A

STATISTICAL DATA ANALYSIS

REPORT

BY

AYODEJI AYOOLA

ON

CARE ONE HOSPITAL (RADIOLOGY WORKFLOW)

Contents

1. PROBLEM STATEMENT

The time taken to receive an ordered X-ray is excessively long in Care One Hospital which gave rise to patients' length of stay at the hospital. The hospital data from 2016 to 2017 needed to be analyzed and applicable solution needed to be suggested.

2. PROJECT ANALYSIS

After reading the data into R, we viewed the structure of each of the variables provided and treated it by converting each of them based on the variable type.

• y which denotes our response indicates Ordered to complete minutes was converted to numerical variable
• x1 which denotes the patients age was converted to numerical variable,
• x2 which denotes the radiology technician was converted to categorical variable
• x3 which denotes the catalog code was converted to categorical variable
• x4 which denotes the in rad.room with the base variable of "1" (if performed in radiology room) when converted to a categorical variable
• x5 which denotes the patient type with a base variable of "IP-in patient" with respect to the other variables when converted to a categorical variable
• x6 which denotes the priority with a base variable of Routine with respect to STAT when converted to a categorical variable
• x7 which denotes the Lock.at.Exam.Complete converted was converted a categorical variable
• x8 which denotes the Exam completed bucket with a base variable of time "12am-8am" with respect to other variable when converted to categorical
• x9 which denotes the Exam room was converted to a categorical variable

Fitting the linear model with all the predictor variables regressed on the response(y)

```
model<-lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9, data=dat)
Plot (model)
```

After fitting the linear model (see Appendix A), we went on to check the model adequacy by observing the plots.
Checking the Residual vs fitted plot, we observed that a pattern exists in the distribution of the data across the graph Indicating the constant variance assumption is questionable.

Checking for the Normal Q-Q plot, we observed we do not have a fairly straight line Indicating the Normality is questionable

The square root of standardized residual vs fitted values also indicates a pattern, nullifying out assumption for normality. The residuals vs leverage plot also indicates we have a pattern, nullifying our assumption for normality

The overall p value for the model is 2.2e-16 indicating it is highly significant. The multiple and Adjusted R values are 0.2561 and 0.252 respectively, which implies the model is not accurate enough.
From the results above, we conclude to use a power transformation (Box-cox) because the plots obtained were questionable.

2.1. Applying the box-cox transformation

```
library(MASS)
b<-boxcox(model)
lambda<-b$x
liklihood<-b$y
which.max(liklihood)
#the liklihood value is 45
lambda[45]
#the Lambda value is  -0.2222222
dat$newy<-log(dat$y)
#Using a log transform because the value of lambda is close to Zero, from the plot
head(dat)
```

Observing the model adequacy plots Checking the Residual vs fitted plot (see Appendix B), we observed that the pattern no longer exists in the distribution of the data across the graph Indicating we have a constant variance (scatter plot) across the data distribution.

Checking for the Normal Q-Q plot, we observed we have a fairly straight-line indicating normality exists
.
The square root of standardized residual vs fitted values also indicates constant variance assumption is satisfied, the residuals vs leverage plot also indicates constant variance assumption is satisfied.

The overall p-value for the model is 2.2e-16 indicating it is highly significant. The multiple and Adjusted R values are 0.5809 and 0.5787 respectively, implying a better R-Square and adj. R-Square values than the first model.

Checking for Outliers,leverage,and and influential points Checking the Residual vs fitted plot, we observed that we have data points (69, 238, 85) which where outliers From the Normal Q-Q plot we had data points (69,238,85) which were outliers The square root of standardized residual vs fitted values outliers with data points (69, 238, 85)

In order to obtain the best fitting model, the stepwise procedures (forward, backward and both) were executed

2.2. Implementing the forward stepwise regression model

```
model3<-lm(newy~1,data=dat)
formula(model3)
step(model3,scope~x1+x2+x3+x4+x5+x6+x7+x8+x9,direction = "forward")
```

The forward stepwise regression gave us the model newy = x6+x9+x8+x3+x2+x7+x5+x1
Arranging the variables in order of ascending AIC values and dropping the variable x4 completely

```
#Dropping predictor variables based on the AIC (from highest to lowest)
summary(lm(newy~x6+x9+x8+x3+x2+x7+x5+x1,data = dat))
# dropping x1 based on AIC obtained from forward stepwise and level of significance
summary(lm(newy~x6+x9+x8+x3+x2+x7+x5,data = dat))
# dropping x5 based on AIC obtained from forward stepwise and level of significance
summary(lm(newy~x6+x9+x8+x3+x2+x7,data = dat))
# dropping x7 based on AIC obtained from forward stepwise and level of significance
summary(lm(newy~x6+x9+x8+x3+x2,data = dat))
# dropping x2 based on AIC obtained from forward stepwise and level of significance
summary(lm(newy~x6+x9+x8+x3,data = dat))
# dropping x3 based on AIC obtained from forward stepwise and level of significance
summary(lm(newy~x6+x9+x8,data = dat))
# dropping x8 based on AIC obtained from forward stepwise and level of significance
summary(lm(newy~x6+x9,data = dat))
```

From the forward stepwise regression, we propose to use two model:

Our first proposed model is newy = x6 + x9 + x8, has an overall p value of < 2.2e − 16 R-Square and adj. R-Square of 0.537 and 0.5368 respectively.

Our second proposed model is newy = x6 + x9, has an overall p value of < 2.2e − 16 R-Square and adj. R-Square of 0.5011 and 0.501 respectively.

2.3. Implementing the backward stepwise regression model

```
model<-lm(newy~x1+x2+x3+x4+x5+x6+x7+x8+x9,data=dat)
formula(model)
step(model,direction = "backward")
```

The backward stepwise regression gave us the model newy = x1 + x5 + x7 + x2 + x8 + x3 + x9 + x6, Arranging the variables in order of ascending AIC values and dropping the variable x4 completely

```
#Dropping predictor variables based on the AIC
summary(lm(newy ~ x1+x5+x7+x2+x8+x3+x9+x6, data = dat))
# dropping x6 based on AIC obtained from backward stepwise and level of significance
summary(lm(newy ~ x1+x5+x7+x2+x8+x3+x9, data = dat))
# dropping x9 based on AIC obtained from backward stepwise and level of significance
summary(lm(newy ~ x1+x5+x7+x2+x8+x3, data = dat))
# dropping x3 based on AIC obtained from backward stepwise and level of significance
summary(lm(newy ~ x1+x5+x7+x2+x8, data = dat))
# dropping x8 based on AIC obtained from backward stepwise and level of significance
summary(lm(newy ~ x1+x5+x7+x2, data = dat))
# dropping x2 based on AIC obtained from backward stepwise and level of significance
summary(lm( newy ~ x1+x5+x7,data = dat))
# dropping x7 based on AIC obtained from backward stepwise and level of significance
summary(lm(newy ~ x1+x5,data = dat))
```

We observe the final model from the both step wise is the same as the model from the backward stepwise which is newy ~ x1+x5

2.4. Using OLSRR

Implementing the olsrr function to obtain a more accurate model, the olsrr function was implemented because it considers major criteria such as the AIC and adjusted R-square and other parameters simultaneously

Note, the olsrr function was implemented on the initial model 1 which contains all predictor variables given which were (x1+x2+x3+x4+x5+x6+x7+x8+x9) The olsrr could not process an output for the large number of data set

Also, the dredge function was implemented on the initial model 1 which contains all predictor variables given which were (x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9) The dredge could not process an output for the large number of data set

Running olsrr with this predictor variables gotten from the stepwise models:

```
library(olsrr)
fullmodel5<-lm(newy~x6+x9+x8,data=dat,na.action = "na.fail")
summary.fit1<-ols_step_best_subset(fullmodel3)
```

From the olsrr output we observed that the predictor variable x6 has significant AIC and a moderate R-square and adjusted R-square values so we regress the response newy on x6

From the olsrr method, x6 seems to be the most significant with an overall p value of 2.2e−16 with R2 and adj. R2 of 0.4239 and 0.4239 respectively.

3. RECOMMENDATION AND CONCLUSION

Reaching a conclusion, for our proposed model we have:
full model1: newy = x6 (proposed model 1)
full model2: newy x6 + x9 (proposed model 2)
full model3: newy x6 + x9 + x8 (proposed model 3)
full model4: newy x1 + x5 (proposed model 4)

After analyzing the data, it was observed that one particular predictor variable (x6 which denotes priority) recurred and has major influence on the hospitals delivery time of radiology results

So, if the hospital was to consider just one predictor variable, they should consider x6 which denotes priority, this is because the data shows that the hospital is giving far more priority to STAT patient than routine from the overall result of analysis done, the variable x6 most importantly contributes to the uptick in patients' length of stay and for hospital to solve this problem, the hospital needs to find a balance between prioritizing the routine and STATS patient.

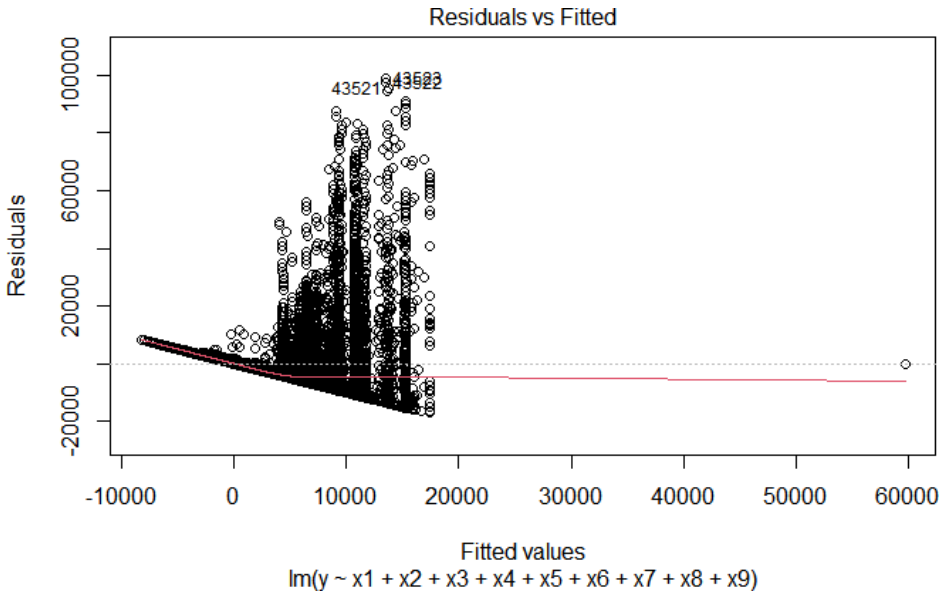
If the hospital was to consider two predictor variables, they should go with the second model. This model shows that the priority(x6) and exam rooms(x9) are the significant variables contributing to the problem A process can be generated to handle each case based on priority. Example, for all STAT cases, portable X-ray machines can be taken to the patient's room by radiology technician While all routine cases can be handled at the radiology room.

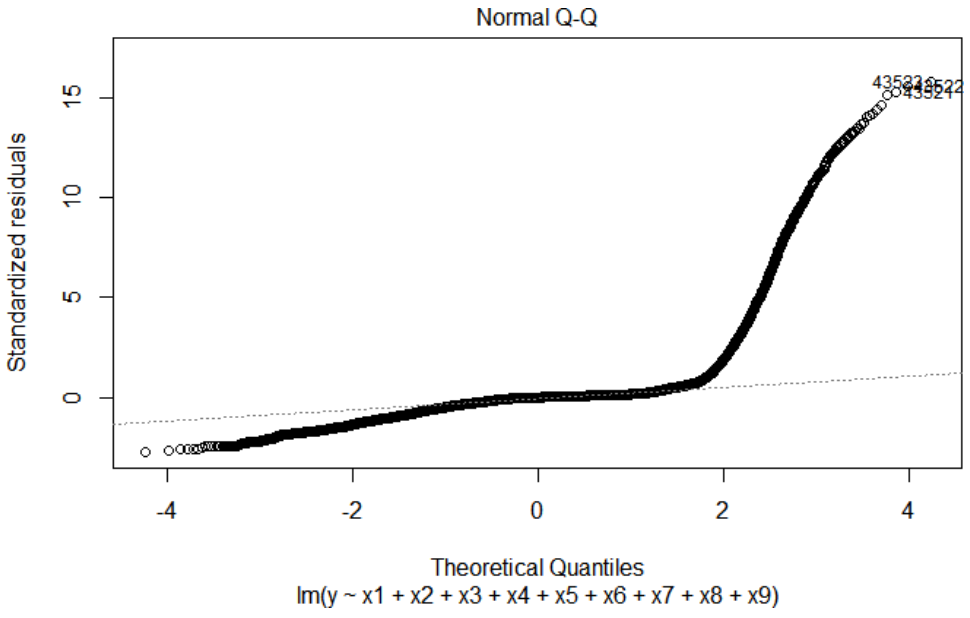
If the hospital was to consider three predictor variable, they should go with the third model This model shows that the priority(x6), exam rooms(x9) and exam completed bucket(x8) are the significant variables Since the time period taken to complete the exam will determine how many of the patients will be attended to over a certain period So the hospital should consider dividing the longer shifts into two and hire technicians for these shifts.

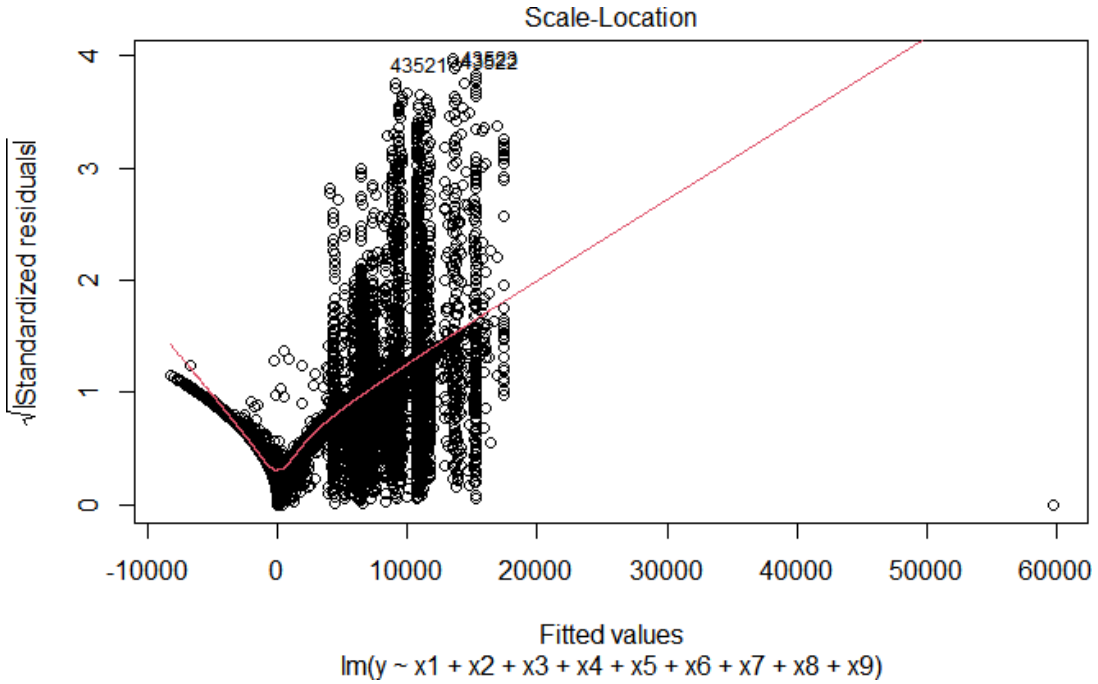
There is also one model that can be considered which is the fourth model. This model shows that the patient age(x1) and patient type(x5) can be causing the patients to stay long because older patients tend to stay longer in the hospital increasing the uptick length of stay at the hospital. Also, there are many patients in (IP, OPEC, OPOBS,OPSRG), which also translates to an increase in the uptick length of stay at the hospital.

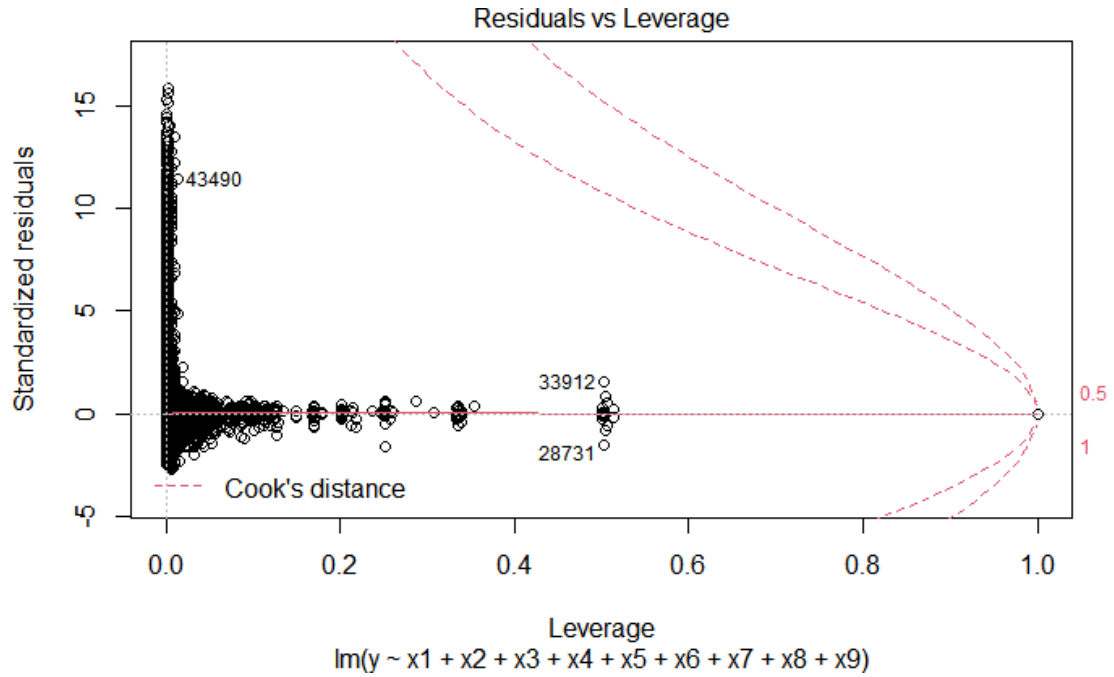
APPENDIX A

Plot of fitted linear model y~x1+x2+x3+x4+x5+x6+x7+x8+x9



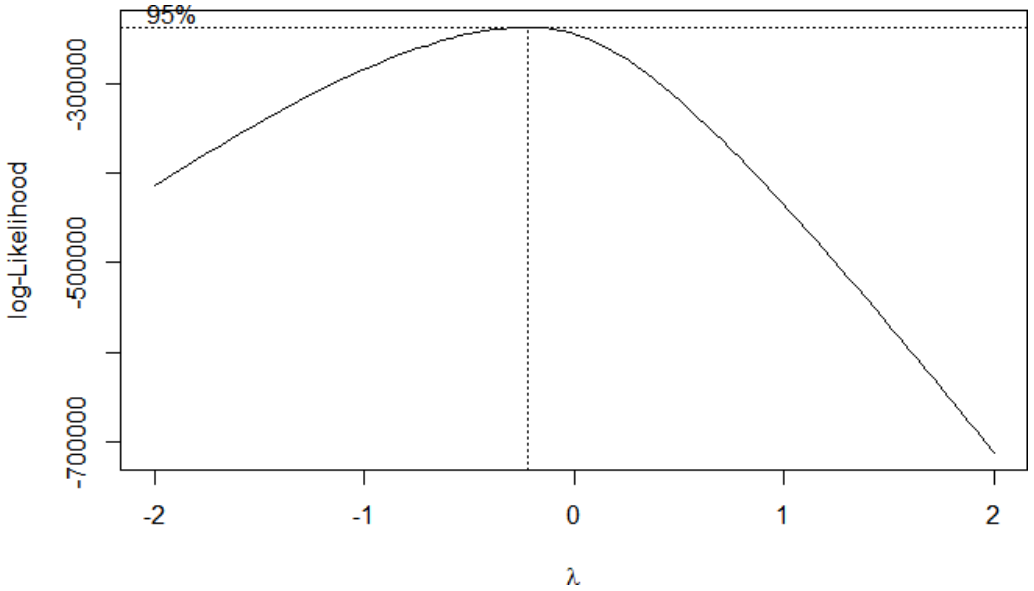


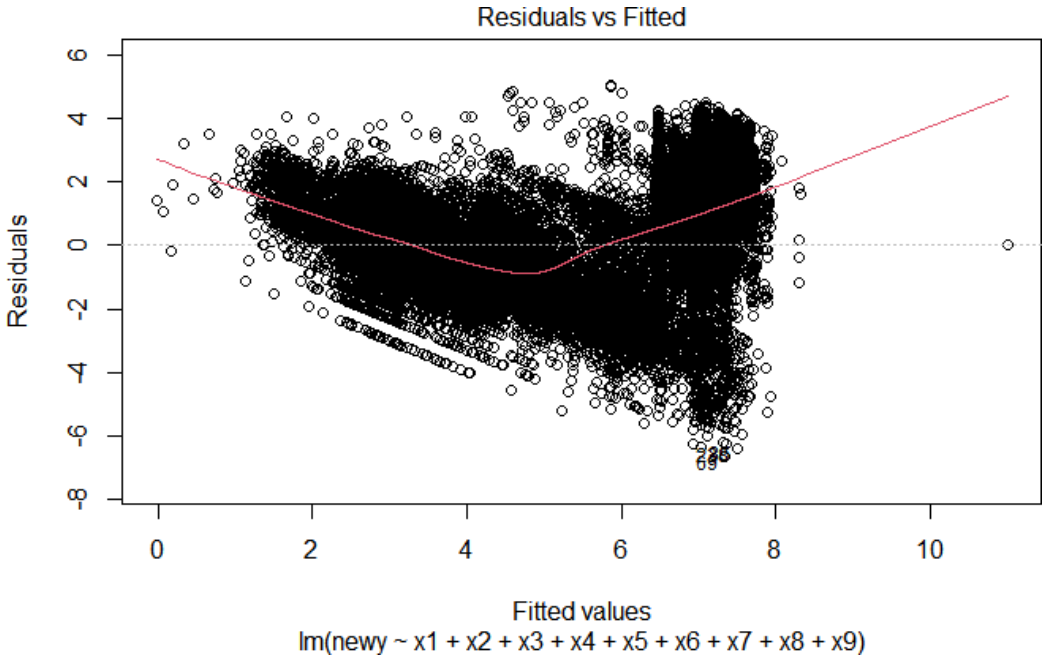

Scale-Location
√|Standardized residuals|
Fitted values
lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9)

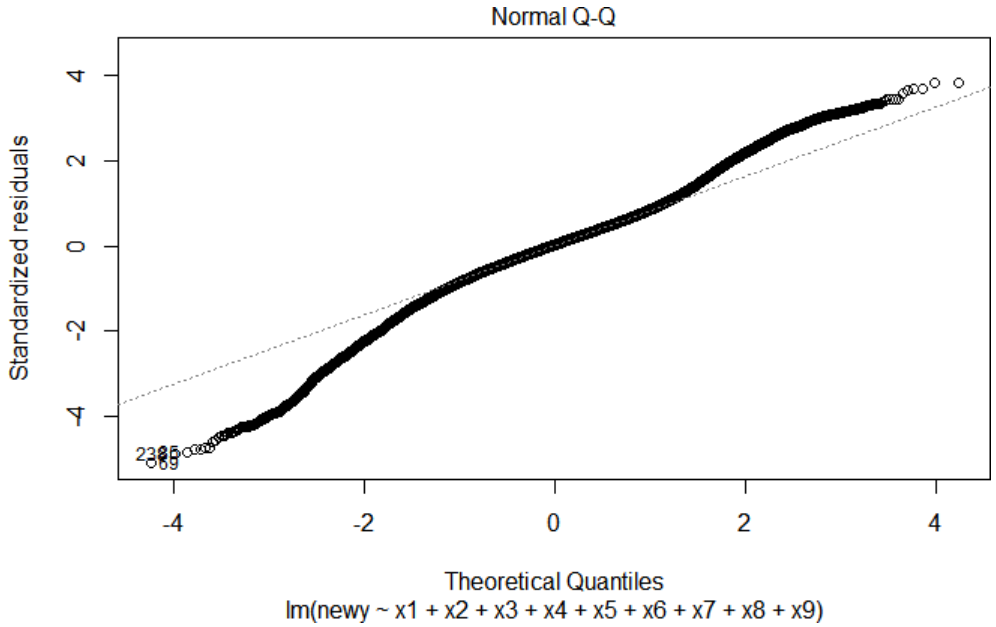

Residuals vs Leverage
Standardized residuals
43490
33912
28731
Cook's distance
0.5
1
Leverage
lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9)

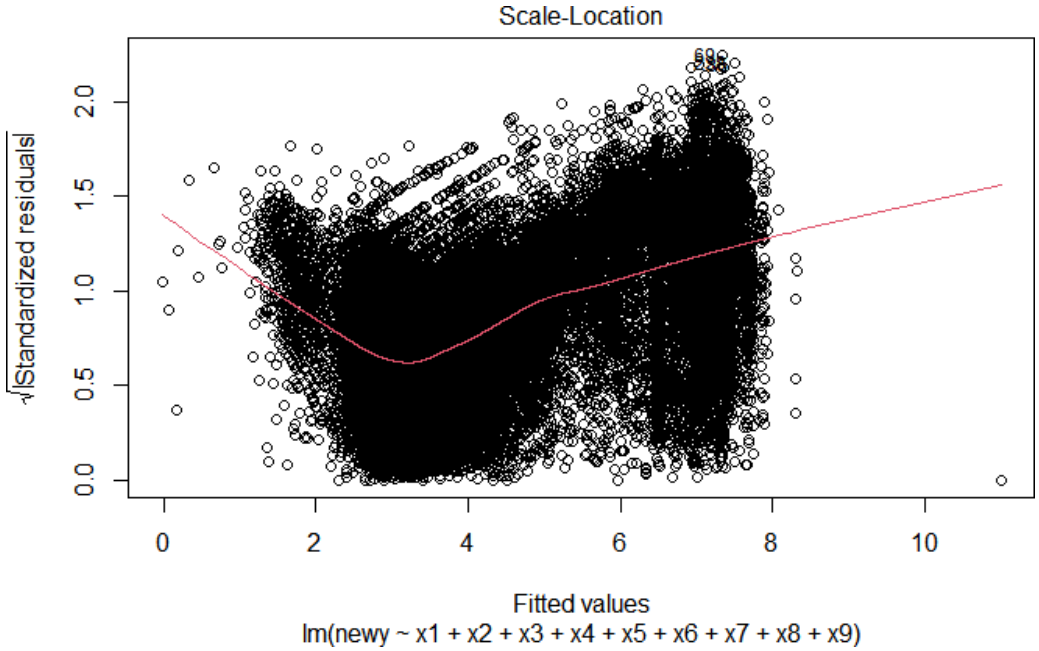
APPENDIX B

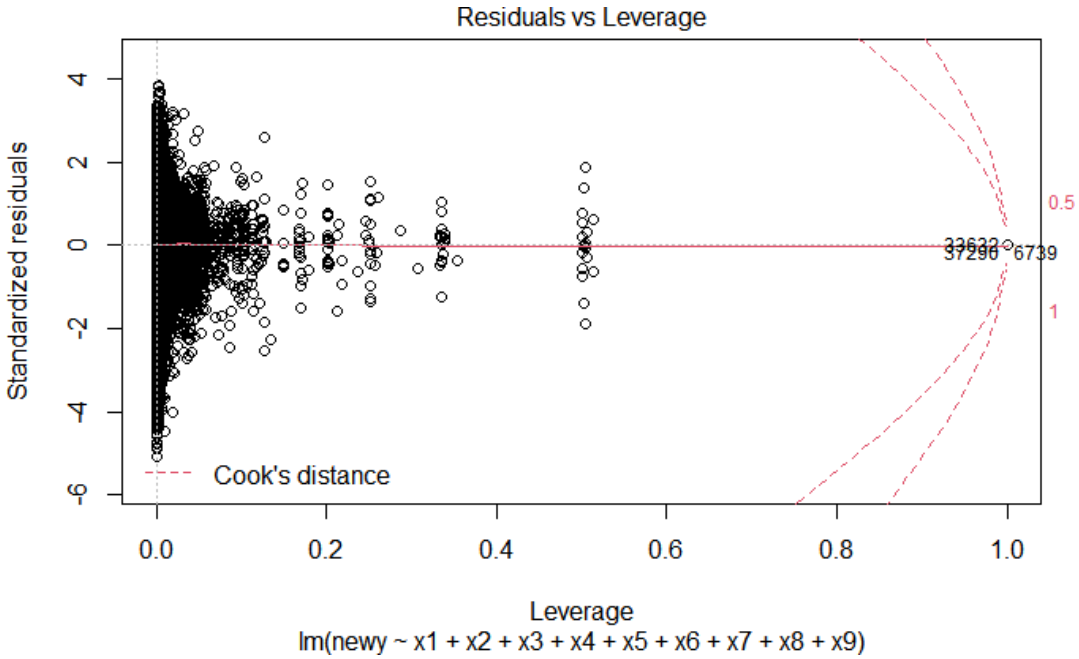
Plot for fitted linear model y~x1+x2+x3+x4+x5+x6+x7+x8+x9 after box-cox transformation






Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(newy ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9)


Scale-Location
√|Standardized residuals|
Fitted values
lm(newy ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9)


Residuals vs Leverage
Standardized residuals
4
2
0
-2
-4
-6
0.0
0.2
0.4
0.6
0.8
1.0
Leverage
lm(newy ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9)
Cook's distance
0.5
1
33632
37290
6739