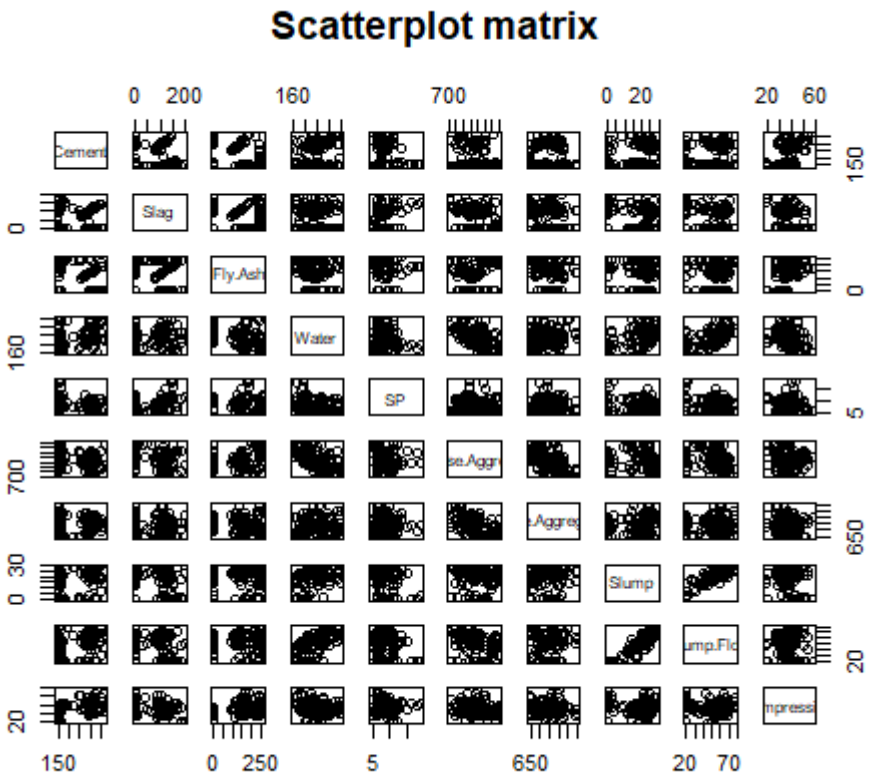
Question-1

1)pairs(df, main = "Scatterplot matrix")



2) fit1<-lm(Slump.Flow~Water,data=df)

summary(fit1)

```
Call:
lm(formula = Slump.Flow ~ Water, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-37.211 -10.836   2.734  11.031  22.163

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -58.72755   13.28635  -4.420 2.49e-05 ***
Water         0.54947    0.06704   8.196 8.10e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.68 on 101 degrees of freedom
Multiple R-squared:  0.3995,   Adjusted R-squared:  0.3935
F-statistic: 67.18 on 1 and 101 DF,  p-value: 8.097e-13
```

fit2<lm(Slump.Flow~Water+Cement+Slag+Fly.Ash+SP+Coarse.Aggregate+Fine.Aggregate,dat
a=d)summary(fit2)

```
Call:
lm(formula = Slump.Flow ~ Water + Cement + Slag + Fly.Ash + SP +
    Coarse.Aggregate + Fine.Aggregate, data = df)

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
Water               0.73180    0.35282   2.074   0.0408 *
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
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
```

fit3<-lm(Slump.Flow~Water+I(Water^2),data=df)

summary(fit3)

```
Call:
lm(formula = Slump.Flow ~ Water + I(Water^2), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-38.357  -9.678   2.865  10.271  21.473

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.173e+02  1.247e+02  -1.742   0.0845 .
Water        2.154e+00  1.256e+00   1.714   0.0896 .
I(Water^2)  -4.015e-03  3.140e-03  -1.279   0.2040
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.64 on 100 degrees of freedom
Multiple R-squared:  0.4091,   Adjusted R-squared:  0.3973
F-statistic: 34.62 on 2 and 100 DF,  p-value: 3.759e-12
```

Regression Diagonistics

confint(fit1)

```
                  2.5 %      97.5 %
(Intercept) -85.0841046 -32.3709993
Water         0.4164861   0.6824575
```

confint(fit2)

```
                      2.5 %      97.5 %
(Intercept)      -947.84451365 442.0951684
Water               0.03136972   1.4322277
Cement             -0.16942710   0.2767133
Slag               -0.31614617   0.3047654
Fly.Ash            -0.16520290   0.2875048
SP                 -1.01716230   1.6138194
Coarse.Aggregate   -0.19454098   0.3418613
Fine.Aggregate     -0.18771010   0.3757443
```

confint(fit3)

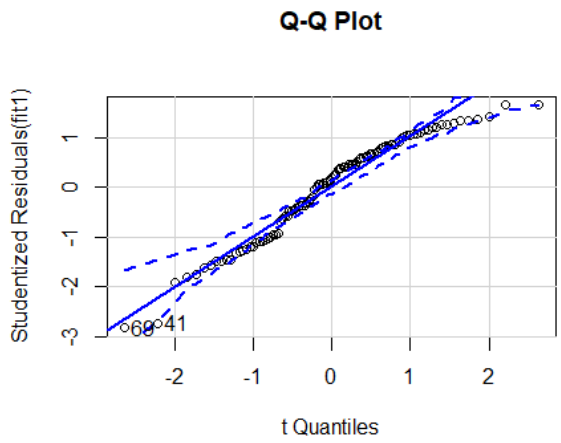
```
                   2.5 %     97.5 %
(Intercept) -464.8282531 30.1569253
Water         -0.3389976  4.6466730
I(Water^2)    -0.0102455  0.0022148
```

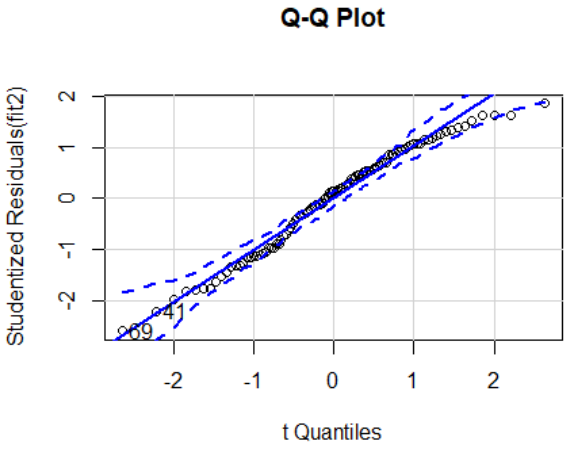
Typical Approach

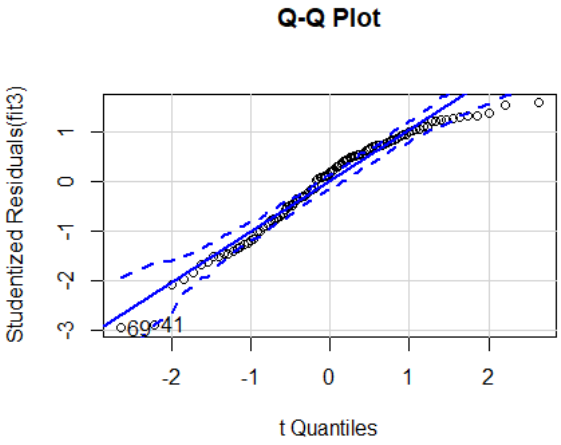
qqPlot(fit1,labels=row.names(df),id.method="identify",simulate=TRUE,main="Q-Q Plot")

qqPlot(fit2,labels=row.names(df),id.method="identify",simulate=TRUE,main="Q-Q Plot")

qqPlot(fit3,labels=row.names(df),id.method="identify",simulate=TRUE,main="Q-Q Plot")





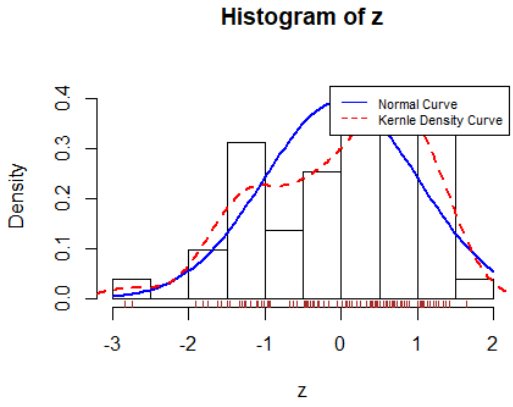


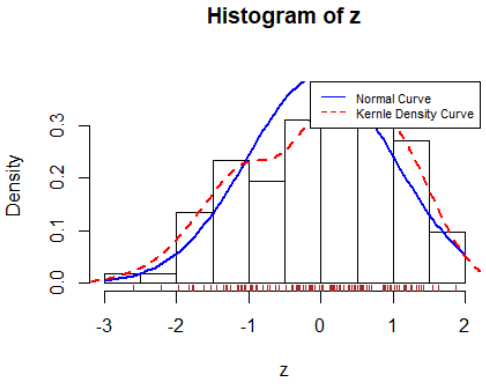
Residual Plots

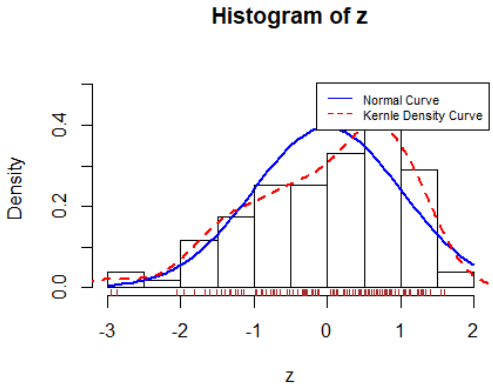
residplot(fit1)

residplot(fit2)

residplot(fit3)







Independence of Errors

durbinWatsonTest(fit1)

durbinWatsonTest(fit2)

durbinWatsonTest(fit3)

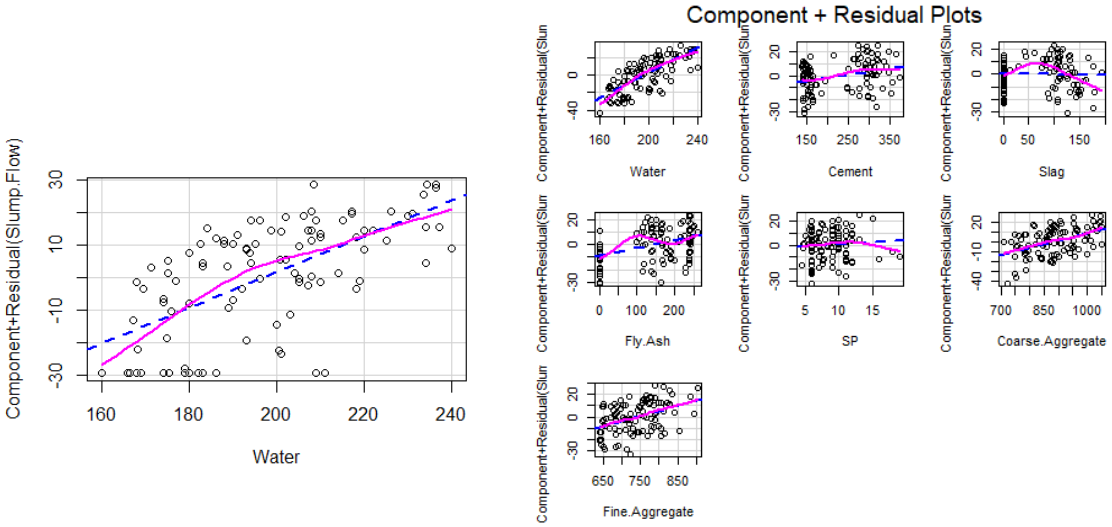
```
lag Autocorrelation D-W Statistic p-value
   1      0.06495095      1.842612    0.42
 Alternative hypothesis: rho != 0

lag Autocorrelation D-W Statistic p-value
   1     -0.01249995      2.009189    0.81
 Alternative hypothesis: rho != 0

lag Autocorrelation D-W Statistic p-value
   1      0.05728419       1.86123   0.444
 Alternative hypothesis: rho != 0
```

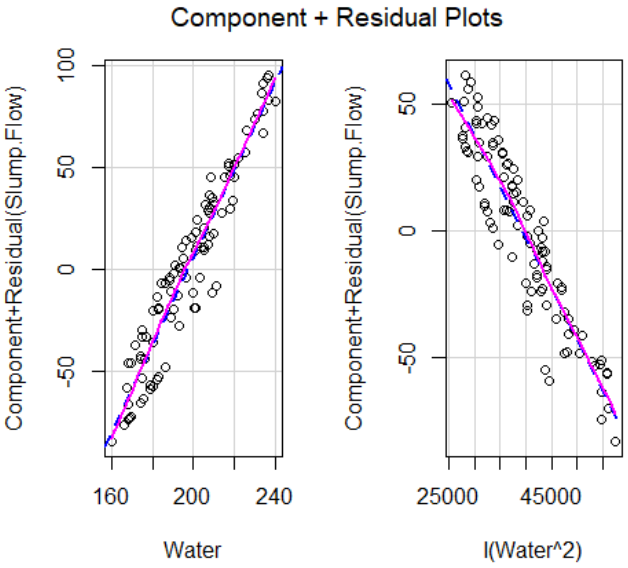
Linearity

crPlots(fit1)

crPlots(fit2)

crPlots(fit3)





Homoscedasticity

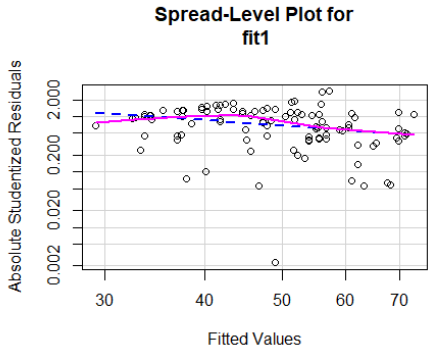
ncvTest(fit1)

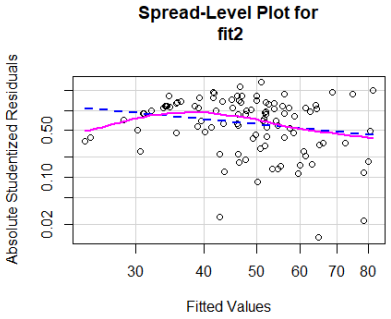
ncvTest(fit2)

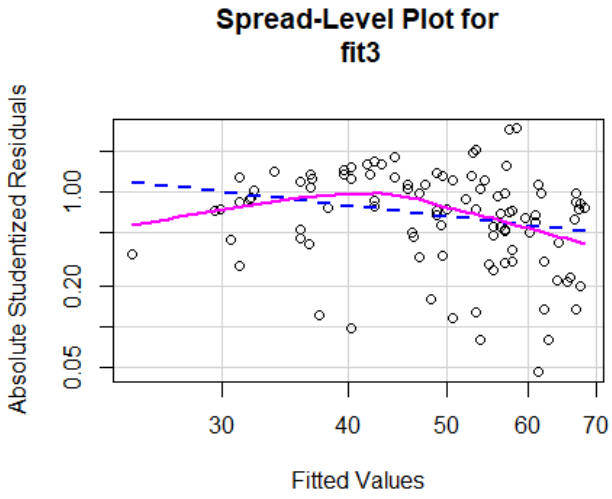
ncvTest(fit3)

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.76085, Df = 1, p = 0.18452
```

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

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.9714059, Df = 1, p = 0.32433

spreadLevelPlot(fit1)
spreadLevelPlot(fit2)
spreadLevelPlot(fit3)
```







Global Test

gvmodel1<-gvlma(fit1)

summary(gvmodel1)

summary(gvmodel2)

gvmodel3<-gvlma(fit3)

summary(gvmodel3)

Multicollinearity

vif(fit1)

vif(fit2)

vif(fit3)

OutLiers

outlierTest(fit1)

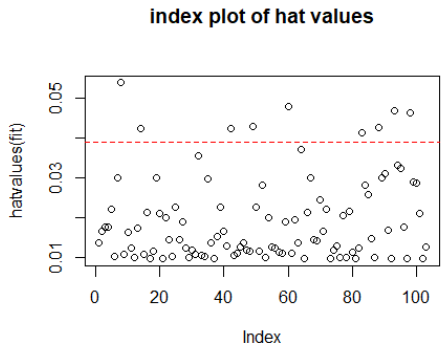
outlierTest(fit2)

outlierTest(fit3)

```
> outlierTest(fit1)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
69 -2.833049          0.0055772      0.57445
> outlierTest(fit2)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
69 -2.603738           0.010717           NA
> outlierTest(fit3)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
69 -2.945746          0.0040164      0.41369
```

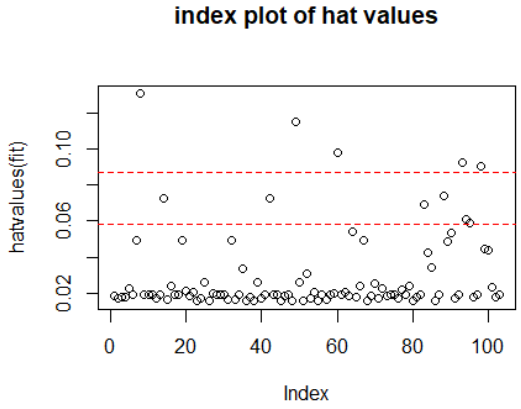
High Leverage Points

hat.plot(fit1)



hat.plot(fit2)

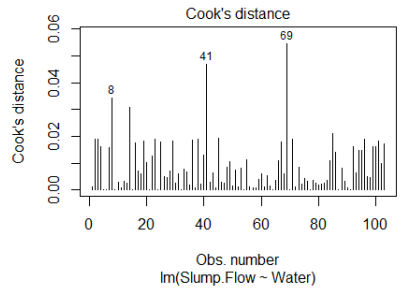
hat.plot(fit3)



Influential observations

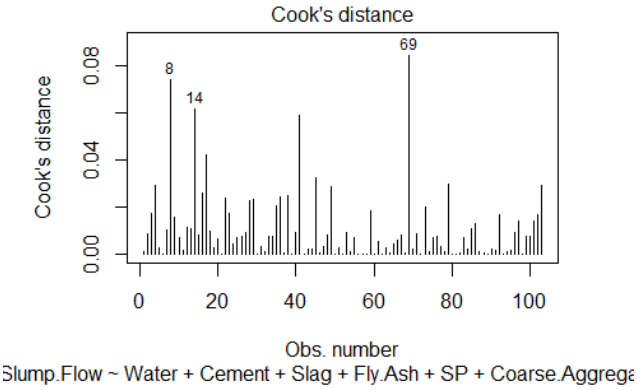
cutoff <- 4/(nrow(df)-length(fit1$coefficients)-2)

plot(fit1,which=4,cook.levels=cutoff)

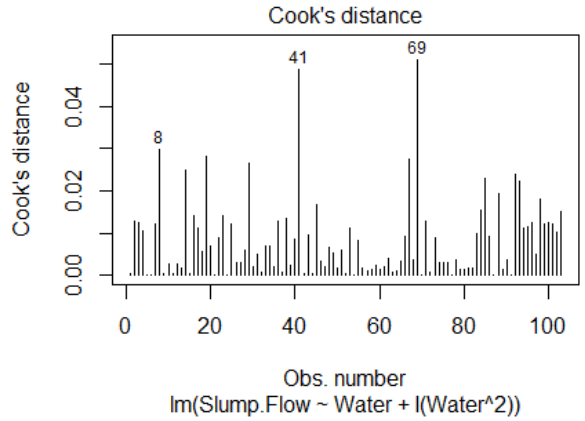


cutoff <- 4/(nrow(df)-length(fit2$coefficients)-2)

plot(fit2,which=4,cook.levels=cutoff)




```
cutoff <- 4/(nrow(df)-length(fit3$coefficients)-2)

plot(fit3,which=4,cook.levels=cutoff)
```



Corrective Measures

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

sqrt(vif(fit2))>2

sqrt(vif(fit3))>2

df<-df[-c(69),]
```

Best regression model

AIC(fit1, fit2, fit3)

Fine Tune

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

stepAIC(Step_fit, direction = "backward")

```
                  Df Sum of Sq   RSS    AIC
- SP                1      0.28 14617 520.43
- Coarse.Aggregate  1      8.17 14625 520.48
- Slag              1     10.14 14627 520.50
- Cement            1     12.20 14629 520.51
- Fly.Ash           1     16.40 14634 520.54
- Fine.Aggregate    1     23.39 14640 520.59
<none>                         14617 522.43
- Water             1    517.67 15135 523.98

Step:  AIC=520.43
Slump.Flow ~ Water + Cement + Slag + Fly.Ash + Coarse.Aggregate +
    Fine.Aggregate

                  Df Sum of Sq   RSS    AIC
- Coarse.Aggregate  1     11.70 14629 518.51
- Cement            1     17.95 14635 518.55
- Slag              1     20.95 14638 518.58
- Fly.Ash           1     25.10 14642 518.60
- Fine.Aggregate    1     35.81 14653 518.68
<none>                         14617 520.43
- Water             1    967.16 15584 524.96

Step:  AIC=518.51
Slump.Flow ~ Water + Cement + Slag + Fly.Ash + Fine.Aggregate

                 Df Sum of Sq   RSS    AIC
- Cement          1      12.1 14641 516.60
- Fly.Ash         1      39.6 14669 516.79
- Fine.Aggregate  1     151.1 14780 517.56
<none>                       14629 518.51
- Slag            1    1135.0 15764 524.13
- Water           1   11704.5 26334 576.47

Step:  AIC=516.6
Slump.Flow ~ Water + Slag + Fly.Ash + Fine.Aggregate

                 Df Sum of Sq   RSS    AIC
- Fly.Ash         1      28.2 14669 514.79
- Fine.Aggregate  1     139.1 14780 515.56
<none>                       14641 516.60
- Slag            1    1834.4 16476 526.64
- Water           1   11811.6 26453 574.93

Step:  AIC=514.79
```

```
Slump.Flow ~ Water + Slag + Fine.Aggregate

                 Df Sum of Sq   RSS    AIC
- Fine.Aggregate  1     112.4 14782 513.57
<none>                       14669 514.79
- Slag            1    2440.9 17110 528.49
- Water           1   12269.9 26939 574.79

Step:  AIC=513.57
Slump.Flow ~ Water + Slag

        Df Sum of Sq   RSS    AIC
<none>               14782 513.57
- Slag   1    2720.4 17502 528.80
- Water  1   12697.3 27479 574.81

Call:
lm(formula = Slump.Flow ~ Water + Slag, data = df)

Coefficients:
(Intercept)        Water         Slag
  -52.54983      0.55369     -0.08574
```

INTERPRETATION

All the above regression models has low Adjusted R values so the simple linear regression and multiple linear regression may not be the best regression model for the given dataset. Normality is violated Independence of errors is not satisfied and there is no correlation between the response variable and the predictor variables. The constant variance assumption is also not met. Considering the above conclusions, second order polynomial equation seems to be the best model among the three models evaluated but it may not be an ideal one.

QUESTION 2

FOREST FIRES DATA

Initialising Data:

Question 2 Forest Fires Data

library(readxl)

forest<-read_xlsx("Forest Fires Data.xlsx")

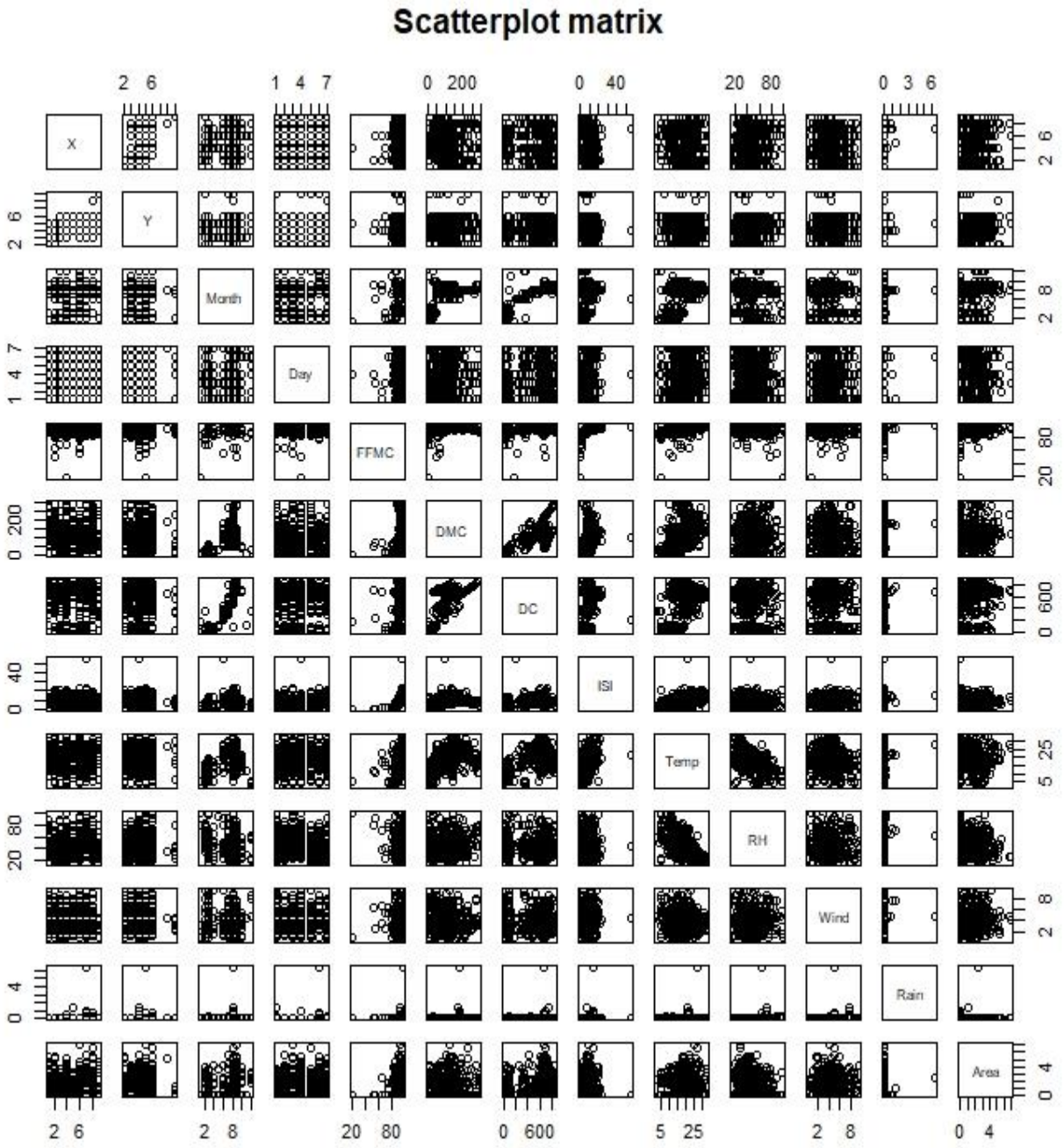
attach(forest)

###log transformation of area

```
Area<- log(Area+1)
###change month and day to numeric
tst <- capitalize(c(Month))
match(tst, month.abb)
Months<-match(tst, month.abb)
Day<-factor(c(Day))
Days<-as.numeric(Day)
forest[["Month"]] <- Months
forest[["Day"]]<-Days
```

SCATTERPLOT

```
pairs(forest, main = "Scatterplot matrix")
```

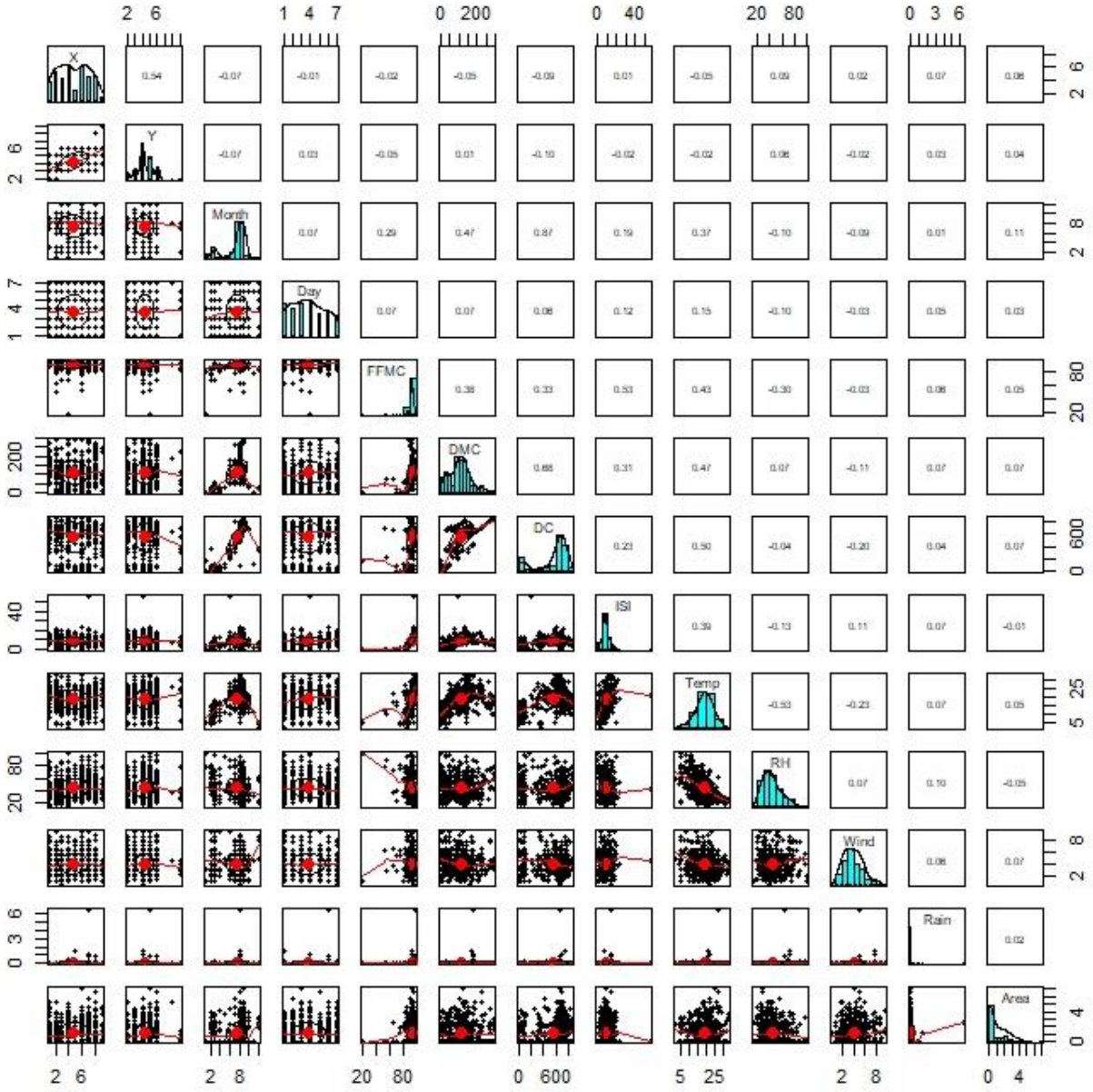


USING PEARSON'S METHOD:

###Using pearson

library(psych)

pairs.panels(forest,method="pearson")



REGRESSION MODELS:

Model 1:

###Model 1 Multiple linear regression

fit1<-
lm(Area~X+Y+FFMC+DMC+DC+ISI+Temp+RH+Wind+Rain+Month+Day,data=forest)

summary(fit1)

```
Call:
lm(formula = Area ~ X + Y + FFMC + DMC + DC + ISI + Temp + RH + 
    Wind + Rain + Month + Day, data = forest)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1728 -1.0819 -0.5324  0.8366  5.5995 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.3210832  1.3863173  -0.232   0.8169
X            0.0426562  0.0316338   1.348   0.1781
Y           -0.0021349  0.0601283  -0.036   0.9717
FFMC         0.0050199  0.0144648   0.347   0.7287
DMC          0.0023859  0.0015417   1.548   0.1223
DC          -0.0011811  0.0007001  -1.687   0.0922 .
ISI         -0.0253822  0.0168888  -1.503   0.1335
Temp         0.0075483  0.0174008   0.434   0.6646
RH          -0.0039050  0.0052393  -0.745   0.4564
Wind         0.0587710  0.0371763   1.581   0.1145
Rain         0.0757787  0.2115808   0.358   0.7204
Month        0.1527399  0.0607600   2.514   0.0123 *
Day          0.0132305  0.0323201   0.409   0.6825
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.388 on 504 degrees of freedom
Multiple R-squared:  0.03799,  Adjusted R-squared:  0.01509 
F-statistic: 1.659 on 12 and 504 DF,  p-value: 0.07269
```

###Model 2 polynomial regression

fit2<-lm(Area~Month + I(Month^2)+I(Month^3)+I(Month^4))

summary(fit2)

```
Call:
lm(formula = Area ~ Month + I(Month^2) + I(Month^3) + I(Month^4))

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7735 -1.0863 -0.6676  0.8929  5.8051 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.9861371  1.5784933   0.625    0.532
Month       -0.2030365  1.2003887  -0.169    0.866
I(Month^2)   0.0851006  0.3052106   0.279    0.780
I(Month^3)  -0.0120766  0.0318249  -0.379    0.704
I(Month^4)   0.0006009  0.0011619   0.517    0.605

Residual standard error: 1.387 on 512 degrees of freedom
```

```
Multiple R-squared:  0.02321,  Adjusted R-squared:  0.01558
F-statistic: 3.042 on 4 and 512 DF,  p-value: 0.017
```

###Model 2 Simple linear regression

Fit3<-lm(Area~Month)

summary(fit3)

```
Call:
lm(formula = Area ~ Month)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3585 -1.1478 -0.7096  0.8982  5.7776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.58610    0.21018   2.789  0.00549 **
Month        0.07022    0.02690   2.611  0.00930 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.391 on 515 degrees of freedom
Multiple R-squared:  0.01306,  Adjusted R-squared:  0.01114
F-statistic: 6.815 on 1 and 515 DF,  p-value: 0.009304
```

###Model 4 Multiple Linear regression

fit4<-lm(Area~Month+X+Wind+DC+DMC,data=forest)

summary(fit4)

```
Call:
lm(formula = Area ~ Month + X + Wind + DC + DMC, data = forest)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0261 -1.0776 -0.5832  0.8711  5.7178

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0268920  0.3011193  -0.089  0.92887
Month        0.1580688  0.0590617   2.676  0.00768 **
X            0.0390629  0.0264425   1.477  0.14022
Wind         0.0438871  0.0356153   1.232  0.21842
DC          -0.0011622  0.0006668  -1.743  0.08195 .
DMC          0.0021135  0.0013984   1.511  0.13131
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.384 on 511 degrees of freedom
Multiple R-squared:  0.02999,  Adjusted R-squared:  0.0205
F-statistic:  3.16 on 5 and 511 DF,  p-value: 0.008065
```

>

REGRESSION DIAGNOSTICS:

Model 1:

```
confint(fit1)
                  2.5 %       97.5 %
(Intercept) -3.0447557887 2.4025893736
X           -0.0194942027 0.1048066677
Y           -0.1202679591 0.1159981545
FFMC        -0.0233987443 0.0334386050
DMC         -0.0006430003 0.0054147528
DC          -0.0025565799 0.0001944675
ISI         -0.0585632945 0.0077989942
Temp        -0.0266387465 0.0417354107
RH          -0.0141985669 0.0063886262
Wind        -0.0142685569 0.1318105050
Rain        -0.3399103203 0.4914677097
Month        0.0333658713 0.2721140049
Day         -0.0502681589 0.0767291058

>
```

Model 2:

```
confint(fit2)
                 2.5 %      97.5 %
(Intercept) -2.114983591 4.087257717
Month       -2.561329813 2.155256864
I(Month^2)  -0.514518685 0.684719942
I(Month^3)  -0.074600164 0.050446866
I(Month^4)  -0.001681816 0.002883588
```

Model 3:

```
confint(fit3)
                2.5 %    97.5 %
(Intercept) 0.17318550 0.9990053
Month       0.01737451 0.1230596
```

Model 4:

```
confint(fit4)
                  2.5 %       97.5 %
(Intercept) -0.6184760756 0.5646921692
```

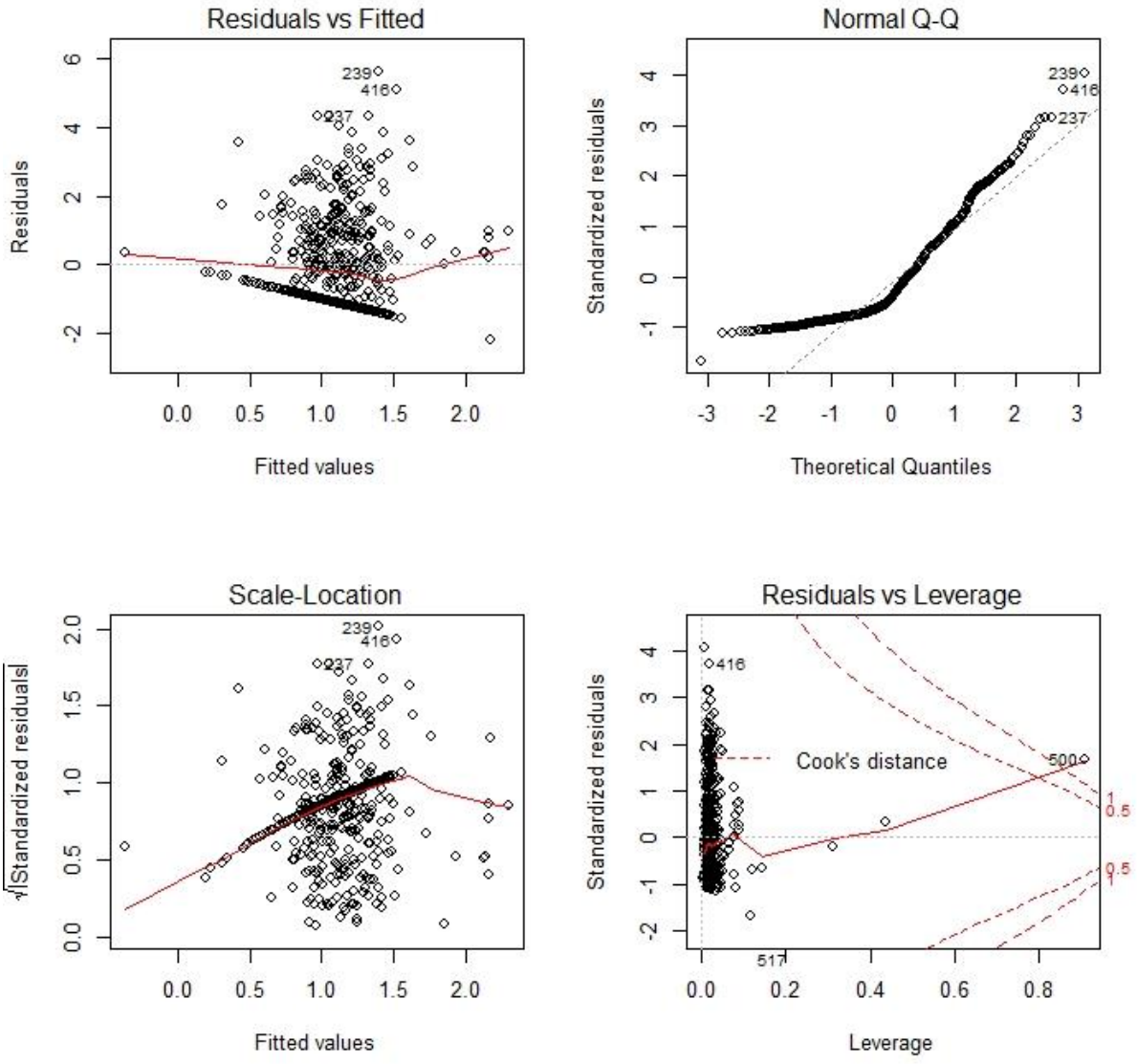
```
Month         0.0420351193 0.2741024355
X            -0.0128864860 0.0910123120
Wind         -0.0260834091 0.1138575981
DC           -0.0024722698 0.0001478544
DMC          -0.0006337772 0.0048607759
```

TYPICAL APPROACH:

Model 1

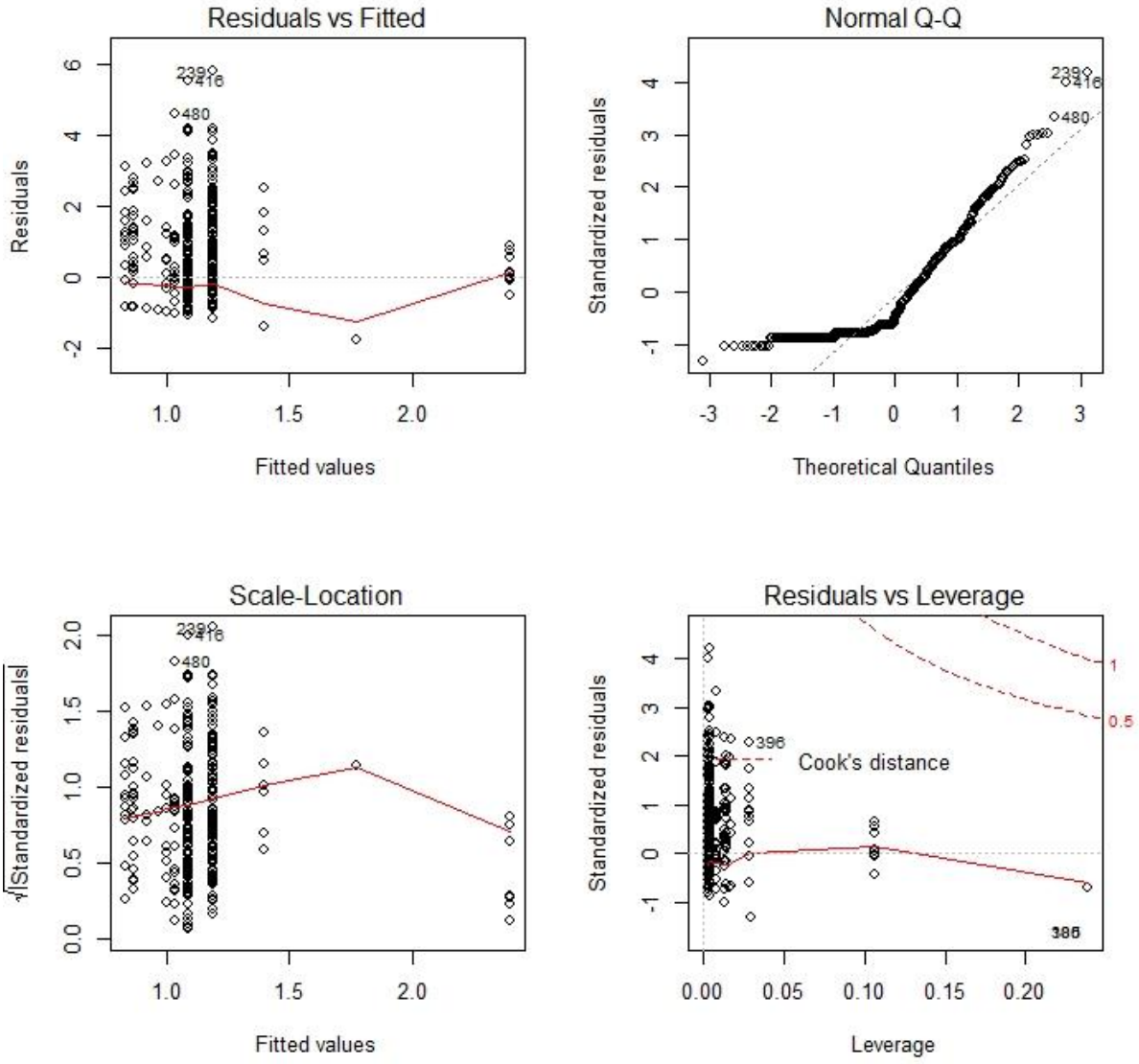
par(mfrow=c(2,2))

plot(fit1)



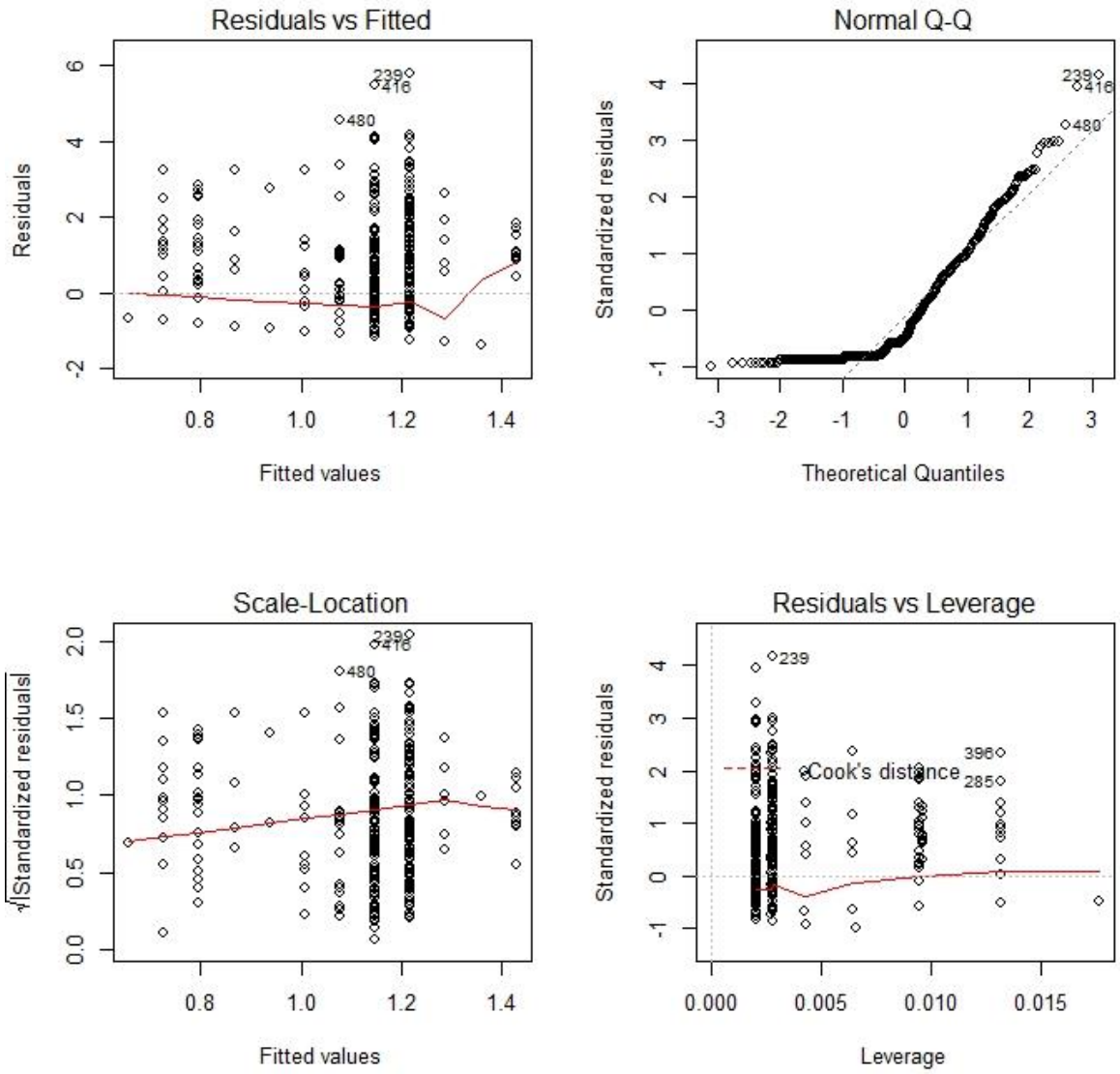
Model 2

plot(fit2)



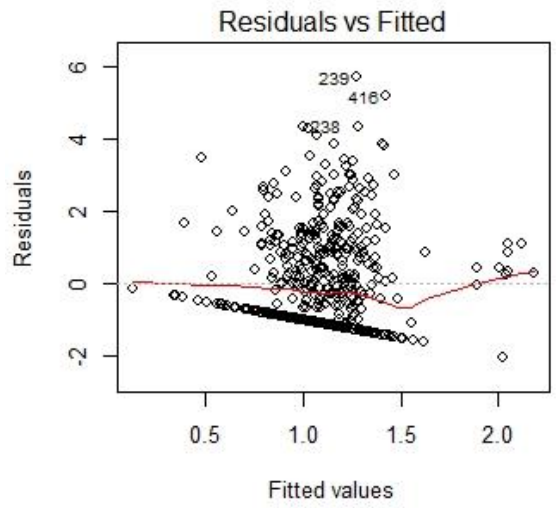
Model 3

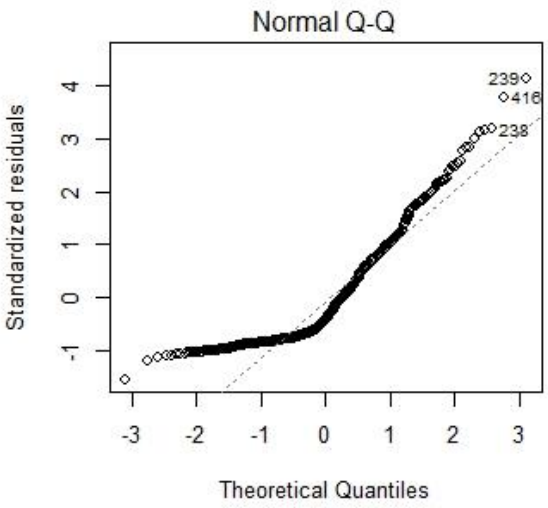
plot(fit3)

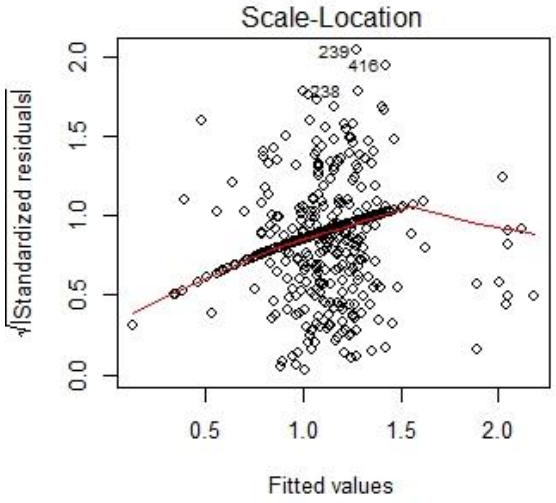


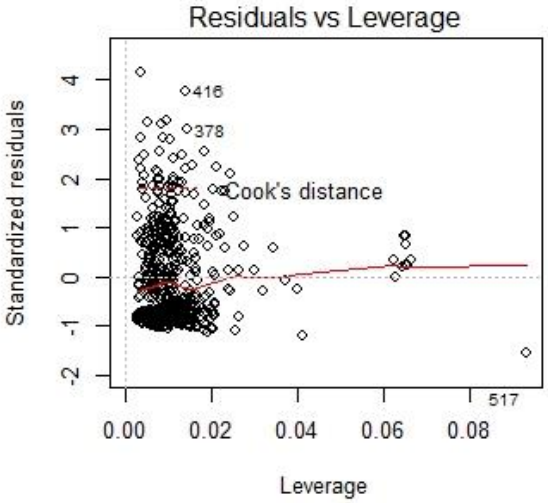
Model 4

plot(fit4)









ENHANCED APPROACH

Model 1

###enhanced approach

par(mfrow=c(1,1))

library(car)

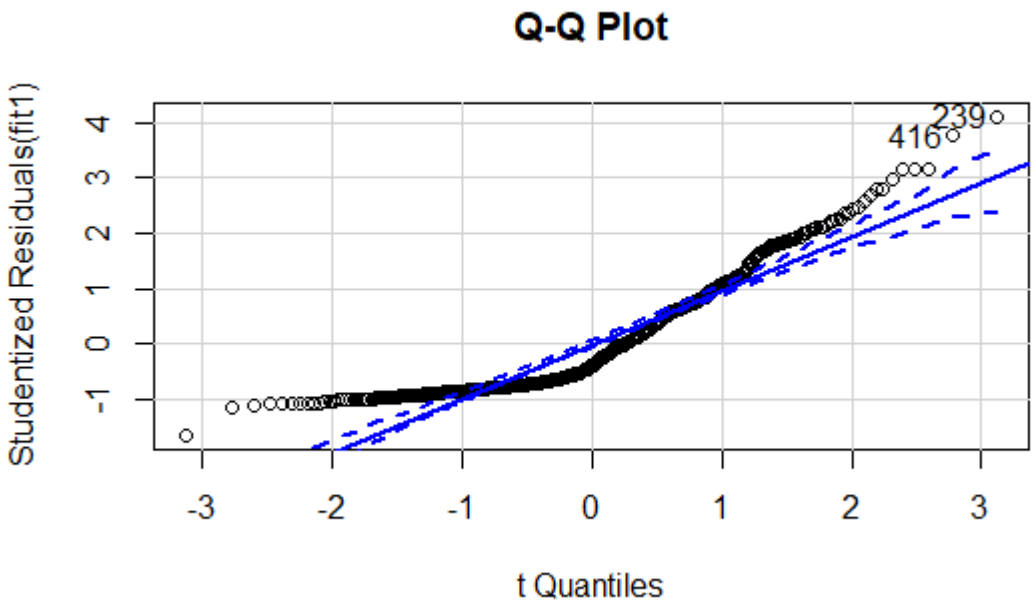
qqPlot(fit1,labels=row.names(forest),id.method="identify",simuate=TRUE, main ="Q-Q Plot")

qqPlot(fit2,labels=row.names(forest),id.method="identify",simuate=TRUE, main ="Q-Q Plot")

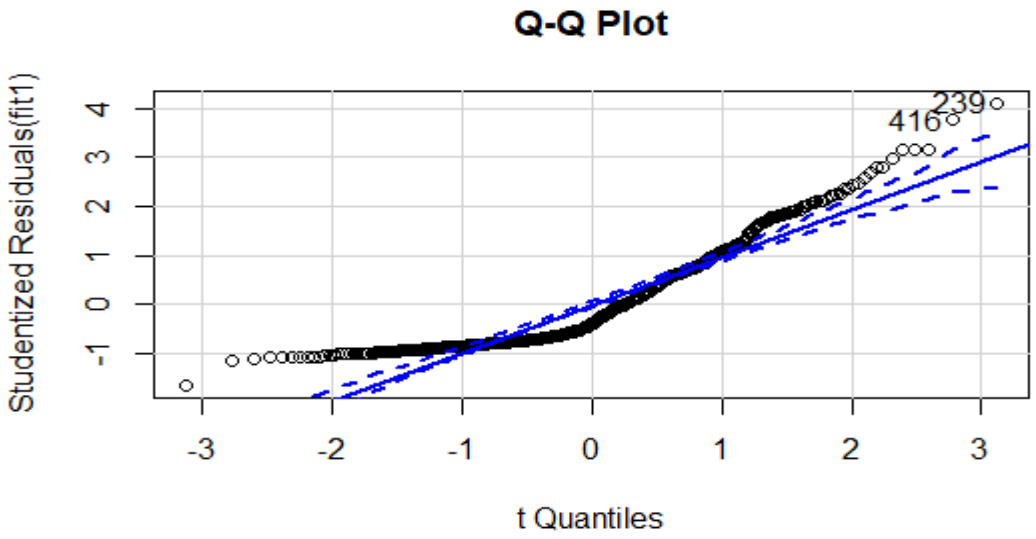
qqPlot(fit3,labels=row.names(forest),id.method="identify",simuate=TRUE, main ="Q-Q Plot")

qqPlot(fit4,labels=row.names(forest),id.method="identify",simuate=TRUE, main ="Q-Q Plot")

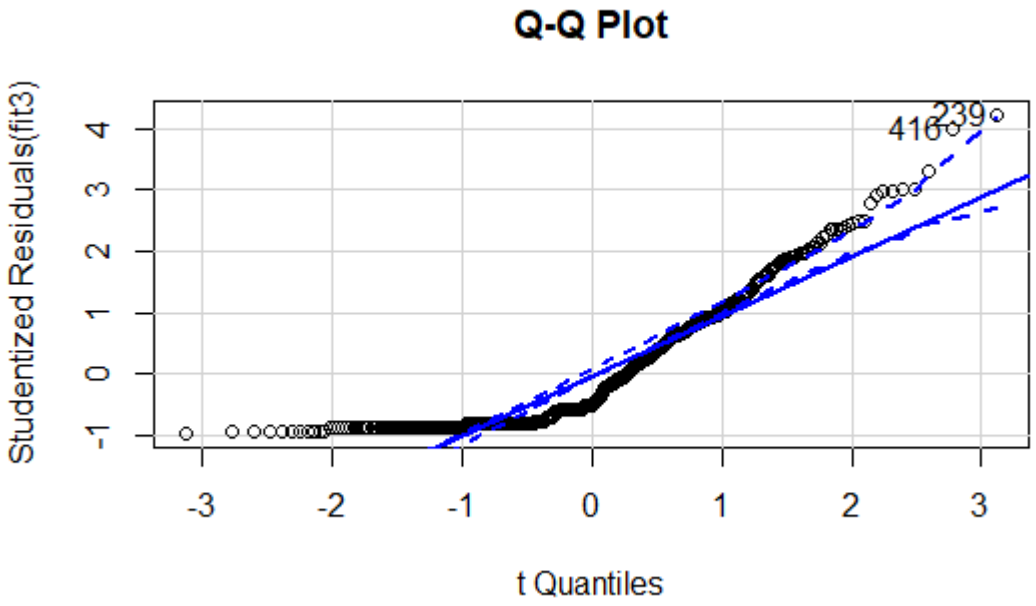
Model 1



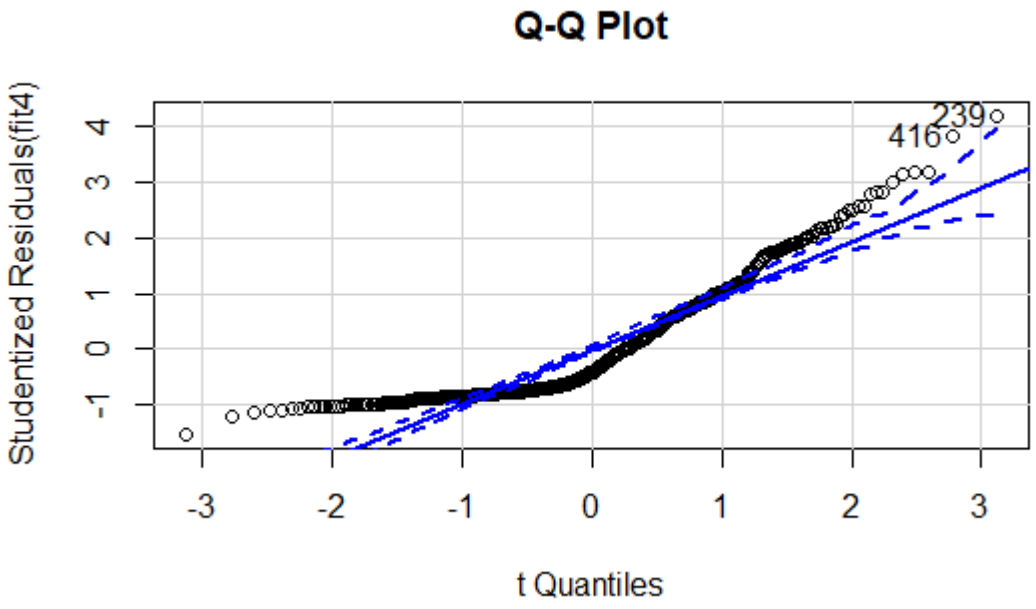
Model 2



Model 3



Model 4

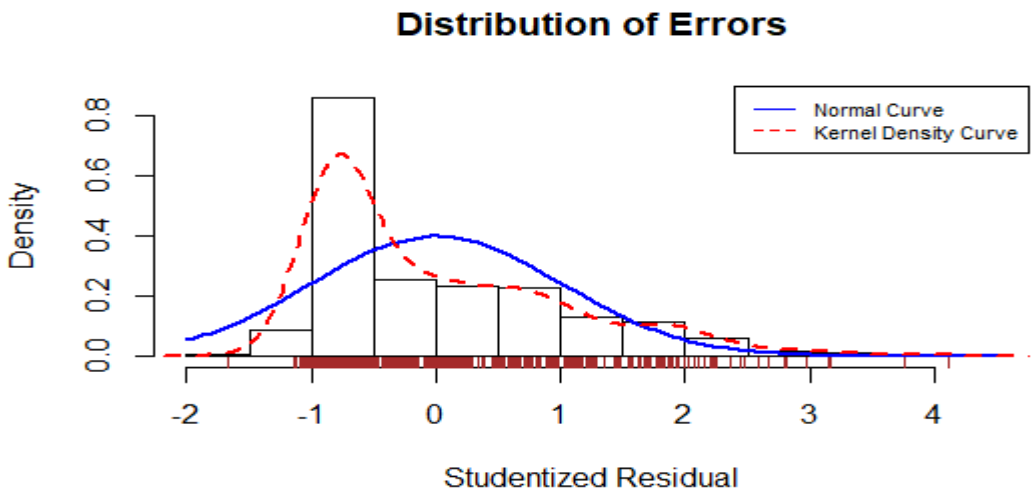


RESIDPLOT FUNCTION:

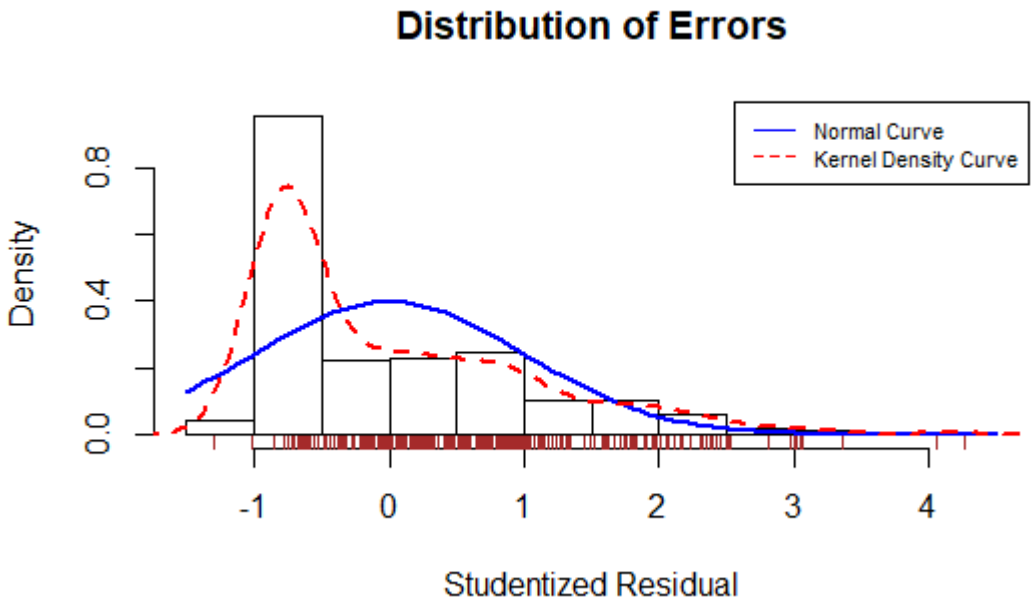
```
###residplot function
residplot <- function(fit, nbreaks=10)
{
  z <- rstudent(fit)
  hist(z, breaks=nbreaks, freq=FALSE,
       xlab="Studentized Residual",
       main="Distribution of Errors")
  rug(jitter(z), col="brown")
  curve(dnorm(x, mean=mean(z), sd=sd(z)),
        add=TRUE, col="blue", lwd=2)
  lines(density(z)$x, density(z)$y,
        col="red", lwd=2, lty=2)
  legend("topright",
         legend = c( "Normal Curve", "Kernel Density Curve"),
```

```
        lty=1:2, col=c("blue","red"), cex=.7)
}
residplot(fit1)
residplot(fit2)
residplot(fit3)
residplot(fit4)
```

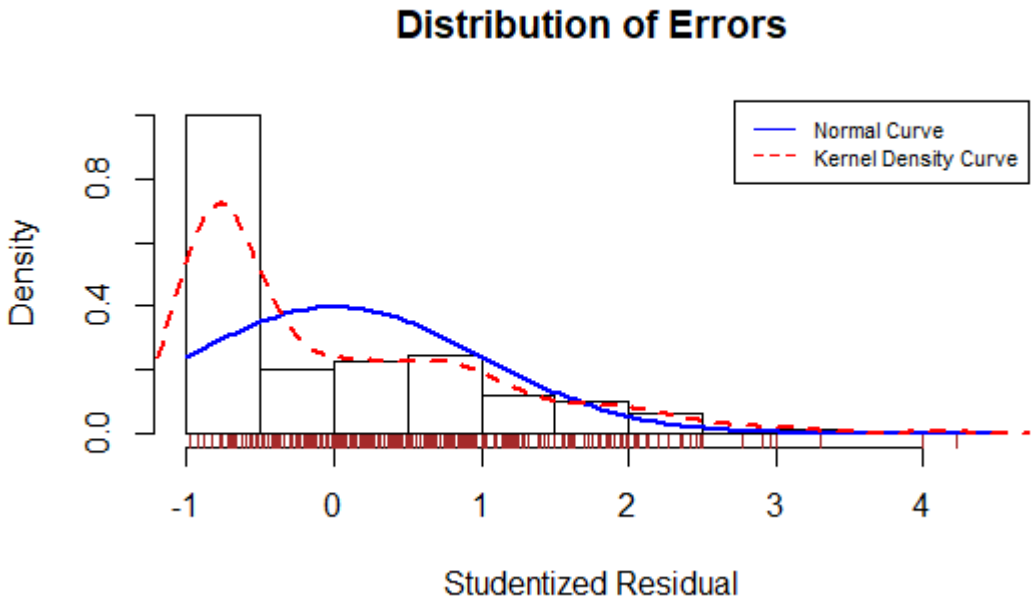
Model 1



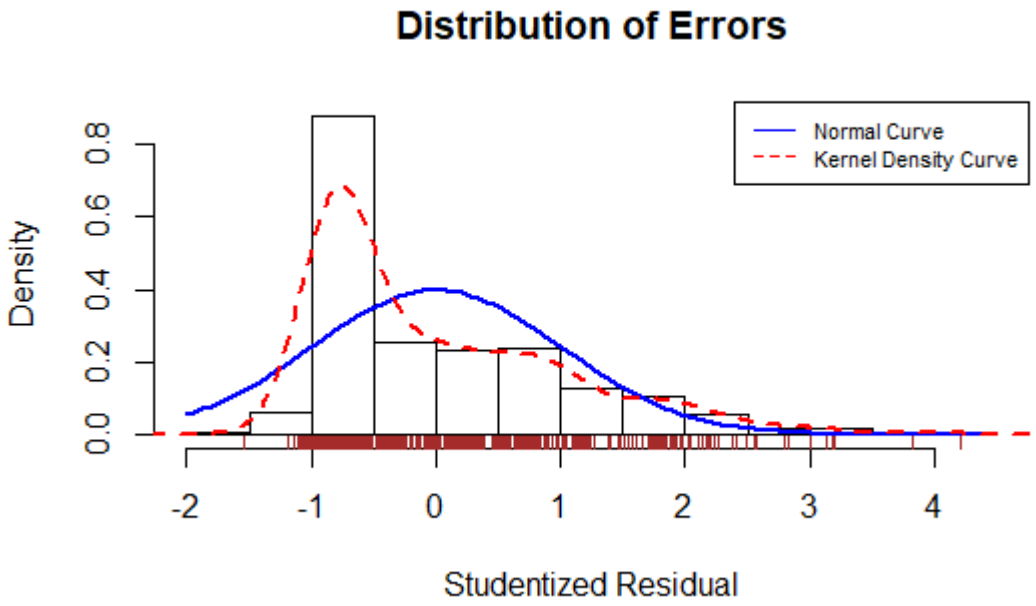
Model 2



Model 3



Model 4



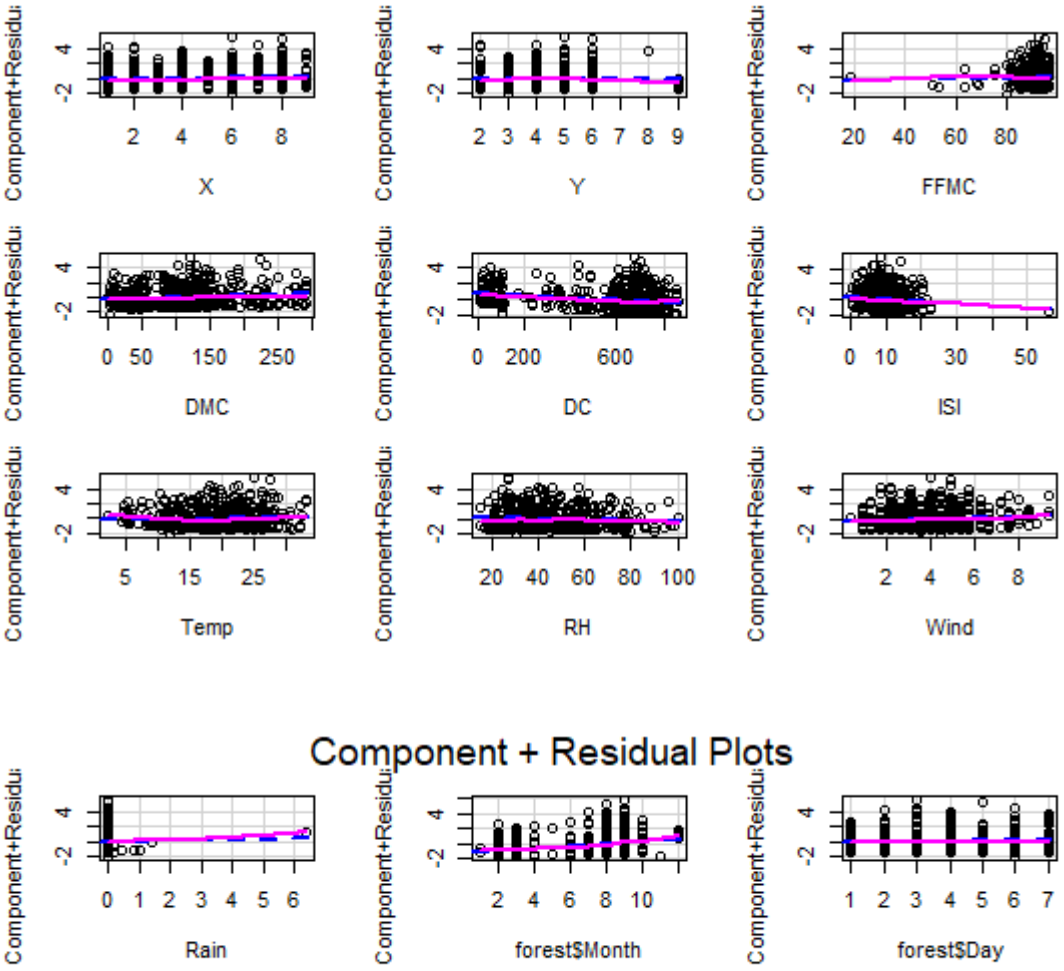
INDEPENDENCE OF ERRORS:

```
durbinWatsonTest(fit1)
 lag Autocorrelation D-W Statistic p-value
   1       0.5234635     0.9472819       0
 Alternative hypothesis: rho != 0
durbinWatsonTest(fit2)
 lag Autocorrelation D-W Statistic p-value
   1       0.5313619     0.9333253       0
 Alternative hypothesis: rho != 0
durbinWatsonTest(fit3)
 lag Autocorrelation D-W Statistic p-value
   1       0.5350754     0.9273588       0
 Alternative hypothesis: rho != 0
durbinWatsonTest(fit4)
 lag Autocorrelation D-W Statistic p-value
   1       0.5239835     0.9468966       0
 Alternative hypothesis: rho != 0
```

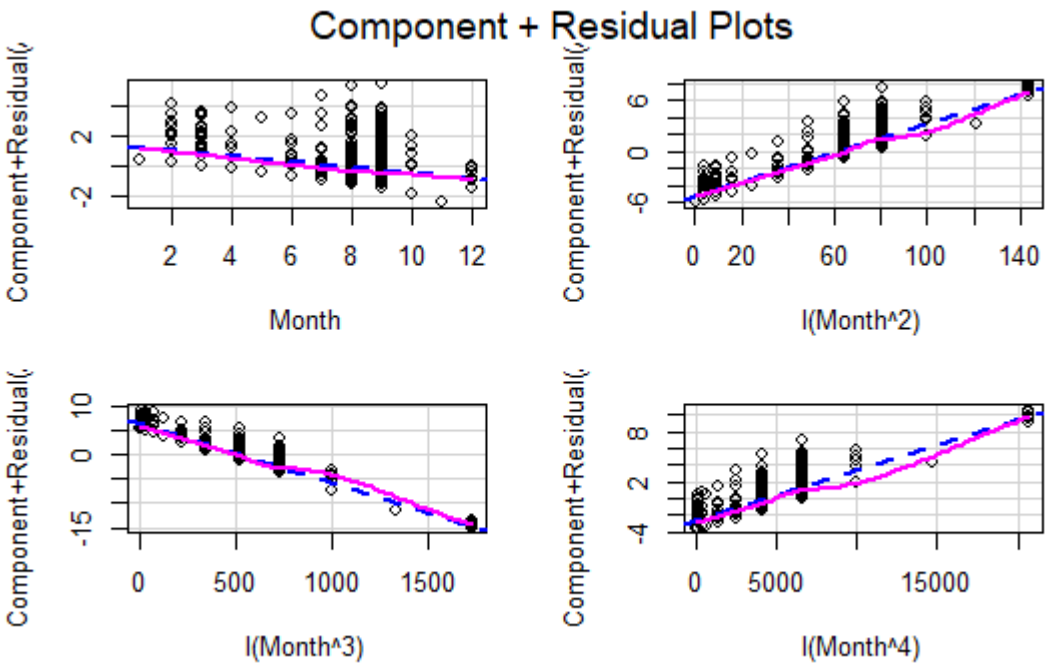
LINEARITY:

```
crPlots(fit1)
crPlots(fit2)
crPlots(fit3)
crPlots(fit4)
```

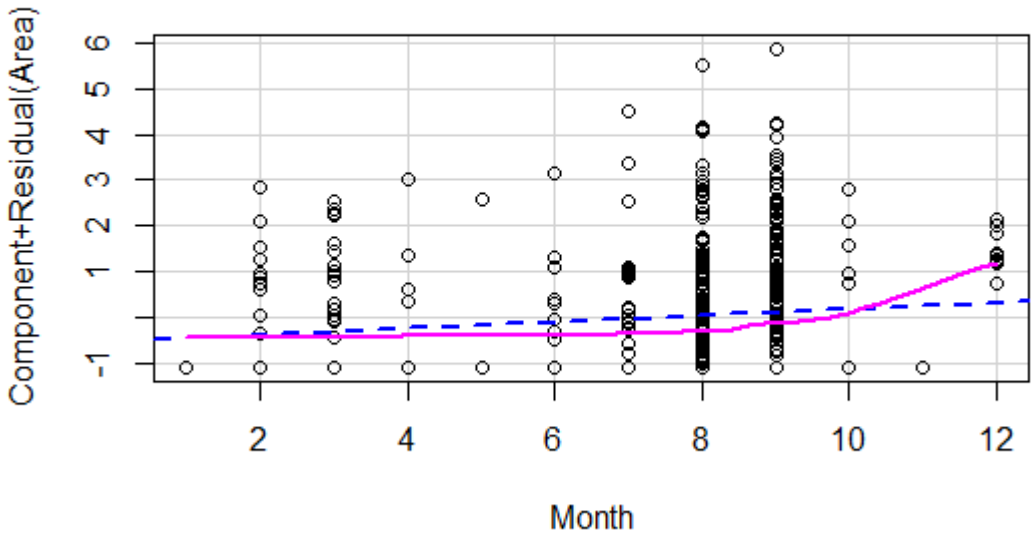
Model 1


Component+Residual
X
Y
FFMC
DMC
DC
ISI
Temp
RH
Wind
Component + Residual Plots
Rain
forest$Month
forest$Day

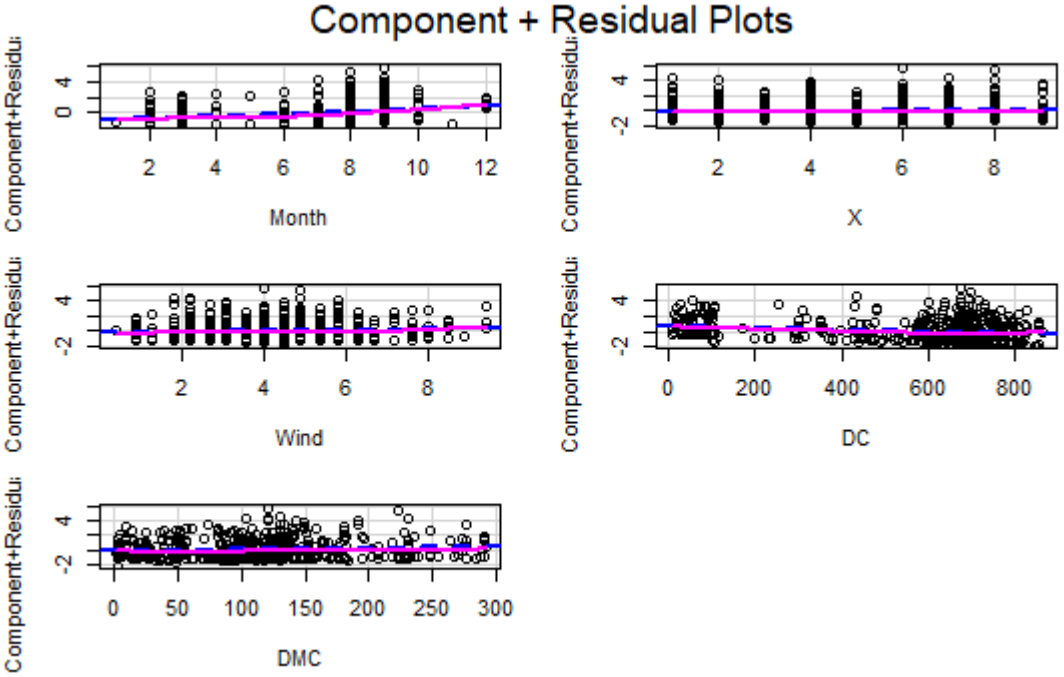
Model 2


Component + Residual Plots
Component+Residual(
Month
I(Month^2)
I(Month^3)
I(Month^4)

Model 3



Model 4



HOMOSCEDASTICITY:

```
ncvTest(fit1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
```

```
Chisquare = 11.03601, Df = 1, p = 0.00089359
ncvTest(fit2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.1060695, Df = 1, p = 0.74466
ncvTest(fit3)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.69401, Df = 1, p = 0.10073
ncvTest(fit4)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 6.399094, Df = 1, p = 0.011418
```

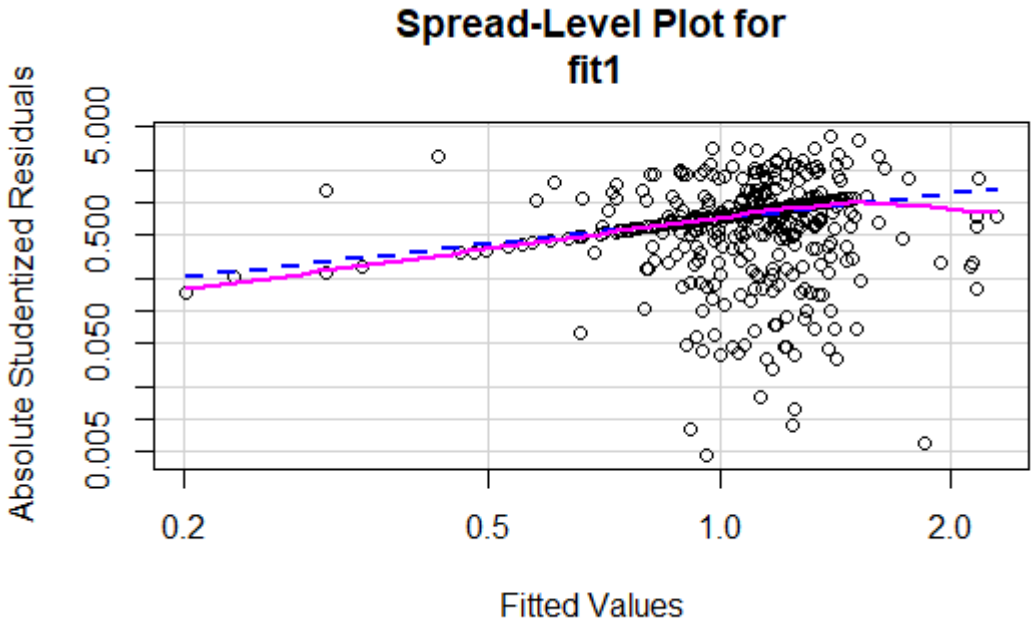
spreadLevelPlot(fit1)

spreadLevelPlot(fit2)

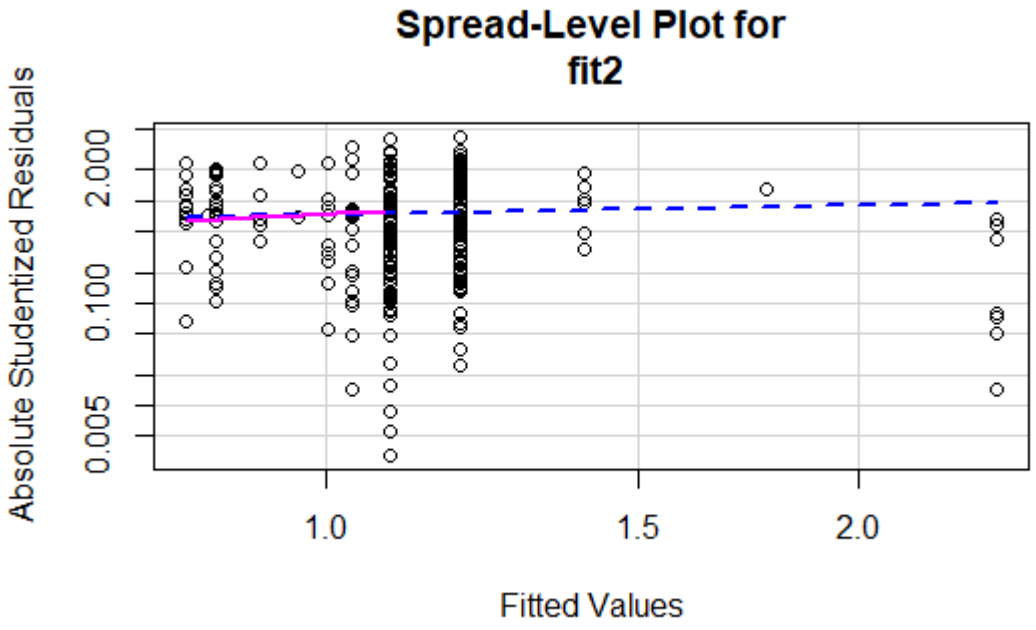
spreadLevelPlot(fit3)

spreadLevelPlot(fit4)

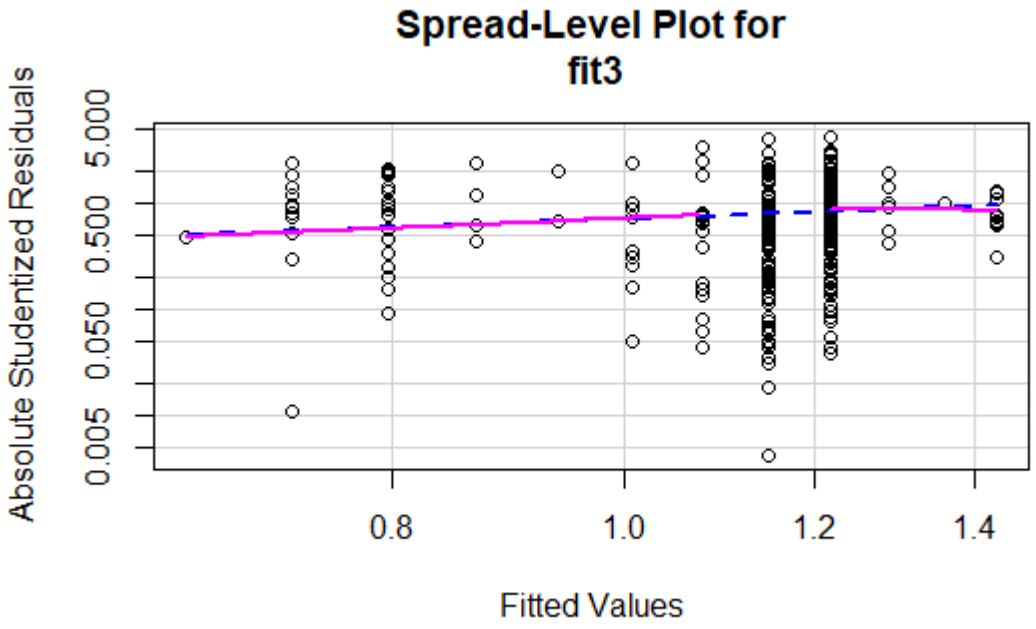
Model 1



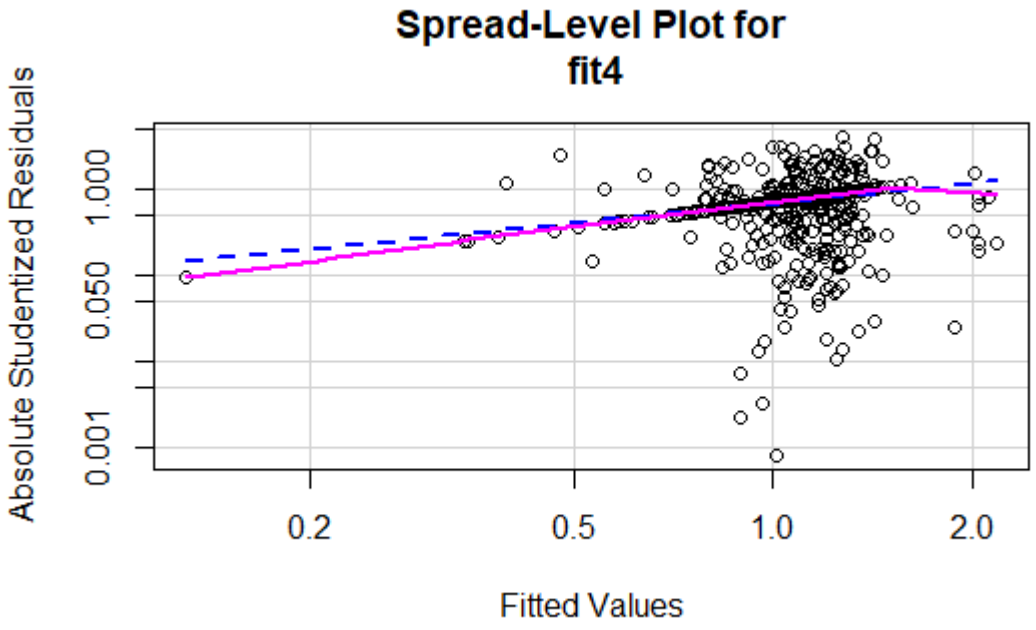
Model 2



Model 3



Model 4



GLOBAL TEST:

Model 1

###Global test

library(gvlma)

globvalmodel1 <- gvlma(fit1)

summary(globvalmodel1)

```
Call:
lm(formula = Area ~ X + Y + FFMC + DMC + DC + ISI + Temp + RH +
    Wind + Rain + forest$Month + forest$Day, data = forest)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1728 -1.0819 -0.5324  0.8366  5.5995

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3210832  1.3863173  -0.232   0.8169
X             0.0426562  0.0316338   1.348   0.1781
Y            -0.0021349  0.0601283  -0.036   0.9717
FFMC          0.0050199  0.0144648   0.347   0.7287
DMC           0.0023859  0.0015417   1.548   0.1223
DC           -0.0011811  0.0007001  -1.687   0.0922 .
ISI          -0.0253822  0.0168888  -1.503   0.1335
Temp          0.0075483  0.0174008   0.434   0.6646
RH           -0.0039050  0.0052393  -0.745   0.4564
Wind          0.0587710  0.0371763   1.581   0.1145
Rain          0.0757787  0.2115808   0.358   0.7204
forest$Month  0.1527399  0.0607600   2.514   0.0123 *
forest$Day    0.0132305  0.0323201   0.409   0.6825
---
```

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

Residual standard error: 1.388 on 504 degrees of freedom
Multiple R-squared:  0.03799,  Adjusted R-squared:  0.01509
F-statistic: 1.659 on 12 and 504 DF,  p-value: 0.07269


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        136.891 0.000000 Assumptions NOT satisfied!
Skewness           113.849 0.000000 Assumptions NOT satisfied!
Kurtosis            12.803 0.000346 Assumptions NOT satisfied!
Link Function        1.988 0.158582    Assumptions acceptable.
Heteroscedasticity   8.251 0.004072 Assumptions NOT satisfied!
```

Model 2

globvalmodel2 <- gvlma(fit2)

summary(globvalmodel2)

```
Call:
lm(formula = Area ~ Month + I(Month^2) + I(Month^3) + I(Month^4))

Residuals:
    Min      1Q  Median      3Q     Max
-1.7735 -1.0863 -0.6676  0.8929  5.8051

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.9861371  1.5784933   0.625    0.532
Month       -0.2030365  1.2003887  -0.169    0.866
I(Month^2)   0.0851006  0.3052106   0.279    0.780
I(Month^3)  -0.0120766  0.0318249  -0.379    0.704
I(Month^4)   0.0006009  0.0011619   0.517    0.605

Residual standard error: 1.387 on 512 degrees of freedom
Multiple R-squared:  0.02321,  Adjusted R-squared:  0.01558
F-statistic: 3.042 on 4 and 512 DF,  p-value: 0.017


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        169.800 0.000e+00 Assumptions NOT satisfied!
Skewness           134.768 0.000e+00 Assumptions NOT satisfied!
Kurtosis            25.037 5.624e-07 Assumptions NOT satisfied!
Link Function        2.041 1.532e-01    Assumptions acceptable.
Heteroscedasticity   7.954 4.798e-03 Assumptions NOT satisfied!
```

Model 3

globvalmodel3 <- gvlma(fit3)

summary(globvalmodel3)

```
Call:
lm(formula = Area ~ Month)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3585 -1.1478 -0.7096  0.8982  5.7776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.58610    0.21018   2.789  0.00549 **
Month        0.07022    0.02690   2.611  0.00930 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.391 on 515 degrees of freedom
Multiple R-squared:  0.01306,  Adjusted R-squared:  0.01114
F-statistic: 6.815 on 1 and 515 DF,  p-value: 0.009304


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        153.426 0.000e+00 Assumptions NOT satisfied!
Skewness           124.382 0.000e+00 Assumptions NOT satisfied!
Kurtosis            17.818 2.431e-05 Assumptions NOT satisfied!
Link Function        3.776 5.201e-02    Assumptions acceptable.
Heteroscedasticity   7.450 6.344e-03 Assumptions NOT satisfied!
```

Model 4

globvalmodel4 <- gvlma(fit4)

summary(globvalmodel4)

```
Call:
lm(formula = Area ~ Month + X + Wind + DC + DMC, data = forest)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0261 -1.0776 -0.5832  0.8711  5.7178

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0268920  0.3011193  -0.089  0.92887
Month        0.1580688  0.0590617   2.676  0.00768 **
X            0.0390629  0.0264425   1.477  0.14022
Wind         0.0438871  0.0356153   1.232  0.21842
DC          -0.0011622  0.0006668  -1.743  0.08195 .
DMC          0.0021135  0.0013984   1.511  0.13131
```

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

Residual standard error: 1.384 on 511 degrees of freedom
Multiple R-squared:  0.02999,  Adjusted R-squared:  0.0205
F-statistic:  3.16 on 5 and 511 DF,  p-value: 0.008065


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = fit4)

                     Value   p-value                   Decision
Global Stat        145.2800 0.000e+00 Assumptions NOT satisfied!
Skewness           119.7451 0.000e+00 Assumptions NOT satisfied!
Kurtosis            16.3122 5.372e-05 Assumptions NOT satisfied!
Link Function        0.6963 4.040e-01    Assumptions acceptable.
Heteroscedasticity   8.5264 3.500e-03 Assumptions NOT satisfied!
```

MULTICOLLINEARITY

```
vif(fit1)
            X             Y          FFMC           DMC            DC
     1.435206      1.465093      1.707992      2.611778      8.080790
          ISI          Temp            RH          Wind          Rain
     1.588530      2.734980      1.958044      1.188519      1.050469
forest$Month   forest$Day
     5.123223      1.037052
vif(fit2)
     Month I(Month^2) I(Month^3) I(Month^4)
  2000.637  19989.676  25481.795   3993.024
vif(fit4)
   Month         X     Wind       DC      DMC
4.867580 1.008344 1.096837 7.370947 2.160726
```

UNUSUAL OBSERVATIONS

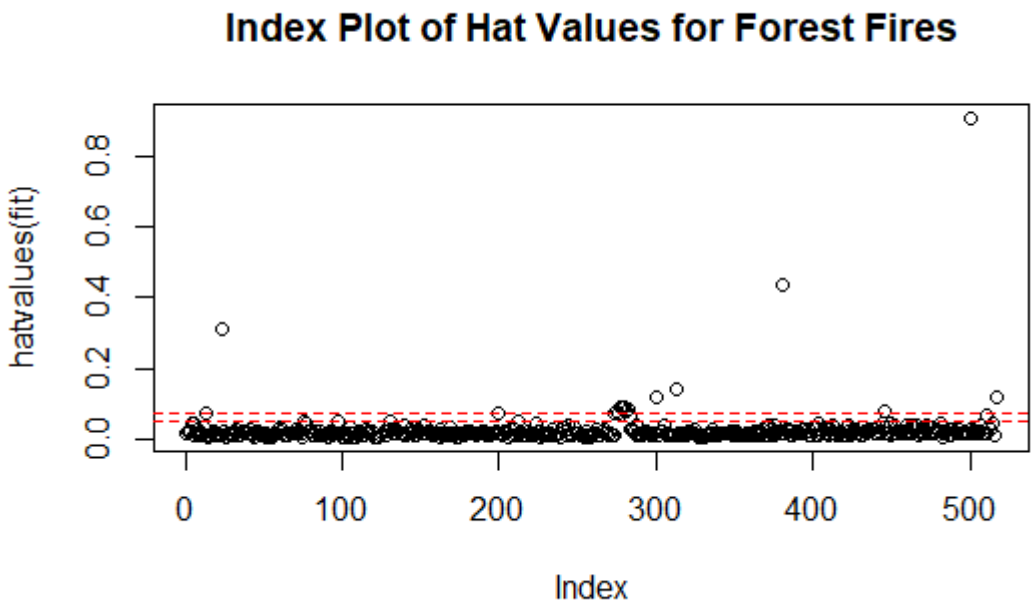
OUTLIERS

```
###Unusual observations- outliers
outlierTest(fit1)
    rstudent unadjusted p-value Bonferonni p
239  4.11376         4.5473e-05      0.02351
outlierTest(fit2)
    rstudent unadjusted p-value Bonferonni p
239 4.262530         2.4076e-05     0.012447
416 4.052022         5.8678e-05     0.030336
outlierTest(fit3)
    rstudent unadjusted p-value Bonferonni p
239 4.228123         2.7895e-05     0.014422
416 3.993213         7.4690e-05     0.038615
outlierTest(fit4)
    rstudent unadjusted p-value Bonferonni p
239 4.205782         3.0736e-05      0.01589
```

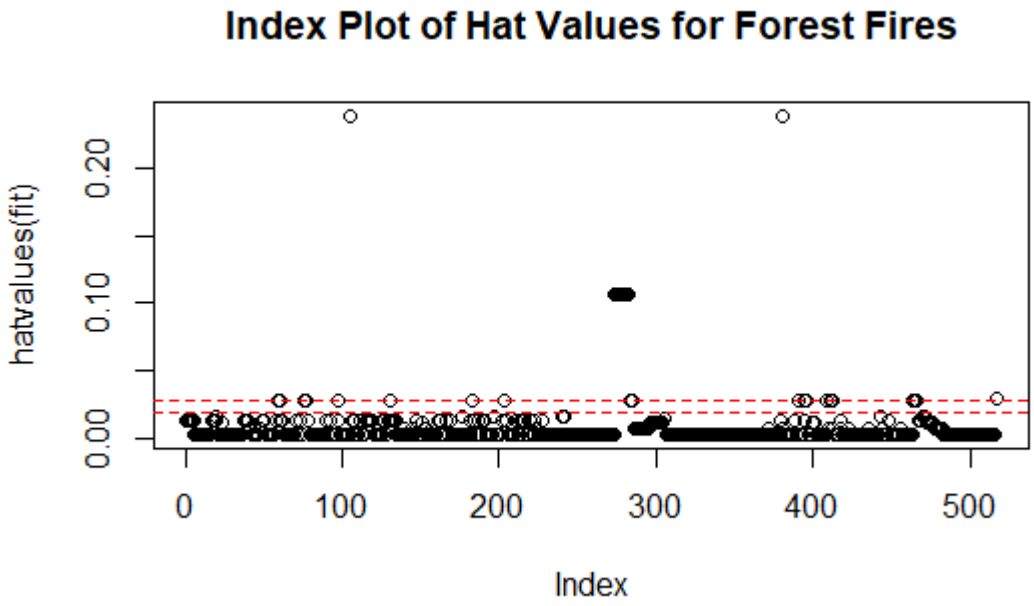
HIGH LEVERAGE POINTS

```
hat.plot <- function(fit){
  p <- length(coefficients(fit))
  n <- length(fitted(fit))
  plot(hatvalues(fit), main = "Index Plot of Hat Values for Forest Fires")
  abline(h=c(2,3)*p/n, col = "red", lty =2)
  identify(1:n, hatvalues(fit), names(hatvalues(fit)))
}
hat.plot(fit1)
hat.plot(fit2)
hat.plot(fit3)
hat.plot(fit4)
```

Model 1

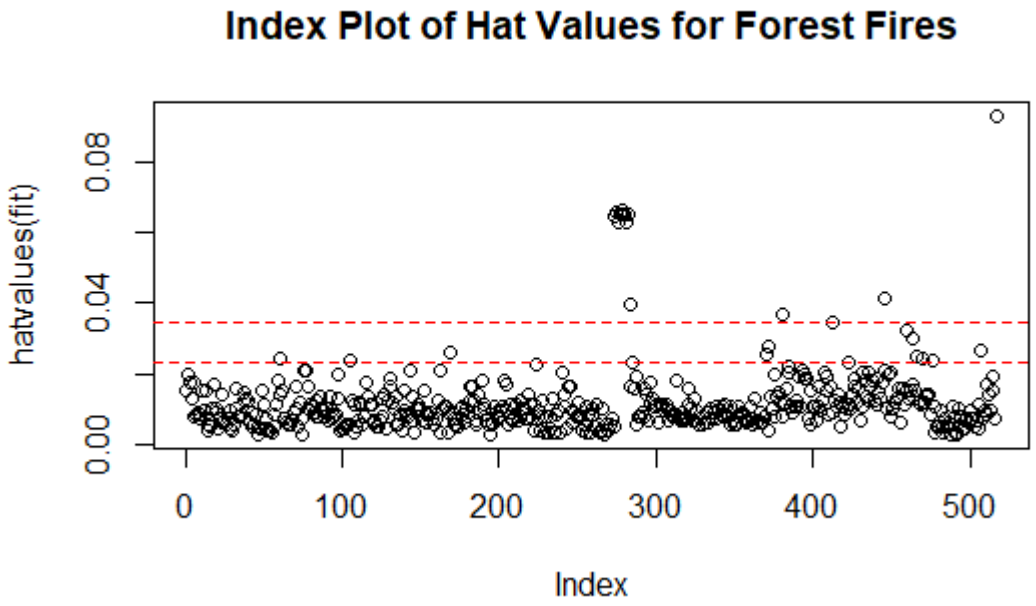


Model 2



Model 3

Model 4

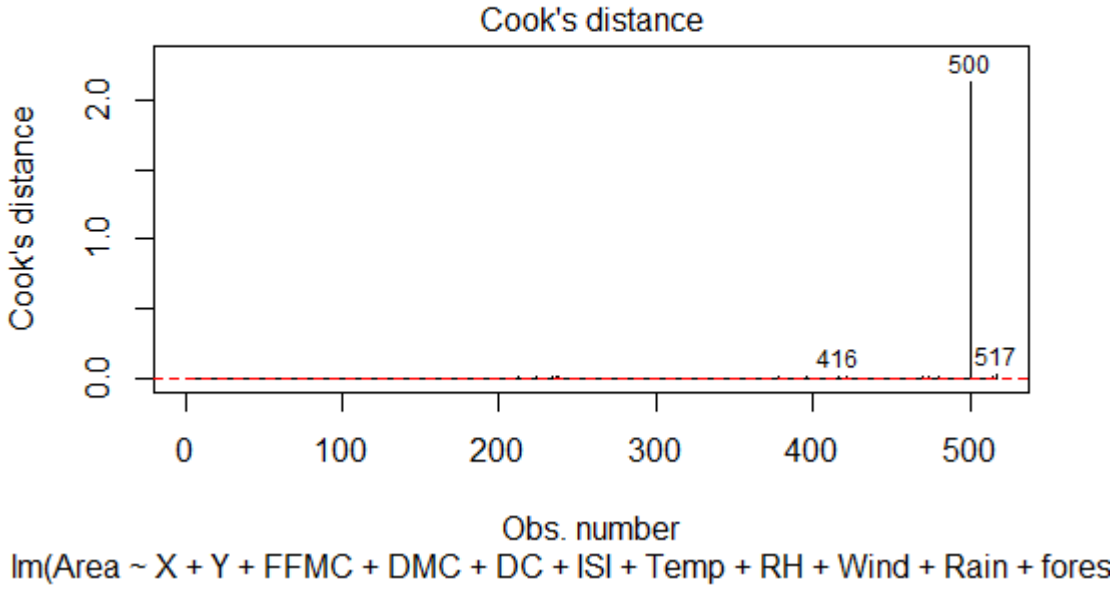


INFLUENTIAL OBSERVATIONS

Model 1

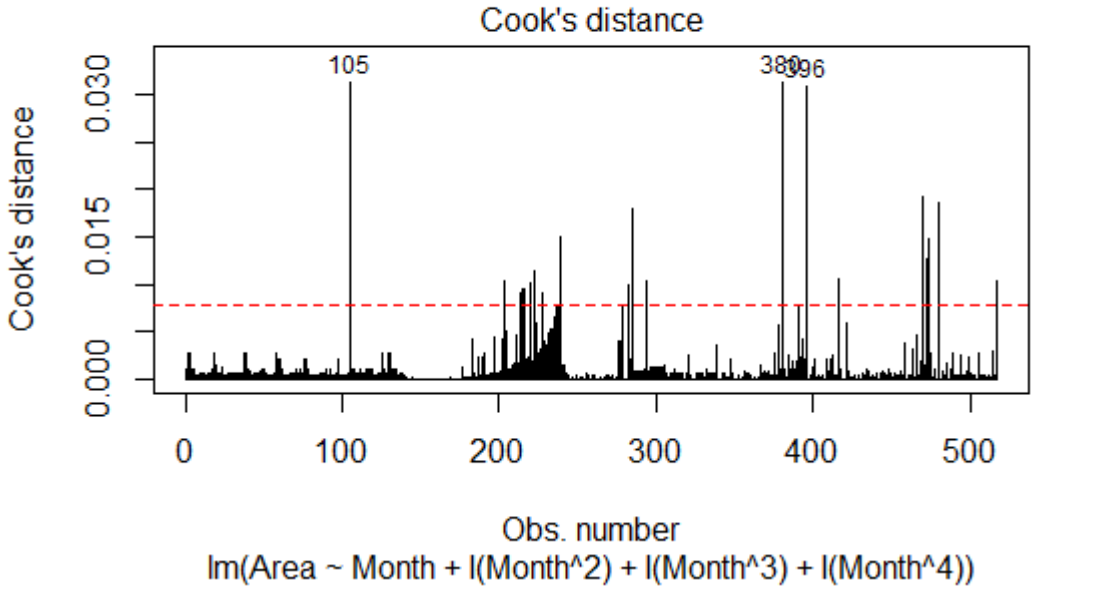
###Influential observations

```
cutoff <- 4/(nrow(forest)-length(fit1$coefficients)-2)
plot(fit1,which=4,cook.levels=cutoff)
abline(h=cutoff, lty=2, col="red")
```



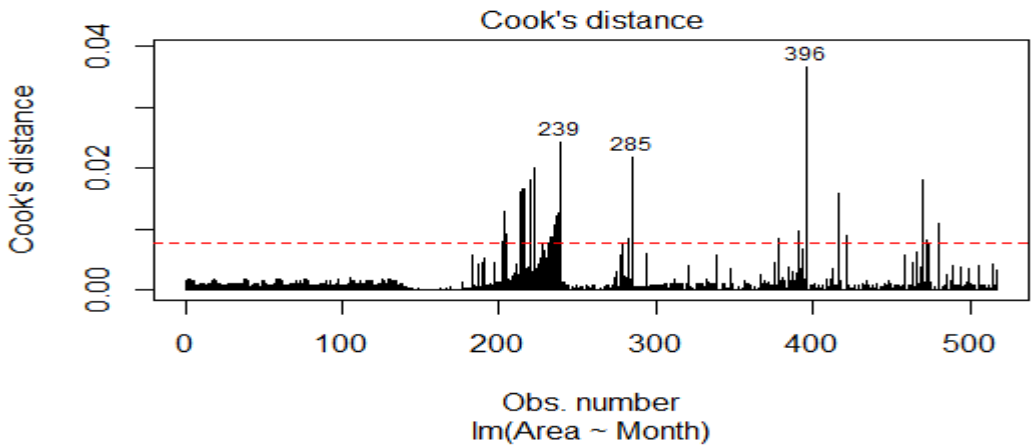
Model 2

```
cutoff <- 4/(nrow(forest)-length(fit2$coefficients)-2)
plot(fit2,which=4,cook.levels=cutoff)
abline(h=cutoff, lty=2, col="red")
```



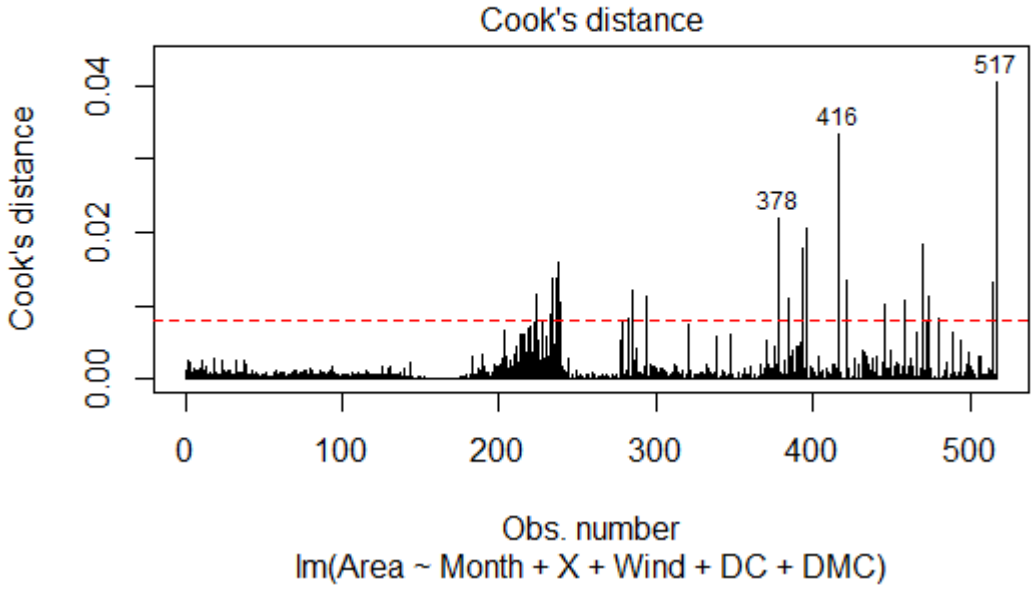
Model 3

```
cutoff <- 4/(nrow(forest)-length(fit3$coefficients)-2)
plot(fit3,which=4,cook.levels=cutoff)
abline(h=cutoff, lty=2, col="red")
```



Model 4

```
cutoff <- 4/(nrow(forest)-length(fit4$coefficients)-2)
plot(fit4,which=4,cook.levels=cutoff)
abline(h=cutoff, lty=2, col="red")
```



INFLUENCE PLOT

###Influence plot

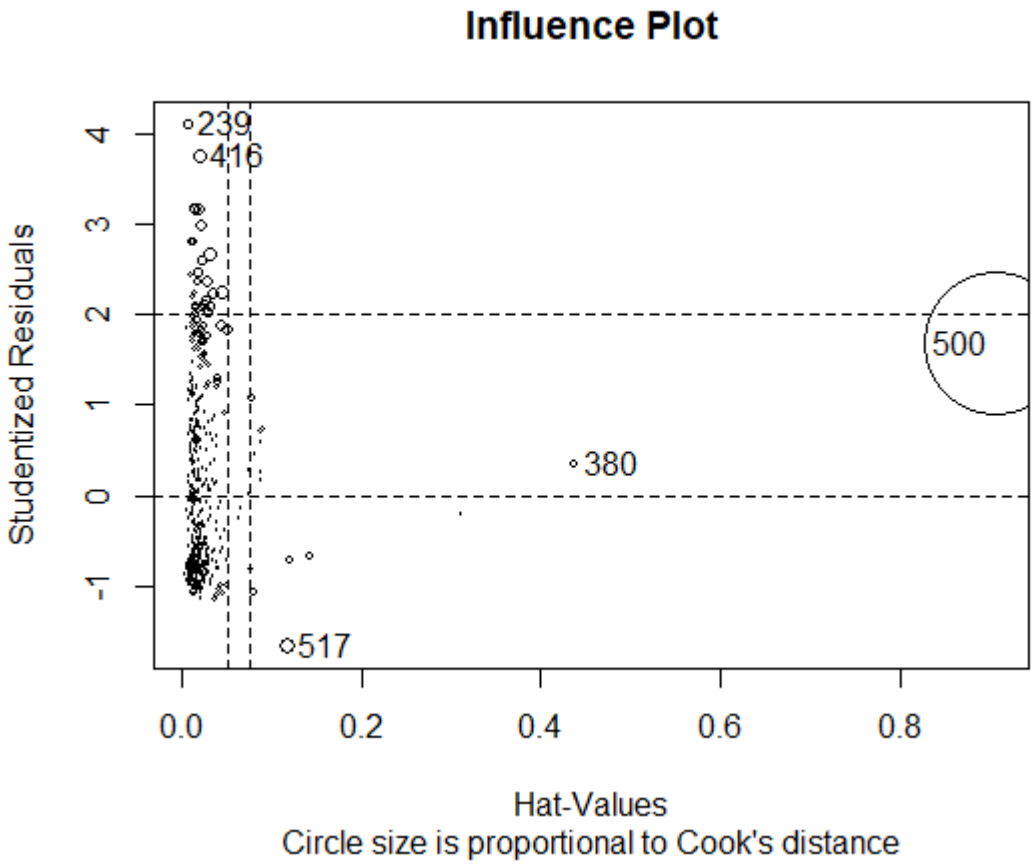
influencePlot(fit1, id.method="identify", main="Influence Plot", sub="Circle size is proportional to Cook's distance")

influencePlot(fit2, id.method="identify", main="Influence Plot", sub="Circle size is proportional to Cook's distance")

influencePlot(fit3, id.method="identify", main="Influence Plot", sub="Circle size is proportional to Cook's distance")

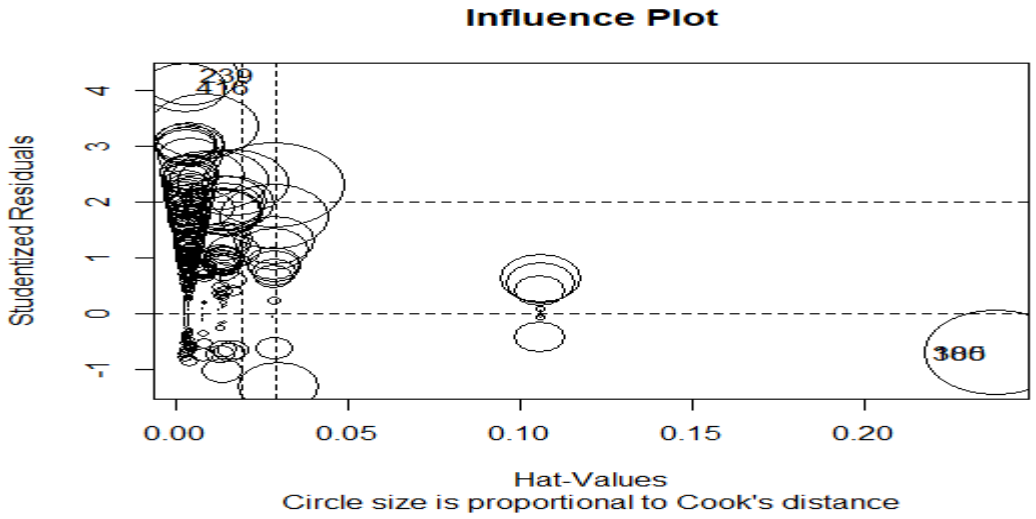
influencePlot(fit4, id.method="identify", main="Influence Plot", sub="Circle size is proportional to Cook's distance")

Model 1



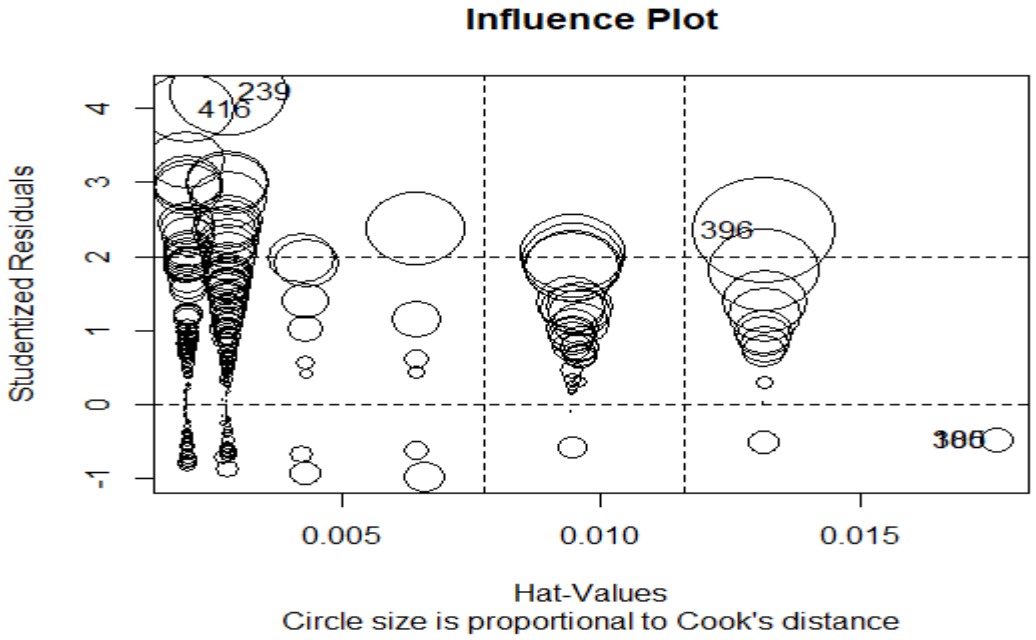

```
StudRes         Hat       CookD
239  4.1137596 0.007694119 0.009784514
380  0.3454901 0.438035056 0.007169467
416  3.7616053 0.019867561 0.021501917
500  1.6834509 0.907359351 2.127442281
517 -1.6701042 0.118153689 0.028645677
```

Model 2




```
StudRes         Hat      CookD
105 -0.7070585 0.238127422 0.03128183
239  4.2625304 0.004257334 0.01503249
380 -0.7070585 0.238127422 0.03128183
416  4.0520218 0.003325258 0.01063555
```

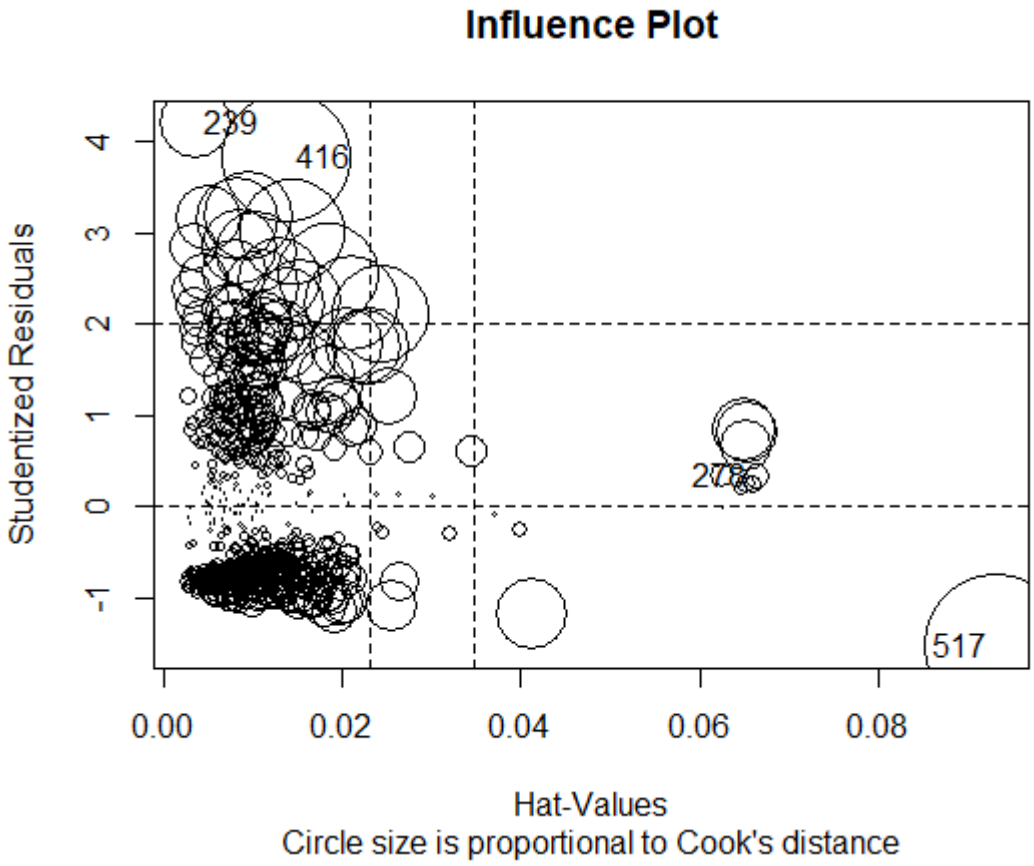
Model 3




```
StudRes         Hat      CookD
105 -0.4758127 0.017623383 0.00203379
239  4.2281233 0.002803358 0.02433095
380 -0.4758127 0.017623383 0.00203379
396  2.3554107 0.013152049 0.03664612
416  3.9932128 0.002037030 0.01581516
```

Model 4

```
StudRes          Hat       CookD
239  4.2057824 0.003599533 0.010313302
278  0.3397338 0.066348900 0.001369393
416  3.8285612 0.013835661 0.033382224
517 -1.5393680 0.093233003 0.040499077
```



CORRECTIVE MEASURES

```
sqrt(vif(fit1))>2
           X            Y         FFMC          DMC           DC
       FALSE        FALSE        FALSE        FALSE         TRUE
         ISI         Temp           RH         Wind         Rain
       FALSE        FALSE        FALSE        FALSE        FALSE
forest$Month   forest$Day
        TRUE        FALSE

sqrt(vif(fit2))>2
     Month I(Month^2) I(Month^3) I(Month^4)
      TRUE       TRUE       TRUE       TRUE

sqrt(vif(fit4))>2
Month     X  Wind    DC   DMC
 TRUE FALSE FALSE  TRUE FALSE
```

forest_data <- forest[-c(510,380,500),]

newfit <- fit4 <- lm(Area~Month+X+Wind+DC+DMC,data=forest[-c(239,416),])

outlierTest(newfit)

```
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
478  3.31012         0.00099882      0.51439
```

BEST REGRESSION MODEL

```
AIC(fit1, fit2, fit3, fit4)
     df      AIC
fit1 14 1820.913
fit2  6 1812.797
fit3  3 1812.141
fit4  7 1773.456
```

FINE TUNE THE SELECTION OF PREDICTOR VARIABLES

###Fine tune

library(MASS)

step_fit <- lm(Area ~ X + Y + FFMC + ISI + DMC + DC + Temp + Wind + Rain, data = forest)

stepAIC(step_fit, direction = "backward")

```
Start:  AIC=353.79
Area ~ X + Y + FFMC + ISI + DMC + DC + Temp + Wind + Rain

       Df Sum of Sq    RSS    AIC
- Rain  1    0.0642 986.08 351.83
- Y     1    0.1220 986.14 351.86
- DMC   1    0.3838 986.40 351.99
- DC    1    1.2836 987.30 352.46
- FFMC  1    1.3935 987.41 352.52
- Temp  1    1.9024 987.92 352.79
- X     1    2.7730 988.79 353.24
<none>              986.02 353.79
- ISI   1    5.2208 991.24 354.52
- Wind  1    9.3708 995.39 356.68

Step:  AIC=351.83
```

```
Area ~ X + Y + FFMC + ISI + DMC + DC + Temp + Wind

       Df Sum of Sq    RSS    AIC
- Y     1    0.1214 986.20 349.89
- DMC   1    0.3995 986.48 350.03
- DC    1    1.2737 987.36 350.49
- FFMC  1    1.3974 987.48 350.56
- Temp  1    1.9407 988.02 350.84
- X     1    2.8335 988.92 351.31
<none>              986.08 351.83
- ISI   1    5.2063 991.29 352.55
- Wind  1    9.5266 995.61 354.80

Step:  AIC=349.89
Area ~ X + FFMC + ISI + DMC + DC + Temp + Wind

       Df Sum of Sq    RSS    AIC
- DMC   1    0.4622 986.67 348.13
- DC    1    1.1925 987.40 348.51
- FFMC  1    1.3672 987.57 348.60
- Temp  1    1.9712 988.17 348.92
<none>              986.20 349.89
- X     1    4.9342 991.14 350.47
- ISI   1    5.2430 991.45 350.63
- Wind  1    9.4501 995.65 352.82

Step:  AIC=348.13
Area ~ X + FFMC + ISI + DC + Temp + Wind

       Df Sum of Sq    RSS    AIC
- FFMC  1    1.5888 988.25 346.96
- Temp  1    2.2683 988.93 347.32
- DC    1    3.2975 989.96 347.86
<none>              986.67 348.13
- X     1    4.9728 991.64 348.73
- ISI   1    5.0409 991.71 348.77
- Wind  1    9.6743 996.34 351.18

Step:  AIC=346.96
Area ~ X + ISI + DC + Temp + Wind

       Df Sum of Sq    RSS    AIC
- Temp  1    3.1757 991.43 346.62
- ISI   1    3.5648 991.82 346.82
<none>              988.25 346.96
- DC    1    4.0689 992.32 347.09
- X     1    4.9595 993.21 347.55
- Wind  1    9.6617 997.92 349.99

Step:  AIC=346.62
Area ~ X + ISI + DC + Wind

       Df Sum of Sq     RSS    AIC
- ISI   1    1.7669  993.20 345.54
<none>              991.43 346.62
- X     1    4.7953  996.23 347.12
- Wind  1    7.7737  999.20 348.66
- DC    1    9.0999 1000.53 349.35

Step:  AIC=345.54
Area ~ X + DC + Wind

       Df Sum of Sq     RSS    AIC
<none>              993.20 345.54
- X     1    4.6437  997.84 345.95
- Wind  1    6.8023 1000.00 347.07
- DC    1    7.6548 1000.85 347.51
```

```
Call:
lm(formula = Area ~ X + DC + Wind, data = forest)

Coefficients:
(Intercept)            X           DC         Wind
  0.3801571    0.0411523    0.0005033    0.0654530
```

INTERPRETATION

All the above regression models has low Adjusted R values so the simple linear regression and multiple linear regression may not be the best regression model for the given dataset. It can be seen that the normality is violated through qqplot. Independence of errors is not satisfied and there is no correlation between the response variable and the predictor variables. The constant variance assumption is also not met. Considering the above conclusions, fit4 seems to be the best model among the four models evaluated but it may not be an ideal one.