**Task 4 Solution:**

**Assignment3\_Q2 <- read.csv("C:\\Class\\Semester 2\\Data Mining in Engineering\\Assignments\\Assignment 3\\Forest Fires Data.csv")**

To read the Forest Fires data.

**str(Assignment3\_Q2)**

To get structure of the data in Forest fire data.

From the structure we find that the mean is 12.85 and median is 0.52 hence its heavily skewed towards the right.

**Console:**

'data.frame': 517 obs. of 13 variables:

$ ï..X :int 7 7 7 8 8 8 8 8 8 7 ...

$ Y :int 5 4 4 6 6 6 6 6 6 5 ...

$ Month: Factor w/ 12 levels "apr","aug","dec",..: 8 11 11 8 8 2 2 2 12 12 ...

$ Day : Factor w/ 7 levels "fri","mon","sat",..: 1 6 3 1 4 4 2 2 6 3 ...

$ FFMC :num 86.2 90.6 90.6 91.7 89.3 92.3 92.3 91.5 91 92.5 ...

$ DMC :num 26.2 35.4 43.7 33.3 51.3 ...

$ DC :num 94.3 669.1 686.9 77.5 102.2 ...

$ ISI :num 5.1 6.7 6.7 9 9.6 14.7 8.5 10.7 7 7.1 ...

$ Temp :num 8.2 18 14.6 8.3 11.4 22.2 24.1 8 13.1 22.8 ...

$ RH :int 51 33 33 97 99 29 27 86 63 40 ...

$ Wind :num 6.7 0.9 1.3 4 1.8 5.4 3.1 2.2 5.4 4 ...

$ Rain :num 0 0 0 0.2 0 0 0 0 0 0 ...

$ Area :num 0 0 0 0 0 0 0 0 0 0 ...

**Assignment3\_Q2$logarea <- log(Assignment3\_Q2$Area+1)**

We are transforming area variable into log(area)

**install.packages("car")**

To install car package.

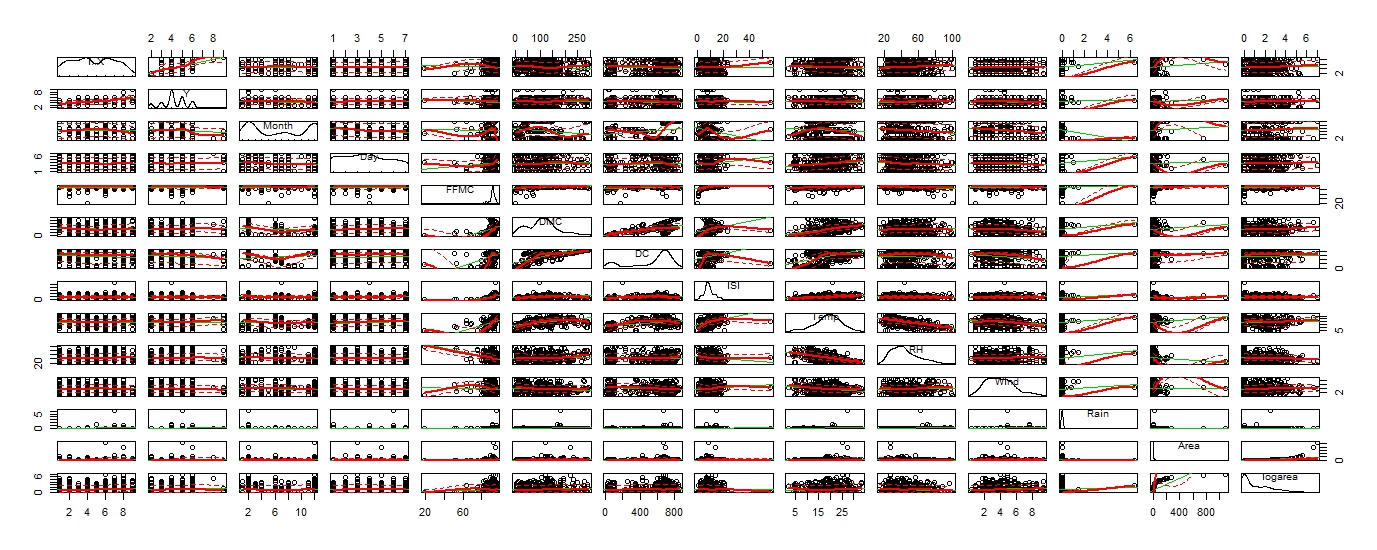
**require(car)**

To load car package.

**scatterplotMatrix(Assignment3\_Q2)**

To construct a scatter plot matrix.

**Graph:**



**MODEL 1**

**Model1 <- lm(logarea~ISI+FFMC+DC,data = Assignment3\_Q2)**

After seeing the scatter plot we are selecting ISI, FFMC and DC as the initial predictors.

**summary(Model1)**

From the summary we can find out R^2 is 0.0013.

**Console:**

Call:

lm(formula = logarea ~ ISI + FFMC + DC, data = Assignment3\_Q2)

Residuals:

Min 1Q Median 3Q Max

-1.2579 -1.1245 -0.6536 0.9159 5.8086

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.1984780 1.1386638 -0.174 0.862

ISI -0.0164462 0.0159686 -1.030 0.304

FFMC 0.0140269 0.0136032 1.031 0.303

DC 0.0003402 0.0002634 1.292 0.197

Residual standard error: 1.397 on 513 degrees of freedom

Multiple R-squared: 0.007151, Adjusted R-squared: 0.001344

F-statistic: 1.232 on 3 and 513 DF, p-value: 0.2975

**dpar<- par(no.readonly = T)**

Defining a default graphic parameter to retain a single graph in the plot window.

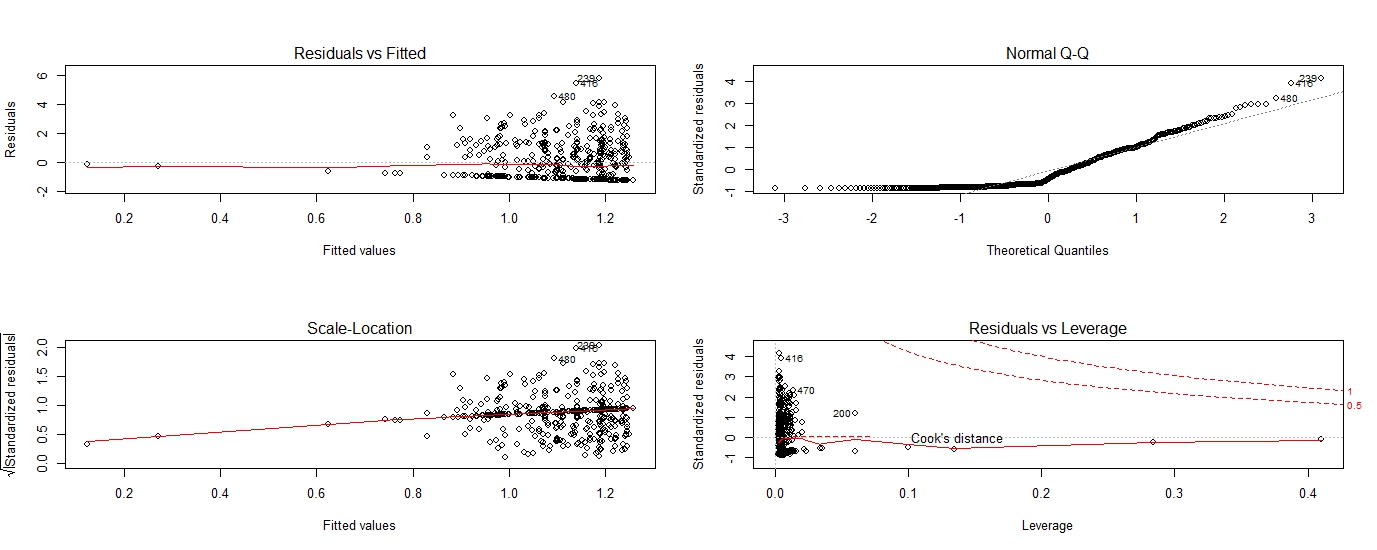
**par(mfrow = c(2,2))**

To generate 2X2 plots.

**plot(Model1)**

Model is plotted, and we can find out that all assumptions are violated.

**Graph:**



**par(dpar)**

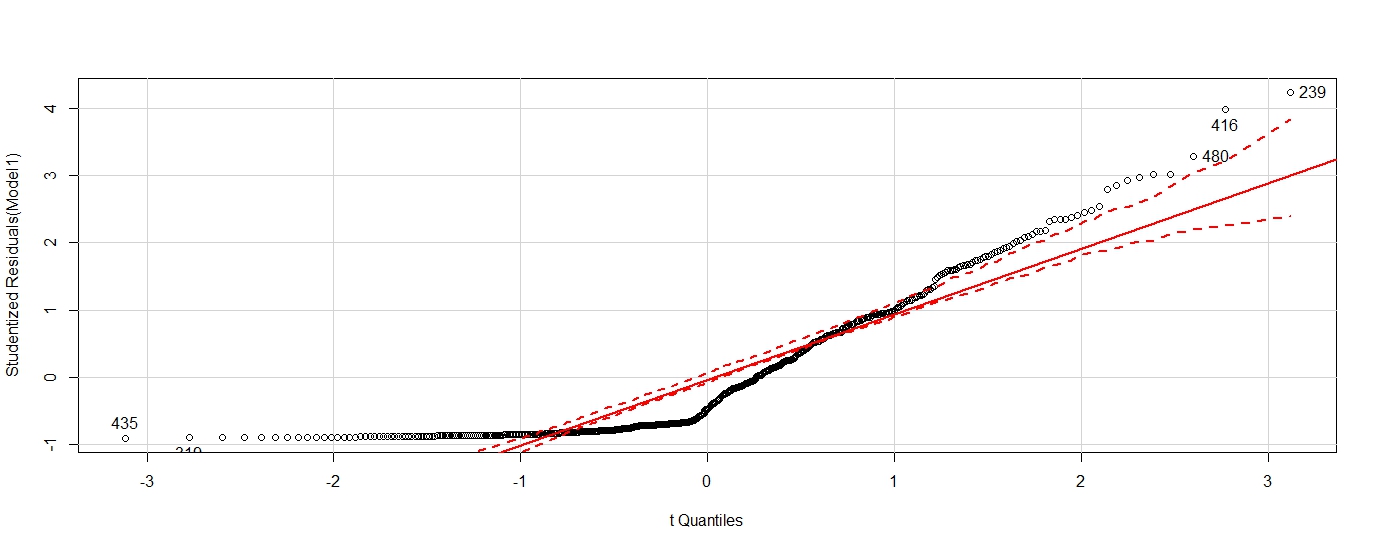
For having the default parameter for viewing 1 graph in the plot window.

**Enhanced approach**

**qqPlot(Model1,simulate = T,id.method = "identify", labels = rownames(Assignment3\_Q2))**

The data lacks normality.

**Graph:**



**durbinWatsonTest(Model1)**

Durbin watson test is done in order to know the independent assumptions of the model since the p value is non-significant hence its independent.

**Console:** lag Autocorrelation D-W Statistic p-value

1 0.5395845 0.9190409 0

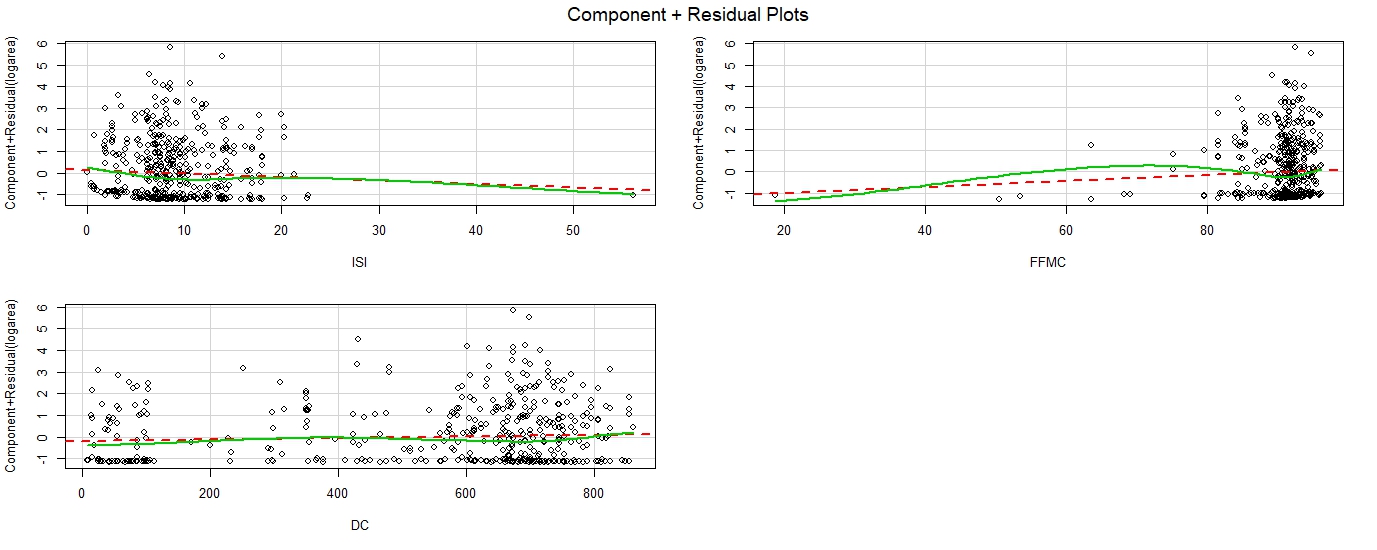
Alternative hypothesis: rho != 0

**crPlots(Model1)**

coplot is done to determine linearity.

From the plot we can find that DC is somewhat linear whereas, FFMC and ISI are not linear.

**Graph:**



**ncvTest(Model1)**

P value is significant, so homoscedasticity is not satisfied.

**Console:**

Non-constant Variance Score Test

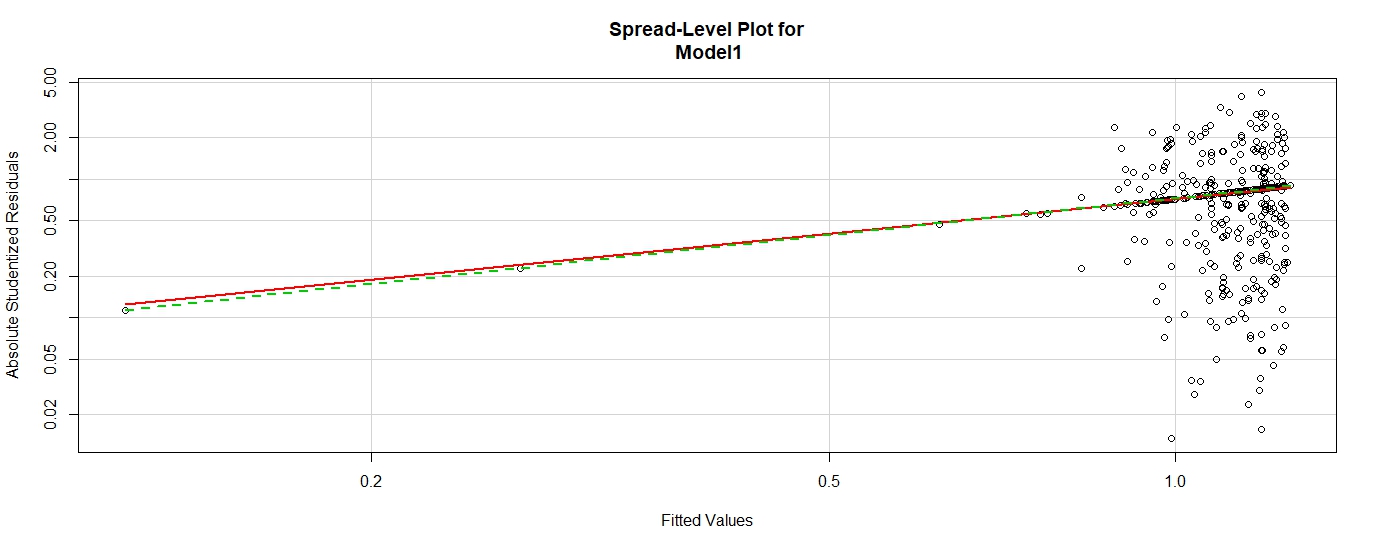
Variance formula: ~ fitted.values

Chisquare = 4.069081 Df = 1 p = 0.04367504

**spreadLevelPlot(Model1)**

The p value is significant and the suggested power transformation for getting a better homoscedasticity is 0.165.

**Graph:**



**Console:**

Suggested power transformation: 0.1657714

**install.packages("gvlma")**

To install gvlma package.

**require(gvlma)**

To load gvlma package.

**summary(gvlma(Model1)**

From the global statistics we can find that 1 out of the 5 assumptions are satisfied.

**Console:**

Call:

lm(formula = logarea ~ ISI + FFMC + DC, data = Assignment3\_Q2)

Residuals:

Min 1Q Median 3Q Max

-1.2579 -1.1245 -0.6536 0.9159 5.8086

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.1984780 1.1386638 -0.174 0.862

ISI -0.0164462 0.0159686 -1.030 0.304

FFMC 0.0140269 0.0136032 1.031 0.303

DC 0.0003402 0.0002634 1.292 0.197

Residual standard error: 1.397 on 513 degrees of freedom

Multiple R-squared: 0.007151, Adjusted R-squared: 0.001344

F-statistic: 1.232 on 3 and 513 DF, p-value: 0.2975

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:

Level of Significance = 0.05

Call:

gvlma(x = Model1)

Value p-value Decision

Global Stat 147.9694 0.000e+00 Assumptions NOT satisfied!

Skewness 123.7061 0.000e+00 Assumptions NOT satisfied!

Kurtosis 16.5351 4.776e-05 Assumptions NOT satisfied!

Link Function 0.2796 5.970e-01 Assumptions acceptable.

Heteroscedasticity 7.4486 6.349e-03 Assumptions NOT satisfied!

**vif(Model1)**

We determine the variance inflation factor for checking multi collinearity error.

**Console:**

ISI FFMC DC

1.400592 1.489786 1.127665

**sqrt(vif(Model1))>2**

Since the square root values are less than 2 it suggests that there is no multi collinearity error.

**Console:**

ISI FFMC DC

FALSE FALSEFALSE

**outlierTest(Model1)**

From the Outlier test we find that the 239th and 416th records are outliers.

**Console:**

rstudent unadjusted p-value Bonferonni p

239 4.230119 2.7675e-05 0.014308

416 3.983778 7.7666e-05 0.040153

**fitted(Model1)[c(238:240,415:417)]**

To see the predicted values for record 239, 416.

**Console:**

238 239 240 415 416 417

1.1984714 1.1870064 0.8954544 1.1943344 1.1403364 1.1513344

**Model1$residuals[c(238:240,415:417)]**

We inferred from the residual error that 239th and 416th has a high residual error.

Since the value is positive we can infer that our model under predicted the actual value of the 239th and 416th record by 5.808 and 5.476 respectively.

**Console:**

238 239 240 415 416 417

4.1669437 5.8086133 -0.8954544 -1.1943344 5.4761035 0.9306040

**Hat.plot<- function(X)**

**{**

**p <- length(coefficients(X))**

**N <- length(fitted(X))**

**plot(hatvalues(X), main = "Plot of Hat Values")**

**abline(h=c(2,3)\*(p/N),col = "red")**

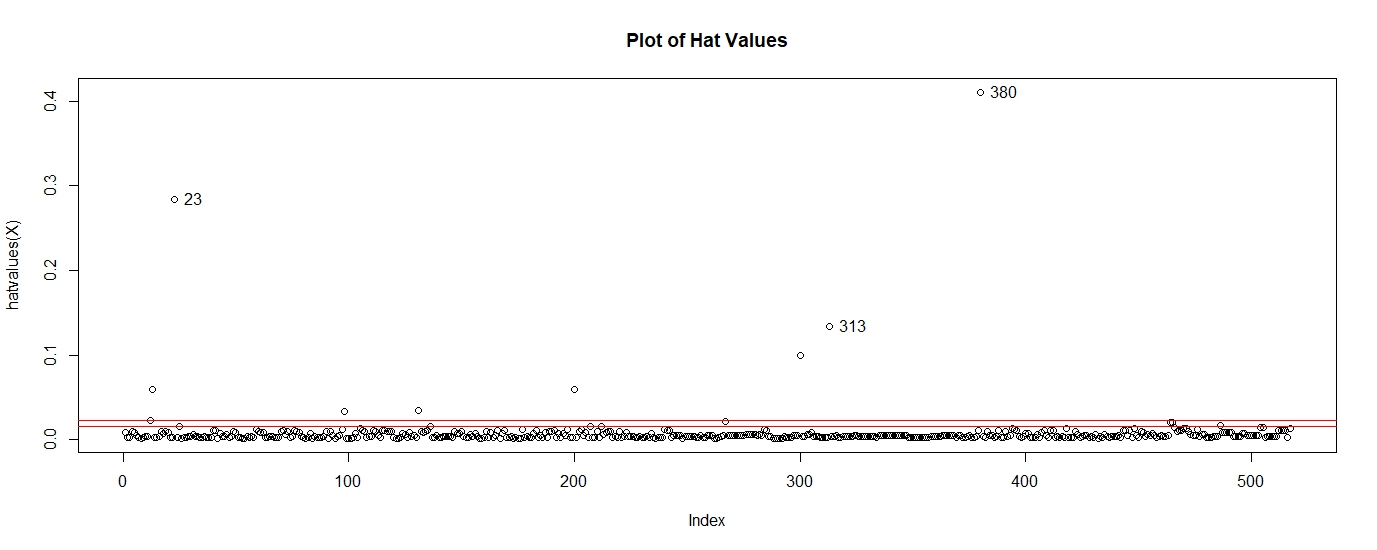
**identify(1:N, hatvalues(X), names(hatvalues(X)))**

**}**

**Hat.plot(Model1)**

From the hat plot we find that there are many points which have high leverage.

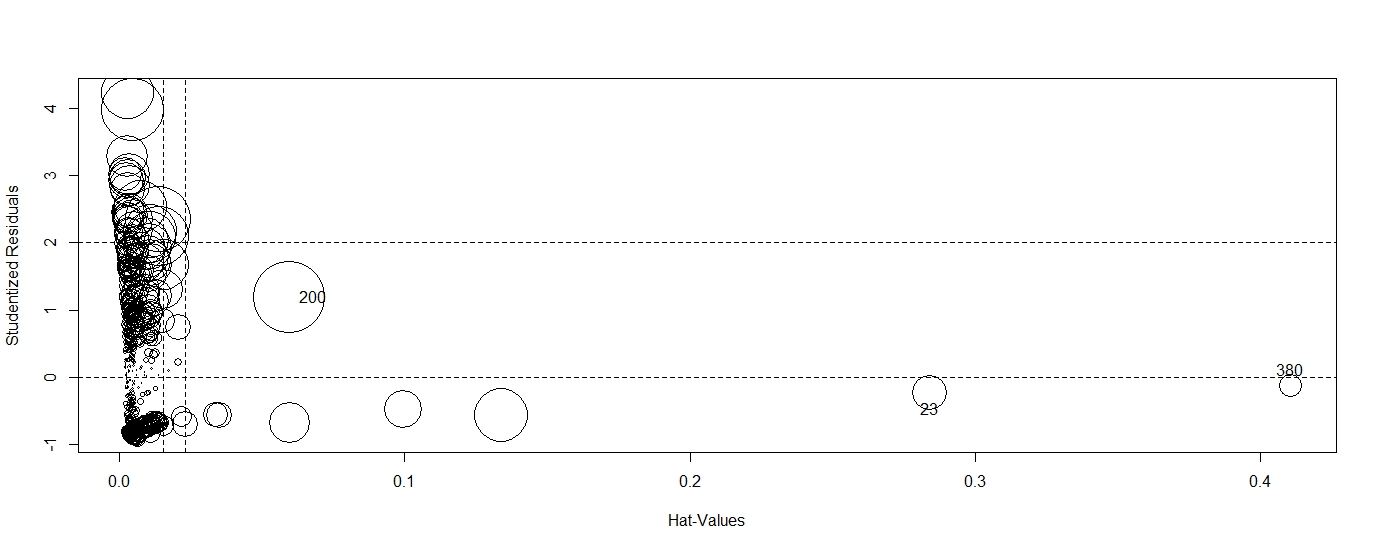
**Graph:**

**influencePlot(Model1, id.method = "identify")**

We find that the circles with bigger radius have larger cooks distance hence are more influential so removing those will alter the model.

In this case 200 is highly influential.

**Graph:**

**boxTidwell(logarea~DC,data = Assignment3\_Q2)**

To correct linearity assumption, non-significant p value suggests that it’s not required.

**Console:**

Score Statistic p-value MLE of lambda

-0.5010811 0.6163141 0.2108823

iterations = 7

**MODEL 2**

**Assignment3\_Q2.2 <- Assignment3\_Q2[-c(200, 239, 416),]**

To remove 200th, 239th and 416th records.

**Model2 <- lm(logarea~ISI+FFMC+DC,data = Assignment3\_Q2.2)**

After seeing the scatter plot we are selecting DC, FFMC and ISI as the initial predictors.

**summary(Model2)**

From the summary we can find out R^2 is 0.0013

**Console:**

Call:

lm(formula = logarea ~ ISI + FFMC + DC, data = Assignment3\_Q2.2)

Residuals:

Min 1Q Median 3Q Max

-1.2212 -1.0958 -0.6220 0.9091 4.5552

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.4029509 1.1333900 -0.356 0.722

ISI -0.0182383 0.0154935 -1.177 0.240

FFMC 0.0165874 0.0135267 1.226 0.221

DC 0.0002733 0.0002565 1.065 0.287

Residual standard error: 1.354 on 510 degrees of freedom

Multiple R-squared: 0.007212, Adjusted R-squared: 0.001372

F-statistic: 1.235 on 3 and 510 DF, p-value: 0.2963

**dpar<- par(no.readonly = T)**

Defining a default graphic parameter to retain a single graph in the plot window.

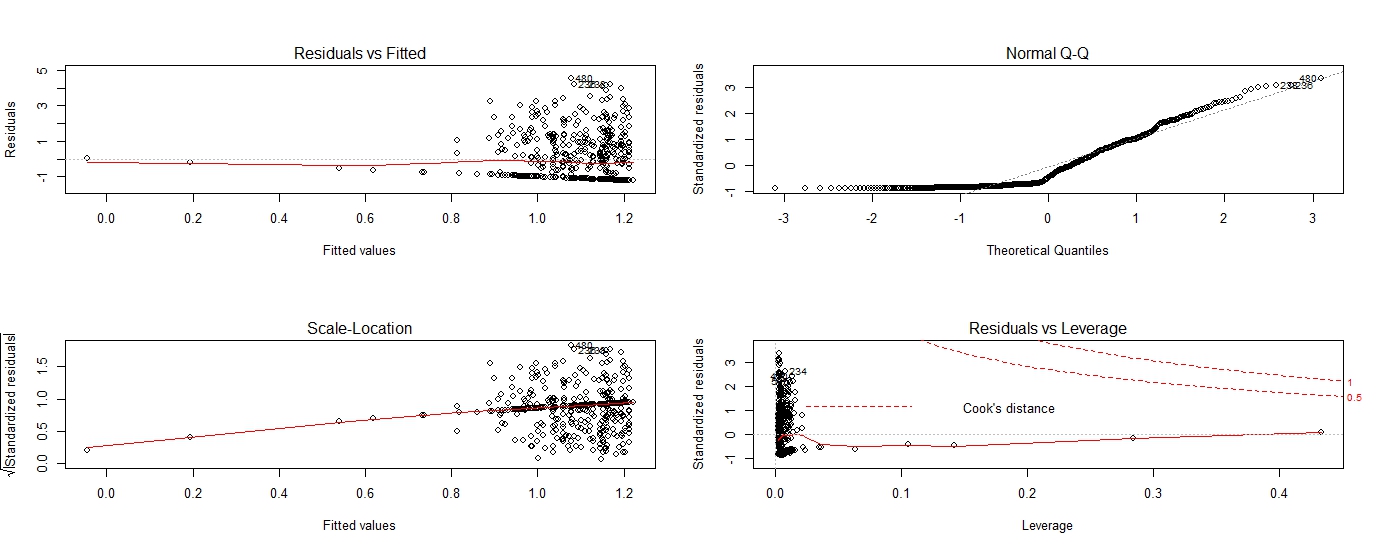
**par(mfrow = c(2,2))**

To generate 2X2 plots.

**plot(Model2)**

The data still lacks normality.

**Graph:**



**par(dpar)**

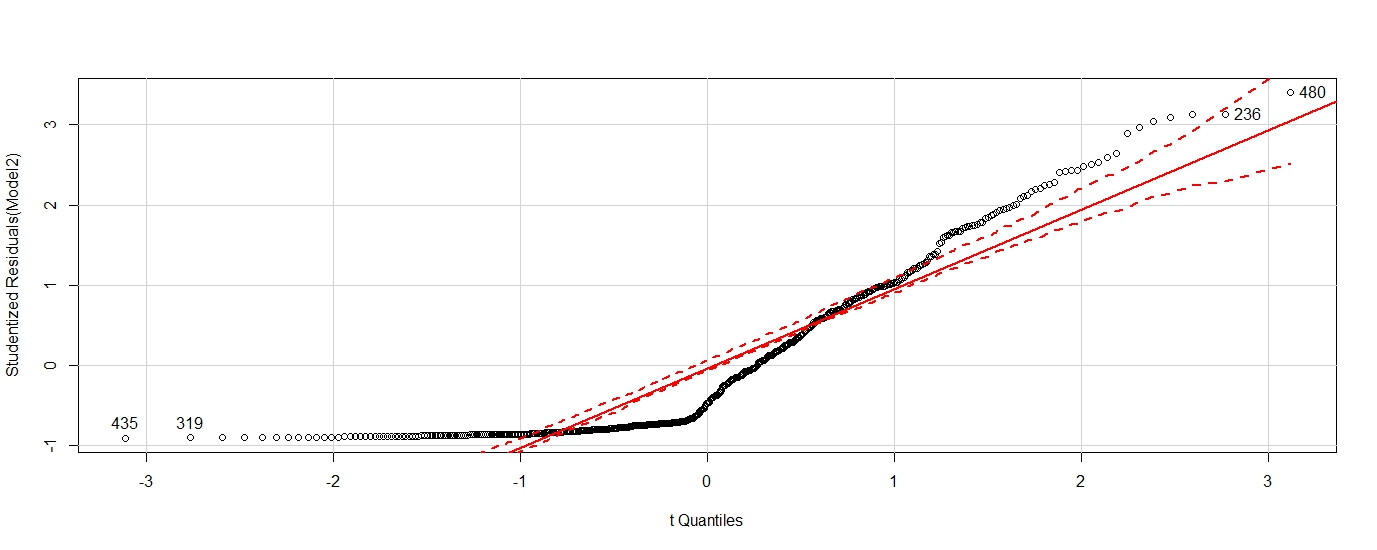
For having the default parameter for viewing 1 graph in the plot window.

**Enhanced approach**

**qqPlot(Model2,simulate = T,id.method = "identify", labels = rownames(Assignment3\_Q2.2))**

Normality is not perfect because several points are deviating from the normal curve.

**Graph:**



**durbinWatsonTest(Model2)**

Durbin watson test is done to know the independent assumptions of the model since the p value is non-significant hence its independent.

**Console:**

lag Autocorrelation D-W Statistic p-value

1 0.5505988 0.8969028 0

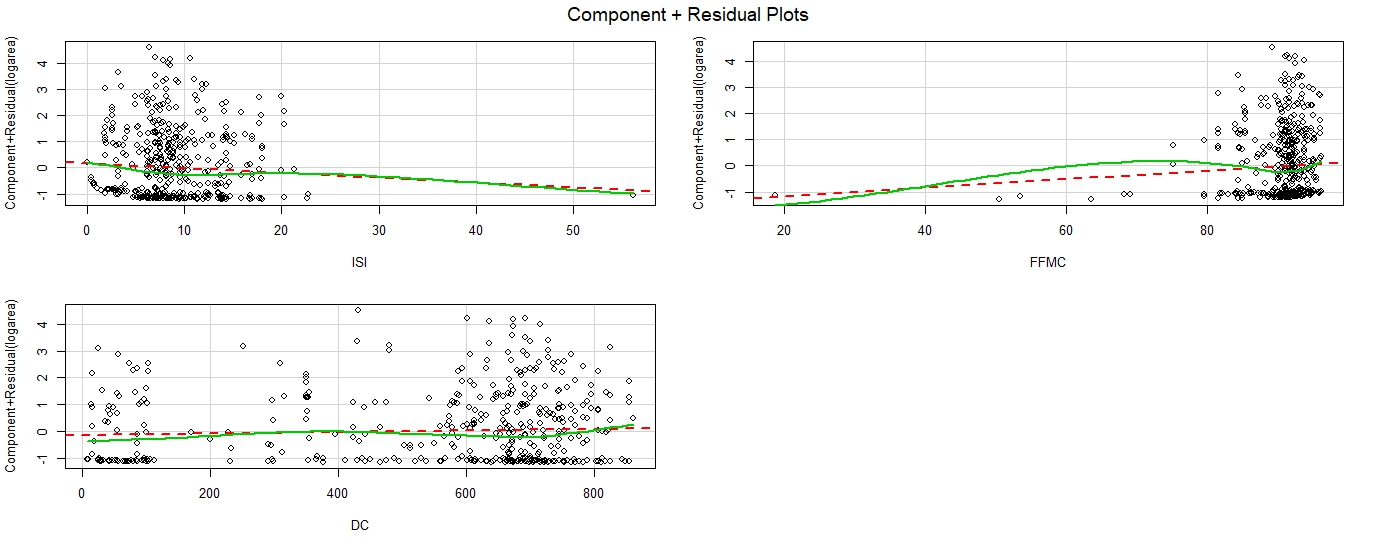
Alternative hypothesis: rho != 0

**crPlots(Model2)**

crplot is done to determine linearity.

From the plot we can find that DC is somewhat linear whereas, ISI and FFMC are not linear.

**Graph:**



**ncvTest(Model2)**

P value is significant, so homoscedasticity is not satisfied.

**Console:**

Non-constant Variance Score Test

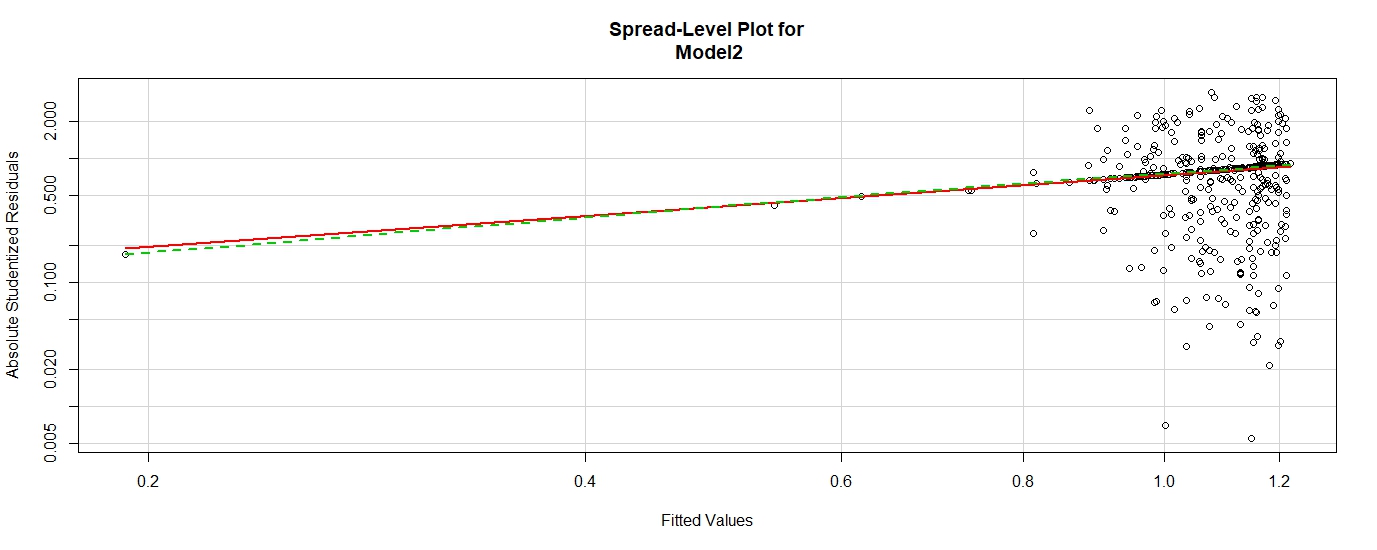
Variance formula: ~ fitted.values

Chisquare = 3.411769 Df = 1 p = 0.06473302

**spreadLevelPlot(Model2)**

The p value is significant and the suggested power transformation for getting a better homoscedasticity is 0.167.

**Graph:**



**Console:**

Suggested power transformation: 0.1671237

**install.packages("gvlma")**

To install gvlma package.

**require(gvlma)**

To load gvlma package.

**summary(gvlma(Model2))**

From the global statistics we can find that 2assumptions out of 5 assumptions are satisfied.

**Console:**

Call:

lm(formula = logarea ~ ISI + FFMC + DC, data = Assignment3\_Q2.2)

Residuals:

Min 1Q Median 3Q Max

-1.2212 -1.0958 -0.6220 0.9091 4.5552

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.4029509 1.1333900 -0.356 0.722

ISI -0.0182383 0.0154935 -1.177 0.240

FFMC 0.0165874 0.0135267 1.226 0.221

DC 0.0002733 0.0002565 1.065 0.287

Residual standard error: 1.354 on 510 degrees of freedom

Multiple R-squared: 0.007212, Adjusted R-squared: 0.001372

F-statistic: 1.235 on 3 and 510 DF, p-value: 0.2963

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:

Level of Significance = 0.05

Call:

gvlma(x = Model2)

Value p-value Decision

Global Stat 110.9027 0.000000 Assumptions NOT satisfied!

Skewness 101.9150 0.000000 Assumptions NOT satisfied!

Kurtosis 1.8988 0.168216 Assumptions acceptable.

Link Function 0.1289 0.719541 Assumptions acceptable.

Heteroscedasticity 6.9600 0.008335 Assumptions NOT satisfied!

**vif(Model2)**

We determine the variance inflation factor for checking multi collinearity error.

**Console:**

ISI FFMC DC

1.392427 1.493386 1.137320

**sqrt(vif(Model2))>2**

Since the square root values are less than 2 it suggests that there is no multi collinearity error.

**Console:**

ISI FFMC DC

FALSE FALSEFALSE

**outlierTest(Model2)**

From the Outlier test we find that the 480th record is an outlier.

**Console:**

No Studentized residuals with Bonferonni p < 0.05

Largest |rstudent|:

rstudent unadjusted p-value Bonferonni p

480 3.403703 0.00071722 0.36865

**fitted(Model2)[477:482]**

To see the predicted values for record 477 to 482.

**Console:**

480 481 482 483 484 485

1.077863 1.122761 1.122761 1.074469 1.074469 1.074469

**Model2$residuals[477:482]**

We inferred from the residual error that 480th has a high residual error.

Since the value is positive we can infer that our model under predicted the actual value of the 480th record by 4.55.

**Console:**

480 481 482 483 484 485

4.5552470 0.1989948 -1.1227611 -0.2459172 -1.0744690 2.2371683

**Hat.plot<- function(X)**

**{**

**p <- length(coefficients(X))**

**N <- length(fitted(X))**

**plot(hatvalues(X), main = "Plot of Hat Values")**

**abline(h=c(2,3)\*(p/N),col = "red")**

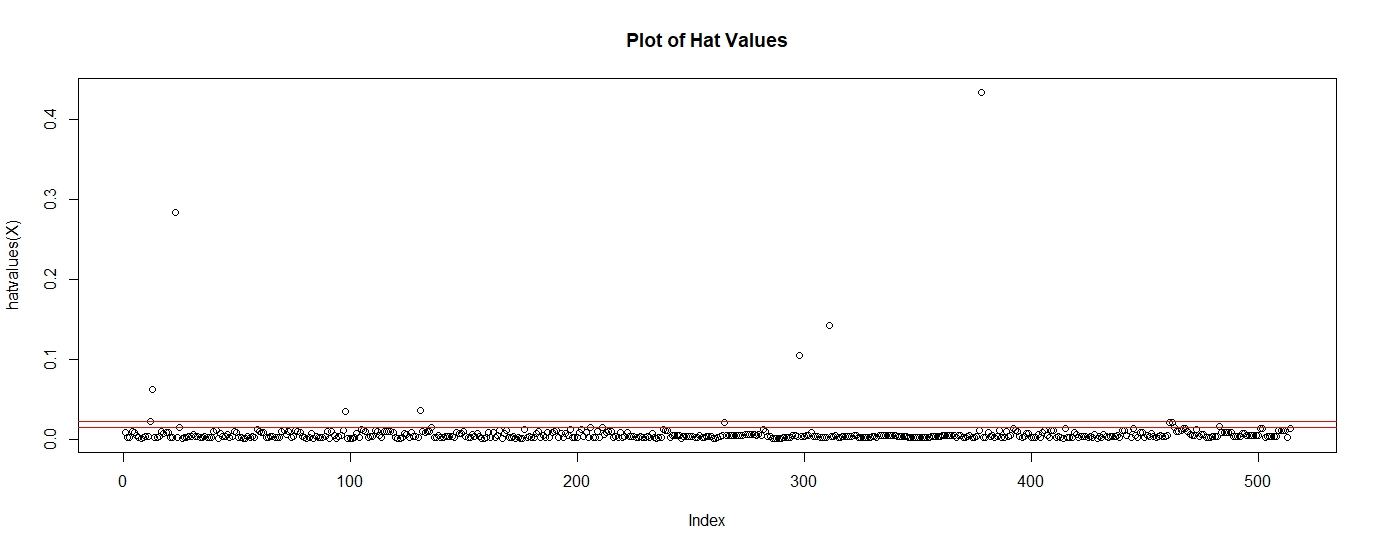
**identify(1:N, hatvalues(X), names(hatvalues(X)))**

**}**

**Hat.plot(Model2)**

From the hat plot we find that many records have high leverage.

**Graph:**

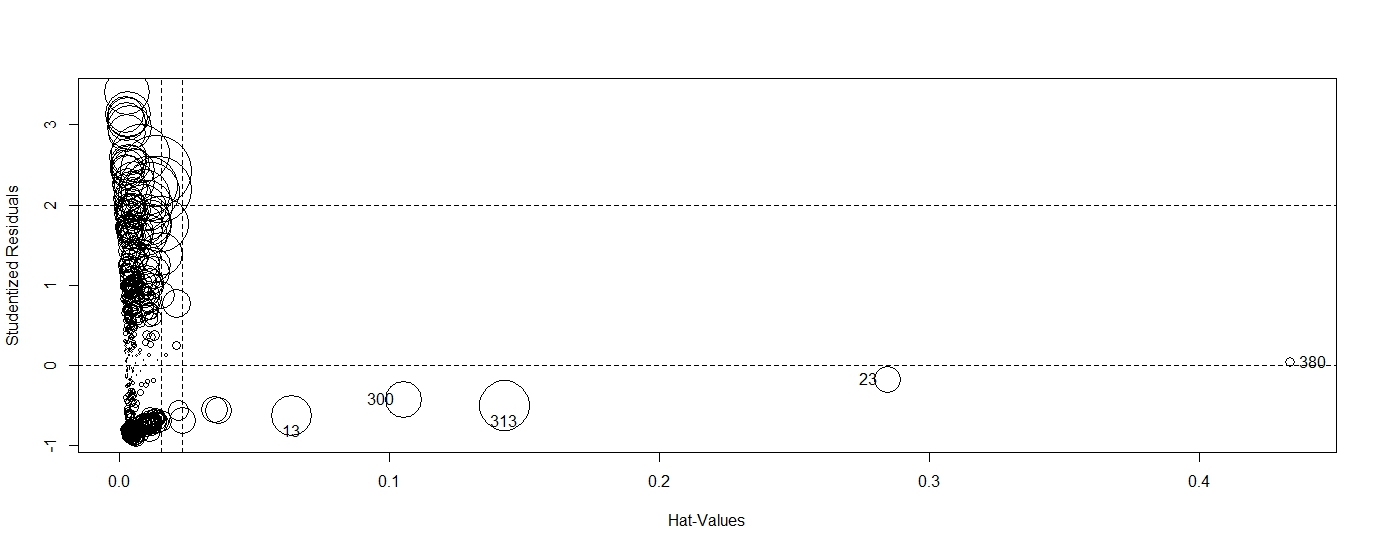


**influencePlot(Model2, id.method = "identify")**

We find that the circles with bigger radius have larger cooks distance hence are more influential so removing those will alter the model.

In this case several points are influential.

**Graph:**



**boxTidwell(logarea~DC,data = Assignment3\_Q2.2)**

To correct linearity assumption, non-significant p value suggests that it’s not required.

**Console:**

Score Statistic p-value MLE of lambda

-0.5009169 0.6164296 0.1205078

iterations = 9

**MODEL 3**

**Assignment3\_Q2.3 <- Assignment3\_Q2.2[-c(480),]**

To remove the 480th record.

**Model3 <- lm(logarea~ISI+FFMC+DC,data = Assignment3\_Q2.3)**

After seeing the scatter plot we are selecting DC, FFMC and ISI as the initial predictors.

**summary(Model3)**

From the summary we can find out R^2 is 0.0013

**Console:**

Call:

lm(formula = logarea ~ ISI + FFMC + DC, data = Assignment3\_Q2.3)

Residuals:

Min 1Q Median 3Q Max

-1.2213 -1.0973 -0.6261 0.9096 4.5550

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.4058135 1.1345751 -0.358 0.721

ISI -0.0181341 0.0155188 -1.169 0.243

FFMC 0.0166160 0.0135404 1.227 0.220

DC 0.0002729 0.0002567 1.063 0.288

Residual standard error: 1.355 on 509 degrees of freedom

Multiple R-squared: 0.007206, Adjusted R-squared: 0.001355

F-statistic: 1.232 on 3 and 509 DF, p-value: 0.2976

**dpar<- par(no.readonly = T)**

Defining a default graphic parameter to retain a single graph in the plot window.

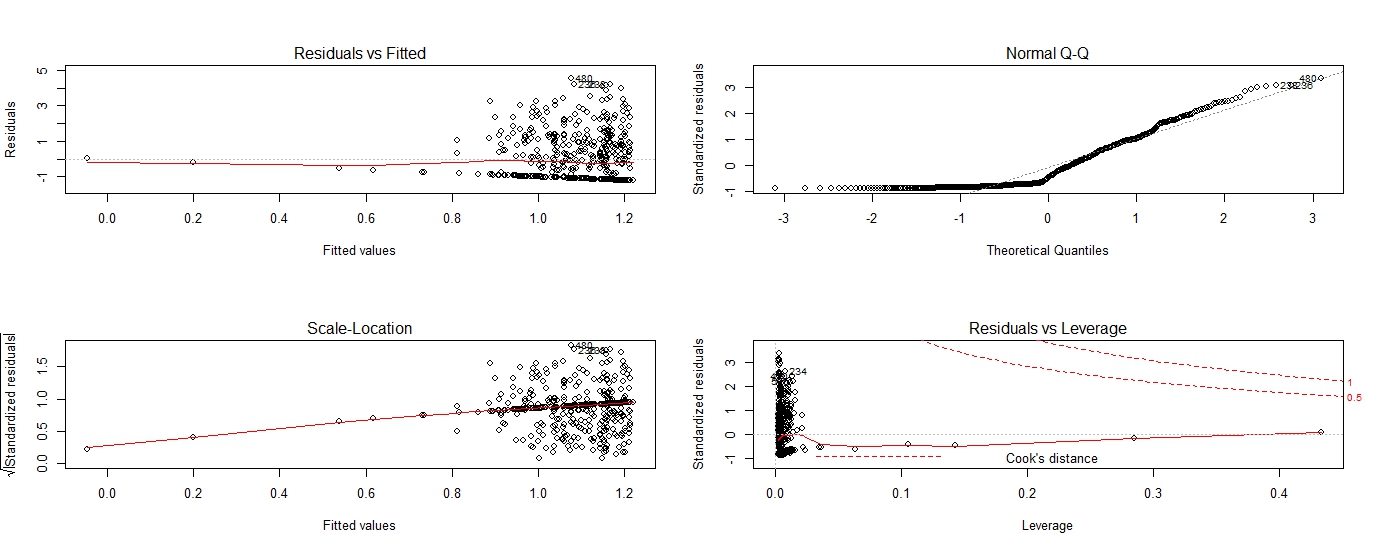
**par(mfrow = c(2,2))**

To generate 2X2 plots.

**plot(Model3)**

Data lacks Normality.

**Graph:**



**par(mfrow = c(1,1))**

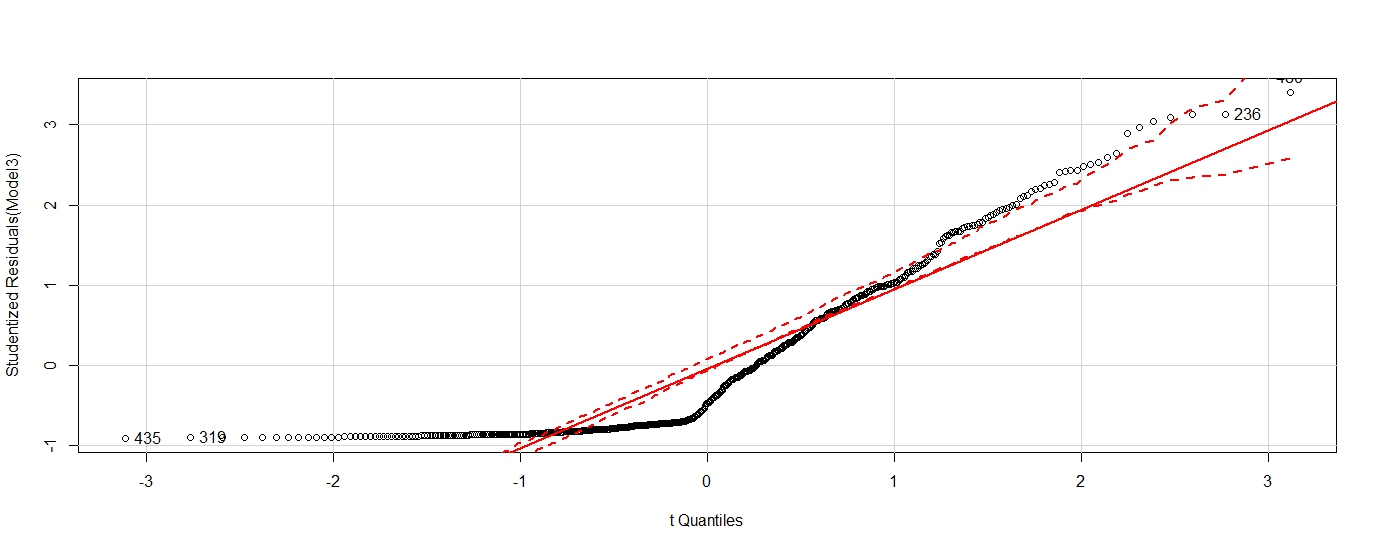
For having the default parameter for viewing 1 graph in the plot window.

**Enhanced approach**

**qqPlot(Model3,simulate = T,id.method = "identify", labels = rownames(Assignment3\_Q2.3))**

Normality is not perfect several points are deviating from the normal curve.

**Graph:**



**durbinWatsonTest(Model3)**

Durbin watson test is done in order to know the independent assumptions of the model since the p value is non-significant hence its independent.

**Console:**

lag Autocorrelation D-W Statistic p-value

1 0.5513265 0.8954481 0

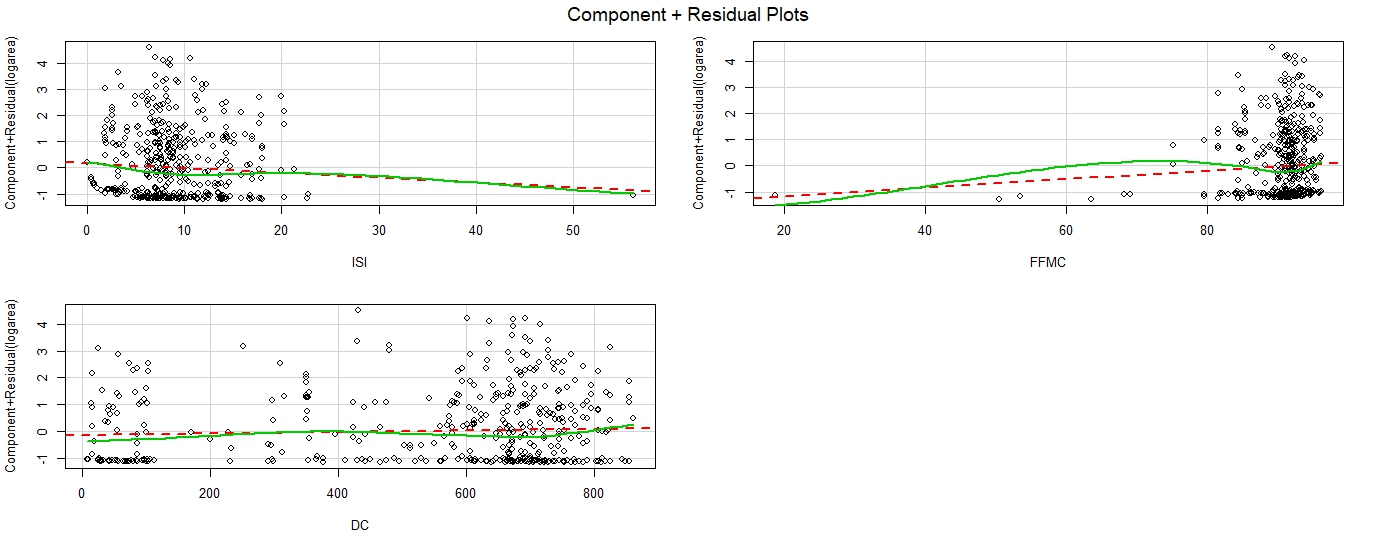
Alternative hypothesis: rho != 0

**crPlots(Model3)**

crplot is done to determine linearity

From the plot we can find that DC is somewhat linear whereas, FFMC and ISI are not linear.

**Graph:**



**ncvTest(Model3)**

P value is non-significant, so homoscedasticity is satisfied.

**Console:**

Non-constant Variance Score Test

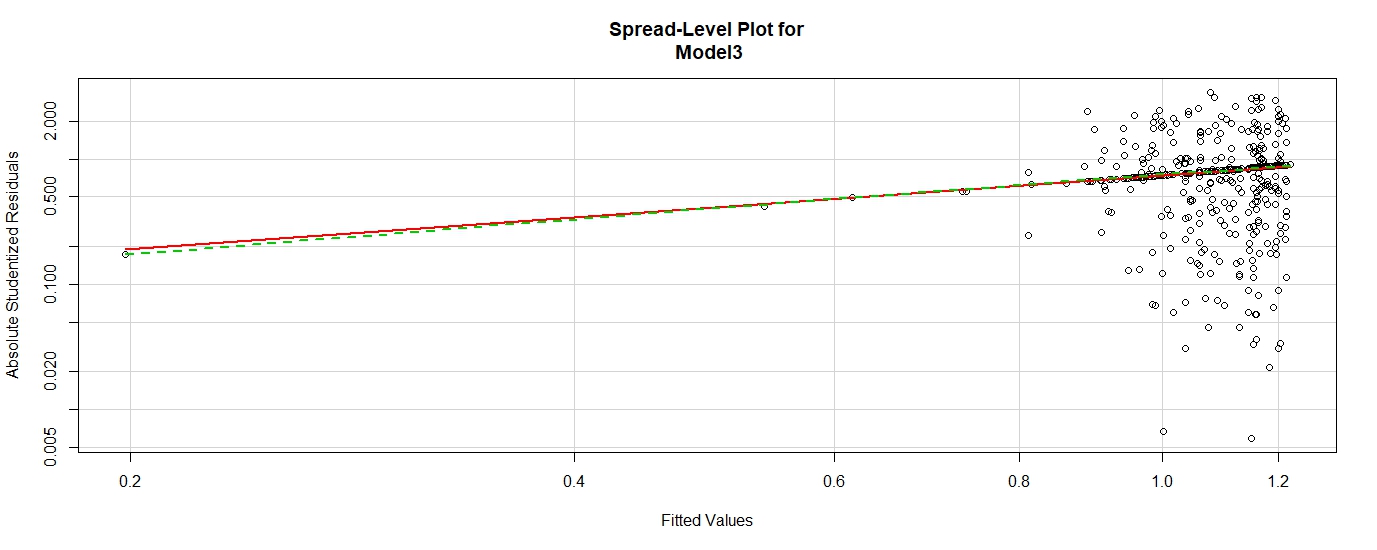
Variance formula: ~ fitted.values

Chisquare = 3.383994 Df = 1 p = 0.06583234

**spreadLevelPlot(Model3)**

The p value is non-significant and the suggested power transformation for getting a better homoscedasticity is 0.1664.

**Graph:**



**Console:**

Suggested power transformation: 0.1664404

**install.packages("gvlma")**

To install gvlma package.

**require(gvlma)**

To load gvlma package.

**summary(gvlma(Model3))**

From the global statistics we can find that 2 out of 5 assumptions are satisfied.

**Console:**

Call:

lm(formula = logarea ~ ISI + FFMC + DC, data = Assignment3\_Q2.3)

Residuals:

Min 1Q Median 3Q Max

-1.2213 -1.0973 -0.6261 0.9096 4.5550

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.4058135 1.1345751 -0.358 0.721

ISI -0.0181341 0.0155188 -1.169 0.243

FFMC 0.0166160 0.0135404 1.227 0.220

DC 0.0002729 0.0002567 1.063 0.288

Residual standard error: 1.355 on 509 degrees of freedom

Multiple R-squared: 0.007206, Adjusted R-squared: 0.001355

F-statistic: 1.232 on 3 and 509 DF, p-value: 0.2976

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:

Level of Significance = 0.05

Call:

gvlma(x = Model3)

Value p-value Decision

Global Stat 110.4730 0.000000 Assumptions NOT satisfied!

Skewness 101.3486 0.000000 Assumptions NOT satisfied!

Kurtosis 1.8013 0.179557 Assumptions acceptable.

Link Function 0.1315 0.716851 Assumptions acceptable.

Heteroscedasticity 7.1916 0.007325 Assumptions NOT satisfied!

**vif(Model3)**

We determine the variance inflation factor for checking multi collinearity error.

**Console:**

ISI FFMC DC

1.390950 1.491811 1.137323

**sqrt(vif(Model3))>2**

Since the square root values are less than 2 it suggests that there is no multi collinearity error.

**Console:**

ISI FFMC DC

FALSE FALSEFALSE

**outlierTest(Model3)**

From the Outlier test we find that the 480th record is an outlier.

**Console:**

No Studentized residuals with Bonferonni p < 0.05

Largest |rstudent|:

rstudent unadjusted p-value Bonferonni p

480 3.400316 0.00072607 0.37247

**fitted(Model3)[475:480]**

To see the predicted values for records from 475 to 480.

**Console:**

478 479 480 481 482 484

1.000087 1.000087 1.078066 1.123358 1.123358 1.075585

**Model3$residuals[475:480]**

We inferred from the residual error that 480th has a high residual error.

Since the value is positive we can infer that our model under predicted the actual value of the 480th record by 4.55.

**Console:**

478 479 480 481 482 484

1.1233713 0.1661838 4.5550433 0.1983977 -1.1233581 -1.0755847

**Hat.plot<- function(X)**

**{**

**p <- length(coefficients(X))**

**N <- length(fitted(X))**

**plot(hatvalues(X), main = "Plot of Hat Values")**

**abline(h=c(2,3)\*(p/N),col = "red")**

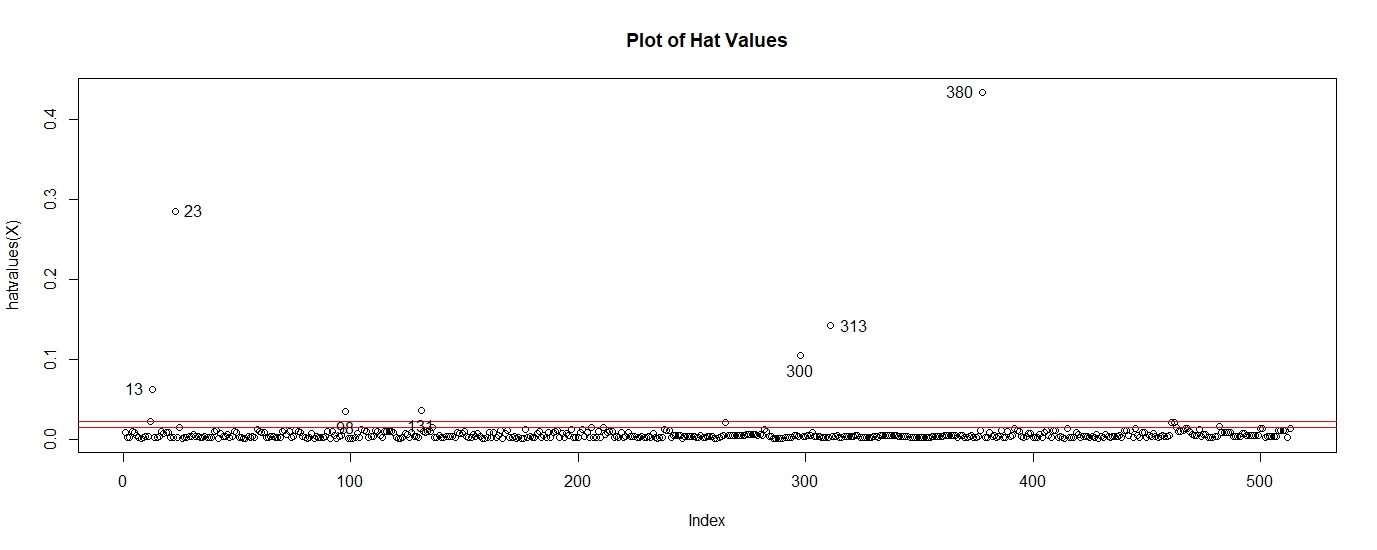
**identify(1:N, hatvalues(X), names(hatvalues(X)))**

**}**

**Hat.plot(Model3)**

From the hat plot we find that many records have high leverage.

**Graph:**

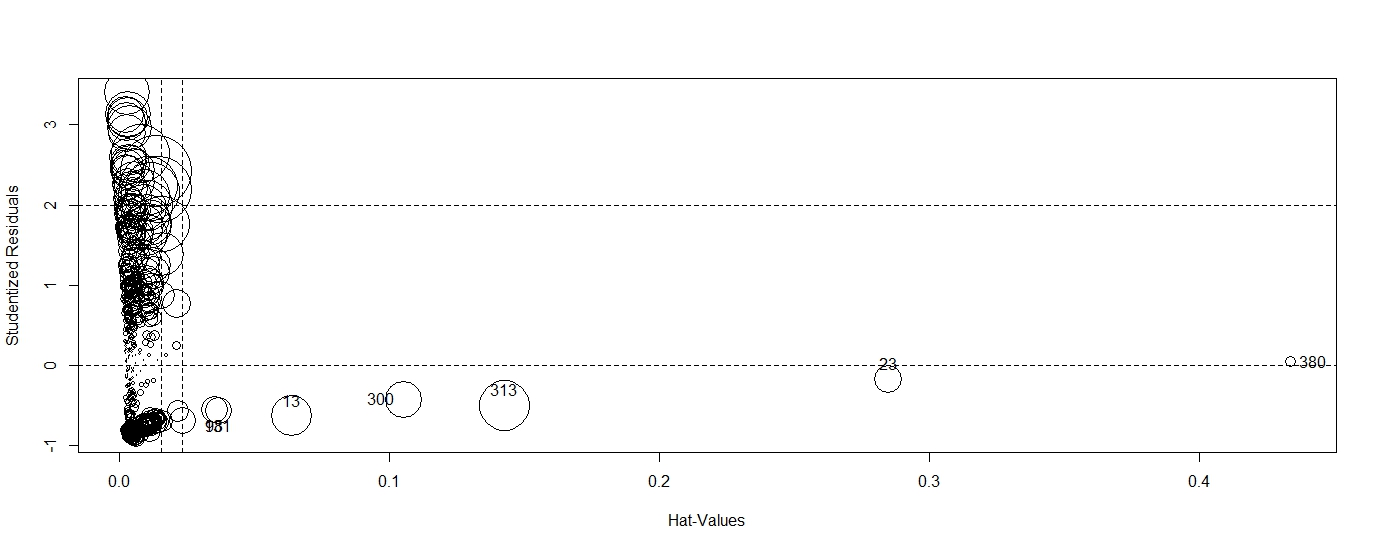


**influencePlot(Model3, id.method = "identify")**

We find that the circles with bigger radius have larger cooks distance hence are more influential so removing those will alter the model.

In this case several records are highly influential.

**Graph:**



**boxTidwell(logarea~DC,data = Assignment3\_Q2.3)**

To correct linearity assumption, non-significant p value suggests that it’s not required.

**Console:**

Score Statistic p-value MLE of lambda

-0.5089962 0.6107549 0.1170065

iterations = 9

**Conclusion:**

**Forest Fire Data has a very poor R^2 value for all the 3 models which we generated hence the data set is not suitable for linear regression and other methods can be performed.**