***Course: IE 7275 16692 – Data Mining in Engineering***

***Group 8***

***Names of group members and NUID: 1) Gandhar Kothari (NUID: 001676935)***

***2) Siddharth Patil (NUID: 001680396)***

***Document name: Homework 3 – Submission***

* Task 3:

**Code:**

**slumpdata <- read.csv(file.choose(), header=T)**

**slumpdata**

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

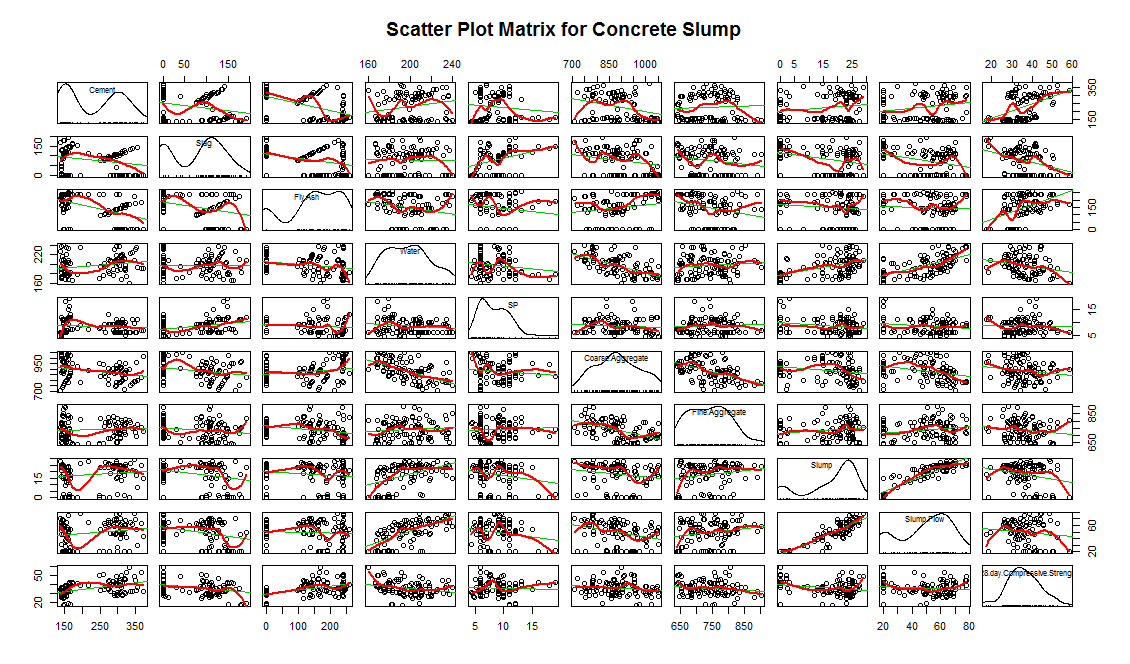
**require(car)**

**slumpd <- as.data.frame(slumpdata)**

**slumpd**

**scatterplotMatrix(slumpd, spread=FALSE, lty.smooth=2, main="scatter plot matrix")**

**Output:**

****

* Cement, Slag, Fly-Ash, Water, SP, Coarse-Aggregate and Fine Aggregate chosen as initial predictor variables. Slump, Slump Flow and 28-day Compressive Strength are response variables.
* All 3 response variables are built with each predictor variables and analysed. Outliers tests are performed and influential plots are generated.

**Code:**

**fit1 <- lm(Slump~Cement+Slag+Fly.Ash+Water+SP+Coarse.Aggregate+Fine.Aggregate, data=slumpd)**

**summary(fit1)**

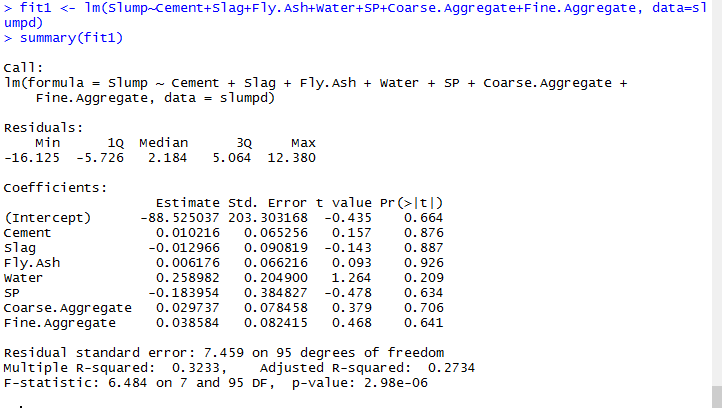
**fit2 <- lm(Slump.Flow~Cement+Slag+Fly.Ash+Water+SP+Coarse.Aggregate+Fine.Aggregate, data=slumpd)**

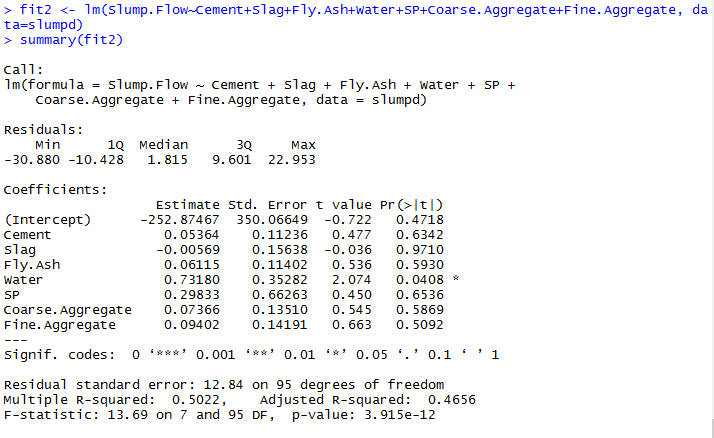
**summary(fit2)**

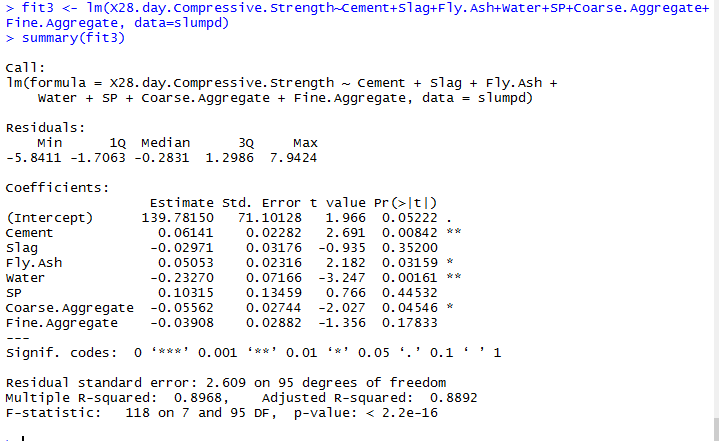
**fit3 <- lm(X28.day.Compressive.Strength~Cement+Slag+Fly.Ash+Water+SP+Coarse.Aggregate+Fine.Aggregate, data=slumpd)**

**summary(fit3)**

**Output:**

****

****

****

1. In 1st model Response variable is Slump.

* Predictor variables are Cement, Slag, Fly-Ash, Water, SP, Coarse-Aggregate and Fine Aggregate.
* The model accounts for 32.33% of the total variance.

1. In 2nd model response variable is slump flow.

* Predictor variables are Cement, Slag, Fly-Ash, Water, SP, Coarse-Aggregate and Fine Aggregate.
* The model accounts for 50.22% of the total variance.

1. In 2nd model response variable is slump flow.

* Predictor variables are Cement, Slag, Fly-Ash, Water, SP, Coarse-Aggregate and Fine Aggregate.
* The model accounts for 89.69% of the total variance.

**Code:**

**plot(fit1, main="fit1")**

**qqplot(fit1, labels=row.names(slumpd), id.method="identify", simulate=T, main="plot of fit1")**

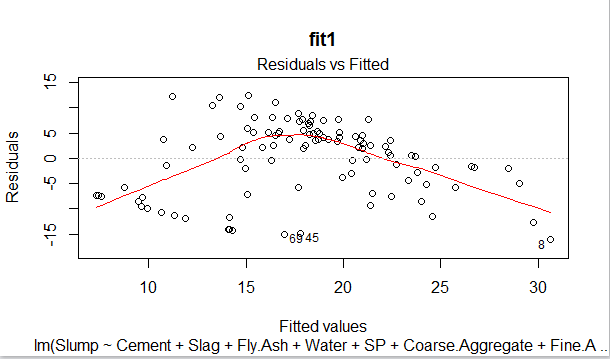
**plot(fit2, main="fit1")**

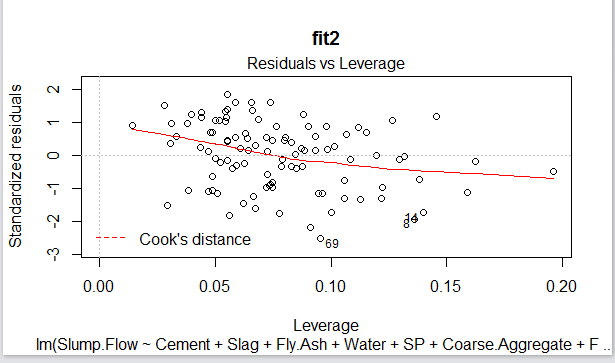
**qqplot(fit2, labels=row.names(slumpd), id.method="identify", simulate=T, main="plot of fit1")**

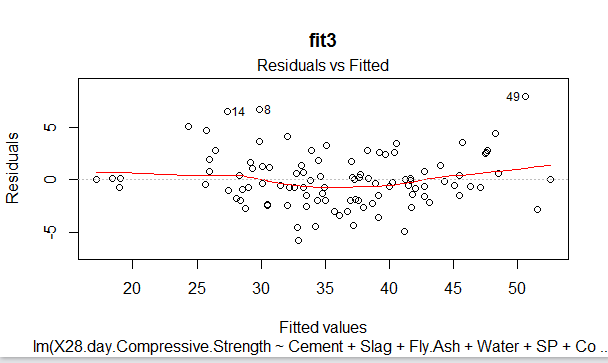
**plot(fit3, main="fit1")**

**qqplot(fit3, labels=row.names(slumpd), id.method="identify", simulate=T, main="plot of fit1")**

**Output:**

****

****

****

**FIT1:**

* The points in the plot shown do not fall on the straight 45 degree line and hence the normality assumption has been violated.
* The predictor variables have no significant connection to each other therefore independence assumption is true in this case.
* We can see there is a curved relation between residual vs fitted values hence the linearity does not hold true here.
* Homoskedasticity assumption is true in this scenario.

**FIT2:**

* The points in the plot shown do not fall on the straight 45 degree line and hence the normality assumption has been violated.
* The predictor variables have no significant connection to each other therefore independence assumption is true in this case.
* We can see there is a curved relation between residual vs fitted values hence the linearity does not hold true here.
* Homoskedasticity assumption is true in this scenario.

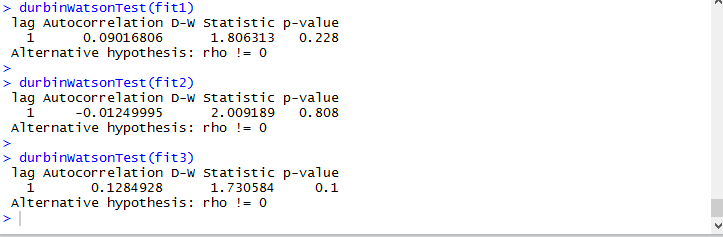
**Code:**

**durbinWatsonTest(fit1)**

**durbinWatsonTest(fit2)**

**durbinWatsonTest(fit3)**

**Output:**

****

* From the results we can see that 1st fit p value is 0.228 hence there is no autocorrelation and errors are independent.
* The p value of 0.1 in fit3 suggests that there is no autocorrelation and the errors are independent.

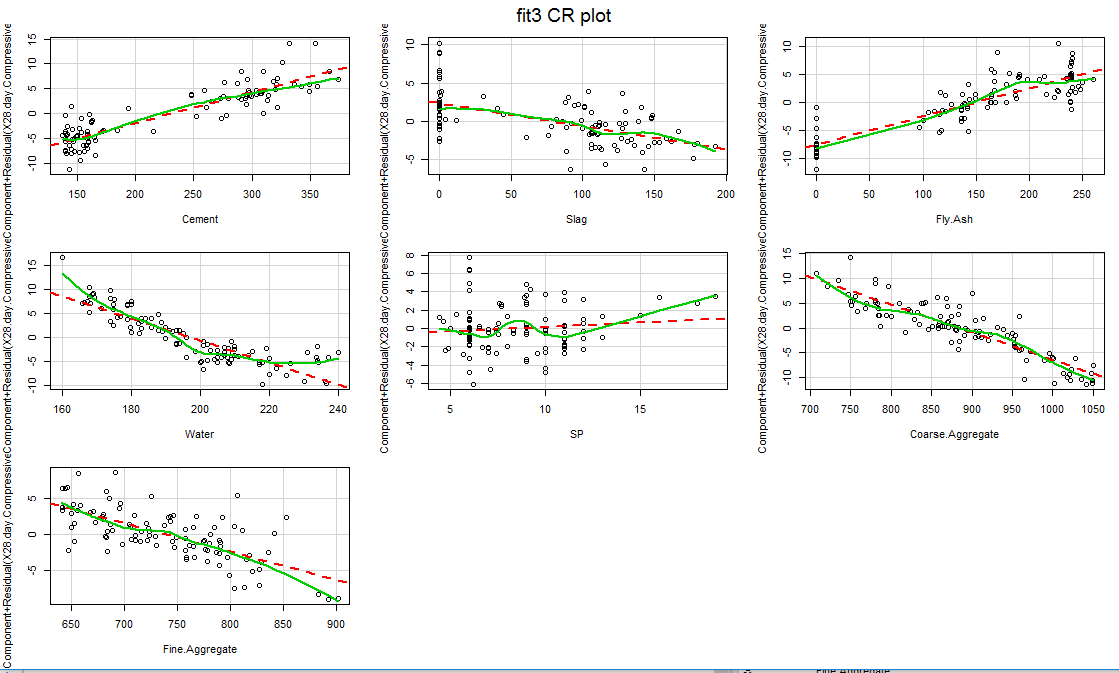
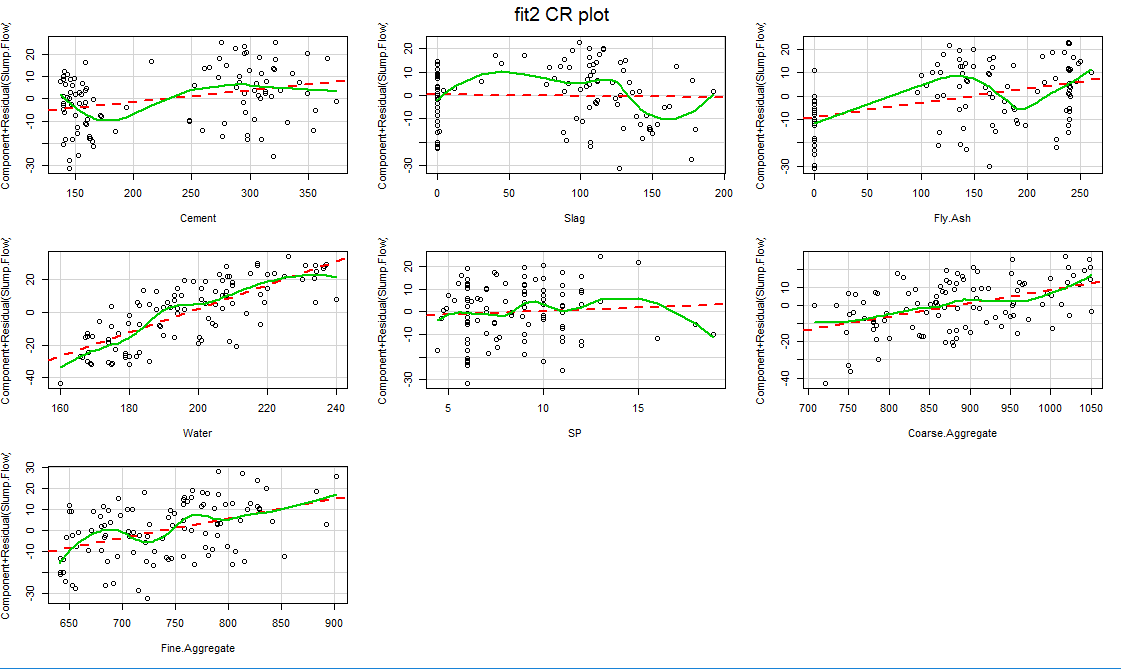
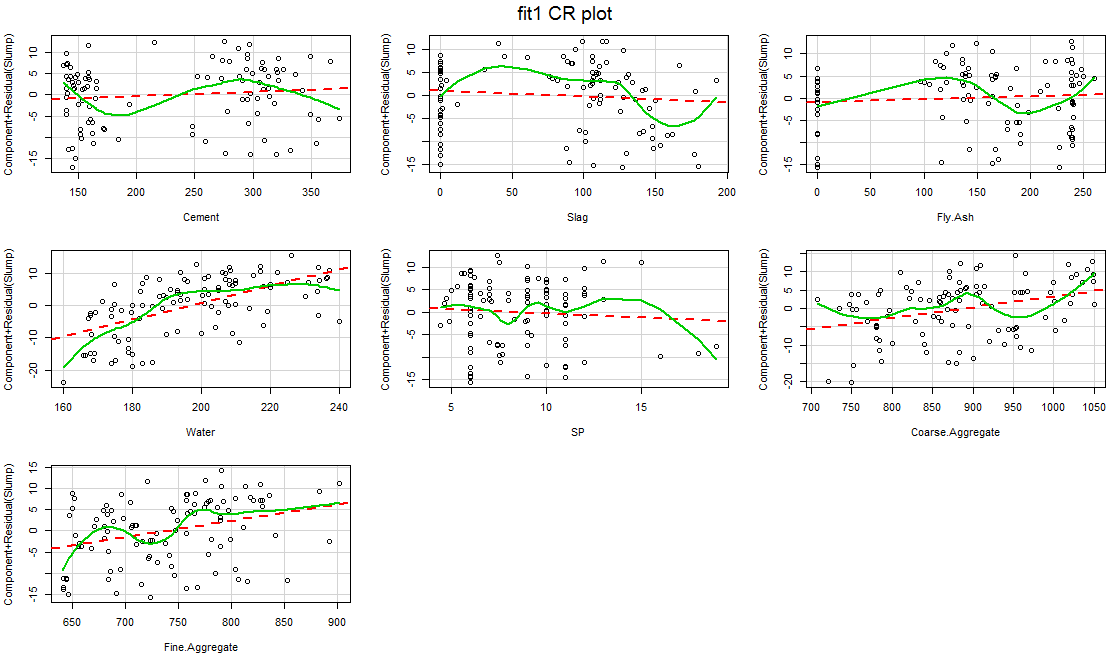
**Code:**

**crPlots(fit1, main="fit1 CR plot")**

**crPlots(fit2, main="fit2 CR plot")**

**crPlots(fit3, main="fit3 CR plot")**

**Output:**

****

* Component plus residual plot of fit1: The curve against the model regression line shows that there is a significant deviation in from the regression line. Which indicates the linearity assumption is not satisfied.
* Component plus residual plot of fit2 and fit3: there is no significance deviation from regression line. Therefore we can conclude that assumption of linearity is satisfied.

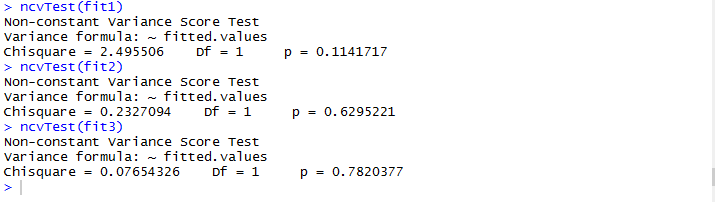
**Code:**

**ncvTest(fit1)**

**ncvTest(fit2)**

**ncvTest(fit3)**

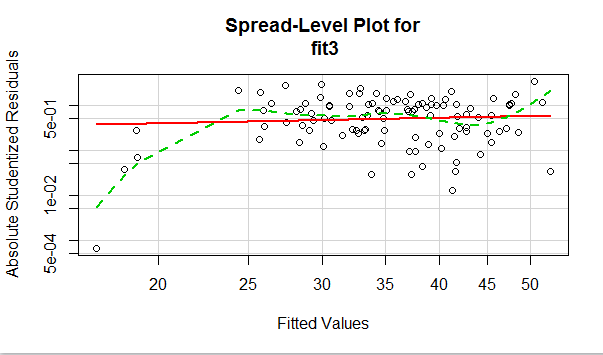
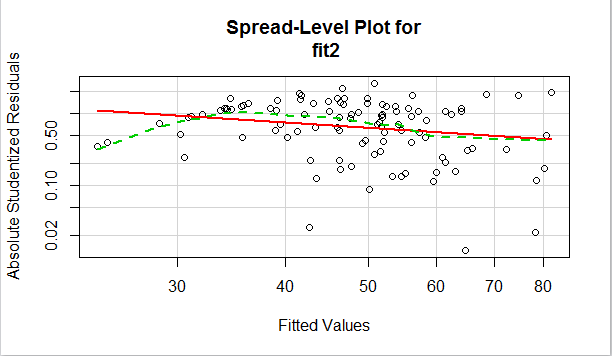
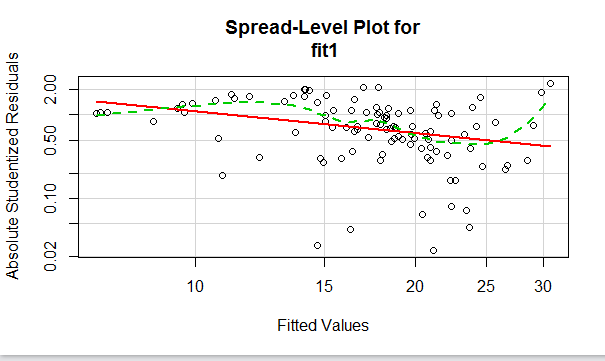
**Output:**

****

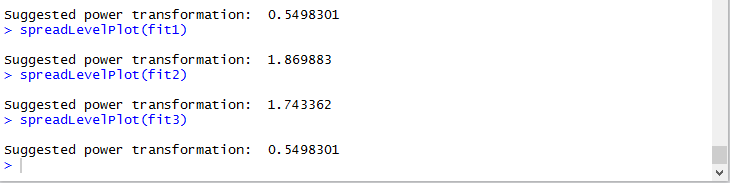
**Code:**

**spreadLevelPlot(fit1)**

**spreadLevelPlot(fit2)**

**spreadLevelPlot(fit3)**

* Suggested power of transformation is obtained through spreadlevelplot:



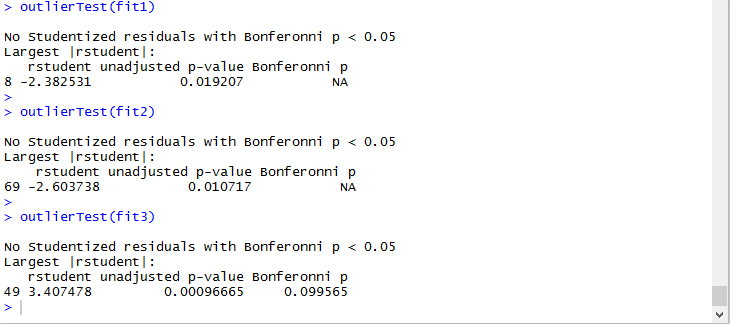
**Code:**

**outlierTest(fit1)**

**outlierTest(fit2)**

**outlierTest(fit3)**

**Output:**

****

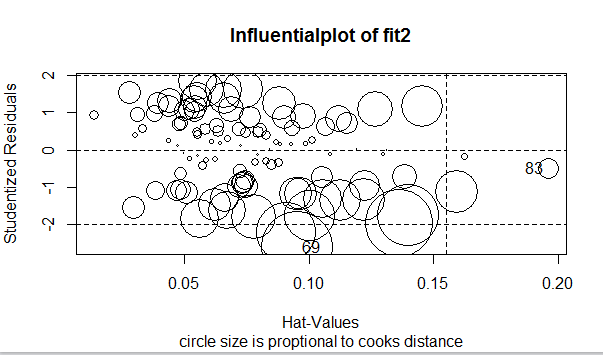
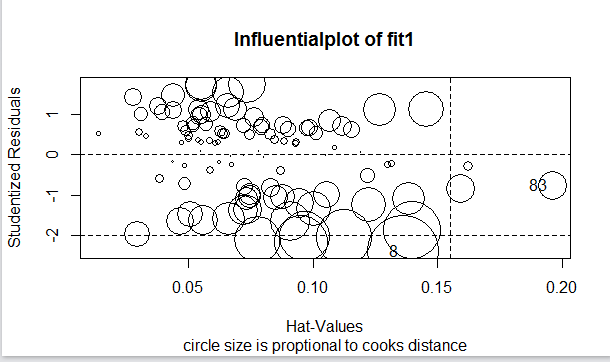
* Outliers test is conducted on every model and from the result it is safe to say that there are no significant outliers.

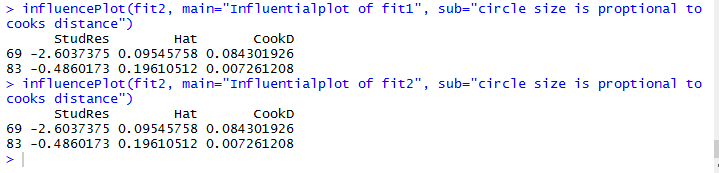
**Code:**

**influencePlot(fit1, main="Influentialplot of fit1", sub="circle size is proptional to cooks distance")**

**influencePlot(fit2, main="Influentialplot of fit2", sub="circle size is proptional to cooks distance")**

**Output:**

****

****

Unusual Observations:

* Outlier test is conducted and found there were no significant outliers.
* Influential plots are generated and following unusual combination were observed:

|  |  |  |
| --- | --- | --- |
| StudRes | Hat | CookD |
| 8 -2.3825309 | 0.1361820 | 0.3265184 |
| 83 -0.7529806 | 0.1961051 | 0.1317879 |

For above unusual observation following corrective measures were taken:

* Regression method is used to identify the best combination of predictors for this model.
* The best predictors are intercept, slag and water.

**Code:**

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

**require(gvlma)**

**gymodelfit = gvlma(fit1)**

**summary(gymodelfit)**

**Output:**

> summary(gymodelfit)

Call:

lm(formula = Slump ~ Cement + Slag + Fly.Ash + Water + SP + Coarse.Aggregate +

Fine.Aggregate, data = slumpd)

Residuals:

Min 1Q Median 3Q Max

-16.125 -5.726 2.184 5.064 12.380

Coefficients:

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

(Intercept) -88.525037 203.303168 -0.435 0.664

Cement 0.010216 0.065256 0.157 0.876

Slag -0.012966 0.090819 -0.143 0.887

Fly.Ash 0.006176 0.066216 0.093 0.926

Water 0.258982 0.204900 1.264 0.209

SP -0.183954 0.384827 -0.478 0.634

Coarse.Aggregate 0.029737 0.078458 0.379 0.706

Fine.Aggregate 0.038584 0.082415 0.468 0.641

Residual standard error: 7.459 on 95 degrees of freedom

Multiple R-squared: 0.3233, Adjusted R-squared: 0.2734

F-statistic: 6.484 on 7 and 95 DF, p-value: 2.98e-06

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

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

Level of Significance = 0.05

Call:

gvlma(x = fit1)

Value p-value Decision

Global Stat 43.8622 6.853e-09 Assumptions NOT satisfied!

Skewness 4.2686 3.882e-02 Assumptions NOT satisfied!

Kurtosis 2.0421 1.530e-01 Assumptions acceptable.

Link Function 36.6293 1.429e-09 Assumptions NOT satisfied!

Heteroscedasticity 0.9222 3.369e-01 Assumptions acceptable.

**Code:**

**gymodelfit2 = gvlma(fit2)**

**summary(gymodelfit2)**

**Output:**

> summary(gymodelfit2)

Call:

lm(formula = Slump.Flow ~ Cement + Slag + Fly.Ash + Water + SP +

Coarse.Aggregate + Fine.Aggregate, data = slumpd)

Residuals:

Min 1Q Median 3Q Max

-30.880 -10.428 1.815 9.601 22.953

Coefficients:

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

(Intercept) -252.87467 350.06649 -0.722 0.4718

Cement 0.05364 0.11236 0.477 0.6342

Slag -0.00569 0.15638 -0.036 0.9710

Fly.Ash 0.06115 0.11402 0.536 0.5930

Water 0.73180 0.35282 2.074 0.0408 \*

SP 0.29833 0.66263 0.450 0.6536

Coarse.Aggregate 0.07366 0.13510 0.545 0.5869

Fine.Aggregate 0.09402 0.14191 0.663 0.5092

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.84 on 95 degrees of freedom

Multiple R-squared: 0.5022, Adjusted R-squared: 0.4656

F-statistic: 13.69 on 7 and 95 DF, p-value: 3.915e-12

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

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

Level of Significance = 0.05

Call:

gvlma(x = fit2)

Value p-value Decision

Global Stat 21.919 2.080e-04 Assumptions NOT satisfied!

Skewness 1.703 1.919e-01 Assumptions acceptable.

Kurtosis 2.382 1.228e-01 Assumptions acceptable.

Link Function 16.433 5.041e-05 Assumptions NOT satisfied!

Heteroscedasticity 1.401 2.365e-01 Assumptions acceptable.

**Code:**

**gymodelfit3 = gvlma(fit3)**

**summary(gymodelfit3)**

**Output:**

> summary(gymodelfit3)

Call:

lm(formula = X28.day.Compressive.Strength ~ Cement + Slag + Fly.Ash +

Water + SP + Coarse.Aggregate + Fine.Aggregate, data = slumpd)

Residuals:

Min 1Q Median 3Q Max

-5.8411 -1.7063 -0.2831 1.2986 7.9424

Coefficients:

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

(Intercept) 139.78150 71.10128 1.966 0.05222 .

Cement 0.06141 0.02282 2.691 0.00842 \*\*

Slag -0.02971 0.03176 -0.935 0.35200

Fly.Ash 0.05053 0.02316 2.182 0.03159 \*

Water -0.23270 0.07166 -3.247 0.00161 \*\*

SP 0.10315 0.13459 0.766 0.44532

Coarse.Aggregate -0.05562 0.02744 -2.027 0.04546 \*

Fine.Aggregate -0.03908 0.02882 -1.356 0.17833

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.609 on 95 degrees of freedom

Multiple R-squared: 0.8968, Adjusted R-squared: 0.8892

F-statistic: 118 on 7 and 95 DF, p-value: < 2.2e-16

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

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

Level of Significance = 0.05

Call:

gvlma(x = fit3)

Value p-value Decision

Global Stat 13.8618 0.007749 Assumptions NOT satisfied!

Skewness 5.2971 0.021361 Assumptions NOT satisfied!

Kurtosis 1.8595 0.172685 Assumptions acceptable.

Link Function 5.8936 0.015196 Assumptions NOT satisfied!

Heteroscedasticity 0.8117 0.367631 Assumptions acceptable.

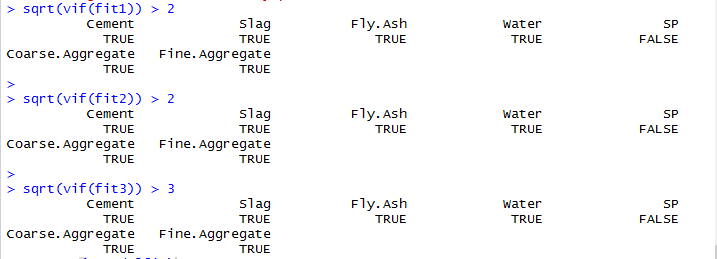
**Code: (Multicolinearity test)**

**sqrt(vif(fit1)) > 2**

**sqrt(vif(fit2)) > 2**

**sqrt(vif(fit3)) > 3**

**Output:**

****

**Code and Output for power transform:**

> summary(powerTransform(slumpd$Slump))

bcPower Transformation to Normality

Est.Power Std.Err. Wald Lower Bound Wald Upper Bound

S$Slump 2.0004 0.2564 1.4978 2.5029

Likelihood ratio tests about transformation parameters

LRT df pval

LR test, lambda = (0) 123.00861 1 0.000000e+00

LR test, lambda = (1) 21.21781 1 4.099378e-06

> summary(powerTransform(slumpd$Slump.Flow))

bcPower Transformation to Normality

Est.Power Std.Err. Wald Lower Bound Wald Upper Bound

csdn$Slump.Flow 1.4678 0.2723 0.9342 2.0015

Likelihood ratio tests about transformation parameters

LRT df pval

LR test, lambda = (0) 31.061871 1 2.499330e-08

LR test, lambda = (1) 3.036391 1 8.141676e-02

> summary(powerTransform(slumpd$X28.day.Compressive.Strength))

bcPower Transformation to Normality

Est.Power Std.Err. Wald Lower Bound Wald Upper Bound

csdn$X28.day.Compressive.Strength 0.753 0.3416 0.0834 1.4226

Likelihood ratio tests about transformation parameters

LRT df pval

LR test, lambda = (0) 5.1208483 1 0.02364006

LR test, lambda = (1) 0.5133552 1 0.47369003

**Code:**

**install.packages("leaps")**

**require(leaps)**

**leaps1 = regsubsets(Slump~Cement+Slag+Fly.Ash+Water+SP+Coarse.Aggregate+Fine.Aggregate, data= slumpd, nbest = 4)**

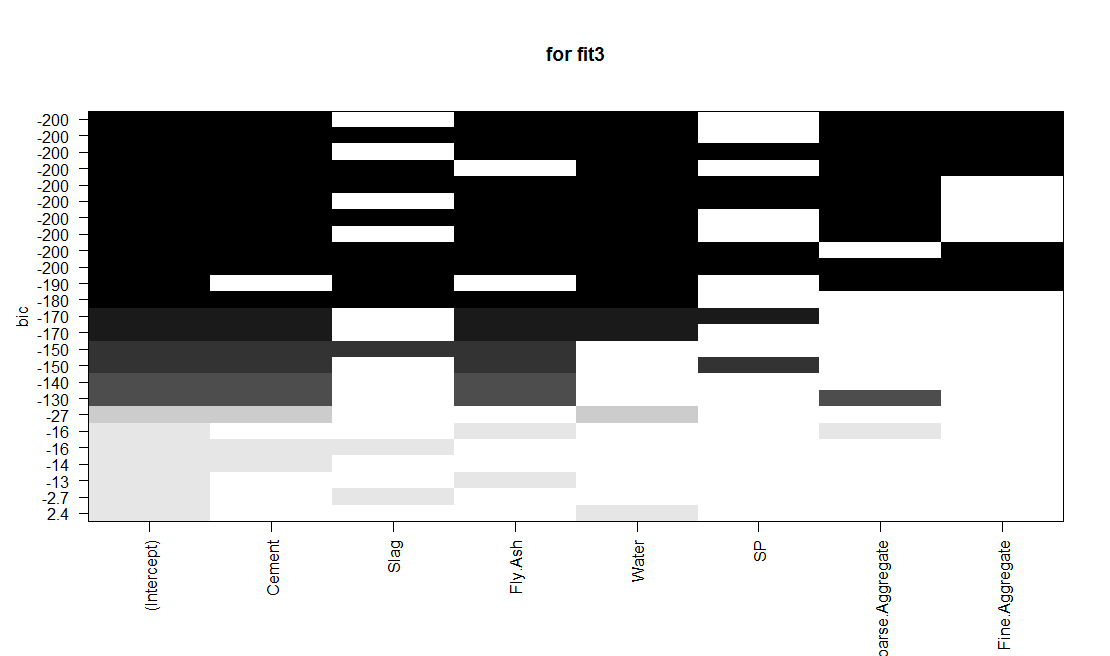
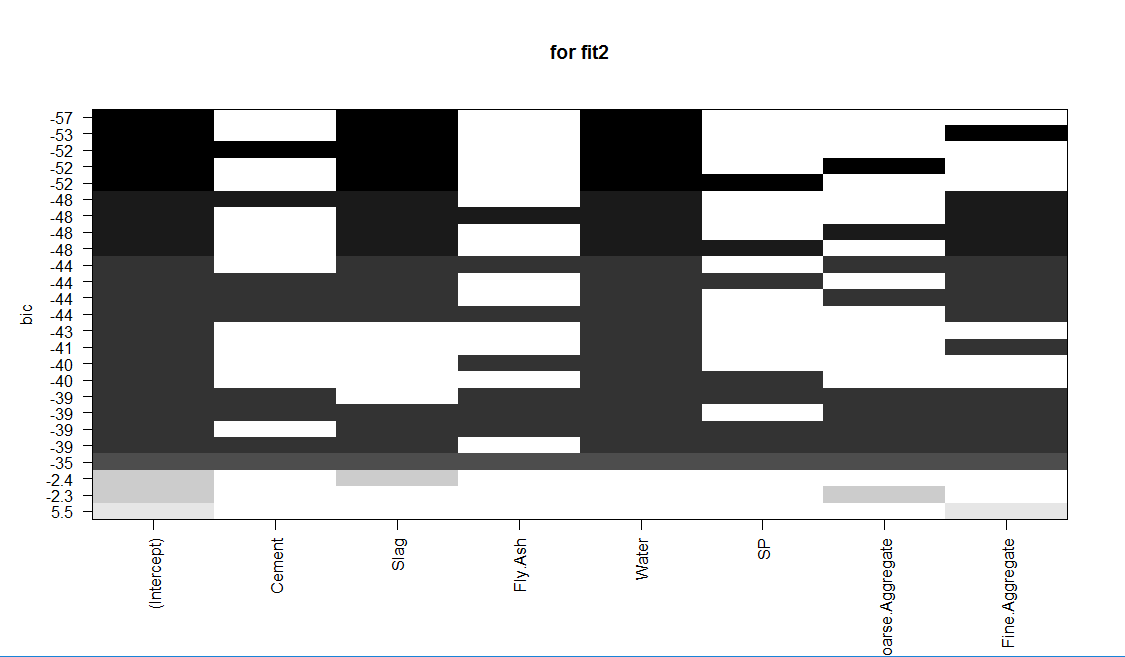
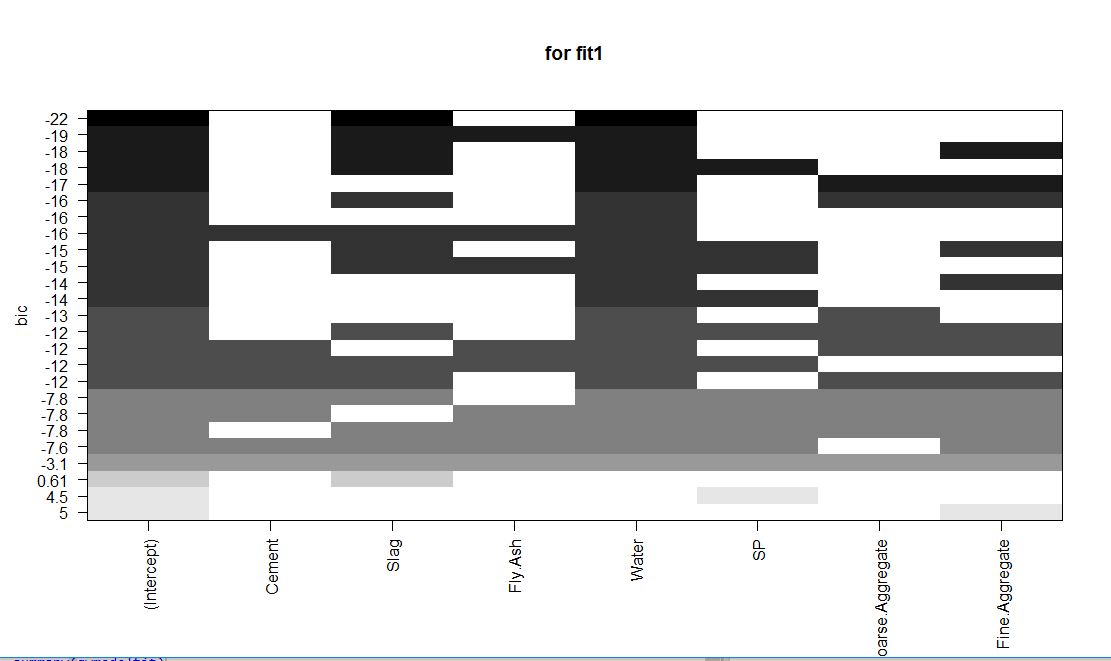
**plot(leaps1, main="for fit1")**

**leaps2 = regsubsets(Slump.Flow~Cement+Slag+Fly.Ash+Water+SP+Coarse.Aggregate+Fine.Aggregate, data= slumpd, nbest = 4)**

**plot(leaps2, main="for fit2")**

**leaps3 = regsubsets(X28.day.Compressive.Strength~Cement+Slag+Fly.Ash+Water+SP+Coarse.Aggregate+Fine.Aggregate, data= slumpd, nbest = 4)**

**plot(leaps3, main="for fit3")**

**Output:**

**FIT3:**

* Response variable is 28-Day Compressive Strength
* Predictor variables are Cement, Slag, Fly-Ash, Water, SP, Coarse-Aggregate and Fine Aggregate
* From summary results we can see that it accounts for 89.68% of the variance.

Assumption tests:

* The points in the plot shown do not fall on the straight 45 degree line and hence the normality assumption has been violated.
* The predictor variables have no significant connection to each other and therefore the independence of errors assumption is satisfied in this scenario.
* We can see a curved relationship in the Residuals vs fitted graph and hence the linearity assumption has been violated.
* Homoscedasticity assumption has been met in this scenario as points in the scale location graph indeed form a random band around the line.

Unusual Observations:

* There were no significant outliers detected in this model through the outlier test.
* In the influence plots the following were identified as unusual observations by the influence plot:

|  |  |  |  |
| --- | --- | --- | --- |
| StudRes | | Hat | CookD |
| 49 | 3.407478 | 0.1124498 | 0.406704 |
| 83 | 1.795464 | 0.1961051 | 0.309922 |

Corrective Measures:

Transformation is not needed for this data. We can see from the plot that the best combination of predictors would be:

Cement, Fly-Ash , water, Slag and Coarse-Aggregate and fine aggregate.

Task 4: Forest fire data

1. Plotting the Scatter Plot Matrix –

CODE -

library(xlsx)

FFdata <- read.xlsx("Forest Fires Data.xlsx", sheetName = "forestfires")

#creating a data frame for further anaysis

FFdataDFr <- as.data.frame(FFdata[,c("X","Y","Month","Day","FFMC","DMC","DC","ISI","Temp","RH","Wind","Rain","Area")])

#creating a scatterplot to understand the correlation amongst predictors

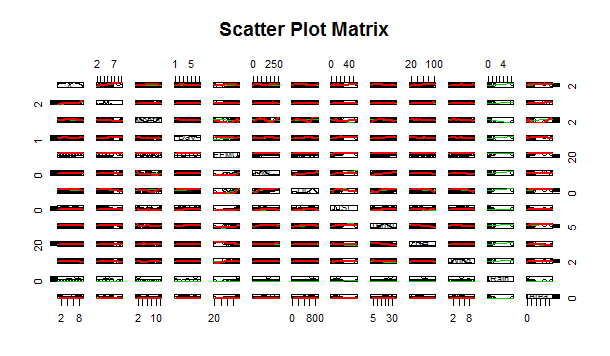
install.packages("car")

require(car)

scatterplotMatrix(FFdataDFr, spread = FALSE, lty.smooth=2,

main = "Scatter Plot Matrix")

RESULT:



The initial set of predictors that we are going to use is:

* FFMC
* DMC
* DC
* ISI
* Temp
* RH
* Wind

1. Building a few regression models to find the best model.

CODE:

fit1 <- lm(Area ~ FFMC + DMC + DC + ISI + Temp + RH + Wind, data = FFdataNN)

summary(fit1)

RESULT:

Call:

lm(formula = Area ~ FFMC + DMC + DC + ISI + Temp + RH + Wind,

data = FFdataNN)

Residuals:

Min 1Q Median 3Q Max

-0.04294 -0.02922 -0.02021 0.01109 0.28802

Coefficients:

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

(Intercept) 2.514e+00 4.688e-02 53.625 <2e-16 \*\*\*

FFMC 1.941e-04 5.016e-04 0.387 0.6990

DMC 5.590e-05 5.076e-05 1.101 0.2713

DC 3.032e-06 1.237e-05 0.245 0.8064

ISI -8.587e-04 5.867e-04 -1.464 0.1439

Temp 3.043e-04 5.924e-04 0.514 0.6077

RH -1.809e-04 1.773e-04 -1.020 0.3081

Wind 2.300e-03 1.265e-03 1.818 0.0696 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04838 on 509 degrees of freedom

Multiple R-squared: 0.01903, Adjusted R-squared: 0.005539

F-statistic: 1.411 on 7 and 509 DF, p-value: 0.1986

CODE:

fit2 <- lm(Area ~ Temp + RH + Wind, data = FFdataNN)

summary(fit2)

RESULT:

Call:

lm(formula = Area ~ Temp + RH + Wind, data = FFdataNN)

Residuals:

Min 1Q Median 3Q Max

-0.04011 -0.02899 -0.02145 0.01135 0.28945

Coefficients:

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

(Intercept) 2.5252314 0.0150980 167.256 <2e-16 \*\*\*

Temp 0.0005574 0.0004432 1.258 0.209

RH -0.0001038 0.0001540 -0.674 0.500

Wind 0.0018916 0.0012233 1.546 0.123

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04839 on 513 degrees of freedom

Multiple R-squared: 0.01055, Adjusted R-squared: 0.004767

F-statistic: 1.824 on 3 and 513 DF, p-value: 0.1418

CODE:

fit3 <- lm(Area ~ FFMC + DMC + DC + ISI, data = FFdataNN)

summary(fit3)

RESULT:

Call:

lm(formula = Area ~ FFMC + DMC + DC + ISI, data = FFdataNN)

Residuals:

Min 1Q Median 3Q Max

-0.03938 -0.02886 -0.02184 0.01109 0.29285

Coefficients:

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

(Intercept) 2.500e+00 3.994e-02 62.600 <2e-16 \*\*\*

FFMC 4.122e-04 4.773e-04 0.864 0.388

DMC 4.782e-05 4.713e-05 1.015 0.311

DC 2.476e-06 1.184e-05 0.209 0.834

ISI -6.225e-04 5.578e-04 -1.116 0.265

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0485 on 512 degrees of freedom

Multiple R-squared: 0.008006, Adjusted R-squared: 0.0002558

F-statistic: 1.033 on 4 and 512 DF, p-value: 0.3896

INTERPREATION FOR 2):

The response variable is Area. The are 3 linear models:

- fit1 is a combination of all predictor variables.

- fit2 is a combination of RH, Wind and Temp.

- fit3 is a combination of FFMC, DC, DMC, ISI.

The variable Rain is not correlated with the response variable and had most of its values as 0.

1. Perform Regression diagnostics using typical and enhanced approach

CODE:

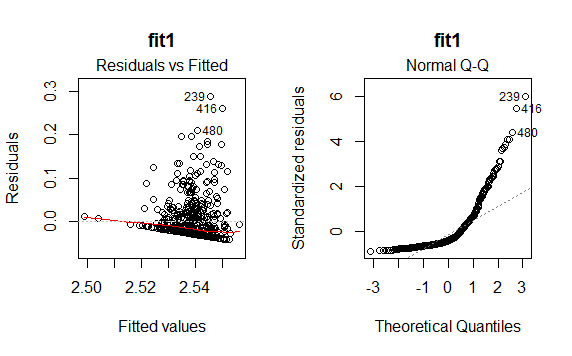
#First we use the typical appraoch

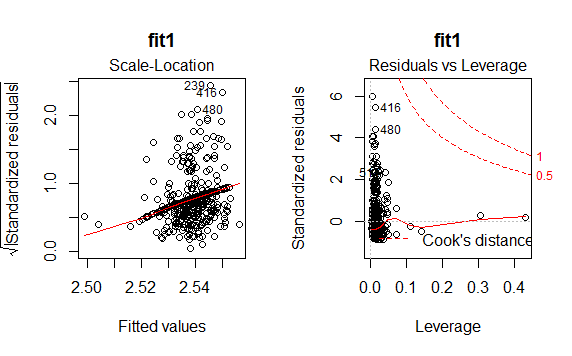
# for fit1

par(mfrow=c(1,2))

plot(fit1, main = "fit1")

RESULT:



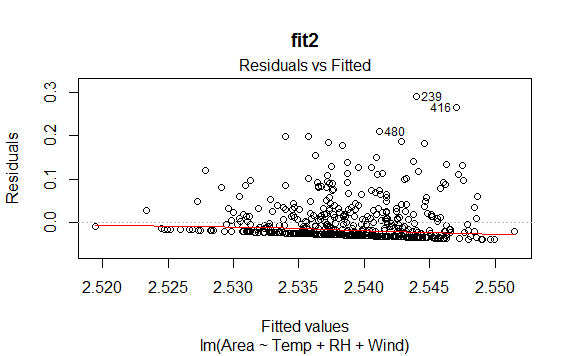


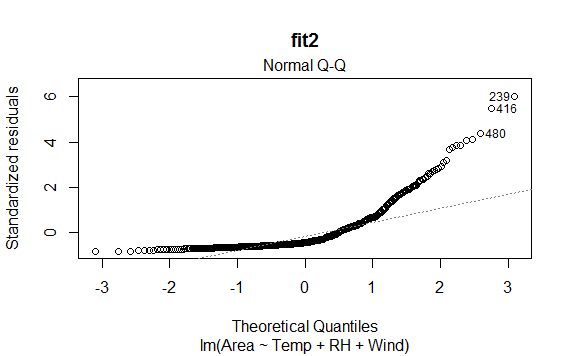
# for fit2

par(mfrow=c(1,2))

plot(fit2, main = "fit2")

RESULT:



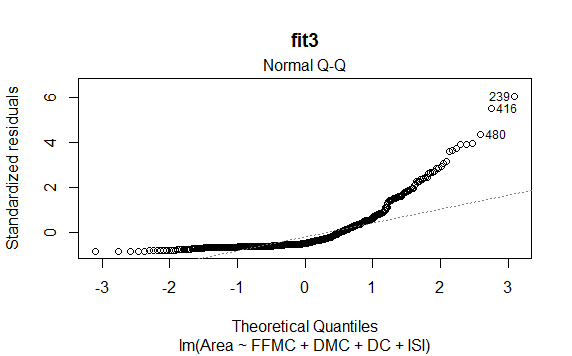


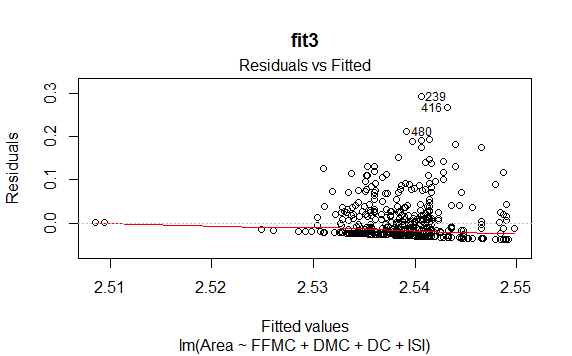
# for fit3

par(mfrow=c(1,2))

plot(fit3, main = "fit3")

RESULT:





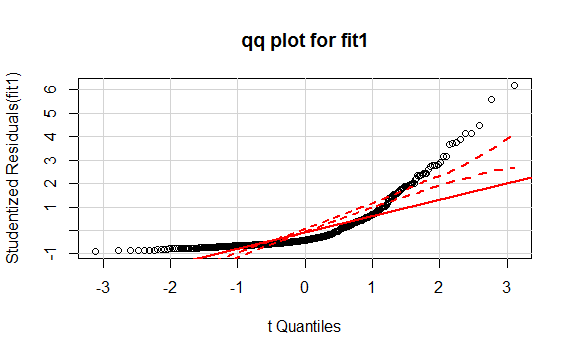
#Using enhanced approach

CODE:

# for fit1

qqPlot(fit1,main="qq plot for fit1")

RESULT:

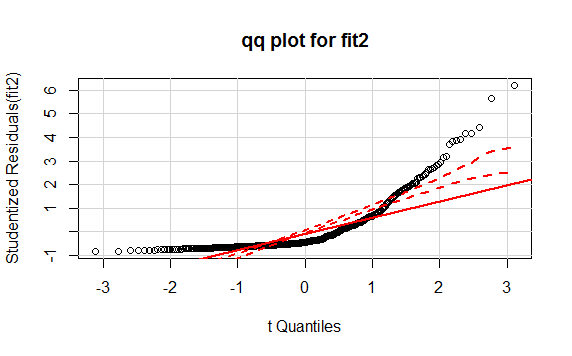


CODE:

# for fit2

qqPlot(fit2,main="qq plot for fit2")

RESULT:

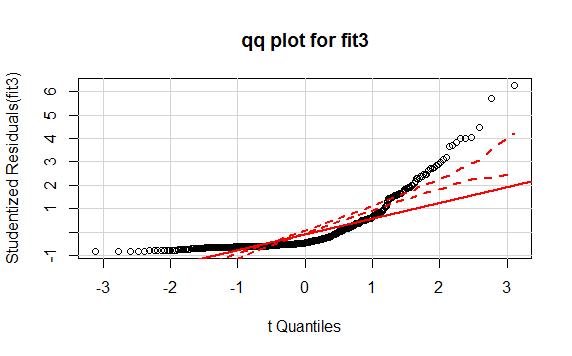


CODE:

# for fit3

qqPlot(fit3,main="qq plot for fit3")

RESULT:



INTERPRETATION for MODEL:

1. Model: fit1

Area is the Response variable.

Predictor variables are FFMC, DMC, DC, ISI, Temp, RH and Wind

The model accounts for 1.93% of the variance as can be seen from the results of the summary.

Testing the Assumptions:

Normality: The points do not fall on a 45degree line in the qq plot hence normality assumption has not been followed.

Independence of errors: The predictor variables have significant connection to each other and therefore the independence of errors assumption does not hold true in this case.

Linearity: We cannot see a linear relationship in the Residuals vs fitted graph and hence the linearity assumption has been violated.

Homoscedasticity: From the spread level graph we see that we have not followed this assumption

QQplot results: The plot the normality assumption has not been followed as there are many points that do not fall on a straight line.

Durbin-Watson test: The result of the DW test reveal that the p value of 0 and an auto correction of 0.498749.

1. Model: fit2

Area is the Response variable.

Predictor variables are Temp, RH and Wind

The model accounts for 1.055% of the variance as can be seen from the results of the summary.

Testing the Assumptions:

Normality: The points do not fall on a 45degree line in the qq plot hence normality assumption has not been followed.

Independence of errors: The predictor variables have significant connection to each other and therefore the independence of errors assumption does not hold true in this case.

Linearity: We cannot see a linear relationship in the Residuals vs fitted graph and hence the linearity assumption has been violated.

Homoscedasticity: From the spread level graph we see that we have not followed this assumption

QQplot results: The plot the normality assumption has not been followed as there are many points that do not fall on a straight line.

Durbin-Watson test: The result of the DW test reveal that the p value of 0 and an auto correction of 0.4964756.

1. Model: fit3

Area is the Response variable.

Predictor variables are FFMC, DMC, DC, ISI

The model accounts for 0.8% of the variance as can be seen from the results of the summary.

Testing the Assumptions:

Normality: The points do not fall on a 45degree line in the qq plot hence normality assumption has not been followed.

Independence of errors: The predictor variables have significant connection to each other and therefore the independence of errors assumption does not hold true in this case.

Linearity: We cannot see a linear relationship in the Residuals vs fitted graph and hence the linearity assumption has been violated.

Homoscedasticity: From the spread level graph we see that we have not followed this assumption

QQplot results: The plot the normality assumption has not been followed as there are many points that do not fall on a straight line.

Durbin-Watson test: The result of the DW test reveal that the p value of 0 and an auto correction of 0.4952255

.

1. Identify unusual observations and take corrective measures

CODE:

#for fit1

durbinWatsonTest(fit1)

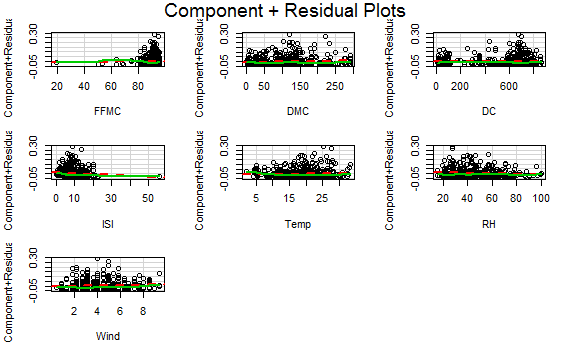
crPlots(fit1)

RESULT:

lag Autocorrelation D-W Statistic p-value

1 0.498749 1.001249 0

Alternative hypothesis: rho != 0



CODE:

#for fit2

durbinWatsonTest(fit2)

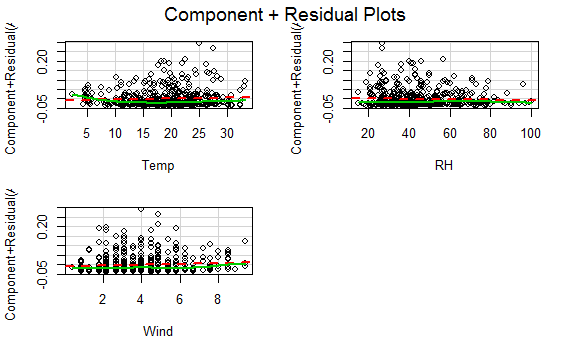
crPlots(fit2)

RESULT:

lag Autocorrelation D-W Statistic p-value

1 0.4964756 1.005805 0

Alternative hypothesis: rho != 0



CODE:

#for fit3

durbinWatsonTest(fit3)

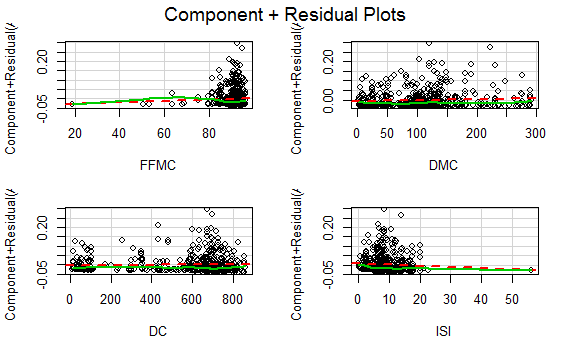
crPlots(fit3)

RESULT:

lag Autocorrelation D-W Statistic p-value

1 0.4952255 1.008615 0

Alternative hypothesis: rho != 0



# Evaluating homoscedasticity

CODE:

#for fit1

ncvTest(fit1)

par(mfrow= c(1,1))

spreadLevelPlot(fit1)

RESULT:

> ncvTest(fit1)

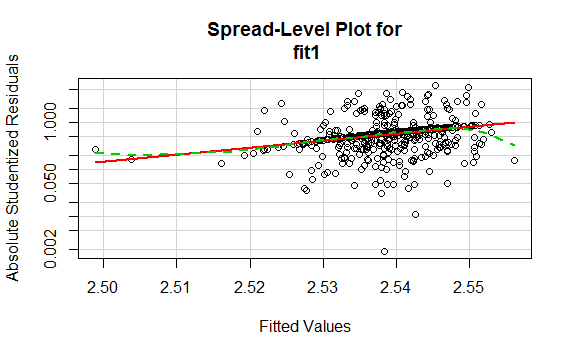
Non-constant Variance Score Test

Variance formula: ~ fitted.values

Chisquare = 38.95867 Df = 1 p = 4.328728e-10

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

> spreadLevelPlot(fit1)



CODE:

#for fit2

ncvTest(fit2)

spreadLevelPlot(fit2)

RESULT:

> #for fit2

> ncvTest(fit2)

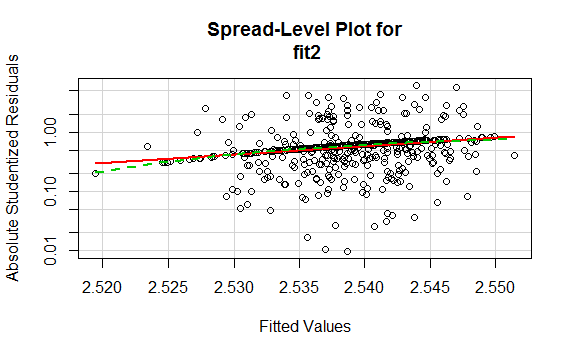
Non-constant Variance Score Test

Variance formula: ~ fitted.values

Chisquare = 25.90837 Df = 1 p = 3.580126e-07

> spreadLevelPlot(fit2)

Suggested power transformation: -83.68075



CODE:

#for P3

ncvTest(fit3)

spreadLevelPlot(fit3)

RESULT:

> #for P3

> ncvTest(fit3)

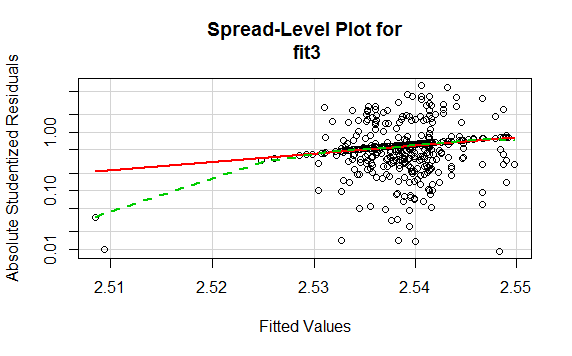
Non-constant Variance Score Test

Variance formula: ~ fitted.values

Chisquare = 20.90934 Df = 1 p = 4.815413e-06

> spreadLevelPlot(fit3)

Suggested power transformation: -82.77377



1. Selecting the best regression model

CODE:

library(gvlma)

# for fit1

gvmodelfit1 <- gvlma(fit1)

summary(gvmodelfit1)

par(ask=F)

par(mar = rep(2, 4))

plot(gvmodelfit1)

RESULT:

> # for fit1

> gvmodelfit1 <- gvlma(fit1)

> summary(gvmodelfit1)

Call:

lm(formula = Area ~ FFMC + DMC + DC + ISI + Temp + RH + Wind,

data = FFdataNN)

Residuals:

Min 1Q Median 3Q Max

-0.04294 -0.02922 -0.02021 0.01109 0.28802

Coefficients:

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

(Intercept) 2.514e+00 4.688e-02 53.625 <2e-16 \*\*\*

FFMC 1.941e-04 5.016e-04 0.387 0.6990

DMC 5.590e-05 5.076e-05 1.101 0.2713

DC 3.032e-06 1.237e-05 0.245 0.8064

ISI -8.587e-04 5.867e-04 -1.464 0.1439

Temp 3.043e-04 5.924e-04 0.514 0.6077

RH -1.809e-04 1.773e-04 -1.020 0.3081

Wind 2.300e-03 1.265e-03 1.818 0.0696 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04838 on 509 degrees of freedom

Multiple R-squared: 0.01903, Adjusted R-squared: 0.005539

F-statistic: 1.411 on 7 and 509 DF, p-value: 0.1986

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

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

Level of Significance = 0.05

Call:

gvlma(x = fit1)

Value p-value Decision

Global Stat 1493.5112 0.000e+00 Assumptions NOT satisfied!

Skewness 487.4149 0.000e+00 Assumptions NOT satisfied!

Kurtosis 986.4684 0.000e+00 Assumptions NOT satisfied!

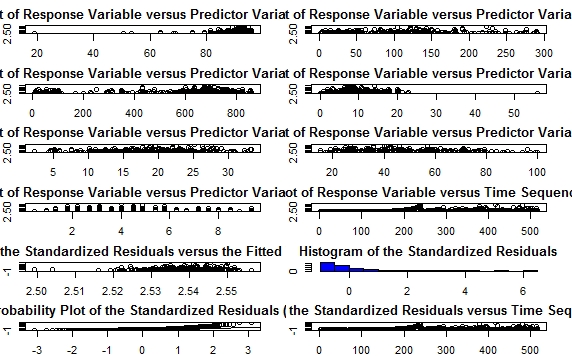
Link Function 0.6117 4.342e-01 Assumptions acceptable.

Heteroscedasticity 19.0163 1.296e-05 Assumptions NOT satisfied!

> par(ask=F)

> par(mar = rep(2, 4))

> plot(gvmodelfit1)



CODE:

# for fit2

gvmodelfit2 <- gvlma(fit2)

summary(gvmodelfit2)

par(ask=F)

par(mar = rep(2, 4))

plot(gvmodelfit2)

RESULT:

> # for fit2

> gvmodelfit2 <- gvlma(fit2)

> summary(gvmodelfit2)

Call:

lm(formula = Area ~ Temp + RH + Wind, data = FFdataNN)

Residuals:

Min 1Q Median 3Q Max

-0.04011 -0.02899 -0.02145 0.01135 0.28945

Coefficients:

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

(Intercept) 2.5252314 0.0150980 167.256 <2e-16 \*\*\*

Temp 0.0005574 0.0004432 1.258 0.209

RH -0.0001038 0.0001540 -0.674 0.500

Wind 0.0018916 0.0012233 1.546 0.123

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04839 on 513 degrees of freedom

Multiple R-squared: 0.01055, Adjusted R-squared: 0.004767

F-statistic: 1.824 on 3 and 513 DF, p-value: 0.1418

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

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

Level of Significance = 0.05

Call:

gvlma(x = fit2)

Value p-value Decision

Global Stat 1.551e+03 0.000e+00 Assumptions NOT satisfied!

Skewness 5.010e+02 0.000e+00 Assumptions NOT satisfied!

Kurtosis 1.033e+03 0.000e+00 Assumptions NOT satisfied!

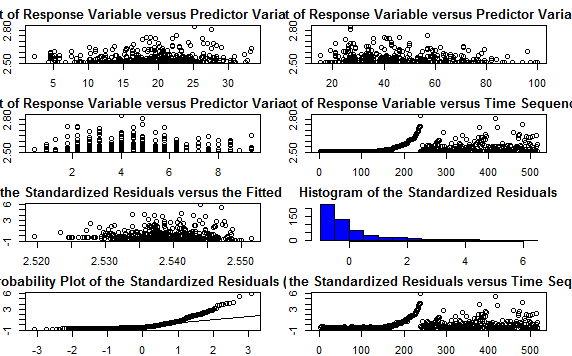
Link Function 2.074e-03 9.637e-01 Assumptions acceptable.

Heteroscedasticity 1.698e+01 3.781e-05 Assumptions NOT satisfied!

> par(ask=F)

> par(mar = rep(2, 4))

> plot(gvmodelfit2)



CODE:

# for fit3

gvmodelfit3 <- gvlma(fit3)

summary(gvmodelfit3)

par(ask=F)

par(mar = rep(2, 4))

plot(gvmodelfit3)

RESULT:

> # for fit3

> gvmodelfit3 <- gvlma(fit3)

> summary(gvmodelfit3)

Call:

lm(formula = Area ~ FFMC + DMC + DC + ISI, data = FFdataNN)

Residuals:

Min 1Q Median 3Q Max

-0.03938 -0.02886 -0.02184 0.01109 0.29285

Coefficients:

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

(Intercept) 2.500e+00 3.994e-02 62.600 <2e-16 \*\*\*

FFMC 4.122e-04 4.773e-04 0.864 0.388

DMC 4.782e-05 4.713e-05 1.015 0.311

DC 2.476e-06 1.184e-05 0.209 0.834

ISI -6.225e-04 5.578e-04 -1.116 0.265

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0485 on 512 degrees of freedom

Multiple R-squared: 0.008006, Adjusted R-squared: 0.0002558

F-statistic: 1.033 on 4 and 512 DF, p-value: 0.3896

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

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

Level of Significance = 0.05

Call:

gvlma(x = fit3)

Value p-value Decision

Global Stat 1.583e+03 0.000e+00 Assumptions NOT satisfied!

Skewness 5.062e+02 0.000e+00 Assumptions NOT satisfied!

Kurtosis 1.056e+03 0.000e+00 Assumptions NOT satisfied!

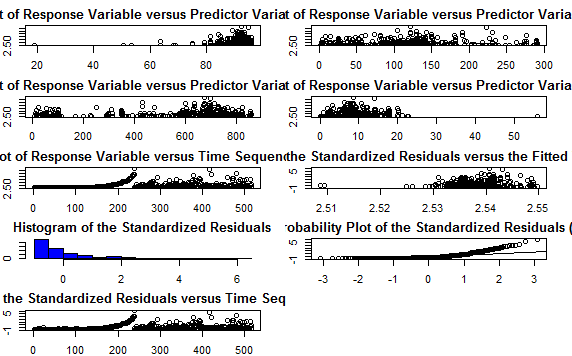
Link Function 2.281e-02 8.799e-01 Assumptions acceptable.

Heteroscedasticity 2.074e+01 5.267e-06 Assumptions NOT satisfied!

> par(ask=F)

> par(mar = rep(2, 4))

> plot(gvmodelfit3)



# We test the model using multicollinearity

CODE: sqrt(vif(fit1))>2

RESULT: > sqrt(vif(fit1))>2

FFMC DMC DC ISI Temp RH Wind

FALSE FALSE FALSE FALSE FALSE FALSE FALSE

CODE: sqrt(vif(fit2))>2

RESULT:

> sqrt(vif(fit2))>2

Temp RH Wind

FALSE FALSE FALSE

CODE: sqrt(vif(fit3))>2

RESULT:

> sqrt(vif(fit3))>2

FFMC DMC DC ISI

FALSE FALSE FALSE FALSE

# We test the model using the outlier test

CODE: outlierTest(fit1)

RESULT: > outlierTest(fit1)

rstudent unadjusted p-value Bonferonni p

239 6.185627 1.2735e-09 6.5839e-07

416 5.598282 3.5476e-08 1.8341e-05

480 4.463636 9.9339e-06 5.1358e-03

237 4.136502 4.1265e-05 2.1334e-02

238 4.130399 4.2339e-05 2.1889e-02

CODE: outlierTest(fit2)

RESULT: > outlierTest(fit2)

rstudent unadjusted p-value Bonferonni p

239 6.211540 1.0868e-09 5.6189e-07

416 5.640357 2.8101e-08 1.4528e-05

480 4.438525 1.1100e-05 5.7386e-03

238 4.188565 3.3057e-05 1.7091e-02

237 4.149903 3.8949e-05 2.0137e-02

CODE: outlierTest(fit3)

RESULT: > outlierTest(fit3)

rstudent unadjusted p-value Bonferonni p

239 6.268286 7.7691e-10 4.0166e-07

416 5.721195 1.8033e-08 9.3231e-06

480 4.462788 9.9599e-06 5.1493e-03

238 4.053346 5.8357e-05 3.0170e-02

237 3.992799 7.4875e-05 3.8710e-02

236 3.980716 7.8663e-05 4.0669e-02

# ANOVA

CODE: anova(fit1,fit2)

RESULT: > anova(fit1,fit2)

Analysis of Variance Table

Model 1: Area ~ FFMC + DMC + DC + ISI + Temp + RH + Wind

Model 2: Area ~ Temp + RH + Wind

Res.Df RSS Df Sum of Sq F Pr(>F)

1 509 1.1912

2 513 1.2015 -4 -0.010293 1.0996 0.356

CODE: anova(fit2,fit3)

RESULT: > anova(fit2,fit3)

Analysis of Variance Table

Model 1: Area ~ Temp + RH + Wind

Model 2: Area ~ FFMC + DMC + DC + ISI

Res.Df RSS Df Sum of Sq F Pr(>F)

1 513 1.2015

2 512 1.2046 1 -0.0030932

CODE: AIC(fit2,fit3)

RESULT: > AIC(fit2,fit3)

df AIC

fit2 5 -1658.161

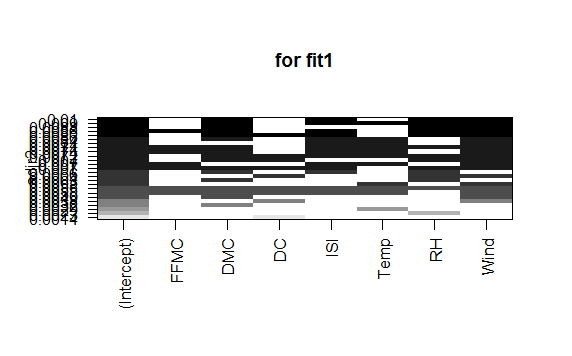
fit3 6 -1654.832

# 6 Fine tuning the slection of predictor variables

CODE: library(leaps)

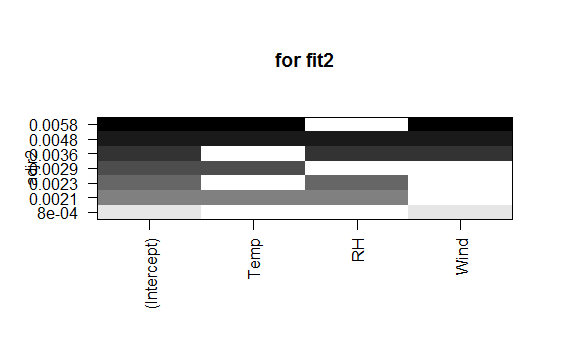
leapsfit1 <- regsubsets(Area~FFMC+DMC+DC+ISI+Temp+RH+Wind, data= FFdataNN, nbest = 4)

plot(leapsfit1,scale = "adjr2", main = "for fit1")

RESULT: 

CODE: leapsfit2 <- regsubsets(Area~Temp+RH+Wind, data= newdata, nbest = 4)

plot(leapsfit2,scale = "adjr2", main = "for fit2")

RESULT: 

CODE: leapsfit3 <- regsubsets(Area~FFMC+DMC+DC+ISI, data= newdata, nbest = 4)

plot(leapsfit3,scale = "adjr2", main = "for fit3")

RESULT: