Regression Analysis Final Project

Kristoffer Larsen

MA 4710

December 12, 2020

Table of Contents

I. Introduction

The goal of the data analysis for this project is to fit a linear regression model and optimize it in order to interpret the model. The dataset (senic) is a collection of medical data from the participating 113 hospitals. We aim to model the average length of stay of patients (Y) by multiple predictor variables: the average age of patients (X1), the percent probability of acquiring an infection (X2), the ratio of the number of cultures performed without signs to acquired infections (X3), the ratio of number of X-rays performed without sign of symptoms times 100 (X4), the number of beds in the hospital (X5), an encoded value with one being associated with a medical school and two being not (X6), an encoded value of the geographic region ranging from one to four for the intermediate directions (X7), the number of patients per day (X8), the number of full-time nurses working (X9), and the percent of potential services that are provided by the hospital (X10). Below are a series of graphical tools and accompanying descriptions for the exploratory analysis of the data.

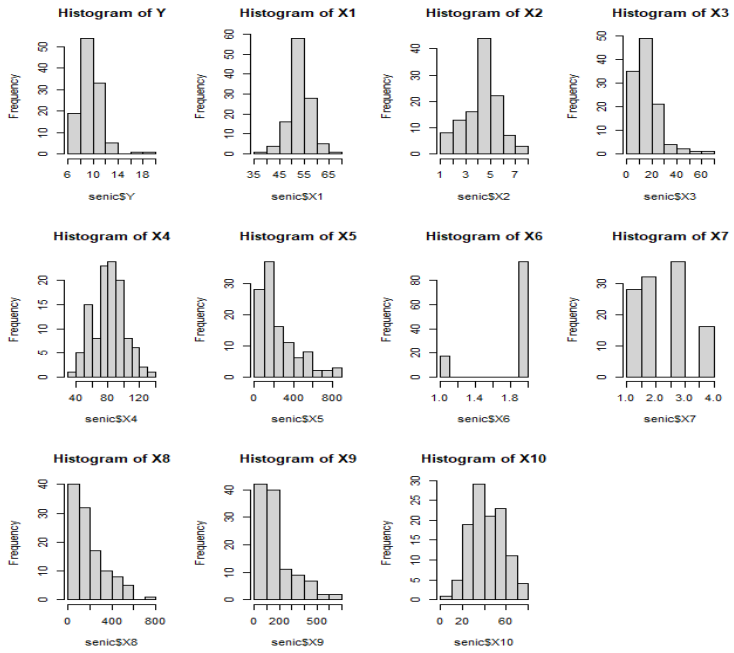


Figure 1. Histograms of response and predictor variables.

The histogram for the dependent variable (Y) has a moderate right skew, with the mode of hospital stay being roughly 8-10 days. The histogram for (X1) shows a strong unimodal distribution with light tails. The histogram for (X2) shows a strong unimodal distribution with normal tails. The histogram for (X3) has a unimodal strong right skew distribution. The histogram for (X4) has a roughly normal distribution with a left tail peak. The histogram for (X5) has a unimodal strong right skew distribution. The histogram for (X6) is unimodal at two indicating the majority of hospitals are not associated with a medical school. The histogram for (X7) is semi-uniformly distributed with a mode of three indicating the south has the most hospitals. The histogram for (X8) and (X9) both show a strong right skew. The histogram for (X10) has a bell shape translating to a roughly normal distribution.

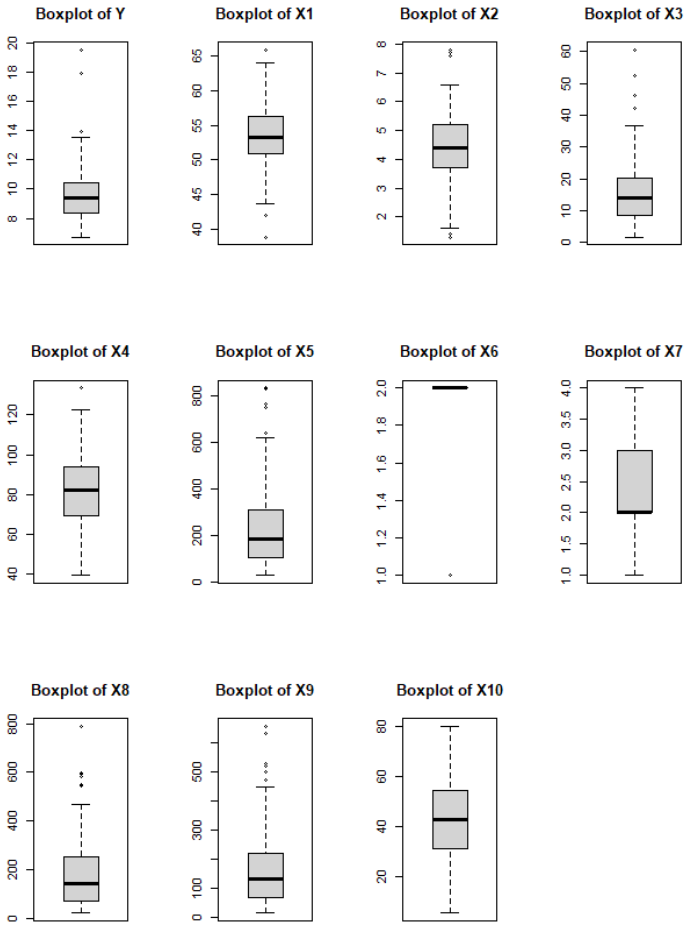


Figure 2. Boxplots of response and predictor variables

The boxplot of the dependent variable (Y) exhibits a right skew with some larger outliers indicating some few hospitals have abnormally long hospital stays. The boxplots of (X1) and (X2) show a fairly symmetric distribution with at least one outlier per side. The boxplot of (X3) shows a textbook right skew with multiple outliers. The boxplot of (X4) has a fairly symmetric distribution with one larger outlier. The boxplot of (X5) displays a right skew with multiple clustered outliers. The boxplot of (X6) again indicates most hospitals are not affiliated with a medical school. The boxplot of (X7) does not provide any meaningful information. The boxplots of (X8) and (X9) exhibit a right skew with multiple outliers each. The boxplot of (X10) is near perfectly symmetric.

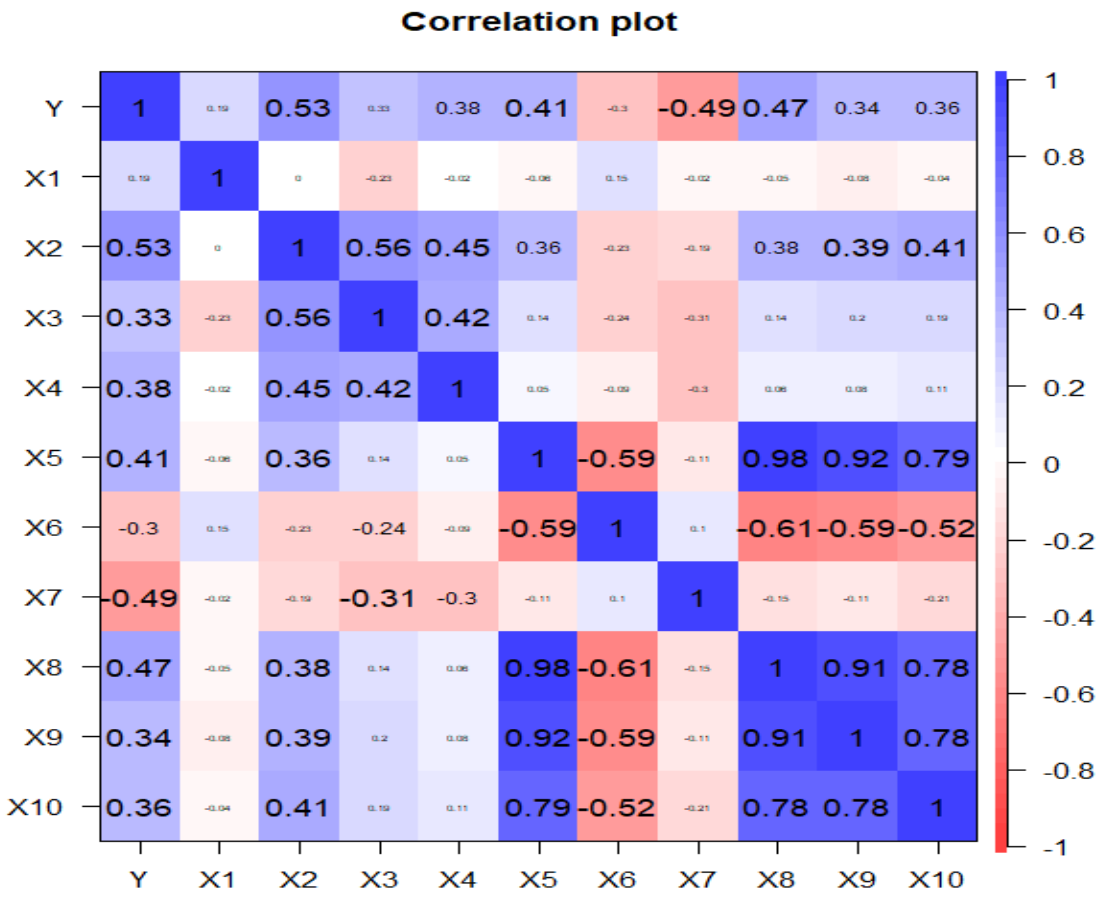


Figure 3. Correlation plot of variables

The correlation matrix displays the value of the correlation between the two variables in each cell. The predictor variables (X8), (X9), and (X10) all have a very strong correlation with each other, likely because they all deal with the size of the hospital. Also, the same predictor variables have a strong correlation with the predictor variable (X5). Moderately correlated include (X6) and the aforementioned strongly correlated variables; (X7) and (X2) against (Y); and between (X2) and (X3).

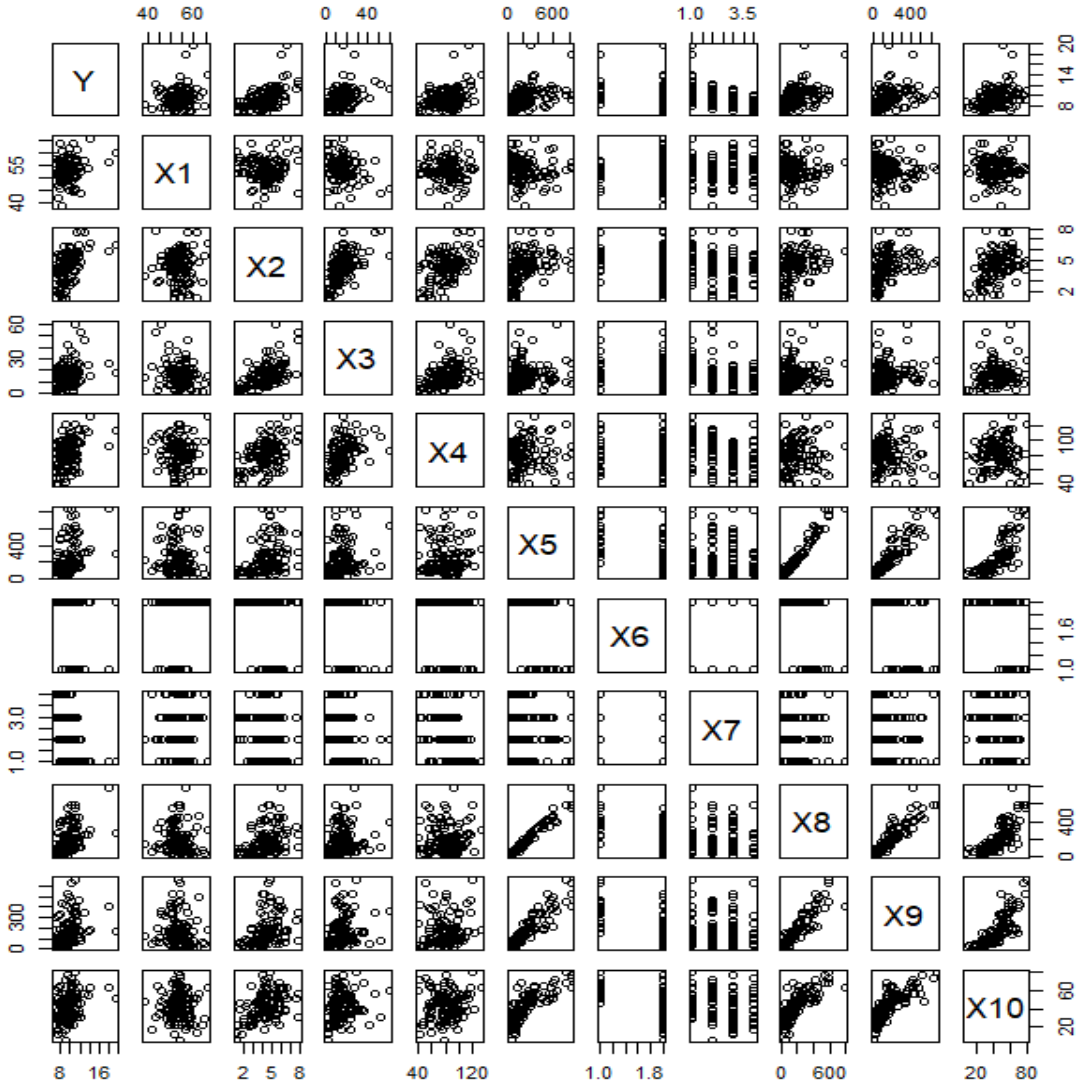
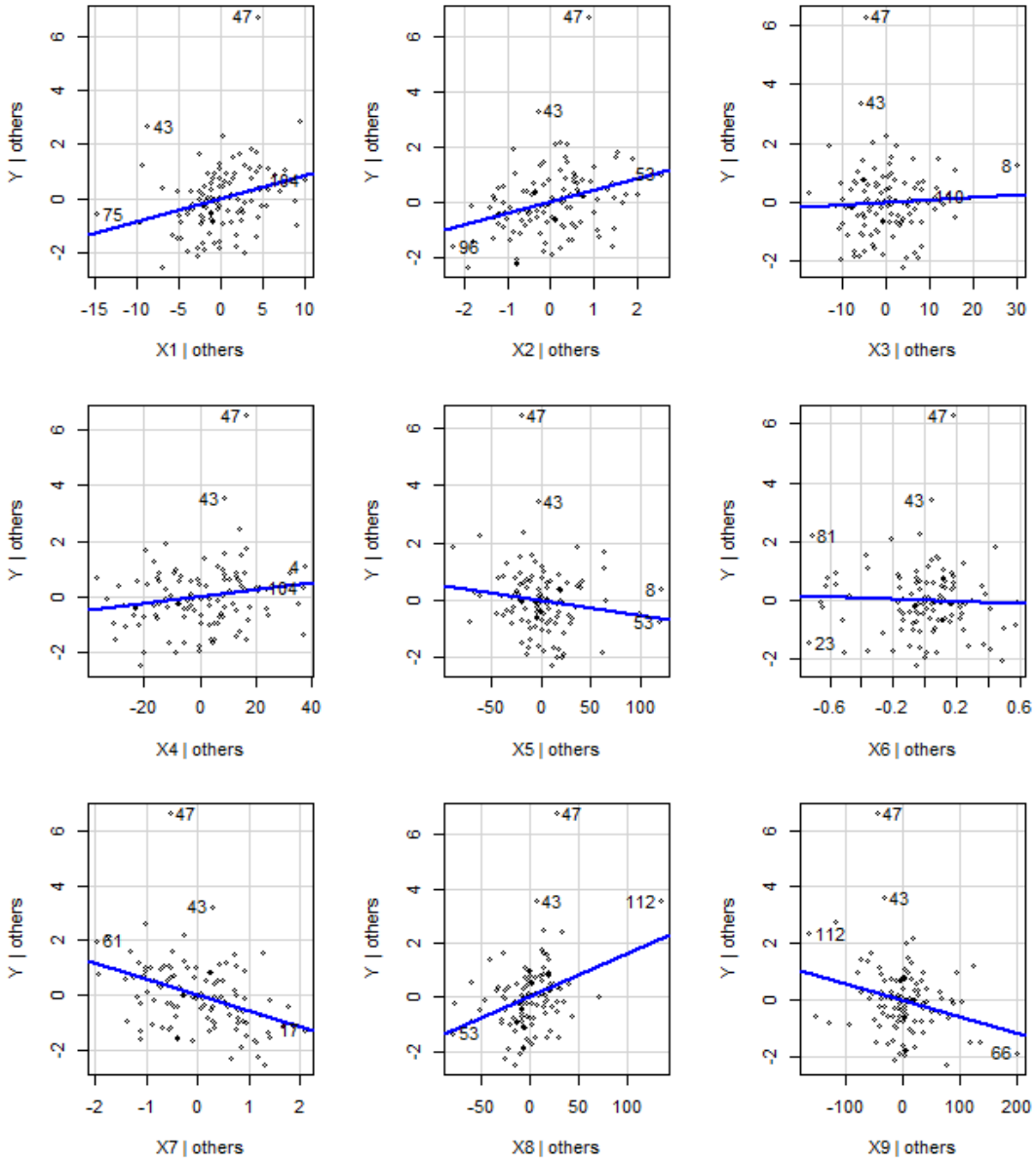


Figure 4. Scatterplot matrix of variables

The scatterplot matrix of the variables supports the findings from figure 3. such as the intensely strong correlation between the three variables (X8), (X9), and (X10). It is hard to see the relationship between (X6) and (X7) with the other variables; however, it is clear that (X5) as mentioned before has a moderate correlation with the strongly correlated variables. All other relationships are not as apparent, appearing in clouds of points. In face of all this strong correlation, the decision was made to standardize all of the predictor variables.



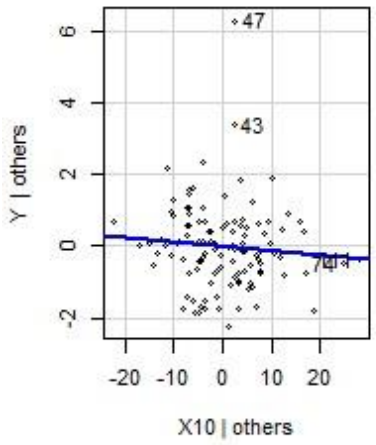


Figure 5. Added variable plots

The added-variable plots show the linear regression line when the select variable is added given that all the others are already included so we can look at the direct relationship between an individual predictor variable and its marginal effect on the response variable (Y). In the added-variable plots of (X1) and (X2), there is a pretty moderate linear relationship with (Y). For the added-variable plots of (X3) and (X4), there is a weak linear relationship because of the cloud shape distribution of points. In the added-variable plots of (X5) and (X6), there appears to be a pretty weak negative linear relationship due to the large cloud of points. The added-variable plots of (X7) and (X9) show a pretty moderate negative linear relationship. While the plot of (X8) shows a pretty moderate positive linear relationship. Finally, the added-variable plot of (X10) exhibits a weak negative linear relationship.

II. Model/Methods

A. Creating Categorical Terms

It is important to note that the variables of (X6) and (X7) are categorical variables. The variable (X6) will encode the value one if the hospital is affiliated with a medical school and the value two for the inverse. The variable (X7) is an indicator of geographic region for the hospital with one being encoded for the northeast, two for north central, three for the south, and four for the west. The rest of the predictor variables contain their respective numerical data and do not need to be encoded.

B. Model Creation and Selection

A simple model without interactions was chosen for simplicity's sake. In addition, some of the predictor variables exhibit multicollinearity so it was not thought apt to include potential problems. Model selection was performed through stepwise selection inorder to satisfy the procedure of the adjusted R2, Mallow's CP, and AIC/BIC. Below are plots showing the number of variables chosen against the model selection criterion. Only the single best model of the possible combinations is shown at each variable level.

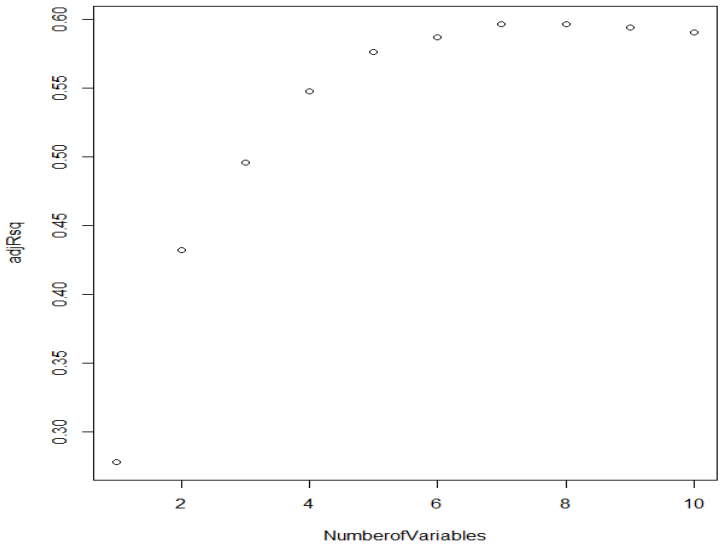


Figure 6. Adjusted R2 plot

The adjusted R2 plot (Figure 6.) shows there is a significant difference in the amount of variation that can be explained going from one predictor variable to two predictor variables. Additionally, the plot shows that the highest R2 comes from a model of 7 predictor variables.

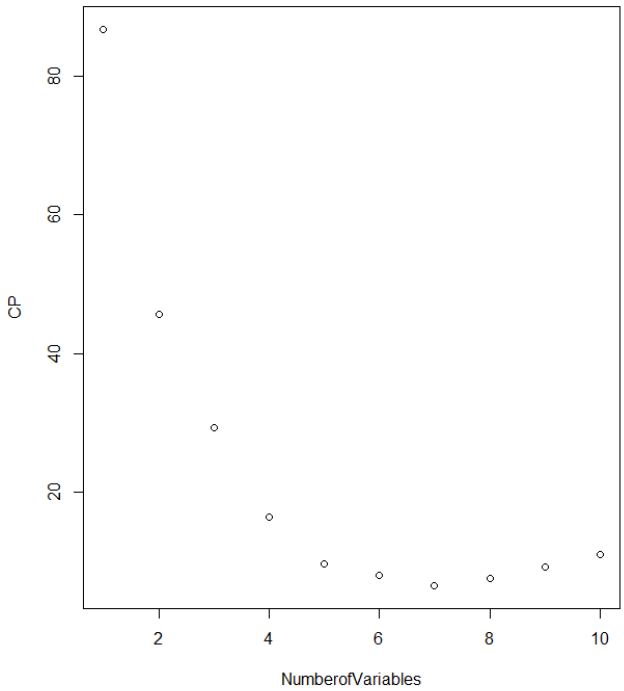


Figure 7. Mallows CP plot

The Mallows CP plot (Figure 7.) choose the number of variables by the number needed to minimize its value (or come close to one plus the number of predictor variables). Moreover, the plot shows that the optimal number of predictor variables is again 7.

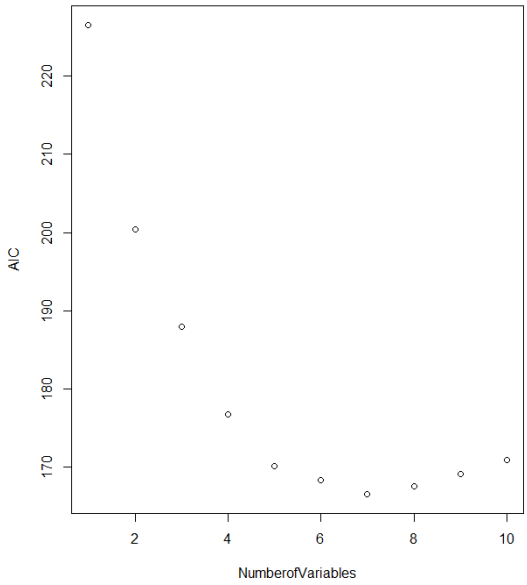


Figure 8. Akaike's information criterion (AIC) plot

The AIC plot (Figure 8.) indicates again that 7 predictor variables are needed to have the smallest value of AIC.

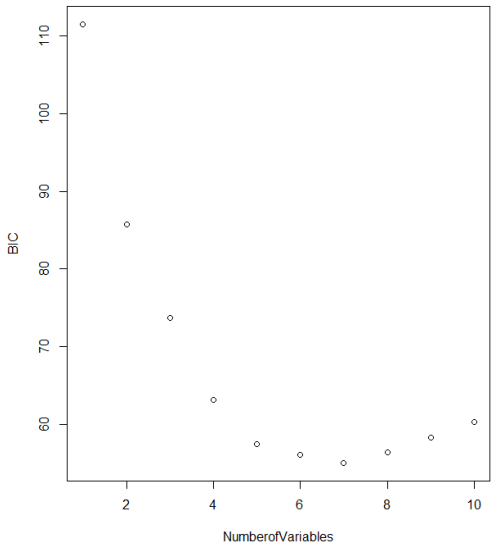


Figure 9. Bayesian information criterion (BIC) plot.

The BIC plot (Figure 9.) indicates again that 7 predictor variables are needed to have the small value of BIC which will seek. After analyzing the optimal number of variables, the actual subset of variables that fulfills the aforementioned figures: (X1), (X2), (X4), (X5), (X7), (X8), and (X9). This model has an adjusted R2 of 0.59662, a Mallow's CP of 6.48305, an AIC value of 166.5292, and a BIC value of 54.98371. Note however, with 6 or 8 predictor variables the values of these procedures are only just very slightly lower. The model is summarized below.

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

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6483     0.1142  84.483  < 2e-16 ***
X1            0.3522     0.1154   3.051  0.00289 **
X2            0.5845     0.1418   4.121 7.54e-05 ***
X4            0.2621     0.1344   1.951  0.05376 .
X5           -1.2075     0.6397  -1.887  0.06185 .
X7           -0.5768     0.1243  -4.639 1.01e-05 ***
X8            2.5482     0.6196   4.113 7.79e-05 ***
X9           -0.8437     0.2919  -2.891  0.00467 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

Table 1. Summary of the model output

As seen in the output of the summary, we can see through the p-value of the overall model that the model is significant compared with the model containing only the intercept. Additionally, approximately 59.66% percent of the variation in the data can be explained by the model.

C. Model Assumptions

In order for our model to be a good choice for the dataset requires the assumptions of linearity, normality, homoscedasticity of error terms. Before this can be done, our predictor variables must not have multicollinearity. It severely cripples the conclusions of the model and the effects of certain predictor variables.

X1	X2	X4	X5	X7	X8	X9
1.012338	1.528244	1.372107	31.102062	1.174799	29.173884	6.474638

Figure 10. Variance inflation factors (VIF) of model predictor variables

As suspected from our exploratory analysis of the data, the predictor variables (X5) and (X8) possess variance inflation above 10 despite being standardized already. This indicates there are multicollinearity problems within our selected model. In the correlation matrix (Figure 3.), the variable (X5) has the most and highest correlation coefficients of our model; with this in mind, the predictor variable (X5) will be removed from our analysis and model selection will be redone.

D. Redone Model Selection

Without the predictor variable (X5) stepwise selection was performed inorder to satisfy the procedure of the adjusted R2, Mallow's CP, and AIC/BIC.

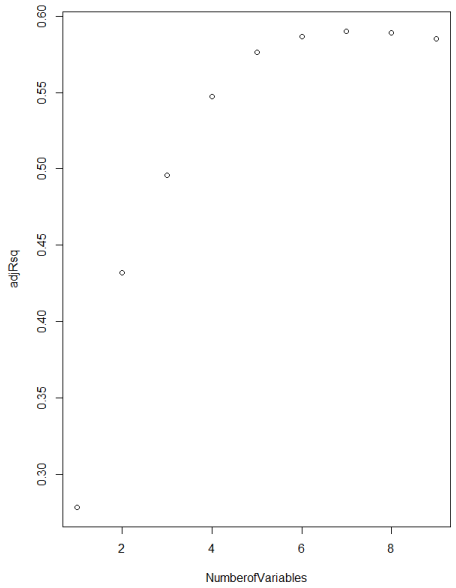


Figure 11. Adjusted R2 plot

The adjusted R2 plot (Figure 11.) shows there is a significant difference in the amount of variation that can be explained going from one predictor variable to two predictor variables. Additionally, the plot shows that the highest R2 comes from a model of 7 predictor variables.

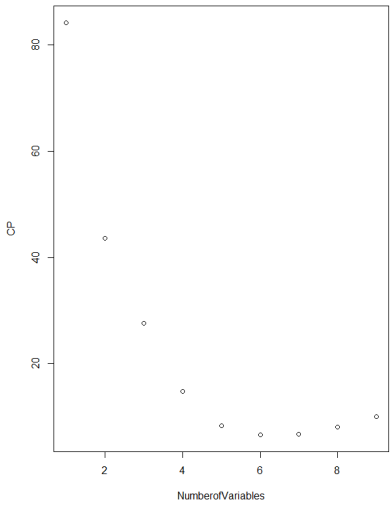


Figure 12. Mallows CP plot

The Mallows CP plot (Figure 12.) shows that the optimal number of predictor variables is again 7, agreeing with the plot of the adjusted R2.

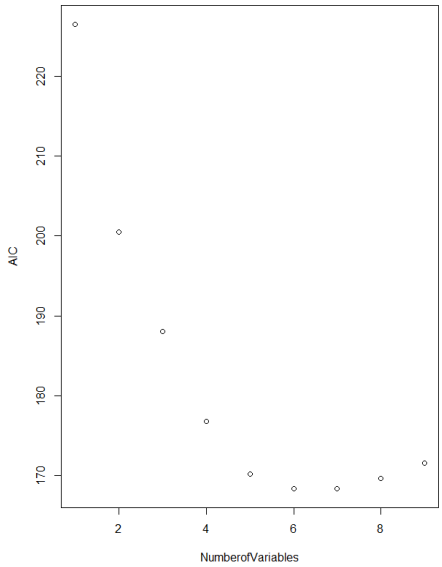


Figure 13. Akaike's information criterion (AIC) plot

The AIC plot (Figure 13.) indicates again that 7 predictor variables are needed to have the smallest value of AIC.

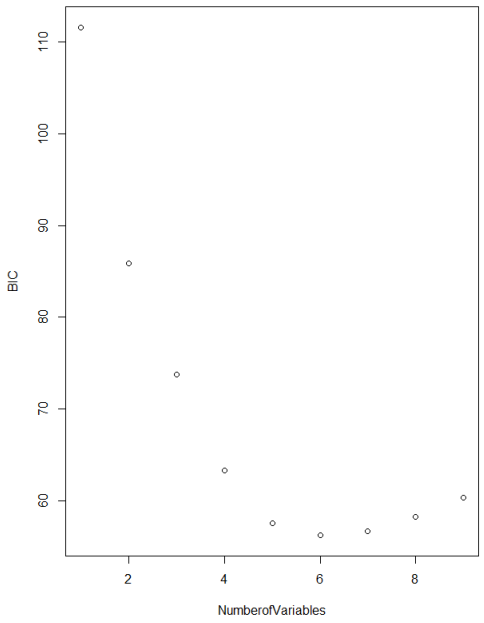


Figure 14. Bayesian information criterion (BIC) plot.

The BIC plot (Figure 14.) indicates again that 7 predictor variables are needed to have the small value of BIC which will seek. After analyzing the optimal number of variables, the actual subset of variables that fulfills the aforementioned figures: (X1), (X2), (X4), (X7), (X8), (X9), and (X10). This model has an adjusted R2 of 0.59024, a Mallow's CP of 6.74530, an AIC value of 168.3038, and a BIC value of 56.71521. The difference between the model selected ignoring (X5) and the model including (X5) do not differ much; the R2 is slightly less, while the remaining criterion are only slightly greater. The model summary is below. As seen in the output of the summary, we can see through the p-value of the overall model that the model is significant compared with the model containing only the intercept. Additionally, approximately 59.02% percent of the variation in the data can be explained by the model.

```
Call:
lm(formula = Y ~ X1 + X2 + X4 + X7 + X8 + X9 + X10, data = senic5)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2412 -0.6533 -0.0504  0.5832  6.3562

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6483     0.1151  83.823  < 2e-16 ***
X1            0.3538     0.1163   3.041  0.00298 **
X2            0.6359     0.1439   4.418 2.43e-05 ***
X4            0.2563     0.1355   1.892  0.06127 .
X7           -0.6503     0.1242  -5.235 8.53e-07 ***
X8            1.5930     0.2871   5.549 2.17e-07 ***
X9           -0.8979     0.2924  -3.071  0.00272 **
X10          -0.2706     0.1978  -1.368  0.17428
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.224 on 105 degrees of freedom
Multiple R-squared:  0.6158,    Adjusted R-squared:  0.5902
F-statistic: 24.05 on 7 and 105 DF,  p-value: < 2.2e-16
```

Table 2. Summary of the model output

E. Redone Model Assumption

In order for our model to be a good choice for the dataset requires the assumptions of linearity, normality, homoscedasticity of error terms. Checking the multicollinearity of our predictor variables first, we find no significant issues as the VIF's are all below 10 indicating no serious multicollinearity issues.

```
      X1       X2       X4       X7       X8       X9      X10
1.012654 1.549749 1.373126 1.154718 6.165637 6.394902 2.928264
```

Figure 11. Variance inflation factors (VIF) of model predictor variables

The removal of (X5) from the variables and selection process definitely decreased the overall amount of multicollinearity, which can be easily seen in the decrease of the VIF of (X8).

a) Linearity

In a model that fulfills the linearity assumption the plots of the response variable against the predictor variables will show a straight line pattern, either positive or negative.

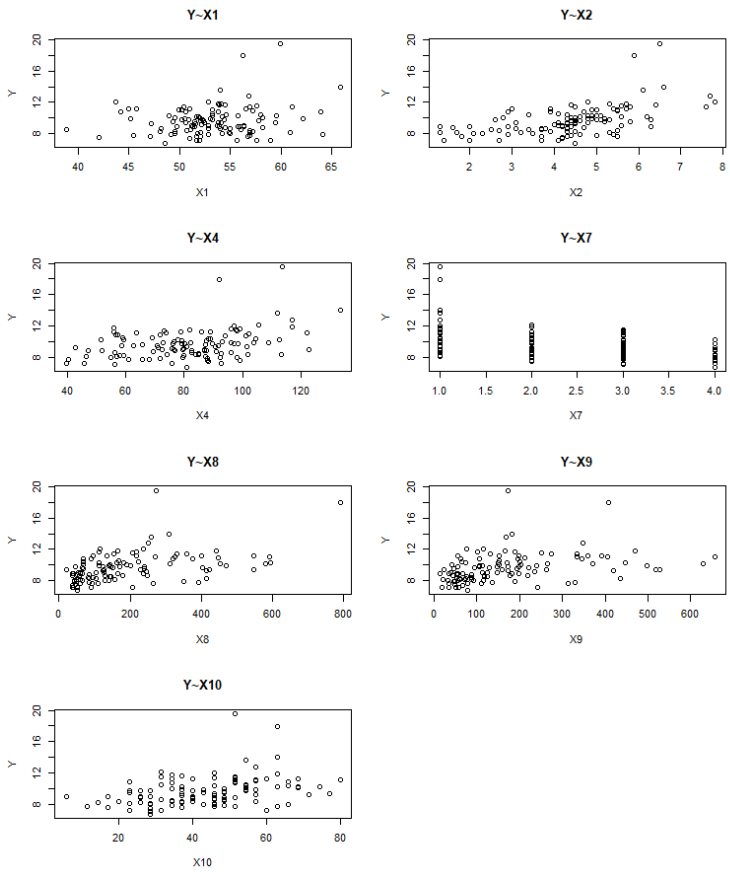


Figure 12. Scatter plot of response variable against predictor variables

At first glance, the predictor variable (X7) sticks out for its multimodal distribution indicating a non-linear trend; however, as mentioned before (X7) is a categorical variable and can be ignored. The rest of the predictor variables with the exception of (X2) appear very linear. The predictor variable (X2) has a slight curvilinear shape, but it is very modest. As such, the linearity assumption is valid.

b) Homoscedasticity

In a model that fulfills the assumption of equal error variances, we would expect that the plot of the residuals against the fitted values and the predictor variables to have a rectangular shape: they also do not possess a cone shape.

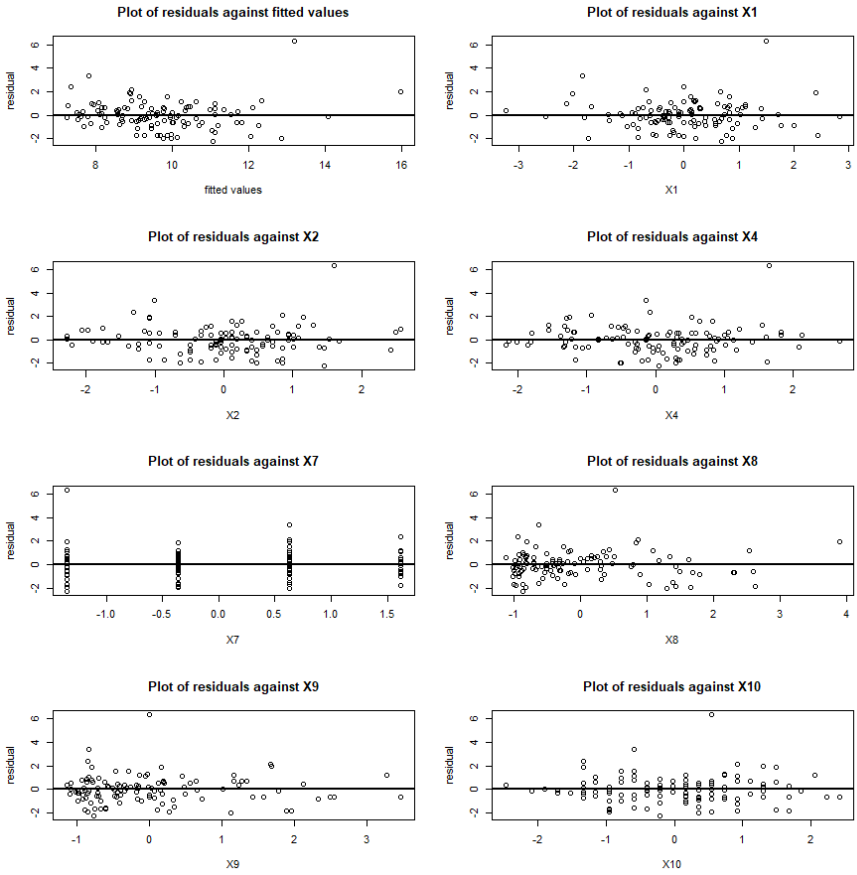


Figure 13. Scatterplot of residuals against fitted values and predictor variables

As shown in (Figure 13.) there appears to be some non-constant error variance evident in the plots of the fitted values, (X1), and (X9) from the slight cone shape. Overall, however there does not seem to be any major problems. To further analyze this assumption, a Breusch-Pagan Test was conducted with the significance level of .05.

```
        studentized Breusch-Pagan test

data:  senic5.lmfit
BP = 8.8676, df = 7, p-value = 0.2623
```

Figure 14. Breusch-Pagan test output

With the p-value of .2623 greater than our significance level, we can retain the null hypothesis that there is a constant error variance and reject the alternative hypothesis which is that there is no constant error variance. Our assumption of homoscedasticity is valid.

c) Normality

In a model that fulfills the normality assumption we would expect the normal qq plot of the error terms to follow the normal distribution line and not seriously deviate.

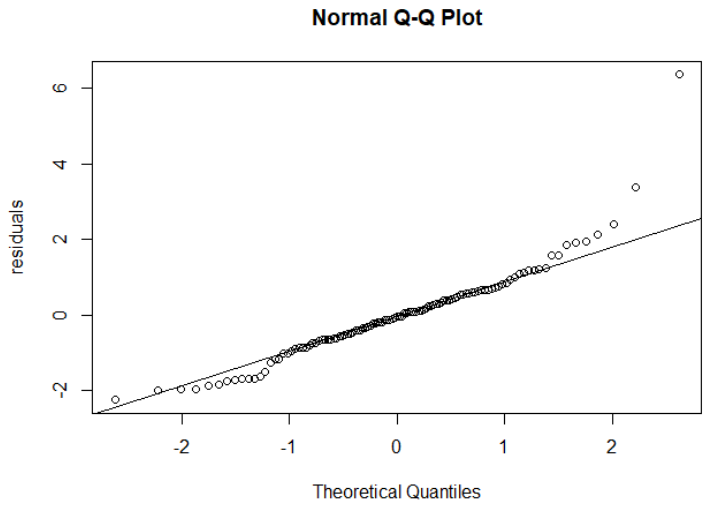


Figure 15. Normal Q-Q Plot

The Q-Q normal plot (Figure 16.) shows serious deviations on the right side, indicating that the error terms are not normally distributed. Despite the error terms for the most part following normal distribution, there are some residuals that are so abnormally large. To further evaluate this assumption, a Shapiro-Wilk test was conducted at the significance level of .05.

```
        Shapiro-Wilk normality test

data:  senic5.res
W = 0.90854, p-value = 1.05e-06
```

Figure 16. Shapiro-Wilk test output

With the p-value of 1.05e-06 extremely less below our significance level, we can reject the null hypothesis that the error terms are distributed normally, and accept the alternative hypothesis that the error terms are not distributed normally. We can conclude then that our assumption of normality is violated.

d) Outliers and Influential Points

We can check outlying (Y) observations through the plot of the jackknife residuals against the fitted values. Additionally, outlying observations of the predictor variables can be checked through the use of leverages.

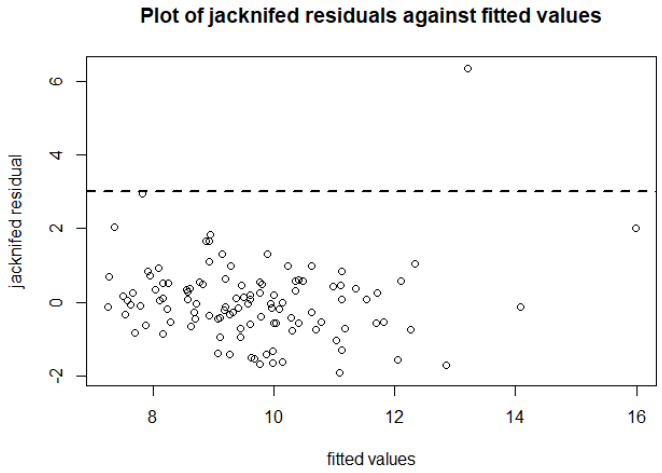


Figure 17. Scatter Plot of jackknife residuals against fitted values

As shown in the above plot (Figure 17.), we can see a single jackknife residual above the cutoff of three standardizations away, which is the value 47. Indicated that it is an outlier in (Y).

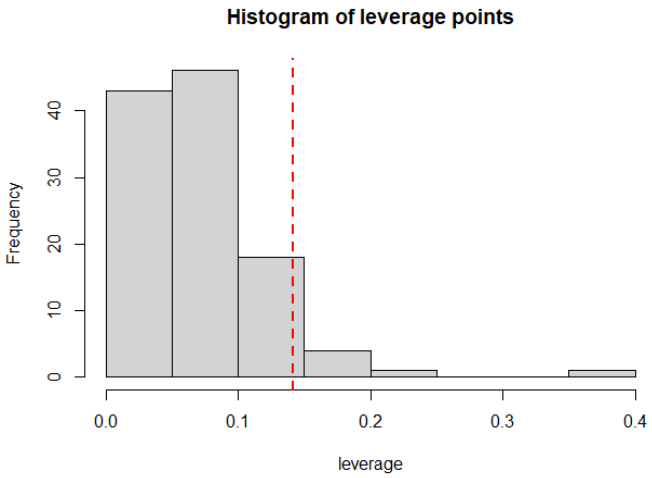


Figure 18. Histogram of leverage points

In the histogram of the leverage points (Figure 18.), we see that the cut off for the leverages is at .14. The high leverage values that indicate the X observations are outlying are 11, 46, 54, 66,

104, 112. For the influential observations we look at the plots of Cook's D, DFFITs, and DFBETAS.

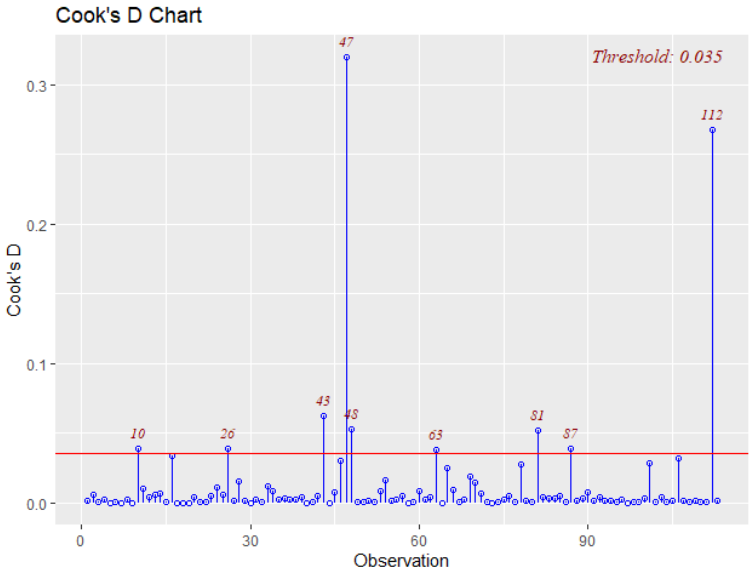


Figure 19. Plot of Cook's D

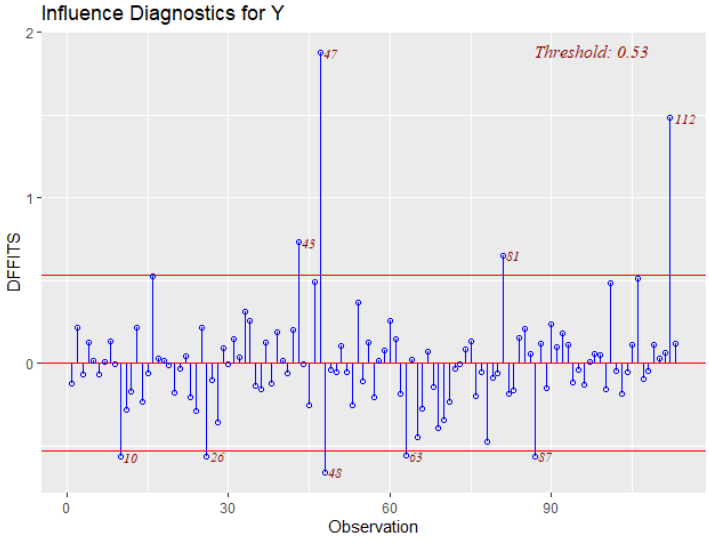


Figure 20. Plot of DFFITS

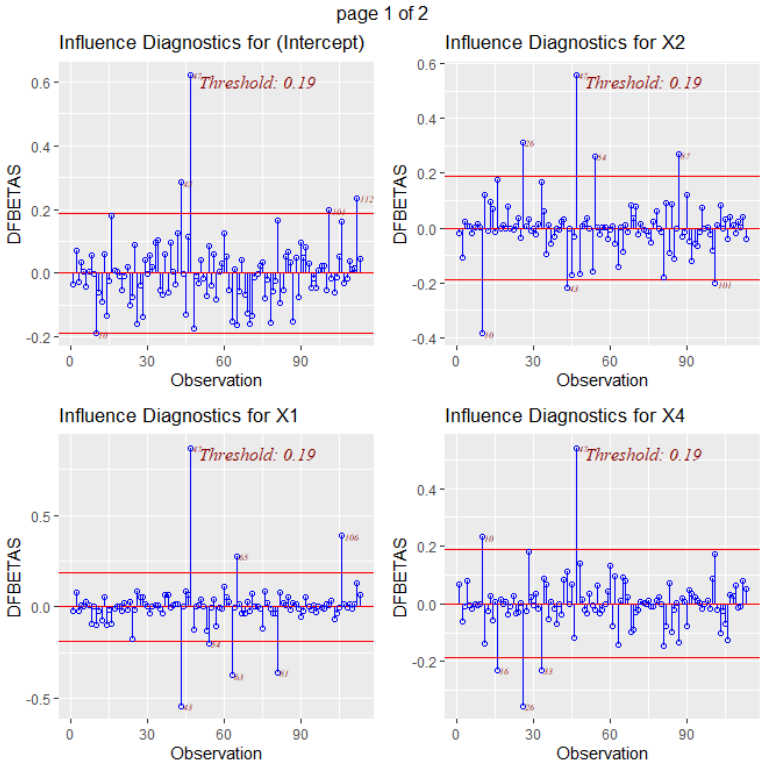


Figure 21. Plot one of DFBETAS

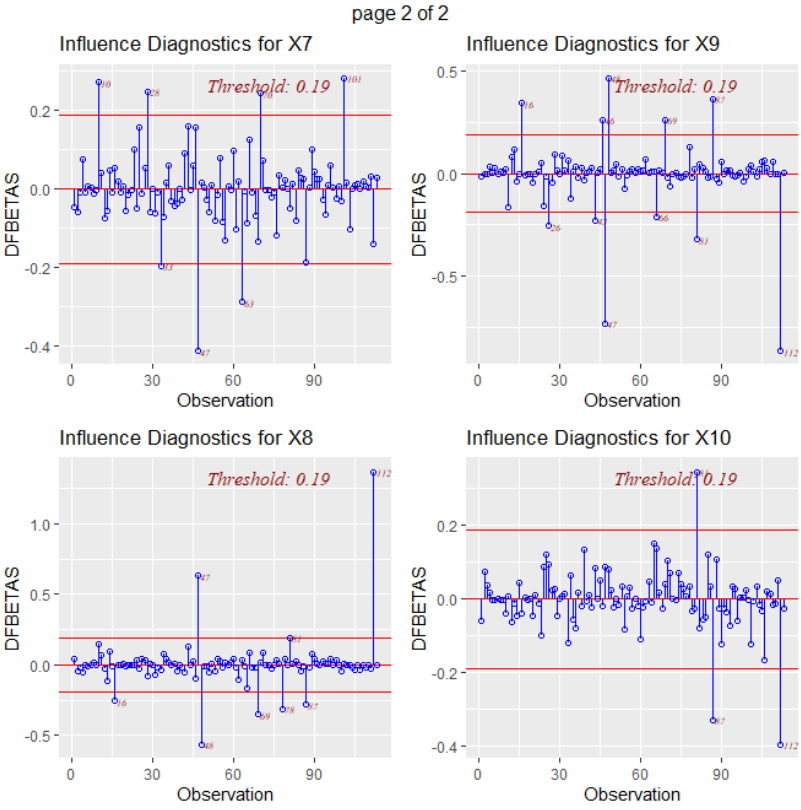


Figure 22. Plot two of DFBETAS

In the plots showing us our outlying values (Figure 19-22.), such as the Cook's D plot which shows us the influence on all the fitted values. The observations 10, 26, 43, 47, 48, 63, 81, 87, and 112 have significant influence on all the fitted values. With the plot of DFFITS we can identify observations that have substantial influence on their respective fitted values. The observations 10, 26, 43, 47, 48, 63, 81, 87, and 112 also have significant influence on their respective fitted values. Through the plots of DFBETAS we can discover what observation has influence on the different predictor variables including the intercept. The observations 10, 43, 47, 101, and 112 have influence on the intercept; 43, 47, 63, 65, 84, and 106 on (X1); 10, 26, 47, 54, 87, and 101 on (X2); 10, 16, 26, 33, and 47 on (X4); 10, 28, 33, 47, 63, 70, and 101 on (X7); 16, 47, 48, 69, 78, 87, and 112 on (X8); 16, 26, 43, 46, 47, 48, 66, 69, 81, 87, and 112 on (X9); 81, 87, and 112 on (X10).

F. Transforming Model

Remedial action is necessary because the assumption of normal error terms was violated. All other assumptions and potential issues like multicollinearity were addressed and do not need to be changed. To remediate the normality violation, a box cox transformation will be employed to determine the best lambda to change the power of the (Y) values.

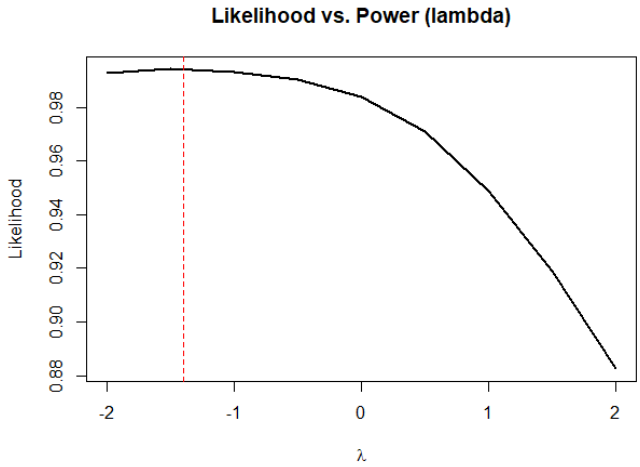


Figure 23. Likelihood values against lambda

In the box cox transformation (Figure 23.), we can see that the best power of the (Y) values is -1.400877 and a new model can be created. The model summary is below.

```
Call:
lm(formula = trans.Y ~ X1 + X2 + X4 + X7 + X8 + X9 + X10, data = senic5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.0200863 -0.0040058 -0.0005342  0.0036149  0.0152468

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0440077  0.0006330  69.524  < 2e-16 ***
X1          -0.0014887  0.0006398  -2.327 0.021896 *
X2          -0.0033640  0.0007915  -4.250 4.64e-05 ***
X4          -0.0010098  0.0007450  -1.355 0.178212
X7           0.0040024  0.0006832   5.858 5.43e-08 ***
X8          -0.0056015  0.0015787  -3.548 0.000582 ***
X9           0.0019825  0.0016078   1.233 0.220320
X10          0.0011658  0.0010880   1.072 0.286389
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.006729 on 105 degrees of freedom
Multiple R-squared:  0.5865,    Adjusted R-squared:  0.559
F-statistic: 21.28 on 7 and 105 DF,  p-value: < 2.2e-16
```

Table 3. Summary of the model output

As seen in the output of the summary, we can see through the p-value of the overall model that the model is significant compared with the model containing only the intercept. Additionally, approximately 55.9% percent of the variation in the data can be explained by the model.

G. Model Assumptions on Transformed Model

In order for our model to be a good choice for the dataset requires the assumptions of linearity, normality, homoscedasticity of error terms. Checking the multicollinearity of our predictor variables first, we find no significant issues as the VIF's are all below 10 indicating no serious multicollinearity issues.

```
> vif(senic5.lmfit)
      X1       X2       X4       X7       X8       X9      X10
1.012654 1.549749 1.373126 1.154718 6.165637 6.394902 2.928264
```

Figure 24. Variance inflation factors (VIF) of model predictor variables

The transformation has almost no effect on the multicollinearity. All the VIF's are below 10 indicating no serious multicollinearity issues.

a) Linearity

In a model that fulfills the linearity assumption the plots of the response variable against the predictor variables will show a straight line pattern, either positive or negative.

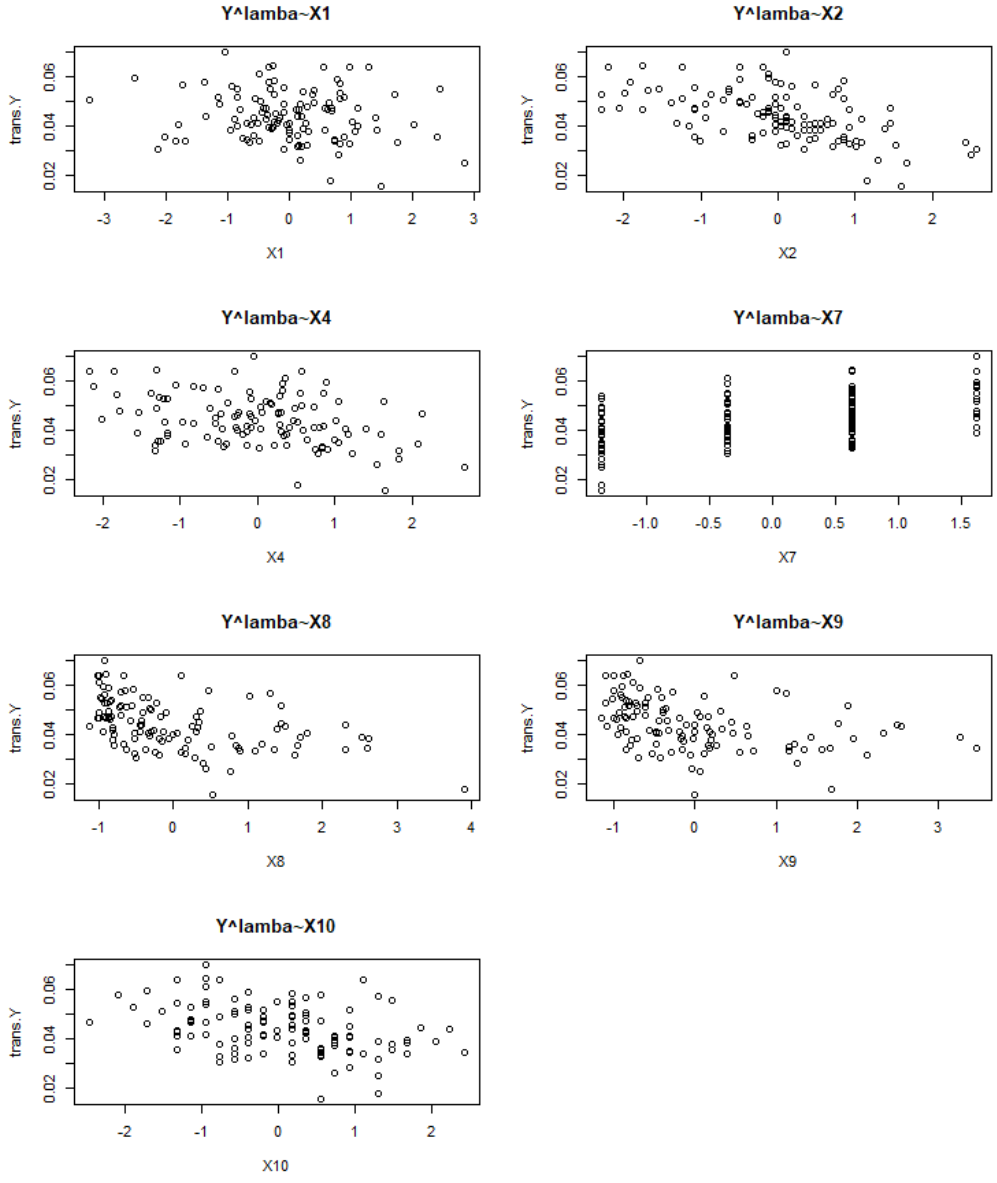


Figure 25. Scatter plot of response variable against predictor variables

At first glance, the predictor variable (X7) sticks out for its multimodal distribution indicating a non-linear trend; however, as mentioned before (X7) is a categorical variable and can be ignored. The rest of the predictor variables with the exception of (X8) appear very linear. The predictor variable (X2) has a slight curvilinear shape, but it is very modest. As such, the linearity assumption is valid.

b) Homoscedasticity

In a model that fulfills the assumption of equal error variances, we would expect that the plot of the residuals against the fitted values and the predictor variables to have a rectangular shape: they also do not possess a cone shape.

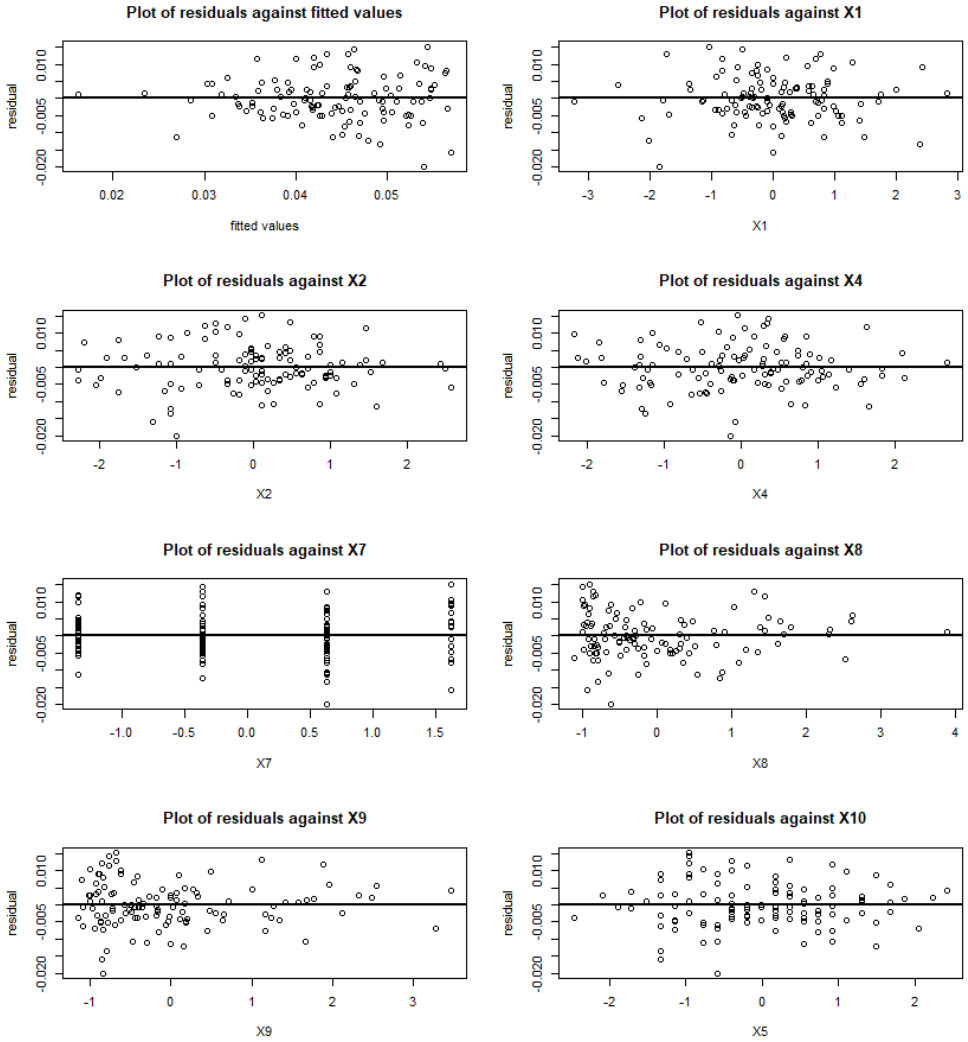


Figure 26. Scatterplot of residuals against fitted values and predictor variables

As shown in (Figure 26.) there appears to be some non-constant error variance evident in the plots of the fitted values, (X8), and (X9) from the slight cone shape. Overall, however there does not seem to be any major problems. To further analyze this assumption, a Breusch-Pagan Test was conducted with the significance level of .05.

```
> bptest(senic5.lmfit)

        studentized Breusch-Pagan test

data:  senic5.lmfit
BP = 7.4581, df = 7, p-value = 0.3828
```

Figure 27. Breusch-Pagan test output

With the p-value of .3823 greater than our significance level, we can retain the null hypothesis that there is a constant error variance and reject the alternative hypothesis which is that there is no constant error variance. Our assumption of homoscedasticity is valid. Note that this p-value is higher than before our transformation was applied, indicating the boxbox lambda has improved our assumption of homoscedasticity.

c) Normality

In a model that fulfills the normality assumption we would expect the normal qq plot of the error terms to follow the normal distribution line and not seriously deviate.

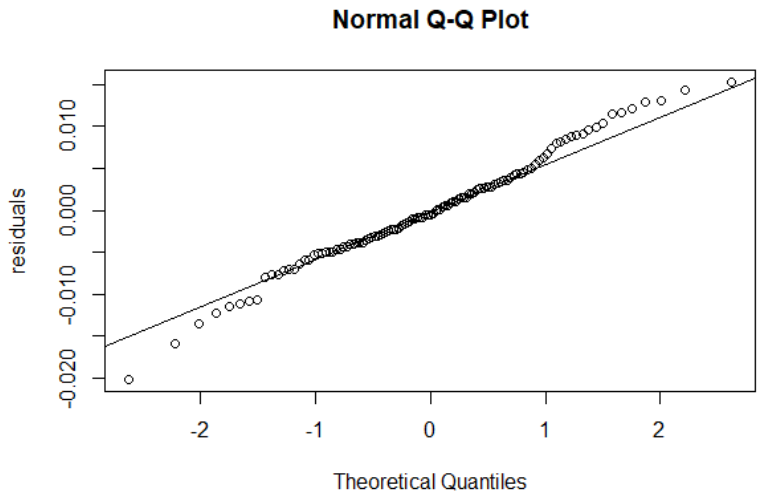


Figure 28. Normal Q-Q Plot

The Q-Q normal plot (Figure 28.) shows some slight deviations on the left side, indicating that the error terms are not solely normally distributed. Despite there being some residuals that are not following normal distribution, the error terms for the most part follow normal distribution. To further evaluate this assumption, a Shapiro-Wilk test was conducted at the significance level of .05.

```
> shapiro.test(senic5.res)

        Shapiro-Wilk normality test

data:  senic5.res
W = 0.98916, p-value = 0.5075
```

Figure 29. Shapiro-Wilk test output

With the p-value of .5075 which is above our significance level, we can retain the null hypothesis that the error terms are distributed normally, and reject the alternative hypothesis that the error terms are not distributed normally. We can conclude then that our assumption of normality is valid.

d) Outliers and Influential Points

We can check outlying (Y) observations through the plot of the jackknife residuals against the fitted values. Additionally, outlying observations of the predictor variables can be checked through the use of leverages.

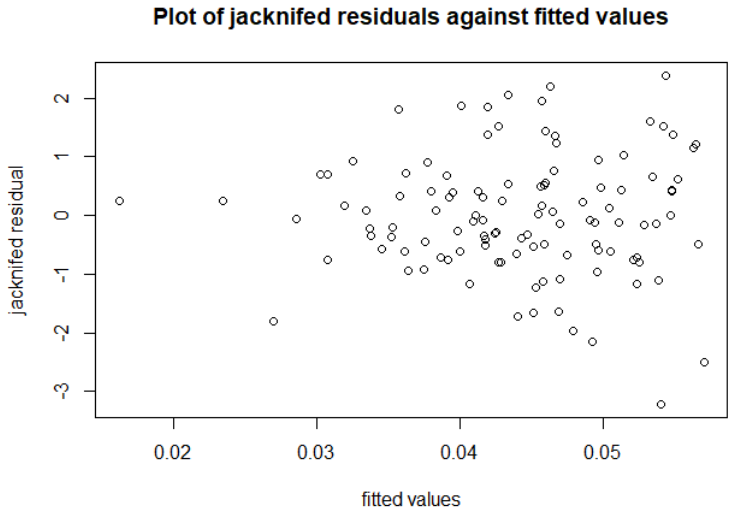


Figure 30. Scatter Plot of jackknife residuals against fitted values

As shown in the above plot (Figure 30.), we can see not a single jackknife residual above the cutoff of three standardizations away, indicating that there wasn't any outliers with respect to (Y).

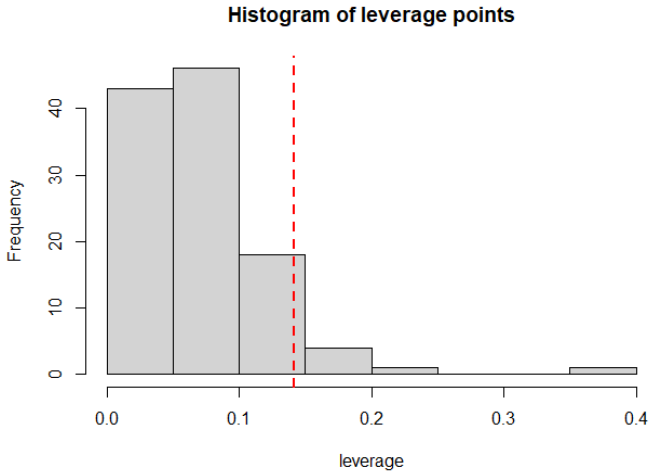


Figure 31. Histogram of leverage points

In the histogram of the leverage points (Figure 31.), we see that the cut off for the leverages is at .1415. The high leverage values that indicate the outlying (X) observations are 11, 46, 54, 66, 104, 112. For the influential observations we look at the plots of Cook's D, DFFITs, and DFBETAS.

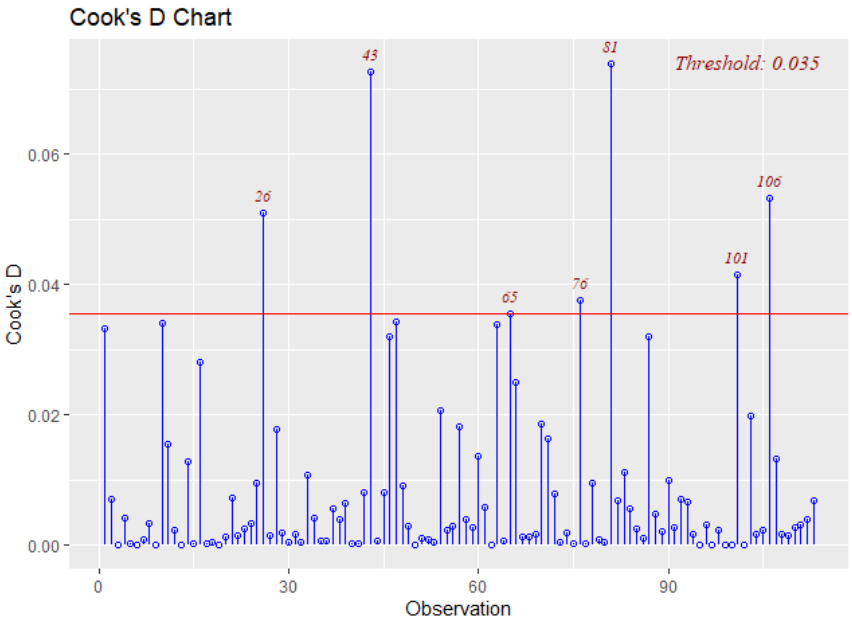


Figure 32. Plot of Cook's D

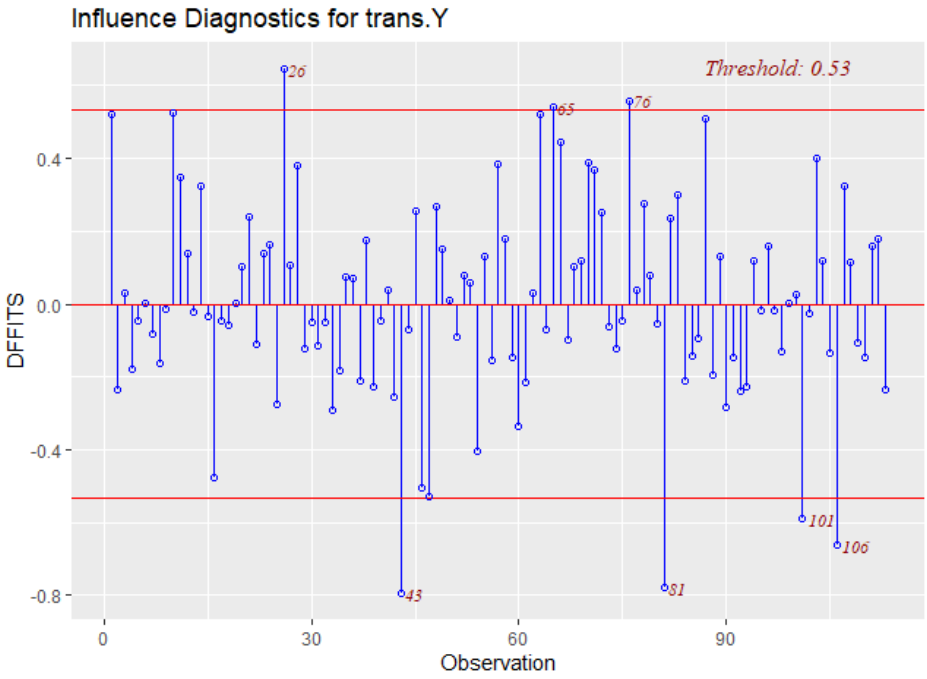


Figure 33. Plot of DFFITS

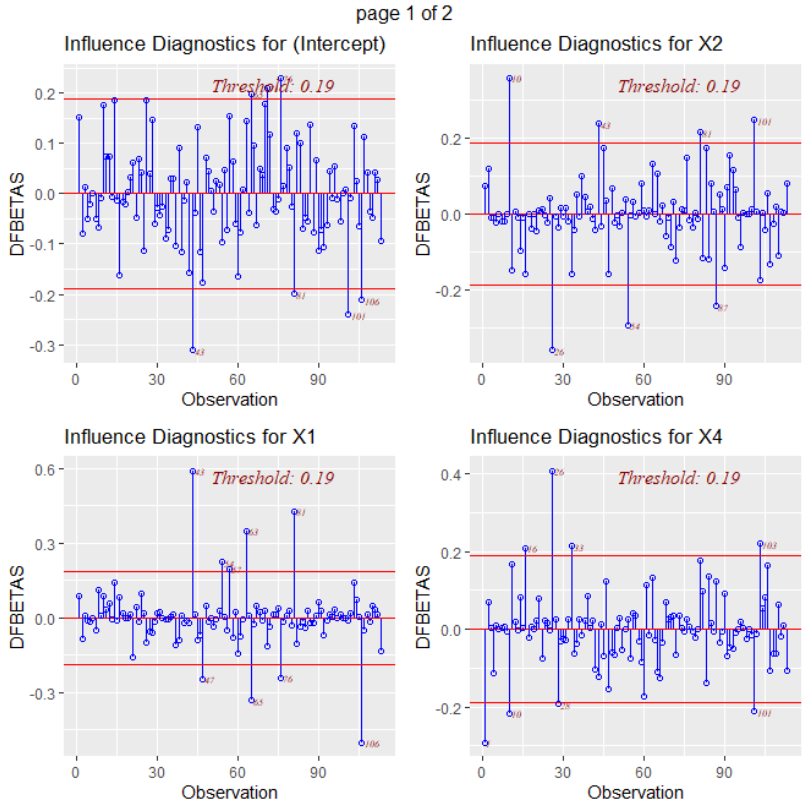


Figure 34. Plot one of DFBETAS

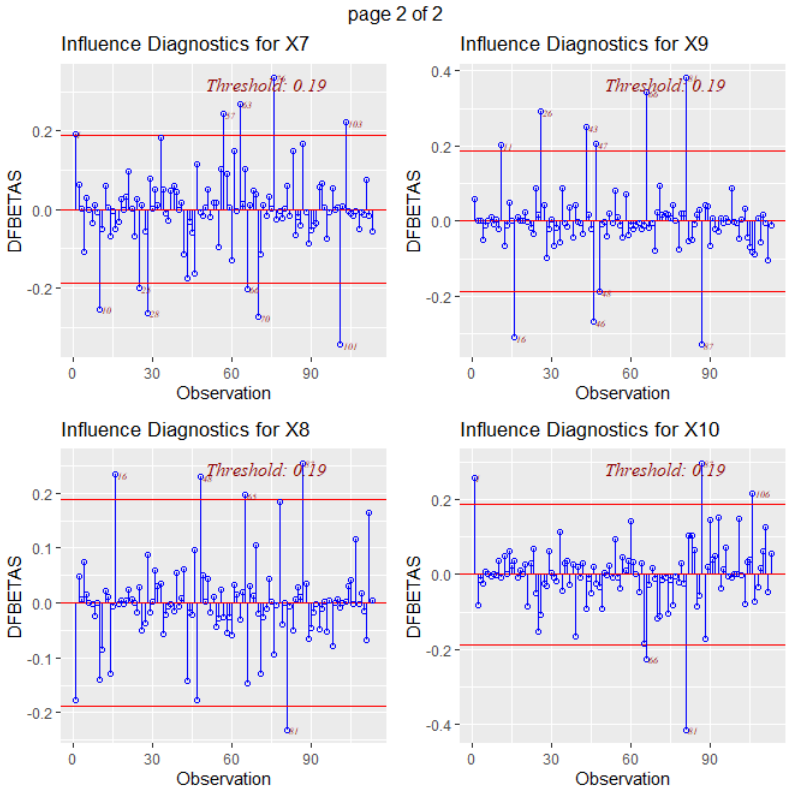


Figure 35. Plot two of DFBETAS

In the plots showing us our outlying values (Figure 19-22.), such as the Cook's D plot which shows us the influence on all the fitted values. The observations 26, 43, 65, 76, 81, 101, and 106 have significant influence on all the fitted values. With the plot of DFFITS we can identify observations that have substantial influence on their respective fitted values. The observations 26, 45, 45, 65, 76, 81, 101, and 106 also have significant influence on their respective fitted values. Through the plots of DFBETAS we can discover what observation has influence on the different predictor variables including the intercept. The observations 43, 65, 71, 76, 81, 101, and 106 have influence on the intercept; 43, 47, 54, 57, 65, 76, 81, and 106 on (X1); 10, 26, 43, 54, 81, 87, and 101 on (X2); 7, 10, 16, 26, 28, 33, 101, and 103 on (X4); 1, 10, 25, 28, 57, 63, 66, 70, 76, 101, and 103 on (X7); 16, 48, 65, 81, and 87 on (X8); 11, 16, 26, 43, 46, 47, 48, 66, 81, and 87 on (X9); 1, 66, 81, 87, and 106 on (X10).

III. Results

The final transformed and redone model that we determined best fit our dataset is below. Additionally, is the ANOVA table of the model.

```
Call:
lm(formula = trans.Y ~ X1 + X2 + X4 + X7 + X8 + X9 + X10, data = senic5)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0200863 -0.0040058 -0.0005342  0.0036149  0.0152468

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0440077  0.0006330  69.524  < 2e-16 ***
X1          -0.0014887  0.0006398  -2.327 0.021896 *
X2          -0.0033640  0.0007915  -4.250 4.64e-05 ***
X4          -0.0010098  0.0007450  -1.355 0.178212
X7           0.0040024  0.0006832   5.858 5.43e-08 ***
X8          -0.0056015  0.0015787  -3.548 0.000582 ***
X9           0.0019825  0.0016078   1.233 0.220320
X10          0.0011658  0.0010880   1.072 0.286389
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.006729 on 105 degrees of freedom
Multiple R-squared:  0.5865,    Adjusted R-squared:  0.559
F-statistic: 21.28 on 7 and 105 DF,  p-value: < 2.2e-16
```

Figure 36. Summary of the model output

```
Analysis of Variance Table

Response: trans.Y
           Df    Sum Sq   Mean Sq F value    Pr(>F)
X1          1 0.0002387 0.0002387  5.2724   0.02365 *
X2          1 0.0033460 0.0033460 73.9031 8.447e-14 ***
X4          1 0.0002423 0.0002423  5.3526   0.02264 *
X7          1 0.0019142 0.0019142 42.2789 2.734e-09 ***
X8          1 0.0008323 0.0008323 18.3823 4.024e-05 ***
X9          1 0.0001178 0.0001178  2.6027   0.10969
X10         1 0.0000520 0.0000520  1.1482   0.28639
Residuals 105 0.0047539 0.0000453
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 37. Summary of the ANOVA model output

In the output of the summary of the model (Figure 36.), we can see through the p-value (2.2e-16) of the overall model that the model is significant compared with the model containing only the intercept through the F-test. Additionally, approximately 55.9% percent of the variation in the data can be explained by the model. In figure 36, we can also see that all predictor variables except (X4), (X9), and (X10) are significant. Likely the reason that (X9) and (X10) are still not

significant in the model is due to their very high correlation with other predictor variables. In the anova table of the model (Figure 37.), we can see the f-values and their respective p-values of the predictor variables, indicating we can conclude a linear relationship between all the predictor variables excluding (X9) and (X10). For the regression coefficients, each predictor variable represents the main effect as there are no interaction terms. The mean value of the average length of stay of the patients increases by .044007 when all other predictor variables are at zero; decreases by .0014887 for every unit increase in the average age of the patient (X1); decreases by .0033640 for every unit increases in the percent probability of acquiring an infection (X2); decreases by .0010098 for every unit increase in the ratio of the number of X-rays performed without sign of symptoms time 100 (X4); increases by .0040024 for unit increase in the geographic region encoding scheme 1=NE, 2= NC, 3=S, and 4=W (X7); decreases by .0056015 for every unit increase in the average number of patients per day (X8); increases by .0019825 for every unit increase the number of full time nurses working (X9); increases by .0011658 for every unit increase in the percent of potential services that are provided by the hospital (X10). Note that these represent the partial effect, because this mean change in (Y) only happens when all other predictor variables are held constant.

VI. Conclusions

In our process of creating a linear regression model for the dataset, we explored visual tools to investigate relationships between the variables in the form of boxplots, histograms, added-variable plots, scatterplot matrices, and correlation plot matrices. After understand the underlying relationships between the response and predictor variables, we conducted model selection through stepwise selection through criterion, such as adjusted R2, Mallow's CP, and

AIC/BIC; however, after starting the assumption checking serious multicollinearity problems were detected prompting us to remove the variable (X5) from selection. Model selection was performed again. The normality assumption was violated and a box cox transformation was conducted to remediate. With the output value of lambda that optimizes the likelihood function, we adjusted the power of our (Y) values. The assumptions of this transformed model were checked and all were valid, additionally, we found no outliers with respect to our fitted values. This transformed model approximately explains 55.9% percent of the variation in the average length of stay of a patient through the predictor variables: (X1), (X2), (X4), (X7), (X8), (X9), (X10). To improve this model, investigation into the non-significant predictor variables should be conducted. Despite no serious multicollinearity issues and the overall significance of the model, the predictor variables (X9) and (X10) are not significant.