Regression on 3 partitions

Suresh Ooty

Table of Contents

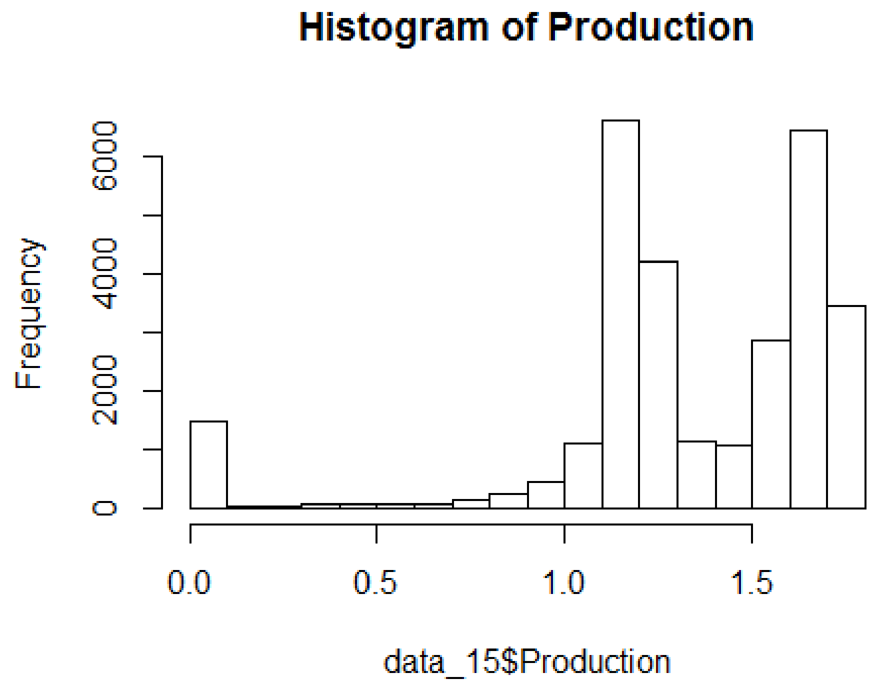
Dataset Reduction & Regression 1
Year 2015 dataset Overview 1
Data Split 2
LOG transformation 4
Iteration 1 Regression Models 6
Set 1 6
Set 14 13
Set 18 22
Conclusions 26

Dataset Reduction & Regression

Year 2015 dataset Overview

Based on the observations made earlier, the dataset was sliced in to three partitions based on the production value to study on the predictor variables and other influences if any, on the response variable *Production*. If not a solid Regression model that is predict worthy is an outcome of this study, this approach helped researcher to stream line the focus on the dataset in a larger extent of behaviour of response variable, outliers, influence of time component and specific patterns. These patterns may be an indication that at different production capacity of the plant, the operating conditions (obvious) were different, as the type of paper manufactured was different. This is yet to validated with the dataset provider.

```r
hist(data_15$Production, main = "Histogram of Production")
```



__Notes:__

1. Production Data is negatively skewed, too many Production values below 0.6 and many of them are '0'
2. The histogram on Production variable indicate that the distribution is bi-modal(above 0.8). There is a posssibility two or more different models required to predict the response (*Production*)
3. There is a possibility that there could be third model on the response value between 0.01 to 0.8

As this is a production unit, it is possible that the production value ranges could differ when the paper types are changed.

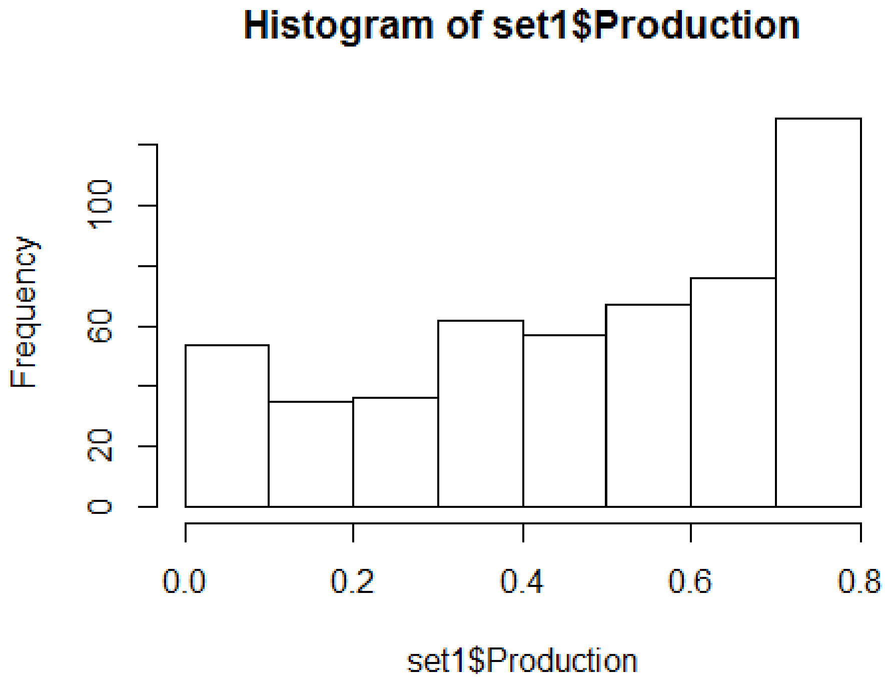
Note : This is assumption, the researcher has no idea about the production value ranges against paper type.

A Data elimination is necessary for ZERP values in the response variable

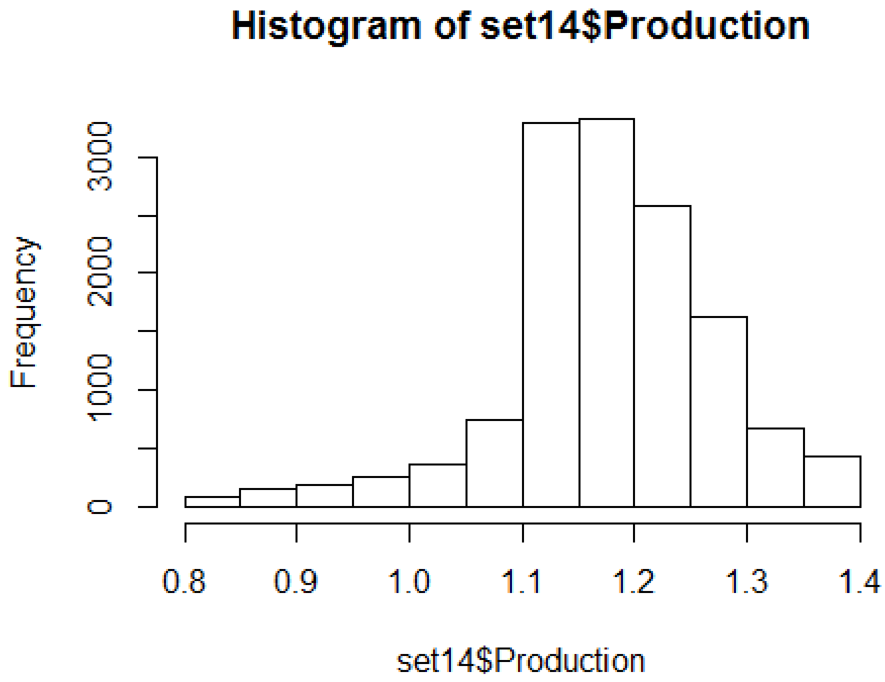
Data Split

```
#splitting the dataset based on production value ranges {0.01 to 0.8},{0.8 to 
1.4} & {above 1.4}
temp<-data_15[,-1]
temp <- cbind("ID" = sprintf("%d", 1:nrow(temp)), temp)
set1 <- temp[which(temp$Production <= 0.8),]
set1 <- set1[which(set1$Production > 0),]
temp <- temp[which(temp$Production > 0.8),]
set14 <-temp[which(temp$Production <= 1.4),]
set18 <-temp[which(temp$Production > 1.4),]
```

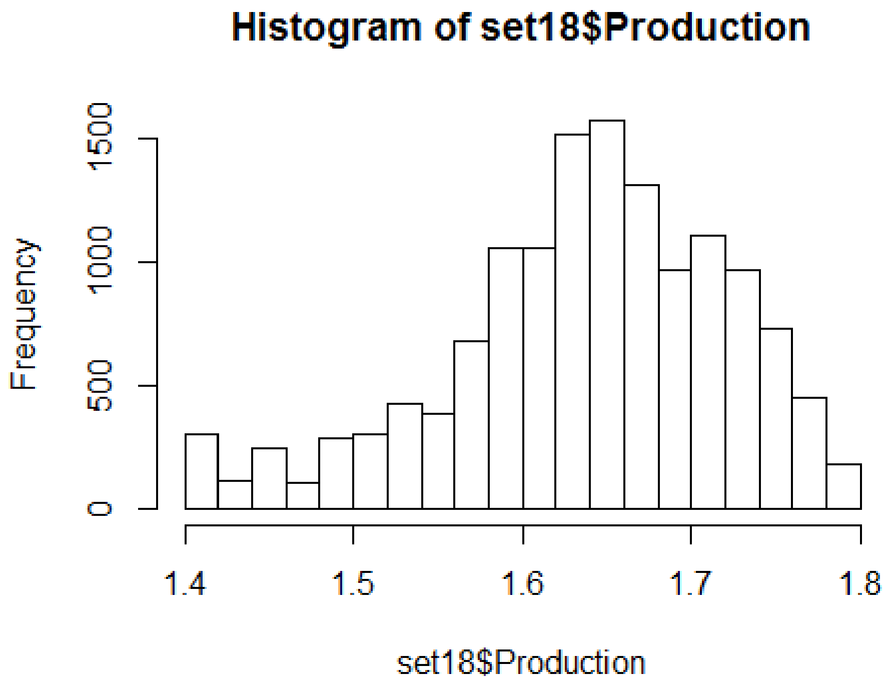
```
hist(set1$Production)
```



```
hist(set14$Production)
```

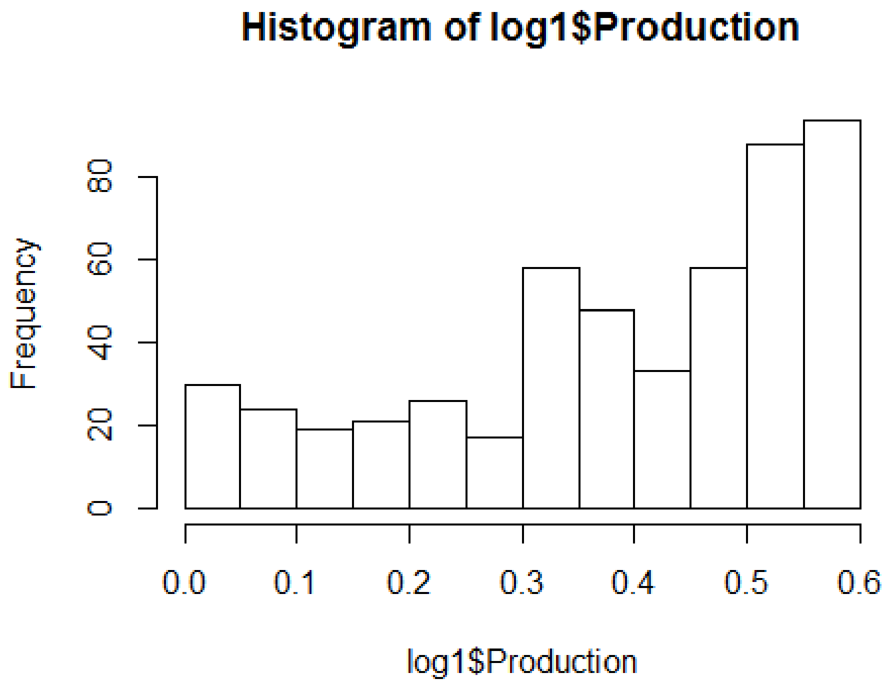


```r
hist(set18$Production)
```

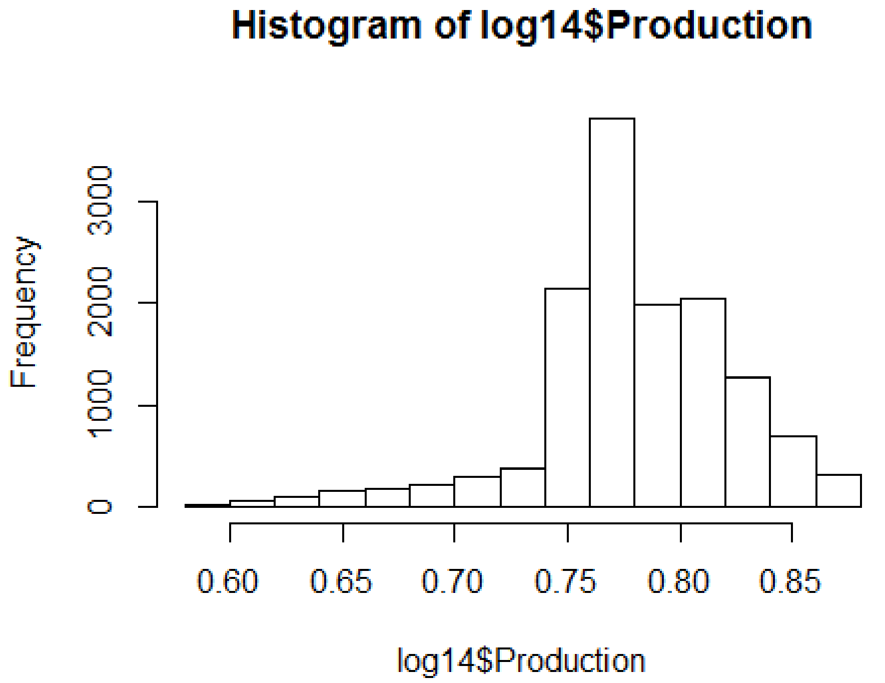


LOG transformation

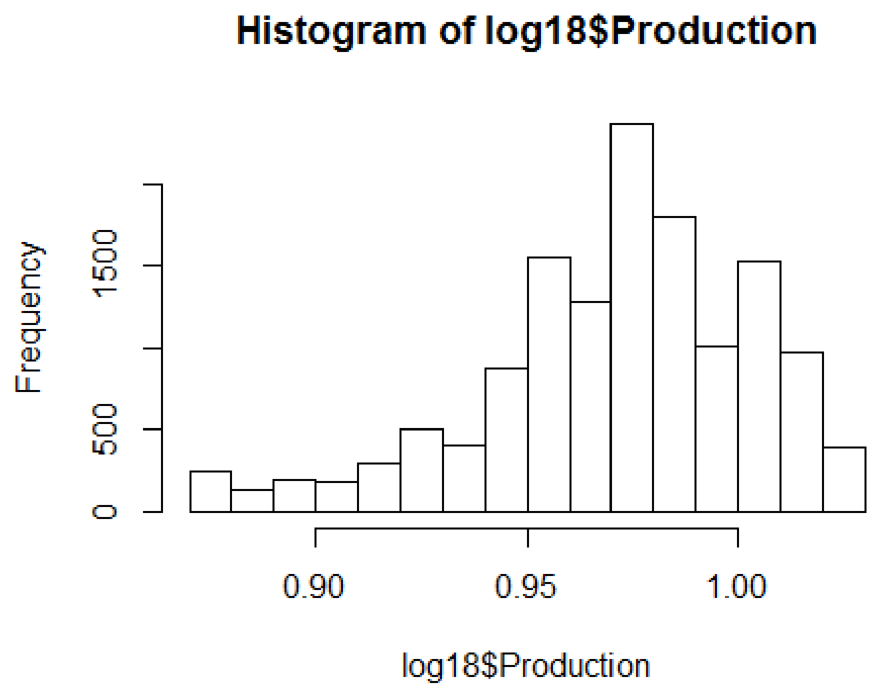
```r
# Log transformation
log1 <- log(set1[,2:10]+1)
log1 <- cbind("ID"=set1$ID,log1)
hist(log1$Production)
```



```
log14 <- log(set14[,2:10]+1)
log14 <- cbind("ID"=set14$ID,log14)
hist(log14$Production)
```



```
log18 <- log(set18[,2:10]+1)
log18 <- cbind("ID"=set18$ID,log18)
hist(log18$Production)
```

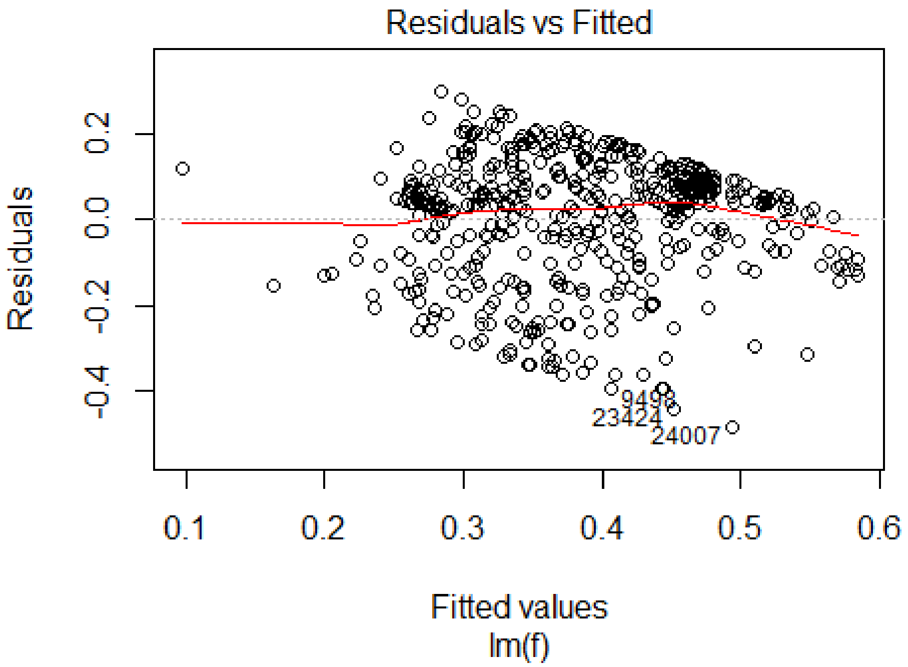


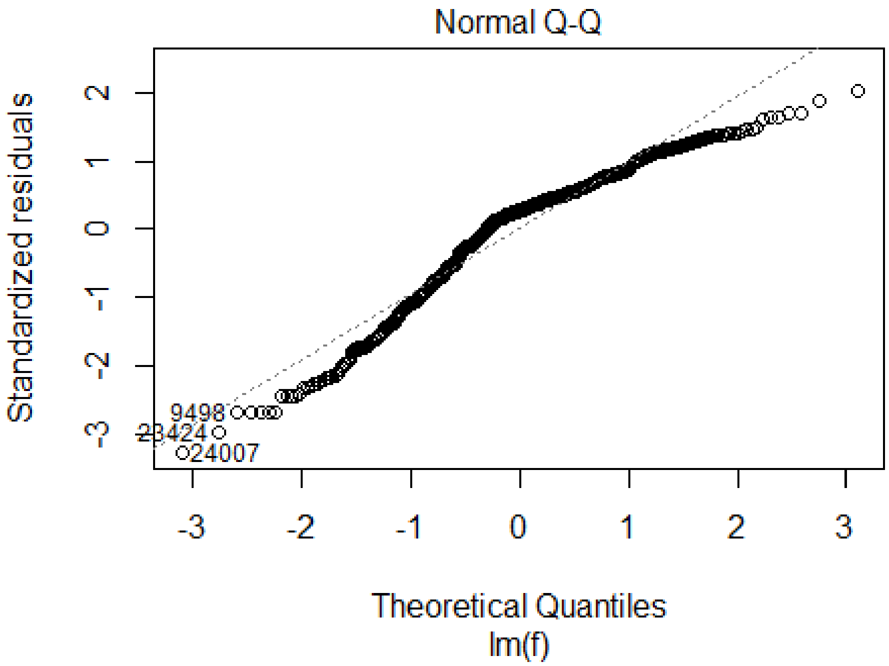
Iteration 1 Regression Models

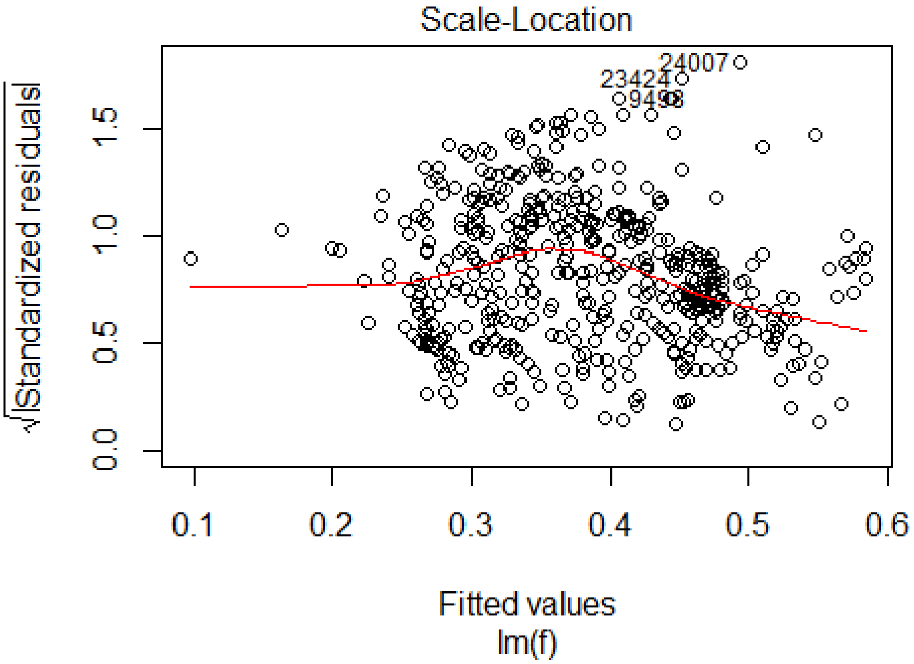
Set 1

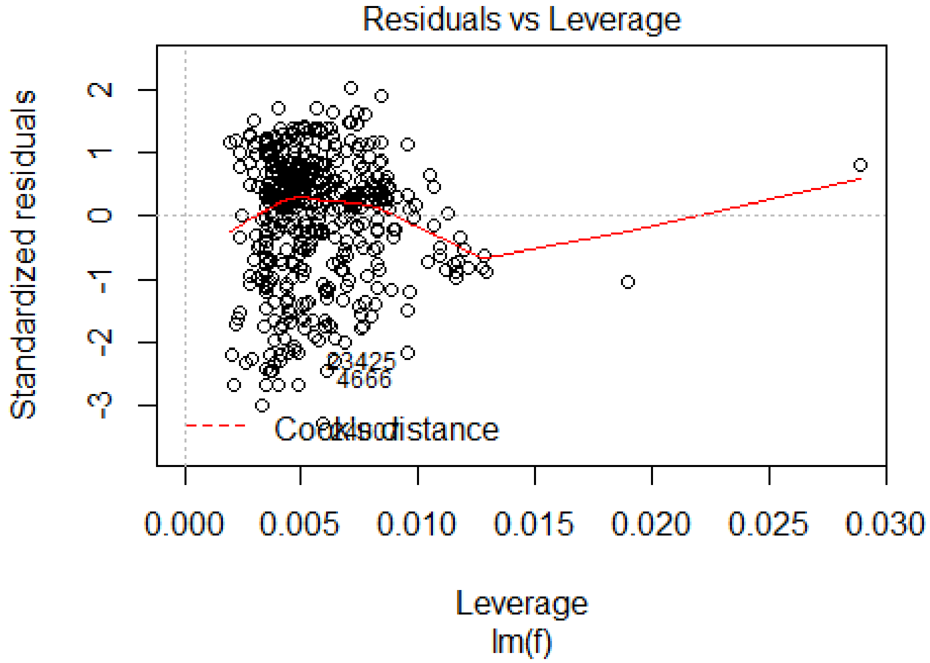
```
n <- names(log1)
# dropping several variables, shown here is the final model
f <- as.formula(paste("Production ~", paste(n[!n %in%
c("Production","WatMCon","CmpACon","SteCon","NatGCon","ID","WatWGen","WatGCon
")], collapse = " + ")))
lin1.mdl <- lm(f,data = log1)
summary(lin1.mdl)

##
## Call:
## lm(formula = f, data = log1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48282 -0.09216  0.04041  0.09982  0.29770
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.142024   0.120583  -9.471  < 2e-16 ***
## Hay_out_waste -0.020085   0.002743  -7.323 9.48e-13 ***
## EleCon         0.235634   0.018385  12.817  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1481 on 513 degrees of freedom
## Multiple R-squared:  0.2527, Adjusted R-squared:  0.2498
## F-statistic: 86.74 on 2 and 513 DF,  p-value: < 2.2e-16
```

```
plot(lin1.mdl)
```


Residuals vs Fitted
Residuals
0.2
0.0
-0.2
-0.4
9498
23424
24007
0.1
0.2
0.3
0.4
0.5
0.6
Fitted values
lm(f)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-3
9498
23424
24007
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(f)





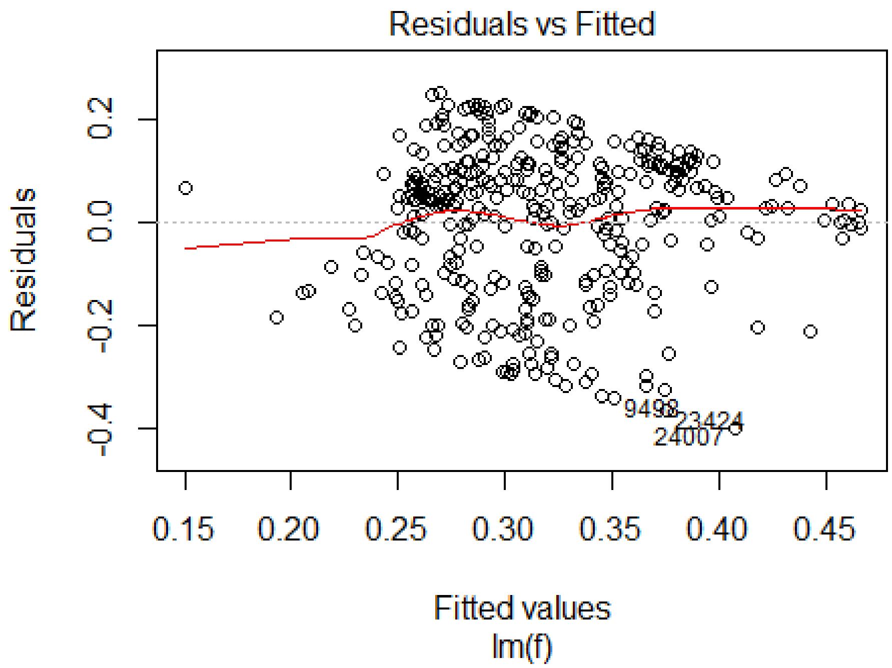
When some lower values were removed from the dataset

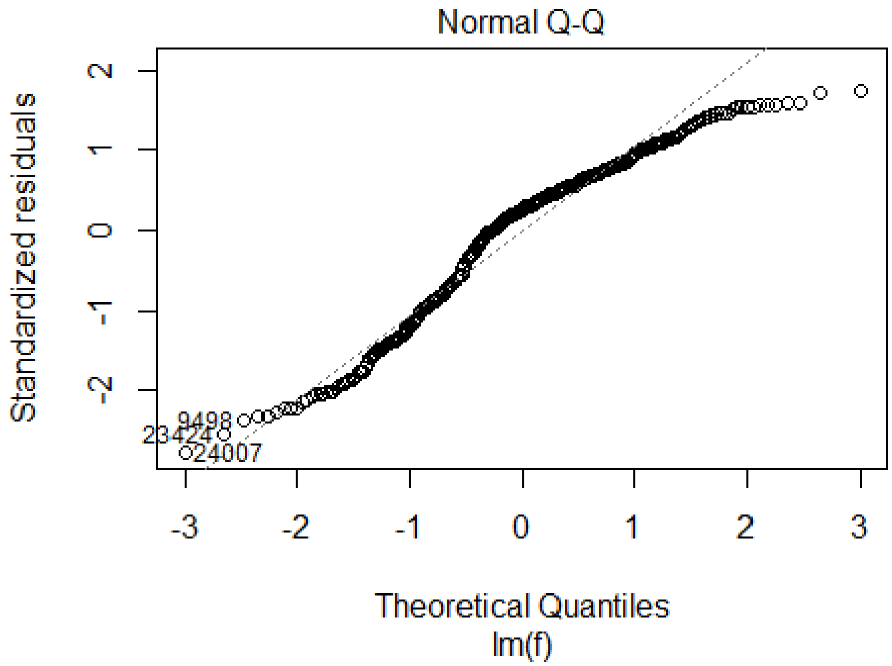
```
temp <- set1[which(set1$Production >= 0.1),]
temp <- set1[which(set1$Production < 0.7),]
```

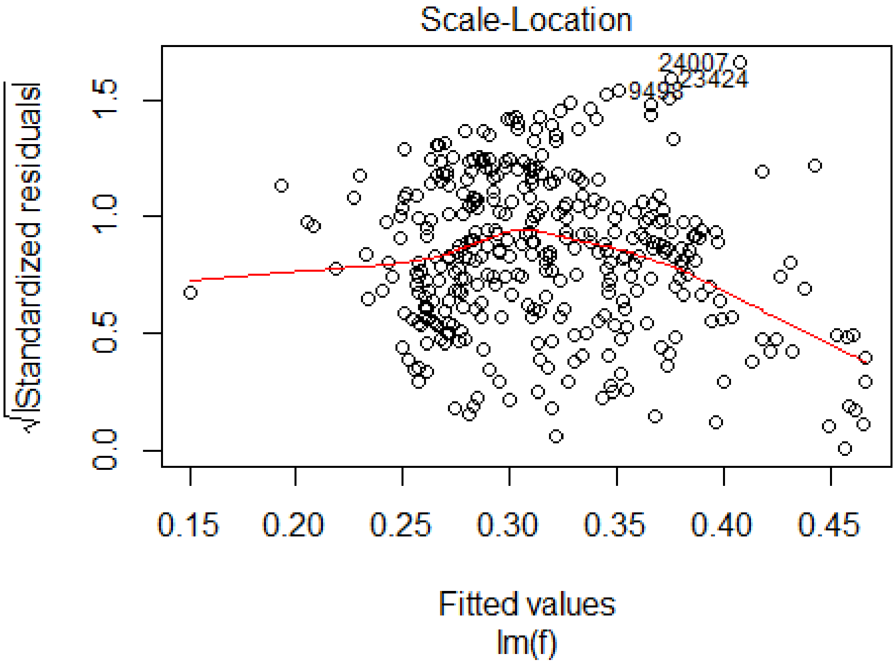
```r
log1 <- log(temp[,2:10]+1)
log1 <- cbind("ID"=temp$ID,log1)

n <- names(log1)
# dropping several variables, shown here is the final model
f <- as.formula(paste("Production ~", paste(n[!n %in%
c("Production","WatMCon","CmpACon","SteCon","NatGCon","ID","WatWGen","WatGCon
")], collapse = " + ")))
lin1_.mdl <- lm(f,data = log1)
summary(lin1_.mdl)
```

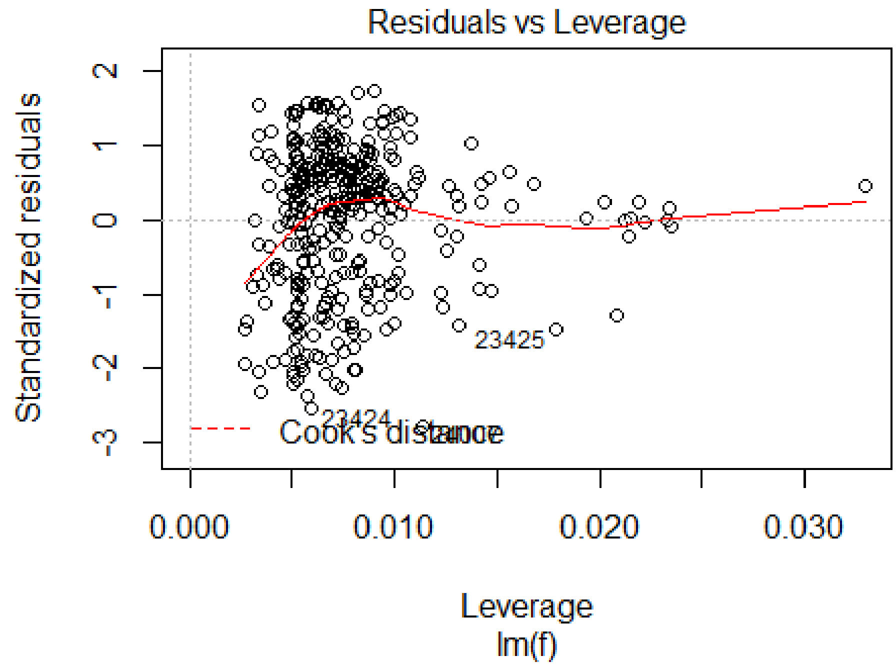
```
##
## Call:
## lm(formula = f, data = log1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39743 -0.10396  0.03673  0.10111  0.24915
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.655091   0.135671  -4.829 2.01e-06 ***
## Hay_out_waste -0.015255   0.003302  -4.620 5.30e-06 ***
## EleCon         0.153141   0.021147   7.242 2.57e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1444 on 372 degrees of freedom
## Multiple R-squared:  0.1272, Adjusted R-squared:  0.1226
## F-statistic: 27.12 on 2 and 372 DF,  p-value: 1.013e-11
```

```r
plot(lin1_.mdl)
```


Residuals vs Fitted
Residuals
0.2
0.0
-0.2
-0.4
0.15
0.20
0.25
0.30
0.35
0.40
0.45
9498
23424
24007
Fitted values
lm(f)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
9498
23424
24007
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(f)





Observations:

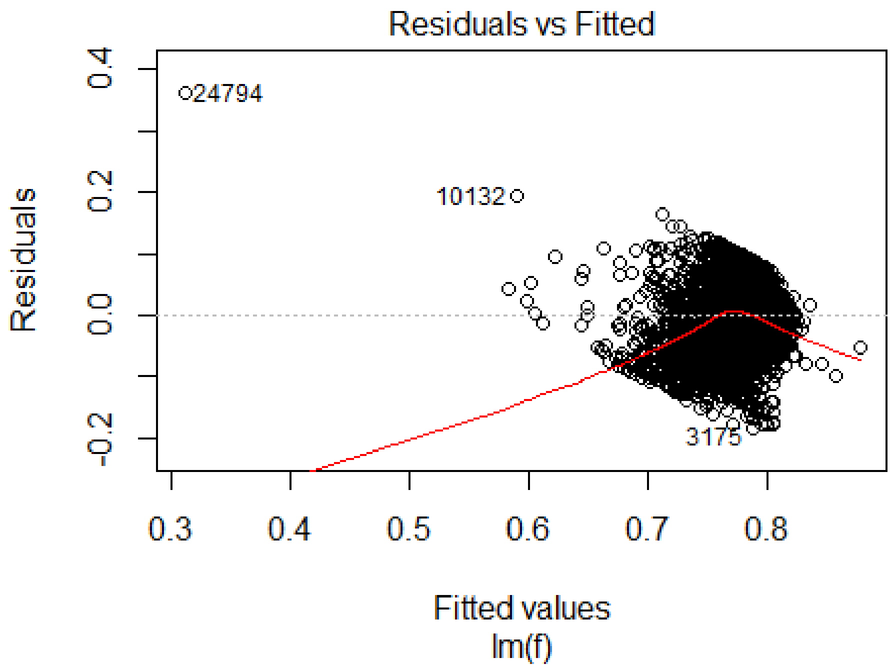
Reducing the dataset to different sizes does not give very clean residual plots. There seems to be an outlier. However, this is the least priority of this study. If required, this dataset shall be analyzed further in detail later.

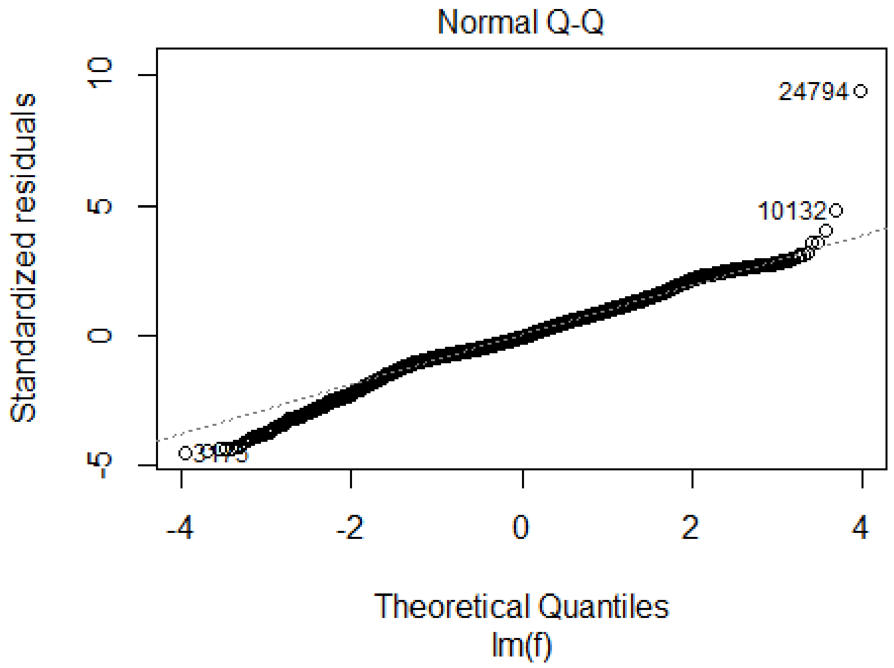
Set 14

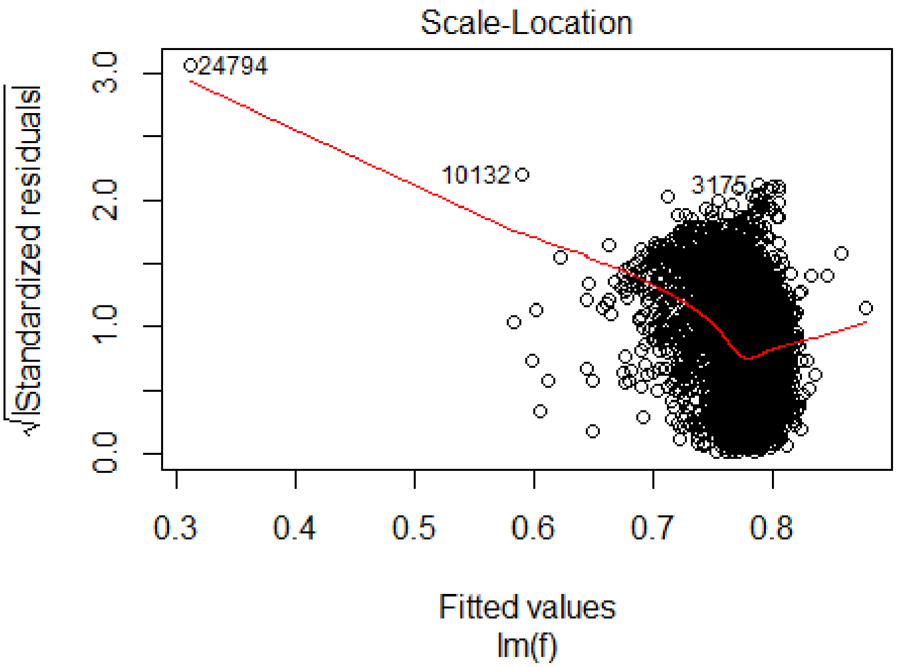
```
n <- names(log14)
# dropping WatMcon, WatWGen during iteration
f <- as.formula(paste("Production ~", paste(n[!n %in%
c("Production","WatMCon","ID","WatWGen")], collapse = " + ")))
lin14.mdl <- lm(f,data = log14)
summary(lin14.mdl)
```

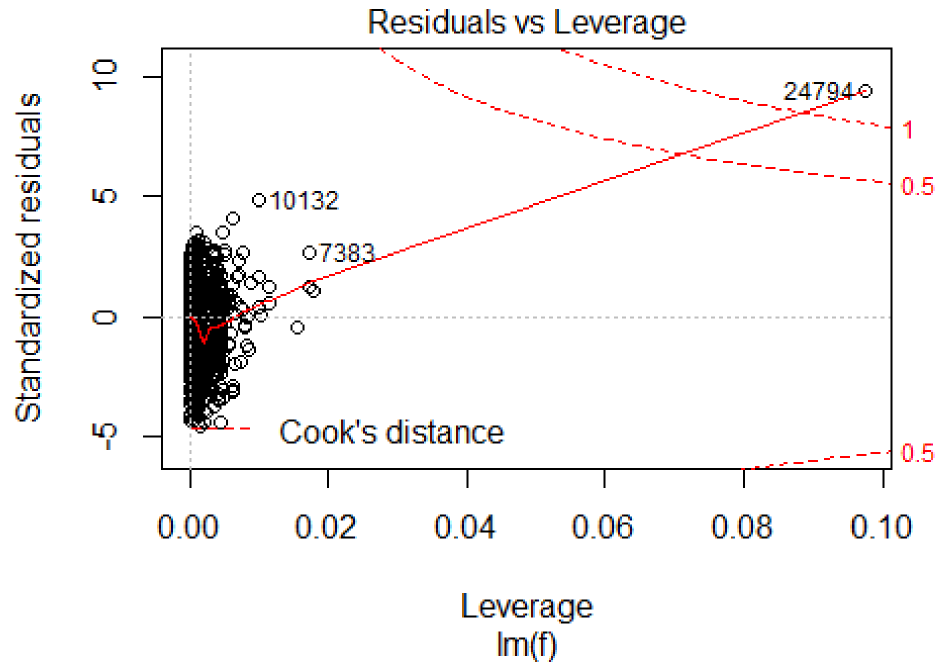
```
##
## Call:
## lm(formula = f, data = log14)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18270 -0.02500 -0.00286  0.02690  0.36114
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.0553446  0.0381582 -27.657  < 2e-16 ***
## Hay_out_waste -0.0023590  0.0001708 -13.809  < 2e-16 ***
## CmpACon        0.0591411  0.0035395  16.709  < 2e-16 ***
## EleCon         0.0967432  0.0044781  21.604  < 2e-16 ***
## NatGCon        0.1071849  0.0036908  29.041  < 2e-16 ***
## SteCon         0.0273152  0.0042442   6.436 1.27e-10 ***
## WatGCon       -0.0028682  0.0003072  -9.337  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04052 on 13717 degrees of freedom
## Multiple R-squared:  0.1731, Adjusted R-squared:  0.1728
## F-statistic: 478.6 on 6 and 13717 DF,  p-value: < 2.2e-16
```

```
plot(lin14.mdl)
```


Residuals vs Fitted
Residuals
0.4
0.2
0.0
-0.2
24794
10132
3175
0.3
0.4
0.5
0.6
0.7
0.8
Fitted values
lm(f)

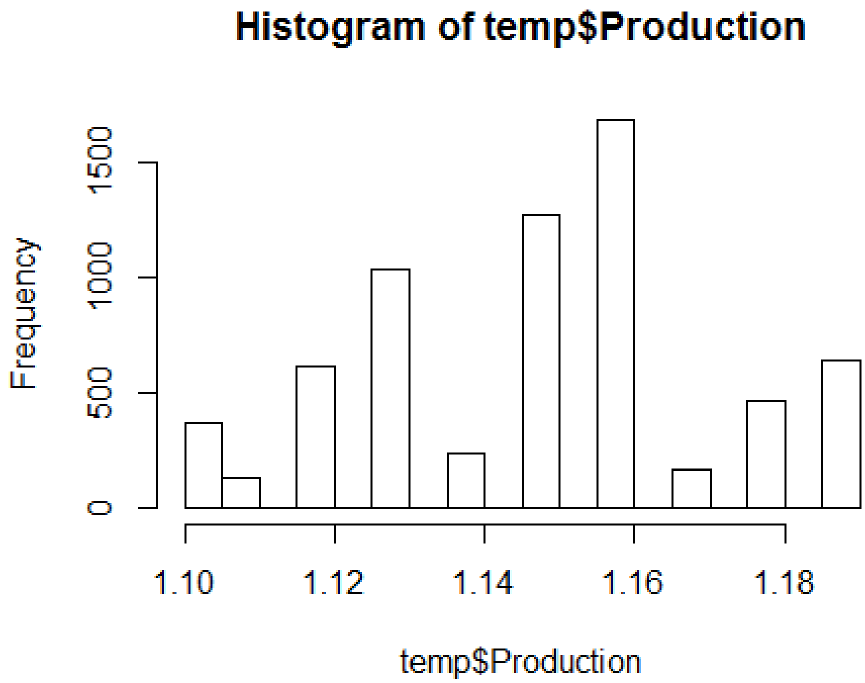

Normal Q-Q
Standardized residuals
10
5
0
-5
24794
10132
3175
-4
-2
0
2
4
Theoretical Quantiles
lm(f)





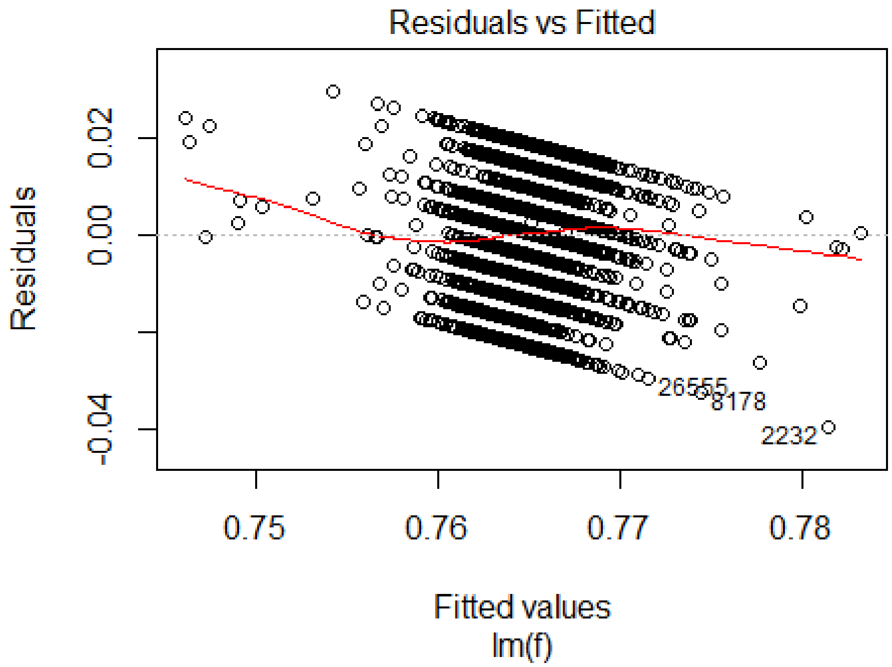
__Notes:__ Though the Q-Q plot indicates normality, the other residual plots indicate that the datapoinst are too many or there are influence of outliers in the model. Earlier, it was noted for set14 dataset that the production values are predominantly maintained between 1.1 to 1.2. A closer look in to this section of data might give some insights.

```
# chosing section between 1.1 to 1.2
temp <- set14[which(set14$Production >= 1.1),]
temp <- temp[which(temp$Production < 1.2),]
hist(temp$Production)
```

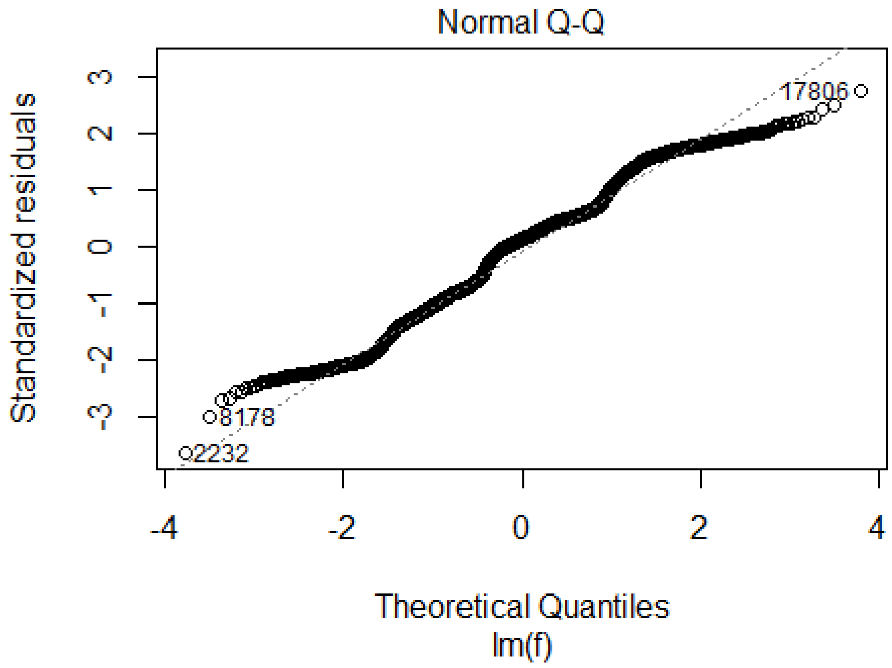


```
log14_ <- log(temp[,2:10]+1)
n <- names(log14_)
# dropping WatMcon, WatWGen during iteration
f <- as.formula(paste("Production ~", paste(n[!n %in%
c("Production","WatMCon","ID","WatWGen","CmpACon","NatGCon","Hay_out_waste")]
, collapse = " + ")))

lin14_.mdl <- lm(f,data = log14_)
summary(lin14_.mdl)

##
## Call:
## lm(formula = f, data = log14_)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.039591 -0.008407  0.001379  0.006244  0.029678
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.7502285  0.0157513  47.630  < 2e-16 ***
## EleCon          0.0246120  0.0021728  11.327  < 2e-16 ***
## SteCon      -0.0222077  0.0017221 -12.896  < 2e-16 ***
## WatGCon     -0.0007141  0.0001382  -5.166 2.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01096 on 6611 degrees of freedom
## Multiple R-squared:  0.03706,    Adjusted R-squared:  0.03663
## F-statistic: 84.82 on 3 and 6611 DF,  p-value: < 2.2e-16
```

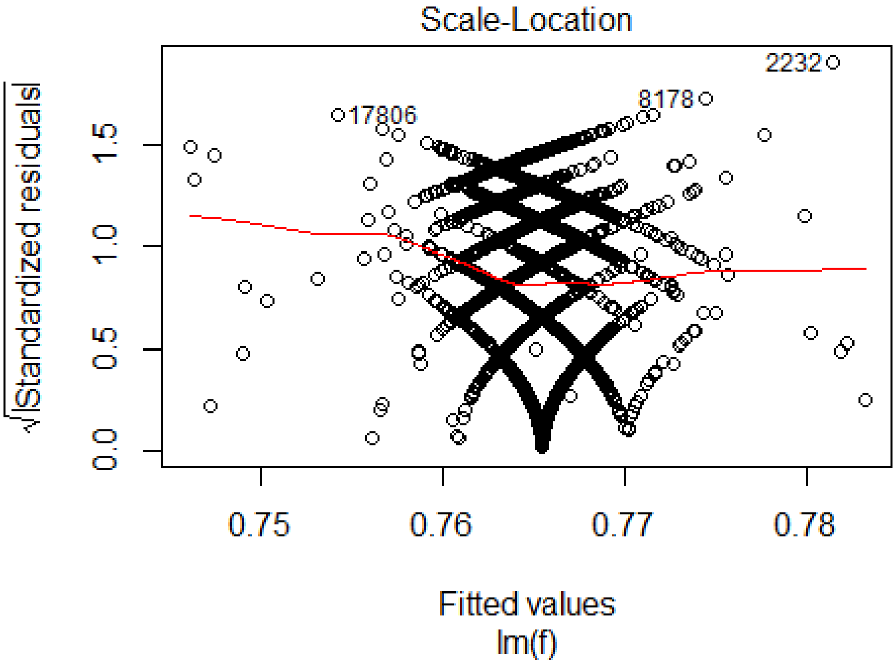
```
plot(lin14_.mdl)
```

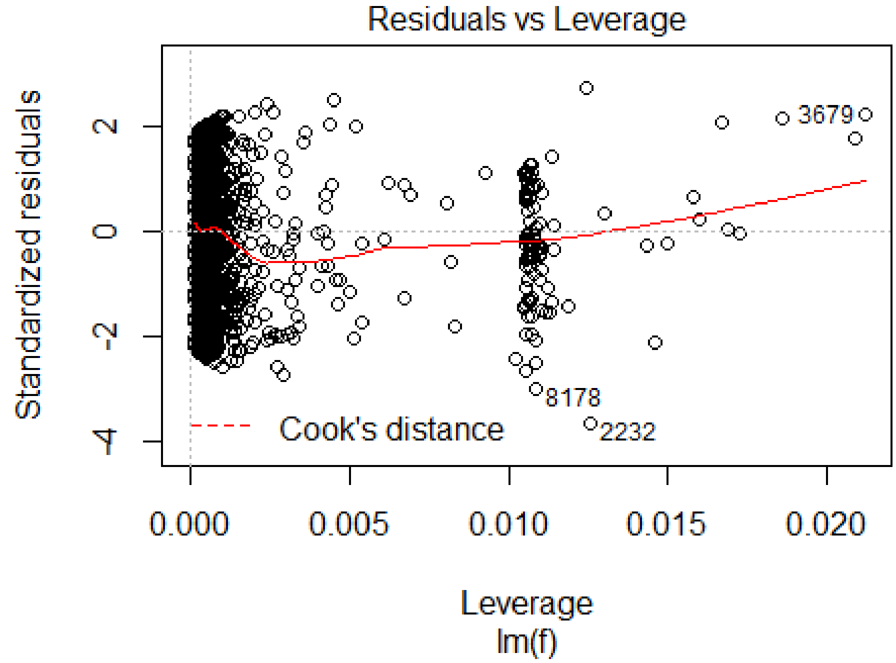


Residuals vs Fitted



Normal Q-Q



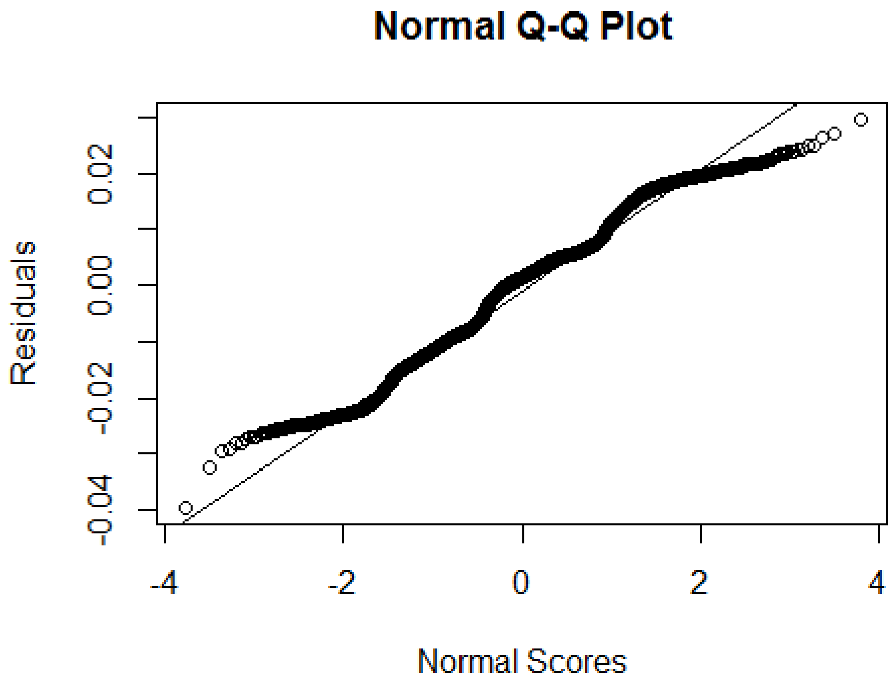


__Notes:__

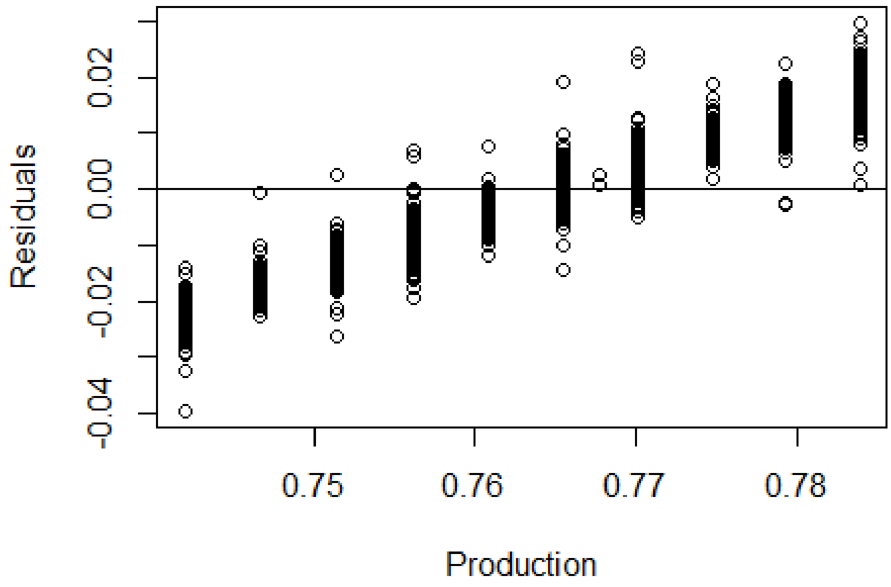
- The regression model reduced to three predictors, the Adjusted R-square is very low, given the number of degrees of freedom

- The histogram indicates that there are gaps in between the response variable from 1.1 to 1.2 MT, this is also reflecting the residual plots, in diagonal patterns.The residuals follow a specific pattern to each of the response values. Each strip of those diagonal lines are the residuals for each specific Production value. A more deep dive is required
- The normal plot indicates that the residuals are deviating from normal at its ends. A smaller dataset (in 100s rather than 1000s) need to be taken for analysis.

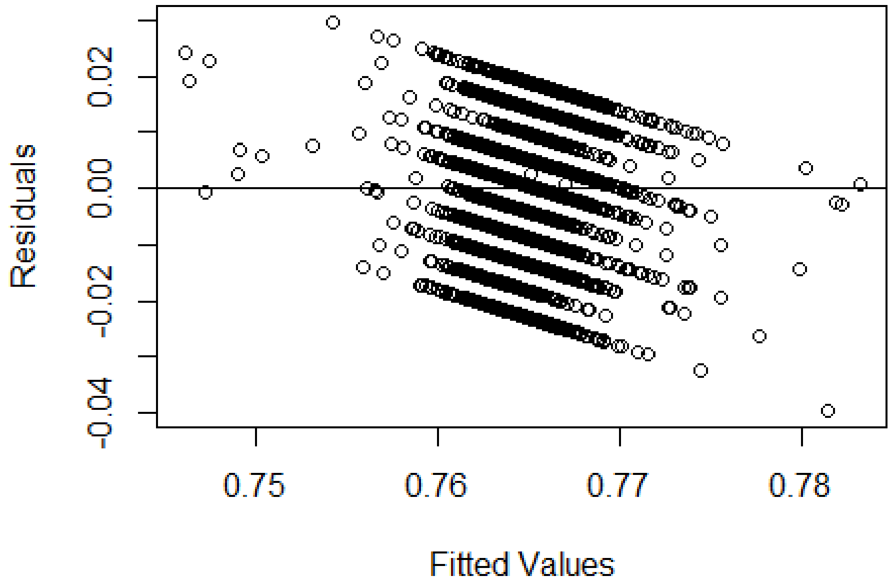
```
res <- resid(lin14_.mdl)
fittedY <- fitted.values(lin14_.mdl)
qqnorm(res,ylab = "Residuals",xlab = "Normal Scores")
qqline(res)
```



```
plot(log14_$Production,res,ylab = "Residuals",xlab = "Production")
abline(0,0)
```



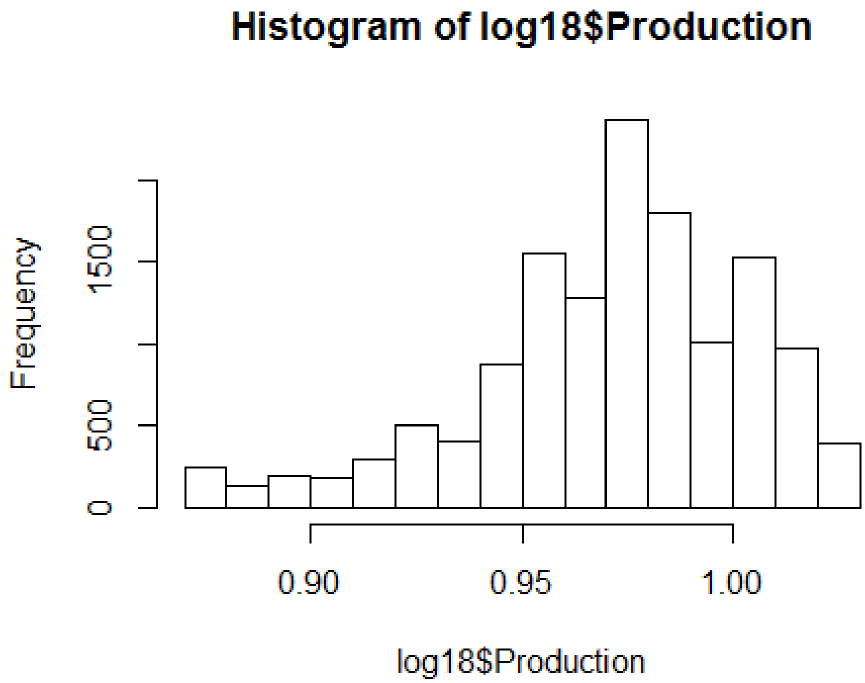
```
plot(fittedY,res,ylab="Residuals",xlab="Fitted Values")
abline(0,0)
```



__Notes:__ The residuals vs Actuals violate the assumption of independence. The plots have a moving mean & moving variance, may be this is because the captured values are in a time order?. Nevertheless, there seems to be an autocorrelation within the residuals itself, which indicate further transformation or pruning of dataset is necessary.

Set 18

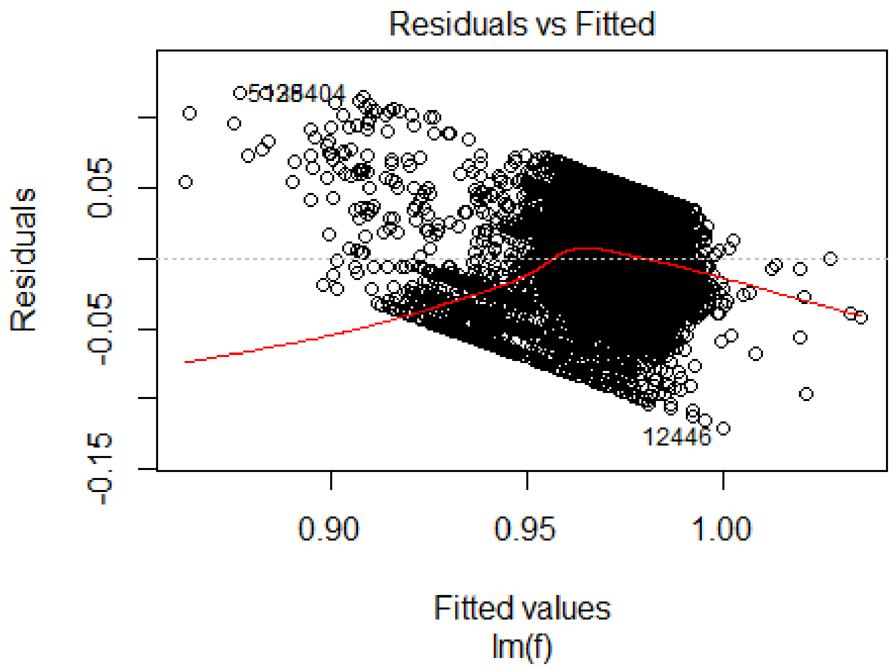
```
n <- names(log18)
hist(log18$Production)
```

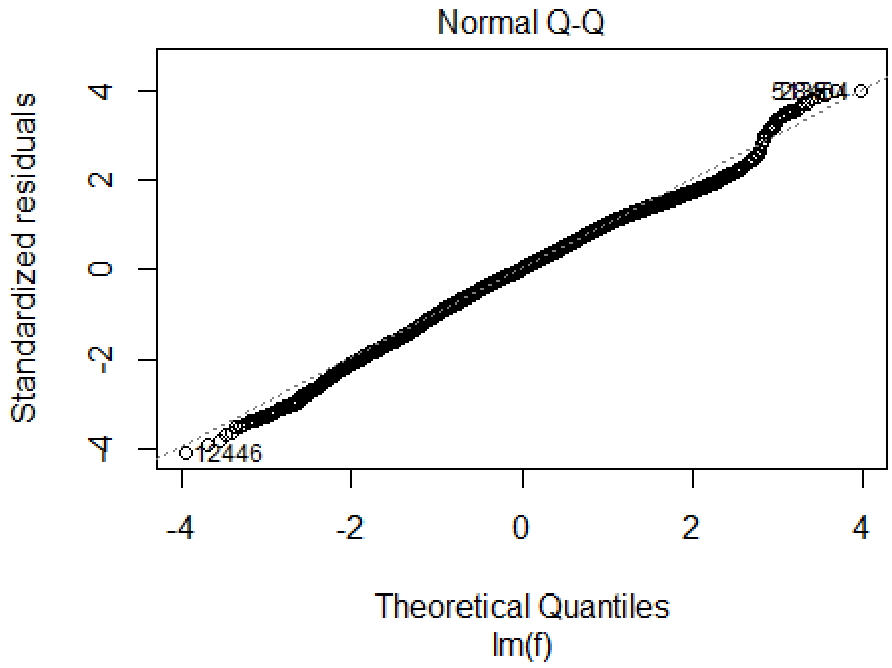


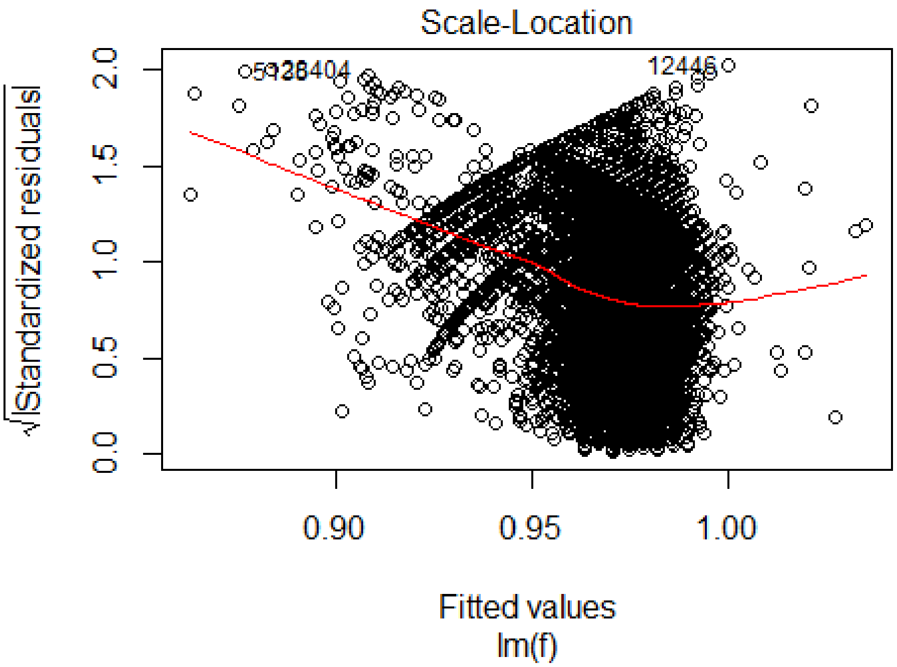
```
# dropping WatMcon, WatWGen during iteration
f <- as.formula(paste("Production ~", paste(n[!n %in%
c("Production","WatMCon","ID","Hay_out_waste","WatWGen")], collapse = " +
")))
lin18.mdl <- lm(f,data = log18)
summary(lin18.mdl)
```

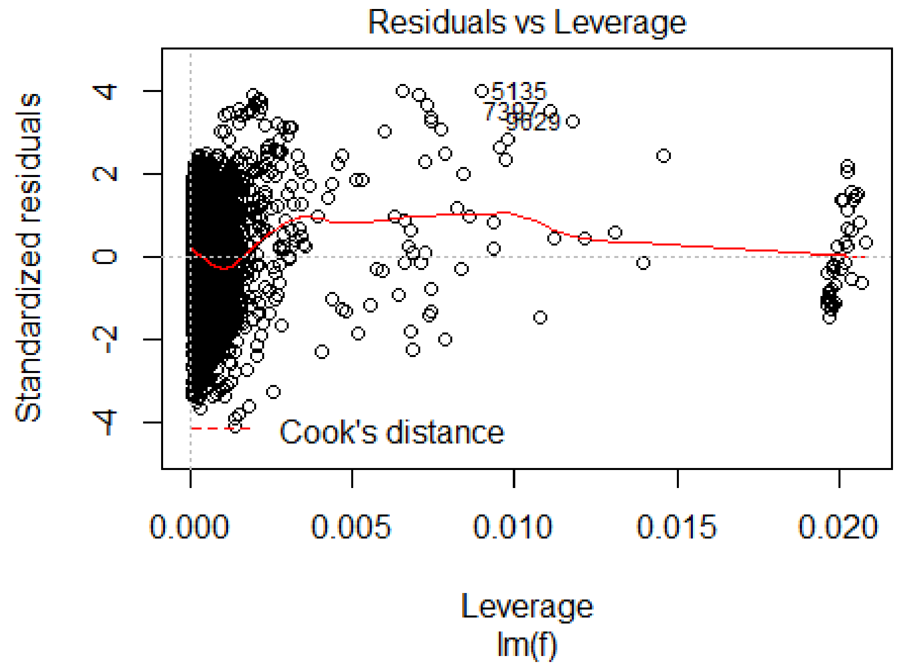
```
##
## Call:
## lm(formula = f, data = log18)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.120427 -0.018660  0.000804  0.020971  0.117093
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.3832672  0.0417367  -9.183  < 2e-16 ***
## CmpACon     -0.0433668  0.0029709 -14.597  < 2e-16 ***
## EleCon       0.1367469  0.0043487  31.446  < 2e-16 ***
## NatGCon     -0.0174304  0.0036491  -4.777  1.8e-06 ***
## SteCon       0.0901014  0.0037048  24.320  < 2e-16 ***
## WatGCon      0.0012378  0.0005012   2.470   0.0135 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02937 on 13749 degrees of freedom
## Multiple R-squared:  0.1687, Adjusted R-squared:  0.1684
## F-statistic:   558 on 5 and 13749 DF,  p-value: < 2.2e-16
```

```
plot(lin18.mdl)
```


Residuals vs Fitted
Residuals
0.05
-0.05
-0.15
0.90
0.95
1.00
Fitted values
lm(f)
5128404
12446


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
-4
-2
0
2
4
Theoretical Quantiles
lm(f)
528404
12446





__Notes:__ Few iterations of reducing the dataset by constraining the Production value ranges indicate that there is a possibility of fitting number of models using more detailed level of datasets. However, this is noted as an observation for further reference.There were no gaps found in the values of production at high capacity (above 1.4 MT), but the pattern of residuals indicate the autocorrelation of residuals caused by the time influence.

Conclusions

It has become clear that not just the dataset needs to be sliced further for patterns, or the dataset need to be associated with machine level attributes (the original dataset had site level & machine level) and different methods (such as temporal study)need to be applied to specifically deal with seasonality and trends.