlated to each other so remove one of them from the model.
reduced.lmfit <- lm(Y ~ x2 + x7 + x2x8 + x8 + x9 + x1 + x4+ x2x4 , data=senic.itact.Std)
vif(reduced.lmfit)
summary(reduced.lmfit)
```

```
##### 4.2 Delete Outliers. #####
model <- lm(Y ~ x2 + x7 + x2x8 + x8 + x9 + x1 + x4+ x2x4, data=senic.itact.Std[-c( 47,43),])
which(abs(data.res) >3)


## Check the Model Overall Significance
summary(model)
model <- lm(Y ~ x2 + x7 + x2x8 + x8 + x9 + x1 + x4, data=senic.itact.Std[-c( 47,43),])
summary(model)
anova(model)



########### 5. Model Assumptions Checking after Remedial Measures. ##############
senic.itact.Std_ot <- senic.itact.Std[-c( 47,43),]
model <- lm(Y ~ x2 + x7 + x2x8 + x8 + x9 + x1 + x4, data=senic.itact.Std_ot)
summary(model)


##### 5.1 Linearity assumption checking. #####
data.fitted <- fitted(model)
data.res <- rstudent(model)                #jackknifed residual

# Residual plot against fitted values.
plot(data.res ~ data.fitted, xlab="fitted values", ylab="residuals", main="residual plot against fitted values")

## Residual plot against predictor variables.
par(mfrow=c(3,3))
plot(data.res ~ senic.itact.Std_ot$x2, xlab="x2", ylab="residuals", main="residual plot against x2")
plot(data.res ~ senic.itact.Std_ot$x7, xlab="x7", ylab="residuals", main="residual plot against x7")
plot(data.res ~ senic.itact.Std_ot$x2x8, xlab="x2x8", ylab="residuals", main="residual plot against x2x8")
plot(data.res ~ senic.itact.Std_ot$x2x5, xlab="x2x5", ylab="residuals", main="residual plot against x2x5")
plot(data.res ~ senic.itact.Std_ot$x8, xlab="x8", ylab="residuals", main="residual plot against x8")
plot(data.res ~ senic.itact.Std_ot$x9, xlab="x9", ylab="residuals", main="residual plot against x9")
plot(data.res ~ senic.itact.Std_ot$x1, xlab="x1", ylab="residuals", main="residual plot against x1")
plot(data.res ~ senic.itact.Std_ot$x4, xlab="x4", ylab="residuals", main="residual plot against x4")
plot(data.res ~ senic.itact.Std_ot$x2x4, xlab="x2x4", ylab="residuals", main="residual plot against x2x4")


##### 5.2 Constant Variance assumption checking.  #####
## Residual plot against fitted values.
plot(data.res ~ data.fitted, xlab="fitted values", ylab="residuals", main="residual plot against fitted values")
abline(h=0,lwd=2)


##### BP test.
library(lmtest)
bptest(model)


##### 5.3 Independence assumption checking. #####
```

```
## Residual plot against time
## Data is not time series related. So, we can't check it.


##### 5.4 Normality assumption checking. #####
## Boxplot of residuals
boxplot(data.res, xlab="residual", main="Boxplot of residuals")

hist(data.res, main="Histogram of residuals", xlab="Residuals",probability=TRUE)
lines(seq(-3,3,length.out = 1000),dnorm(seq(-3,3,length.out = 1000)))

## QQ Plot
qqnorm(data.res); qqline(data.res)


##### Shapiro-Wilk test
shapiro.test(data.res)


##### 5.5 Outliers checking. #####
## Residual plot against fitted values
plot(data.res ~ data.fitted, xlab="fitted values", ylab="residuals", main="residual plot against fitted values")
abline(h=0,lwd=2)

which(abs(data.res) >3)

library(car)
leveragePlots(model)


############ Detection methods of influential points
install.packages("olsrr")
library(olsrr)

## 1. DFFITS
ols_plot_dffits(model)

## 2. Cook's D
ols_plot_cooksd_chart(model)


##### 5.6 Diagnose Multicollinearity. #####
summary(model)
library(dplyr)
install.packages("corrplot")
library("corrplot")
corrplot(cor(select(senic.itact.Std_ot,Y, x2 , x7 , x2x8 , x8 , x9 , x1 , x4, x2x4)))
anova(model)

# Check Variance Inflation Factor(VIF).
library(car,carData)
```

```
vif(model)


################################ 3. Result ########################################
summary(model)
anova(model)
```