Homework 4

Group 2
Harshish Vataliya
Sagar Pancholi

617-982-4228
857-763-8362

vataliya.h@husky.neu.edu
pancholi.sa@husky.neu.edu

Percentage of Effort Contributed by Student 1: 70%

Percentage of Effort Contributed by Student 2: 30%

Signature of Student 1: 

Signature of Student 2: 

Submission Date: 11/05/2019

Problem 1

Loading of Required Packages

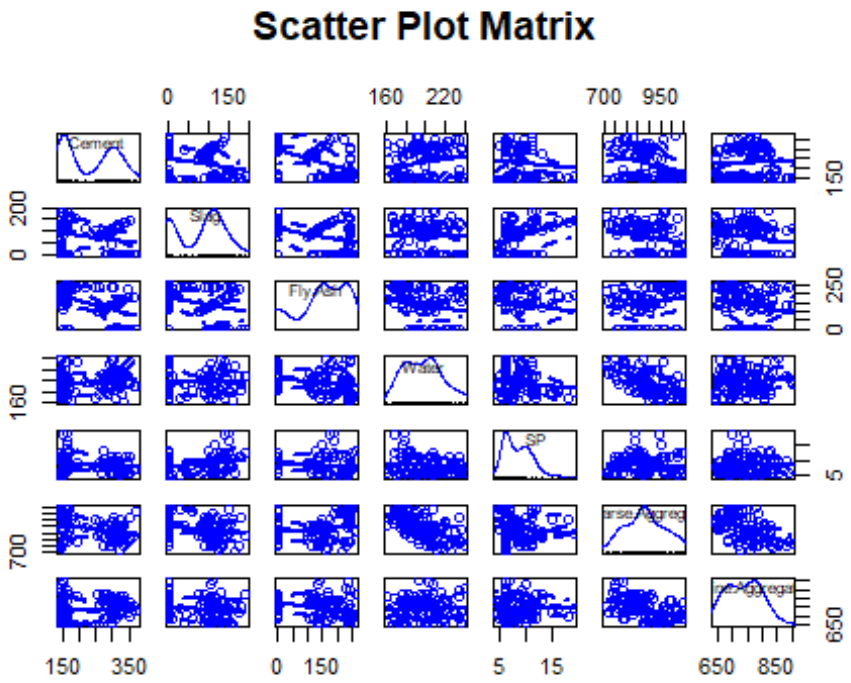
```
library(gvlma)
library(MASS)
library(readxl)
library(car)

## Loading required package: carData

library(carData)
library(ggplot2)
library(lattice)
library(caret)
```

Loadind of Dataset

```
cdata <- read_excel("Concrete Slump Test Data.xlsx")
cor(cdata[,c(2,3,4,5,6,7,8)])

##                       Cement        Slag    Fly Ash       Water
## Cement            1.00000000 -0.24355253 -0.4865353  0.22109124
## Slag             -0.24355253  1.00000000 -0.3226191 -0.02677464
## Fly Ash          -0.48653529 -0.32261907  1.0000000 -0.24132061
## Water             0.22109124 -0.02677464 -0.2413206  1.00000000
## SP               -0.10638679  0.30650431 -0.1435080 -0.15545589
## Coarse Aggregate -0.30985683 -0.22379245  0.1726200 -0.60220129
## Fine Aggregate    0.05695887 -0.18352199 -0.2828543  0.11459095
##                           SP Coarse Aggregate Fine Aggregate
## Cement           -0.10638679       -0.3098568     0.05695887
## Slag              0.30650431       -0.2237924    -0.18352199
## Fly Ash          -0.14350798        0.1726200    -0.28285429
## Water            -0.15545589       -0.6022013     0.11459095
## SP                1.00000000       -0.1041594     0.05829047
## Coarse Aggregate -0.10415943        1.0000000    -0.48853677
## Fine Aggregate    0.05829047       -0.4885368     1.00000000
```

Scatter Plot Matrix

```
scatterplotMatrix(cdata[,c(2,3,4,5,6,7,8)], main="Scatter Plot Matrix")
```



Initial Set of predictor variables: Cement, Slag, Fly Ash, Water, SP, Coarse Aggregate, Fine Aggregate

Initial Response Variables Slump, Slump Flow, 28 Day Compressive Strength

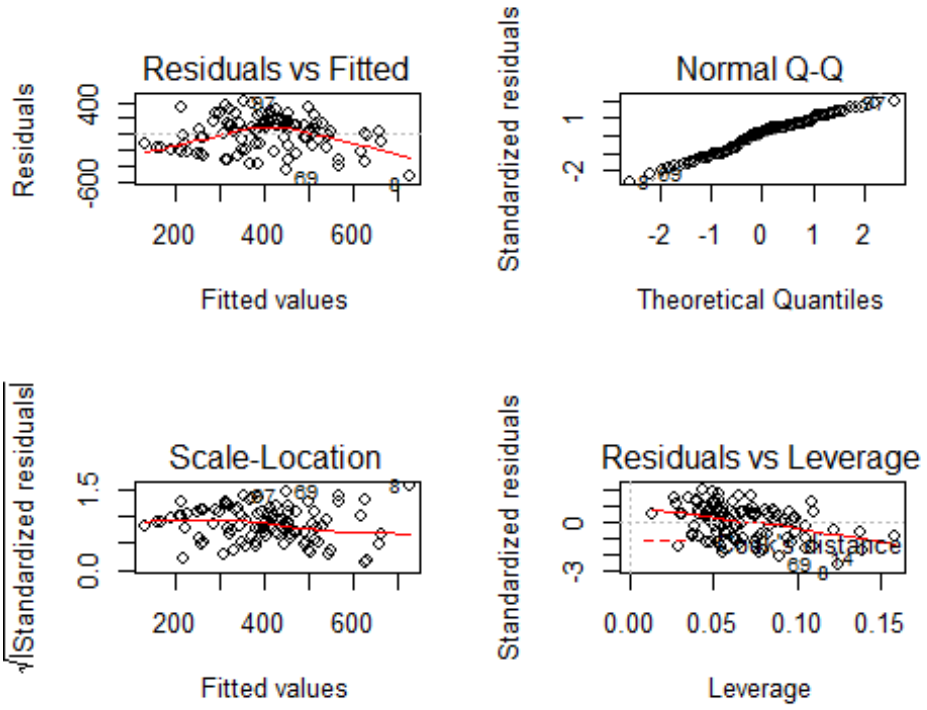
Potential Regression Model for Slump:

```
cdata1 <- as.data.frame(cdata[,c("No","Cement","Slag","Fly
Ash","Water","SP","Coarse Aggregate","Fine Aggregate","Slump","Slump
Flow","28-day Compressive Strength")])
fit1 <- lm( Slump^2~cdata1$Slag+cdata1$`Fly
Ash`+cdata1$Water+cdata1$SP+cdata1$`Coarse Aggregate`+cdata1$`Fine
Aggregate`, data = cdata1)
summary(fit1)

##
## Call:
## lm(formula = Slump^2 ~ cdata1$Slag + cdata1$`Fly Ash` + cdata1$Water +
##     cdata1$SP + cdata1$`Coarse Aggregate` + cdata1$`Fine Aggregate`,
##     data = cdata1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -517.00 -173.38   37.32  137.99  402.47
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                -837.23624 1017.49177  -0.823  0.41264
## cdata1$Slag                  -0.58301    0.48679  -1.198  0.23400
## cdata1$`Fly Ash`             -0.01148    0.31874  -0.036  0.97135
## cdata1$Water                  4.96753    1.64799   3.014  0.00329 **
## cdata1$SP                    -7.35215    8.20228  -0.896  0.37231
## cdata1$`Coarse Aggregate`     0.02585    0.44369   0.058  0.95366
## cdata1$`Fine Aggregate`       0.46847    0.50910   0.920  0.35978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 214 on 96 degrees of freedom
## Multiple R-squared:  0.2536, Adjusted R-squared:  0.2069
## F-statistic: 5.436 on 6 and 96 DF,  p-value: 7.102e-05
```

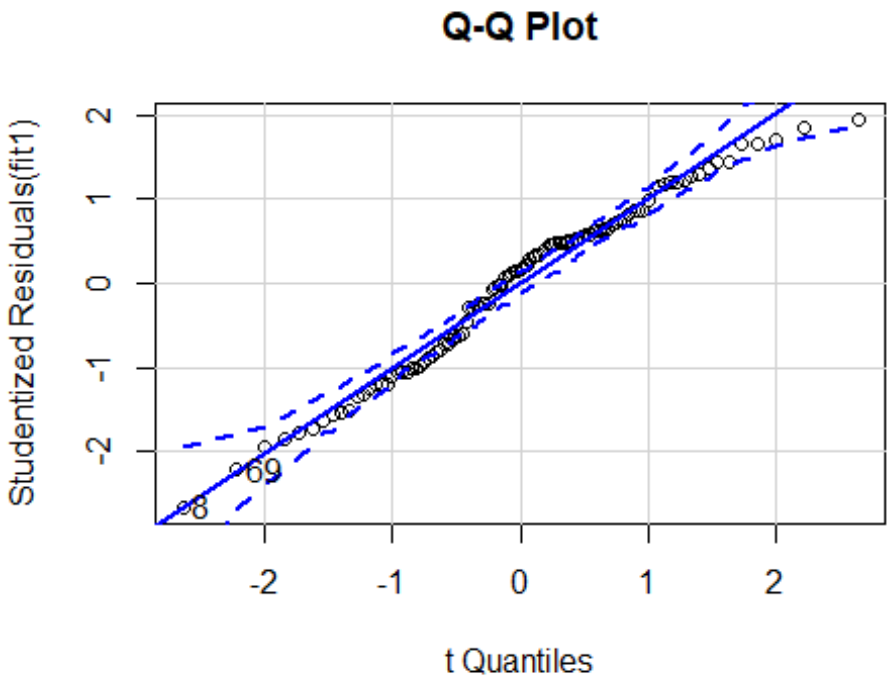
Performance Diagnostics using Typical Approach

```
par(mfrow=c(2,2))
plot(fit1)
```



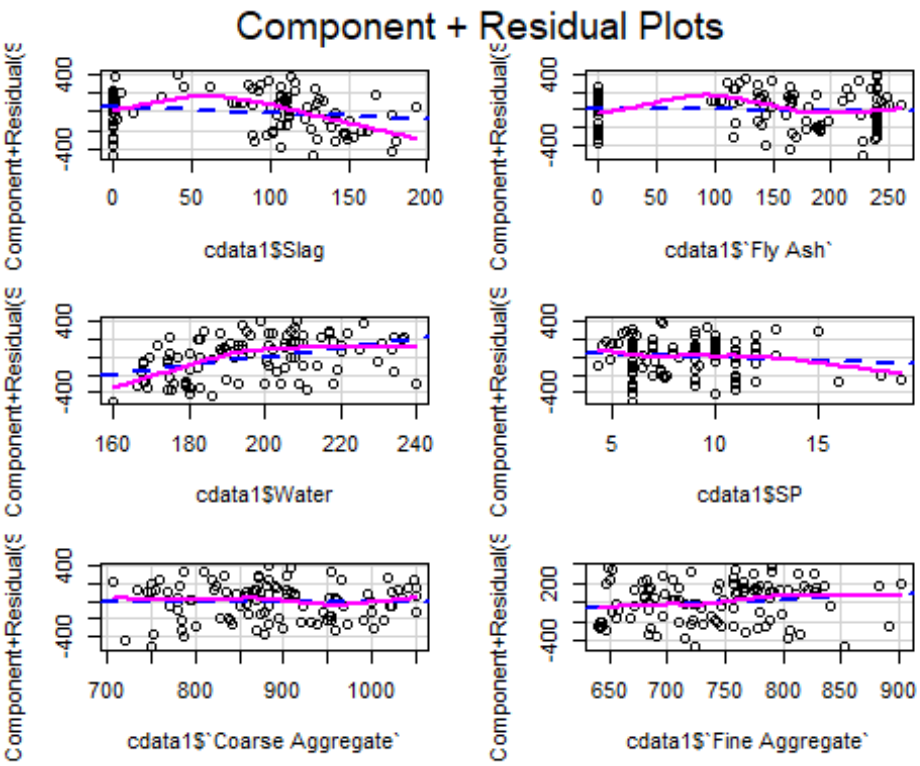
Performance Diagnostics using Enhanced Approach

```
#Normality#
par(mfrow=c(1,1))
qqPlot(fit1, labels=row.names(cdata1), id.method="identify", simulate=T,
main="Q-Q Plot")
```

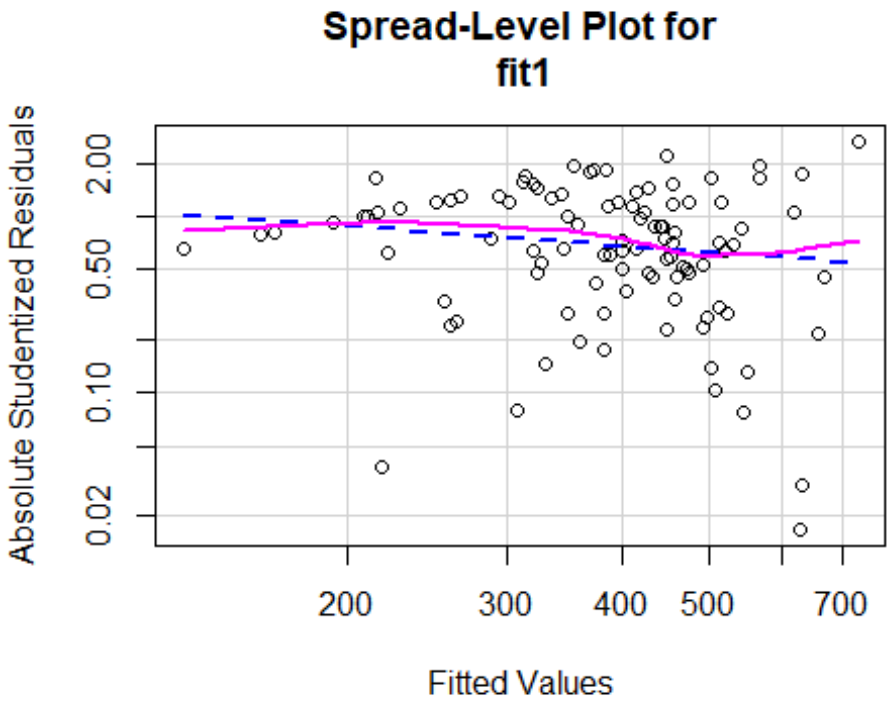



```
## [1]  8 69
```

```
#Linearity#
crPlots(fit1)
```




```
#Homoskedasticity#
ncvTest(fit1)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.1794061, Df = 1, p = 0.67188

spreadLevelPlot(fit1)
```




```
##
## Suggested power transformation:  1.372681

#Global Validation#
model1 <- gvlma(fit1)
summary(model1)

##
## Call:
## lm(formula = Slump^2 ~ cdata1$Slag + cdata1$`Fly Ash` + cdata1$Water +
##     cdata1$SP + cdata1$`Coarse Aggregate` + cdata1$`Fine Aggregate`,
##     data = cdata1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -517.00 -173.38   37.32  137.99  402.47
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -837.23624 1017.49177  -0.823  0.41264
## cdata1$Slag                   -0.58301    0.48679  -1.198  0.23400
## cdata1$`Fly Ash`              -0.01148    0.31874  -0.036  0.97135
## cdata1$Water                   4.96753    1.64799   3.014  0.00329 **
## cdata1$SP                     -7.35215    8.20228  -0.896  0.37231
## cdata1$`Coarse Aggregate`      0.02585    0.44369   0.058  0.95366
## cdata1$`Fine Aggregate`        0.46847    0.50910   0.920  0.35978
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 214 on 96 degrees of freedom
## Multiple R-squared:  0.2536, Adjusted R-squared:  0.2069
## F-statistic: 5.436 on 6 and 96 DF,  p-value: 7.102e-05
##
##
## 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        25.4156 4.150e-05 Assumptions NOT satisfied!
## Skewness            1.3218 2.503e-01    Assumptions acceptable.
## Kurtosis            2.0948 1.478e-01    Assumptions acceptable.
## Link Function      21.2936 3.940e-06 Assumptions NOT satisfied!
## Heteroscedasticity  0.7053 4.010e-01    Assumptions acceptable.
```

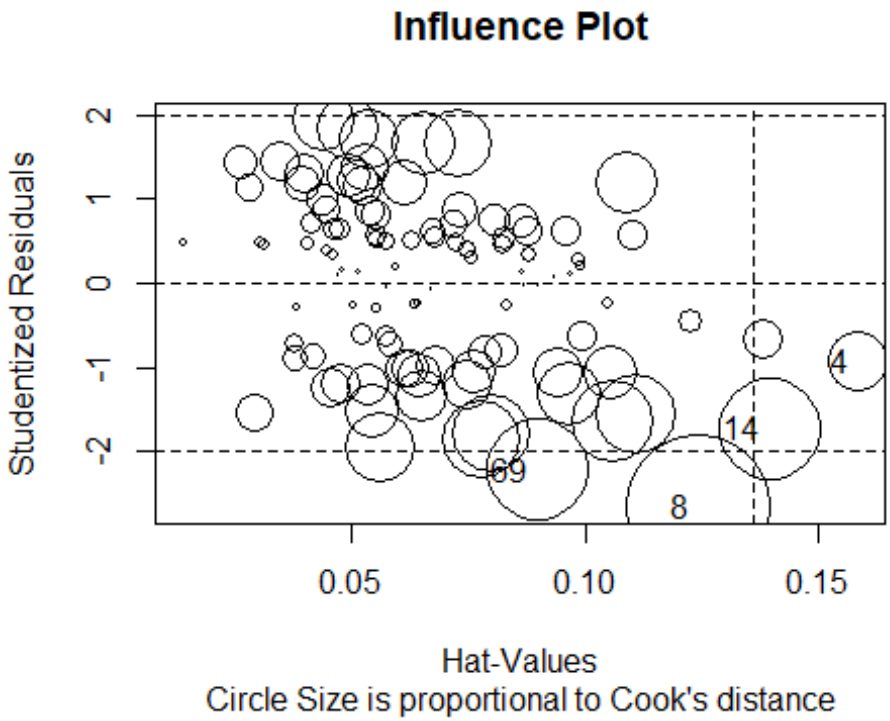
```
#Multicolinearity#
sqrt(vif(fit1))>2
```

```
##               cdata1$Slag          cdata1$`Fly Ash`
##                     FALSE                     FALSE
##              cdata1$Water                cdata1$SP
##                     FALSE                     FALSE
## cdata1$`Coarse Aggregate`   cdata1$`Fine Aggregate`
##                     FALSE                     FALSE
```

```
#Unusual Observations#
outlierTest(fit1)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 8 -2.661944          0.0091236      0.93973
```

```
#High Leverage Points#
influencePlot(fit1, main="Influence Plot", sub="Circle Size is proportional
to Cook's distance")
```




```
##       StudRes        Hat      CookD
## 4  -0.9345057 0.15840228 0.02351236
## 8  -2.6619441 0.12424111 0.13504729
## 14 -1.7332009 0.13946116 0.06812556
## 69 -2.2193390 0.08953457 0.06647703
```

Selection of Best Model Using Stepwise Regression with direction as Backward

```
step(fit1, direction = "backward")
```

```
## Start:  AIC=1112.16
## Slump^2 ~ cdata1$Slag + cdata1$`Fly Ash` + cdata1$Water + cdata1$SP +
##     cdata1$`Coarse Aggregate` + cdata1$`Fine Aggregate`
##
##                              Df Sum of Sq     RSS    AIC
## - cdata1$`Fly Ash`            1        59 4397128 1110.2
## - cdata1$`Coarse Aggregate`   1       156 4397224 1110.2
## - cdata1$SP                   1     36800 4433869 1111.0
## - cdata1$`Fine Aggregate`     1     38783 4435852 1111.1
## - cdata1$Slag                 1     65698 4462767 1111.7
## <none>                                    4397069 1112.2
## - cdata1$Water                1    416163 4813232 1119.5
##
## Step:  AIC=1110.16
## Slump^2 ~ cdata1$Slag + cdata1$Water + cdata1$SP + cdata1$`Coarse
Aggregate` +
##     cdata1$`Fine Aggregate`
```

```
##
##                              Df Sum of Sq     RSS    AIC
## - cdata1$`Coarse Aggregate`  1       293 4397422 1108.2
## - cdata1$SP                  1     36884 4434012 1109.0
## - cdata1$`Fine Aggregate`    1     54534 4451662 1109.4
## <none>                                   4397128 1110.2
## - cdata1$Slag                1     86370 4483499 1110.2
## - cdata1$Water               1    530123 4927251 1119.9
##
## Step:  AIC=1108.17
## Slump^2 ~ cdata1$Slag + cdata1$Water + cdata1$SP + cdata1$`Fine Aggregate`
##
##                             Df Sum of Sq     RSS    AIC
## - cdata1$SP                  1     38093 4435514 1107.0
## - cdata1$`Fine Aggregate`    1     79679 4477101 1108.0
## <none>                                   4397422 1108.2
## - cdata1$Slag                1    111444 4508866 1108.7
## - cdata1$Water               1    963702 5361124 1126.6
##
## Step:  AIC=1107.05
## Slump^2 ~ cdata1$Slag + cdata1$Water + cdata1$`Fine Aggregate`
##
##                             Df Sum of Sq     RSS    AIC
## - cdata1$`Fine Aggregate`    1     66112 4501626 1106.6
## <none>                                   4435514 1107.0
## - cdata1$Slag                1    176962 4612477 1109.1
## - cdata1$Water               1   1061188 5496702 1127.2
##
## Step:  AIC=1106.58
## Slump^2 ~ cdata1$Slag + cdata1$Water
##
##                Df Sum of Sq     RSS    AIC
## <none>                       4501626 1106.6
## - cdata1$Slag   1    225915 4727541 1109.6
## - cdata1$Water  1   1135280 5636906 1127.7

##
## Call:
## lm(formula = Slump^2 ~ cdata1$Slag + cdata1$Water, data = cdata1)
##
## Coefficients:
##  (Intercept)   cdata1$Slag  cdata1$Water
##    -567.4150       -0.7787        5.2225
```

Applying corrective measures by removing outliers to attain normality. Based on the Step AIC approach reformulating the model with the given attributes and removing the insignificant attributes

```
cdata1 <- cdata1[-c(8,69),]
modfit1 <- lm( cdata1$Slump^2~cdata1$Slag+cdata1$Water)
summary(modfit1)
```

```
##
## Call:
## lm(formula = cdata1$Slump^2 ~ cdata1$Slag + cdata1$Water)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -428.96 -163.41   14.03  148.67  364.96
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -688.2590   204.0320  -3.373  0.00106 **
## cdata1$Slag    -0.8219     0.3373  -2.437  0.01662 *
## cdata1$Water    5.9001     1.0231   5.767 9.42e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 203.6 on 98 degrees of freedom
## Multiple R-squared:  0.2865, Adjusted R-squared:  0.2719
## F-statistic: 19.67 on 2 and 98 DF,  p-value: 6.559e-08
```

Potential Regression Model for Slump Flow:

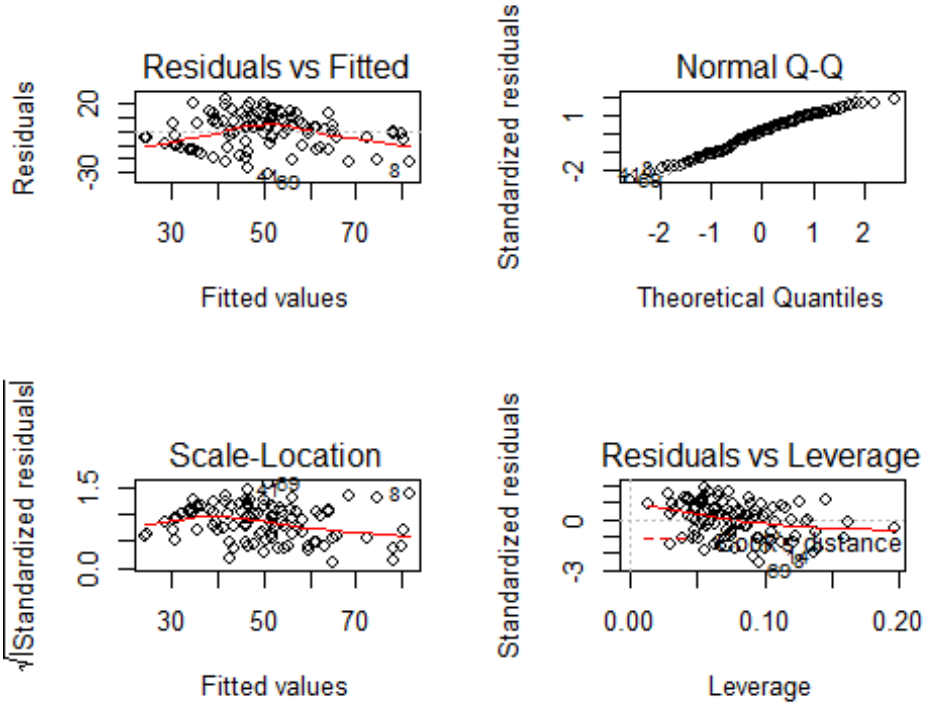
```
cdata1 <- as.data.frame(cdata[,c("No","Cement","Slag","Fly
Ash","Water","SP","Coarse Aggregate","Fine Aggregate","Slump","Slump
Flow","28-day Compressive Strength")])
fit2 <- lm( cdata1$`Slump Flow`~cdata1$Cement+cdata1$Slag+cdata1$`Fly
Ash`+cdata1$Water+cdata1$SP+cdata1$`Coarse Aggregate`+cdata1$`Fine
Aggregate`)
summary(fit2)
```

```
##
## Call:
## lm(formula = cdata1$`Slump Flow` ~ cdata1$Cement + cdata1$Slag +
##     cdata1$`Fly Ash` + cdata1$Water + cdata1$SP + cdata1$`Coarse
Aggregate` +
##     cdata1$`Fine Aggregate`)
##
## 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
## cdata1$Cement               0.05364    0.11236   0.477   0.6342
## cdata1$Slag                -0.00569    0.15638  -0.036   0.9710
## cdata1$`Fly Ash`            0.06115    0.11402   0.536   0.5930
```

```
## cdata1$Water                  0.73180    0.35282   2.074   0.0408 *
## cdata1$SP                     0.29833    0.66263   0.450   0.6536
## cdata1$`Coarse Aggregate`     0.07366    0.13510   0.545   0.5869
## cdata1$`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
```

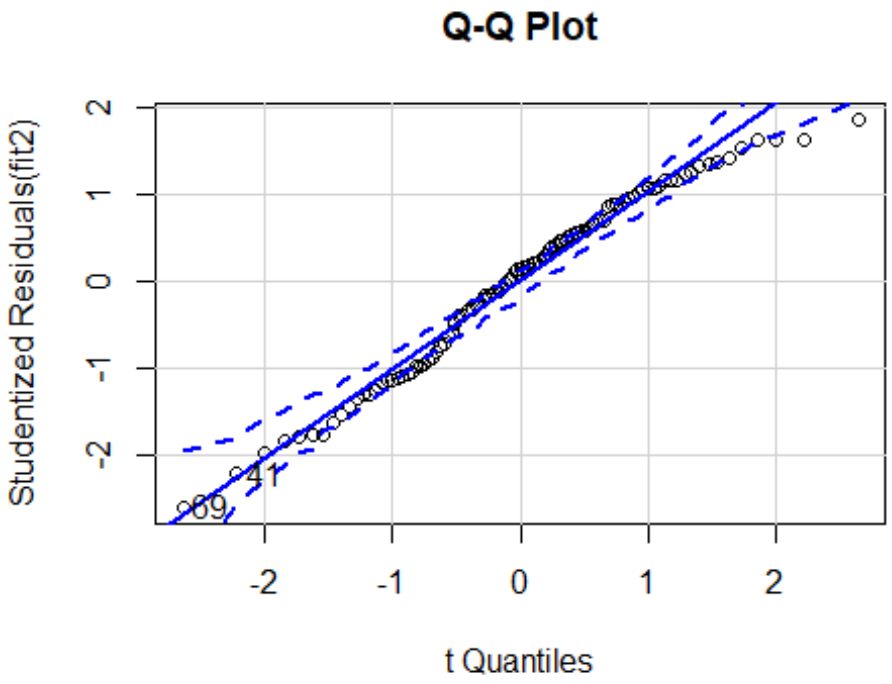
Performance Diagnostics using Typical Approach

```
par(mfrow=c(2,2))
plot(fit2)
```

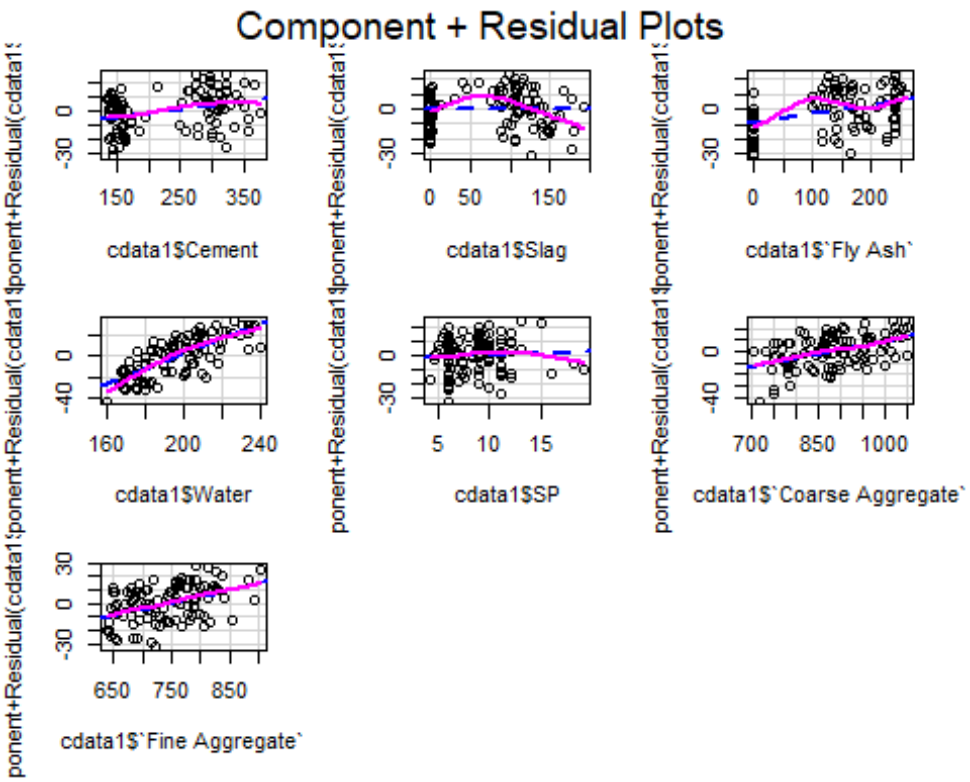


Performance Diagnostics using Enhanced Approach

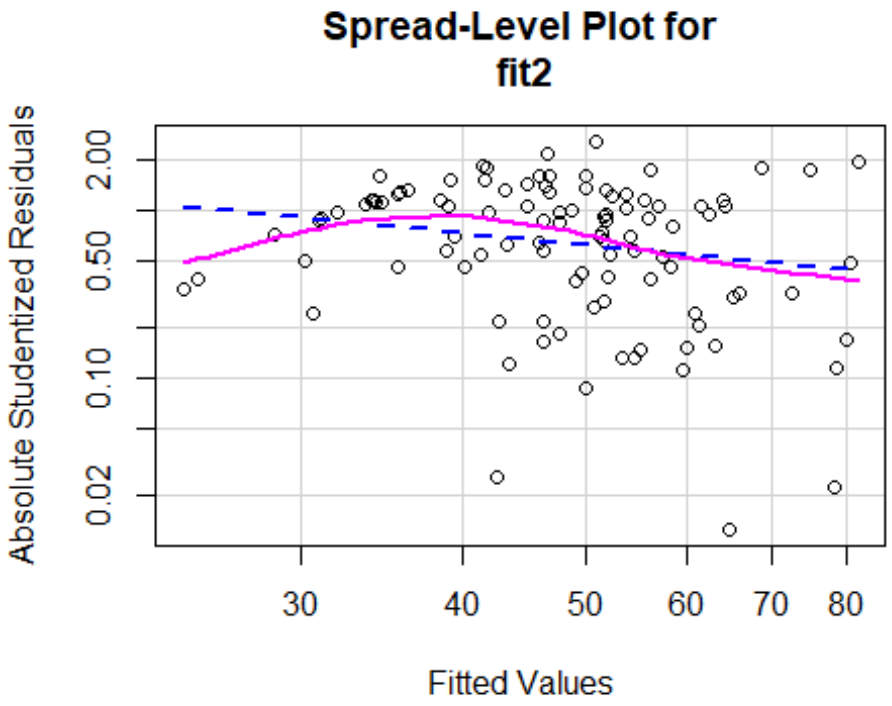
```
#Normality#
par(mfrow=c(1,1))
qqPlot(fit2, labels=row.names(cdata1), id.method="identify", simulate=T,
main="Q-Q Plot")
```




```
## [1] 41 69

#Linearity#
crPlots(fit2)
```




```
#Homoskedasticity#
ncvTest(fit2)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.2327094, Df = 1, p = 0.62952

spreadLevelPlot(fit2)
```




```
##
## Suggested power transformation:  1.743362
```

```
#Global Validation#
model2 <- gvlma(fit2)
summary(model2)
```

```
##
## Call:
## lm(formula = cdata1$`Slump Flow` ~ cdata1$Cement + cdata1$Slag +
##     cdata1$`Fly Ash` + cdata1$Water + cdata1$SP + cdata1$`Coarse
Aggregate` +
##     cdata1$`Fine Aggregate`)
##
## 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
## cdata1$Cement                 0.05364    0.11236   0.477   0.6342
## cdata1$Slag                  -0.00569    0.15638  -0.036   0.9710
## cdata1$`Fly Ash`              0.06115    0.11402   0.536   0.5930
## cdata1$Water                  0.73180    0.35282   2.074   0.0408 *
## cdata1$SP                     0.29833    0.66263   0.450   0.6536
## cdata1$`Coarse Aggregate`     0.07366    0.13510   0.545   0.5869
```

```
## cdata1$`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.
```

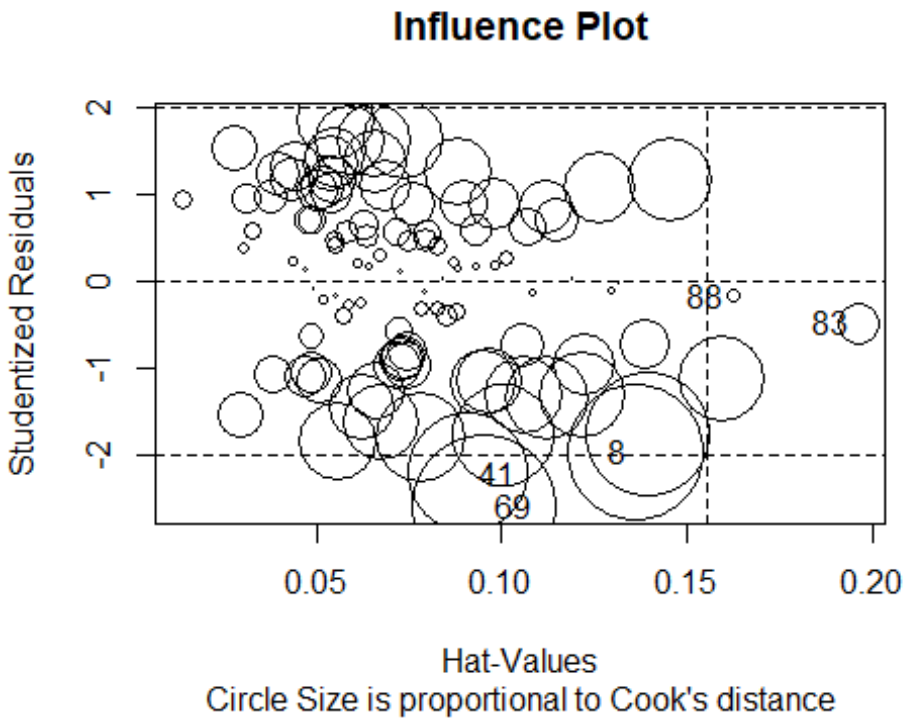
```
#Multicolinearity#
sqrt(vif(fit2))>2
```

```
##             cdata1$Cement                cdata1$Slag
##                      TRUE                       TRUE
##          cdata1$`Fly Ash`               cdata1$Water
##                      TRUE                       TRUE
##                cdata1$SP cdata1$`Coarse Aggregate`
##                     FALSE                       TRUE
##   cdata1$`Fine Aggregate`
##                      TRUE
```

```
#Unusual Observations#
outlierTest(fit2)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 69 -2.603738           0.010717           NA
```

```
#High Leverage Points#
influencePlot(fit2, main="Influence Plot", sub="Circle Size is proportional
to Cook's distance")
```




```
##       StudRes        Hat        CookD
## 8  -1.9682860 0.13618195 0.0741036039
## 41 -2.2181634 0.09091319 0.0590686474
## 69 -2.6037375 0.09545758 0.0843019258
## 83 -0.4860173 0.19610512 0.0072612082
## 88 -0.1726772 0.16233785 0.0007297749
```

Selection of Best Model Using Stepwise Regression with direction as Backward

```
step(fit2, direction = "backward")
```

```
## Start:  AIC=533.56
## cdata1$`Slump Flow` ~ cdata1$Cement + cdata1$Slag + cdata1$`Fly Ash` +
##     cdata1$Water + cdata1$SP + cdata1$`Coarse Aggregate` + cdata1$`Fine
Aggregate`
##
##                              Df Sum of Sq   RSS    AIC
## - cdata1$Slag                 1      0.22 15672 531.56
## - cdata1$SP                   1     33.44 15705 531.78
## - cdata1$Cement               1     37.60 15709 531.81
## - cdata1$`Fly Ash`            1     47.45 15719 531.87
## - cdata1$`Coarse Aggregate`   1     49.04 15720 531.88
## - cdata1$`Fine Aggregate`     1     72.40 15744 532.03
## <none>                                   15671 533.56
## - cdata1$Water                1    709.69 16381 536.12
##
## Step:  AIC=531.56
```

```
## cdata1$`Slump Flow` ~ cdata1$Cement + cdata1$`Fly Ash` + cdata1$Water +
##     cdata1$SP + cdata1$`Coarse Aggregate` + cdata1$`Fine Aggregate`
##
##                             Df Sum of Sq   RSS    AIC
## - cdata1$SP                  1      62.1 15734 529.97
## <none>                                   15672 531.56
## - cdata1$Cement              1    1244.7 16916 537.43
## - cdata1$`Coarse Aggregate`  1    1679.4 17351 540.05
## - cdata1$`Fly Ash`           1    1759.2 17431 540.52
## - cdata1$`Fine Aggregate`    1    2292.3 17964 543.62
## - cdata1$Water               1   10877.0 26549 583.86
##
## Step:  AIC=529.97
## cdata1$`Slump Flow` ~ cdata1$Cement + cdata1$`Fly Ash` + cdata1$Water +
##     cdata1$`Coarse Aggregate` + cdata1$`Fine Aggregate`
##
##                             Df Sum of Sq   RSS    AIC
## <none>                                   15734 529.97
## - cdata1$Cement              1    1193.1 16927 535.50
## - cdata1$`Coarse Aggregate`  1    1678.8 17412 538.41
## - cdata1$`Fly Ash`           1    1746.5 17480 538.81
## - cdata1$`Fine Aggregate`    1    2237.1 17971 541.66
## - cdata1$Water               1   11947.4 27681 586.16

##
## Call:
## lm(formula = cdata1$`Slump Flow` ~ cdata1$Cement + cdata1$`Fly Ash` +
##     cdata1$Water + cdata1$`Coarse Aggregate` + cdata1$`Fine Aggregate`)
##
## Coefficients:
##               (Intercept)              cdata1$Cement
##                -249.50866                    0.05366
##          cdata1$`Fly Ash`                cdata1$Water
##                   0.06101                    0.72313
## cdata1$`Coarse Aggregate`    cdata1$`Fine Aggregate`
##                   0.07291                    0.09554
```

Applying corrective measures by removing outliers to attain normality. Based on the Step AIC approach reformulating the model with the given attributes and removing the insignificant attributes

```
cdata1 <- cdata1[-c(41,69),]
modfit2 <- lm(  cdata1$`Slump Flow` ~ cdata1$Cement + cdata1$`Fly Ash` +
cdata1$Water + cdata1$`Coarse Aggregate` + cdata1$`Fine Aggregate`)
summary(modfit2)

##
## Call:
## lm(formula = cdata1$`Slump Flow` ~ cdata1$Cement + cdata1$`Fly Ash` +
##     cdata1$Water + cdata1$`Coarse Aggregate` + cdata1$`Fine Aggregate`)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.219  -8.978   1.896   9.219  22.121
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -205.15672   47.84578  -4.288 4.33e-05 ***
## cdata1$Cement                0.04898    0.01898   2.581 0.011384 *
## cdata1$`Fly Ash`             0.06148    0.01758   3.497 0.000718 ***
## cdata1$Water                 0.68902    0.08019   8.593 1.67e-13 ***
## cdata1$`Coarse Aggregate`    0.04783    0.02251   2.125 0.036183 *
## cdata1$`Fine Aggregate`      0.07688    0.02486   3.093 0.002605 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.04 on 95 degrees of freedom
## Multiple R-squared:  0.5363, Adjusted R-squared:  0.5119
## F-statistic: 21.97 on 5 and 95 DF,  p-value: 1.447e-14
```

```
model2 <- gvlma(modfit2)
summary(model2)
```

```
##
## Call:
## lm(formula = cdata1$`Slump Flow` ~ cdata1$Cement + cdata1$`Fly Ash` +
##     cdata1$Water + cdata1$`Coarse Aggregate` + cdata1$`Fine Aggregate`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.219  -8.978   1.896   9.219  22.121
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -205.15672   47.84578  -4.288 4.33e-05 ***
## cdata1$Cement                0.04898    0.01898   2.581 0.011384 *
## cdata1$`Fly Ash`             0.06148    0.01758   3.497 0.000718 ***
## cdata1$Water                 0.68902    0.08019   8.593 1.67e-13 ***
## cdata1$`Coarse Aggregate`    0.04783    0.02251   2.125 0.036183 *
## cdata1$`Fine Aggregate`      0.07688    0.02486   3.093 0.002605 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.04 on 95 degrees of freedom
## Multiple R-squared:  0.5363, Adjusted R-squared:  0.5119
## F-statistic: 21.97 on 5 and 95 DF,  p-value: 1.447e-14
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
```

```
##
## Call:
##  gvlma(x = modfit2)
##
##                     Value   p-value                   Decision
## Global Stat        21.187 2.907e-04 Assumptions NOT satisfied!
## Skewness            1.180 2.774e-01    Assumptions acceptable.
## Kurtosis            3.466 6.265e-02    Assumptions acceptable.
## Link Function      15.310 9.125e-05 Assumptions NOT satisfied!
## Heteroscedasticity  1.232 2.670e-01    Assumptions acceptable.
```

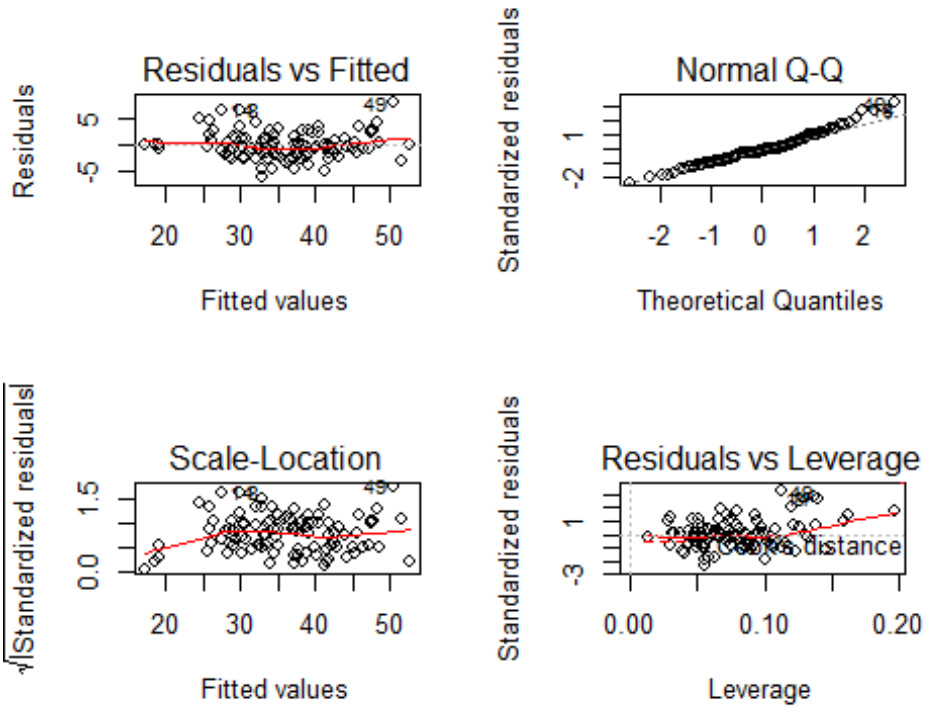
Potential Regression Model for 28 Days Compressive Strength:

```
cdata1 <- as.data.frame(cdata[,c("No","Cement","Slag","Fly
Ash","Water","SP","Coarse Aggregate","Fine Aggregate","Slump","Slump
Flow","28-day Compressive Strength")])
fit3 <- lm( cdata1$`28-day Compressive
Strength`~cdata1$Cement+cdata1$Slag+cdata1$`Fly
Ash`+cdata1$Water+cdata1$SP+cdata1$`Coarse Aggregate`+cdata1$`Fine
Aggregate`)
summary(fit3)
```

```
##
## Call:
## lm(formula = cdata1$`28-day Compressive Strength` ~ cdata1$Cement +
##     cdata1$Slag + cdata1$`Fly Ash` + cdata1$Water + cdata1$SP +
##     cdata1$`Coarse Aggregate` + cdata1$`Fine Aggregate`)
##
## 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 .
## cdata1$Cement               0.06141    0.02282   2.691  0.00842 **
## cdata1$Slag                -0.02971    0.03176  -0.935  0.35200
## cdata1$`Fly Ash`            0.05053    0.02316   2.182  0.03159 *
## cdata1$Water               -0.23270    0.07166  -3.247  0.00161 **
## cdata1$SP                   0.10315    0.13459   0.766  0.44532
## cdata1$`Coarse Aggregate`  -0.05562    0.02744  -2.027  0.04546 *
## cdata1$`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
```

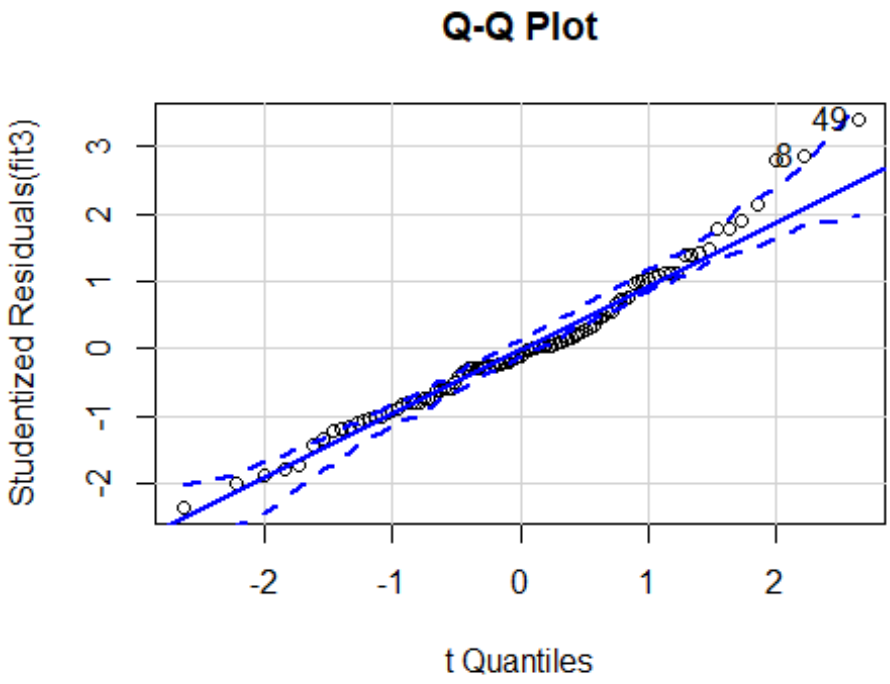
Performance Diagnostics using Typical Approach

```
par(mfrow=c(2,2))
plot(fit3)
```



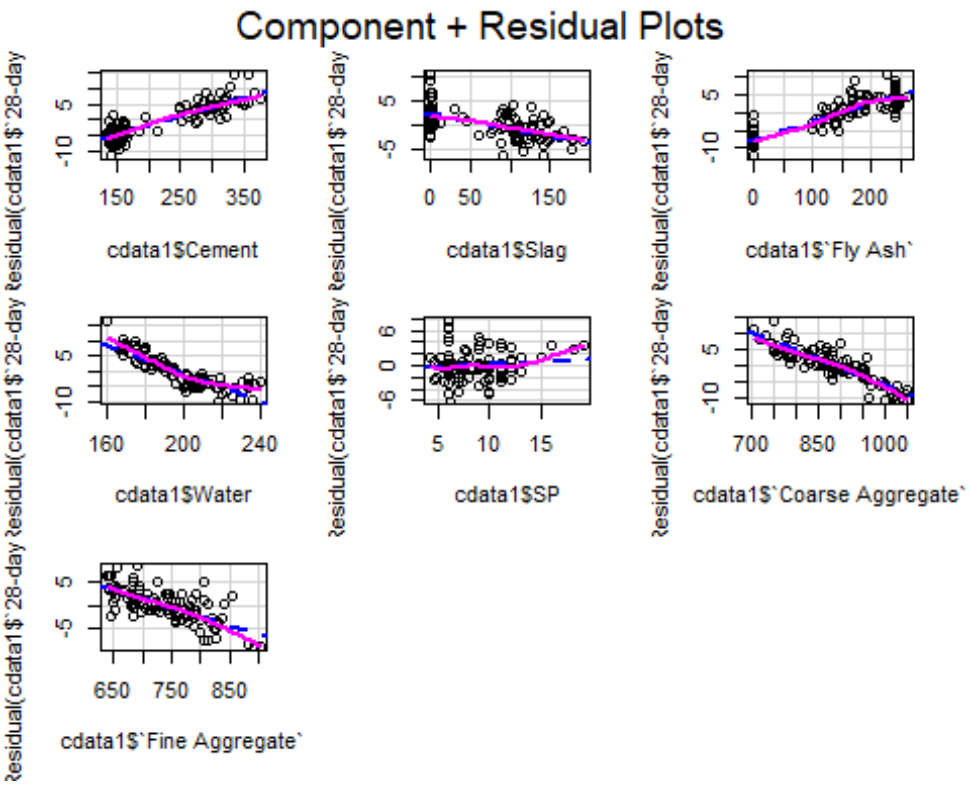
Performance Diagnostics using Enhanced Approach

```
#Normality#
par(mfrow=c(1,1))
qqPlot(fit3, labels=row.names(cdata1), id.method="identify", simulate=T,
main="Q-Q Plot")
```

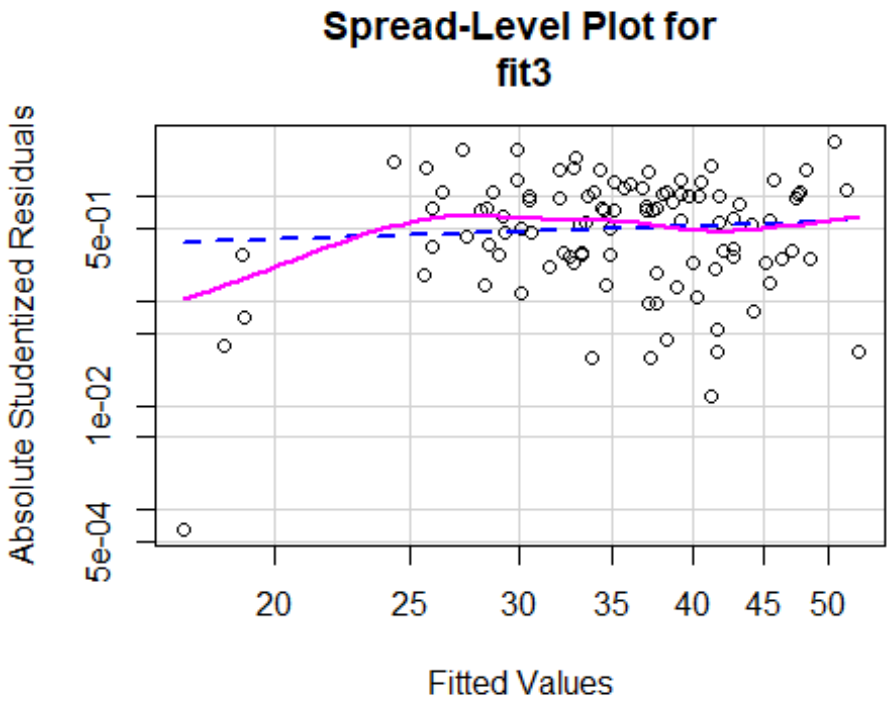



```
## [1]  8 49
```

```
#Linearity#
crPlots(fit3)
```




```
#Homoskedasticity#
ncvTest(fit3)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.07654326, Df = 1, p = 0.78204

spreadLevelPlot(fit3)
```




```
##
## Suggested power transformation:  0.5498301

#Global Validation#
model3 <- gvlma(fit3)
summary(model3)

##
## Call:
## lm(formula = cdata1$`28-day Compressive Strength` ~ cdata1$Cement +
##     cdata1$Slag + cdata1$`Fly Ash` + cdata1$Water + cdata1$SP +
##     cdata1$`Coarse Aggregate` + cdata1$`Fine Aggregate`)
##
## 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 .
## cdata1$Cement                0.06141    0.02282   2.691  0.00842 **
## cdata1$Slag                 -0.02971    0.03176  -0.935  0.35200
## cdata1$`Fly Ash`             0.05053    0.02316   2.182  0.03159 *
## cdata1$Water                -0.23270    0.07166  -3.247  0.00161 **
## cdata1$SP                    0.10315    0.13459   0.766  0.44532
## cdata1$`Coarse Aggregate`   -0.05562    0.02744  -2.027  0.04546 *
## cdata1$`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.
```

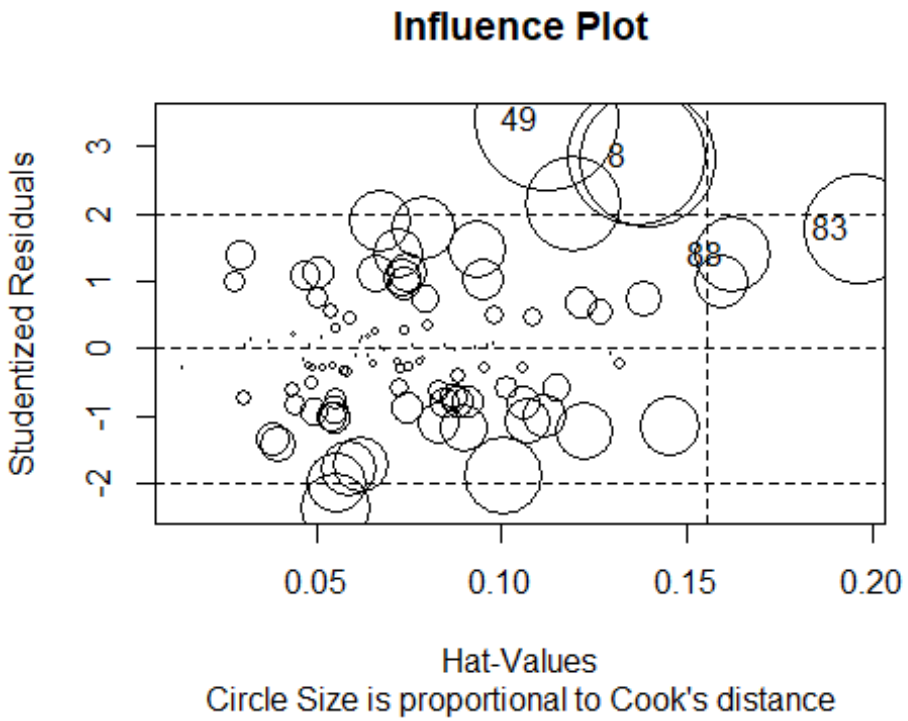
```
#Multicolinearity#
sqrt(vif(fit3))>2
```

```
##             cdata1$Cement               cdata1$Slag
##                      TRUE                      TRUE
##          cdata1$`Fly Ash`               cdata1$Water
##                      TRUE                      TRUE
##                cdata1$SP cdata1$`Coarse Aggregate`
##                     FALSE                      TRUE
##   cdata1$`Fine Aggregate`
##                      TRUE
```

```
#Unusual Observations#
outlierTest(fit3)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 49 3.407478         0.00096665     0.099565
```

```
#High Leverage Points#
influencePlot(fit3, main="Influence Plot", sub="Circle Size is proportional
to Cook's distance")
```




```
##     StudRes       Hat      CookD
## 8  2.873893 0.1361820 0.15120617
## 49 3.407478 0.1124498 0.16540816
## 83 1.795464 0.1961051 0.09605162
## 88 1.403472 0.1623378 0.04723434
```

Selection of Best Model Using Stepwise Regression with direction as Backward

```
step(fit3, direction = "backward")
```

```
## Start:  AIC=205.19
## cdata1$`28-day Compressive Strength` ~ cdata1$Cement + cdata1$Slag +
##     cdata1$`Fly Ash` + cdata1$Water + cdata1$SP + cdata1$`Coarse
Aggregate` +
##     cdata1$`Fine Aggregate`
##
##                               Df Sum of Sq    RSS    AIC
## - cdata1$SP                    1     3.997 650.48 203.83
## - cdata1$Slag                  1     5.953 652.44 204.14
## - cdata1$`Fine Aggregate`      1    12.512 659.00 205.17
## <none>                                     646.48 205.19
## - cdata1$`Coarse Aggregate`    1    27.963 674.45 207.55
## - cdata1$`Fly Ash`             1    32.394 678.88 208.23
## - cdata1$Cement                1    49.278 695.76 210.76
## - cdata1$Water                 1    71.756 718.24 214.03
##
## Step:  AIC=203.83
```

```
## cdata1$`28-day Compressive Strength` ~ cdata1$Cement + cdata1$Slag +
##     cdata1$`Fly Ash` + cdata1$Water + cdata1$`Coarse Aggregate` +
##     cdata1$`Fine Aggregate`
##
##                             Df Sum of Sq    RSS    AIC
## <none>                                    650.48 203.83
## - cdata1$Slag                1    23.123 673.60 205.43
## - cdata1$`Fly Ash`           1    34.509 684.99 207.15
## - cdata1$`Fine Aggregate`    1    41.289 691.77 208.17
## - cdata1$Cement              1    58.491 708.97 210.70
## - cdata1$`Coarse Aggregate`  1    81.493 731.97 213.99
## - cdata1$Water               1   184.744 835.23 227.58

##
## Call:
## lm(formula = cdata1$`28-day Compressive Strength` ~ cdata1$Cement +
##     cdata1$Slag + cdata1$`Fly Ash` + cdata1$Water + cdata1$`Coarse
Aggregate` +
##     cdata1$`Fine Aggregate`)
##
## Coefficients:
##               (Intercept)              cdata1$Cement
##                 177.11354                    0.04970
##               cdata1$Slag           cdata1$`Fly Ash`
##                  -0.04519                    0.03859
##              cdata1$Water  cdata1$`Coarse Aggregate`
##                  -0.27055                   -0.06986
##   cdata1$`Fine Aggregate`
##                  -0.05358
```

Applying corrective measures by removing outliers to attain normality. Based on the Step AIC approach reformulating the model with the given attributes and removing the insignificant attributes

```
cdata1 <- cdata1[-c(8,49),]
modfit3 <- lm( cdata1$`28-day Compressive Strength` ~ cdata1$Cement  +
cdata1$`Fly Ash` + cdata1$Water + cdata1$`Coarse Aggregate` +
    cdata1$`Fine Aggregate` )
summary(modfit3)
```

```
##
## Call:
## lm(formula = cdata1$`28-day Compressive Strength` ~ cdata1$Cement +
##     cdata1$`Fly Ash` + cdata1$Water + cdata1$`Coarse Aggregate` +
##     cdata1$`Fine Aggregate`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4141 -1.4201 -0.2588  1.1861  6.6110
##
## Coefficients:
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 89.304204   9.472228   9.428 2.76e-15 ***
## cdata1$Cement                0.078231   0.003848  20.333  < 2e-16 ***
## cdata1$`Fly Ash`             0.066516   0.003600  18.476  < 2e-16 ***
## cdata1$Water                -0.182945   0.016699 -10.955  < 2e-16 ***
## cdata1$`Coarse Aggregate` -0.034693   0.004323  -8.026 2.66e-12 ***
## cdata1$`Fine Aggregate`   -0.019709   0.005035  -3.914  0.00017 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.424 on 95 degrees of freedom
## Multiple R-squared:  0.903,  Adjusted R-squared:  0.8979
## F-statistic:   177 on 5 and 95 DF,  p-value: < 2.2e-16
```

```
model31 <- gvlma(modfit3)
summary(model31)
```

```
##
## Call:
## lm(formula = cdata1$`28-day Compressive Strength` ~ cdata1$Cement +
##     cdata1$`Fly Ash` + cdata1$Water + cdata1$`Coarse Aggregate` +
##     cdata1$`Fine Aggregate`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4141 -1.4201 -0.2588  1.1861  6.6110
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 89.304204   9.472228   9.428 2.76e-15 ***
## cdata1$Cement                0.078231   0.003848  20.333  < 2e-16 ***
## cdata1$`Fly Ash`             0.066516   0.003600  18.476  < 2e-16 ***
## cdata1$Water                -0.182945   0.016699 -10.955  < 2e-16 ***
## cdata1$`Coarse Aggregate` -0.034693   0.004323  -8.026 2.66e-12 ***
## cdata1$`Fine Aggregate`   -0.019709   0.005035  -3.914  0.00017 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.424 on 95 degrees of freedom
## Multiple R-squared:  0.903,  Adjusted R-squared:  0.8979
## F-statistic:   177 on 5 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 = modfit3)
##
```

```
##                      Value p-value                Decision
## Global Stat        4.3964  0.3550 Assumptions acceptable.
## Skewness           1.5744  0.2096 Assumptions acceptable.
## Kurtosis           0.2137  0.6439 Assumptions acceptable.
## Link Function      2.5685  0.1090 Assumptions acceptable.
## Heteroscedasticity 0.0399  0.8417 Assumptions acceptable.
```

Hence, it can be inferred that the given set of predictors can be used to predict the value of 28 Days Compressive Strength. For the other two response variables, some assumptions do not hold. If corrective measures are applied on that then there are chances of overfitting of model and loss of some important data.

Problem 2

Loading of Required Packages

```
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(ggplot2)
library(car)

## Loading required package: carData

library(carData)
library(lattice)
library(MASS)
library(gvlma)
library(readxl)
```

Importing Data to R

```
firedata <- read_excel("Forest Fires Data.xlsx")
str(firedata)

## Classes 'tbl_df', 'tbl' and 'data.frame':    517 obs. of  13 variables:
##  $ X    : num  7 7 7 8 8 8 8 8 8 7 ...
##  $ Y    : num  5 4 4 6 6 6 6 6 6 5 ...
##  $ Month: chr  "mar" "oct" "oct" "mar" ...
##  $ Day  : chr  "fri" "tue" "sat" "fri" ...
##  $ 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   : num  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 ...
```

```
summary(firedata)
```

```
##        X               Y          Month               Day
##  Min.   :1.000   Min.   :2.0   Length:517         Length:517
##  1st Qu.:3.000   1st Qu.:4.0   Class :character   Class :character
##  Median :4.000   Median :4.0   Mode  :character   Mode  :character
##  Mean   :4.669   Mean   :4.3
##  3rd Qu.:7.000   3rd Qu.:5.0
##  Max.   :9.000   Max.   :9.0
##       FFMC            DMC              DC             ISI
##  Min.   :18.70   Min.   :  1.1   Min.   :  7.9   Min.   : 0.000
##  1st Qu.:90.20   1st Qu.: 68.6   1st Qu.:437.7   1st Qu.: 6.500
##  Median :91.60   Median :108.3   Median :664.2   Median : 8.400
##  Mean   :90.64   Mean   :110.9   Mean   :547.9   Mean   : 9.022
##  3rd Qu.:92.90   3rd Qu.:142.4   3rd Qu.:713.9   3rd Qu.:10.800
##  Max.   :96.20   Max.   :291.3   Max.   :860.6   Max.   :56.100
##       Temp             RH              Wind            Rain
##  Min.   : 2.20   Min.   : 15.00   Min.   :0.400   Min.   :0.00000
##  1st Qu.:15.50   1st Qu.: 33.00   1st Qu.:2.700   1st Qu.:0.00000
##  Median :19.30   Median : 42.00   Median :4.000   Median :0.00000
##  Mean   :18.89   Mean   : 44.29   Mean   :4.018   Mean   :0.02166
##  3rd Qu.:22.80   3rd Qu.: 53.00   3rd Qu.:4.900   3rd Qu.:0.00000
##  Max.   :33.30   Max.   :100.00   Max.   :9.400   Max.   :6.40000
##       Area
##  Min.   :   0.00
##  1st Qu.:   0.00
##  Median :   0.52
##  Mean   :  12.85
##  3rd Qu.:   6.57
##  Max.   :1090.84
```

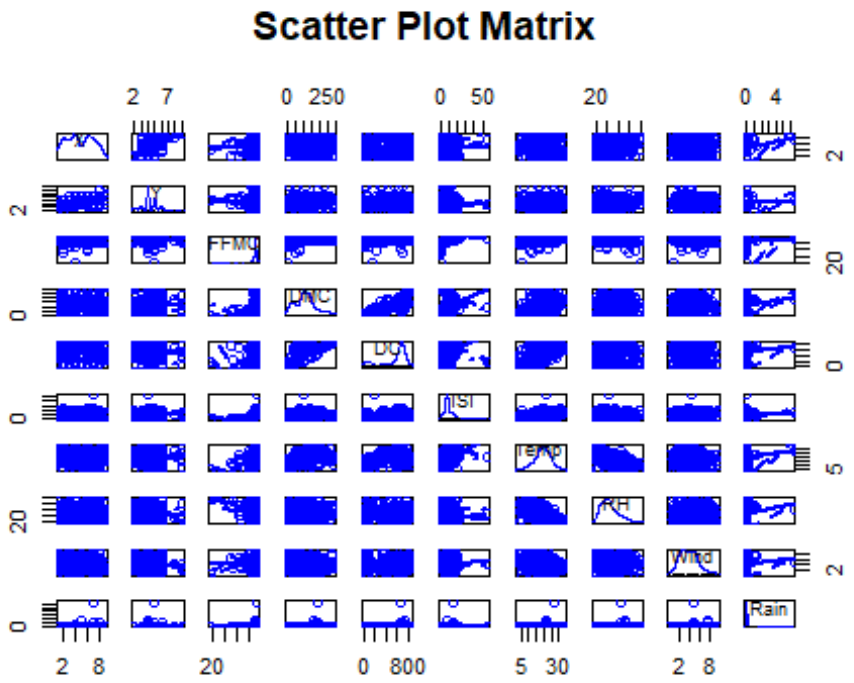
```
firedata$Month <- as.factor(firedata$Month)
firedata$Day <- as.factor(firedata$Day)
```

Scatter Plot Matrix

```
firedata <- firedata[-c(512,105,466,469,475,238),]
firedata1 <- firedata
scatterplotMatrix(firedata[,-c(3,4,13)], main="Scatter Plot Matrix")
```

```
## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit smooth

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit smooth
```

```
## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit smooth

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit smooth

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit smooth

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit smooth

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit smooth

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit smooth

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit smooth
## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit negative part of the spread

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit negative part of the spread

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit negative part of the spread

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit negative part of the spread

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit negative part of the spread

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit negative part of the spread

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit negative part of the spread

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit negative part of the spread

## Warning in smoother(x[subs], y[subs], col = smoother.args$col[i], log.x =
## FALSE, : could not fit negative part of the spread
```



Initial Model

```
model1 <- lm(log(Area+1)~., data=firedata)
summary(model1)

##
## Call:
## lm(formula = log(Area + 1) ~ ., data = firedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9521 -1.0304 -0.4938  0.8163  5.1621
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.1362554  1.8516345  -0.614  0.53974
## X            0.0550214  0.0322380   1.707  0.08852 .
## Y            0.0027591  0.0609737   0.045  0.96393
## Monthaug     0.2022533  0.8205437   0.246  0.80541
## Monthdec     2.0791198  0.7947320   2.616  0.00917 **
## Monthfeb     0.1264146  0.5614196   0.225  0.82194
## Monthjan     0.3760459  1.9029676   0.198  0.84343
## Monthjul     0.0009529  0.7116443   0.001  0.99893
## Monthjun    -0.4571192  0.6575228  -0.695  0.48726
## Monthmar    -0.3850491  0.5046445  -0.763  0.44583
## Monthmay     0.6867385  1.0940447   0.628  0.53049
## Monthnov    -1.0166474  1.4696156  -0.692  0.48941
## Monthoct     0.6199547  0.9794197   0.633  0.52705
## Monthsep     0.7980814  0.9206232   0.867  0.38643
```

```
## Daymon       0.1511698  0.2251898   0.671  0.50235
## Daysat       0.2972334  0.2173396   1.368  0.17207
## Daysun       0.2126592  0.2106128   1.010  0.31314
## Daythu       0.0127272  0.2410556   0.053  0.95791
## Daytue       0.2476444  0.2350040   1.054  0.29251
## Daywed       0.1856563  0.2450496   0.758  0.44904
## FFMC         0.0118313  0.0193116   0.613  0.54040
## DMC          0.0036070  0.0018853   1.913  0.05631 .
## DC          -0.0016806  0.0012717  -1.322  0.18695
## ISI         -0.0170988  0.0184954  -0.924  0.35569
## Temp         0.0381376  0.0223465   1.707  0.08853 .
## RH           0.0011245  0.0062229   0.181  0.85667
## Wind         0.0730418  0.0386023   1.892  0.05907 .
## Rain         0.0238409  0.2133732   0.112  0.91108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.372 on 483 degrees of freedom
## Multiple R-squared:  0.07714,    Adjusted R-squared:  0.02555
## F-statistic: 1.495 on 27 and 483 DF,  p-value: 0.05389
```

Performance Diagnostics using Typical Approach

```
stepAIC(model1)

## Start:  AIC=350.37
## log(Area + 1) ~ X + Y + Month + Day + FFMC + DMC + DC + ISI +
##     Temp + RH + Wind + Rain
##
##         Df Sum of Sq    RSS    AIC
## - Day    6     5.527 914.60 341.46
## - Y      1     0.004 909.08 348.37
## - Rain   1     0.023 909.10 348.38
## - RH     1     0.061 909.14 348.40
## - FFMC   1     0.706 909.78 348.76
## - ISI    1     1.609 910.69 349.27
## - DC     1     3.287 912.36 350.21
## <none>               909.08 350.37
## - Month 11    40.674 949.75 350.73
## - Temp   1     5.482 914.56 351.44
## - X      1     5.483 914.56 351.44
## - Wind   1     6.739 915.82 352.14
## - DMC    1     6.889 915.97 352.22
##
## Step:  AIC=341.46
## log(Area + 1) ~ X + Y + Month + FFMC + DMC + DC + ISI + Temp +
##     RH + Wind + Rain
##
##         Df Sum of Sq    RSS    AIC
## - Y      1     0.001 914.61 339.46
```

```
## - Rain   1     0.040 914.64 339.49
## - RH     1     0.161 914.77 339.55
## - FFMC   1     0.467 915.07 339.73
## - ISI    1     1.576 916.18 340.34
## <none>               914.60 341.46
## - Month 11    40.364 954.97 341.53
## - DC     1     3.805 918.41 341.59
## - X      1     5.663 920.27 342.62
## - Wind   1     6.356 920.96 343.00
## - Temp   1     6.736 921.34 343.21
## - DMC    1     6.833 921.44 343.27
##
## Step:  AIC=339.46
## log(Area + 1) ~ X + Month + FFMC + DMC + DC + ISI + Temp + RH +
##     Wind + Rain
##
##         Df Sum of Sq    RSS    AIC
## - Rain   1     0.040 914.65 337.49
## - RH     1     0.163 914.77 337.56
## - FFMC   1     0.466 915.07 337.73
## - ISI    1     1.578 916.18 338.35
## <none>               914.61 339.46
## - DC     1     3.907 918.51 339.64
## - Month 11    41.113 955.72 339.93
## - Wind   1     6.390 920.99 341.02
## - Temp   1     6.775 921.38 341.24
## - DMC    1     7.004 921.61 341.36
## - X      1     8.044 922.65 341.94
##
## Step:  AIC=337.49
## log(Area + 1) ~ X + Month + FFMC + DMC + DC + ISI + Temp + RH +
##     Wind
##
##         Df Sum of Sq    RSS    AIC
## - RH     1     0.205 914.85 335.60
## - FFMC   1     0.485 915.13 335.76
## - ISI    1     1.581 916.23 336.37
## <none>               914.65 337.49
## - DC     1     3.879 918.52 337.65
## - Month 11    41.256 955.90 338.03
## - Wind   1     6.492 921.14 339.10
## - DMC    1     6.973 921.62 339.37
## - Temp   1     7.202 921.85 339.50
## - X      1     8.131 922.78 340.01
##
## Step:  AIC=335.6
## log(Area + 1) ~ X + Month + FFMC + DMC + DC + ISI + Temp + Wind
##
##         Df Sum of Sq    RSS    AIC
## - FFMC   1     0.389 915.24 333.82
```

```
## - ISI    1     1.524 916.37 334.45
## <none>               914.85 335.60
## - DC     1     3.976 918.83 335.82
## - Month 11    42.762 957.61 336.95
## - Wind   1     6.504 921.36 337.22
## - DMC    1     7.833 922.68 337.96
## - X      1     8.456 923.31 338.30
## - Temp   1    11.458 926.31 339.96
##
## Step:  AIC=333.82
## log(Area + 1) ~ X + Month + DMC + DC + ISI + Temp + Wind
##
##         Df Sum of Sq    RSS    AIC
## - ISI    1     1.135 916.37 332.45
## <none>               915.24 333.82
## - DC     1     4.045 919.28 334.07
## - Wind   1     6.220 921.46 335.28
## - Month 11    43.444 958.68 335.52
## - X      1     8.463 923.70 336.52
## - DMC    1     8.699 923.94 336.65
## - Temp   1    12.655 927.89 338.84
##
## Step:  AIC=332.45
## log(Area + 1) ~ X + Month + DMC + DC + Temp + Wind
##
##         Df Sum of Sq    RSS    AIC
## - DC     1     3.540 919.91 332.42
## <none>               916.37 332.45
## - Wind   1     5.430 921.81 333.47
## - X      1     8.231 924.61 335.02
## - DMC    1     8.308 924.68 335.06
## - Month 11    46.686 963.06 335.84
## - Temp   1    11.664 928.04 336.92
##
## Step:  AIC=332.42
## log(Area + 1) ~ X + Month + DMC + Temp + Wind
##
##         Df Sum of Sq    RSS    AIC
## <none>               919.91 332.42
## - DMC    1     4.785 924.70 333.07
## - Wind   1     5.934 925.85 333.71
## - X      1     8.058 927.97 334.88
## - Month 11    45.182 965.10 334.92
## - Temp   1    11.576 931.49 336.81

##
## Call:
## lm(formula = log(Area + 1) ~ X + Month + DMC + Temp + Wind, data =
firedata)
##
```

```
## Coefficients:
## (Intercept)            X     Monthaug     Monthdec     Monthfeb
##   -0.007182     0.055788    -0.580814     1.603589     0.113904
##    Monthjan     Monthjul     Monthjun     Monthmar     Monthmay
##   -0.464342    -0.486861    -0.814234    -0.381075     0.688372
##    Monthnov     Monthoct     Monthsep          DMC         Temp
##   -1.053036    -0.336896    -0.197171     0.002238     0.035970
##        Wind
##    0.065407
```

```
model1 <- lm(log(Area + 1) ~ Month + DMC + Temp, data = firedata)
summary(model1)
```

```
##
## Call:
## lm(formula = log(Area + 1) ~ Month + DMC + Temp, data = firedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7360 -1.0628 -0.5897  0.8191  5.5726
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.674118   0.488850   1.379   0.1685
## Monthaug    -0.664678   0.523469  -1.270   0.2048
## Monthdec     1.694085   0.655383   2.585   0.0100 *
## Monthfeb     0.029175   0.555608   0.053   0.9581
## Monthjan    -0.839700   1.448410  -0.580   0.5624
## Monthjul    -0.544299   0.552169  -0.986   0.3247
## Monthjun    -0.778559   0.592124  -1.315   0.1892
## Monthmar    -0.417457   0.495123  -0.843   0.3996
## Monthmay     0.641513   1.072358   0.598   0.5500
## Monthnov    -1.051052   1.445078  -0.727   0.4674
## Monthoct    -0.390698   0.583488  -0.670   0.5034
## Monthsep    -0.323983   0.502363  -0.645   0.5193
## DMC          0.002363   0.001401   1.687   0.0922 .
## Temp         0.031343   0.014354   2.184   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.371 on 497 degrees of freedom
## Multiple R-squared:  0.05191,    Adjusted R-squared:  0.02711
## F-statistic: 2.093 on 13 and 497 DF,  p-value: 0.01331
```

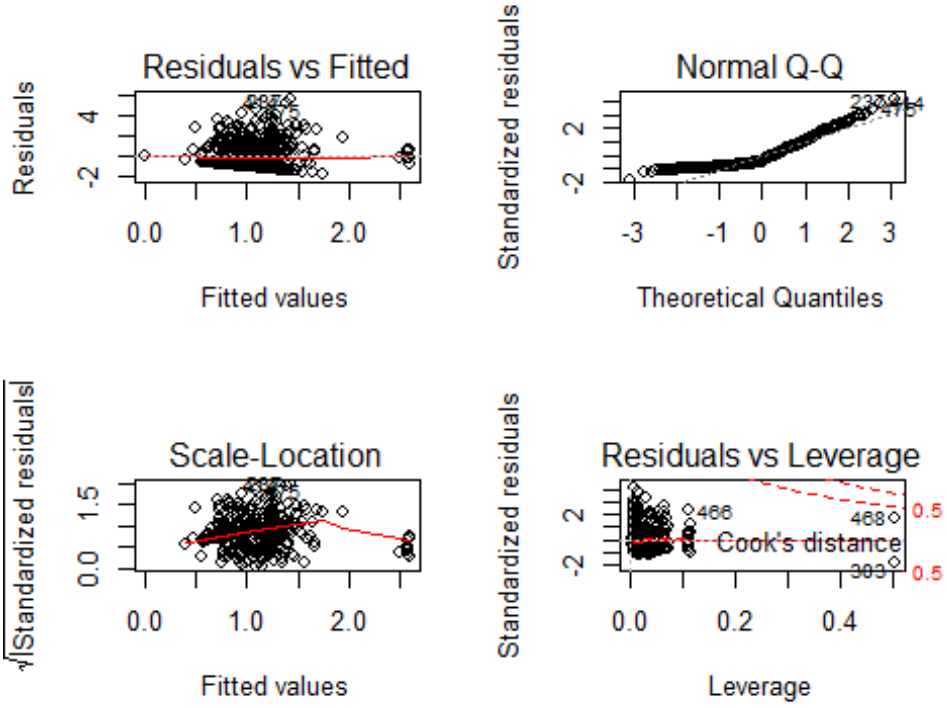
```
stepAIC(model1)
```

```
## Start:  AIC=336.15
## log(Area + 1) ~ Month + DMC + Temp
##
##         Df Sum of Sq    RSS    AIC
## <none>               933.92 336.15
```

```
## - DMC    1     5.348 939.27 337.06
## - Month 11    45.918 979.84 338.67
## - Temp   1     8.960 942.88 339.03

##
## Call:
## lm(formula = log(Area + 1) ~ Month + DMC + Temp, data = firedata)
##
## Coefficients:
## (Intercept)     Monthaug     Monthdec     Monthfeb     Monthjan
##    0.674118    -0.664678     1.694085     0.029175    -0.839700
##    Monthjul     Monthjun     Monthmar     Monthmay     Monthnov
##   -0.544299    -0.778559    -0.417457     0.641513    -1.051052
##    Monthoct     Monthsep          DMC          Temp
##   -0.390698    -0.323983     0.002363     0.031343
```

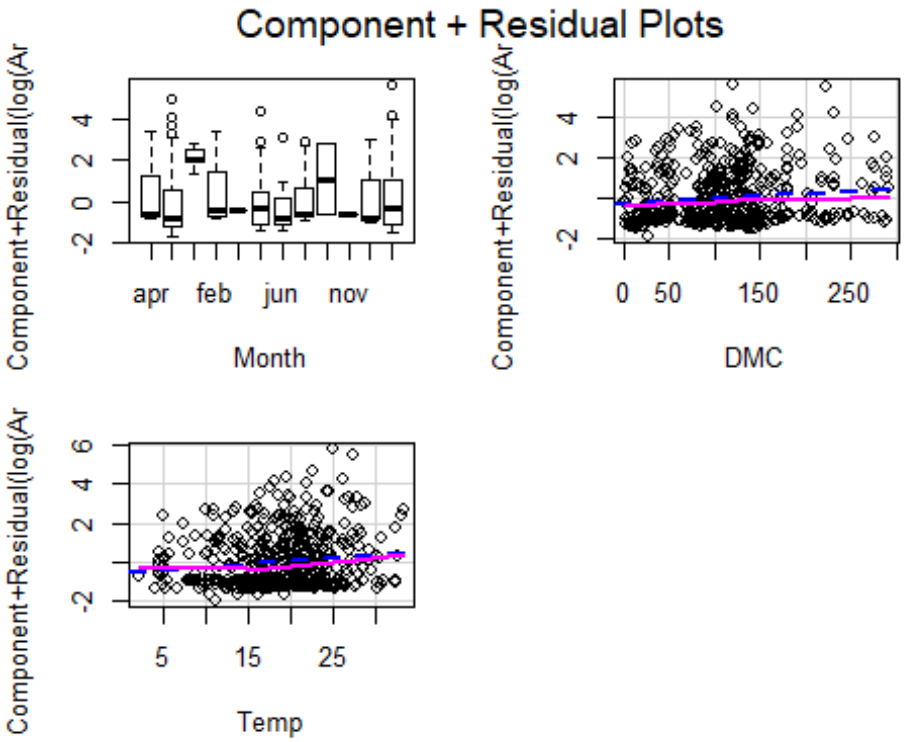
```
par(mfrow=c(2,2))
plot(model1)
```

```
## Warning: not plotting observations with leverage one:
##   378, 511

## Warning: not plotting observations with leverage one:
##   378, 511
```



Performance Diagnostics using Enhanced Approach

```r
par(mfrow=c(1,1))

#qqPlot(model1, labels=row.names(cdata1), id.method="identify", simulate=T,
main="Q-Q Plot")



crPlots(model1)
```




```r
ncvTest(model1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 3.885861, Df = 1, p = 0.048694
```

```r
fit1 <- gvlma(model1)
summary(fit1)
```

```
##
## Call:
## lm(formula = log(Area + 1) ~ Month + DMC + Temp, data = firedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7360 -1.0628 -0.5897  0.8191  5.5726
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.674118   0.488850   1.379   0.1685
## Monthaug    -0.664678   0.523469  -1.270   0.2048
## Monthdec     1.694085   0.655383   2.585   0.0100 *
## Monthfeb     0.029175   0.555608   0.053   0.9581
## Monthjan    -0.839700   1.448410  -0.580   0.5624
## Monthjul    -0.544299   0.552169  -0.986   0.3247
## Monthjun    -0.778559   0.592124  -1.315   0.1892
## Monthmar    -0.417457   0.495123  -0.843   0.3996
## Monthmay     0.641513   1.072358   0.598   0.5500
## Monthnov    -1.051052   1.445078  -0.727   0.4674
## Monthoct    -0.390698   0.583488  -0.670   0.5034
## Monthsep    -0.323983   0.502363  -0.645   0.5193
## DMC          0.002363   0.001401   1.687   0.0922 .
## Temp         0.031343   0.014354   2.184   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.371 on 497 degrees of freedom
## Multiple R-squared:  0.05191,    Adjusted R-squared:  0.02711
## F-statistic: 2.093 on 13 and 497 DF,  p-value: 0.01331
##
##
## 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        155.737 0.000e+00 Assumptions NOT satisfied!
## Skewness           123.975 0.000e+00 Assumptions NOT satisfied!
## Kurtosis            20.721 5.314e-06 Assumptions NOT satisfied!
## Link Function        1.500 2.207e-01    Assumptions acceptable.
## Heteroscedasticity   9.542 2.008e-03 Assumptions NOT satisfied!
```

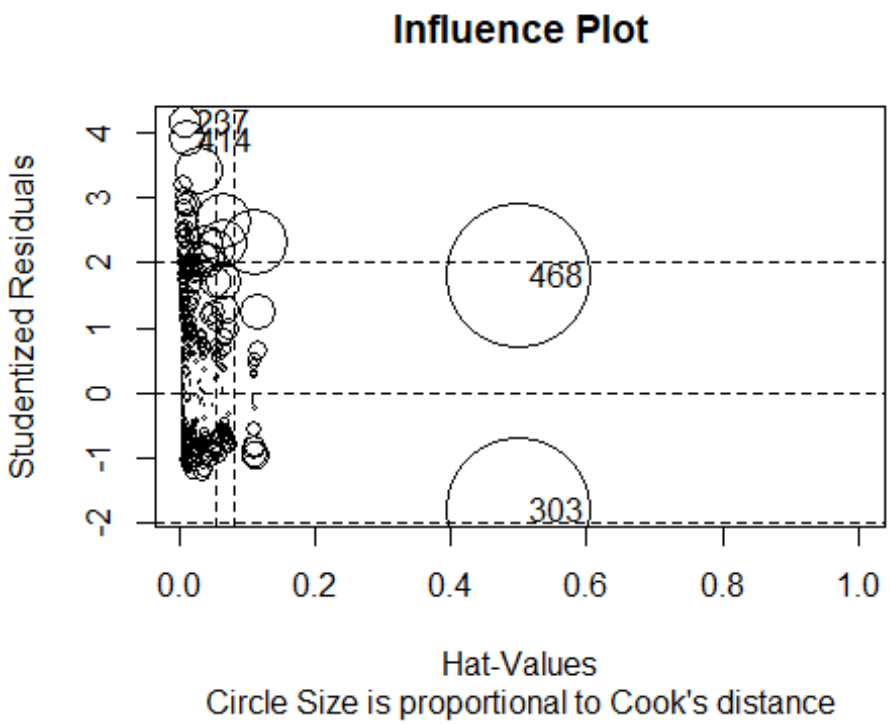
```
sqrt(vif(model1))>2
```

```
##        GVIF    Df GVIF^(1/(2*Df))
## Month FALSE  TRUE           FALSE
## DMC   FALSE FALSE           FALSE
## Temp  FALSE FALSE           FALSE
```

```
outlierTest(model1)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 237 4.150024         3.9124e-05     0.019914
```

```
influencePlot(model1, main="Influence Plot", sub="Circle Size is proportional
to Cook's distance")
```




```
##       StudRes         Hat      CookD
## 237  4.150024 0.009143559 0.01099330
## 303 -1.797168 0.501234681 0.23080746
## 378       NaN 1.000000000        NaN
## 414  3.888458 0.013904743 0.01480827
## 468  1.797168 0.501234681 0.23080746
## 511       NaN 1.000000000        NaN
```

After removing outliers and selecting parameters given by the stepAIC approach to create new model

```
model2 <- lm(log(Area + 1) ~ Month + DMC + Temp, data = firedata1)
summary(model2)
```

```
##
## Call:
## lm(formula = log(Area + 1) ~ Month + DMC + Temp, data = firedata1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7360 -1.0628 -0.5897  0.8191  5.5726
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.674118   0.488850   1.379   0.1685
## Monthaug    -0.664678   0.523469  -1.270   0.2048
## Monthdec     1.694085   0.655383   2.585   0.0100 *
```

```
## Monthfeb     0.029175   0.555608   0.053   0.9581
## Monthjan    -0.839700   1.448410  -0.580   0.5624
## Monthjul    -0.544299   0.552169  -0.986   0.3247
## Monthjun    -0.778559   0.592124  -1.315   0.1892
## Monthmar    -0.417457   0.495123  -0.843   0.3996
## Monthmay     0.641513   1.072358   0.598   0.5500
## Monthnov    -1.051052   1.445078  -0.727   0.4674
## Monthoct    -0.390698   0.583488  -0.670   0.5034
## Monthsep    -0.323983   0.502363  -0.645   0.5193
## DMC          0.002363   0.001401   1.687   0.0922 .
## Temp         0.031343   0.014354   2.184   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.371 on 497 degrees of freedom
## Multiple R-squared:  0.05191,    Adjusted R-squared:  0.02711
## F-statistic: 2.093 on 13 and 497 DF,  p-value: 0.01331
```

Based on the regression analysis of the given data, it can be said that the model generated is not good enough to use for predictions of Area. Because some of the assumptions of linear regression model is not getting satisfied without removing major chunk of data. Hence it is better to use some other model or one can use some other attributes.