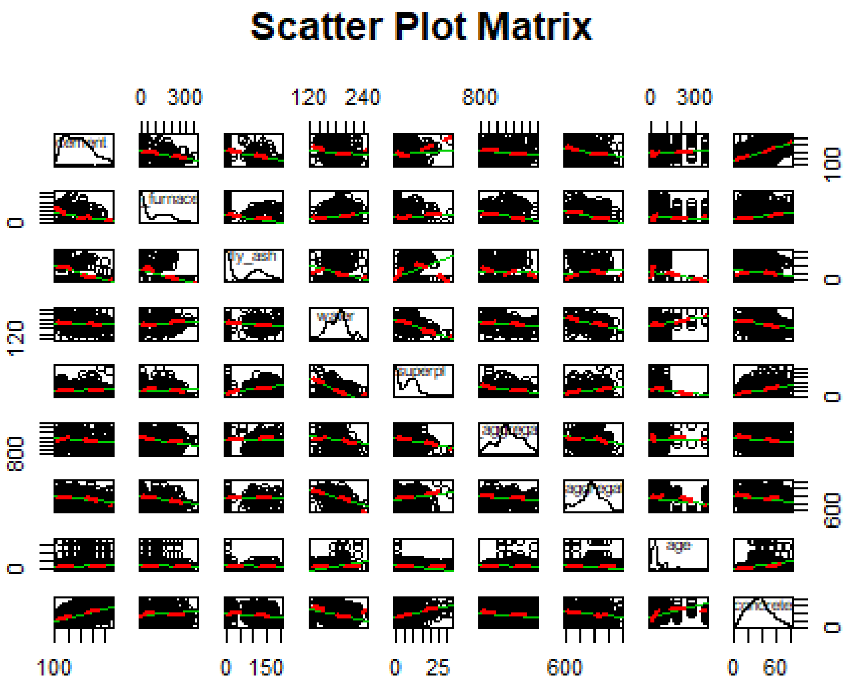
# R Notebook

```r
library(readxl)
library(tidyverse)

## -- Attaching packages ---------------------------------------------------------------
--------------- tidyverse 1.2.1 --

## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.1     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0

## -- Conflicts ------------------------------------------------------------------------
--------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(car)

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

Concrete_Data <- read_excel("C:/Users/HP DV6/Downloads/Concrete_Data.xls")

# Rename column name to make it easier to operate on.
Concrete_Data <- rename(Concrete_Data,
  concrete = `Concrete compressive strength(MPa, megapascals)`,
  cement = `Cement (component 1)(kg in a m^3 mixture)`,
  b_furnace = `Blast Furnace Slag (component 2)(kg in a m^3 mixture)`,
  water = `Water  (component 4)(kg in a m^3 mixture)`,
  superpl = `Superplasticizer (component 5)(kg in a m^3 mixture)`,
  c_aggregate = `Coarse Aggregate  (component 6)(kg in a m^3 mixture)`,
  f_aggregate = `Fine Aggregate (component 7)(kg in a m^3 mixture)`,
  age = `Age (day)`,
  fly_ash = `Fly Ash (component 3)(kg in a m^3 mixture)`



)
train <- Concrete_Data[1:900,]
test <- Concrete_Data[901:1030,]
```

Graphical analysis of concrete data set.

Examining bivariate relationship between dependent variables to determine if interaction effect exist.

```
cor(Concrete_Data)

##                  cement   b_furnace      fly_ash       water     superpl
## cement       1.00000000 -0.27519344 -0.397475440 -0.08154361  0.09277137
## b_furnace   -0.27519344  1.00000000 -0.323569468  0.10728594  0.04337574
## fly_ash     -0.39747544 -0.32356947  1.000000000 -0.25704400  0.37733956
## water       -0.08154361  0.10728594 -0.257043997  1.00000000 -0.65746444
## superpl      0.09277137  0.04337574  0.377339559 -0.65746444  1.00000000
## c_aggregate -0.10935604 -0.28399823 -0.009976788 -0.18231167 -0.26630276
## f_aggregate -0.22272017 -0.28159326  0.079076351 -0.45063498  0.22250149
## age          0.08194726 -0.04424580 -0.154370165  0.27760443 -0.19271652
## concrete     0.49783272  0.13482445 -0.105753348 -0.28961348  0.36610230
##              c_aggregate f_aggregate          age   concrete
## cement      -0.109356039 -0.22272017  0.081947264  0.4978327
## b_furnace   -0.283998230 -0.28159326 -0.044245801  0.1348244
## fly_ash     -0.009976788  0.07907635 -0.154370165 -0.1057533
## water       -0.182311668 -0.45063498  0.277604429 -0.2896135
## superpl     -0.266302755  0.22250149 -0.192716518  0.3661023
## c_aggregate  1.000000000 -0.17850575 -0.003015507 -0.1649278
## f_aggregate -0.178505755  1.00000000 -0.156094049 -0.1672490
## age         -0.003015507 -0.15609405  1.000000000  0.3288770
## concrete    -0.164927821 -0.16724896  0.328876976  1.0000000

scatterplotMatrix(Concrete_Data, spread=FALSE, smoother.args=list(lty=2),
main="Scatter Plot Matrix")
```



Fitting a muitlple linear regression on the data set.

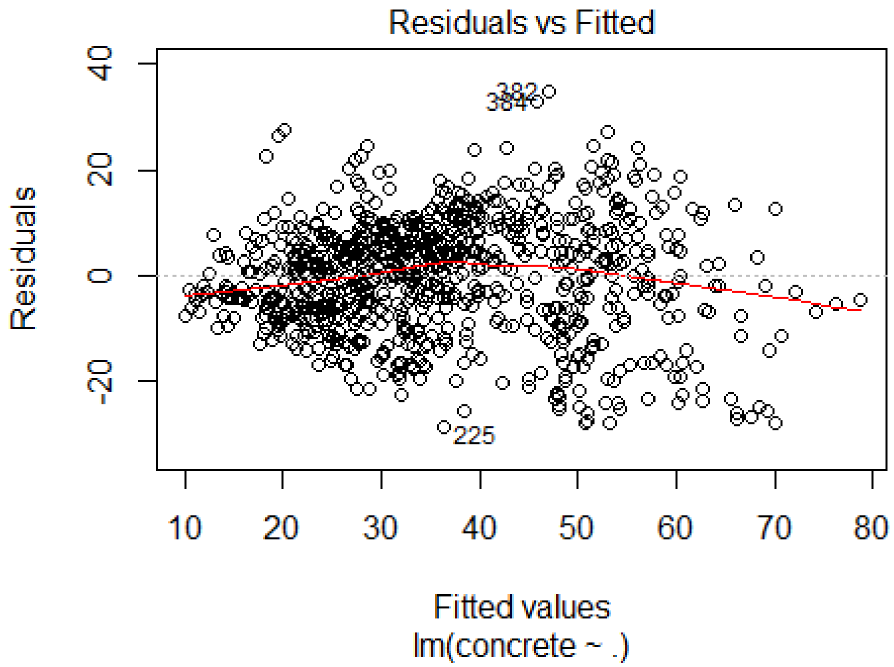
```
lm_fit <- lm(concrete ~ ., data = train)
summary(lm_fit)

##
## Call:
## lm(formula = concrete ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.596  -6.665   0.849   7.090  34.769
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.390738  29.311282  -0.116  0.90793
## cement       0.110993   0.009559  11.612  < 2e-16 ***
## b_furnace    0.094837   0.011159   8.499  < 2e-16 ***
## fly_ash      0.086214   0.013920   6.193 8.98e-10 ***
## water       -0.175662   0.044412  -3.955 8.25e-05 ***
## superpl      0.311767   0.104489   2.984  0.00293 **
## c_aggregate  0.013301   0.010350   1.285  0.19908
## f_aggregate  0.010455   0.011924   0.877  0.38085
## age          0.115364   0.005688  20.281  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

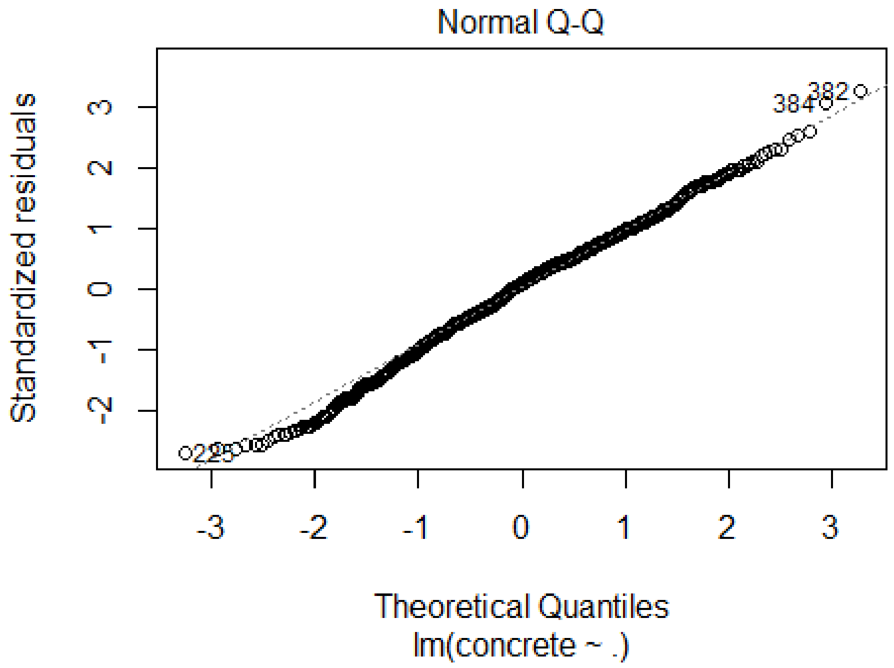
```
## Residual standard error: 10.76 on 891 degrees of freedom
## Multiple R-squared:  0.6127, Adjusted R-squared:  0.6092
## F-statistic: 176.2 on 8 and 891 DF,  p-value: < 2.2e-16
```

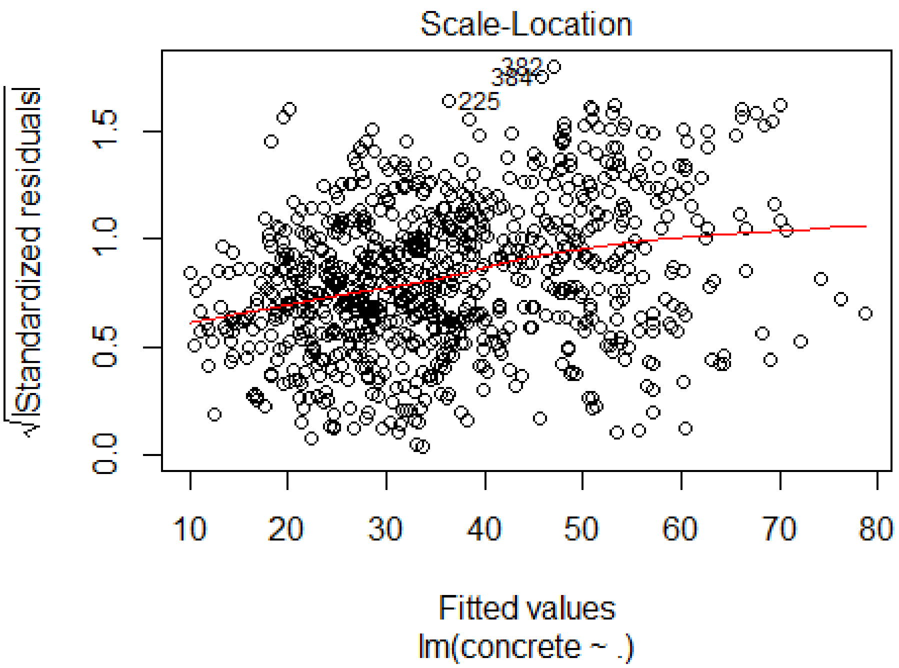
There is a relationship between the response and predicator variables with a F-statistic of 176.2 and a p-value of < 2.2e-16. Most of the predicator variables are statistically significant save two, c_aggregate and f_aggregate.Overall, this model accounts for 60.92% of concrete's variance.

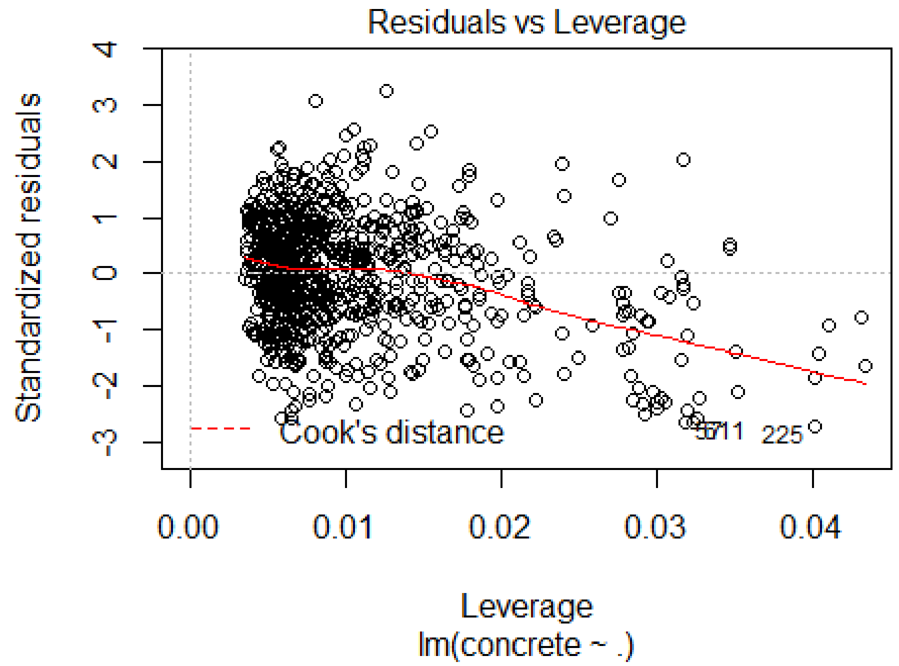
Diagnosing the model.

```
plot(lm_fit)
```


Residuals vs Fitted
Residuals
40
20
0
-20
382
384
225
10
20
30
40
50
60
70
80
Fitted values
lm(concrete ~ .)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
384
382
225
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(concrete ~ .)





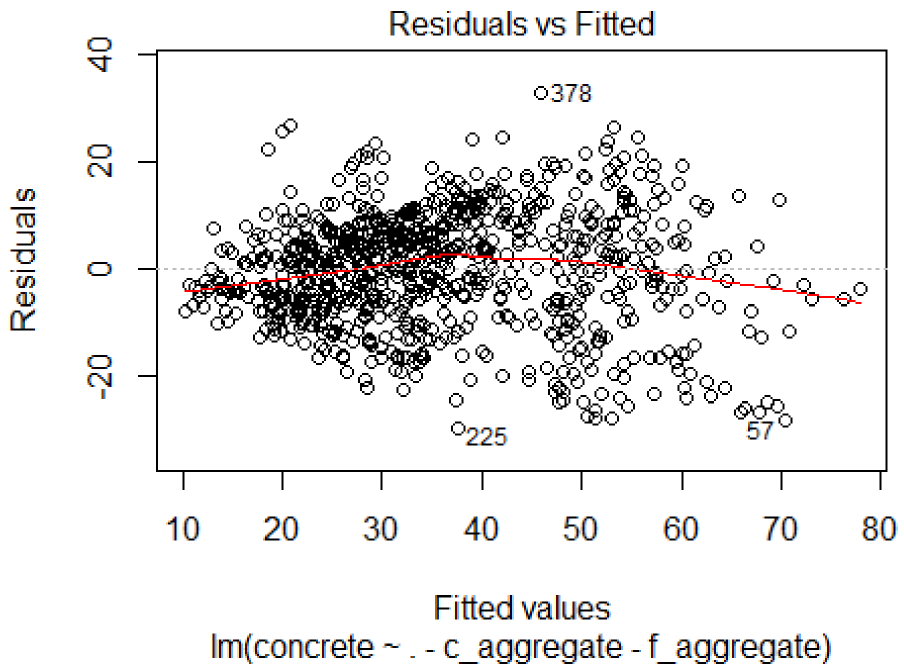
The Residuals vs Fitted plot shows elements of heteroscedasticity. The plot also shows 382, 384 and 225 are outliers. The Residuals vs Leverage shows 225 ,711, and 55 have high leverages.

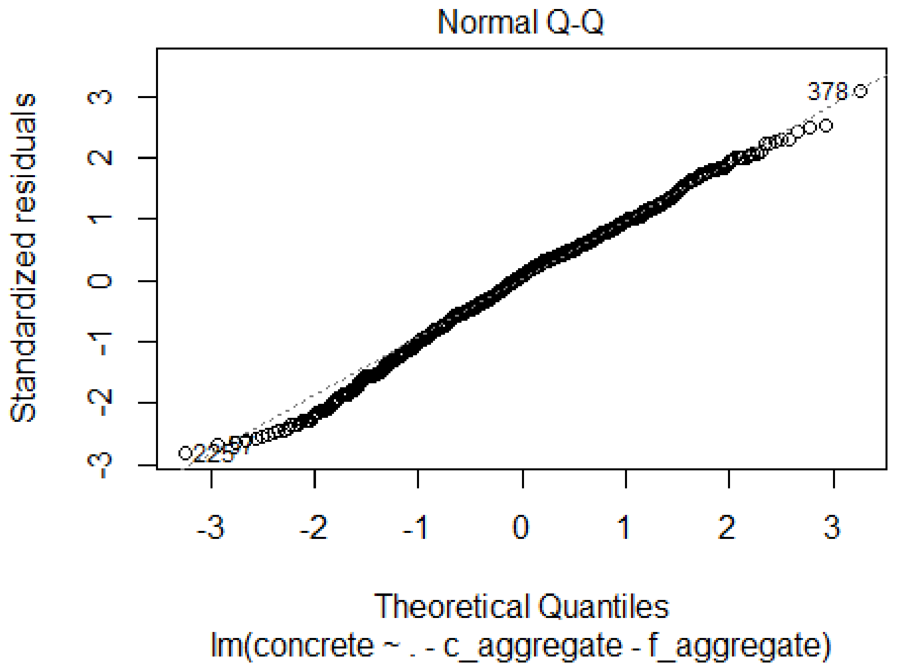
Removing outliers, variables with high leverage, and also log transforming the response variable.

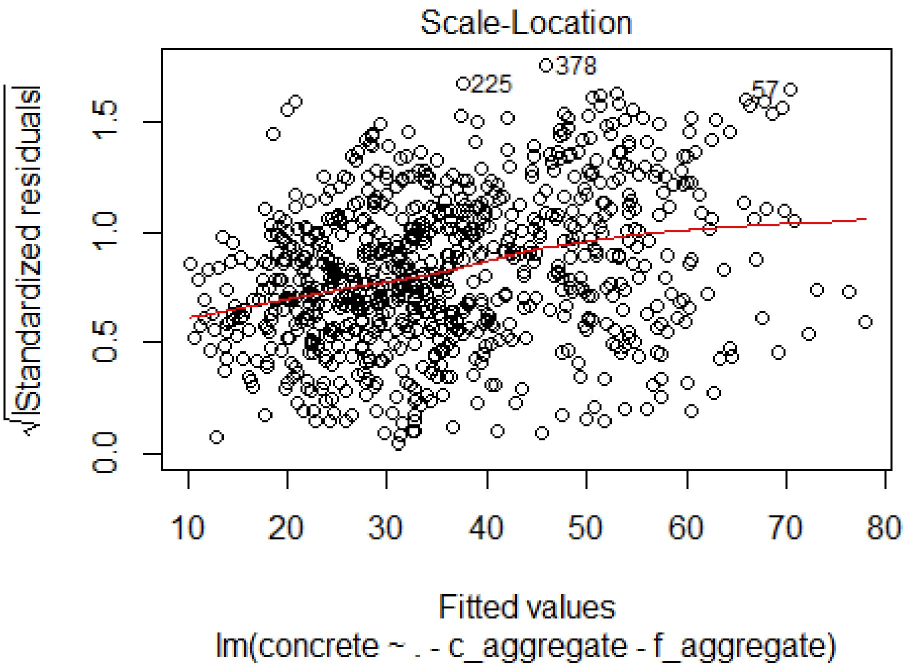
```
train <- filter(train, !cement %in% c(315,516,212.52,305.3,189.6))
lm_fit2 <- lm(concrete~.-c_aggregate-f_aggregate, train)
summary(lm_fit2)
```

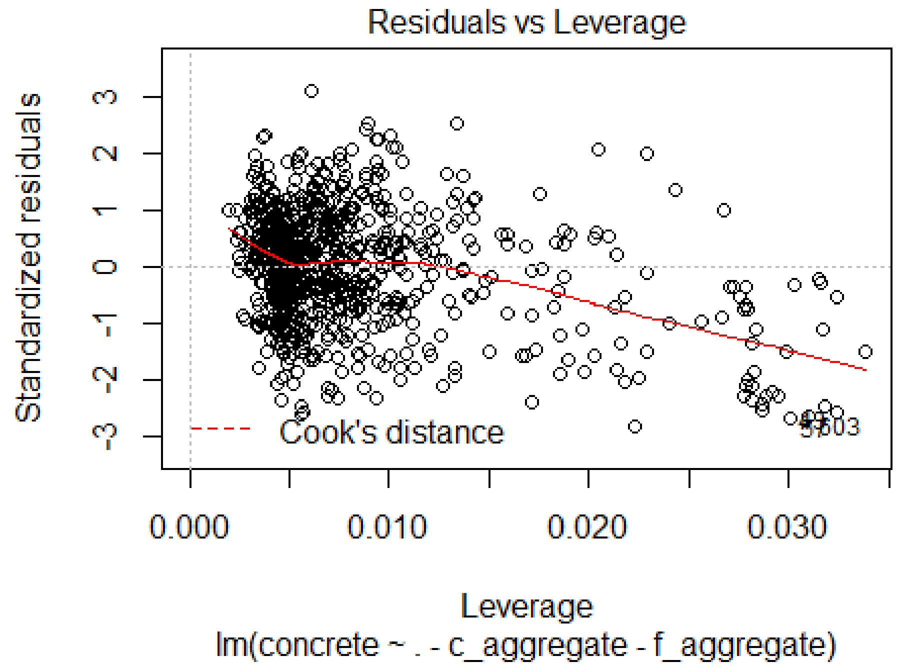
```
##
## Call:
## lm(formula = concrete ~ . - c_aggregate - f_aggregate, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.771  -6.622   0.981   7.030  32.924
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.818373   4.952071   5.618  2.6e-08 ***
## cement       0.103219   0.004679  22.061  < 2e-16 ***
## b_furnace    0.084665   0.005572  15.195  < 2e-16 ***
## fly_ash      0.076602   0.008819   8.686  < 2e-16 ***
## water       -0.210730   0.025044  -8.414  < 2e-16 ***
## superpl      0.274275   0.095609   2.869  0.00422 **
## age          0.113958   0.005635  20.223  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.7 on 881 degrees of freedom
## Multiple R-squared:  0.6155, Adjusted R-squared:  0.6129
## F-statistic: 235.1 on 6 and 881 DF,  p-value: < 2.2e-16
```

```
plot(lm_fit2)
```


Residuals vs Fitted
Residuals
40
20
0
-20
378
225
57
10
20
30
40
50
60
70
80
Fitted values
lm(concrete ~ . - c_aggregate - f_aggregate)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
378
225
57
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(concrete ~ . - c_aggregate - f_aggregate)

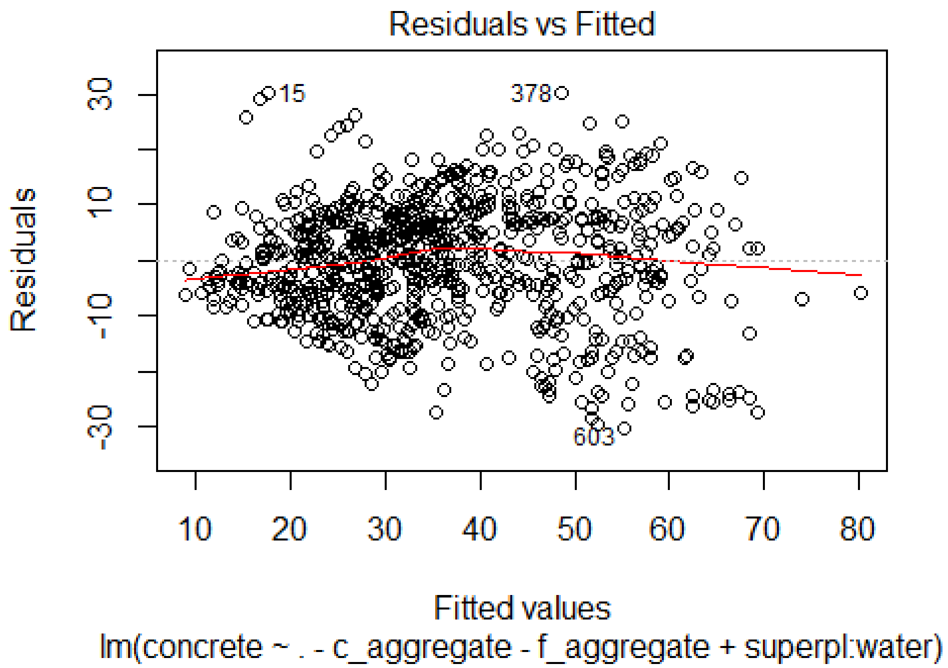


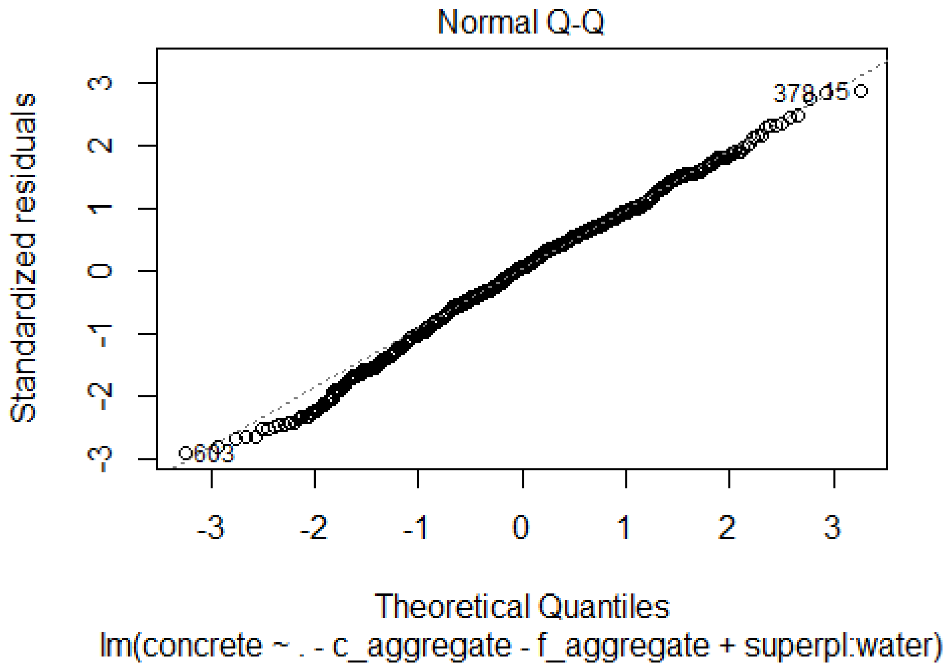


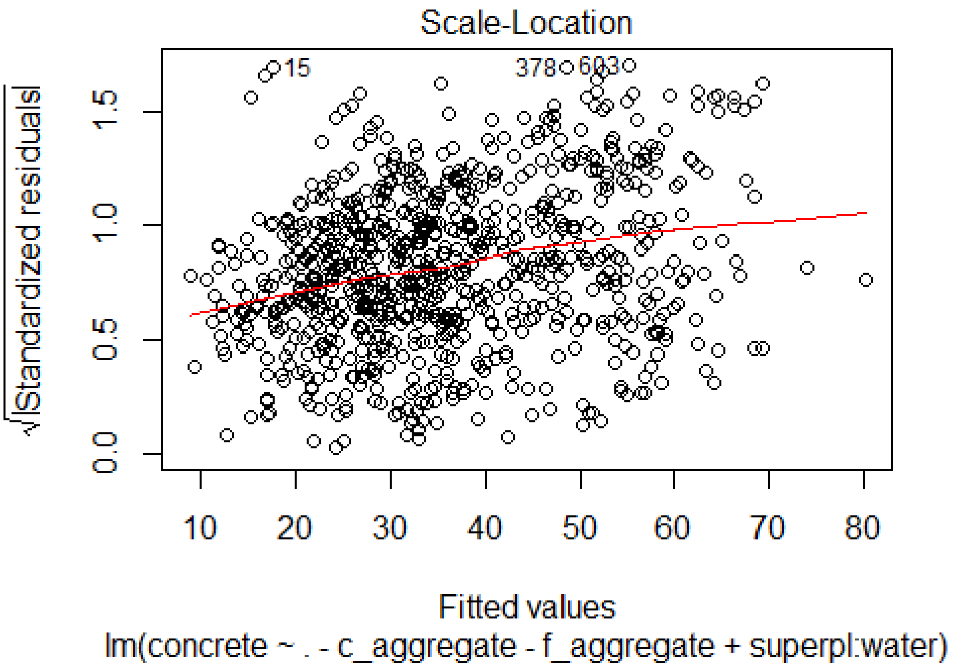
After removing outliers the model R-squared increased from 0.6092 to 0.6129, and the RSE reduces from 10.76 to 10.7.

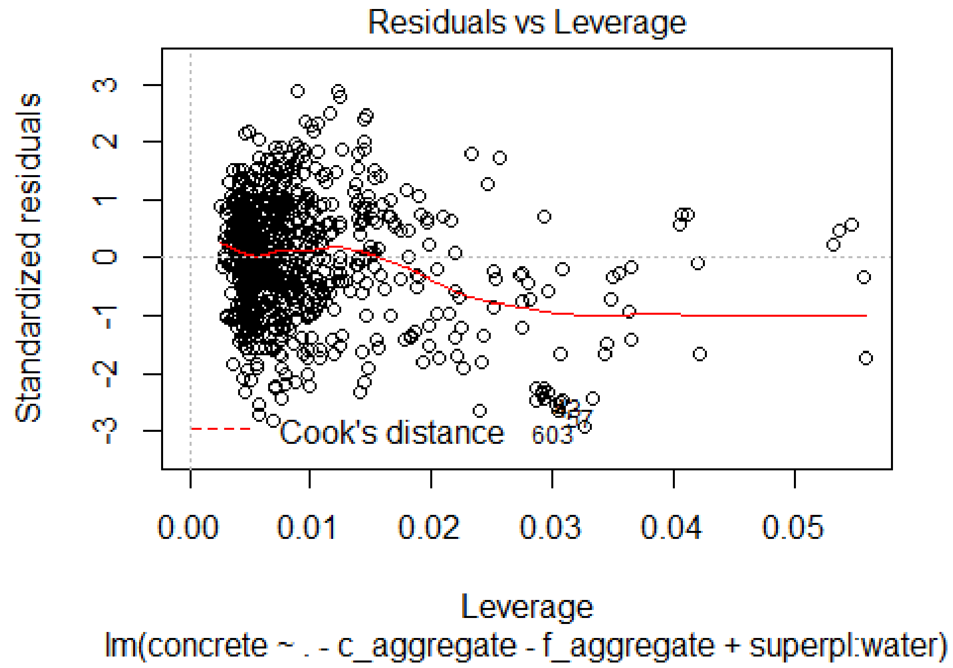
Fitting interactions into the model.

```
lm_fit3 <- lm(concrete~.-c_aggregate -f_aggregate + superpl : water    ,
train)
summary(lm_fit3)

##
## Call:
## lm(formula = concrete ~ . - c_aggregate - f_aggregate + superpl:water,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.2176  -6.3321   0.5695   7.1764  30.1684
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   42.554198   5.677898   7.495 1.62e-13 ***
## cement         0.100766   0.004639  21.721  < 2e-16 ***
## b_furnace      0.079259   0.005597  14.162  < 2e-16 ***
## fly_ash        0.053509   0.009810   5.455 6.38e-08 ***
## water         -0.284806   0.028670  -9.934  < 2e-16 ***
## superpl       -2.372928   0.528795  -4.487 8.17e-06 ***
## age            0.120365   0.005698  21.123  < 2e-16 ***
## water:superpl  0.016709   0.003284   5.088 4.43e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.55 on 880 degrees of freedom
## Multiple R-squared:  0.6265, Adjusted R-squared:  0.6236
## F-statistic: 210.9 on 7 and 880 DF,  p-value: < 2.2e-16

plot(lm_fit3)
```


Residuals vs Fitted
Residuals
30
10
-10
-30
15
378
603
10
20
30
40
50
60
70
80
Fitted values
lm(concrete ~ . - c_aggregate - f_aggregate + superpl:water)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
378
15
603
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(concrete ~ . - c_aggregate - f_aggregate + superpl:water)





After fitting the interaction effect into the dataset. The model accounts for 62.36 % of the variance in concrete compressive strength.

Testing the model and compute the R^2 of the predicted values.

```r
predicted_values <- predict(lm_fit3, test)
predicted_values <- data.frame(predicted_values)
predicted_values <- cbind(predicted_values, test$concrete)

SSE <- sum((predicted_values$`test$concrete` -
predicted_values$predicted_values   ) ^ 2)
SST  <- sum((predicted_values$`test$concrete` -
mean(predicted_values$`test$concrete`)) ^ 2)


print(1 - SSE/ SST)
```

```
## [1] 0.5640707
```