HW 3 Regression Analysis

Anushka Tak Chandan Manjunath

February 16, 2019

Contribution of each student

Anushka Tak 50%

Chandan Manjunath 50%

Signature

Student 1 : Anushka

Student 2 : Chandan

Problem2

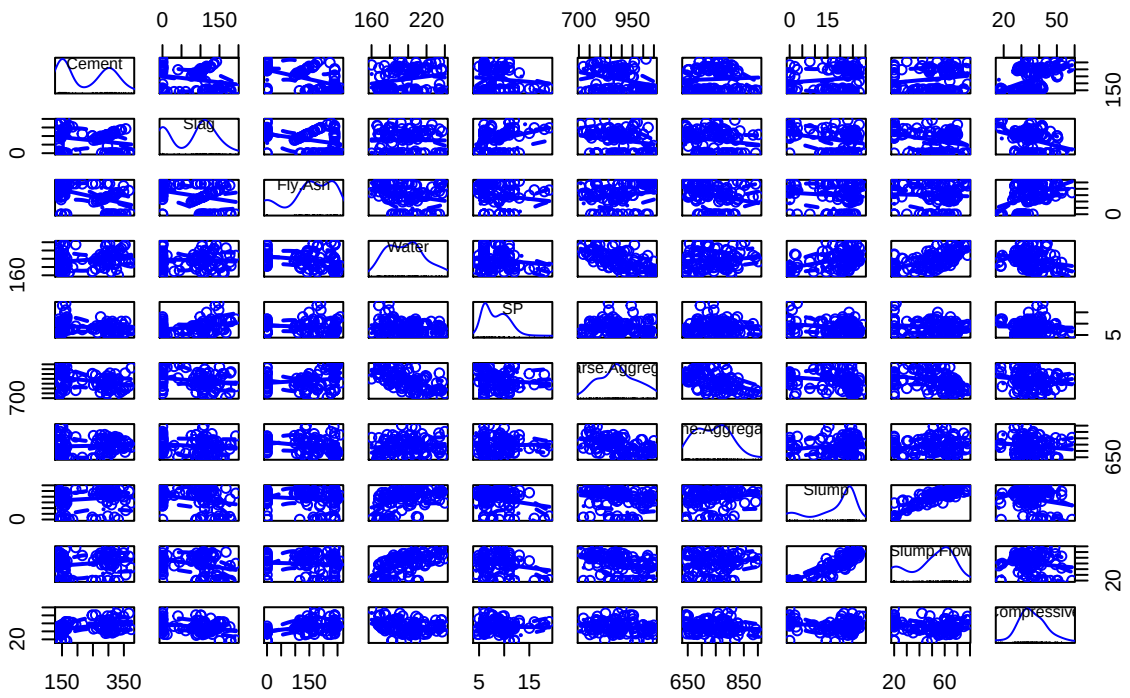
Task 3: Concrete Slump Test Data

Predictor Variables . Cement . Slag . Fly-ash . Water . SP . Coarse-Aggregate . Fine-Aggregate

Response Variables . Slump . Slump Flow . 28 Day Compressive Strength

Scatter Plot Matrix Removing first No(id like) variable

```
# install.packages("readxl")
# install.packages("car")
library(car)
library(readxl)
con_test=read_excel("C:/Users/kkavi/Desktop/anushka/Concrete_Slump_Test_Data.xlsx")
scatterplotMatrix(con_test[-1],spread=FALSE,lty.smooth=2,main="Scatter Plot Matrix")
```



Almost all the graphs of predictors are skewed except for Coarse Aggregate and 28 Day Compressive Strength, which resembles a normal distribution

. Cement increases linearly with 28 Day Compressive Strength, water, fine aggregate, slump, and slump flow.

. Slag increases linearly with SP.

. Fly ash increases linearly with coarse aggregates and 28 Day Compressive Strength.

. Water has a linear relationship with cement, fine aggregate, slump and slump flow.

. Fine Aggregates has a weak linear relationship with slump and slump flow.

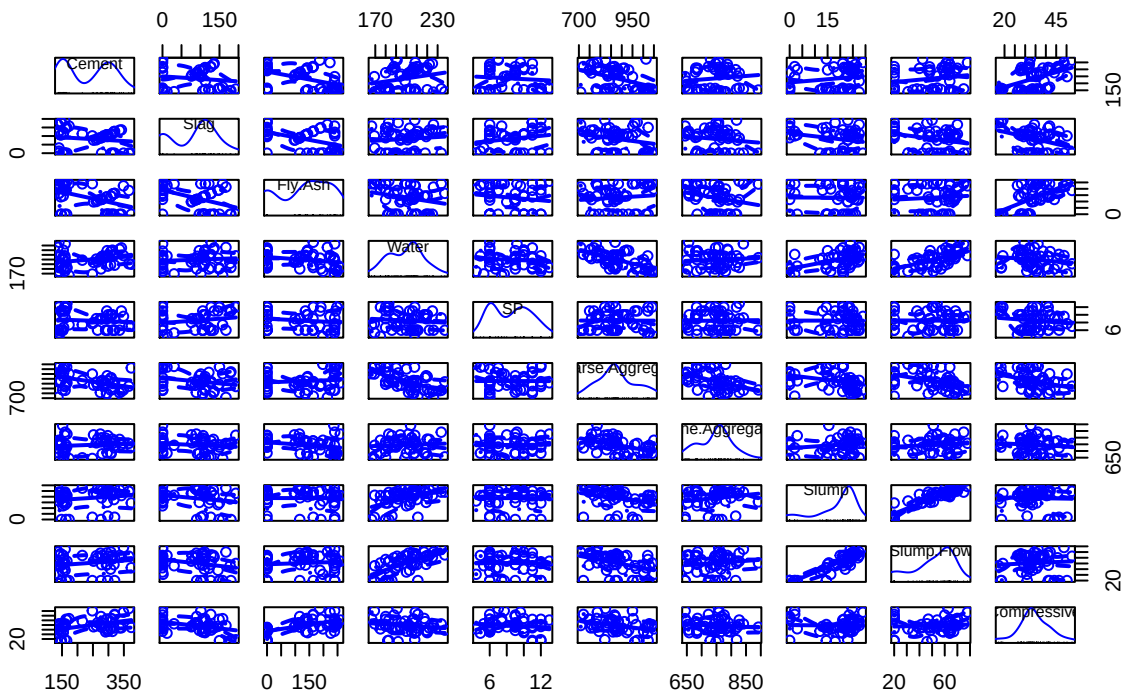
. Slump shows linear relationship with cement(weak), slump flow, water and fine aggregate(weak).

. Slump flow shows linear relationship with slump, water, cement and fine aggregate(weak).

. 28 Day Compressive length shows a linear relationship with fly ash(weak) and Cement.

The principal diagonal contains density and rug plots for each variable.

```
mysample <- con_test[sample(1:nrow(con_test), 50,
                            replace=FALSE),]
scatterplotMatrix(mysample[-1],spread=FALSE,lty.smooth=2,main="Scatter Plot Matrix of a subset")
```



In addition to the ones already mentioned,

. We find Slag to be linearly related to water(weak).

. We find fly ash to be linearly related toslump and slump flow.

From this, we use {water, cement, fine aggregate and slump} as our initial set of predictors to predict slump flow in our dataset.

Typical Approach

Building Potential Regression Models Model 1: With all the predictors Model 2: Potential Predictors from Scatterplot Matrix deductions {water, cement, fine aggregateand slump} Model 3: Model with potential predictors with interactions

```
library(data.table)
setnames(con_test, old=c("Slump Flow","Fine Aggregate","Coarse Aggregate"), new=c("slump_flow", "fine_ag
model1<- lm(slump_flow~.,data=con_test)
summary(model1)
```

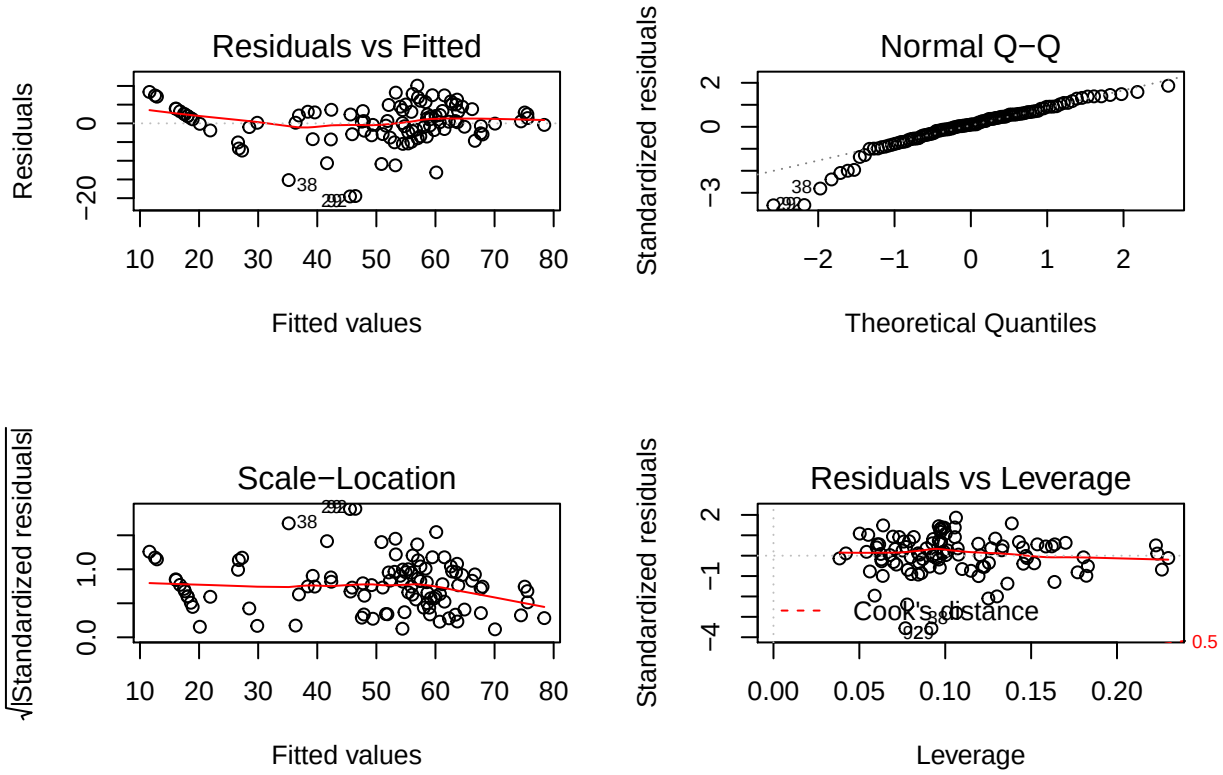
```
##
## Call:
## lm(formula = slump_flow ~ ., data = con_test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -19.5414  -2.5321   0.5258   3.2530  10.0769
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1.765e+02  2.050e+02  -0.861   0.3915
## No                            -2.797e-04  3.019e-02  -0.009   0.9926
## Cement                         1.016e-02  7.011e-02   0.145   0.8851
## Slag                           2.887e-02  9.370e-02   0.308   0.7587
## `Fly Ash`                      2.902e-02  7.155e-02   0.406   0.6860
## Water                          4.265e-01  2.062e-01   2.068   0.0414 *
## SP                             5.447e-01  3.066e-01   1.777   0.0789 .
## coarse_aggregate               5.199e-02  8.030e-02   0.647   0.5190
## fine_aggregate                 5.084e-02  8.200e-02   0.620   0.5368
## Slump                          1.588e+00  8.337e-02  19.052   <2e-16 ***
## `28-day Compressive Strength`  4.506e-01  2.370e-01   1.901   0.0604 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.716 on 92 degrees of freedom
## Multiple R-squared:  0.9045, Adjusted R-squared:  0.8941
## F-statistic: 87.15 on 10 and 92 DF,  p-value: < 2.2e-16
```

```
confint(model1)
```

```
##                                    2.5 %       97.5 %
## (Intercept)                  -583.70332450 230.69767663
## No                             -0.06023576   0.05967643
## Cement                         -0.12908197   0.14940937
## Slag                           -0.15721535   0.21496156
## `Fly Ash`                      -0.11308872   0.17112106
## Water                           0.01695076   0.83612935
## SP                             -0.06422406   1.15365967
## coarse_aggregate               -0.10749837   0.21147726
## fine_aggregate                 -0.11201904   0.21369351
## Slump                           1.42272750   1.75386905
## `28-day Compressive Strength`  -0.02005622   0.92134593
```

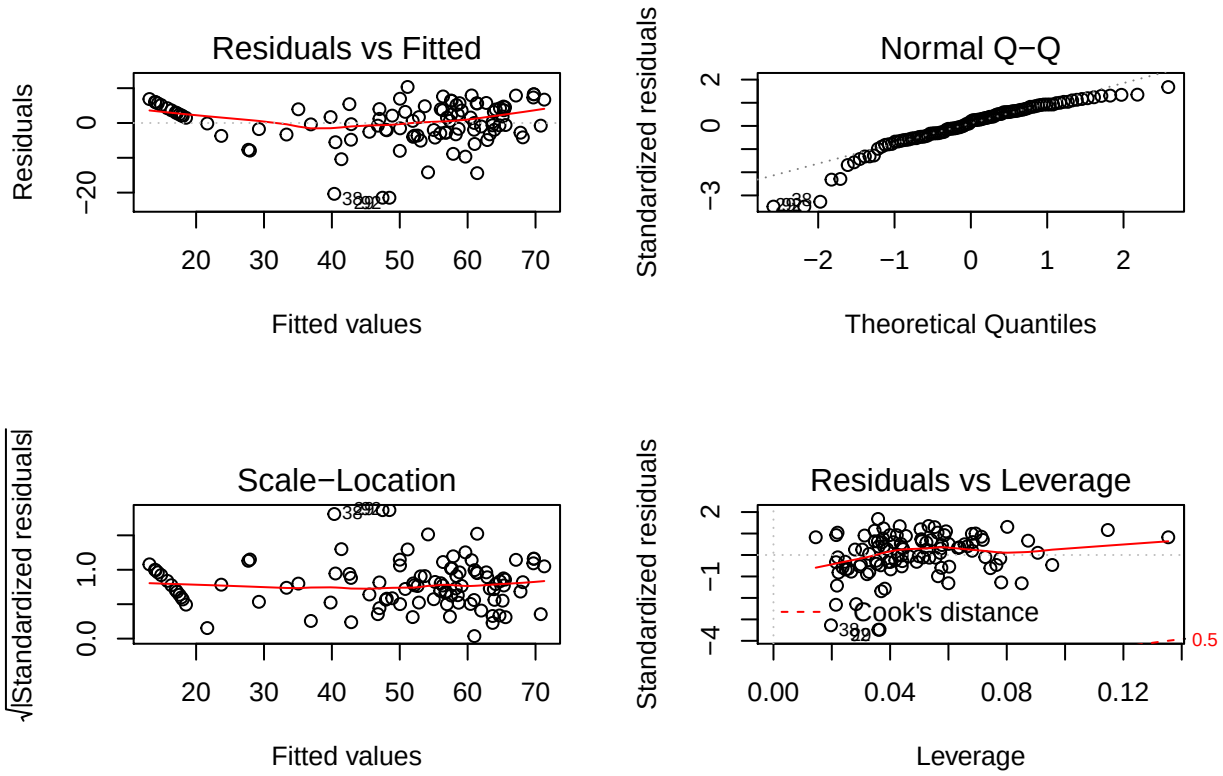
```
par(mfrow = c(2,2))
plot(model1)
```




```
model2<- lm(slump_flow~Water+Cement+fine_aggregate+Slump
,data=con_test)
summary(model2)
```

```
##
## Call:
## lm(formula = slump_flow ~ Water + Cement + fine_aggregate + Slump,
##     data = con_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.477  -2.890   1.186   4.242  10.344
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.504e+01  9.461e+00  -2.647  0.00946 **
## Water           2.301e-01  3.534e-02   6.512 3.19e-09 ***
## Cement          3.118e-03  8.096e-03   0.385  0.70093
## fine_aggregate  3.898e-04  1.003e-02   0.039  0.96908
## Slump           1.567e+00  8.160e-02  19.199  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.28 on 98 degrees of freedom
## Multiple R-squared:  0.8772, Adjusted R-squared:  0.8722
## F-statistic: 175.1 on 4 and 98 DF,  p-value: < 2.2e-16
```

```r
plot(model2)
```

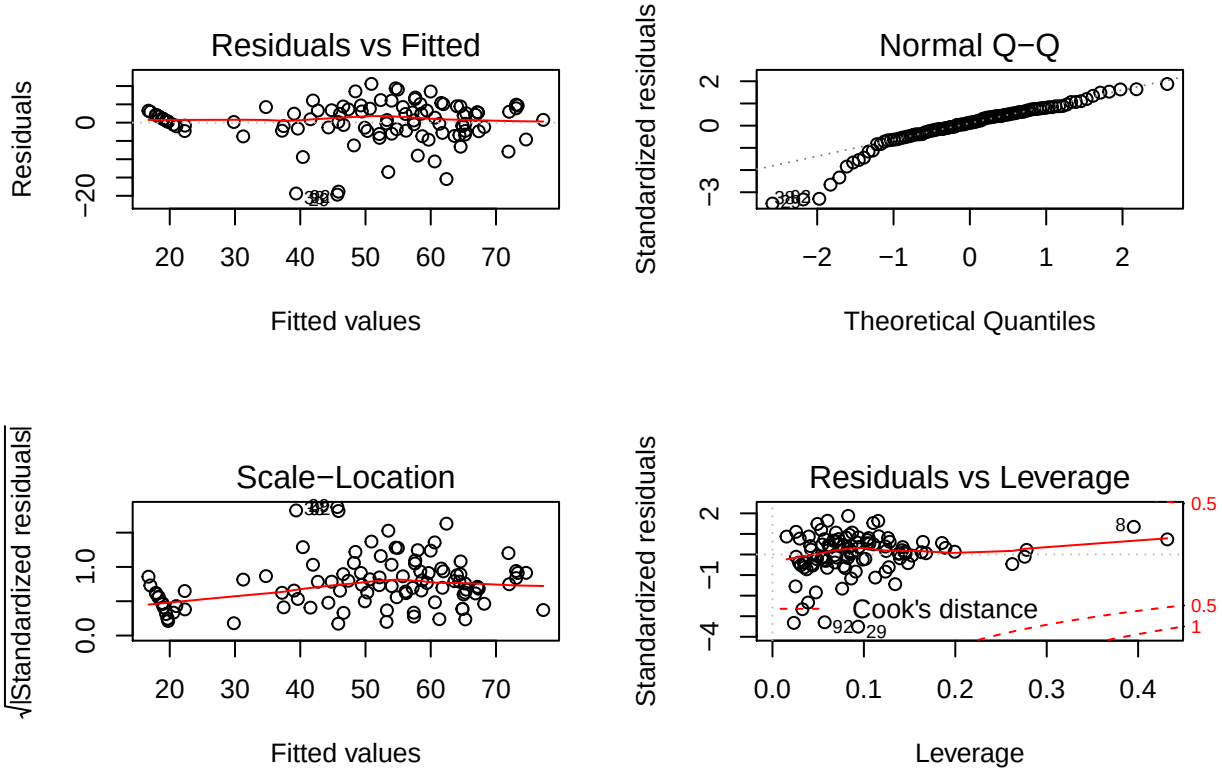



```r
model3<- lm(slump_flow~Water+Cement+fine_aggregate+Slump+Cement:Water+Cement:fine_aggregate+Cement:Slump
summary(model3)
```

```
##
## Call:
## lm(formula = slump_flow ~ Water + Cement + fine_aggregate + Slump +
##     Cement:Water + Cement:fine_aggregate + Cement:Slump + Water:fine_aggregate +
##     Water:Slump, data = con_test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.6867  -2.2413   0.7109   3.3030  10.5900
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9.325e+01  7.724e+01   1.207  0.23036
## Water                  -4.632e-01  3.667e-01  -1.263  0.20960
## Cement                  6.596e-02  1.434e-01   0.460  0.64669
## fine_aggregate         -8.750e-02  1.046e-01  -0.837  0.40489
## Slump                  -1.417e+00  7.842e-01  -1.806  0.07411 .
## Water:Cement           -5.747e-05  4.181e-04  -0.137  0.89098
## Cement:fine_aggregate -8.255e-05  1.454e-04  -0.568  0.57157
## Cement:Slump            6.816e-04  1.030e-03   0.662  0.50961
```

```
## Water:fine_aggregate   5.534e-04  4.836e-04   1.145  0.25535
## Water:Slump            1.533e-02  4.166e-03   3.681  0.00039 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.894 on 93 degrees of freedom
## Multiple R-squared:  0.8974, Adjusted R-squared:  0.8874
## F-statistic: 90.35 on 9 and 93 DF,  p-value: < 2.2e-16
```

```
plot(model3)
```



Model 1:

Normality: Most of the observations lie on the 45 degree line in Q-Q plot but still a significant number does not lie on the line, suggesting our assumption of normality was violated of dependent variable linearly related to the predictors.

Independence: From the scatterplots, we have found some independent variables related to others.

Linearity:There is a systematic relationship between residuals and fitted value(more like a quadratic relationship) thereby we conclude the model used is not an effective one. This hints to adding a quadratic curve to our model yet, again.

Homoscedasticity: The points in the Scale-Location graph seem to be a random band around the horizontal line therefore meeting constant variance assumption.

Outlier: Observations 29,38 and 92 to have large negative residuals from Fitted-Residuals graph, they are potential outliers.

Influential Observations: Observations 29,92 appear to be influential having the lagest Cook's distance.

Model 2:

Normality: Most of the observations lie on the 45 degree line in Q-Q plot but still a significant number does not lie on the line, suggesting our assumption of normality was violated of dependent variable linearly related to the predictors.

Independence: From the scatterplots, we have found some independent variables related to others. In this model, we do not acknowledge their interactions among themselves.

Linearity:There is less of a systematic relationship between residuals and fitted value in comparison with Model 1 yet a weak curve can be observed, thereby intimating a quadratic term to our model.

Homoscedasticity: The points in the Scale-Location graph seem to be a random band around the horizontal line therefore meeting constant variance assumption.

Outlier: Observations 29,38, and 92 to have large negative residuals from Fitted-Residuals graph, they are potential outliers.

Influential Observations: Observations 29,38 and 92 appear to be influential having the lagest Cook's distance.

Model 3:

Linearity:There is a systematic relationship between residuals and fitted value(more like a quadratic relationship) thereby we conclude the model used is not an effective one. This hints to adding a quadratic curve to our model yet, again.

Independence: We have considered predcitors with interactions among them on this model.

Linearity:There is no systematic relationship between residuals and fitted value suggesting this is a good model to capture the relationship of slump flow with the taken predictors.

Homoscedasticity: The points in the Scale-Location graph seem to be a random band around the horizontal line therefore meeting constant variance assumption.

Outlier: Observations 29, 38 and 92 to have large negative residuals from Fitted-Residuals graph, they are potential outliers.

Influential Observations: Observations 29,92 appear to be influential having the lagest Cook's distance.

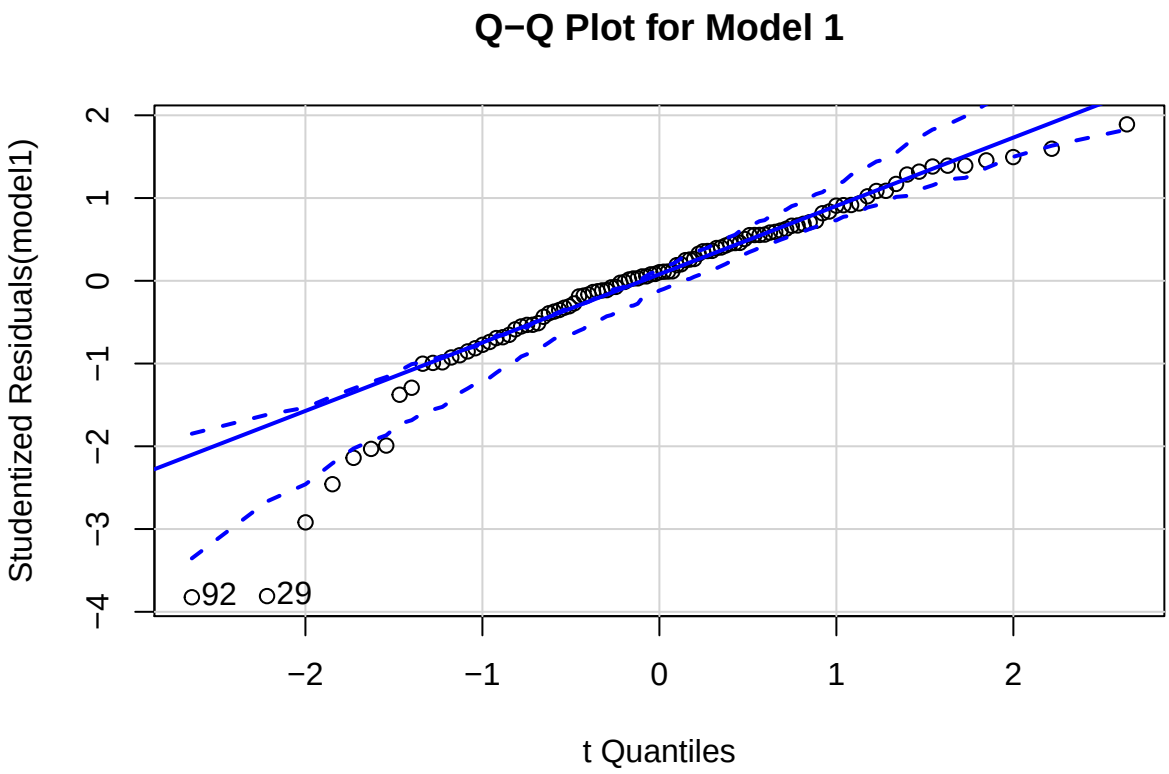
In the model, the Pr(>|t|) column shows that the interactions between predictors is significant. A significant interaction between 2 predictor variables tells that the relationship between one predictor and the response variable depends on the level of the other predictor. Here, the residual standard error has decreased, total variance has increased and adjusted R-squared has also increased.

In model 1, we can say with 95% confidence that the interval [1.422,1.75] conatins true change in slump flow for a 1% change in slump. And other such conclusions can be drawn from Summary and conflint results.

Enhanced Approach

Normality

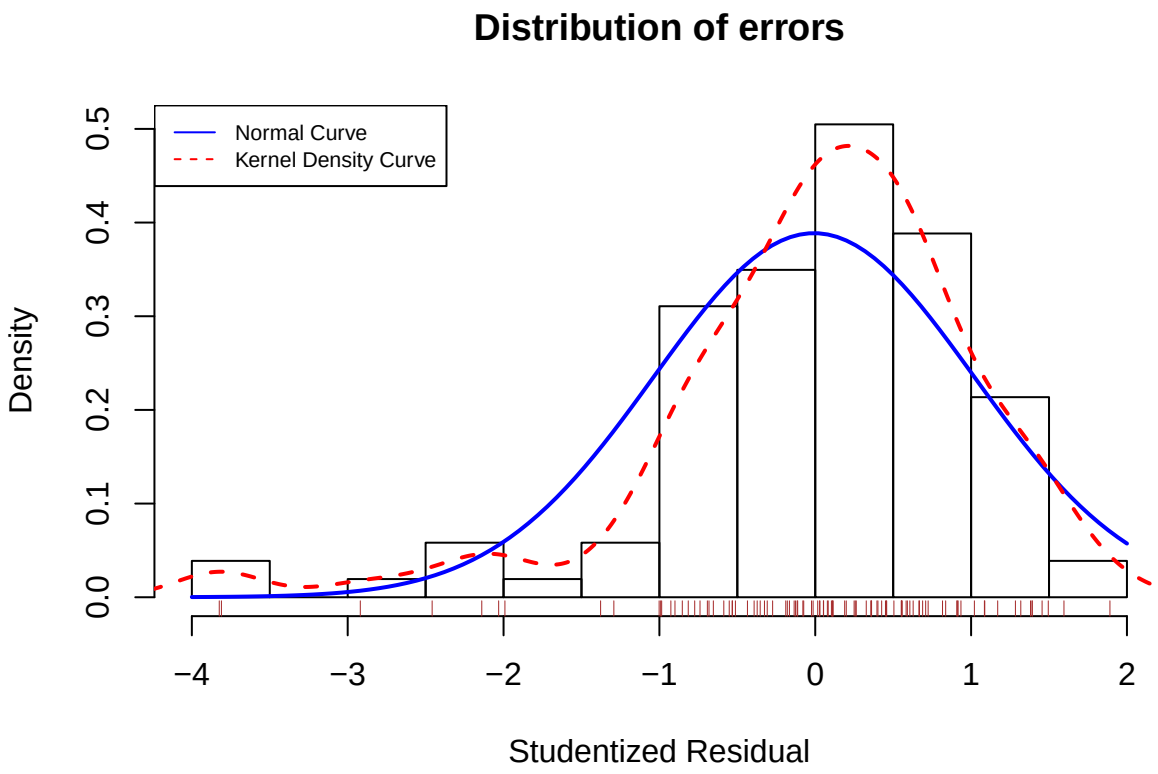
```
qqPlot(model1, labels = row.names(conc_slump), id.method = "identify",  simulate = TRUE, main = "Q-Q Pl
```




```
## [1] 29 92
```

In this model, all the points, except 29 and 92 fall close to the line and are within the confidence envelope, suggesting that the normality assumption is met fairly well.

```
residplot <- function(model1, nbreaks = 10)
{
  z <- rstudent(model1)
  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("topleft", legend = c("Normal Curve", "Kernel Density Curve"), lty = 1:2, col = c("blue", "red
}
residplot(model1)
```



The residplot() function generates a histogram of the studentized residuals and superimposes a normal curve, kernel density curve and rug plot. It is clear that the errors follow the normal distribution quite well, with no outliers.

```
durbinWatsonTest(model1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1757677      2.347453   0.192
##  Alternative hypothesis: rho != 0
```

```
durbinWatsonTest(model2)
```

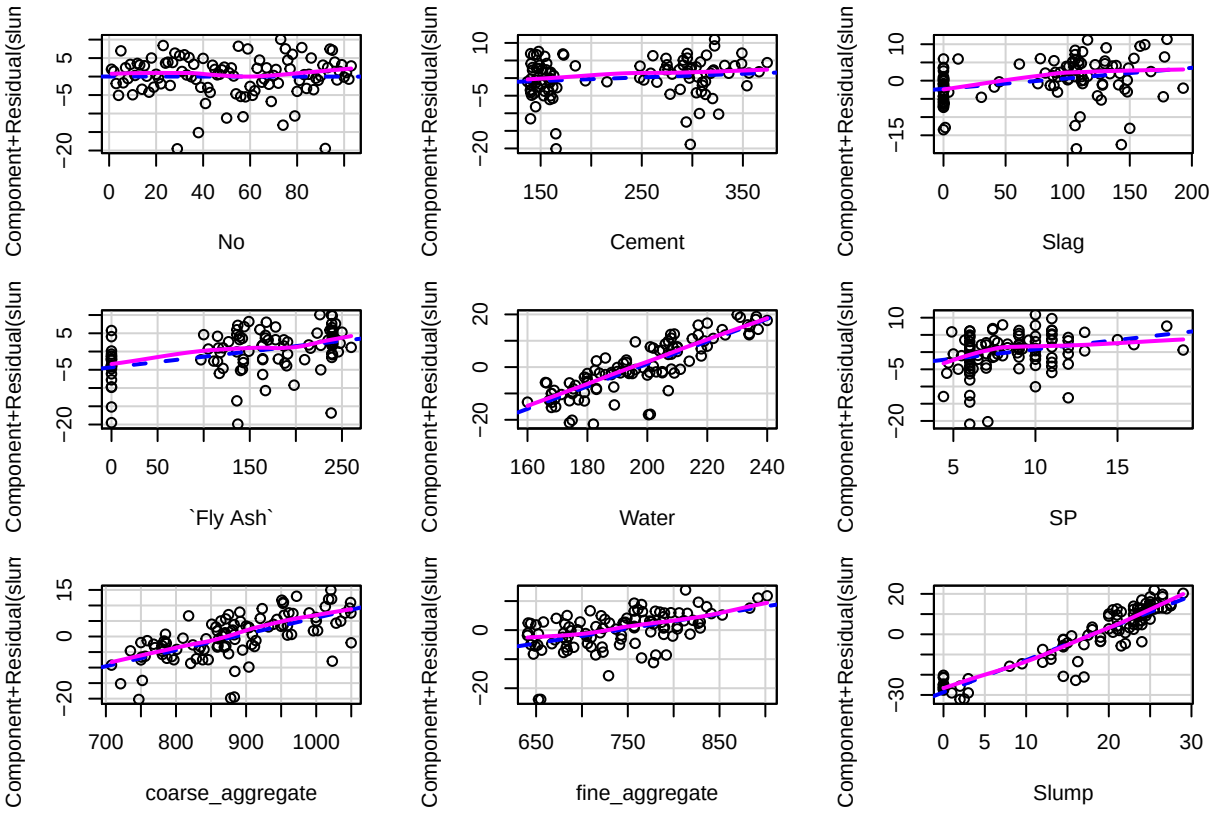
```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.08496791        2.1517   0.524
##  Alternative hypothesis: rho != 0
```

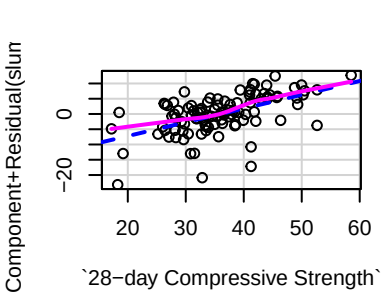
```
durbinWatsonTest(model3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1215336      2.235756    0.31
##  Alternative hypothesis: rho != 0
```

The car package provides the function Durbin-Watson test to detect the serially correlated errors. The non-significant p-values- 0.162 for model 1, 0.53 for model 2, and 0.32 for model 3, suggests a lack of autocorrelation, and conversely an independence of errors. The lag value 1 in all these cases indicate that each observation is being compared with the one next in the dataset.

```
crPlots(model1)
```





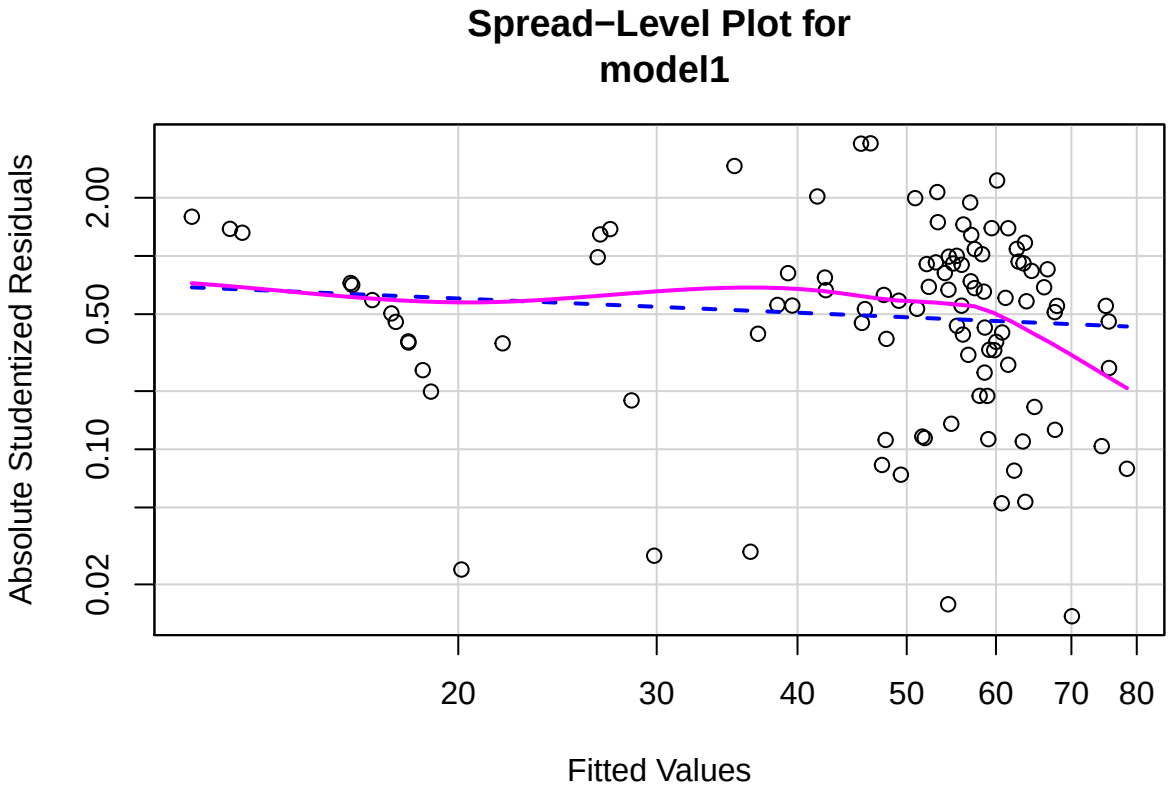
Component plus residual plots (also known as partial residual plots) help to identify the non-linearity in the relationship between the dependent variable and the independent variables. Non-linearity in any of these plots suggest that the functional form of that predictor in the regression is not adequately modeled. For model 1, the component plus residual plot confirms that we have met the linearity assumption except for slag, SP and fly ash.

Homoscedasticity

```
ncvTest(model1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.341524, Df = 1, p = 0.24677
```

```
spreadLevelPlot(model1)
```




```
##
## Suggested power transformation:  1.244087
```

```
#install.packages("gvlma")
library(gvlma)
gvlmodel1<-gvlma(model1)
summary(gvlmodel1)
```

```
##
## Call:
## lm(formula = slump_flow ~ ., data = con_test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.5414  -2.5321   0.5258   3.2530  10.0769
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1.765e+02  2.050e+02  -0.861   0.3915
## No                          -2.797e-04  3.019e-02  -0.009   0.9926
## Cement                       1.016e-02  7.011e-02   0.145   0.8851
## Slag                         2.887e-02  9.370e-02   0.308   0.7587
## `Fly Ash`                    2.902e-02  7.155e-02   0.406   0.6860
## Water                        4.265e-01  2.062e-01   2.068   0.0414 *
## SP                           5.447e-01  3.066e-01   1.777   0.0789 .
```

```
## coarse_aggregate               5.199e-02  8.030e-02   0.647   0.5190
## fine_aggregate                 5.084e-02  8.200e-02   0.620   0.5368
## Slump                          1.588e+00  8.337e-02  19.052   <2e-16 ***
## `28-day Compressive Strength`  4.506e-01  2.370e-01   1.901   0.0604 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.716 on 92 degrees of freedom
## Multiple R-squared:  0.9045, Adjusted R-squared:  0.8941
## F-statistic: 87.15 on 10 and 92 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 = model1)
##
##                      Value   p-value                   Decision
## Global Stat        72.1772 7.883e-15 Assumptions NOT satisfied!
## Skewness           25.2635 5.001e-07 Assumptions NOT satisfied!
## Kurtosis           24.6714 6.798e-07 Assumptions NOT satisfied!
## Link Function      21.3439 3.838e-06 Assumptions NOT satisfied!
## Heteroscedasticity  0.8984 3.432e-01    Assumptions acceptable.
```

```
vif(model1)
```

```
##                           No                       Cement
##                     2.539422                     95.465319
##                         Slag                    `Fly Ash`
##                   100.178295                   116.599474
##                        Water                           SP
##                    54.216261                    2.313020
##             coarse_aggregate               fine_aggregate
##                   157.272897                    84.211822
##                        Slump `28-day Compressive Strength`
##                     1.661293                    10.772227
```

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

```
##                           No                       Cement
##                        FALSE                         TRUE
##                         Slag                    `Fly Ash`
##                         TRUE                         TRUE
##                        Water                           SP
##                         TRUE                        FALSE
##             coarse_aggregate               fine_aggregate
##                         TRUE                         TRUE
##                        Slump `28-day Compressive Strength`
##                        FALSE                         TRUE
```

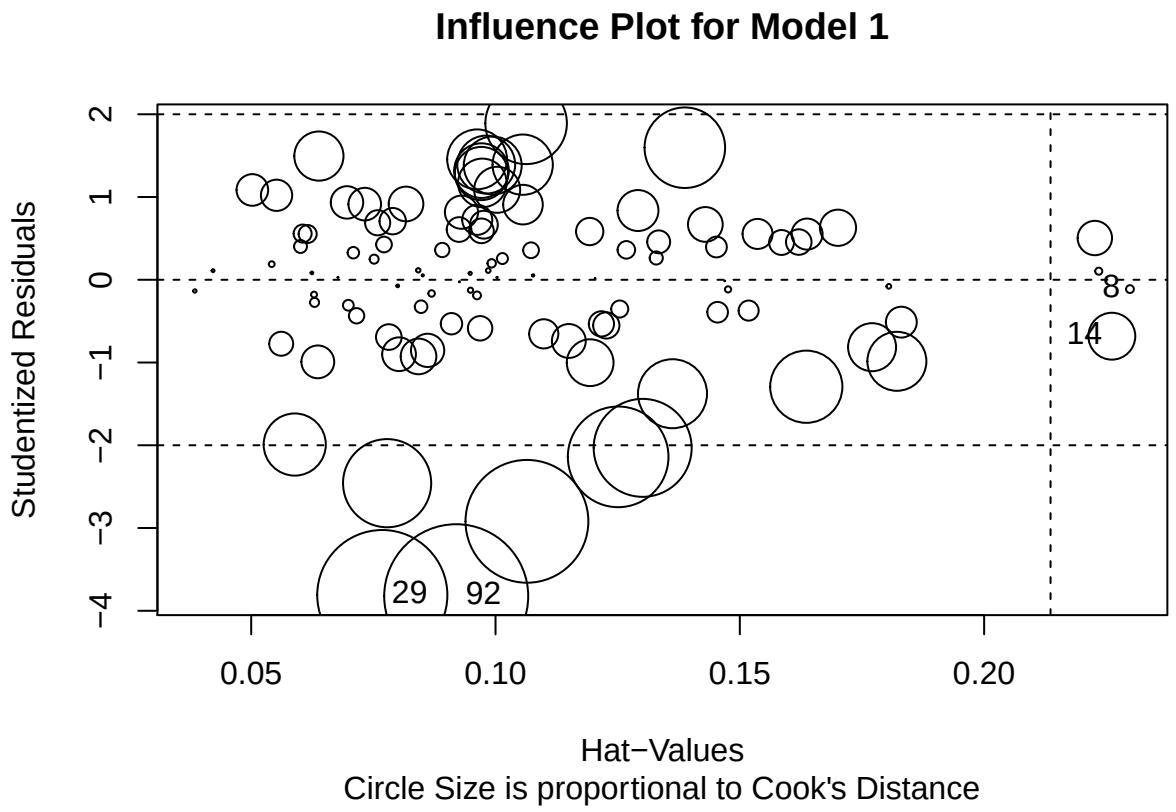
Thus, model 1 indicates the multicollinearity problem since the sqrt(vif(model1)) > 2 for 7 of the 9 predictors.

```r
outlierTest(model1)
```

```
##     rstudent unadjusted p-value Bonferonni p
## 92 -3.824090         0.00024013     0.024734
## 29 -3.810423         0.00025186     0.025942
```

As expected, 92 and 29 are observed as outliers with indicated significance.

```r
hat.plot1<-function(model1){
  p<- length(coefficients(model1))
  n<- length(fitted(model1))
  plot(hatvalues(model1), main ="Index values of Hat values")
  abline(h=c(2,3)*p/n,col="red",lty=2)
  identify(1:n,hatvalues(model1),names(hatvalues(model1)))
}
influencePlot(model1, id.method = "identify", main = "Influence Plot for Model 1", sub = "Circle Size is
```

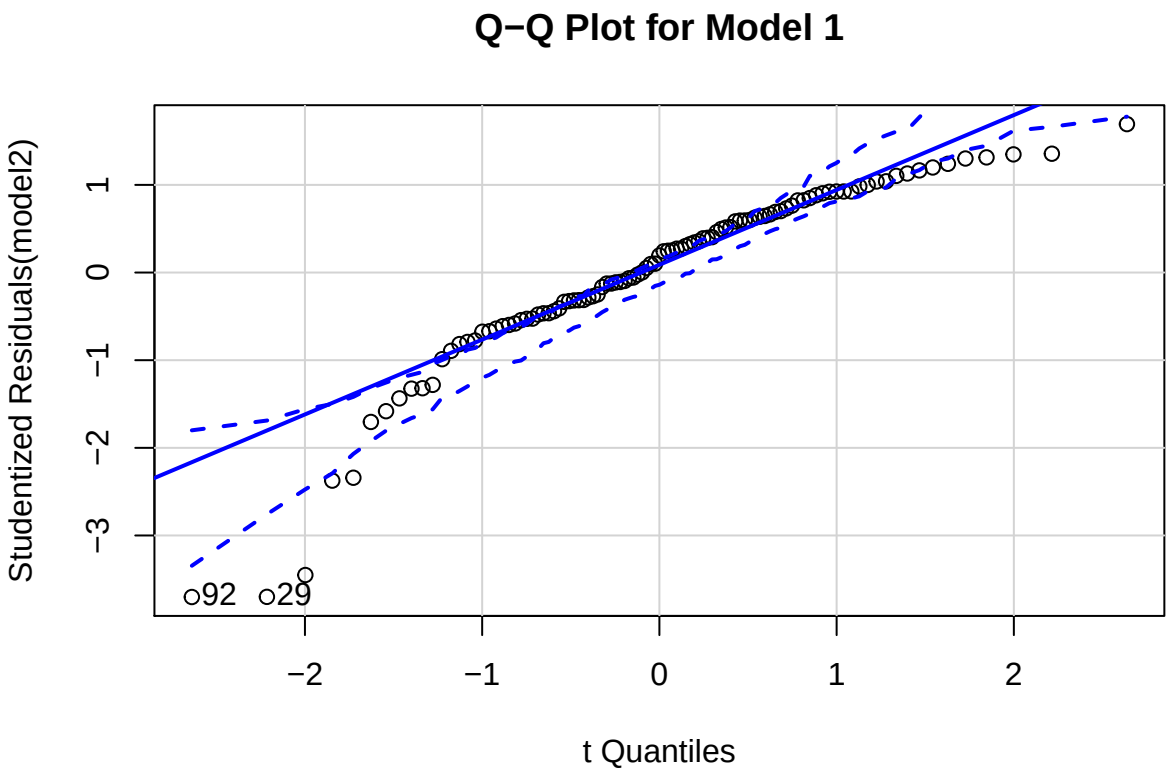



```
##       StudRes        Hat        CookD
## 8  -0.1128021 0.22984514 0.0003489673
## 14 -0.6818485 0.22613708 0.0124229319
## 29 -3.8104228 0.07682913 0.0957752508
## 92 -3.8240902 0.09195396 0.1172607707
```

Model 2

Normality

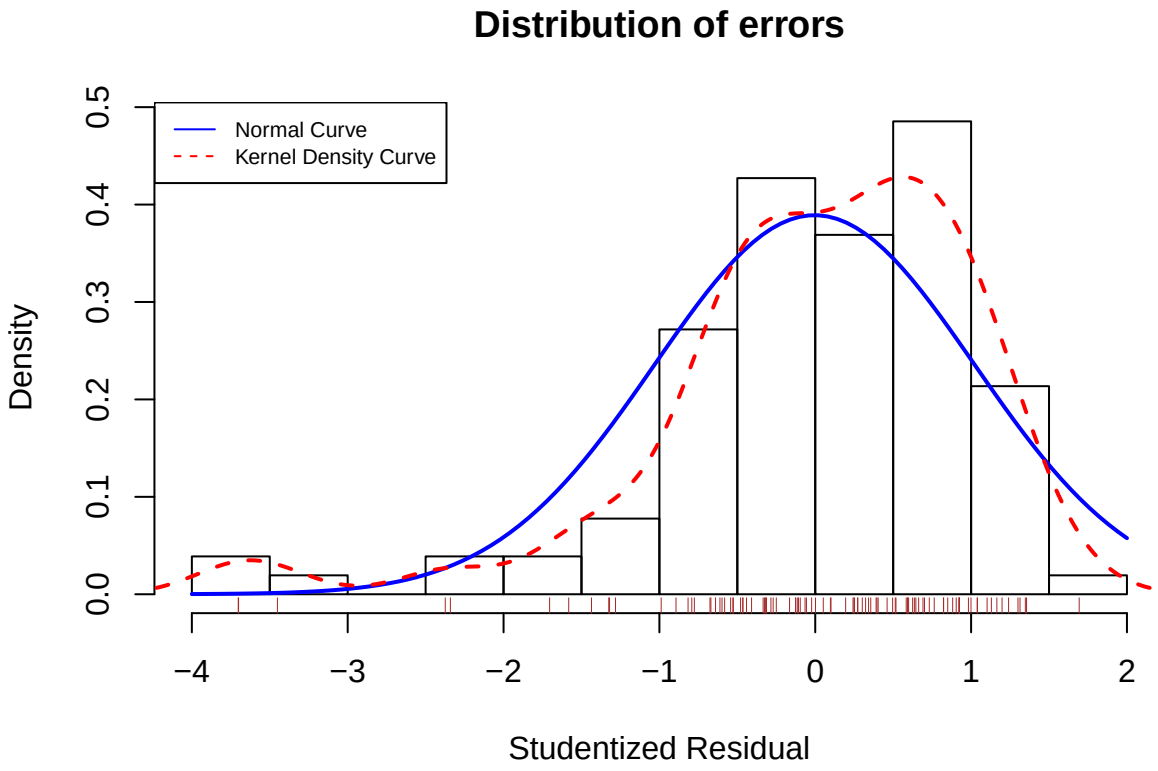
```
qqPlot(model2, labels = row.names(conc_slump), id.method = "identify",  simulate = TRUE, main = "Q-Q Pl
```




```
## [1] 29 92
```

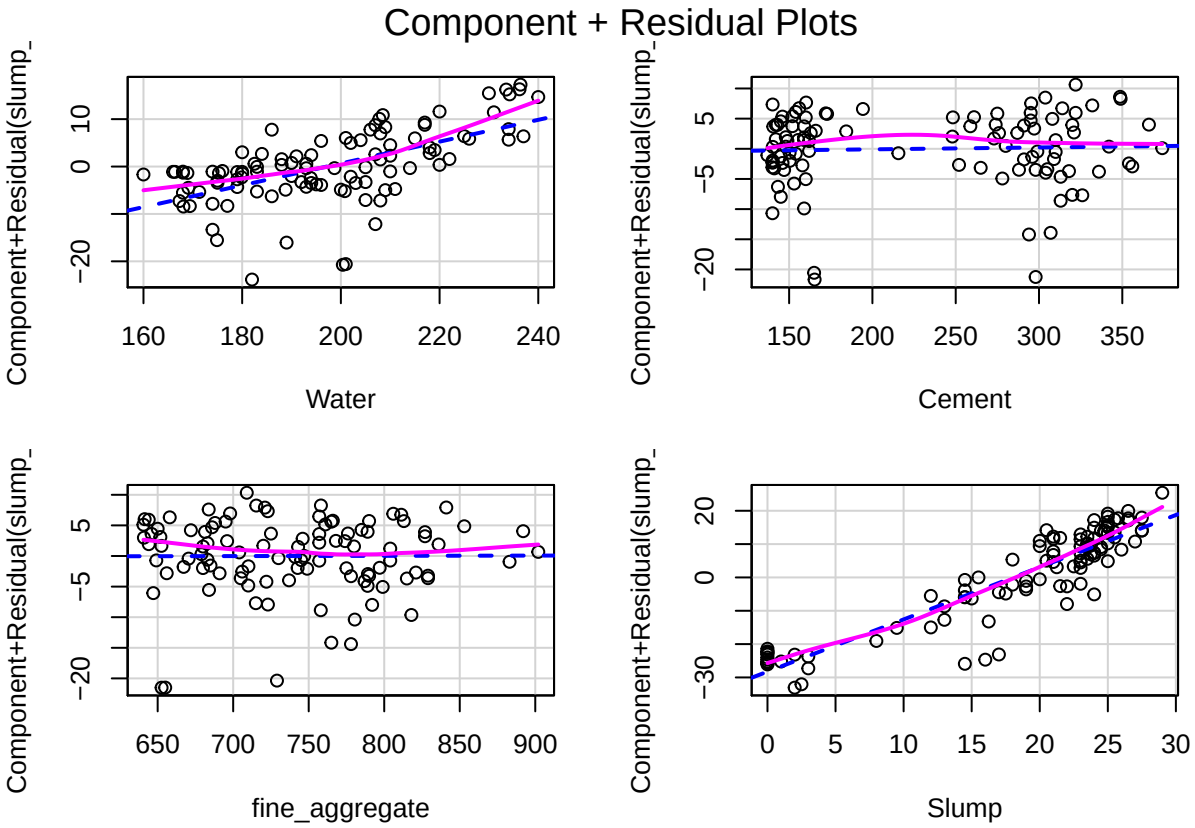
In this model, we observe Obs29 and 92 to be distant from the 45 degree line, others are on the line hence the normality assumption is fairly met.

```
residplot <- function(model2, nbreaks = 10)
{
  z <- rstudent(model2)
  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("topleft", legend = c("Normal Curve", "Kernel Density Curve"), lty = 1:2, col = c("blue", "red
}
residplot(model2)
```



The residplot() function generates a histogram of the studentized residuals and superimposes a normal curve, kernel density curve and rug plot. It is clear that the errors follow the normal distribution quite well, with no outliers.

```
crPlots(model2)
```

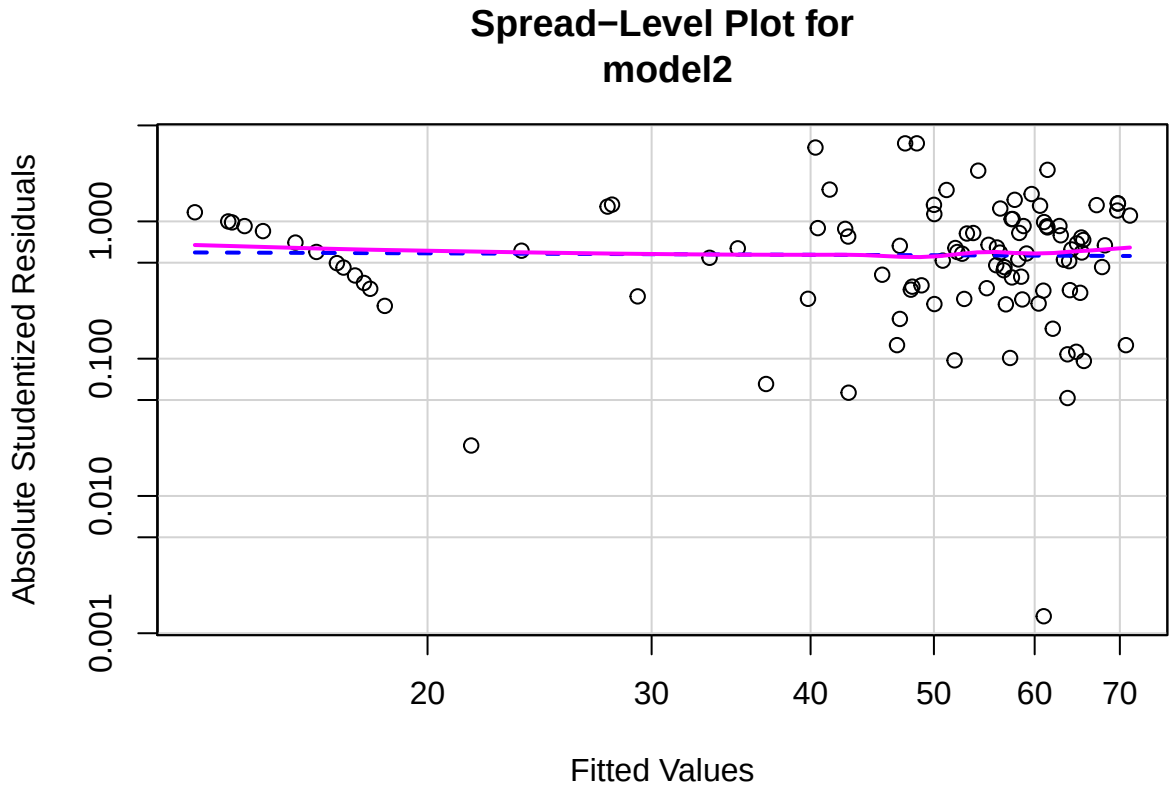


For model2, the linearity assumption is violated for cement, water and fine_aggregate, suggesting for a polynomial regression.'

```
ncvTest(model2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.0007898203, Df = 1, p = 0.97758
```

```
spreadLevelPlot(model2)
```




```
##
## Suggested power transformation:  1.034788
```

```
gvlmodel2<-gvlma(model2)
summary(gvlmodel2)
```

```
##
## Call:
## lm(formula = slump_flow ~ Water + Cement + fine_aggregate + Slump,
##     data = con_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.477  -2.890   1.186   4.242  10.344
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.504e+01  9.461e+00  -2.647  0.00946 **
## Water           2.301e-01  3.534e-02   6.512 3.19e-09 ***
## Cement          3.118e-03  8.096e-03   0.385  0.70093
## fine_aggregate  3.898e-04  1.003e-02   0.039  0.96908
## Slump           1.567e+00  8.160e-02  19.199  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6.28 on 98 degrees of freedom
## Multiple R-squared:  0.8772, Adjusted R-squared:  0.8722
## F-statistic: 175.1 on 4 and 98 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 = model2)
##
##                      Value   p-value                   Decision
## Global Stat        76.530 9.992e-16 Assumptions NOT satisfied!
## Skewness           29.832 4.712e-08 Assumptions NOT satisfied!
## Kurtosis           24.774 6.446e-07 Assumptions NOT satisfied!
## Link Function      19.911 8.113e-06 Assumptions NOT satisfied!
## Heteroscedasticity  2.013 1.560e-01    Assumptions acceptable.
```

```
vif(model2)
```

```
##          Water         Cement fine_aggregate          Slump
##       1.319040       1.054623       1.043861       1.318641
```

```
sqrt(vif(model2))>2
```

```
##          Water         Cement fine_aggregate          Slump
##          FALSE          FALSE          FALSE          FALSE
```

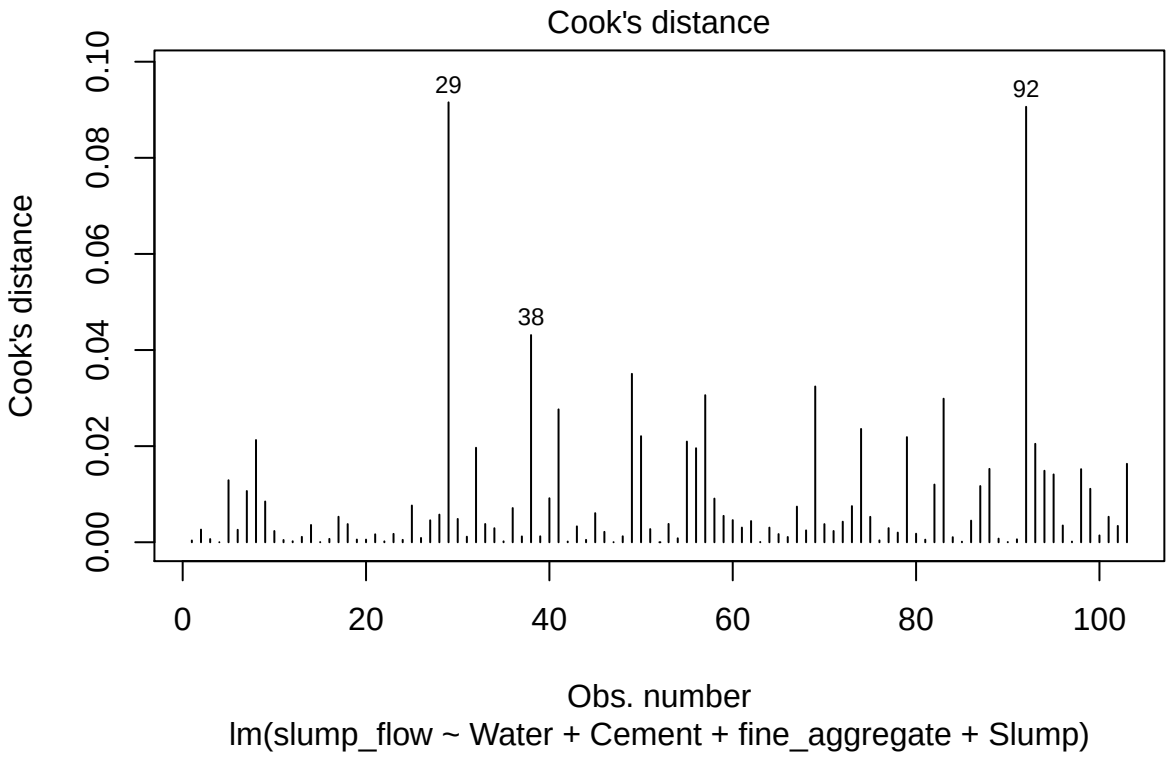
Thus, model 2 doesn't indicate the multicollinearity problem since the sqrt(vif(model2)) > 2 for none of the 4 predictors.

```
outlierTest(model2)
```

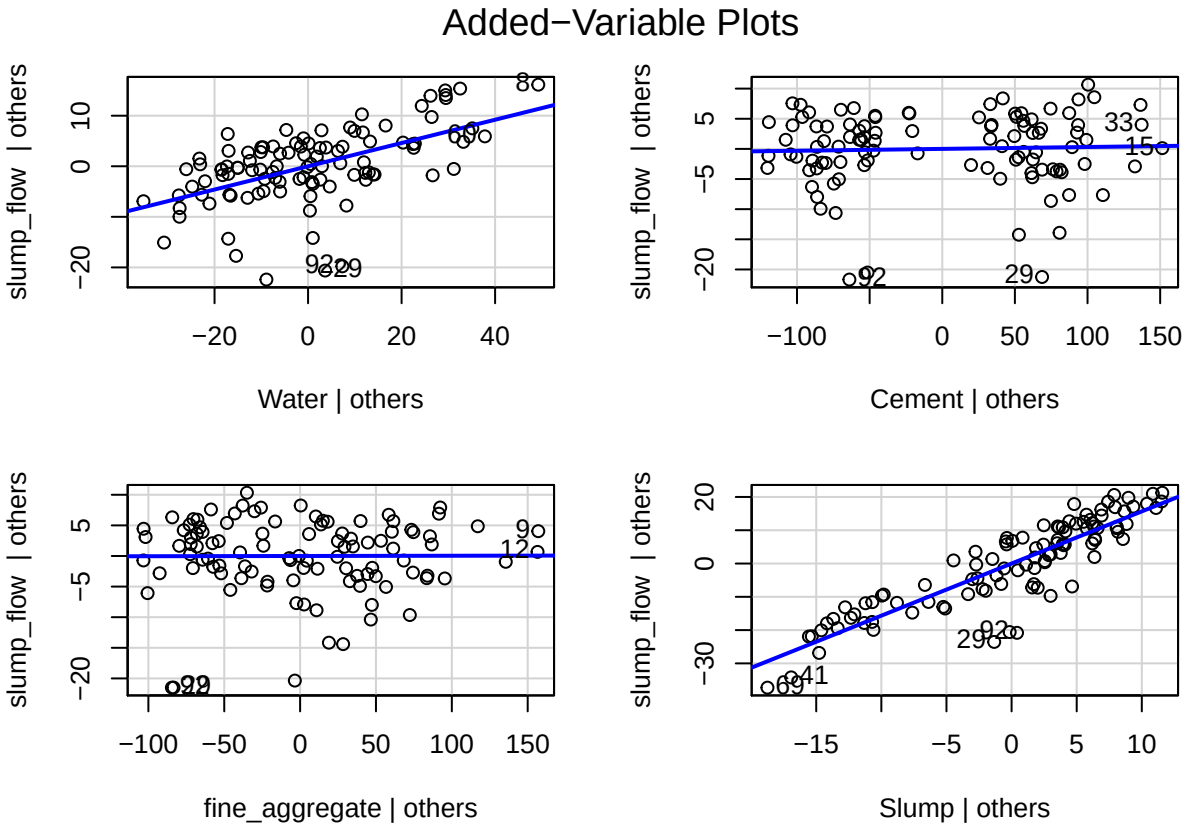
```
##     rstudent unadjusted p-value Bonferonni p
## 92 -3.702039         0.00035520     0.036586
## 29 -3.700182         0.00035749     0.036821
```

As expected, 92 and 29 are observed as outliers with indicated significance.

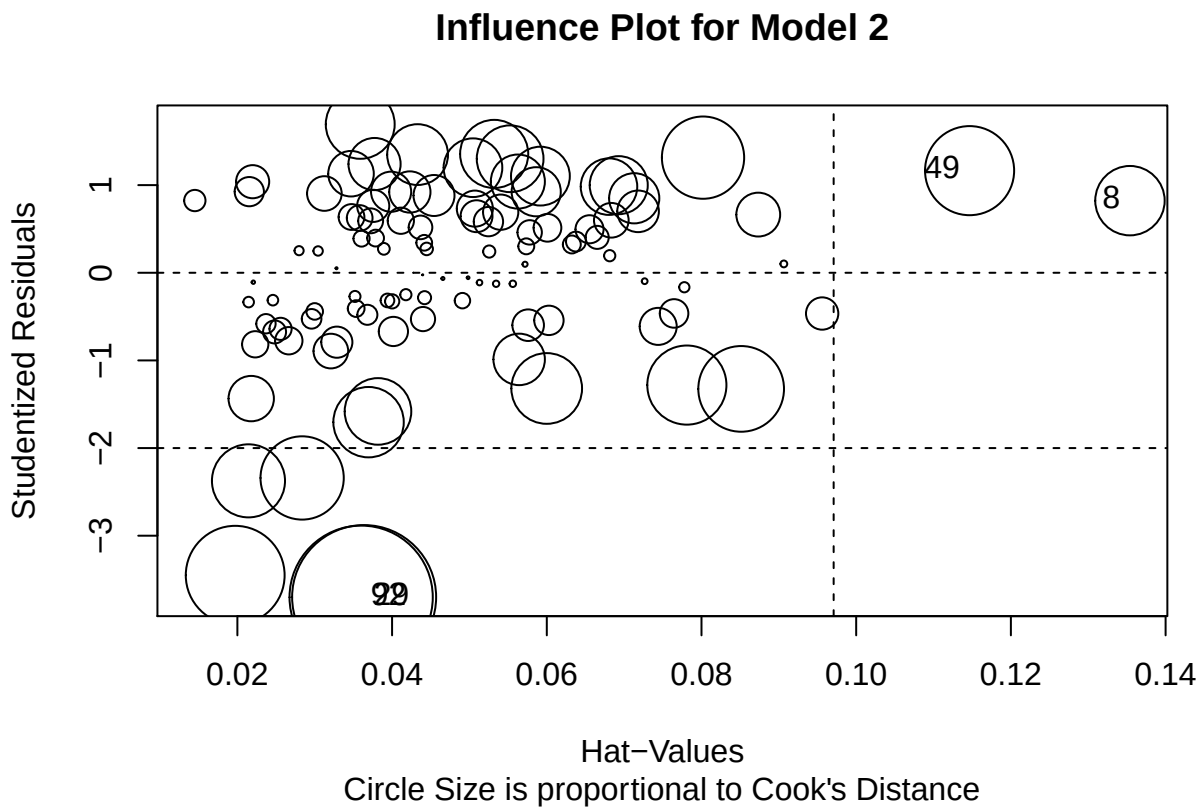
```
hat.plot2<-function(model2){
  p<- length(coefficients(model2))
  n<- length(fitted(model2))
  plot(hatvalues(model2), main ="Index values of Hat values")
  abline(h=c(2,3)*p/n,col="red",lty=2)
  identify(1:n,hatvalues(model1),names(hatvalues(model2)))
}

plot(model2, which = 4)
```




```
avPlots(model2, ask = FALSE, onepage = TRUE, id.method = "identify")
```




```
influencePlot(model2, id.method = "identify", main = "Influence Plot for Model 2", sub = "Circle Size is
```

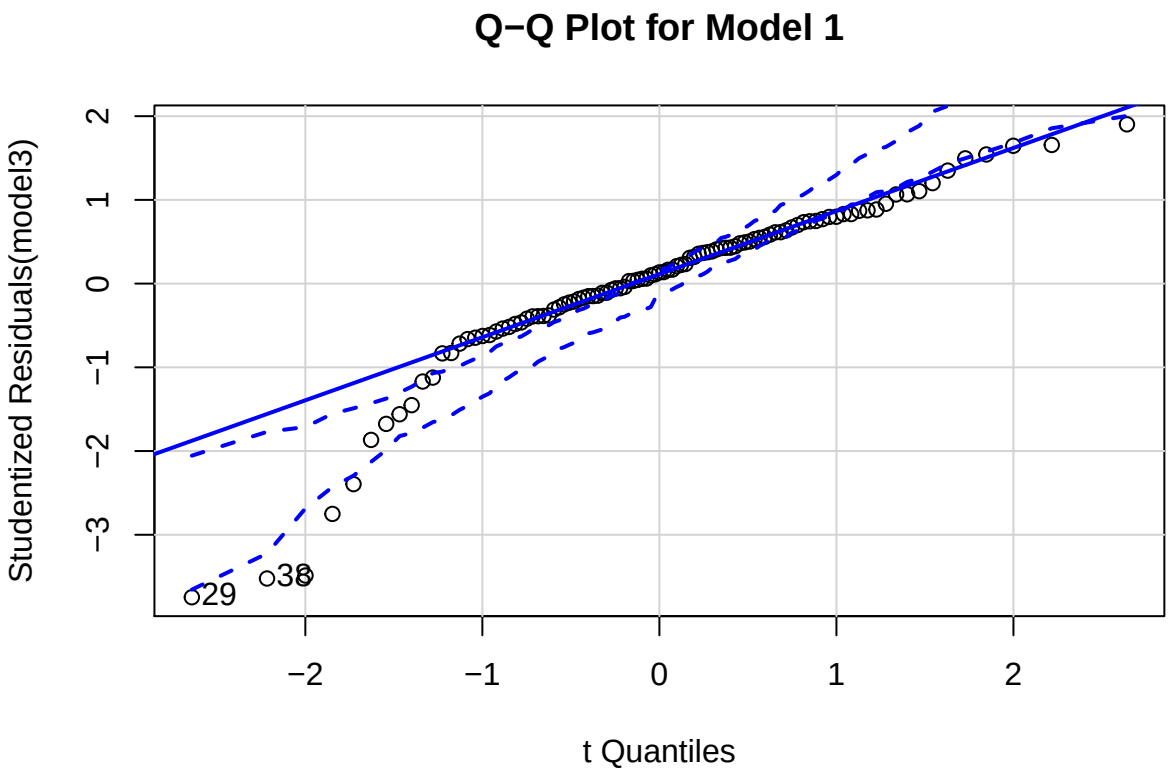



```
##       StudRes        Hat      CookD
## 8   0.8227131 0.13539340 0.02126866
## 29 -3.7001819 0.03638755 0.09154592
## 49  1.1652545 0.11466086 0.03504243
## 92 -3.7020391 0.03600049 0.09061542
```

Model 3

Normality

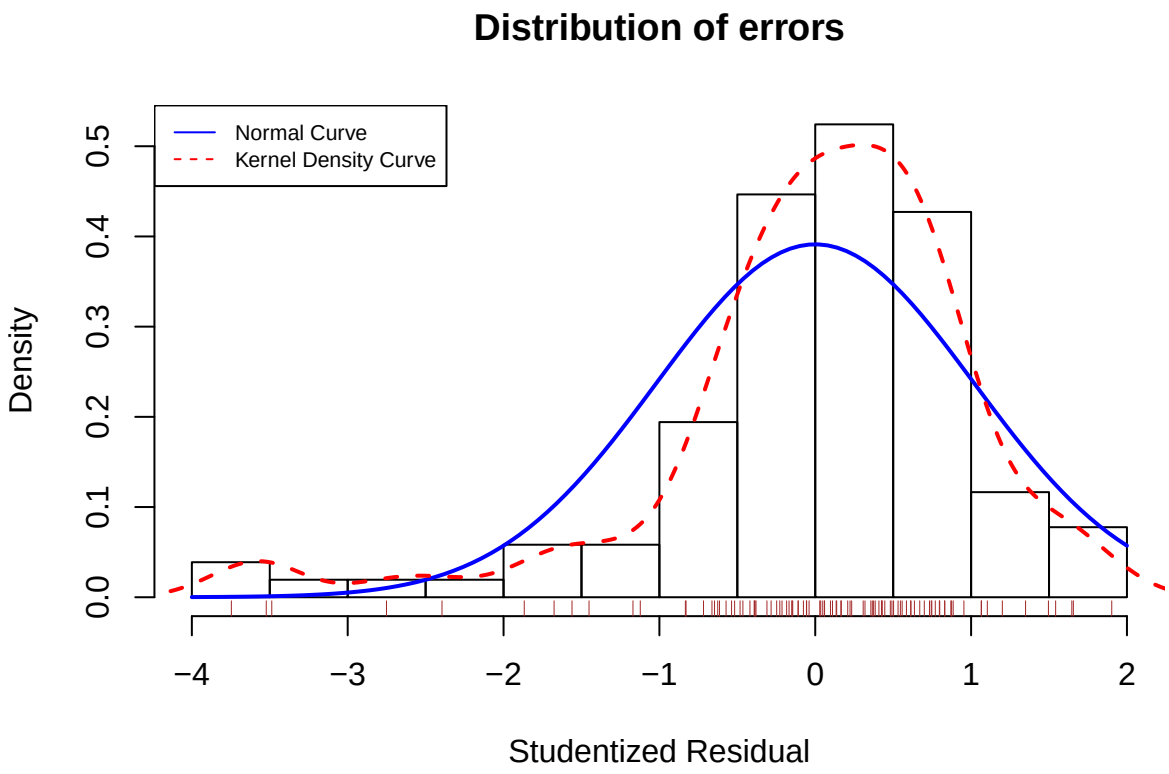
```
qqPlot(model3, labels = row.names(conc_slump), id.method = "identify",  simulate = TRUE, main = "Q-Q Pl
```




```
## [1] 29 38
```

In this model, obs 29 and 38 appear to be far away from the 45 degree line, but others are on it, threby implying normality assumption is right.

```
residplot <- function(model3, nbreaks = 10)
{
  z <- rstudent(model3)
  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("topleft", legend = c("Normal Curve", "Kernel Density Curve"), lty = 1:2, col = c("blue", "red
}
residplot(model3)
```



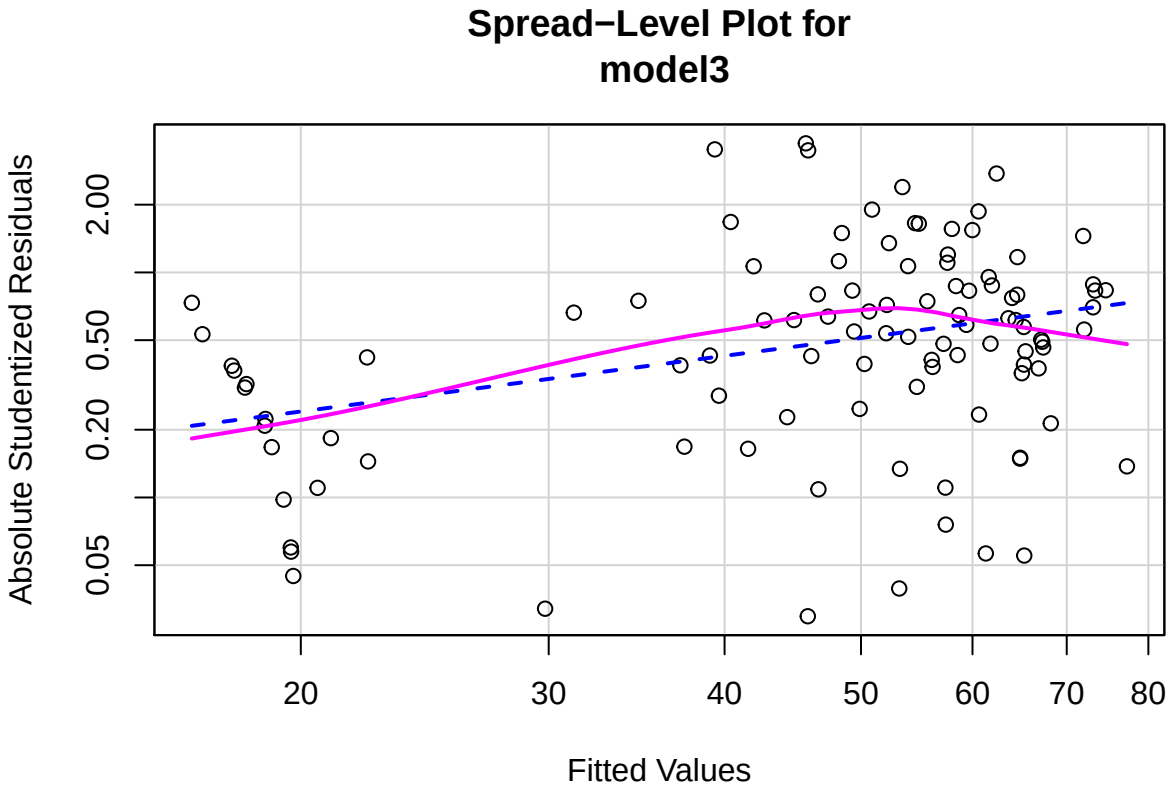
The residplot() function generates a histogram of the studentized residuals and superimposes a normal curve, kernel density curve and rug plot. It is clear that the errors follow the normal distribution quite well, with no outliers.

The component plus residual plot is not available for models with interactions.

```
ncvTest(model3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.6664697, Df = 1, p = 0.41429
```

```
spreadLevelPlot(model3)
```




```
##
## Suggested power transformation:  0.1782448
```

```
gvlmodel3<-gvlma(model3)
summary(gvlmodel3)
```

```
##
## Call:
## lm(formula = slump_flow ~ Water + Cement + fine_aggregate + Slump +
##     Cement:Water + Cement:fine_aggregate + Cement:Slump + Water:fine_aggregate +
##     Water:Slump, data = con_test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.6867  -2.2413   0.7109   3.3030  10.5900
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              9.325e+01  7.724e+01   1.207  0.23036
## Water                   -4.632e-01  3.667e-01  -1.263  0.20960
## Cement                   6.596e-02  1.434e-01   0.460  0.64669
## fine_aggregate          -8.750e-02  1.046e-01  -0.837  0.40489
## Slump                   -1.417e+00  7.842e-01  -1.806  0.07411 .
## Water:Cement            -5.747e-05  4.181e-04  -0.137  0.89098
## Cement:fine_aggregate -8.255e-05  1.454e-04  -0.568  0.57157
```

```
## Cement:Slump           6.816e-04  1.030e-03   0.662  0.50961
## Water:fine_aggregate   5.534e-04  4.836e-04   1.145  0.25535
## Water:Slump            1.533e-02  4.166e-03   3.681  0.00039 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.894 on 93 degrees of freedom
## Multiple R-squared:  0.8974, Adjusted R-squared:  0.8874
## F-statistic: 90.35 on 9 and 93 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 = model3)
##
##                      Value   p-value                   Decision
## Global Stat        71.5405 1.077e-14 Assumptions NOT satisfied!
## Skewness           33.2027 8.304e-09 Assumptions NOT satisfied!
## Kurtosis           35.7668 2.224e-09 Assumptions NOT satisfied!
## Link Function       2.1073 1.466e-01    Assumptions acceptable.
## Heteroscedasticity  0.4637 4.959e-01    Assumptions acceptable.
```

```
vif(model3)
```

```
##                 Water                Cement        fine_aggregate
##             161.17335             375.76886             128.82349
##                 Slump          Water:Cement Cement:fine_aggregate
##             138.27054             154.34081             231.49831
##          Cement:Slump  Water:fine_aggregate           Water:Slump
##              22.05104             287.51030             172.35272
```

```
sqrt(vif(model3))>2
```

```
##                 Water                Cement        fine_aggregate
##                  TRUE                  TRUE                  TRUE
##                 Slump          Water:Cement Cement:fine_aggregate
##                  TRUE                  TRUE                  TRUE
##          Cement:Slump  Water:fine_aggregate           Water:Slump
##                  TRUE                  TRUE                  TRUE
```

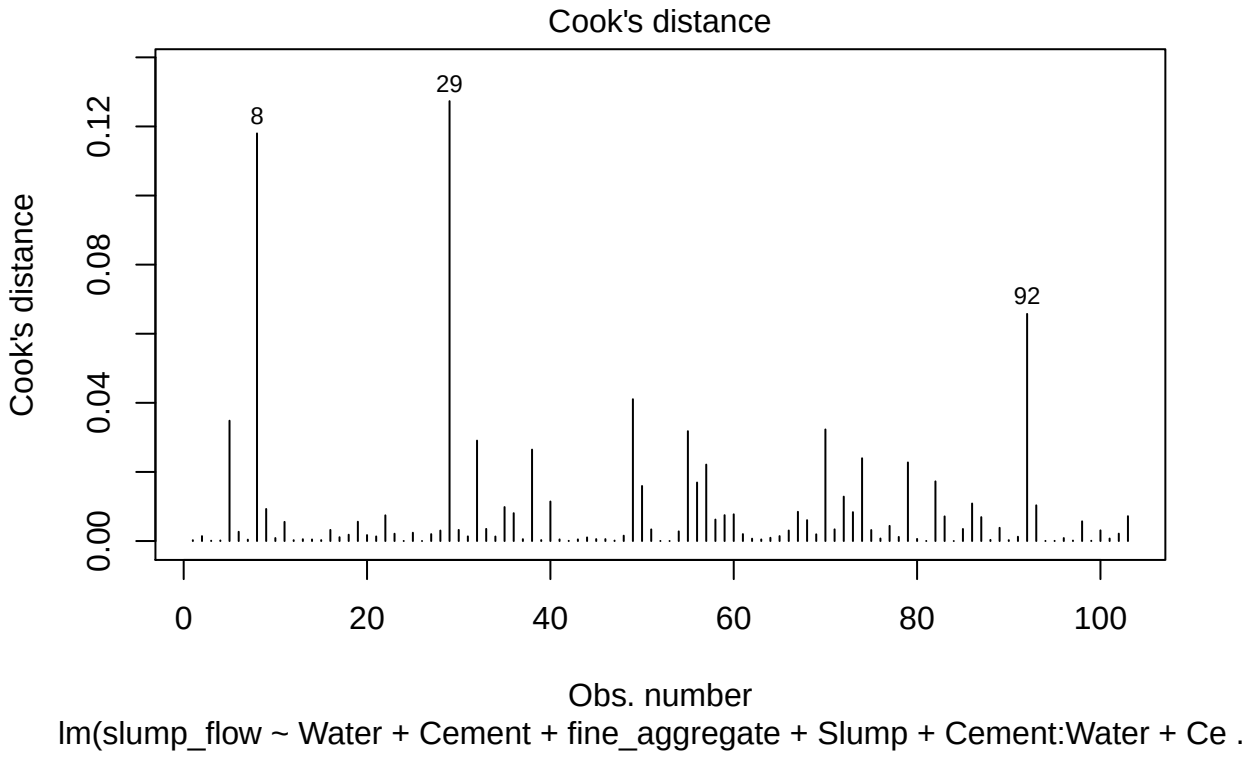
Thus, model 3 indicates the multicollinearity problem since the sqrt(vif(fit3)) > 2 all predictors.

```
outlierTest(model3)
```

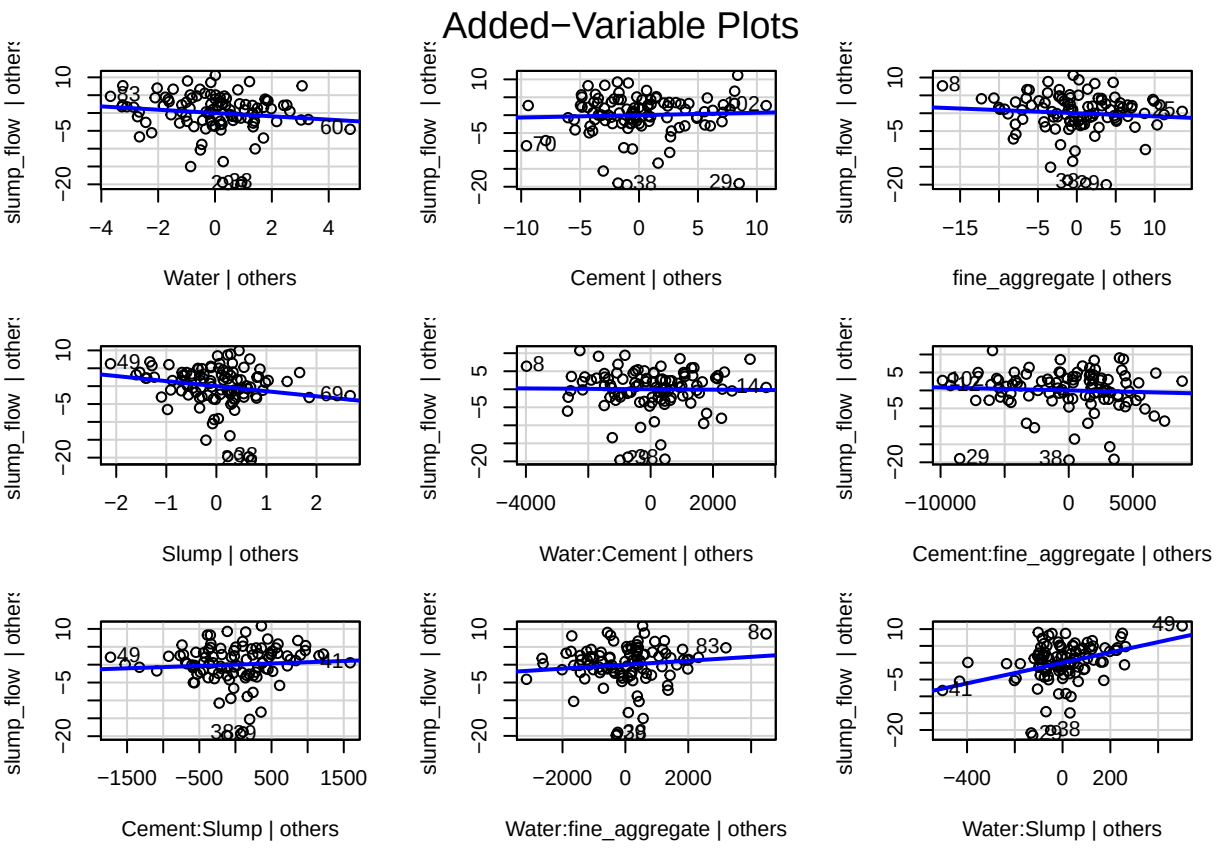
```
##     rstudent unadjusted p-value Bonferonni p
## 29 -3.746221          0.0003129     0.032229
```

Here, 29 are observed as outliers with indicated significance.

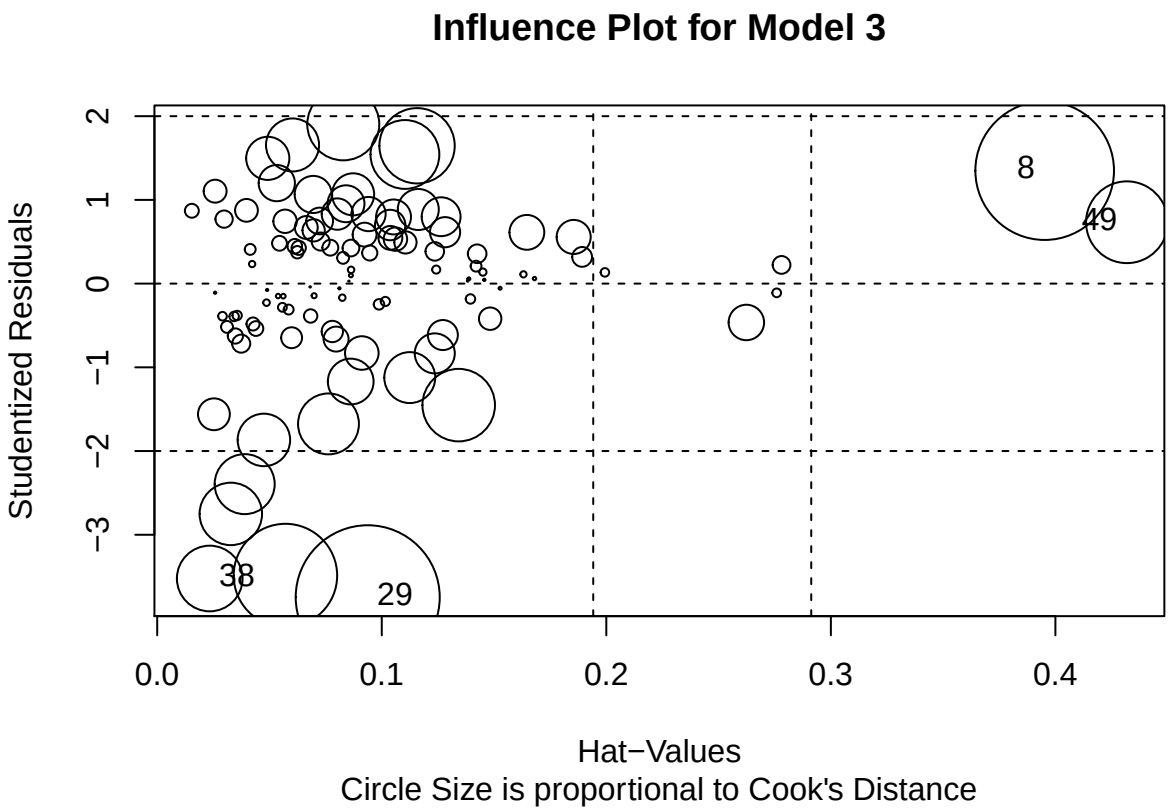
```r
hat.plot3<-function(model3){
  p<- length(coefficients(model3))
  n<- length(fitted(model3))
  plot(hatvalues(model3), main ="Index values of Hat values")
  abline(h=c(2,3)*p/n,col="red",lty=2)
  identify(1:n,hatvalues(model3),names(hatvalues(model3)))
}
plot(model3, which = 4)
```




```r
avPlots(model3, ask = FALSE, onepage = TRUE, id.method = "identify")
```




```
influencePlot(model3, id.method = "identify", main = "Influence Plot for Model 3", sub = "Circle Size is
```




```
##       StudRes        Hat      CookD
## 8   1.3495815 0.39522301 0.11798488
## 29 -3.7462208 0.09374670 0.12732986
## 38 -3.5224457 0.02338156 0.02645989
## 49  0.7329929 0.43186403 0.04104503
```

Corrective Measures

Removing outliers and Influential observations

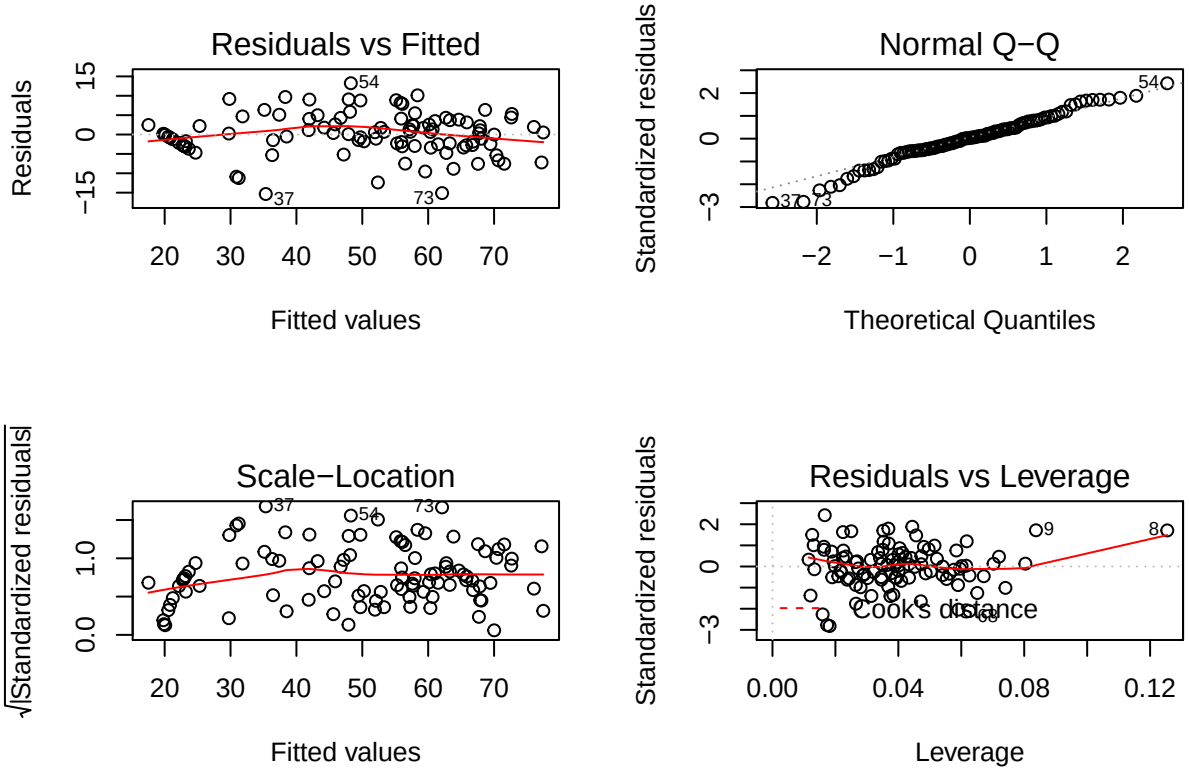
```
con_test<-con_test[-c(29,92),]

model2<- lm(slump_flow~Water+Cement+fine_aggregate+Slump
            ,data=con_test)


model3<- lm(slump_flow~Water+Cement+fine_aggregate+Slump+Cement:Water+Cement:fine_aggregate+Cement:Slump
```

Introducing a ploynomial regression model

```
model4 <- lm(slump_flow~Water+fine_aggregate+I(Slump^2),data=con_test)
summary(model4)
```

```
##
## Call:
## lm(formula = slump_flow ~ Water + fine_aggregate + I(Slump^2),
##     data = con_test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3477  -2.7772   0.1942   3.4659  13.2384
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -20.063737   8.328495  -2.409   0.0179 *
## Water            0.257095   0.030171   8.521 2.07e-13 ***
## fine_aggregate  -0.004378   0.008910  -0.491   0.6243
## I(Slump^2)       0.056181   0.002566  21.892  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.498 on 97 degrees of freedom
## Multiple R-squared:  0.9035, Adjusted R-squared:  0.9005
## F-statistic: 302.8 on 3 and 97 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model4)
```



```r
summary(model1)
```

```
##
## Call:
## lm(formula = slump_flow ~ ., data = con_test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.5414  -2.5321   0.5258   3.2530  10.0769
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.765e+02  2.050e+02  -0.861   0.3915
## No                           -2.797e-04  3.019e-02  -0.009   0.9926
## Cement                        1.016e-02  7.011e-02   0.145   0.8851
## Slag                          2.887e-02  9.370e-02   0.308   0.7587
## `Fly Ash`                     2.902e-02  7.155e-02   0.406   0.6860
## Water                         4.265e-01  2.062e-01   2.068   0.0414 *
## SP                            5.447e-01  3.066e-01   1.777   0.0789 .
## coarse_aggregate              5.199e-02  8.030e-02   0.647   0.5190
## fine_aggregate                5.084e-02  8.200e-02   0.620   0.5368
## Slump                         1.588e+00  8.337e-02  19.052   <2e-16 ***
## `28-day Compressive Strength` 4.506e-01  2.370e-01   1.901   0.0604 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.716 on 92 degrees of freedom
## Multiple R-squared:  0.9045, Adjusted R-squared:  0.8941
## F-statistic: 87.15 on 10 and 92 DF,  p-value: < 2.2e-16
```

```r
summary(model2)
```

```
##
## Call:
## lm(formula = slump_flow ~ Water + Cement + fine_aggregate + Slump,
##     data = con_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.781  -3.064   0.759   4.137   9.683
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -18.924068   8.330595  -2.272   0.0253 *
## Water            0.238053   0.030892   7.706 1.18e-11 ***
## Cement           0.003298   0.007121   0.463   0.6443
## fine_aggregate  -0.009368   0.008923  -1.050   0.2964
## Slump            1.563055   0.071260  21.935  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.483 on 96 degrees of freedom
```

```
## Multiple R-squared:  0.905,  Adjusted R-squared:  0.9011
## F-statistic: 228.7 on 4 and 96 DF,  p-value: < 2.2e-16
```

```
summary(model3)
```

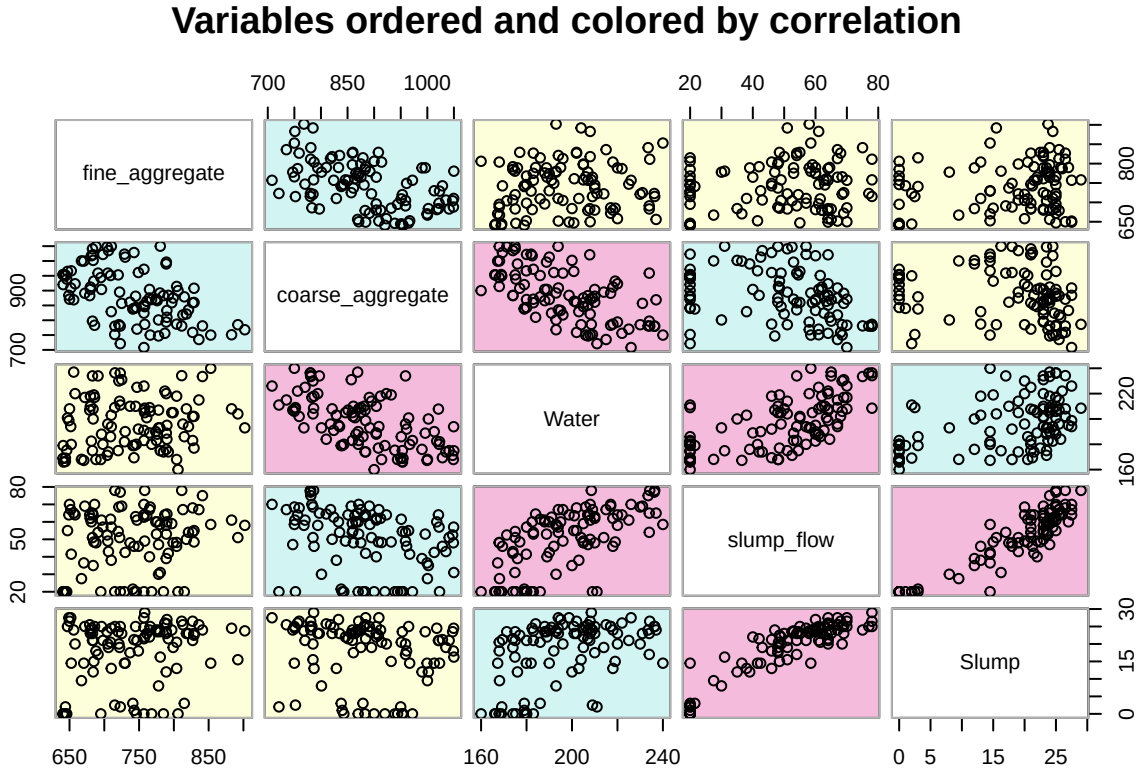
```
##
## Call:
## lm(formula = slump_flow ~ Water + Cement + fine_aggregate + Slump +
##     Cement:Water + Cement:fine_aggregate + Cement:Slump + Water:fine_aggregate +
##     Water:Slump, data = con_test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8225  -2.3811  -0.0069   3.4731   9.0819
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)           61.6335953 67.8629601   0.908 0.366168
## Water                 -0.3111410  0.3215323  -0.968 0.335768
## Cement                 0.1554527  0.1283487   1.211 0.228965
## fine_aggregate        -0.0688139  0.0916203  -0.751 0.454544
## Slump                 -0.9130385  0.6912064  -1.321 0.189835
## Water:Cement          -0.0002408  0.0003669  -0.656 0.513229
## Cement:fine_aggregate -0.0001529  0.0001305  -1.171 0.244666
## Cement:Slump           0.0006663  0.0009000   0.740 0.460976
## Water:fine_aggregate   0.0004770  0.0004227   1.128 0.262123
## Water:Slump            0.0126215  0.0036723   3.437 0.000889 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.15 on 91 degrees of freedom
## Multiple R-squared:  0.9206, Adjusted R-squared:  0.9127
## F-statistic: 117.2 on 9 and 91 DF,  p-value: < 2.2e-16
```

```
summary(model4)
```

```
##
## Call:
## lm(formula = slump_flow ~ Water + fine_aggregate + I(Slump^2),
##     data = con_test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3477  -2.7772   0.1942   3.4659  13.2384
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -20.063737   8.328495  -2.409   0.0179 *
## Water            0.257095   0.030171   8.521 2.07e-13 ***
## fine_aggregate  -0.004378   0.008910  -0.491   0.6243
## I(Slump^2)       0.056181   0.002566  21.892  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 5.498 on 97 degrees of freedom
## Multiple R-squared:  0.9035, Adjusted R-squared:  0.9005
## F-statistic: 302.8 on 3 and 97 DF,  p-value: < 2.2e-16
```

```
library(gclus)
data1<- con_test[c(5,7,8,9,10)]
data1.corr<-abs(cor(data1))
mycol <- dmat.color(data1.corr)
myord <- order.single(data1.corr)
cpairs(data1, myord, panel.colors = mycol, gap= 0.5, main= "Variables ordered and colored by correlatio
```



From the graph, we see relations between several pairs, so the model may be suffering from multicollinearity problem.

Although the Adjusted R squared value decreases, we have tried to reduce multicollinearity in order to avoid overfitting.

Model1 has the best R sq but that may be because of overfitting. Next best is Model3 .

Therefore we conclude to include fine aggregate, slump, water as our final predictors.

```
AIC(model1,model2,model3,model4)
```

```
##        df      AIC
## model1 12 663.7916
## model2  6 637.2493
```

```
## model3 11 629.1786
## model4  5 636.8345
```

```
library(MASS)
stepAIC(model3, direction = "backward")
```

```
## Start:  AIC=340.55
## slump_flow ~ Water + Cement + fine_aggregate + Slump + Cement:Water +
##     Cement:fine_aggregate + Cement:Slump + Water:fine_aggregate +
##     Water:Slump
##
##                          Df Sum of Sq    RSS    AIC
## - Water:Cement            1    11.428 2425.1 339.03
## - Cement:Slump            1    14.539 2428.2 339.16
## - Water:fine_aggregate    1    33.772 2447.5 339.96
## - Cement:fine_aggregate   1    36.369 2450.1 340.06
## <none>                                2413.7 340.55
## - Water:Slump             1   313.322 2727.0 350.88
##
## Step:  AIC=339.03
## slump_flow ~ Water + Cement + fine_aggregate + Slump + Cement:fine_aggregate +
##     Cement:Slump + Water:fine_aggregate + Water:Slump
##
##                          Df Sum of Sq    RSS    AIC
## - Cement:Slump            1      7.93 2433.1 337.36
## - Cement:fine_aggregate   1     30.97 2456.1 338.31
## - Water:fine_aggregate    1     32.12 2457.2 338.36
## <none>                                2425.1 339.03
## - Water:Slump             1    334.43 2759.6 350.08
##
## Step:  AIC=337.36
## slump_flow ~ Water + Cement + fine_aggregate + Slump + Cement:fine_aggregate +
##     Water:fine_aggregate + Water:Slump
##
##                          Df Sum of Sq    RSS    AIC
## - Cement:fine_aggregate   1     30.91 2464.0 336.64
## - Water:fine_aggregate    1     35.22 2468.3 336.81
## <none>                                2433.1 337.36
## - Water:Slump             1    346.17 2779.2 348.80
##
## Step:  AIC=336.64
## slump_flow ~ Water + Cement + fine_aggregate + Slump + Water:fine_aggregate +
##     Water:Slump
##
##                          Df Sum of Sq    RSS    AIC
## - Cement                  1     17.70 2481.7 335.36
## - Water:fine_aggregate    1     45.49 2509.4 336.48
## <none>                                2464.0 336.64
## - Water:Slump             1    339.49 2803.4 347.67
##
## Step:  AIC=335.36
## slump_flow ~ Water + fine_aggregate + Slump + Water:fine_aggregate +
##     Water:Slump
##
```

```
##                         Df Sum of Sq    RSS    AIC
## - Water:fine_aggregate  1     36.16 2517.8 334.82
## <none>                              2481.7 335.36
## - Water:Slump           1    340.70 2822.4 346.35
##
## Step:  AIC=334.82
## slump_flow ~ Water + fine_aggregate + Slump + Water:Slump
##
##                   Df Sum of Sq    RSS    AIC
## - fine_aggregate  1      7.37 2525.2 333.11
## <none>                        2517.8 334.82
## - Water:Slump     1    375.21 2893.0 346.85
##
## Step:  AIC=333.11
## slump_flow ~ Water + Slump + Water:Slump
##
##               Df Sum of Sq    RSS    AIC
## <none>                     2525.2 333.11
## - Water:Slump  1    400.29 2925.5 345.98

##
## Call:
## lm(formula = slump_flow ~ Water + Slump + Water:Slump, data = con_test)
##
## Coefficients:
## (Intercept)        Water        Slump  Water:Slump
##    22.64961     -0.02409     -1.00704      0.01381
```

We start with all the predictors in the model. For each step, the AIC column provides the model AIC resulting from the deletion of the variable listed in that row. When deletion of any more variables increases the AIC, the process stops.

Here, AIC in the last step is 326.88 and the selected list of predictors is as indicated in the last step.

Models with smaller AIC() values indicate adequate fit with fewer parameters and hence are preferred. Here, AIC values for model 3 is the lesser than 1, and hence are preferred. Our conclusion would be based on all the analysis done until this pint to conclude Model3 is the best with the least AIC and most R sq.

Problem2

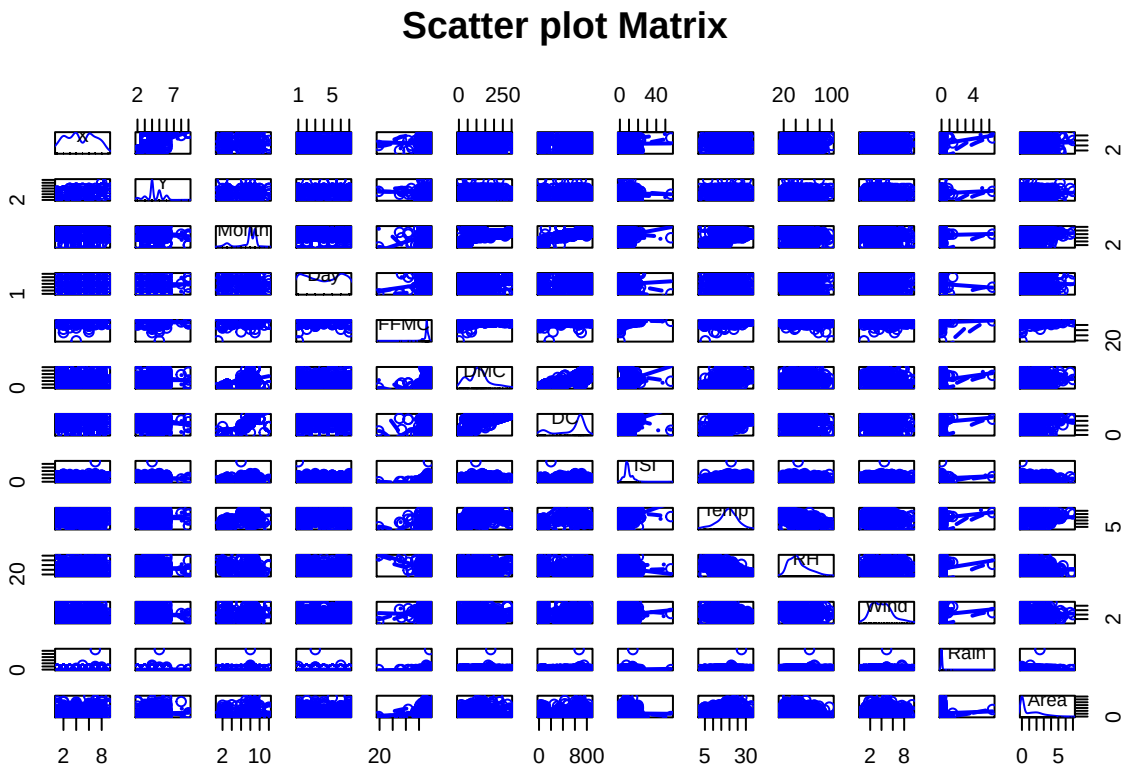
Task 3: Forest Fire Data

Predictor Variables . X Y spatial coordinate . Month . Day . Indices from FWI system like FFMC, DMC, DC, ISI . Temperature . Relative humidity . Wind . Rain

Response Variables . Area

```
library(readxl)
ForestData <- read_excel("C:/Users/kkavi/Desktop/anushka/Forest_fires_data.xlsx")

x<-factor(ForestData$Day)
x <- factor(x, levels = c("sun", "mon", "tue", "wed", "thu", "fri", "sat"), ordered = TRUE)
ForestData$Day <- as.integer(x, ForestData$Day)
month_num <- factor(ForestData$Month)
month_num <- factor(month_num, levels = c("jan", "feb", "mar", "apr", "may", "jun", "jul","aug","sep","
```

```
ForestData$Month <- as.integer(month_num,ForestData$Month)
ForestData$Area <- log(1+ForestData$Area)
library(car)
scatterplotMatrix(ForestData, spread =FALSE,smooth = 2, main= "Scatter plot Matrix")
```



Almost all the graphs of predictors are skewed except for Temperature and wind, which resembles a normal distribution Some meaningful relationships observed,

. DMC increases linearly with rain, month.

. DC increases linearly with month, DMC, temp.

. Temperature increases linearly with rain.

. Water has a linear relationship with cement, fine aggregate, slump and slump flow.

. Area increases linearly with rain, X and Y coordinates.

The principal diagonal contains density and rug plots for each variable.

Modelling based on Domain Knowledge and Initial Analysis

```
model1<-lm(Area ~ X + Y + Month + Day + FFMC + DMC + DC + ISI, data=ForestData)
summary(model1)
```

```
##
## Call:
## lm(formula = Area ~ X + Y + Month + Day + FFMC + DMC + DC + ISI,
```

```
##     data = ForestData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1836 -1.0964 -0.6152  0.8838  5.6943
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.5422144  1.1763492  -0.461  0.64505
## X            0.0413122  0.0314864   1.312  0.19009
## Y           -0.0058830  0.0598975  -0.098  0.92180
## Month        0.1739028  0.0585276   2.971  0.00311 **
## Day          0.0006033  0.0286936   0.021  0.98323
## FFMC         0.0093595  0.0137524   0.681  0.49645
## DMC          0.0024683  0.0014616   1.689  0.09188 .
## DC          -0.0014070  0.0006505  -2.163  0.03100 *
## ISI         -0.0186148  0.0159767  -1.165  0.24452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.388 on 508 degrees of freedom
## Multiple R-squared:  0.02976,    Adjusted R-squared:  0.01448
## F-statistic: 1.948 on 8 and 508 DF,  p-value: 0.05115
```

```
model2<- lm(Area ~ X + Y + Month + Day + Temp   +RH + Wind +    Rain, data = ForestData)
summary(model2)
```

```
##
## Call:
## lm(formula = Area ~ X + Y + Month + Day + Temp + RH + Wind +
##     Rain, data = ForestData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5192 -1.0932 -0.6405  0.8766  5.6376
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.728e-01  5.183e-01   0.526   0.5988
## X            3.784e-02  3.155e-02   1.199   0.2310
## Y            1.991e-02  5.923e-02   0.336   0.7369
## Month        7.410e-02  2.928e-02   2.530   0.0117 *
## Day         -3.544e-03  2.869e-02  -0.124   0.9017
## Temp        -3.262e-05  1.389e-02  -0.002   0.9981
## RH          -4.833e-03  4.546e-03  -1.063   0.2882
## Wind         6.194e-02  3.531e-02   1.754   0.0800 .
## Rain         8.349e-02  2.114e-01   0.395   0.6930
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.39 on 508 degrees of freedom
## Multiple R-squared:  0.02712,    Adjusted R-squared:  0.0118
## F-statistic:  1.77 on 8 and 508 DF,  p-value: 0.08051
```

```r
model3 <- lm(Area ~ FFMC + DMC + DC + ISI, data=ForestData)
summary(model3)
```

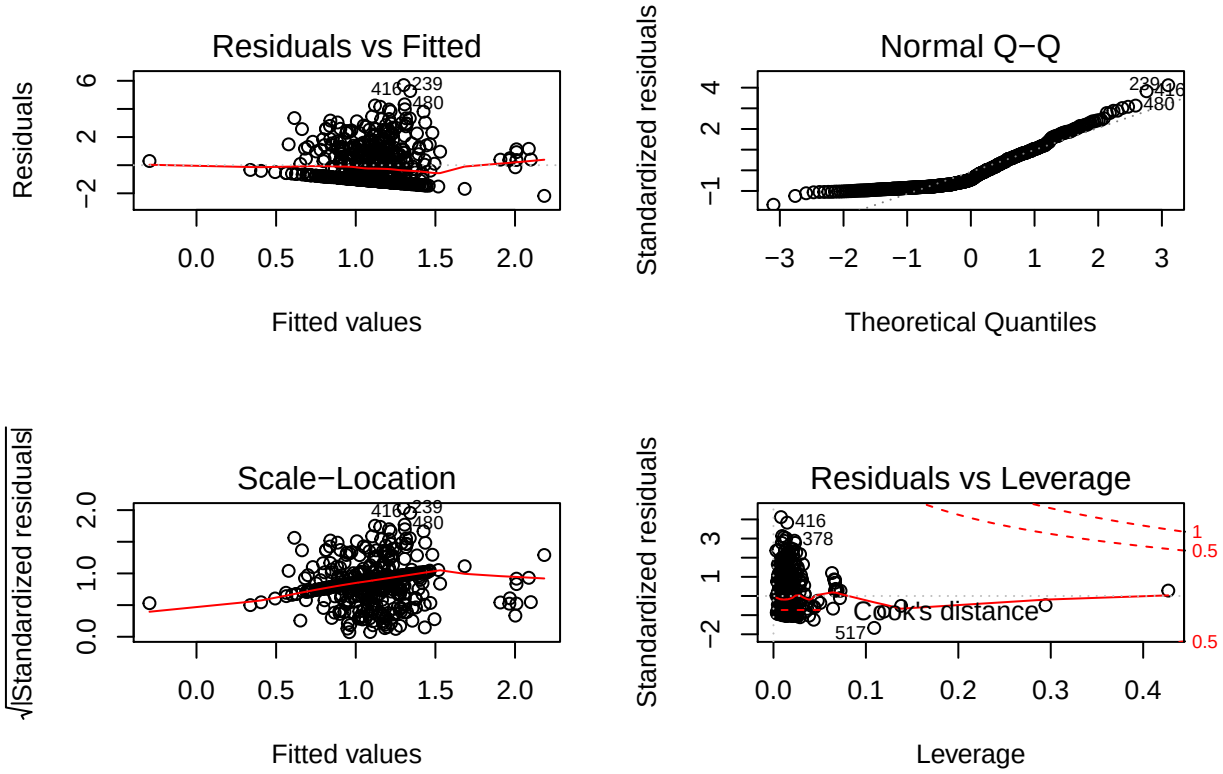
```
##
## Call:
## lm(formula = Area ~ FFMC + DMC + DC + ISI, data = ForestData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3703 -1.1242 -0.6145  0.8882  5.8198
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0851719  1.1513950  -0.074    0.941
## FFMC         0.0126619  0.0137577   0.920    0.358
## DMC          0.0009234  0.0013587   0.680    0.497
## DC           0.0001928  0.0003412   0.565    0.572
## ISI         -0.0176878  0.0160811  -1.100    0.272
##
## Residual standard error: 1.398 on 512 degrees of freedom
## Multiple R-squared:  0.008046,   Adjusted R-squared:  0.0002959
## F-statistic: 1.038 on 4 and 512 DF,  p-value: 0.3869
```

```r
model4 <- lm(Area ~Temp +RH + Wind +    Rain, data = ForestData)
summary(model4)
```

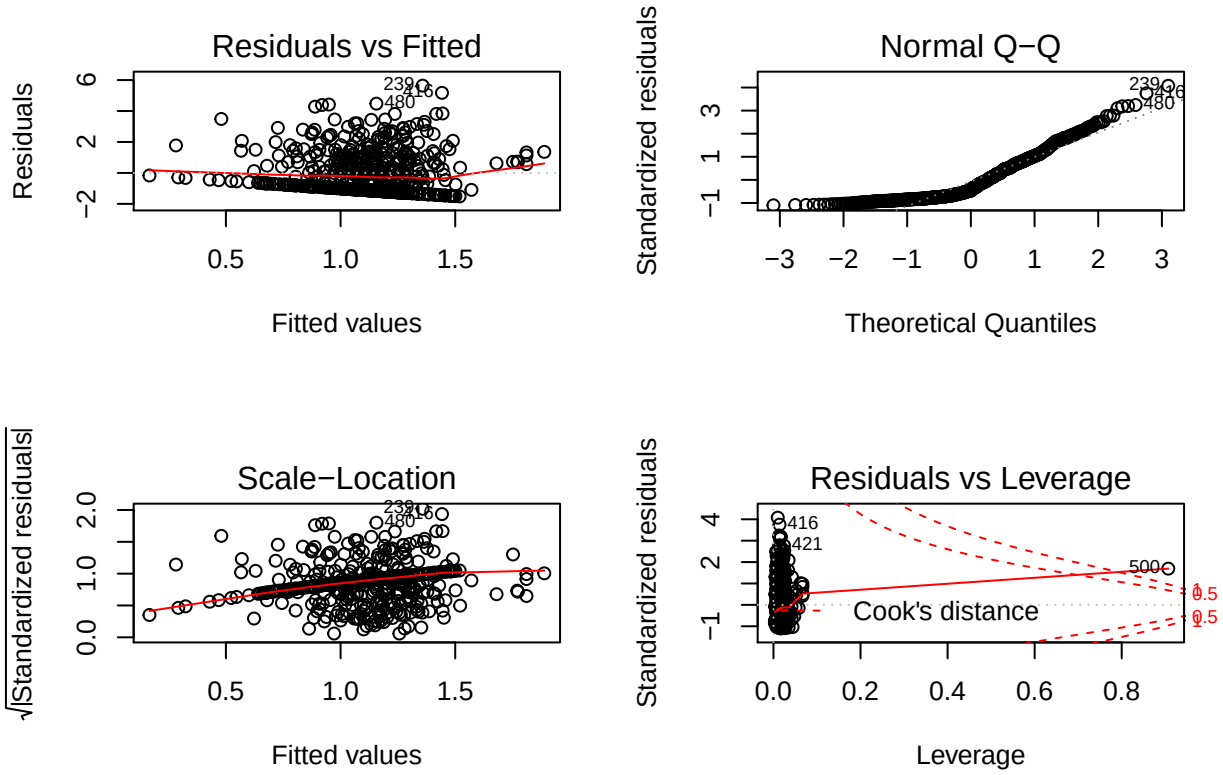
```
##
## Call:
## lm(formula = Area ~ Temp + RH + Wind + Rain, data = ForestData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3993 -1.0978 -0.7081  0.9121  5.7593
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.742041   0.443148   1.674   0.0946 .
## Temp         0.012766   0.012958   0.985   0.3250
## RH          -0.002834   0.004506  -0.629   0.5296
## Wind         0.062603   0.035446   1.766   0.0780 .
## Rain         0.085167   0.211852   0.402   0.6878
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.397 on 512 degrees of freedom
## Multiple R-squared:  0.0104, Adjusted R-squared:  0.002671
## F-statistic: 1.345 on 4 and 512 DF,  p-value: 0.2519
```

Typical Approach

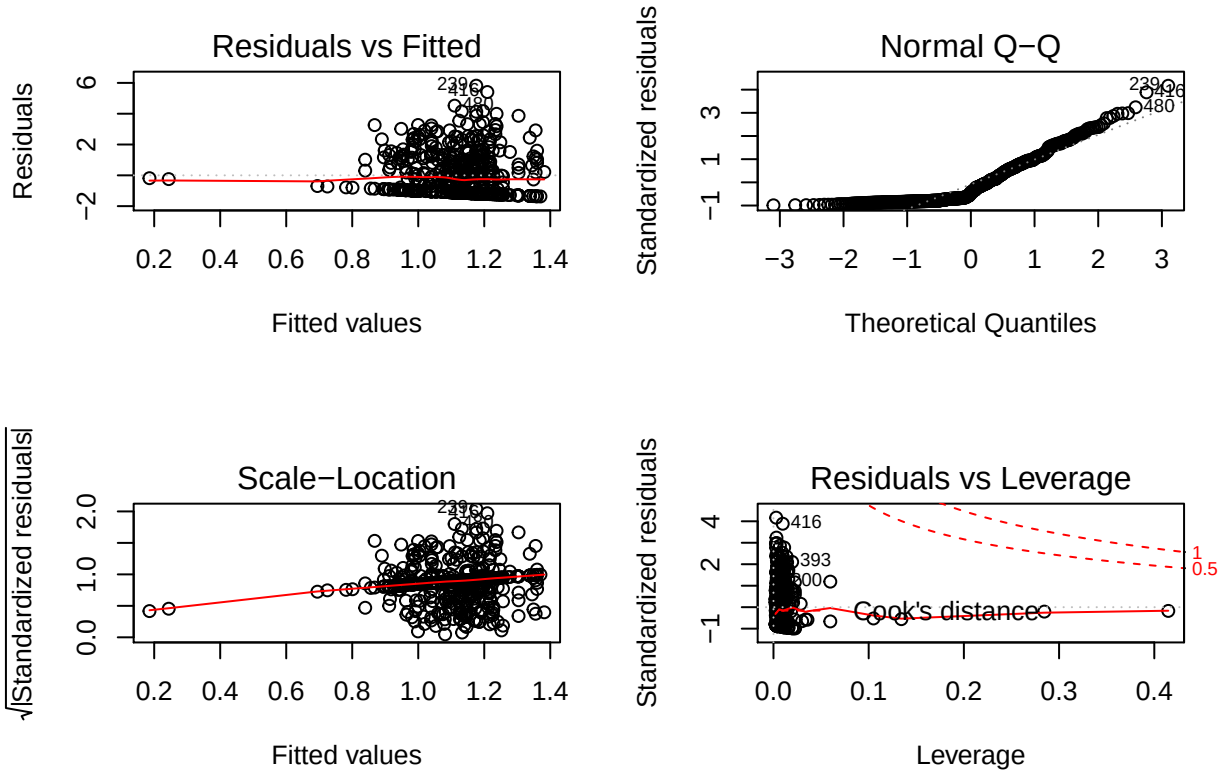
```
par(mfrow=c(2,2))
plot(model1)
```



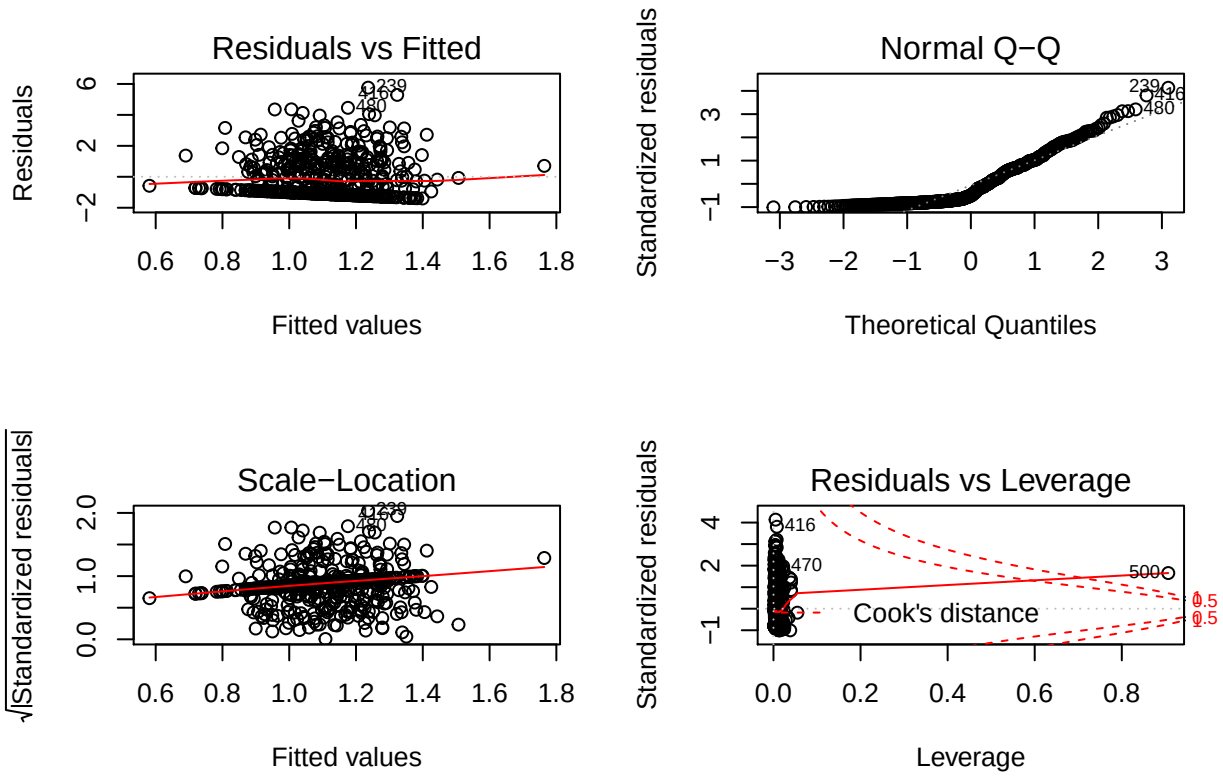

```
par(mfrow=c(2,2))
plot(model2)
```


Residuals vs Fitted
Residuals
Fitted values
239
416
480
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
416
421
500
0.5
1

```
plot(model3)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

```
plot(model4)
```



Normality: Most of the observations do not lie on the 45 degree line in Q-Q plot, thereby violating our assumption of normality in all models.

Independence: From the scatterplots, we have found random distribution hence intimating independence with respect to others. In all models, we do not acknowledge their interactions among themselves.

Linearity:There is random distribution between residuals and fitted value in all models.

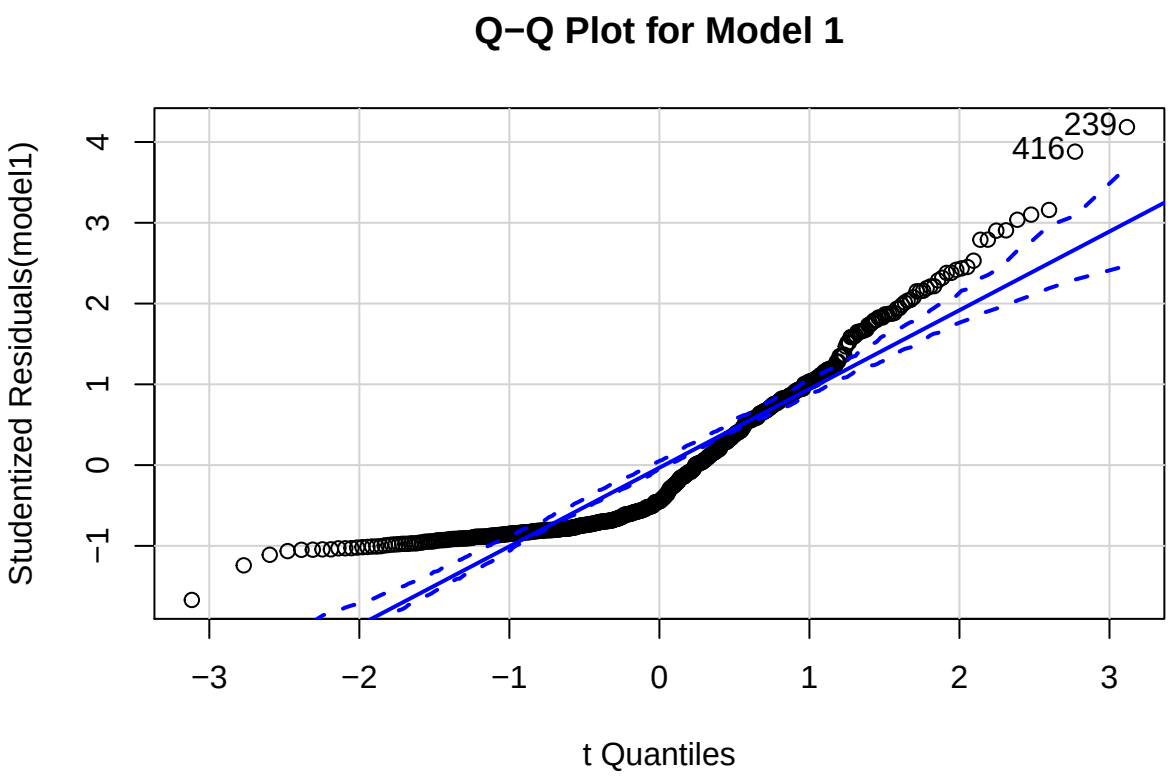
Homoscedasticity: The points in the Scale-Location graph seem to be a linear relationship around the horizontal line therefore violating constant variance assumption.

Outlier: No potential outliers observed. Except in model 1 where observations: 416,239, 480.

Influential Observations: Observations 416, 378 for model2; 416, 421 for model2; 416, 393, 300 for model3 appear to be influential having the lagest Cook's distance.

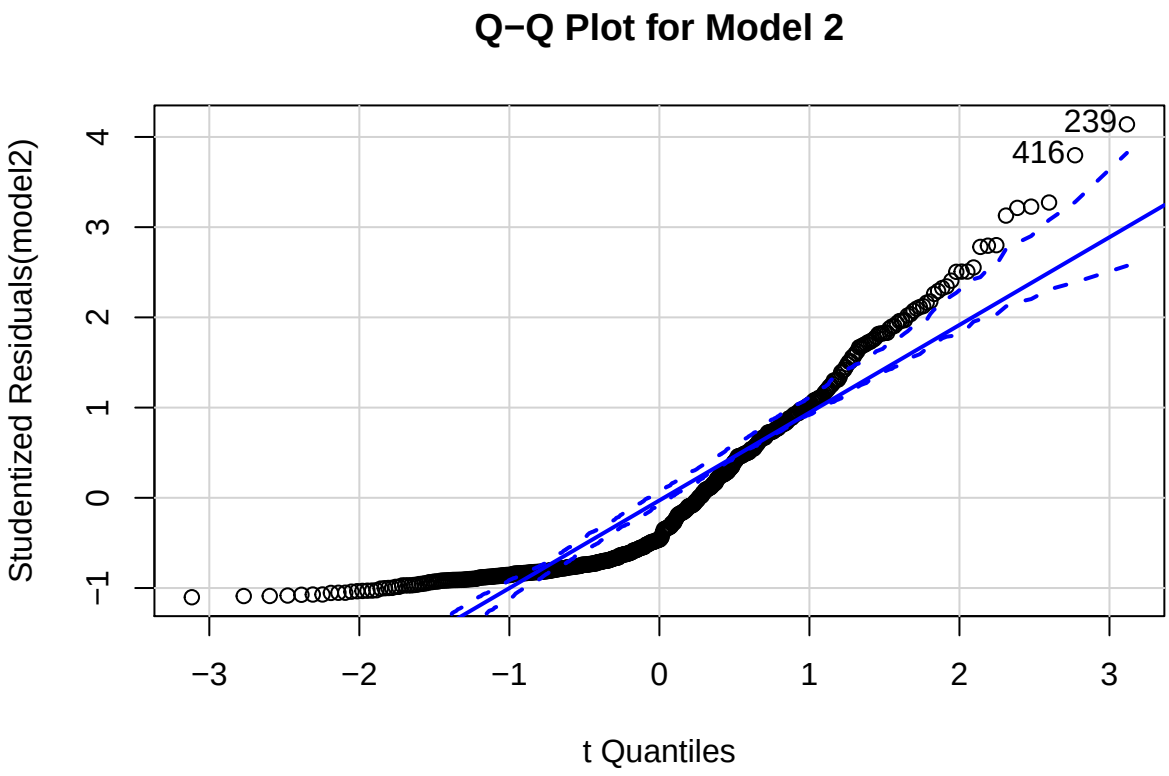
Enhanced Approach

```
qqPlot(model1, labels = row.names(Area), id.method = "identify",  simulate = TRUE, main = "Q-Q Plot for
```



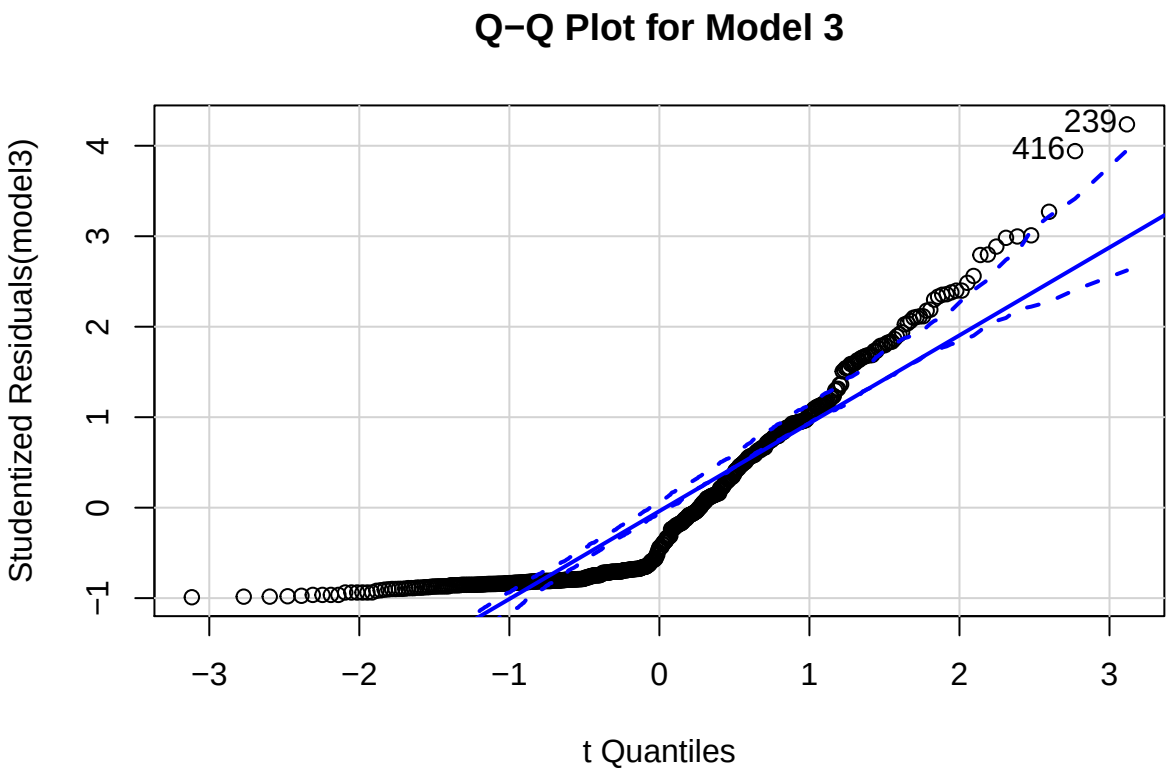

```
## [1] 239 416
```

```
qqPlot(model2, labels = row.names(Area), id.method = "identify",  simulate = TRUE, main = "Q-Q Plot for
```



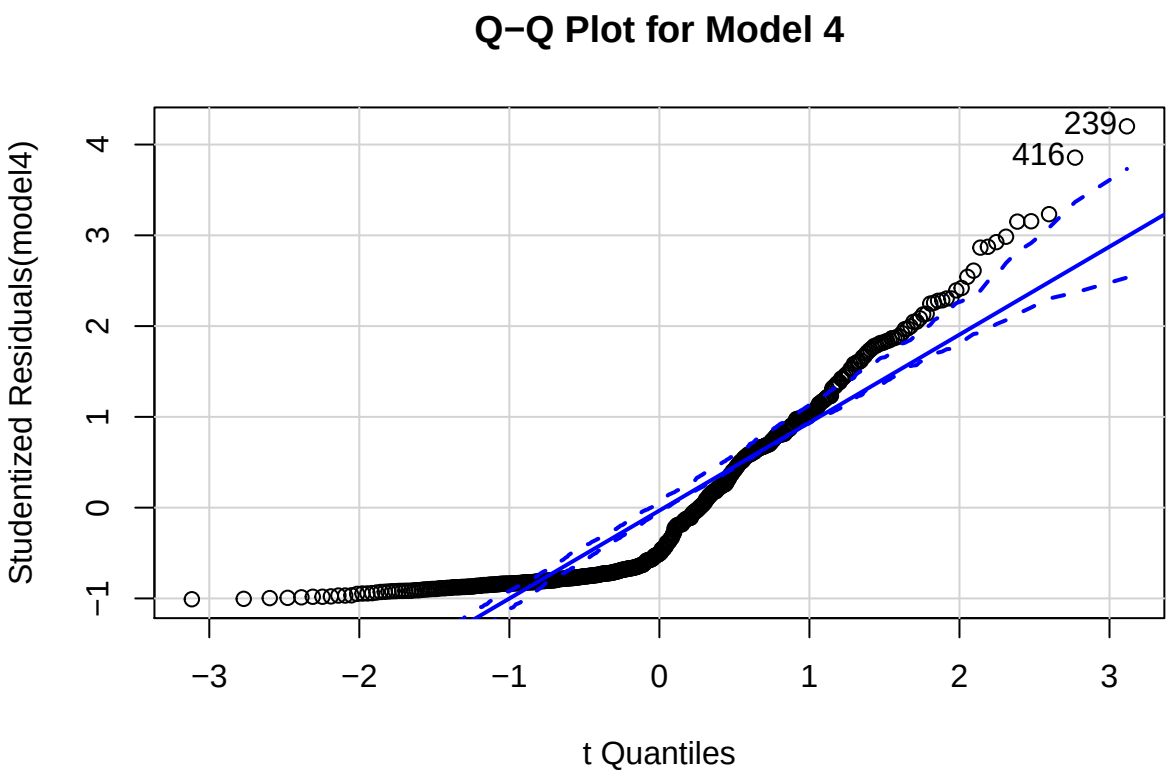

```
## [1] 239 416
```

```
qqPlot(model3, labels = row.names(Area), id.method = "identify",  simulate = TRUE, main = "Q-Q Plot for
```




```
## [1] 239 416
```

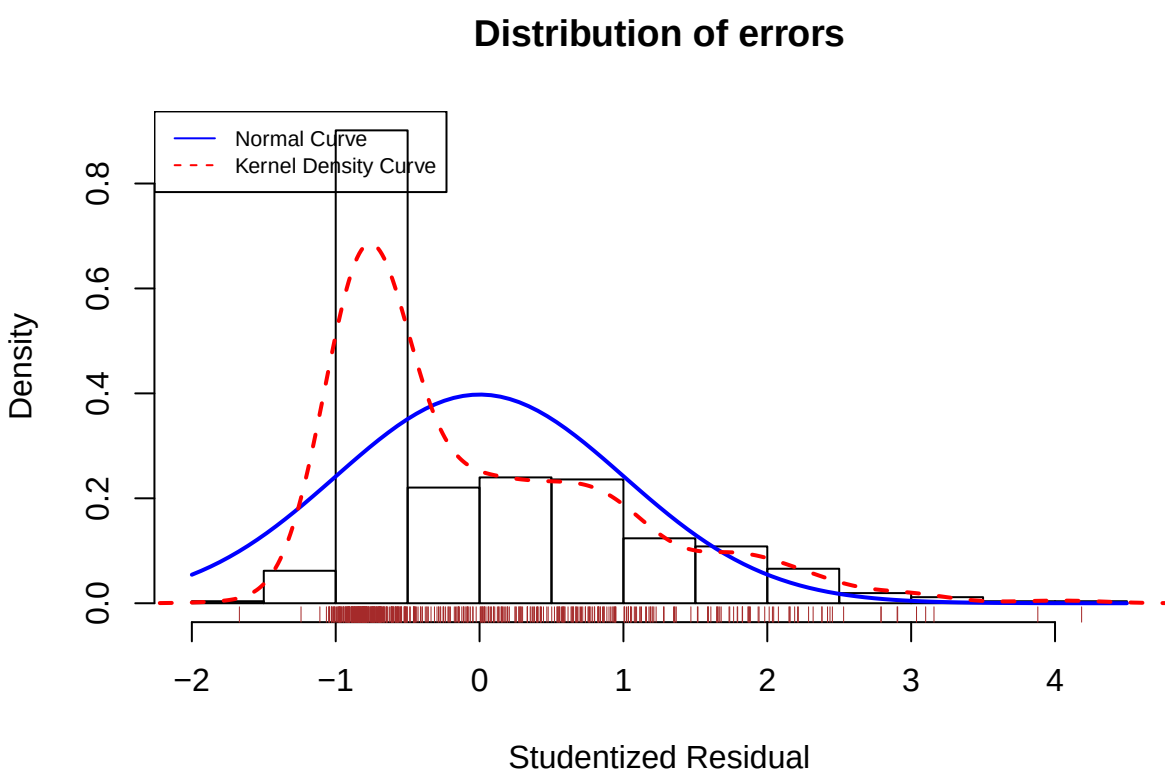
```
qqPlot(model4, labels = row.names(Area), id.method = "identify",  simulate = TRUE, main = "Q-Q Plot for
```



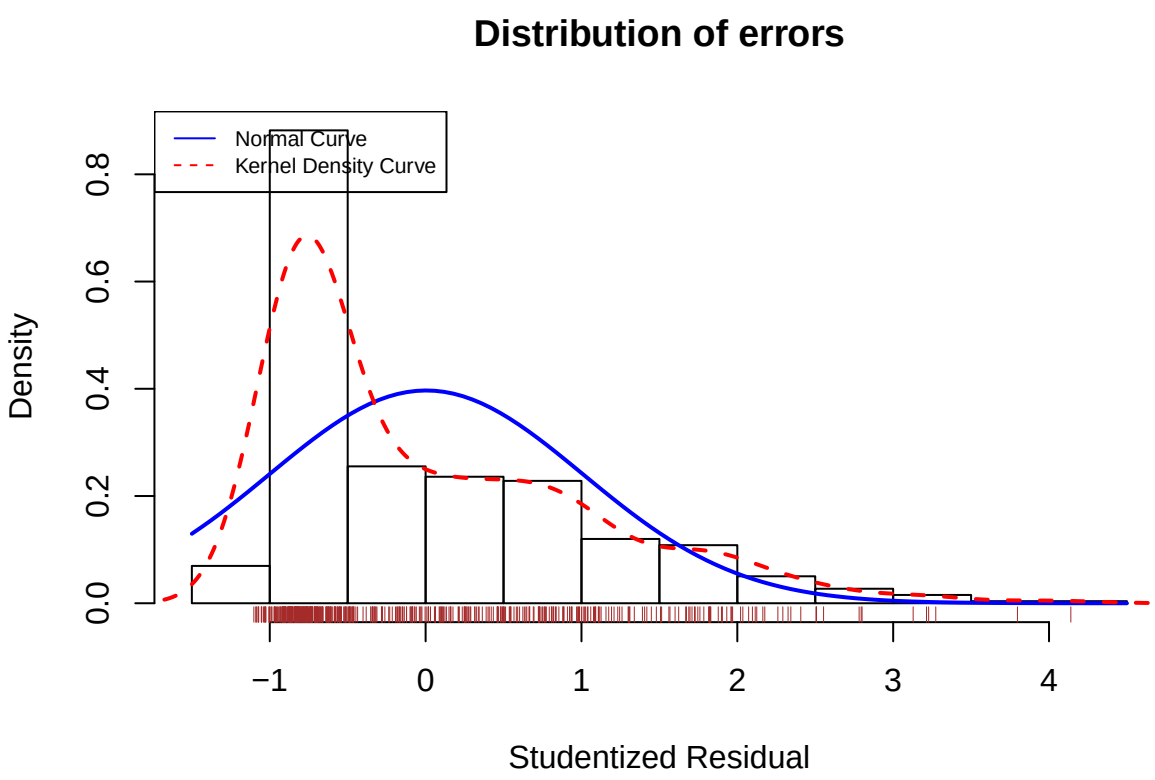

```
## [1] 239 416
```

In this model, majority of the points do not lie on the line suggesting that the normality assumption is violated.

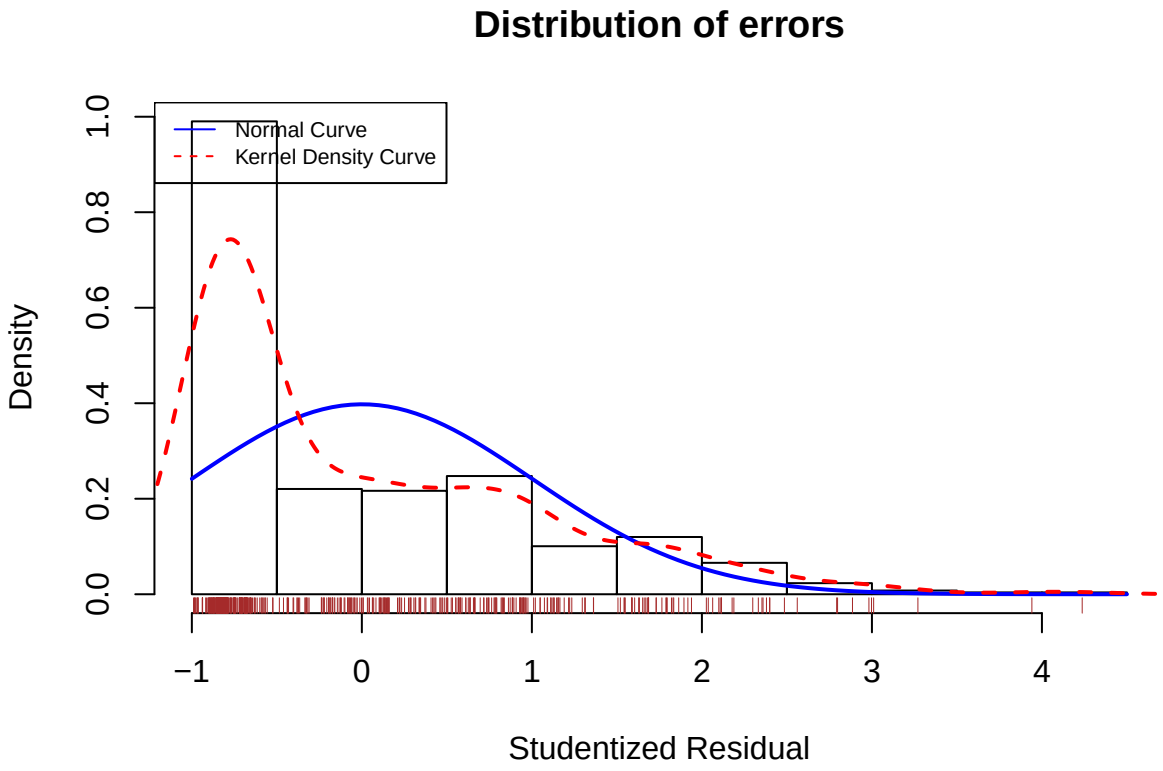
```
residplot <- function(model1, nbreaks = 10)
{
  z <- rstudent(model1)
  hist(z, breaks = nbreaks, freq = FALSE, xlab = "Studentized Residual", main = "Distribution of errors
  rug(jitter(z), col = "brown")
  curve(dnorm(x, mean = mean(z), sd = sd(z)), add = TRUE, col = "blue", lwd = 2)
  lines(density(z)$x, density(z)$y, col = "red", lwd = 2, lty = 2)
  legend("topleft", legend = c("Normal Curve", "Kernel Density Curve"), lty = 1:2, col = c("blue", "red
}
residplot(model1)
```



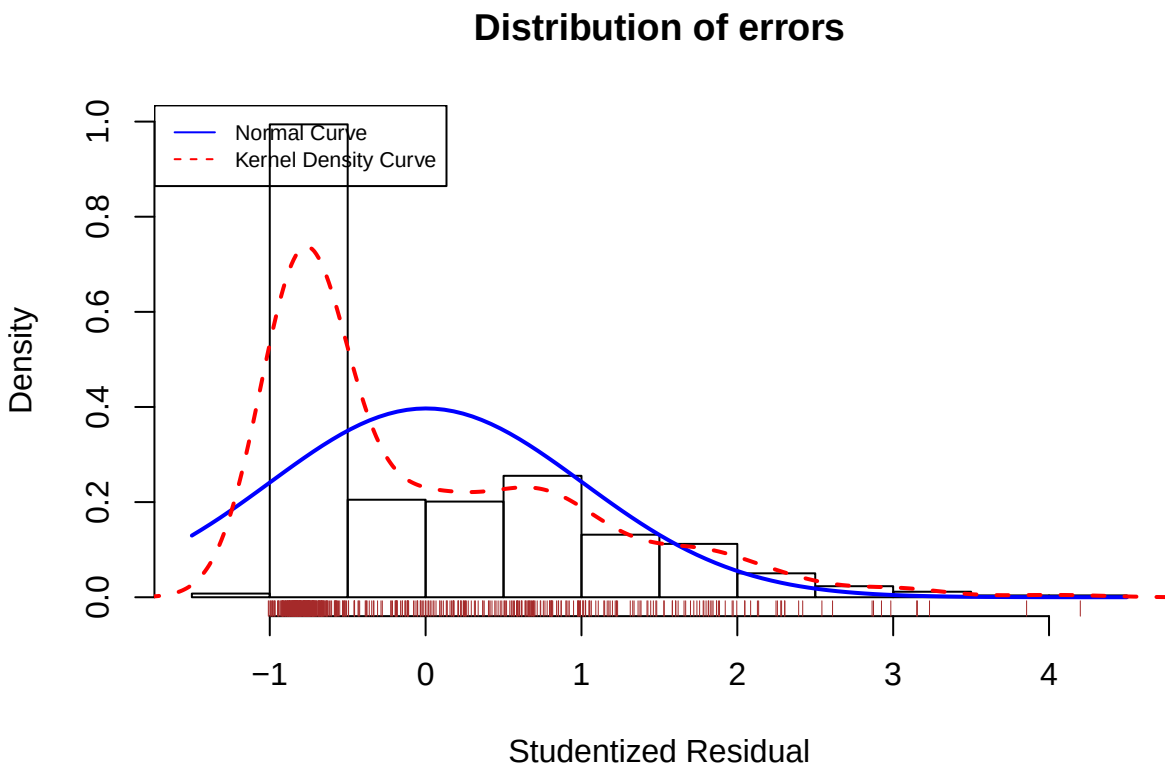

```
residplot(model2)
```




```
residplot(model3)
```




```
residplot(model4)
```



The residplot() function generates a histogram of the studentized residuals and superimposes a normal curve, kernel density curve and rug plot. It is clear that the errors follow the normal distribution quite well, with no outliers.

Independence of error

```
durbinWatsonTest(model2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.5260741     0.9445505       0
##  Alternative hypothesis: rho != 0
```

```
durbinWatsonTest(model1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.5233572     0.9476124       0
##  Alternative hypothesis: rho != 0
```

```
durbinWatsonTest(model3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.5397034       0.91882       0
##  Alternative hypothesis: rho != 0
```

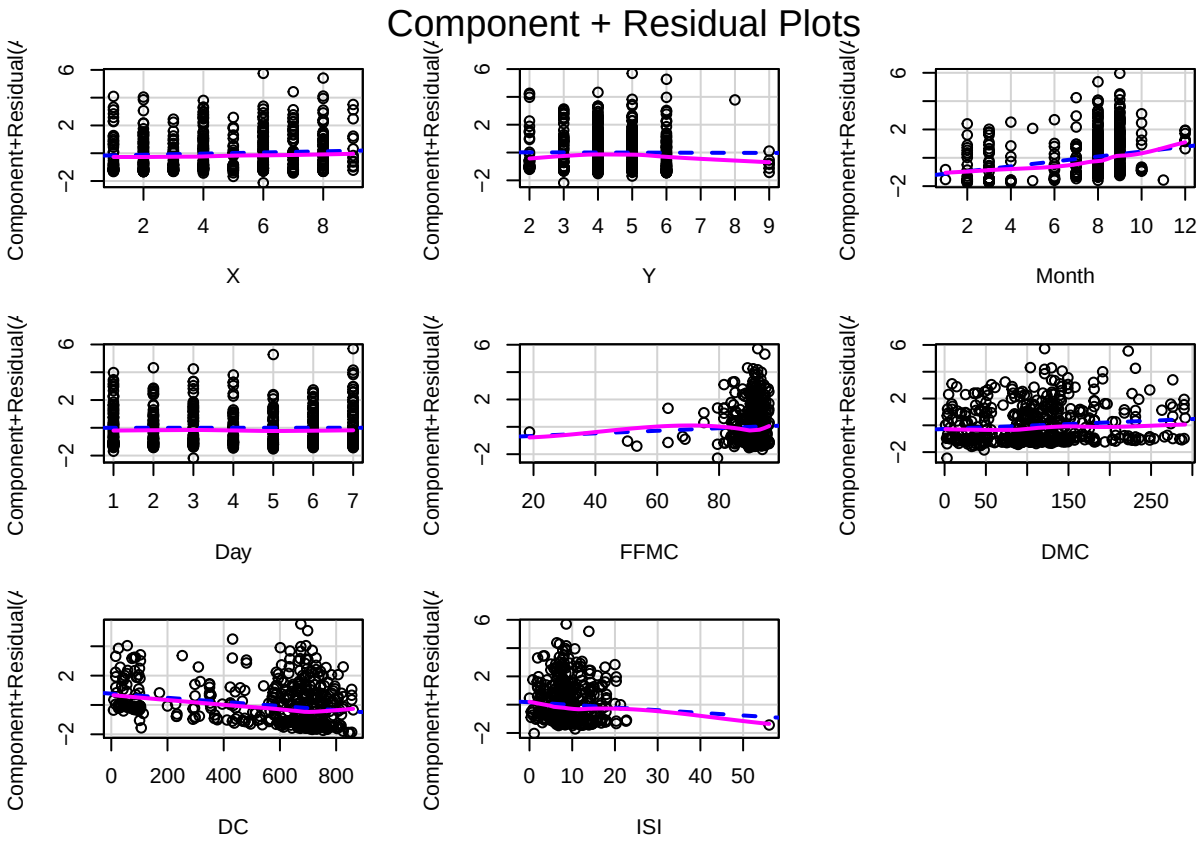
```
durbinWatsonTest(model4)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.5365163     0.9245255       0
##  Alternative hypothesis: rho != 0
```

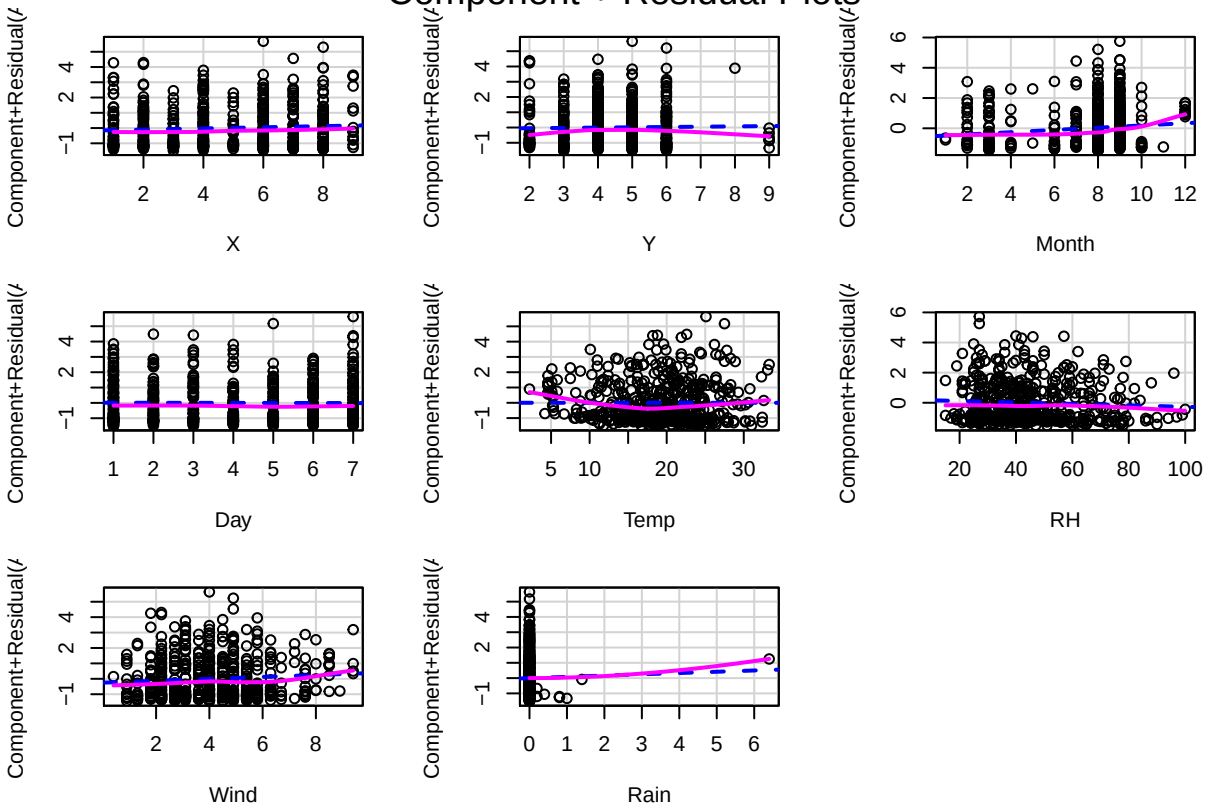
The car package provides the function Durbin-Watson test to detect the serially correlated errors. The non-significant p-values- 0 for model 1, 0 for model 2, and 0 for model 3, suggests a lack of autocorrelation, and conversely an independence of errors. The lag value 1 in all these cases indicate that each observation is being compared with the one next in the dataset.

Linearity

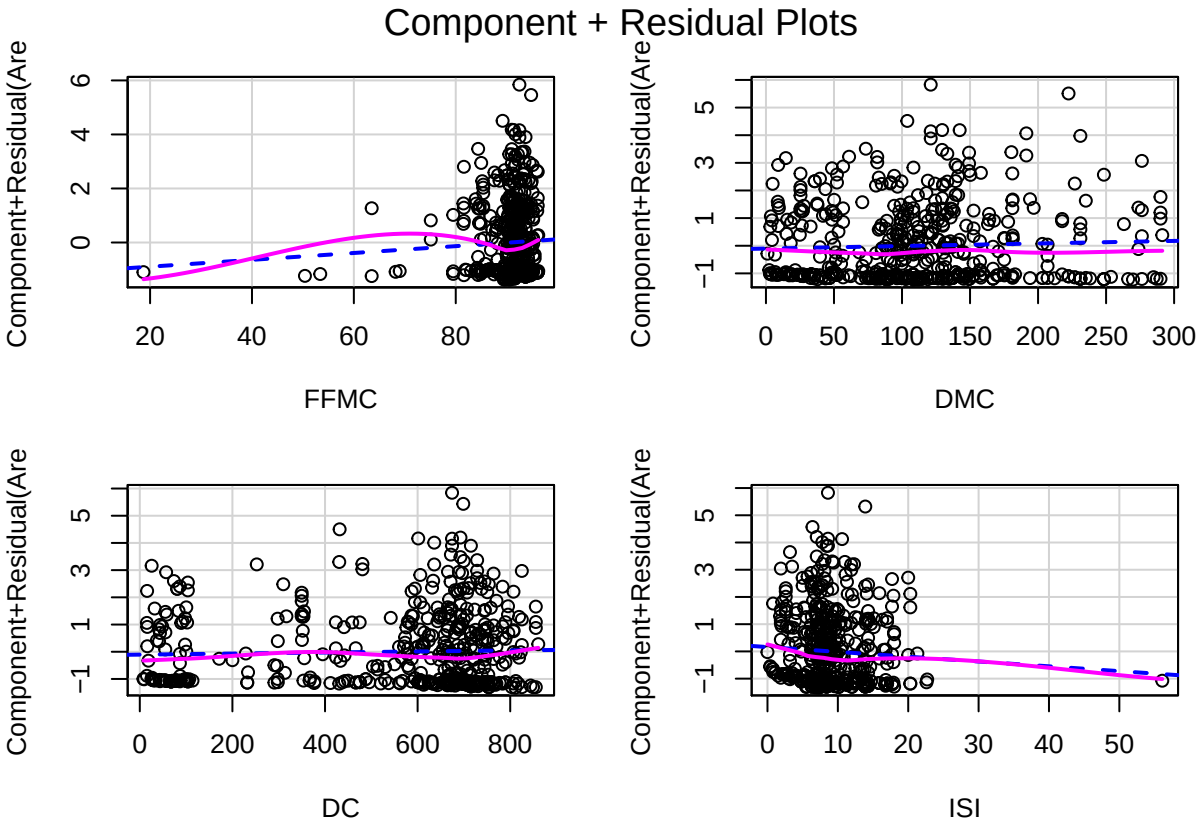
```
crPlots(model1)
```



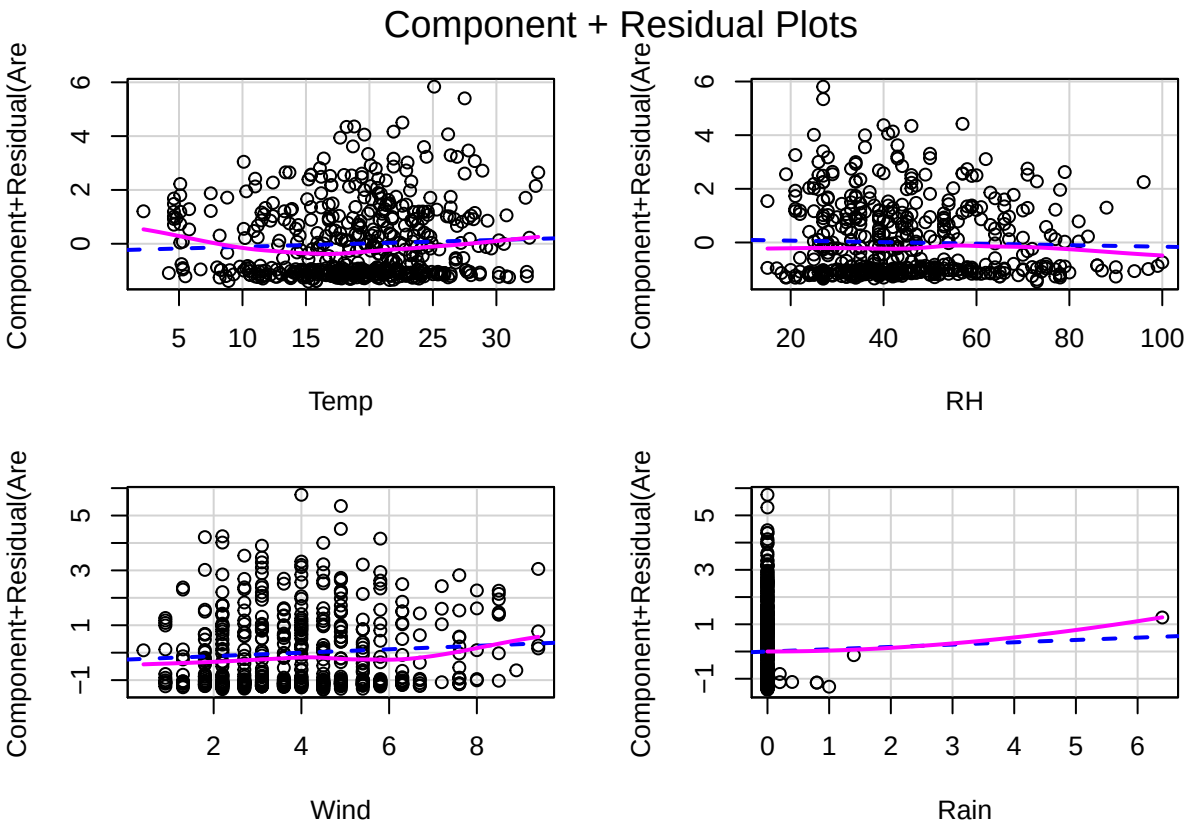

```
crPlots(model2)
```




```
crPlots(model3)
```




```
crPlots(model4)
```



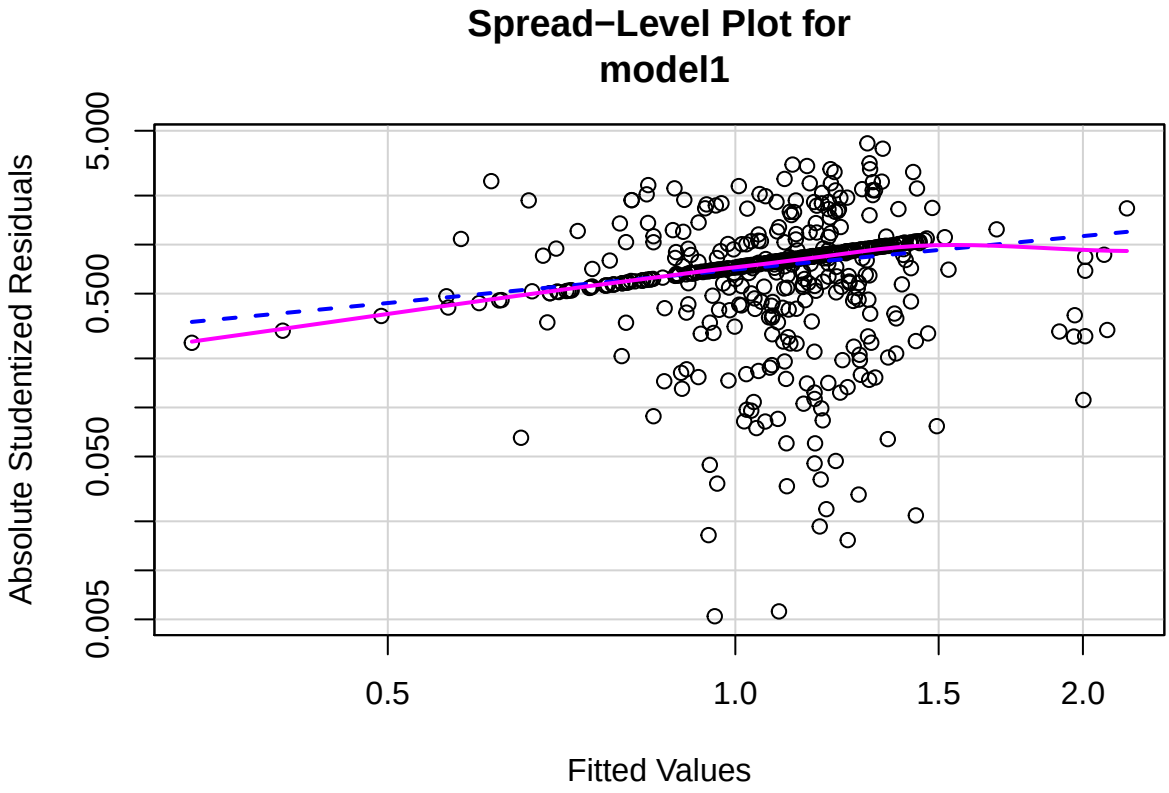
Component plus residual plots (also known as partial residual plots) help to identify the non-linearity in the relationship between the dependent variable and the independent variables. No strong evidence of Non-linearity in any of these plots in any model suggest adequate modelling.

Homoscedasticity

```
ncvTest(model1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 9.341384, Df = 1, p = 0.0022404
```

```
spreadLevelPlot(model1)
```

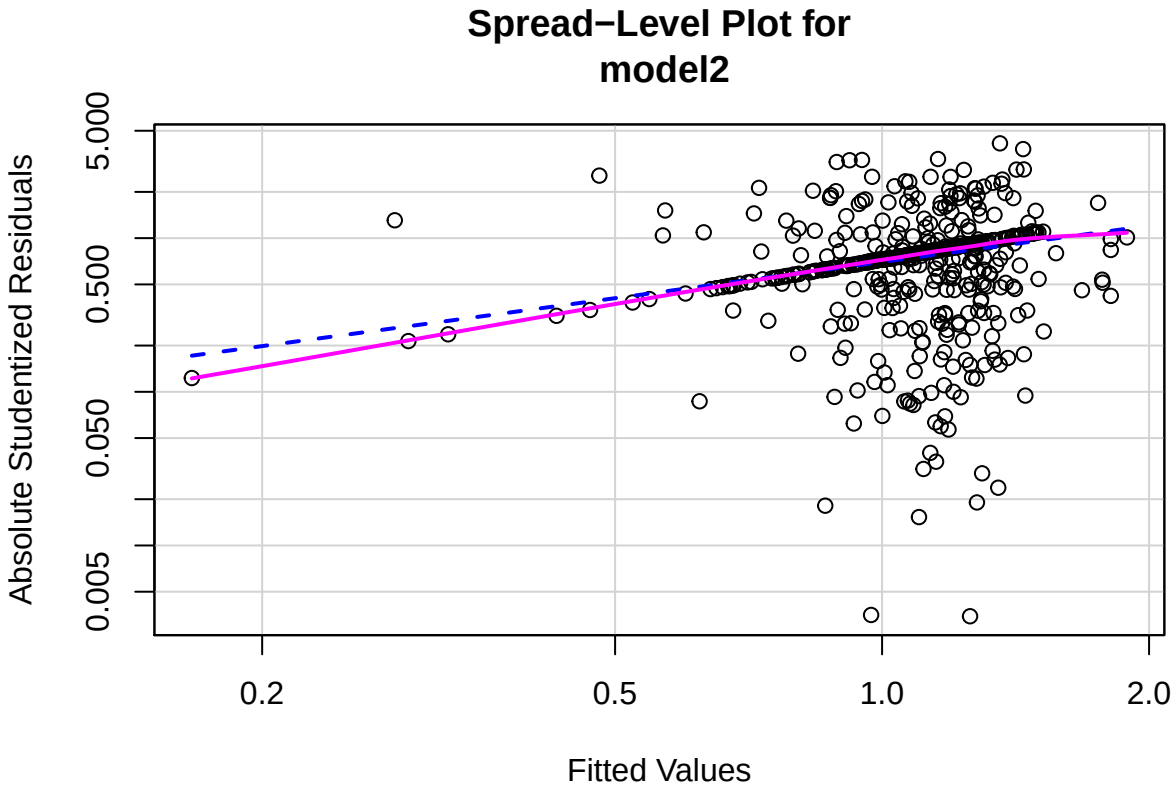



```
##
## Suggested power transformation:  0.3173921
```

```
ncvTest(model2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 5.245158, Df = 1, p = 0.022008
```

```
spreadLevelPlot(model2)
```

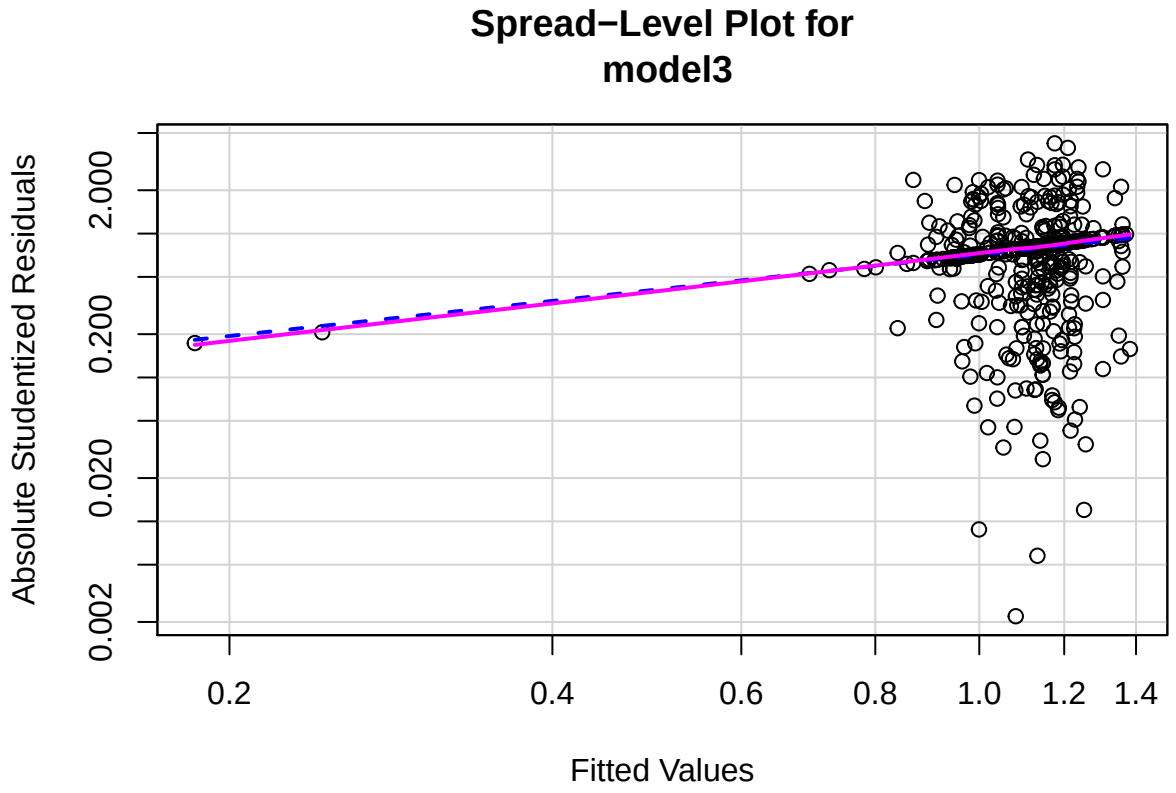



```
##
## Suggested power transformation:  0.2164176
```

```
ncvTest(model3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 5.812039, Df = 1, p = 0.015917
```

```
spreadLevelPlot(model3)
```

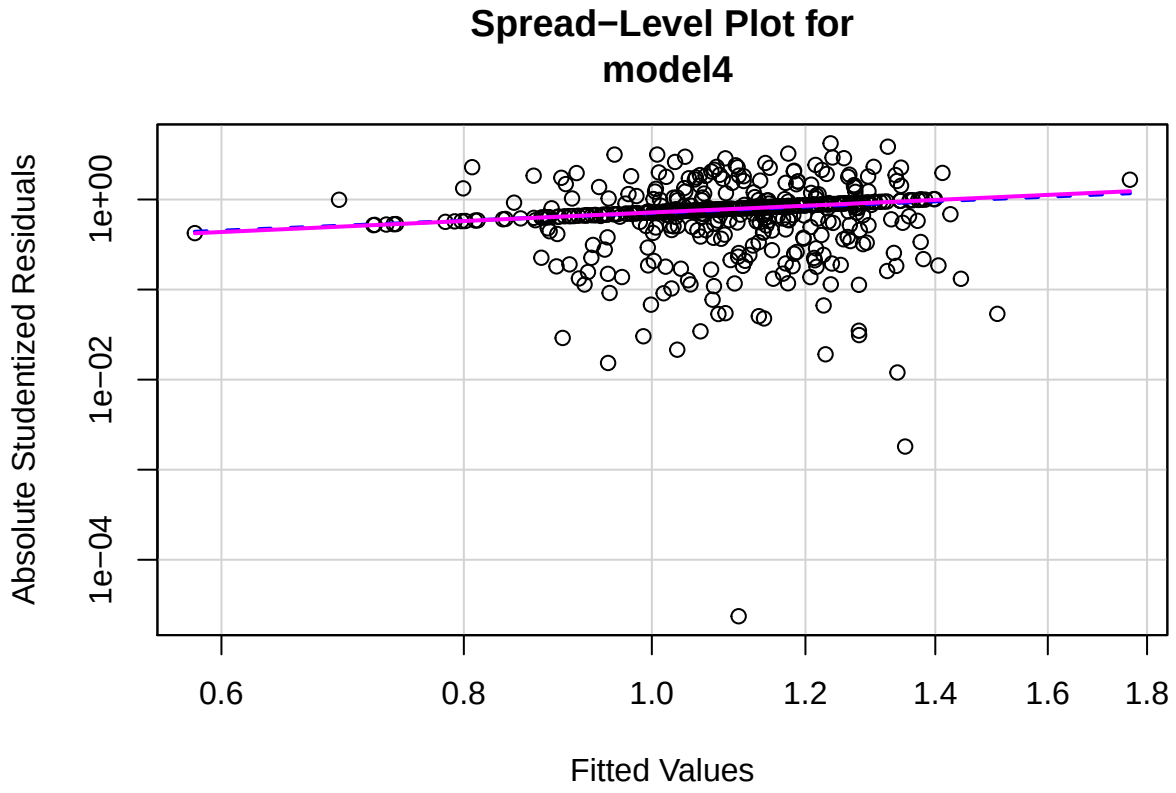



```
##
## Suggested power transformation:  0.1904659
```

```
ncvTest(model4)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 5.59294, Df = 1, p = 0.018033
```

```
spreadLevelPlot(model4)
```




```
##
## Suggested power transformation:  0.113547
```

```
library(gvlma)

gvlm1 <- gvlma(model1)
summary(gvlm1)
```

```
##
## Call:
## lm(formula = Area ~ X + Y + Month + Day + FFMC + DMC + DC + ISI,
##     data = ForestData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1836 -1.0964 -0.6152  0.8838  5.6943
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.5422144  1.1763492  -0.461  0.64505
## X            0.0413122  0.0314864   1.312  0.19009
## Y           -0.0058830  0.0598975  -0.098  0.92180
## Month        0.1739028  0.0585276   2.971  0.00311 **
## Day          0.0006033  0.0286936   0.021  0.98323
## FFMC         0.0093595  0.0137524   0.681  0.49645
```

```
## DMC          0.0024683  0.0014616   1.689  0.09188 .
## DC          -0.0014070  0.0006505  -2.163  0.03100 *
## ISI         -0.0186148  0.0159767  -1.165  0.24452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.388 on 508 degrees of freedom
## Multiple R-squared:  0.02976,    Adjusted R-squared:  0.01448
## F-statistic: 1.948 on 8 and 508 DF,  p-value: 0.05115
##
##
## 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        141.3437 0.0000000 Assumptions NOT satisfied!
## Skewness           116.8635 0.0000000 Assumptions NOT satisfied!
## Kurtosis            14.5626 0.0001356 Assumptions NOT satisfied!
## Link Function        0.3672 0.5445117    Assumptions acceptable.
## Heteroscedasticity   9.5503 0.0019991 Assumptions NOT satisfied!
```

```
gvlm2 <- gvlma(model2)
summary(gvlm2)
```

```
##
## Call:
## lm(formula = Area ~ X + Y + Month + Day + Temp + RH + Wind +
##     Rain, data = ForestData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5192 -1.0932 -0.6405  0.8766  5.6376
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.728e-01  5.183e-01   0.526   0.5988
## X            3.784e-02  3.155e-02   1.199   0.2310
## Y            1.991e-02  5.923e-02   0.336   0.7369
## Month        7.410e-02  2.928e-02   2.530   0.0117 *
## Day         -3.544e-03  2.869e-02  -0.124   0.9017
## Temp        -3.262e-05  1.389e-02  -0.002   0.9981
## RH          -4.833e-03  4.546e-03  -1.063   0.2882
## Wind         6.194e-02  3.531e-02   1.754   0.0800 .
## Rain         8.349e-02  2.114e-01   0.395   0.6930
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.39 on 508 degrees of freedom
## Multiple R-squared:  0.02712,    Adjusted R-squared:  0.0118
## F-statistic:  1.77 on 8 and 508 DF,  p-value: 0.08051
```

```
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = model2)
##
##                      Value   p-value                   Decision
## Global Stat        147.676 0.0000000 Assumptions NOT satisfied!
## Skewness           120.833 0.0000000 Assumptions NOT satisfied!
## Kurtosis            16.467 0.0000495 Assumptions NOT satisfied!
## Link Function        3.744 0.0529907    Assumptions acceptable.
## Heteroscedasticity   6.632 0.0100146 Assumptions NOT satisfied!
```

```
gvlm3 <- gvlma(model3)
summary(gvlm3)
```

```
##
## Call:
## lm(formula = Area ~ FFMC + DMC + DC + ISI, data = ForestData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3703 -1.1242 -0.6145  0.8882  5.8198
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0851719  1.1513950  -0.074    0.941
## FFMC         0.0126619  0.0137577   0.920    0.358
## DMC          0.0009234  0.0013587   0.680    0.497
## DC           0.0001928  0.0003412   0.565    0.572
## ISI         -0.0176878  0.0160811  -1.100    0.272
##
## Residual standard error: 1.398 on 512 degrees of freedom
## Multiple R-squared:  0.008046,   Adjusted R-squared:  0.0002959
## F-statistic: 1.038 on 4 and 512 DF,  p-value: 0.3869
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = model3)
##
##                       Value   p-value                   Decision
## Global Stat        146.1021 0.0000000 Assumptions NOT satisfied!
## Skewness           122.1974 0.0000000 Assumptions NOT satisfied!
## Kurtosis            15.4841 0.0000832 Assumptions NOT satisfied!
## Link Function        0.2369 0.6264190    Assumptions acceptable.
## Heteroscedasticity   8.1836 0.0042270 Assumptions NOT satisfied!
```

```r
gvm4 <- gvlma(model4)
summary(gvm4)
```

```
##
## Call:
## lm(formula = Area ~ Temp + RH + Wind + Rain, data = ForestData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3993 -1.0978 -0.7081  0.9121  5.7593
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.742041   0.443148   1.674   0.0946 .
## Temp         0.012766   0.012958   0.985   0.3250
## RH          -0.002834   0.004506  -0.629   0.5296
## Wind         0.062603   0.035446   1.766   0.0780 .
## Rain         0.085167   0.211852   0.402   0.6878
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.397 on 512 degrees of freedom
## Multiple R-squared:  0.0104, Adjusted R-squared:  0.002671
## F-statistic: 1.345 on 4 and 512 DF,  p-value: 0.2519
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = model4)
##
##                       Value   p-value                   Decision
## Global Stat        143.06714 0.000e+00 Assumptions NOT satisfied!
## Skewness           121.77664 0.000e+00 Assumptions NOT satisfied!
## Kurtosis            15.46861 8.389e-05 Assumptions NOT satisfied!
## Link Function        0.05692 8.114e-01    Assumptions acceptable.
## Heteroscedasticity   5.76497 1.635e-02 Assumptions NOT satisfied!
```

Outliers test

```r
outlierTest(model1)
```

```
##     rstudent unadjusted p-value Bonferonni p
## 239 4.184935         3.3625e-05     0.017384
```

```r
outlierTest(model2)
```

```
##     rstudent unadjusted p-value Bonferonni p
## 239 4.139766         4.0714e-05     0.021049
```

```
outlierTest(model3)
```

```
##     rstudent unadjusted p-value Bonferonni p
## 239 4.236704         2.6911e-05     0.013913
## 416 3.940781         9.2520e-05     0.047833
```

```
outlierTest(model4)
```

```
##     rstudent unadjusted p-value Bonferonni p
## 239  4.20067         3.1404e-05     0.016236
```

Model 1: Observation 239 is indicated to be an outlier. Similar results mentioned above.

High Leverage Points and Influential Observations

```
hat.plot <- function(model1)
{p <- length(coefficients(model1))
n <- length(fitted(model1))
plot(hatvalues(model1), main = "Index Plot of Hat Values")
abline(h=c(2,3)*p/n, col = "red", lty = 2)
identify(1:n, hatvalues(model1), names(hatvalues(model1)))}
 vif(model1)
```

```
##        X        Y    Month      Day     FFMC      DMC       DC      ISI
## 1.420984 1.452972 4.750748 1.013132 1.542944 2.346162 6.970648 1.420706
```

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

```
##     X     Y Month   Day  FFMC   DMC    DC   ISI
## FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE
```

```
 vif(model2)
```

```
##        X        Y    Month      Day     Temp       RH     Wind     Rain
## 1.423191 1.417051 1.186074 1.010307 1.736402 1.469112 1.068352 1.044938
```

```
 sqrt(vif(model2))>2
```

```
##     X     Y Month   Day  Temp    RH  Wind  Rain
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
 vif(model3)
```

```
##     FFMC      DMC       DC      ISI
## 1.522230 1.998603 1.891342 1.418904
```

```r
sqrt(vif(model3))>2
```

```
##  FFMC   DMC    DC   ISI
## FALSE FALSE FALSE FALSE
```

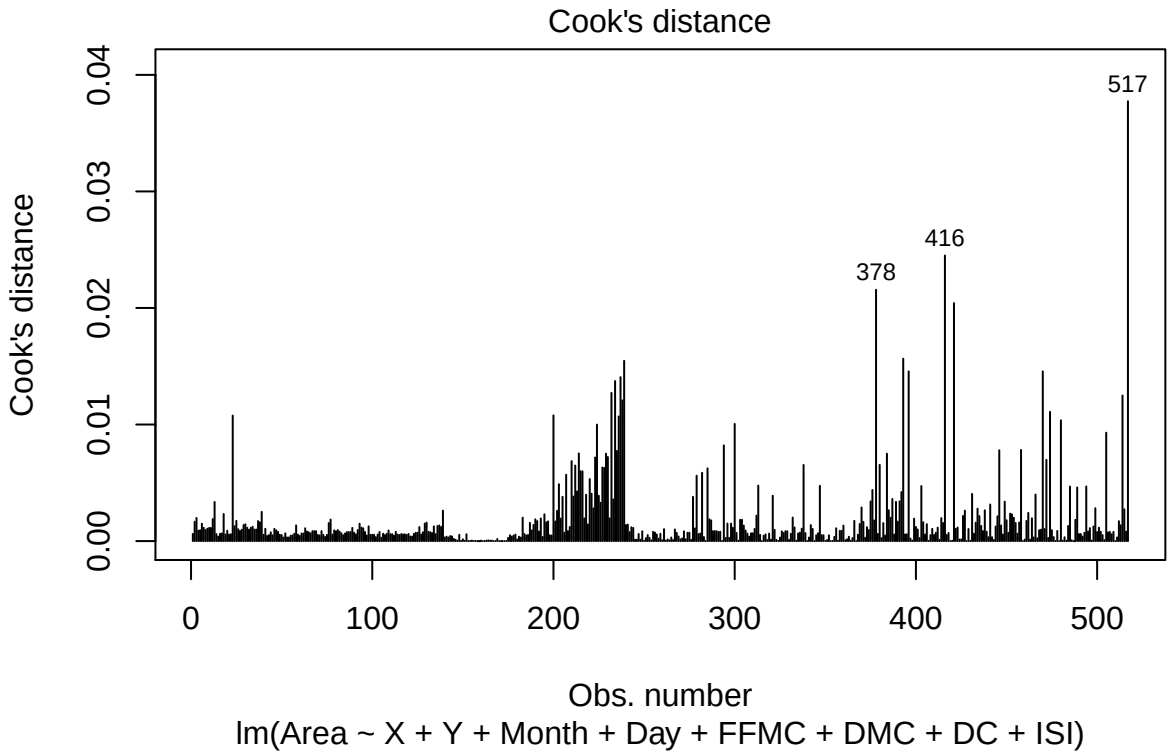
```r
vif(model4)
```

```
##     Temp       RH     Wind     Rain
## 1.497878 1.430149 1.067008 1.040048
```

```r
sqrt(vif(model4))>2
```

```
##  Temp    RH  Wind  Rain
## FALSE FALSE FALSE FALSE
```

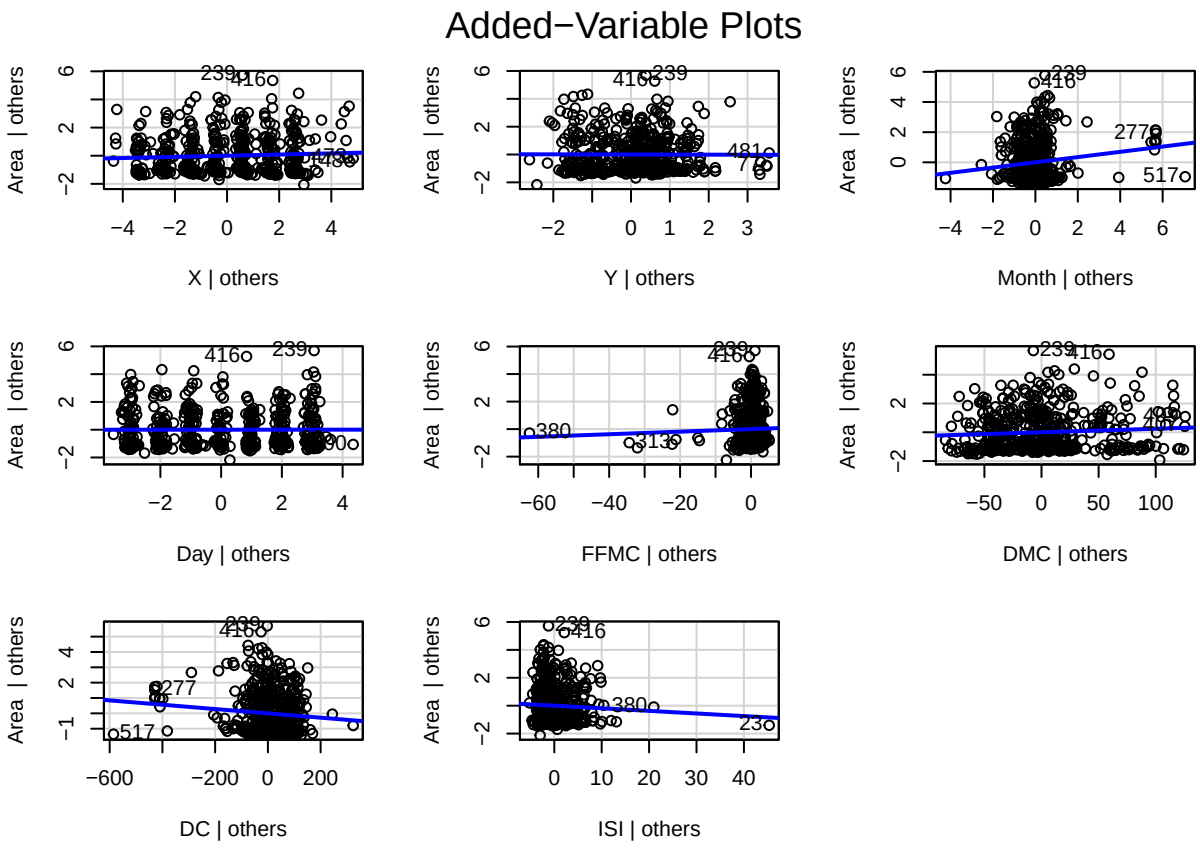
```r
plot(model1, which = 4)
```



All false show no multicollinearity problems.

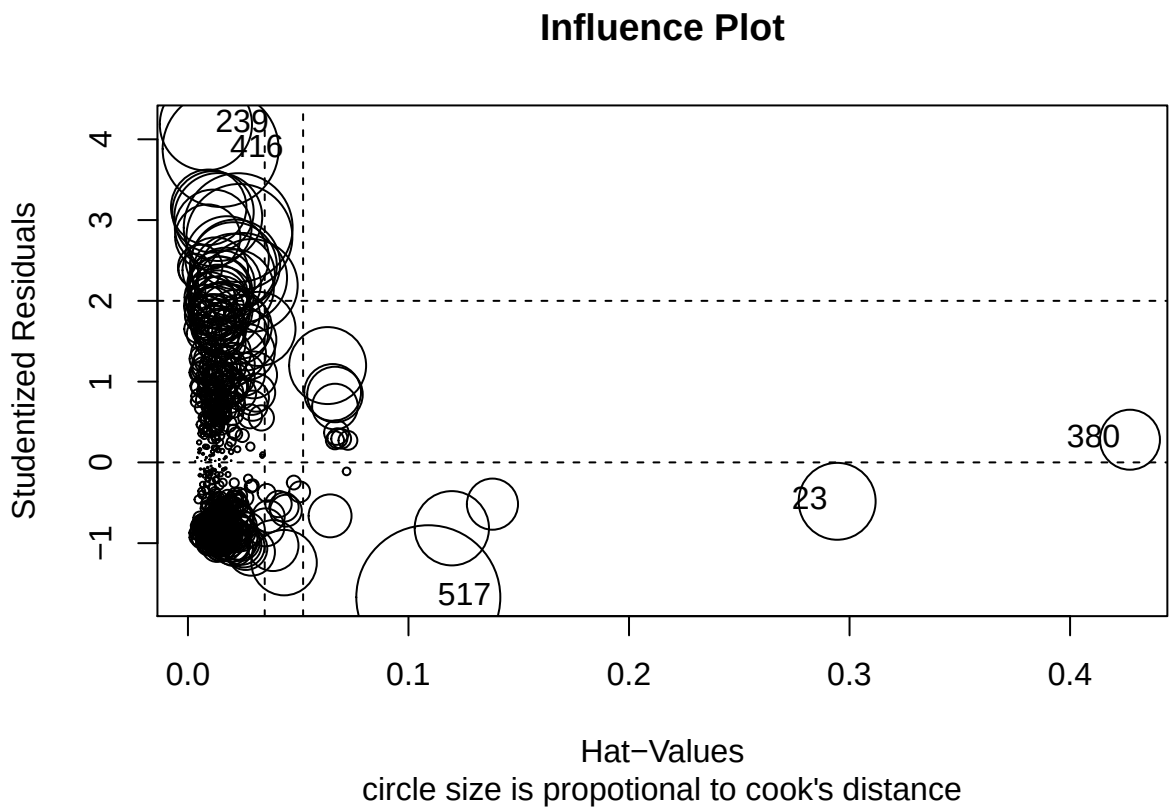
. Added Variable Plots method

```r
avPlots(model1, ask = FALSE, onepage = TRUE, id.method = "identify")
```



#Influence Plots

```
influencePlot(model1, method = "identity", main = 'Influence Plot', sub = "circle size is propotional t
```




```
##        StudRes         Hat       CookD
## 23  -0.4816201 0.294385218 0.010768921
## 239  4.1849350 0.008140585 0.015468469
## 380  0.2807794 0.427151610 0.006543628
## 416  3.8805224 0.014825431 0.024500593
## 517 -1.6692786 0.109000120 0.037743298
```

Corrective Measures

Deleting Observations: Observation 517 is the influential observation. We delete this variable to check the impact of this deletion on the model.

```
ForestData_del <- ForestData[-c(517),]
model1_del <- lm(Area ~ X + Y + Month + Day + FFMC + DMC + DC + ISI, data=ForestData_del)
summary(model1_del)
```

```
##
## Call:
## lm(formula = Area ~ X + Y + Month + Day + FFMC + DMC + DC + ISI,
##     data = ForestData_del)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1.8134 -1.1042 -0.6062  0.8621  5.6842
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.4371379  1.1759721  -0.372 0.710253
## X            0.0450284  0.0315099   1.429 0.153614
## Y           -0.0169325  0.0601577  -0.281 0.778467
## Month        0.2046721  0.0612637   3.341 0.000897 ***
## Day          0.0009054  0.0286438   0.032 0.974795
## FFMC         0.0076913  0.0137646   0.559 0.576562
## DMC          0.0027495  0.0014687   1.872 0.061782 .
## DC          -0.0017220  0.0006762  -2.547 0.011170 *
## ISI         -0.0195838  0.0159592  -1.227 0.220349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.386 on 507 degrees of freedom
## Multiple R-squared:  0.03388,    Adjusted R-squared:  0.01864
## F-statistic: 2.223 on 8 and 507 DF,  p-value: 0.02462
```

```
ForestData_Transform <- ForestData
ForestData_Transform$Area <- ForestData$Area + 1
summary(powerTransform(ForestData_Transform$Area))
```

```
## bcPower Transformation to Normality
##                            Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## ForestData_Transform$Area   -0.7143       -0.71      -0.9145      -0.5142
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df       pval
## LR test, lambda = (0) 53.88499  1 2.1261e-13
##
## Likelihood ratio test that no transformation is needed
##                             LRT df       pval
## LR test, lambda = (1) 348.2111  1 < 2.22e-16
```

```
ForestData_Transform[,13] = ForestData_Transform[,13]^(-0.71)
model1_transform <- lm(Area ~ X + Y + Month + Day + FFMC + DMC + DC + ISI, data= ForestData_Transform)
summary(model1_transform)
```

```
##
## Call:
## lm(formula = Area ~ X + Y + Month + Day + FFMC + DMC + DC + ISI,
##     data = ForestData_Transform)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50665 -0.26285  0.03833  0.26411  0.43768
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0987423  0.2313672   4.749 2.66e-06 ***
```

```
## X           -0.0063275  0.0061928  -1.022  0.30739
## Y           -0.0041720  0.0117808  -0.354  0.72338
## Month       -0.0309074  0.0115113  -2.685  0.00749 **
## Day          0.0002333  0.0056435   0.041  0.96704
## FFMC        -0.0021556  0.0027048  -0.797  0.42585
## DMC         -0.0003020  0.0002875  -1.050  0.29403
## DC           0.0002102  0.0001279   1.643  0.10092
## ISI          0.0028984  0.0031423   0.922  0.35677
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.273 on 508 degrees of freedom
## Multiple R-squared:  0.02739,    Adjusted R-squared:  0.01207
## F-statistic: 1.788 on 8 and 508 DF,  p-value: 0.07692
```

After changing the response variable, we have modified the original model. We can see there is an effect in the residual value of the model after this transformation.

Deduction

```
AIC(model4,model3, model1, model2)
```

```
##        df      AIC
## model4  6 1819.532
## model3  6 1820.762
## model1 10 1817.317
## model2 10 1818.724
```

AIC indicate that model1 and model2 are the best models with a negligible or a very low difference between the two.

Fine tune the selection of the predictor variables

Stepwise regression

```
library(MASS)
stepAIC(model1, direction="backward")
```

```
## Start:  AIC=348.13
## Area ~ X + Y + Month + Day + FFMC + DMC + DC + ISI
##
##         Df Sum of Sq    RSS    AIC
## - Day    1    0.0009 979.07 346.14
## - Y      1    0.0186 979.09 346.14
## - FFMC   1    0.8927 979.96 346.61
## - ISI    1    2.6163 981.68 347.51
## - X      1    3.3179 982.39 347.88
## <none>               979.07 348.13
## - DMC    1    5.4964 984.56 349.03
## - DC     1    9.0176 988.09 350.87
```

```
## - Month  1   17.0154 996.08 355.04
##
## Step:  AIC=346.14
## Area ~ X + Y + Month + FFMC + DMC + DC + ISI
##
##         Df Sum of Sq    RSS    AIC
## - Y      1    0.0187 979.09 344.15
## - FFMC   1    0.9010 979.97 344.61
## - ISI    1    2.6168 981.69 345.52
## - X      1    3.3218 982.39 345.89
## <none>               979.07 346.14
## - DMC    1    5.4957 984.56 347.03
## - DC     1    9.0295 988.10 348.88
## - Month  1   17.0963 996.17 353.09
##
## Step:  AIC=344.15
## Area ~ X + Month + FFMC + DMC + DC + ISI
##
##         Df Sum of Sq    RSS    AIC
## - FFMC   1    0.9129 980.00 342.63
## - ISI    1    2.6076 981.70 343.52
## <none>               979.09 344.15
## - X      1    4.3057 983.39 344.41
## - DMC    1    5.5245 984.61 345.05
## - DC     1    9.0957 988.18 346.93
## - Month  1   17.1361 996.22 351.12
##
## Step:  AIC=342.63
## Area ~ X + Month + DMC + DC + ISI
##
##         Df Sum of Sq    RSS    AIC
## - ISI    1    1.7448 981.75 341.55
## <none>               980.00 342.63
## - X      1    4.2678 984.27 342.87
## - DMC    1    6.4709 986.47 344.03
## - DC     1    9.2024 989.20 345.46
## - Month  1   17.9221 997.92 350.00
##
## Step:  AIC=341.55
## Area ~ X + Month + DMC + DC
##
##         Df Sum of Sq    RSS    AIC
## <none>               981.75 341.55
## - X      1    4.1453 985.89 341.72
## - DMC    1    5.3619 987.11 342.36
## - DC     1    8.9994 990.75 344.26
## - Month  1   17.4060 999.15 348.63

##
## Call:
## lm(formula = Area ~ X + Month + DMC + DC, data = ForestData)
##
## Coefficients:
## (Intercept)            X        Month          DMC           DC
```

```
##    0.133650     0.038898     0.173818     0.002322    -0.001389
```

Here, AIC in the last step is 341.55 and the selected list of predictors is as indicated in the last step i.e. X, Month, DMC, and DC

Interpret the prediction results

Model1 is suggested to be the best model with the predictors indicated in the above steps.

Let us see how the models improve the fit of our dataset.

The final model with the indicated set of predictors is-

```
bestfit <- lm(Area ~ X + Month + DMC + DC, data = ForestData)
summary(bestfit)
```

```
##
## Call:
## lm(formula = Area ~ X + Month + DMC + DC, data = ForestData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1378 -1.0966 -0.5795  0.8991  5.7198
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1336499  0.2716132   0.492  0.62289
## X            0.0388981  0.0264556   1.470  0.14209
## Month        0.1738185  0.0576914   3.013  0.00272 **
## DMC          0.0023223  0.0013888   1.672  0.09509 .
## DC          -0.0013891  0.0006412  -2.166  0.03074 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.385 on 512 degrees of freedom
## Multiple R-squared:  0.02711,    Adjusted R-squared:  0.01951
## F-statistic: 3.567 on 4 and 512 DF,  p-value: 0.006986
```

We can see that the above model improves the fit of the dataset and thus the best fit model.

Results from the scatterplot matrix don't clearly show linear correlation of the predictor variables of the area. Since our data has more than one predictor we will have to conduct multiple linear regression. The journal "Data Mining approaches to predicting forest fires using Meteorological data" suggests 4 different models with different feature selection setups namely STFWI, STM, FWI and M, corresponding to Model1,2,3,and 4 respectively.

Typical approach to regression model shows 95% certainity that change of 1% in predictor variables will cause a change in the response variable Area, between the ranges described in each variable while keeping the other variables constant. In the normal plot since there are points that deviate from the 45 degree angle it is inferred that it violates the normality rule. To obey the linearity rule the Residuals vs Fitted plot needs to have a flat curve which seems to be the case with the graph above and hence doesn't require a quadratic term. The scale-location plot shows points in a random band around the horizontal line hence it has homoscedasticity. Residuals vs leverage plot gives information about the outliers and high leverage points which is corrected in the latter steps. In the enhanced approach OLS regression is used where the 4

assumptions, Normality, Independence, Linearity and Homoscedasticity are tested. Most results in the QQ plot are not within the confidence envelope hence suggesting non-normality. Distribute or errors suggests data is mostly left skewed. Significant p value in the Durbin-watson test suggests an autocorrection feature in the dataset so no independence of errors. The component plus residual plots show no linearity with the dependent variable. Significant chi quare test value and spread level plot show presence of homoskedasticity.

Outlier points 239 and 416 are given in the outlier test for the 4 models. They can be deleted for corrective measures but we only delete influential observation 517 giving errors to improve regression diagnostics. Cooks Distance plot is shown. Transformation required because the models didn't meet normality and linearity assumptions and also because deleting observations did not result in a better fit of the model as observed in the Adjusted R squared value.. Lambda value of 1 was used. p value for all models are above 0.9 suggesting a normal distributing of the dependent variables. Since variables were changed to factors in the beginning of the code it affects multicollinearity as seen in sqrt VIF value table. Only DC has a value greater than 2 hence it was dropped to maintain multicollinearity. Based on AIC and Adjusted R squared transformed model1 and transformed model appears to be the best. Stepwise methods and all subsets regression methods are 2 popular approaches when it comes to selecting the final set of predictor variables from the chosen model T. Stepwise methodindicate X.fac, Y.fac, Month.fac and Wind being the optimal model.