# Study of the SENIC dataset

# Contents

# 1 Introduction

## 1.1 Goal of the project

The study of the SENIC dataset has characteristics of hospitals participating in the study of the SENIC. The project's goals are to identify the best (optimal) model fitting to the SENIC.csv dataset, statistically interpret the model, and make a conclusion connected to the application based on the procedure of regression analysis.

## 1.2 Description of Dataset

The dataset consists of a sample of 113 hospitals (rows) and 11 variables. The description of each variable is as follows in ascending order:

1. **Length of Stay (Y):** Average length of stay of all patients in hospital (days).
2. **Age (X1):** Average age of patients (in years).
3. **Infection Risk (X2):** Estimated probability of acquiring infection in hospital (percent).
4. **Routine Culturing Ratio (X3):** Ratio of a number of cultures performed to the number of patients without signs or symptoms of hospital-acquired infection, times 100.
5. **Routine X-ray Ratio (X4):** Ratio of number of X-rays performed to a number of patients without signs or symptoms of pneumonia, times 100.
6. **Number of Beds (X5):** Number of beds in the hospital during the study period.
7. **Medical School (X6):** Indicator of whether the hospital is associated with a medical school (1 = Yes, 2 = No).
8. **Region (X7):** Indicator of the geographic region for hospital (1 = NE, 2 = NC, 3 = S, 4 = W).
9. **Average Census (X8):** Number of patients per day in hospital during the study period.
10. **Number Nurses (X9):** Number of full-time equivalents registered and licensed practical nurses during the study period (number of full times plus one half the number of part-time).
11. **Available Facilities (X10):** Percent of 35 potential facilities and services provided by the hospital.

As mentioned above, the dataset contains one response variable (Length of Stay: Y) against the ten predictor variables (Age, Infection Risk, Routine Culturing Ratio, Routine X-ray Ratio, Number of Beds, Medical School, Region, Average Census, Number Nurses, Available Facilities: X1, X2, X3, X4, X5, X6, X7, X8, X9, and X10). Here the representation of the response variable is shown below in the form of a box plot:

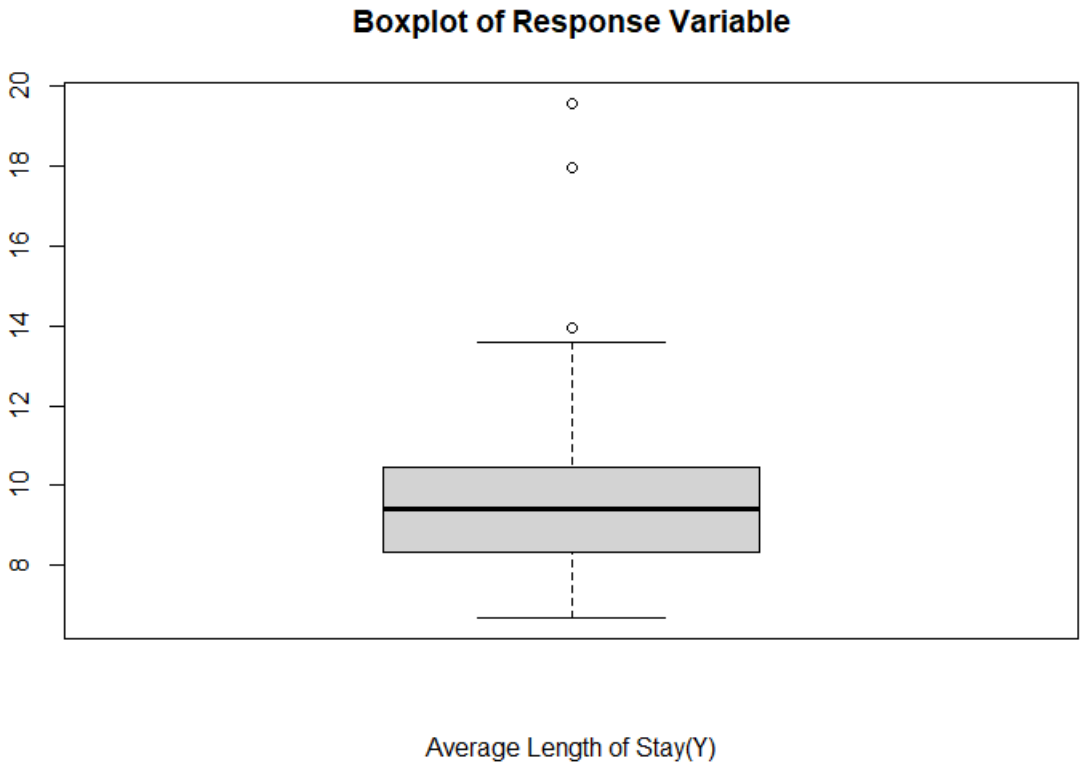



**Figure 1.1: Boxplot of Response Variable: Length of Stay (Y)**

The model is built based on the variables which provide optimal fitting on this measure of the response variable (Length of Stay), interpretation through statistically.

# 1.3 Exploratory Data Analysis

Let's analyze the data to study each predictor variable against the response variable through visualizations and statistical analysis. Here we considered the statement indicating which predictor variables are significant on the response variable at the significance level = 0.05. At the end of the study, we fit the model based on hypothesis testing and compare it with the model-fitting methods at the end of the project.

The histogram (Figure 1.2) of the variables explained the distribution, either symmetric or asymmetric and either right or left-skewed.

- The data of response variable (Y) and predictor variables (X3, X5, X8 and X9) are **right skewed**, i.e., the mean is larger than the median. Here for X8, fewer larger values bring the mean upwards, which doesn't affect the median much.
- The histogram of predictor variables (X6 and X7) isn't normally distributed.

- The histogram of predictor variables (X1 and X2) is normally distributed (symmetric) with low spread (variance).
- The histogram of predictor variables (X4 and X10) is symmetric with more spread (variance).

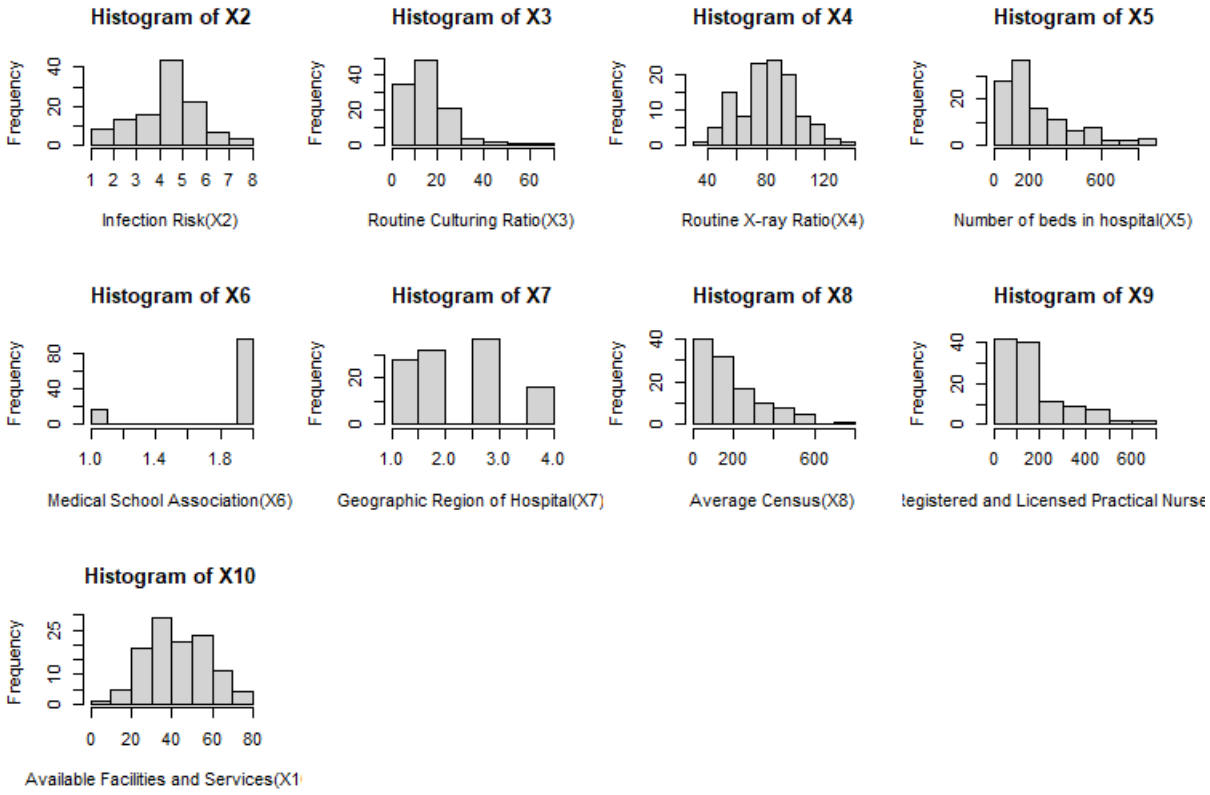



**Figure 1.2: Histogram of Predictor and Response Variables**

The box plot (Figure 1.3) visualizes the outliers which are far away from the whiskers of the response variable (Y) and predictor variables (X1, X2, X3, X4, X5, X7, X8, X9). We conclude from the histogram that the predictor variable (X7) doesn't follow any distribution. Still, the box plot visualization shows that it follows right-skewed distribution, i.e., the mean is larger than the median.

The summary statistics (Table 1.1) indicate the above analysis in terms of values (min, median, max, and quartiles).

```
> summary(senic_df)
       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  
```

**Table 1.1: Summary Statistics of SENIC dataset**

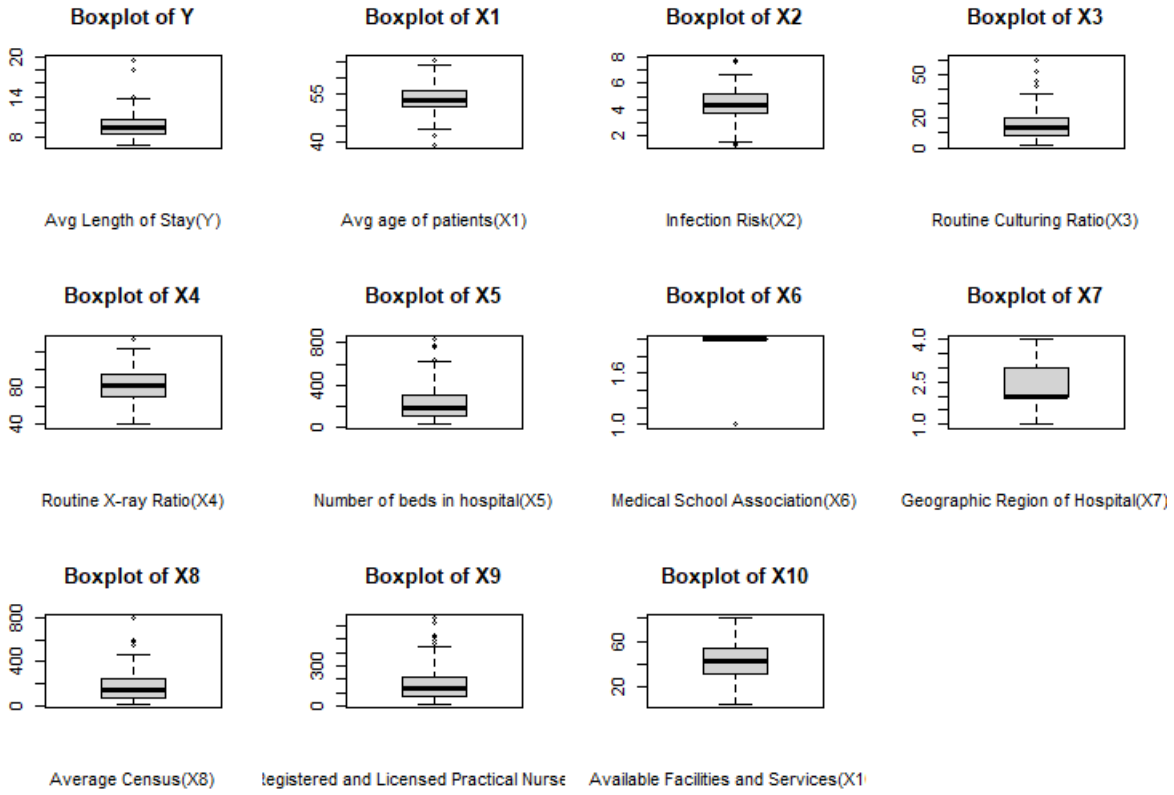



**Figure 1.3: Boxplot of Predictor and Response Variables**

From the pair plot (Figure 1.4), we can check the multicollinearity among the numeric variables. The pair plots of SENIC data show there is some risk from variables in the bottom right. The same analysis is confirmed using the added variable plots (Figure 1.5) and variance inflation factor (VIF) model fitting chapter.

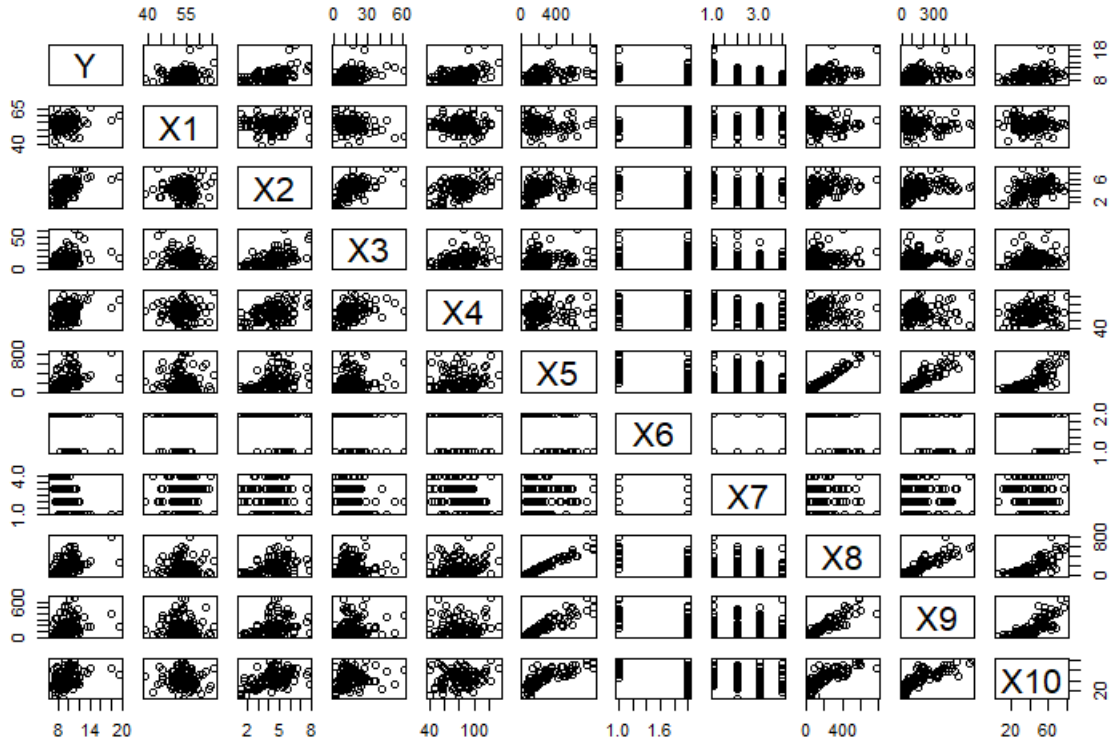
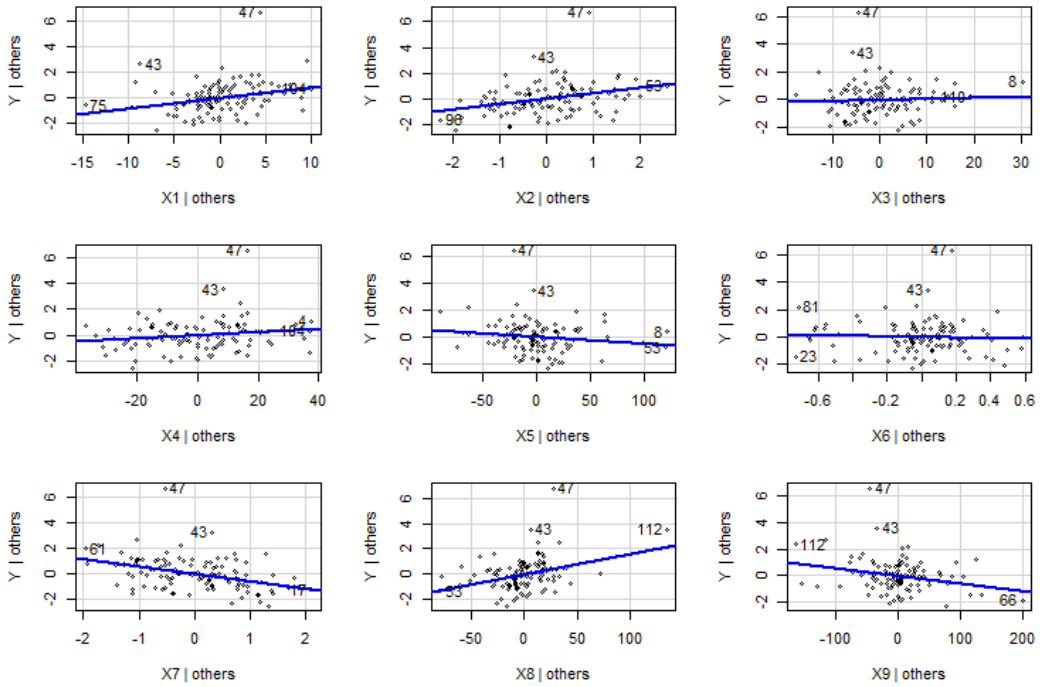



**Figure 1.4: Pair plot of numeric variables**



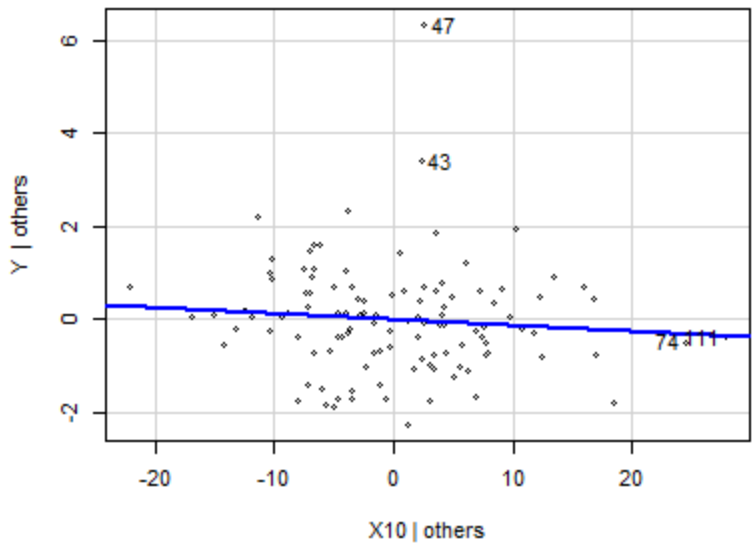



**Figure 1.5: Added variable plot of numeric variables**

# 1.4 Correlation Matrix

Correlation measures the linear relation between two variables and values between -1 and 1 where:

- -1 indicates a perfect negative linear correlation between the two variables
- 0 indicates a no linear correlation between two variables
- 1 indicates a perfect positive linear correlation between the two variables

From the correlation plot (Figure 1.6), we can say that multicollinearity exists among X5, X8, X9, and X10; there is a perfect negative linear correlation between Y and X6 & X7. The figure has an indication bar of the strength of correlation between each variable.

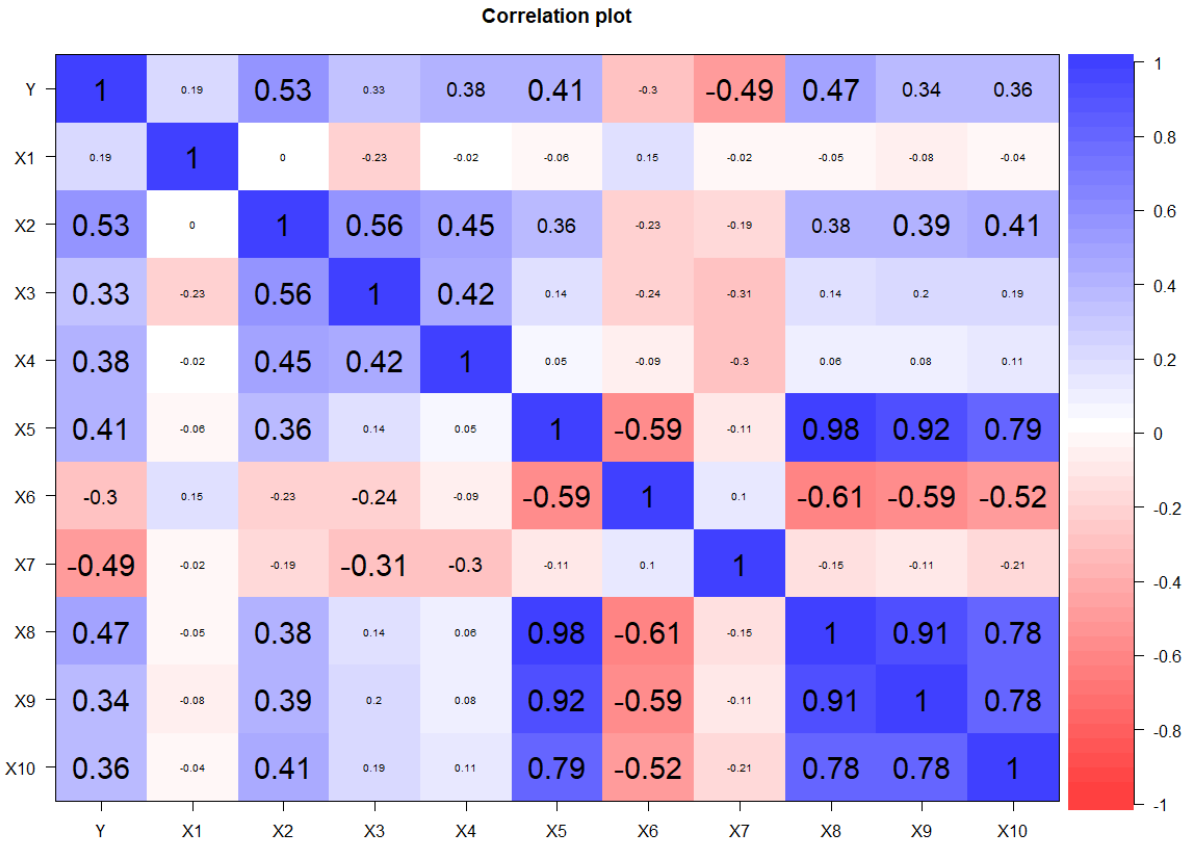



**Figure 1.6: Correlation Matrix of Numeric Variables**

# 1.5 Standardization or Centralization

Standardization of a dataset means to scale all values in the dataset to the mean value is 0, and the standard deviation is 1. And the centralization is centering a variable around its mean.

There are no null values in the dataset and verified by checking the mean and standard deviation of predictor and response variables. The results are available in Table 1.2. Also, standardization or centralization are not required for the SENIC dataset.

There is multicollinearity exists among the predictor variables but standardization is not necessary as it is not polynomial functon.

| | Y | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | 9.648 | 53.232 | 4.355 | 15.79 | 81.63 | 252.2 | 1.8496 | 2.363 | 191.4 | 173.2 | 43.16 |
| sd | 1.911 | 4.462 | 1.341 | 10.23 | 19.36 | 192.8 | 0.3591 | 1.009 | 153.8 | 139.3 | 15.20 |

**Table 1.2: Mean and Standard Deviation of SENIC dataset Variables**

# 2 Model/Methods

## 2.1 Model including all variables

A model including all predictor variables and responses between the categorical and numeric variables were created. The summary of the results is provided below:

```
> summary(senic_df.lmfit)

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

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
```

**Table 2.1: Summary Statistics of Full Model**

The fitted model including all variables is:

Y = 3.7204 + 0.0851X1 + 0.4264X2 + 0.0079X3 + 0.0125X4 - 0.0054X5 – 0.2041X6 - 0.5801X7 + 0.0159X8 - 0.0058X9 - 0.0126X10

Now perform the hypotheses testing and verify the significant linear relationship between response and predictor variables using p-value at the considerable level of 0.05.

$$H0: \beta1 = \beta2 = \beta3 = \beta4 = \beta5 = \beta6 = \beta7 = \beta8 = \beta9 = \beta10 = 0 \; vs. \; Ha: \; not \; H0$$

When it comes to the hypothesis testing for overall relations of the predictors to the response variable, because the p. value from the F-test is less than 0.05, we can conclude that the model that we consider has a better fit than an intercept-only model. (p-value: < 2.2e-16)

$$H0: \beta j = 0 \; vs. \; Ha: \; not \; H0, \; j = 0,1,2,3,4,5,6,7,8,9,10$$

From the summary, X3, X4, X5, X6, and X10 are insignificant to the average length of stay because the associated p. value is greater than 0.05. On the other hand, X1, X2, X7, and X8, X9 have a significant linear relationship with the average length of stay.

$R^2_{adj}$ = 0.5907; 59.07% of the total variation in Y can be explained by the predictor variables.

**Reduced Model**

The reduced model based on hypothesis testing is:

Y = 0.0851X1 + 0.4264X2 - 0.5801X7 + 0.0159X8 - 0.0058X9

- $b1$**:** The average length of stay would increase by 0.0851 as the age increases by one year when the other predictor variables are held constant.
- $b2$**:** The average length of stay would increase by 0.4264 as the infection risk increases by one unit when the other predictor variables are held constant.
- $b7$**:** The average length of stay would decrease by 0.5801 as the geographic region increases by one unit when the other predictor variables are held constant.
- $b8$**:** The average length of stay would increase by 0.0159 as the average census increases by a unit when the other predictor variables are held constant.
- $b9$**:** The average length of stay would decrease by 0.00558 as the number of nurses increases by a unit when the other predictor variables are held constant.

The Adjusted R-squared value for the reduced model has 57.65 % of the total variability in Y can be explained by the regression model using the reduced model (X1, X2, X7, X8, and X9) using p-value at the significant level of 0.05

## 2.2 Model Selection

In the last section, we rejected a few variables based on hypothesis testing at a significant level of 0.05. We performed the model selection using a stepwise function to select the optimal variables based on Adjusted $R^2$, Mallows' C*p*, AIC, and BIC. The analysis is provided below for each criterion.

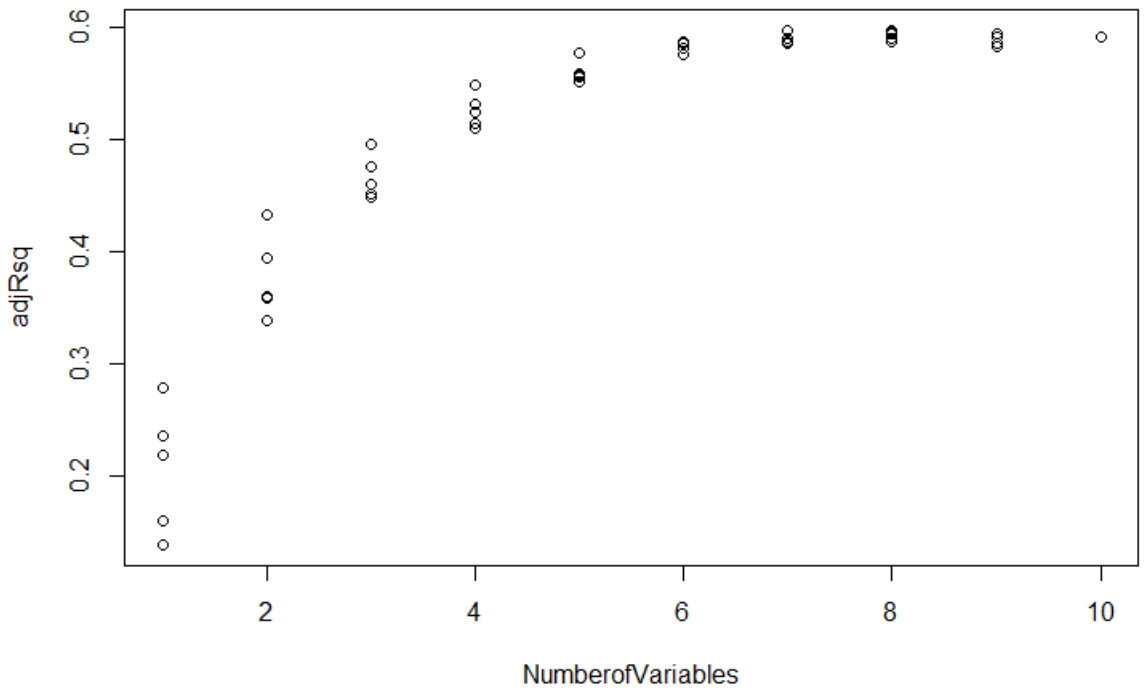



**Figure 2.1: Adjusted $R^2$ plot**

The adjusted $R^2$ plot (Figure 2.1) concludes that eight predictor variables (X1 X2 X4 X5 X7 X8 X9) should include in our model because that is where the points in the plot begin to level out. Also, the model has a high value of 0.5946.

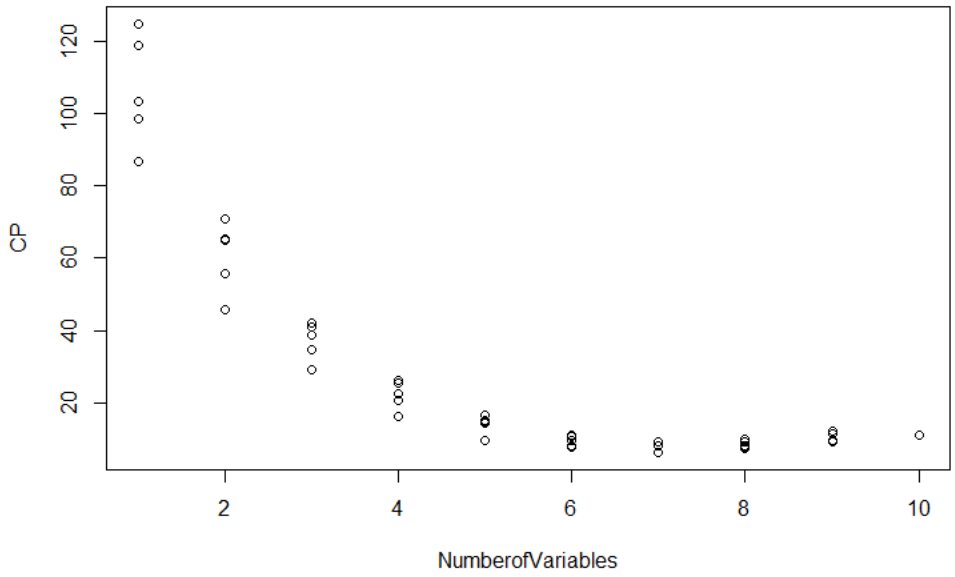



**Figure 2.2: Mallows' C*p* plot**

The Mallows' C*p* plot (Figure 2.2) concludes that six predictor variables (X1 X2 X4 X5 X7 X8 X9) should include in our model because that is where the points in the plot begin to level out. Also, the model has a low value of 6.4830

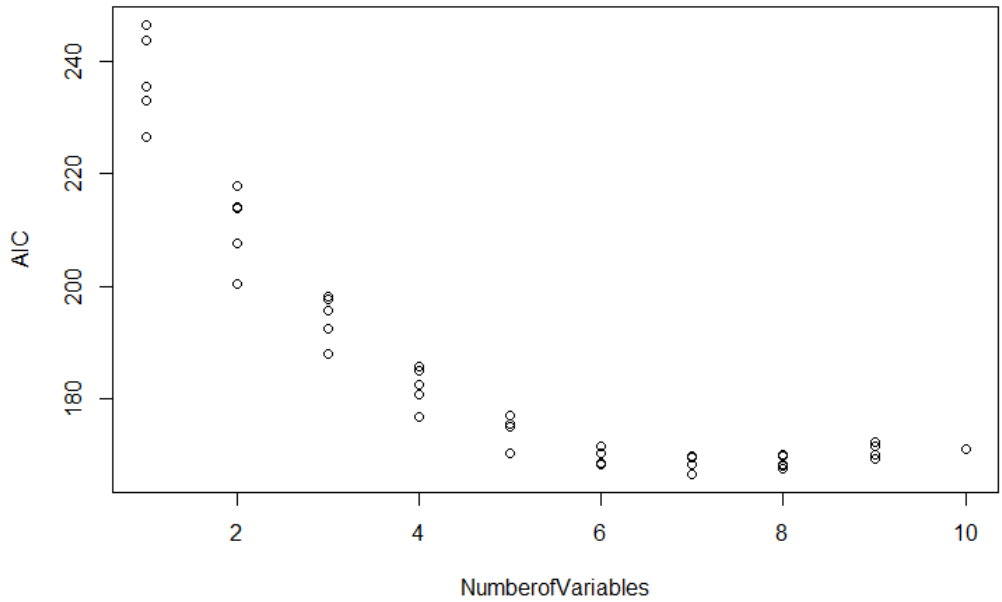



**Figure 2.3: AIC plot**

The AIC plot (Figure 2.3) concludes that six predictor variables (X1 X2 X4 X5 X7 X8 X9) should include in our model because that is where the points in the plot begin to level out. Also, the model has a low value of 169.4667

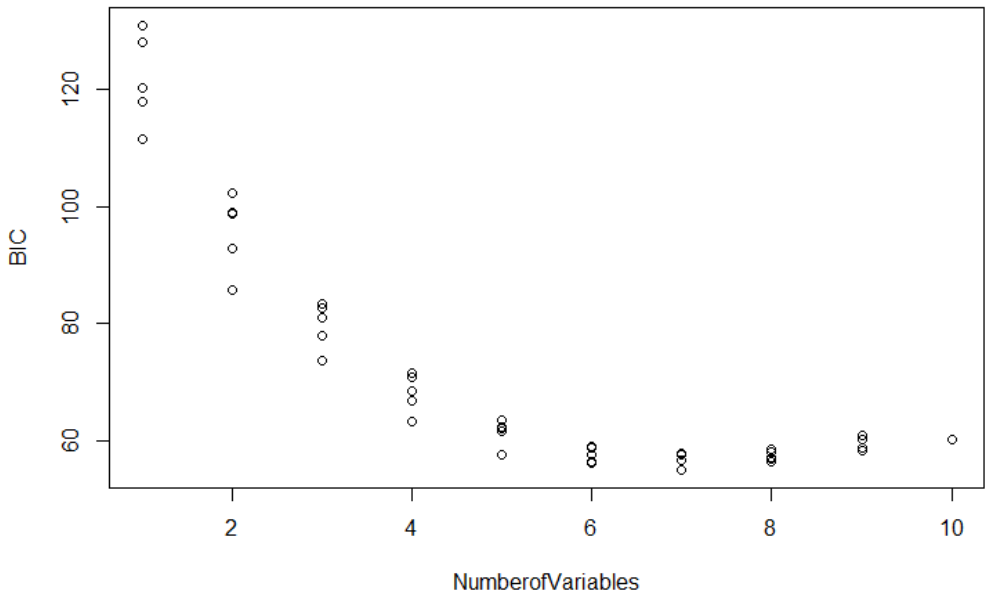



**Figure 2.4: BIC plot**

The BIC plot (Figure 2.4) concludes that six predictor variables (X1 X2 X4 X5 X7 X8 X9) should include in our model because that is where the points in the plot begin to level out. Also, the model has a low value of 54.9837

# 2.3 Best Subset Model

Based on testing the model selection criteria, the model with seven predictor variables (X1 X2 X4 X5 X7 X8 X9) is more appropriate to fit the data. Between the model selection based on p-value at the significant level of 0.05 and including seven predictor variables, the optimal model including seven predictor variables is more appropriate. This model has seven predictor variables with an Adjusted $R^2$ of 59.66%, and then the individual stepwise models have an Adjusted $R^2$ of 57.65%. Even though the p. value of X4 and X5 are greater than the significant level, there is a correlation that exists between Y, X4, and X5 (Refer figure 1.6).

Therefore, the model including seven predictor variables is the optimal model, and the fitted equation for the final model is:

Y = 3.2506 + 0.4358X2 - 0.5714X7 + 0.0165X8 - 0.0060X9 + 0.0789X1 + 0.0135X4 - 0.0062X5

# 2.4 Regression Diagnostics

In this topic, I have provided the details for checking the adequacy of a regression model. It includes methods for detecting outliers, influential points, normality, constant variance, multi-collinearity, Homoscedasticity and non-independence.

## 2.4.1 Homoscedasticity

The plot of the residuals against the approximation shows that the regression model is appropriate. This is because the model has a uniform and random distribution. Category variables are a little more complicated; determine if the distribution is random. To determine if homoscedasticity exists, The Breusch Pagan test has been run.

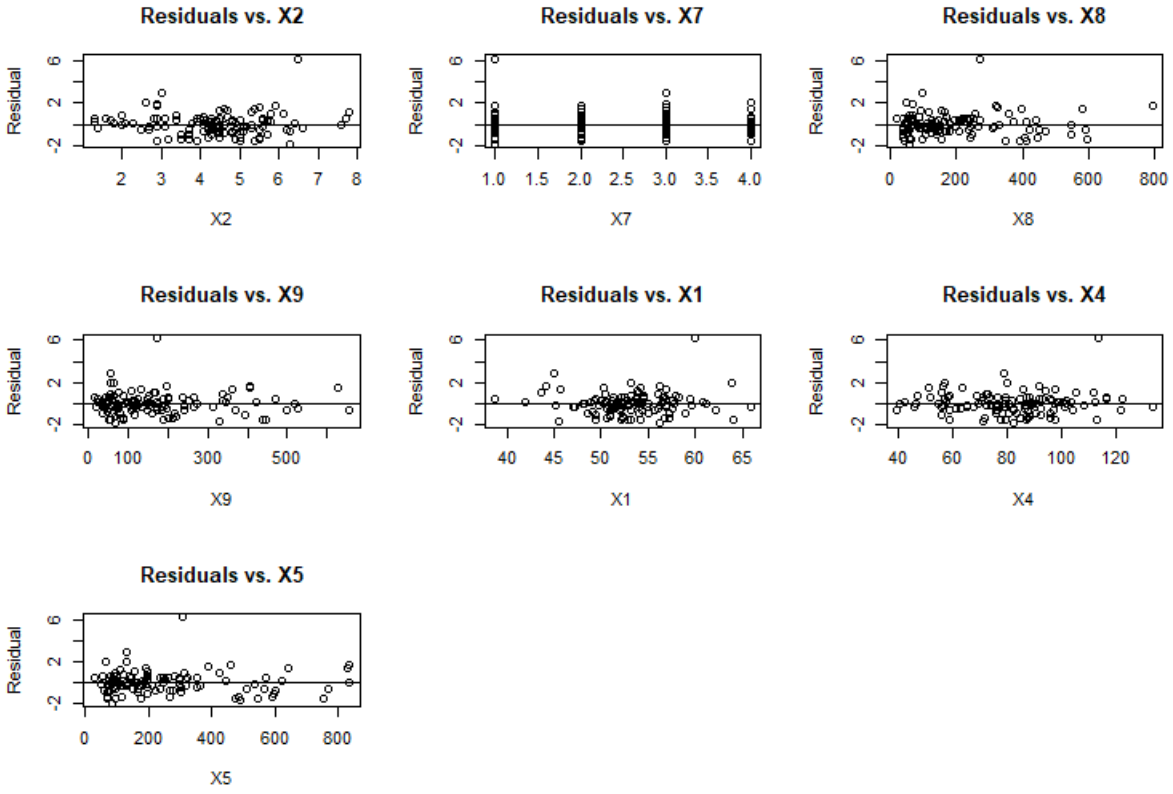



**Figure 2.5: Plots of residuals against the predictor variables.**

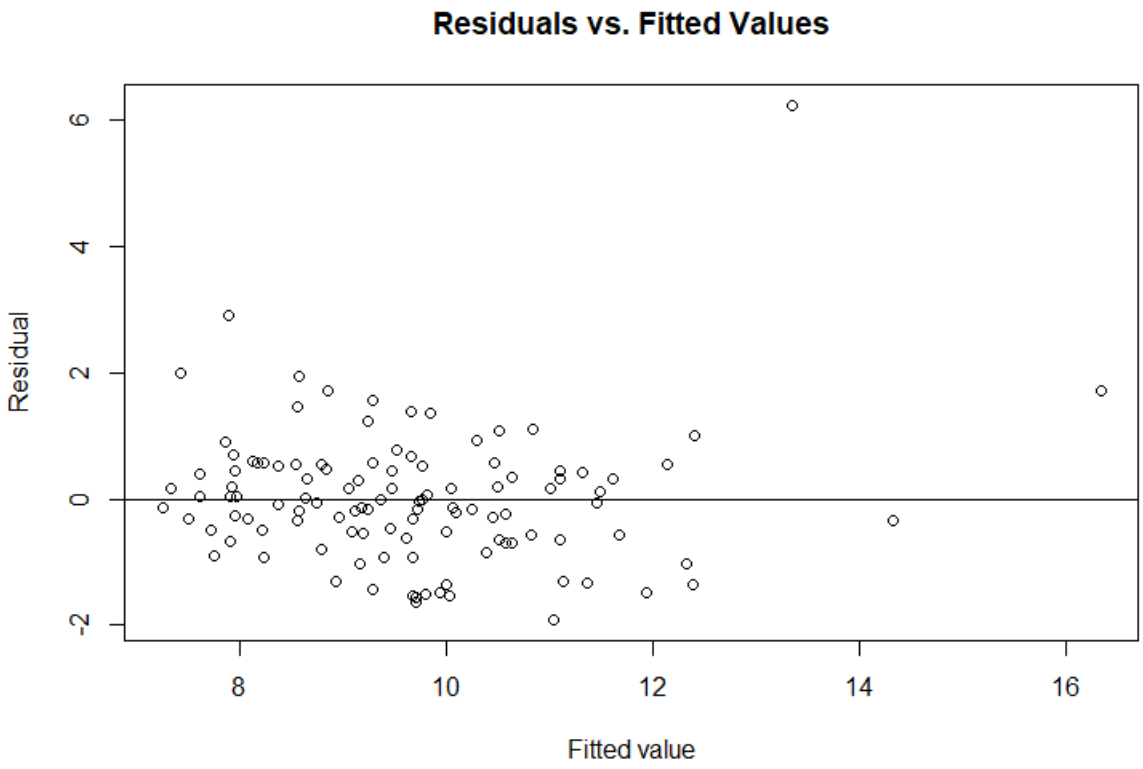



**Figure 2.6: Plot of residuals against the fitted response values.**

In the Breusch-Pagan test, the null hypothesis has a particular error variance; The alternative hypothesis is that there is no constant variance. The decision rule is p. If the value is less than the significance level of 0.05, reject the null hypothesis and draw a conclusion; error distribution is not constant. If the p-value is more significant than the significance level of 0.05, we do not reject the null hypothesis and conclude that the error variance is constant. We fail to reject the null hypothesis because we calculated the test statistic of 7.9695 and the p-value of 0.3353. Make a hypothesis and conclude that the error variance is constant (Results are attached in Appendix)

## 2.4.2 Assumption of non-independence

We can see that the residuals are randomly scattered around zero. The error terms of residuals are assumed to be independent and checked only for time series data. There is constant variance among the residuals, and we can use the residual plot for checking independence.

It is not rational to use the residual plot for checking the independence because the data is not a time-series data.

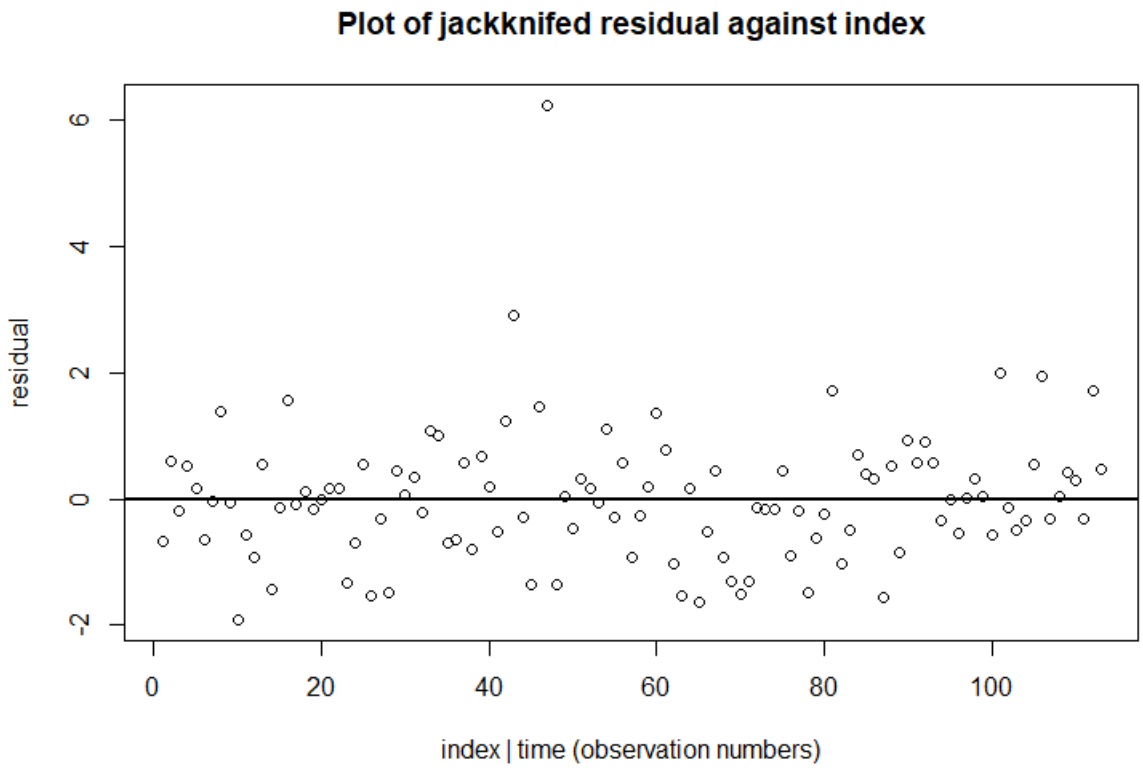



**Figure 2.7: Plot of jackknifed residual against index**

## 2.4.3 Linearity

The residuals are randomly scattered (Figure 2.5 & 2.6) and don't satisfy the linearity assumption. The plots show that the spreads of the residuals do not depend on the fitted values and values of the predictor variables.

The residuals which are far away from the center are the outlying observations. The assumptions don't hold a distribution of points or any limit. Having super-inferential points doesn't break the model, but we have to look into it for resolution.

## 2.4.4 Multi-Collinearity

A rule of thumb for interpreting the variance inflation factor greater than 10 indicates multicollinearity is a significant problem. Here X8 and X5 possess a problem of multicollinearity.

```
> vif(reduced.lmfit2)
       X2        X7        X8        X9        X1        X4        X5
 1.528244  1.174799 29.173884  6.474638  1.012338  1.372107 31.102062
```

**Table 2.2: VIF analysis of Reduced Model**

### 2.4.5 Normality

From the QQ plot (Figure 2.8), we can conclude that the QQ plot follows symmetric distribution and is normally distributed, assigned with slight deviation.

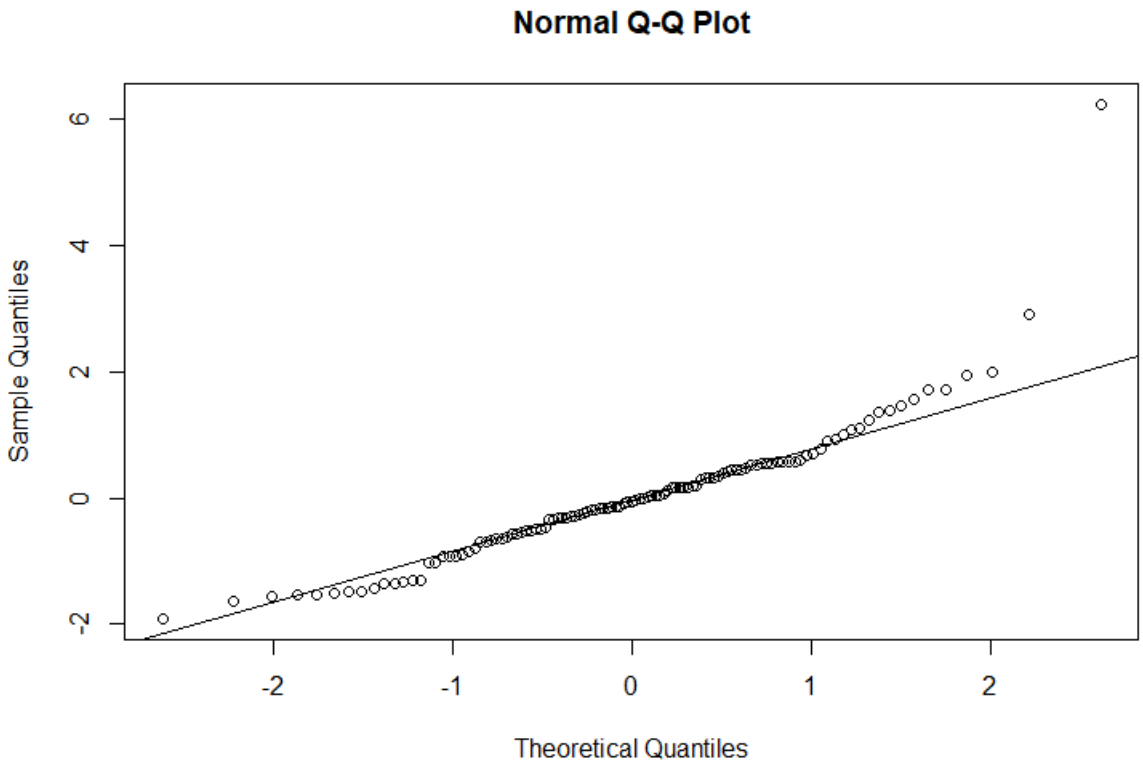



**Figure 2.8: Normal Q-Q plot**

In the Shapiro-Wilk test, the null hypothesis shows that the error terms are normally distributed. The alternative hypothesis, on the other hand, shows that the error terms are not normally distributed. In the decision rule, if the test statistic is small and the p-value is less than significance 0.05, we need to reject the null hypothesis. The test statistic is significant, and the p-value is greater than the significance level, we fail to reject the null hypothesis. We calculated a test statistic of 0.8769 and a p-value of 3.23e-08. Therefore, the null hypothesis can be rejected. From this, we conclude that the error terms are not normally distributed.

The QQ plot and Shapiro-Wilk test provide different results and will check other assumptions and apply remedial measures later in the project.

## 2.4.6 Detection of Influential points and Outliers

The leverage plots (Figure 2.9) indicate outliers for most of the predictor variables and interaction terms. Most outliers appear in the middle section of each graph, above and below the line of best fit. These are all outliers because they are far away from the line of best fit and the cluster of points. The histogram shows that a high leverage point is a point with leverage above 0.10.

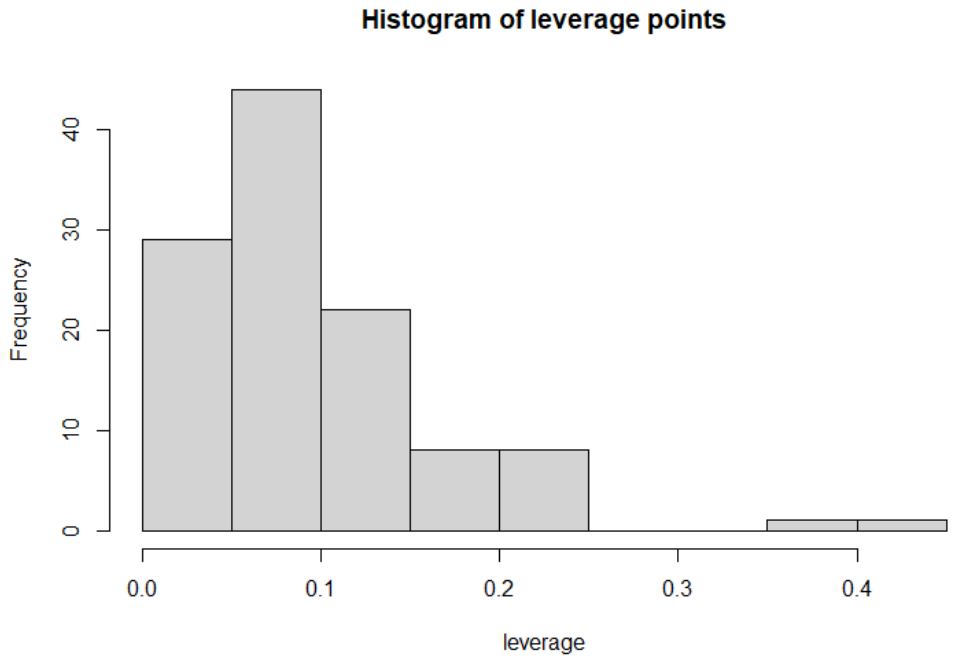



**Figure 2.9: Histogram of Leverage points**

The DFFITS plot (Figure 2.10) shows that there are outliers outside the threshold level. The points that fall outside of DFFITS, which influence a single fitted value (the red line), are considered outliers. Here the threshold value is 0.53.

The Cook's D plot (Figure 2.11) shows that there are outliers outside the threshold level. If any points in this plot fall outside of Cook's distance (the red line), it is influential observation. Here the threshold value is 0.035

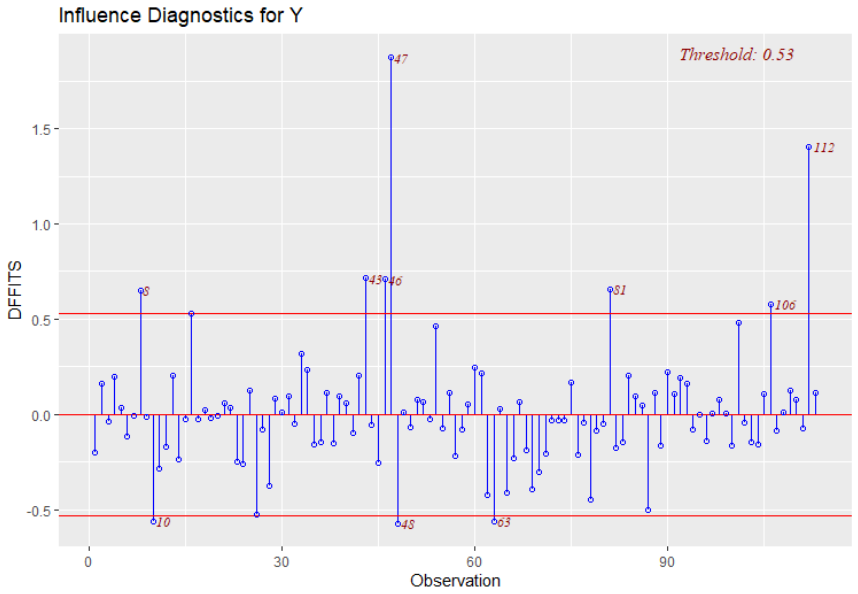



**Figure 2.10: DFFITS plot**

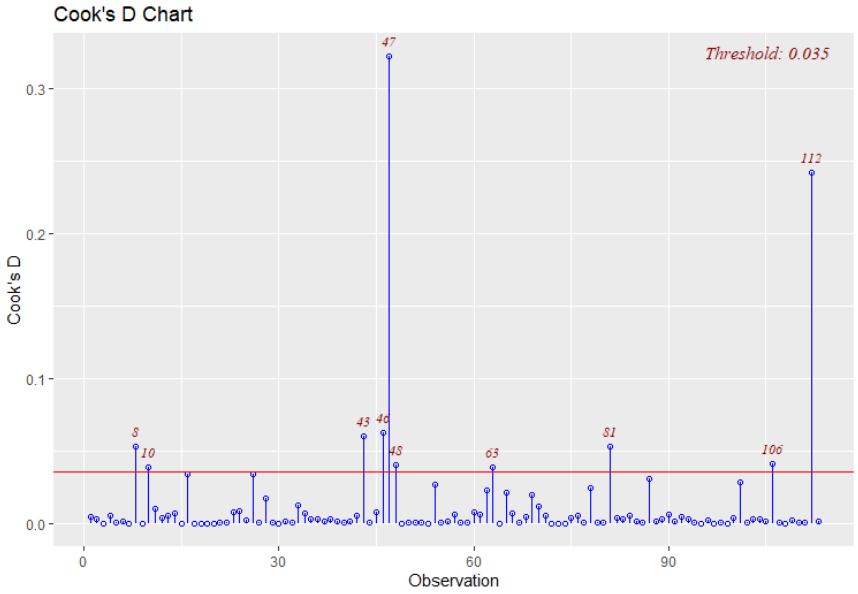



**Figure 2.11: Cook's D plot**

The DFBETAS plot (Figure 2.12) shows that there are outliers outside the threshold level 0.19. The points that fall outside of DFBETAS, which is the influence on single fitted value (the red line), are considered outliers; estimated by MSE. Here the threshold value is calculated for individual predictor variables.

From the residual plots (Figure 2.10, 2.11 & 2.12), we can see that there are several suspicious outliers (or influential points) outside the range of the requirements, the 8, 10, 43, 46, 47, 48, 63, 81, 106 and 112 observations were identified in the DFFITS, Cook's D and DFBETAS plots.

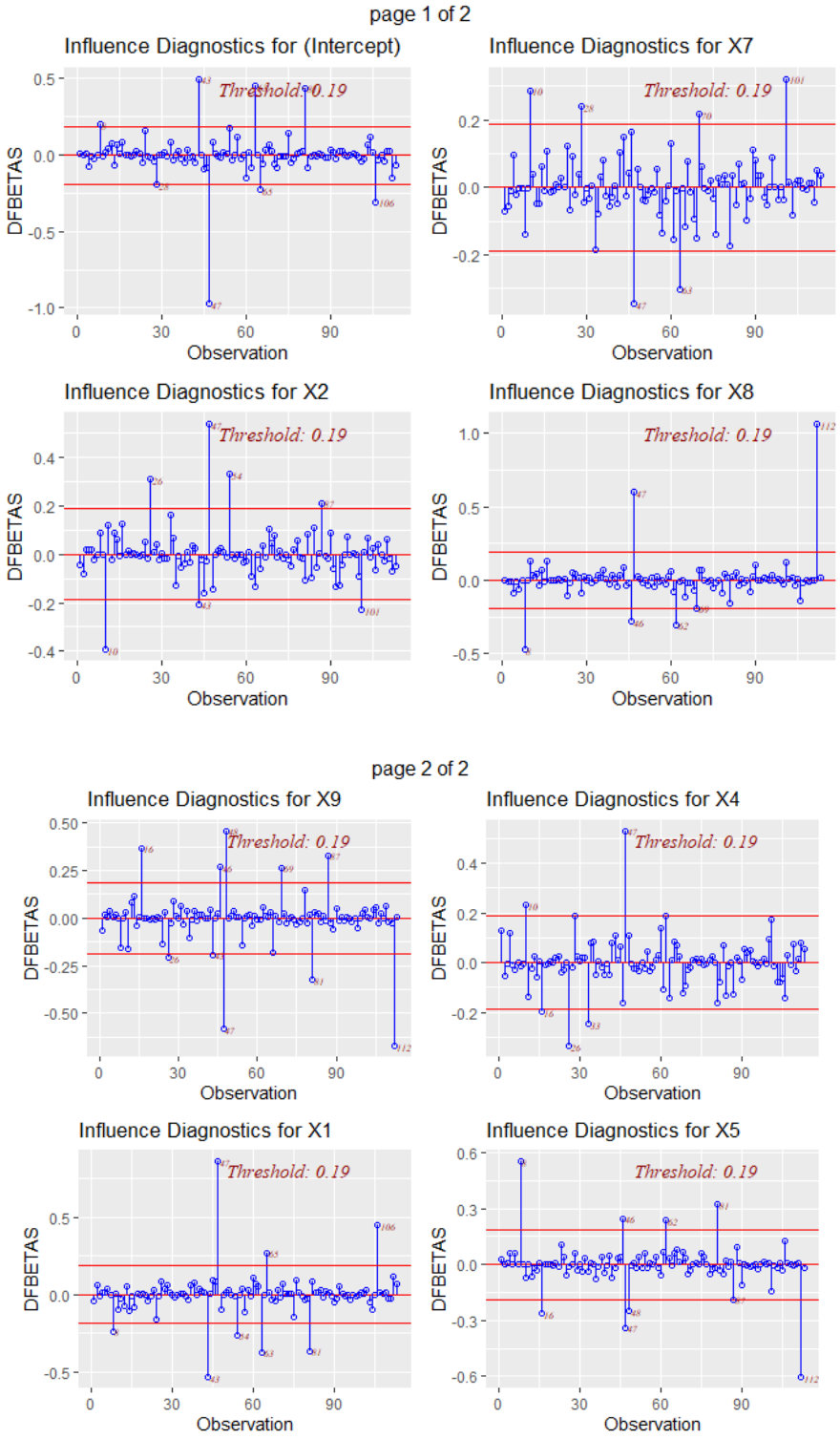



**Figure 2.12: DFBETAS plots**

# 2.5 Model Transformation

The reduced model shows non-linearity and homoscedasticity, and a Boxcox transformation is required to make the model successful. Boxcox does not modify the previously discovered multicollinearity. You can use R's Boxcox function to determine the value of lambda needed, which is -1.3963. Now the new model can be created through the transformation of the response variable. The fitted equation of the new model is:

Y = 7.14e-02 - 2.40e-03X2 + 3.78e-03X7 - 4.75e-05X8 + 1.55e-05X9 - 3.33e-04X1 - 5.37e-05X4 + 1.23e-05X5

```
> summary(boxcox.lmfit)

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

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 ***
X2          -2.401e-03  5.921e-04  -4.054 9.67e-05 ***
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
X1          -3.339e-04  1.448e-04  -2.305   0.0231 *
X4          -5.376e-05  3.885e-05  -1.384   0.1694
X5           1.233e-05  1.857e-05   0.664   0.5082
---
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
```

**Table 2.3: Summary Statistics of Transformed Model**

# 2.6 Diagnostics of Transformed Model

It is required to verify the assumptions for detecting outliers, influential points, normality, constant variance, multi-collinearity, Homoscedasticity and non-independence.; to ensure the transformed model is appropriate.

## 2.6.1 Homoscedasticity

The plot of the residuals against the approximation shows that the regression model is appropriate. This is because the model has a uniform and random distribution. Category variables are a little more complicated; determine if the distribution is random. To determine if homoscedasticity exists, The Breusch Pagan test has been run.

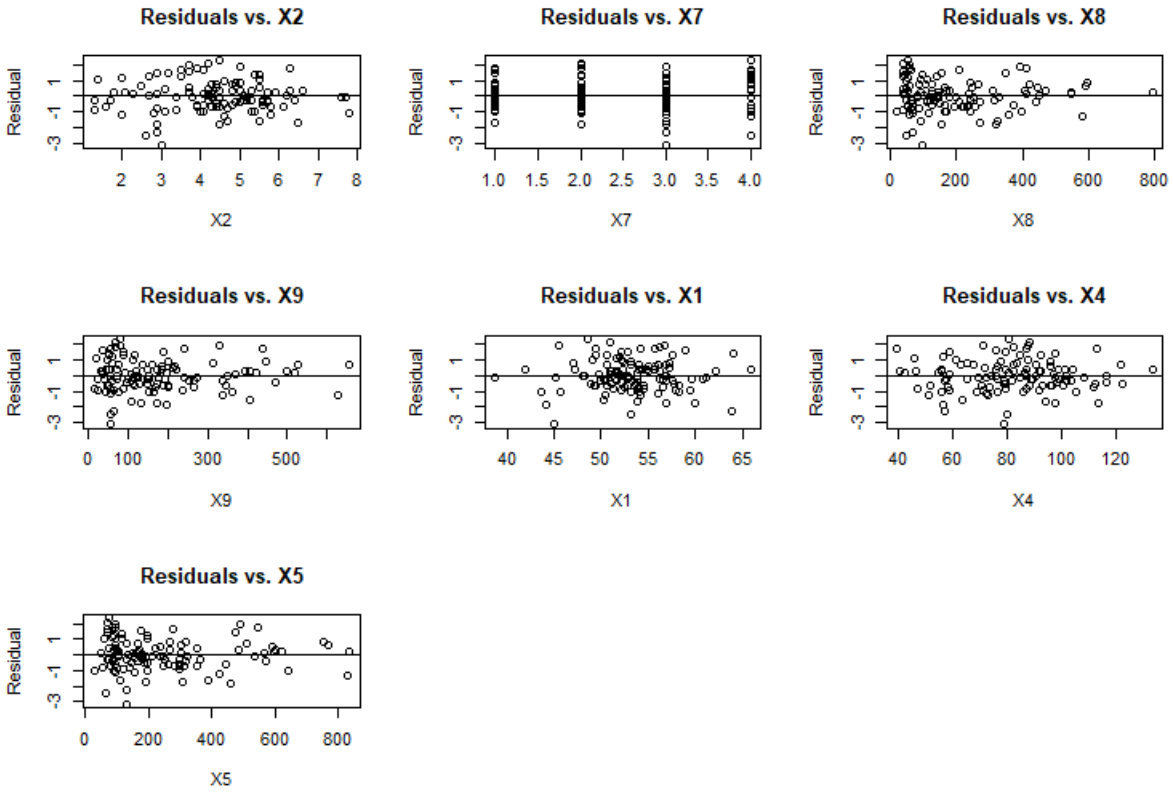



**Figure 2.13: Plots of residuals against the predictor variables.**

```
> bptest(boxcox.lmfit)

        studentized Breusch-Pagan test

data:  boxcox.lmfit
BP = 7.6815, df = 7, p-value = 0.3615
```

**Table 2.4: Summary Statistics of Transformed Model 'bptest'**

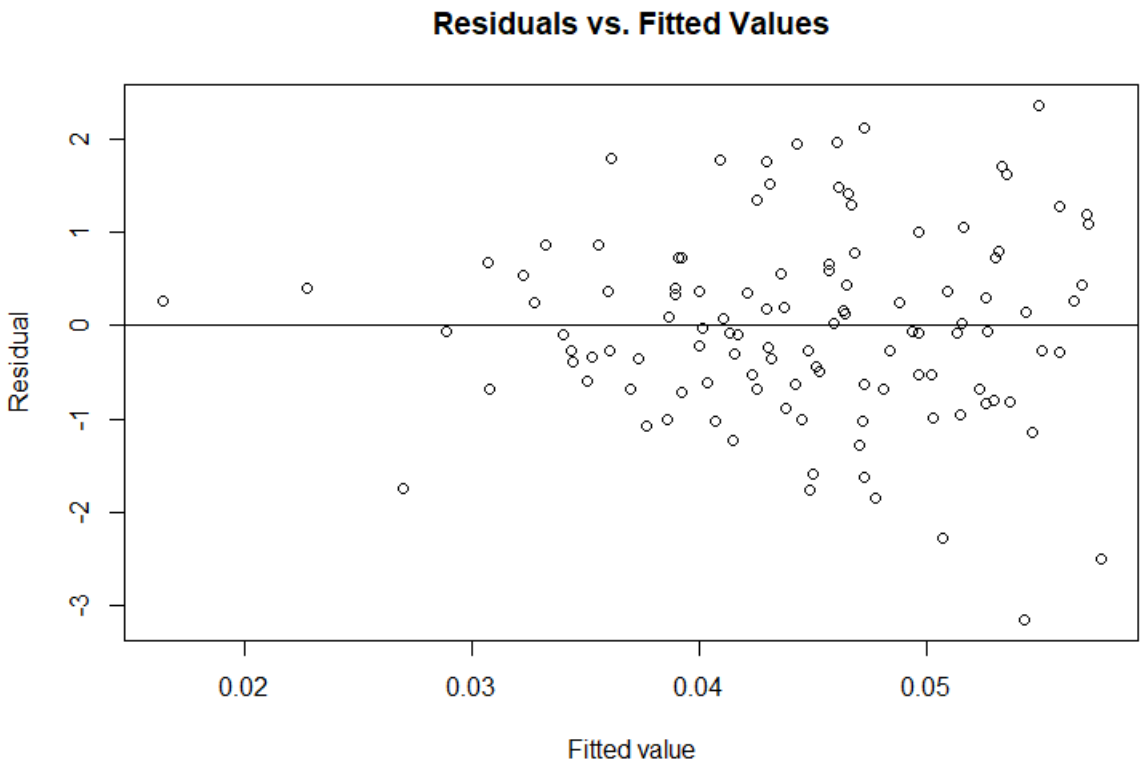



**Figure 2.14: Plot of residuals against the fitted response values.**

In the Breusch-Pagan test, the null hypothesis has a particular error variance; The alternative hypothesis is that there is no constant variance. The decision rule is p. If the value is less than the significance level of 0.05, reject the null hypothesis and draw a conclusion; error distribution is not constant. If the p-value is more significant than the significance level of 0.05, we do not reject the null hypothesis and conclude that the error variance is constant. We fail to reject the null hypothesis because we calculated the test statistic of 7.6815 and the p-value of 0.3615. Make a hypothesis and conclude that the error variance is constant (Table 2.4).

## 2.6.2 Linearity

The residuals (Figure 2.13 & 2.14) are near the center and depend on the fitted values and values of the predictor variables. And satisfy the linearity assumption.

## 2.6.3 Multi-Collinearity

```
> vif(boxcox.lmfit)
       x2        x7        x8        x9        x1        x4        x5
1.528244  1.174799 29.173884  6.474638  1.012338  1.372107 31.102062
```

**Table 2.5: VIF analysis of Transformed Model**

A rule of thumb for interpreting the variance inflation factor greater than 10 indicates multicollinearity is a significant problem. Here X8 and X5 possess a problem of multicollinearity.

## 2.6.4 Normality

From the QQ plot (Figure 2.15), we can conclude that the QQ plot follows symmetric distribution and is normally distributed, assigned with slight deviation.

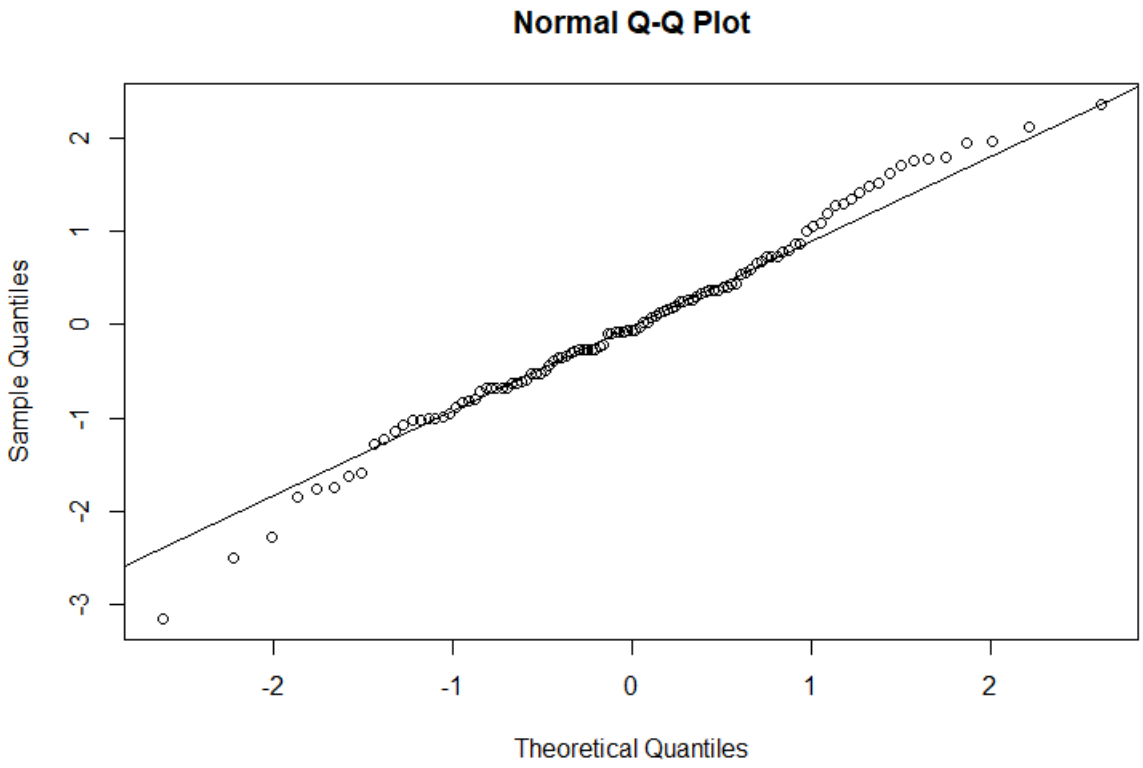



**Figure 2.15: Normal Q-Q plot**

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

        Shapiro-Wilk normality test

data:  boxcox.res
W = 0.98873, p-value = 0.473
```

**Table 2.6: Summary Statistics of Shapiro-Wilk Normality Test**

In the Shapiro-Wilk test, the null hypothesis shows that the error terms are normally distributed. The alternative hypothesis, on the other hand, shows that the error terms are not normally distributed. In the decision rule, if the test statistic is small and the p-value is less than significance 0.05, we need to reject the null hypothesis. The test statistic is significant, and the p-value is greater than the significance level, we fail to reject the null hypothesis. We calculated a test statistic of 0.98873 and a p-value of 0.473. Therefore, the null hypothesis cannot be rejected. From this, we conclude that the error terms are normally distributed.

# 2.7 Reduced Model vs Transformed Model

The reduced model doesn't satisfy the normality assumption, and a multicollinearity problem exists. The transformed model met the normality assumption, and still, the multicollinearity problem persists. But the multicollinearity doesn't cause a severe threat to the fitting of a model (Refer Figure 1.6). There is a reduction is Adjusted $R^2$ value to 55.62% as compared to reduced model 57.65%.

The transformed model with five predictor variables (X1 X2 X7 X8 X9) has an adjusted $R^2$ (55.47%) than the transformed model with seven predictor variables (55.62%). Also, it satisfies all assumptions, including normality, and doesn't possess a threat of multicollinearity unlike transformed and reduced model with seven predictor variables.

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

        Shapiro-Wilk normality test

data:  boxcox.res
W = 0.98612, p-value = 0.2971
```

```
> vif(boxcox.lmfit)
      X2       X7       X8       X9       X1
1.218988 1.053939 5.797580 5.856268 1.010906
```

**Table 2.7: Summary Statistics to verify Normality and Multicollinearity**

The transformed model with five predictor variables (X1 X2 X7 X8 X9) is the better model than the reduced model with five & seven predictor variables and the transformed model with seven predictor variables. Also, the model with interactions and polynomial models either provides the same result or worse than the transformed model with five predictor variables.

# 3 Results

The summary of the final optimal model is:

Y = 6.74e-02 - 2.79e-03X2 + 4.11e-03X7 - 3.30e-05X8 + 1.78e-05X9 - 3.25e-04X1

The predictor variables of the model are Infection Risk, Medical School, Region, Number of Nurses and Age.

## 3.1 F-test

The value of the F-statistic is 28.9 on five predictor variables, and the p-value is 2.2e-16; Therefore, there is an overall significant relationship between the response variable and the predictor variables.

## 3.2 Adjusted R2

The model produced 55.47 % (Adjusted R-squared) of the total variability in Y, which can be explained by the regression model using the entire model.

## 3.3 Significance of the Individual Predictors

From the summary statistics (Table 3.1) and ANOVA table (Table 3.2), the results of each p-value are less than the significant level 0.05 except X9. Even including the X9 variable, we obtained the optimal model with significant improvement in Adjusted R-squared. Also, the relation between response and predictor variables is linear and follows normality.

Table 3.1 shows the t-values of each statistic and table 3.2 shows the F-statistic of each predictor variable.

```
> summary(final.lmfit)

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

Residuals:
       Min         1Q     Median         3Q        Max
-0.0203629 -0.0041742 -0.0002586  0.0043695  0.0149823

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.747e-02  8.287e-03   8.142 7.68e-13 ***
X2          -2.796e-03  5.297e-04  -5.278 6.90e-07 ***
X7           4.117e-03  6.543e-04   6.293 6.98e-09 ***
X8          -3.302e-05  1.007e-05  -3.278  0.00141 **
X9           1.786e-05  1.118e-05   1.597  0.11317
X1          -3.259e-04  1.450e-04  -2.248  0.02665 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.006809 on 107 degrees of freedom
Multiple R-squared:  0.5746,    Adjusted R-squared:  0.5547
F-statistic:  28.9 on 5 and 107 DF,  p-value: < 2.2e-16
```

**Table 3.1: Summary of final model**

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

Response: trans.Y
           Df    Sum Sq   Mean Sq F value    Pr(>F)
X2          1 0.0033961 0.0033961 73.2613 9.065e-14 ***
X7          1 0.0021940 0.0021940 47.3285 4.212e-10 ***
X8          1 0.0007273 0.0007273 15.6900 0.0001348 ***
X9          1 0.0001478 0.0001478  3.1876 0.0770306 .
X1          1 0.0002342 0.0002342  5.0518 0.0266515 *
Residuals 107 0.0049601 0.0000464
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Table 3.2: Analysis of Variance (ANOVA) of final model**

## 3.4 Interpretations of Regression Coefficients

**β2:** As the infection risk increases, the length of stay of a patient would decreases by 2.79e-03 times on average when the other variables are held constant.

**β7:** As the association of medical school with the hospital increases, a patient's length of stay would increase by 4.11e-03 times on average when the other variables are held constant.

**β8:** As the geographic region changes, a patient's length of stay would decrease by 3.30-05 times on average when the other variables are held constant.

**β9:** The number of licensed or registered nurses might not affect the length of stay of a patient as the p-value is greater than the significant value of 0.05. So, we can either include or exclude it as it doesn't affect our model and increases the optimality.

**β1:** As the age increases, the length of stay of a patient would decrease by 3.25e-04 times on average when the other variables are held constant.

# 4 Conclusion

To find the optimal model for the SENIC dataset, we performed exploratory data analysis and reduced the model based on the significant level of 0.05. To validate it, we chose the method for model selection and applied diagnostic measures to ensure the model is fitted perfectly. Later, to improvise it further, we transformed the model with seven predictor variables and chosen the transformed model with five predictor variables, and there is a significant improvement in the Adjusted R-squared of the final model.

We chose the transformed model with five predictor variables, which has the better improvement. This model can be used to identify the length of stay of the patients based on characteristics of hospitals with the predictor variables (Infection Risk, Medical School, Region, Number of Nurses and Age).

The predictor variable Number of Nurses (X9), which has a significant value greater than 0.05, is still included in the model. Even though the inclusion of X9 makes the model slightly better, it's better to remove the X9 variable for time optimization with optimal fit. It's easy to mitigate the fewer variables causing length of stay of patients in the hospitals.

As standardization and centralization are not required based on the exploratory data analysis, there might be a chance to find a better fit if we perform either standardization or centralization.