Section S2: Multiple Regression Analysis Supplemental 141A Final Project

Zheyuan

12/14/2020

```
# --- Data processing and viz ---
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

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

library(broom)
library(gridExtra)

##
## Attaching package: 'gridExtra'

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

library(RColorBrewer)
# --- Stats---
library(corrplot)

## corrplot 0.84 loaded

library(boot)
library(mclust)

## Package 'mclust' version 5.4.7
## Type 'citation("mclust")' for citing this R package in publications.

##
## Attaching package: 'mclust'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
library(PCAtools)
```

```
## Loading required package: ggrepel
```

```
##
## Attaching package: 'PCAtools'
```

```
## The following objects are masked from 'package:stats':
##
##     biplot, screeplot
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'
```

```
## The following object is masked from 'package:boot':
##
##     melanoma
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:boot':
##
##     aml
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(caret)
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:survival':
##
##     cluster
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
# --- Spatial Analysis ---> Let's simplify our life haha
library(tmap)
library(leaflet)
#library(sp)
library(sf)
```

```
## Linking to GEOS 3.8.0, GDAL 3.0.4, PROJ 6.3.1
```

— Step 0: Packages to mess with –

```
    if (!requireNamespace('BiocManager', quietly = TRUE))
        install.packages('BiocManager')

    BiocManager::install('PCAtools')
```

— Step 1: Data loading and procressing —

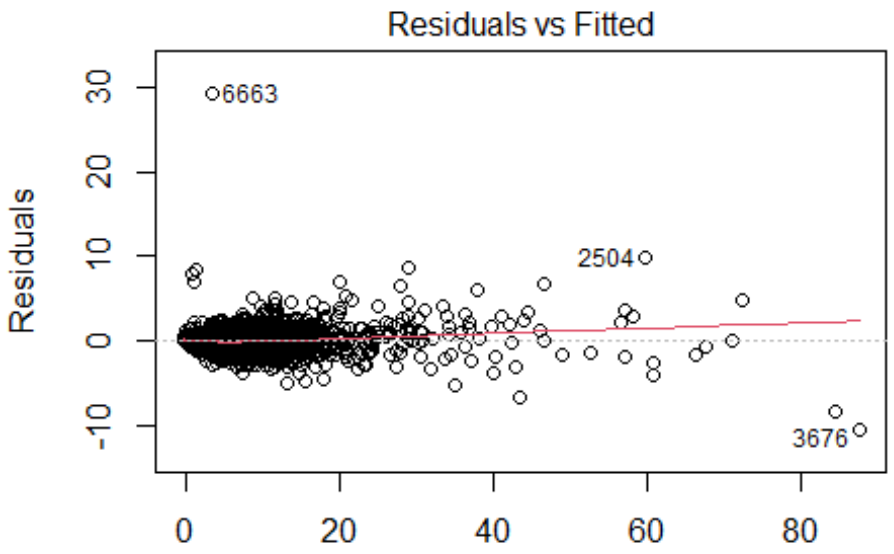
```
## --- Part a: Upload Metadata for samples ---
#path_data<-file.path(getwd(),"data")
path_data = "C:/Users/yzy/OneDrive/Documents/stats141A-FinalProject/data"
META_DATA<-as_tibble(read.csv(file.path(path_data,"IMPROVE_metadata.csv")))
## --- Filter samples from Korea and Canada ---
US_META<-META_DATA %>% filter(Country %nin% c("KR","CA"))



## --- Filter stats not in continental US ---
US_META<-META_DATA %>% filter(State %nin% c("HI","AK","VI"))

## -- Use Mississippi River as a dividing point for WEst-East US --
MR_coords<-c(47.239722, -95.2075)
POS_Sampler<-as.numeric(US_META$Longitude <MR_coords[2])
# --- 1 are WEst US, 0 are East
US_META<-add_column(US_META,WE_US = POS_Sampler)

## --- Part b: Load samples data ---
```

```
DATA<-
as_tibble(read.csv(file.path(path_data,"IMPROVE_2015_data_w_UNC_v2.csv")))

## --- Part c: Select samples from SW given site identifiers from SW_META
table ("Code")
US_DATA_all<-as_tibble(DATA %>% filter(SiteCode %in% US_META$Code))

# Let's identify any samples that (grossly) violate PM2.5 mass balances
# PM2.5 (=Y) cannot be negative!
# Since there's some probability that PM2.5 is negative due to errors at low
concentration, we may use PM2.5 uncertainties to remove samples that fall
outside -3*PM2.5_UNC.
# In this way, we don't risk censoring the data but do remove likely
erroneous data.
US_DATA_all<-US_DATA_all %>% dplyr::filter(PM2.5 > -3*PM2.5_UNC)

exclude<-
c("PM10","POC","ammNO3","ammSO4","SOIL","SeaSalt","OC1","OC2","OC3","OC4","EC
1","EC2","EC3","fAbs_MDL","fAbs")
US_DATA_LRG<- US_DATA_all %>% dplyr::select(!contains(exclude) &
!matches("_UNC") | matches("PM2.5_UNC"))
any(is.na(US_DATA_LRG))

## [1] TRUE

US_DATA_LRG<-US_DATA_LRG[which(complete.cases(US_DATA_LRG)),]
any(is.na(US_DATA_LRG))

## [1] FALSE

## --- Instead of random partitioning, I will partition by first sorting
samples by SiteCode and DATE (already done) and place every other sample in
the test set.
# --- This data has seasonality. Sorting by date therefore ensures
seasonality is equivalent between datasets
n<-nrow(US_DATA_LRG)
ind_test<-seq(1,n,2)
US_DATA_LRG_test<-US_DATA_LRG[ind_test,]
US_DATA_LRG<-US_DATA_LRG[-ind_test,]
```

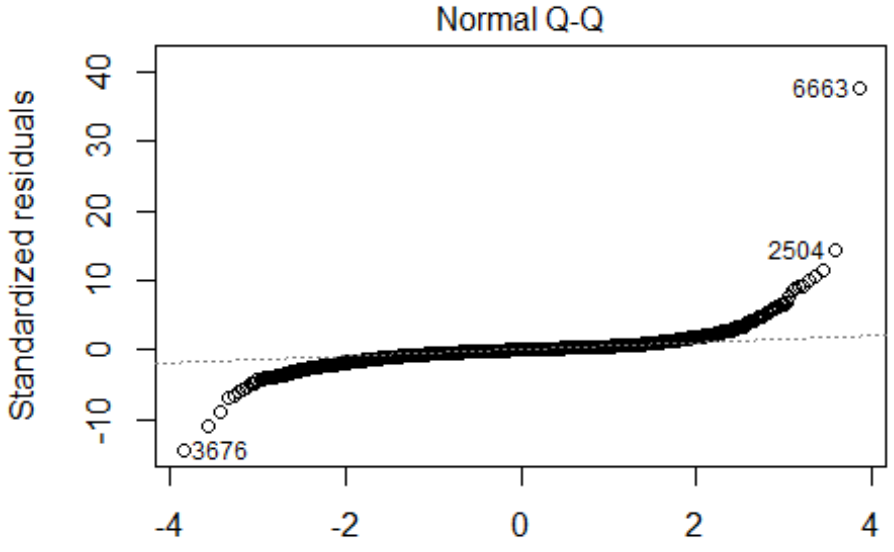
#Rgression Analysis

```
#First order model
fit = lm(PM2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL + CR + CU + FE + PB +
MG + MN + NI + N2 + P + K + RB + SE + SI + NA. + SR + S + TI + V + ZN + ZR +
NO3 + SO4, data = US_DATA_LRG)
summary(fit)

##
## Call:
## lm(formula = PM2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL +
##     CR + CU + FE + PB + MG + MN + NI + N2 + P + K + RB + SE +
```

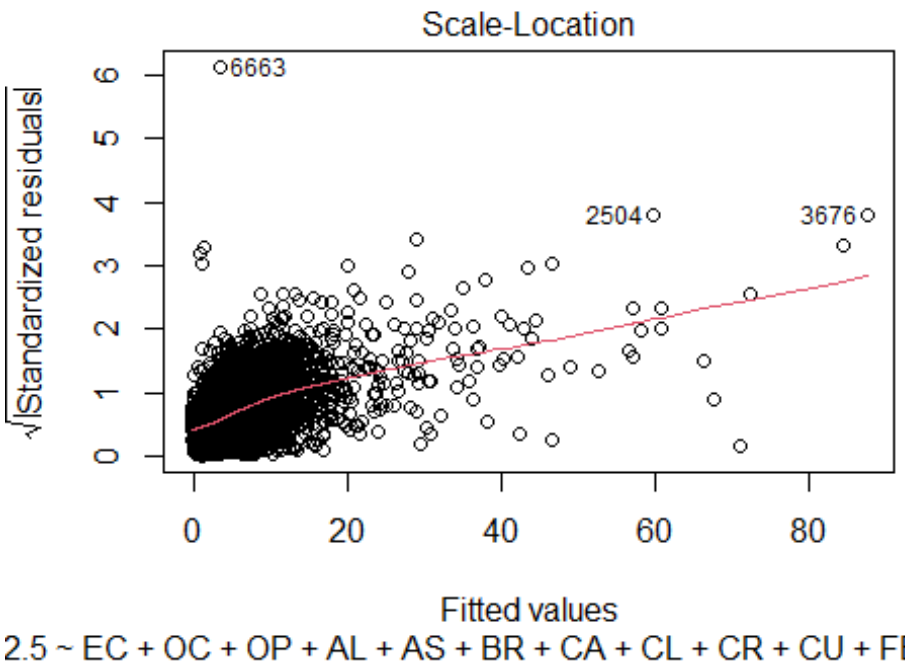
```
##     SI + NA. + SR + S + TI + V + ZN + ZR + NO3 + SO4, data = US_DATA_LRG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3934  -0.2615   0.0169   0.2508  29.1786
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.21256    0.01491 -14.256  < 2e-16 ***
## EC            -0.11101    0.08255  -1.345  0.17874
## OC             1.93476    0.01933 100.074  < 2e-16 ***
## OP             0.22448    0.05640   3.980 6.94e-05 ***
## AL            -0.58137    0.49521  -1.174  0.24043
## AS            16.62336   16.38978   1.014  0.31049
## BR             5.48740    6.90699   0.794  0.42694
## CA             1.91323    0.25723   7.438 1.12e-13 ***
## CL             3.45763    0.10375  33.325  < 2e-16 ***
## CR          -148.03756   58.86488  -2.515  0.01193 *
## CU           -26.39870    5.83226  -4.526 6.08e-06 ***
## FE             3.65516    0.75521   4.840 1.32e-06 ***
## PB            24.95061    5.76302   4.329 1.51e-05 ***
## MG            -0.03643    0.76443  -0.048  0.96199
## MN           -20.57501   10.10126  -2.037  0.04169 *
## NI            49.78920   78.64428   0.633  0.52669
## N2             0.04225    0.33241   0.127  0.89886
## P             44.97508    9.19560   4.891 1.02e-06 ***
## K              2.96531    0.29044  10.210  < 2e-16 ***
## RB            62.93135   42.41997   1.484  0.13797
## SE           144.02218   36.45795   3.950 7.87e-05 ***
## SI             2.99262    0.26512  11.288  < 2e-16 ***
## NA.            0.11404    0.17508   0.651  0.51484
## SR           -14.43224    5.46132  -2.643  0.00824 **
## S              3.92478    0.16637  23.591  < 2e-16 ***
## TI            17.30420    5.30792   3.260  0.00112 **
## V             25.86340   21.44322   1.206  0.22780
## ZN            -0.55866    1.58575  -0.352  0.72462
## ZR             9.27636   10.84616   0.855  0.39243
## NO3            1.22060    0.01208 101.025  < 2e-16 ***
## SO4            0.39605    0.05671   6.984 3.09e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7764 on 8616 degrees of freedom
## Multiple R-squared:  0.9763, Adjusted R-squared:  0.9762
## F-statistic: 1.182e+04 on 30 and 8616 DF,  p-value: < 2.2e-16

#Assumption check
plot(fit)
```



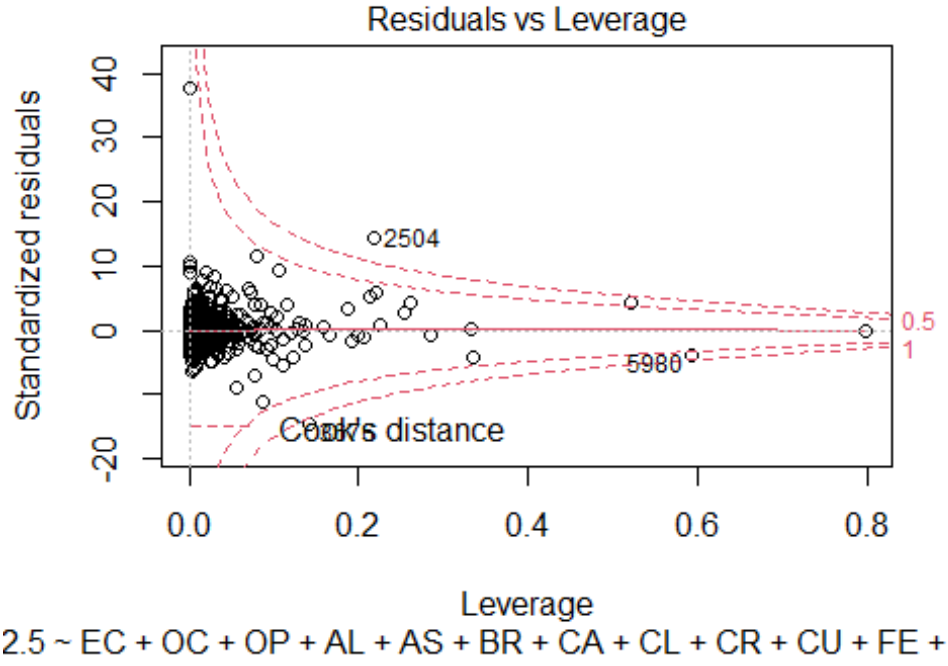
2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL + CR + CU + FE + P



2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL + CR + CU + FE + P



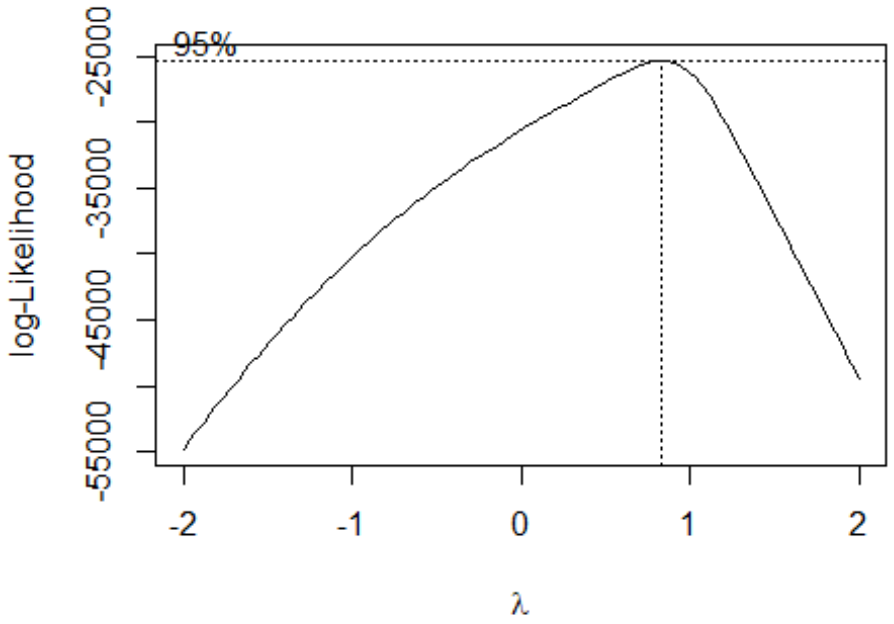
2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL + CR + CU + FE + P



2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL + CR + CU + FE + P

```
#Box Cox Procedure
min(US_DATA_LRG$PM2.5)

## [1] -0.093
```

```
fit.b = lm(PM2.5 + 0.26 ~ EC + OC + OP + AL + AS + BR + CA + CL + CR + CU +
FE + PB + MG + MN + NI + N2 + P + K + RB + SE + SI + NA. + SR + S + TI + V +
ZN + ZR + NO3 + SO4, data = US_DATA_LRG)
boxcox(fit.b)
```



#The QQ plot looks strange, but that just because there are several outliers. The lambda value in Box Cox procedure is very close to 1, which means we do not need to transform PM2.5 to make it more normal. The assumption of homoscedasticity and nonlinearity are valid, too.

```
#model selection
fit0 = lm(PM2.5 ~ 1, data = US_DATA_LRG)
#forward selection on AIC
mod1 = stepAIC(fit0, scope = list(upper = fit, lower = fit0), direction =
"forward", k = 2, trace = FALSE)
#backward elimination on AIC
mod2 = stepAIC(fit, scope = list(upper = fit, lower = fit0), direction =
"backward", k = 2, trace = FALSE)
#forward stepwise on AIC
mod3 = stepAIC(fit0, scope = list(upper = fit, lower = fit0), direction =
"both", k = 2, trace = FALSE)
#backward stepwise on AIC
mod4 = stepAIC(fit, scope = list(upper = fit, lower = fit0), direction =
"forward", k = 2, trace = FALSE)
#forward selection on BIC
mod5 = stepAIC(fit0, scope = list(upper = fit, lower = fit0), direction =
"forward", k = log(n), trace = FALSE)
#backward elimination on BIC
```

```r
mod6 = stepAIC(fit, scope = list(upper = fit, lower = fit0), direction =
"backward", k = log(n), trace = FALSE)
#forward stepwise on BIC
mod7 = stepAIC(fit0, scope = list(upper = fit, lower = fit0), direction =
"both", k = log(n), trace = FALSE)
#backward stepwise on BIC
mod8 = stepAIC(fit, scope = list(upper = fit, lower = fit0), direction =
"forward", k = log(n), trace = FALSE)
summary(mod1)

##
## Call:
## lm(formula = PM2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K +
##     CA + CU + PB + P + OP + TI + SE + V + CR + SR + MN + RB,
##     data = US_DATA_LRG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3521  -0.2623   0.0182   0.2524  29.1815
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.20458    0.01387 -14.754  < 2e-16 ***
## OC             1.92315    0.01473 130.522  < 2e-16 ***
## SO4            0.39926    0.05640   7.079 1.56e-12 ***
## FE             3.71932    0.69901   5.321 1.06e-07 ***
## NO3            1.22098    0.01175 103.883  < 2e-16 ***
## CL             3.53683    0.05907  59.880  < 2e-16 ***
## SI             2.76513    0.13729  20.141  < 2e-16 ***
## S              3.91824    0.16341  23.978  < 2e-16 ***
## K              2.95832    0.27518  10.750  < 2e-16 ***
## CA             2.02114    0.22873   8.836  < 2e-16 ***
## CU           -26.07097    5.64571  -4.618 3.93e-06 ***
## PB            26.11871    4.86269   5.371 8.02e-08 ***
## P             45.10306    9.17202   4.917 8.93e-07 ***
## OP             0.23829    0.05158   4.619 3.90e-06 ***
## TI            15.02221    4.34099   3.461 0.000542 ***
## SE           146.77176   36.11467   4.064 4.87e-05 ***
## V             38.32761   11.37525   3.369 0.000757 ***
## CR          -154.91486   53.38461  -2.902 0.003719 **
## SR           -15.10595    5.36982  -2.813 0.004917 **
## MN           -22.34431    9.61670  -2.323 0.020176 *
## RB            63.64680   42.02435   1.515 0.129930
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7763 on 8626 degrees of freedom
## Multiple R-squared:  0.9763, Adjusted R-squared:  0.9762
## F-statistic: 1.774e+04 on 20 and 8626 DF,  p-value: < 2.2e-16
```

```r
summary(mod2)
```

```
##
## Call:
## lm(formula = PM2.5 ~ OC + OP + CA + CL + CR + CU + FE + PB +
##     MN + P + K + RB + SE + SI + SR + S + TI + V + NO3 + SO4,
##     data = US_DATA_LRG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3521  -0.2623   0.0182   0.2524  29.1815
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.20458    0.01387 -14.754  < 2e-16 ***
## OC             1.92315    0.01473 130.522  < 2e-16 ***
## OP             0.23829    0.05158   4.619 3.90e-06 ***
## CA             2.02114    0.22873   8.836  < 2e-16 ***
## CL             3.53683    0.05907  59.880  < 2e-16 ***
## CR          -154.91486   53.38461  -2.902 0.003719 **
## CU           -26.07097    5.64571  -4.618 3.93e-06 ***
## FE             3.71932    0.69901   5.321 1.06e-07 ***
## PB            26.11871    4.86269   5.371 8.02e-08 ***
## MN           -22.34431    9.61670  -2.323 0.020176 *
## P             45.10306    9.17202   4.917 8.93e-07 ***
## K              2.95832    0.27518  10.750  < 2e-16 ***
## RB            63.64680   42.02435   1.515 0.129930
## SE           146.77176   36.11467   4.064 4.87e-05 ***
## SI             2.76513    0.13729  20.141  < 2e-16 ***
## SR           -15.10595    5.36982  -2.813 0.004917 **
## S              3.91824    0.16341  23.978  < 2e-16 ***
## TI            15.02221    4.34099   3.461 0.000542 ***
## V             38.32761   11.37525   3.369 0.000757 ***
## NO3            1.22098    0.01175 103.883  < 2e-16 ***
## SO4            0.39926    0.05640   7.079 1.56e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7763 on 8626 degrees of freedom
## Multiple R-squared:  0.9763, Adjusted R-squared:  0.9762
## F-statistic: 1.774e+04 on 20 and 8626 DF,  p-value: < 2.2e-16
```

```r
summary(mod3)
```

```
##
## Call:
## lm(formula = PM2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K +
##     CA + CU + PB + P + OP + TI + SE + V + CR + SR + MN + RB,
##     data = US_DATA_LRG)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3521  -0.2623   0.0182   0.2524  29.1815
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.20458    0.01387 -14.754  < 2e-16 ***
## OC             1.92315    0.01473 130.522  < 2e-16 ***
## SO4            0.39926    0.05640   7.079 1.56e-12 ***
## FE             3.71932    0.69901   5.321 1.06e-07 ***
## NO3            1.22098    0.01175 103.883  < 2e-16 ***
## CL             3.53683    0.05907  59.880  < 2e-16 ***
## SI             2.76513    0.13729  20.141  < 2e-16 ***
## S              3.91824    0.16341  23.978  < 2e-16 ***
## K              2.95832    0.27518  10.750  < 2e-16 ***
## CA             2.02114    0.22873   8.836  < 2e-16 ***
## CU           -26.07097    5.64571  -4.618 3.93e-06 ***
## PB            26.11871    4.86269   5.371 8.02e-08 ***
## P             45.10306    9.17202   4.917 8.93e-07 ***
## OP             0.23829    0.05158   4.619 3.90e-06 ***
## TI            15.02221    4.34099   3.461 0.000542 ***
## SE           146.77176   36.11467   4.064 4.87e-05 ***
## V             38.32761   11.37525   3.369 0.000757 ***
## CR          -154.91486   53.38461  -2.902 0.003719 **
## SR           -15.10595    5.36982  -2.813 0.004917 **
## MN           -22.34431    9.61670  -2.323 0.020176 *
## RB            63.64680   42.02435   1.515 0.129930
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7763 on 8626 degrees of freedom
## Multiple R-squared:  0.9763, Adjusted R-squared:  0.9762
## F-statistic: 1.774e+04 on 20 and 8626 DF,  p-value: < 2.2e-16
```

```
summary(mod4)
```

```
##
## Call:
## lm(formula = PM2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL +
##     CR + CU + FE + PB + MG + MN + NI + N2 + P + K + RB + SE +
##     SI + NA. + SR + S + TI + V + ZN + ZR + NO3 + SO4, data = US_DATA_LRG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3934  -0.2615   0.0169   0.2508  29.1786
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.21256    0.01491 -14.256  < 2e-16 ***
## EC            -0.11101    0.08255  -1.345  0.17874
```

```
## OC             1.93476    0.01933 100.074  < 2e-16 ***
## OP             0.22448    0.05640   3.980 6.94e-05 ***
## AL            -0.58137    0.49521  -1.174  0.24043
## AS            16.62336   16.38978   1.014  0.31049
## BR             5.48740    6.90699   0.794  0.42694
## CA             1.91323    0.25723   7.438 1.12e-13 ***
## CL             3.45763    0.10375  33.325  < 2e-16 ***
## CR          -148.03756   58.86488  -2.515  0.01193 *
## CU           -26.39870    5.83226  -4.526 6.08e-06 ***
## FE             3.65516    0.75521   4.840 1.32e-06 ***
## PB            24.95061    5.76302   4.329 1.51e-05 ***
## MG            -0.03643    0.76443  -0.048  0.96199
## MN           -20.57501   10.10126  -2.037  0.04169 *
## NI            49.78920   78.64428   0.633  0.52669
## N2             0.04225    0.33241   0.127  0.89886
## P             44.97508    9.19560   4.891 1.02e-06 ***
## K              2.96531    0.29044  10.210  < 2e-16 ***
## RB            62.93135   42.41997   1.484  0.13797
## SE           144.02218   36.45795   3.950 7.87e-05 ***
## SI             2.99262    0.26512  11.288  < 2e-16 ***
## NA.            0.11404    0.17508   0.651  0.51484
## SR           -14.43224    5.46132  -2.643  0.00824 **
## S              3.92478    0.16637  23.591  < 2e-16 ***
## TI            17.30420    5.30792   3.260  0.00112 **
## V             25.86340   21.44322   1.206  0.22780
## ZN            -0.55866    1.58575  -0.352  0.72462
## ZR             9.27636   10.84616   0.855  0.39243
## NO3            1.22060    0.01208 101.025  < 2e-16 ***
## SO4            0.39605    0.05671   6.984 3.09e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7764 on 8616 degrees of freedom
## Multiple R-squared:  0.9763, Adjusted R-squared:  0.9762
## F-statistic: 1.182e+04 on 30 and 8616 DF,  p-value: < 2.2e-16
```

```
summary(mod5)
```

```
##
## Call:
## lm(formula = PM2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K +
##     CA + CU + PB + P + OP + TI + SE, data = US_DATA_LRG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2069  -0.2578   0.0215   0.2512  29.2244
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.20680    0.01373 -15.065  < 2e-16 ***
```

```
## OC            1.93015    0.01466 131.678  < 2e-16 ***
## SO4           0.42689    0.05485   7.783 7.92e-15 ***
## FE            2.26146    0.58252   3.882 0.000104 ***
## NO3           1.22387    0.01164 105.154  < 2e-16 ***
## CL            3.53975    0.05901  59.987  < 2e-16 ***
## SI            2.99026    0.13068  22.882  < 2e-16 ***
## S             3.86775    0.16145  23.957  < 2e-16 ***
## K             2.43625    0.22489  10.833  < 2e-16 ***
## CA            1.91137    0.22327   8.561  < 2e-16 ***
## CU          -29.52848    5.33513  -5.535 3.21e-08 ***
## PB           24.50111    4.41828   5.545 3.02e-08 ***
## P            47.15010    9.16384   5.145 2.73e-07 ***
## OP            0.22895    0.05157   4.439 9.14e-06 ***
## TI           18.45628    4.19983   4.395 1.12e-05 ***
## SE          141.35127   35.87357   3.940 8.20e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7777 on 8631 degrees of freedom
## Multiple R-squared:  0.9762, Adjusted R-squared:  0.9761
## F-statistic: 2.357e+04 on 15 and 8631 DF,  p-value: < 2.2e-16
```

```
summary(mod6)
```

```
##
## Call:
## lm(formula = PM2.5 ~ OC + OP + CA + CL + CR + CU + FE + PB +
##     P + K + SE + SI + S + TI + V + NO3 + SO4, data = US_DATA_LRG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2598  -0.2620   0.0178   0.2521  29.1899
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.20364    0.01382 -14.737  < 2e-16 ***
## OC             1.92904    0.01465 131.712  < 2e-16 ***
## OP             0.23141    0.05158   4.486 7.34e-06 ***
## CA             1.87043    0.22448   8.332  < 2e-16 ***
## CL             3.53383    0.05896  59.932  < 2e-16 ***
## CR          -170.07023   53.03200  -3.207 0.001346 **
## CU           -25.86852    5.46109  -4.737 2.20e-06 ***
## FE             3.00058    0.61958   4.843 1.30e-06 ***
## PB            25.30822    4.47082   5.661 1.56e-08 ***
## P             45.47777    9.16990   4.959 7.20e-07 ***
## K              2.56550    0.22996  11.156  < 2e-16 ***
## SE           141.90823   35.99421   3.943 8.13e-05 ***
## SI             2.89006    0.13273  21.774  < 2e-16 ***
## S              3.89311    0.16287  23.903  < 2e-16 ***
## TI            15.80104    4.25261   3.716 0.000204 ***
```

```
## V             37.01644   11.37785   3.253 0.001145 **
## NO3            1.22619    0.01165 105.220  < 2e-16 ***
## SO4            0.40972    0.05620   7.290 3.38e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7769 on 8629 degrees of freedom
## Multiple R-squared:  0.9762, Adjusted R-squared:  0.9762
## F-statistic: 2.084e+04 on 17 and 8629 DF,  p-value: < 2.2e-16
```

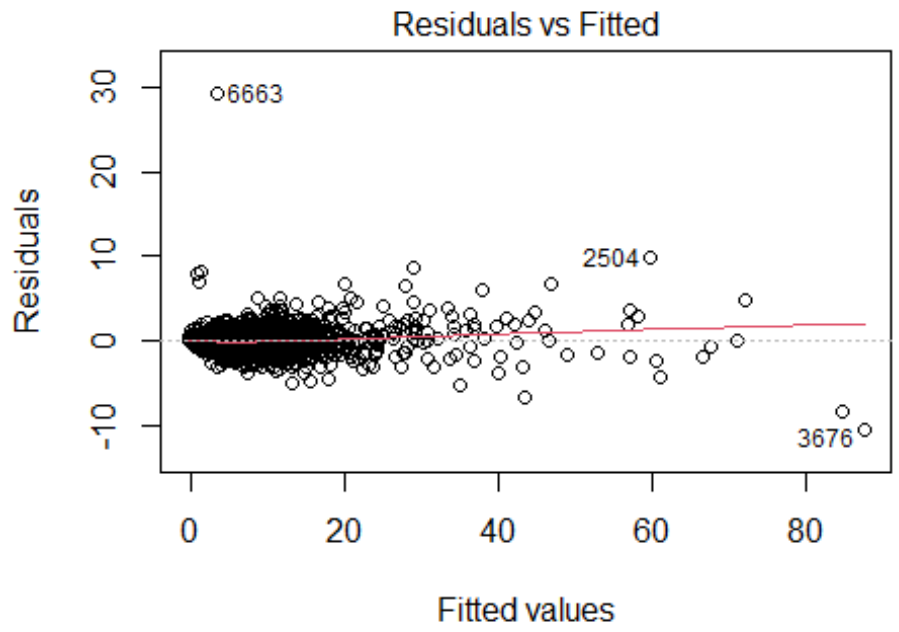
```
summary(mod7)
```

```
##
## Call:
## lm(formula = PM2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K +
##     CA + CU + PB + P + OP + TI + SE, data = US_DATA_LRG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2069  -0.2578   0.0215   0.2512  29.2244
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.20680    0.01373 -15.065  < 2e-16 ***
## OC            1.93015    0.01466 131.678  < 2e-16 ***
## SO4           0.42689    0.05485   7.783 7.92e-15 ***
## FE            2.26146    0.58252   3.882 0.000104 ***
## NO3           1.22387    0.01164 105.154  < 2e-16 ***
## CL            3.53975    0.05901  59.987  < 2e-16 ***
## SI            2.99026    0.13068  22.882  < 2e-16 ***
## S             3.86775    0.16145  23.957  < 2e-16 ***
## K             2.43625    0.22489  10.833  < 2e-16 ***
## CA            1.91137    0.22327   8.561  < 2e-16 ***
## CU          -29.52848    5.33513  -5.535 3.21e-08 ***
## PB           24.50111    4.41828   5.545 3.02e-08 ***
## P            47.15010    9.16384   5.145 2.73e-07 ***
## OP            0.22895    0.05157   4.439 9.14e-06 ***
## TI           18.45628    4.19983   4.395 1.12e-05 ***
## SE          141.35127   35.87357   3.940 8.20e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7777 on 8631 degrees of freedom
## Multiple R-squared:  0.9762, Adjusted R-squared:  0.9761
## F-statistic: 2.357e+04 on 15 and 8631 DF,  p-value: < 2.2e-16
```

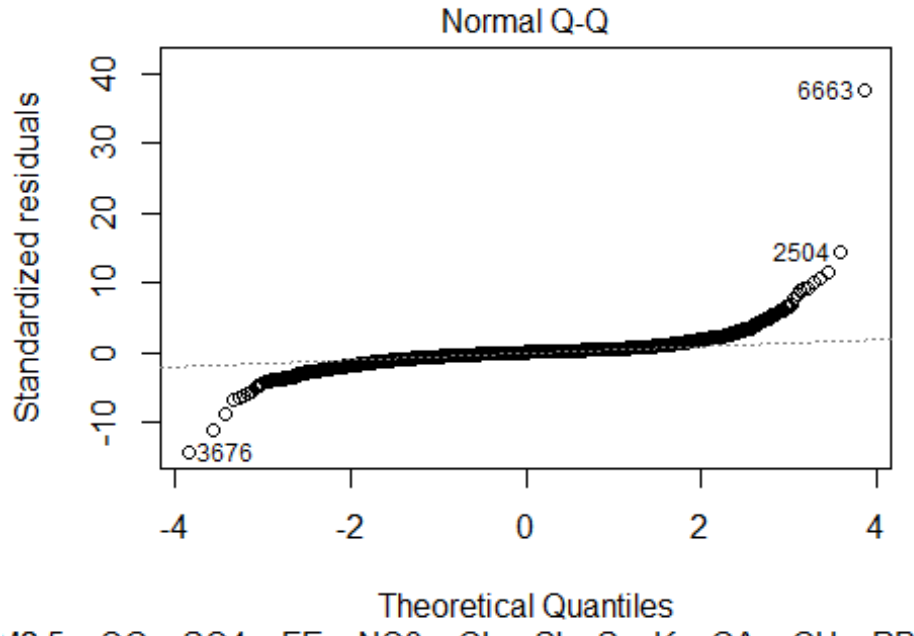
```
summary(mod8)
```

```
##
## Call:
## lm(formula = PM2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL +
```

```
##     CR + CU + FE + PB + MG + MN + NI + N2 + P + K + RB + SE +
##     SI + NA. + SR + S + TI + V + ZN + ZR + NO3 + SO4, data = US_DATA_LRG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3934  -0.2615   0.0169   0.2508  29.1786
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.21256    0.01491 -14.256  < 2e-16 ***
## EC            -0.11101    0.08255  -1.345  0.17874
## OC             1.93476    0.01933 100.074  < 2e-16 ***
## OP             0.22448    0.05640   3.980 6.94e-05 ***
## AL            -0.58137    0.49521  -1.174  0.24043
## AS            16.62336   16.38978   1.014  0.31049
## BR             5.48740    6.90699   0.794  0.42694
## CA             1.91323    0.25723   7.438 1.12e-13 ***
## CL             3.45763    0.10375  33.325  < 2e-16 ***
## CR          -148.03756   58.86488  -2.515  0.01193 *
## CU           -26.39870    5.83226  -4.526 6.08e-06 ***
## FE             3.65516    0.75521   4.840 1.32e-06 ***
## PB            24.95061    5.76302   4.329 1.51e-05 ***
## MG            -0.03643    0.76443  -0.048  0.96199
## MN           -20.57501   10.10126  -2.037  0.04169 *
## NI            49.78920   78.64428   0.633  0.52669
## N2             0.04225    0.33241   0.127  0.89886
## P             44.97508    9.19560   4.891 1.02e-06 ***
## K              2.96531    0.29044  10.210  < 2e-16 ***
## RB            62.93135   42.41997   1.484  0.13797
## SE           144.02218   36.45795   3.950 7.87e-05 ***
## SI             2.99262    0.26512  11.288  < 2e-16 ***
## NA.            0.11404    0.17508   0.651  0.51484
## SR           -14.43224    5.46132  -2.643  0.00824 **
## S              3.92478    0.16637  23.591  < 2e-16 ***
## TI            17.30420    5.30792   3.260  0.00112 **
## V             25.86340   21.44322   1.206  0.22780
## ZN            -0.55866    1.58575  -0.352  0.72462
## ZR             9.27636   10.84616   0.855  0.39243
## NO3            1.22060    0.01208 101.025  < 2e-16 ***
## SO4            0.39605    0.05671   6.984 3.09e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7764 on 8616 degrees of freedom
## Multiple R-squared:  0.9763, Adjusted R-squared:  0.9762
## F-statistic: 1.182e+04 on 30 and 8616 DF,  p-value: < 2.2e-16

plot(mod1, which = c(1,2))
```

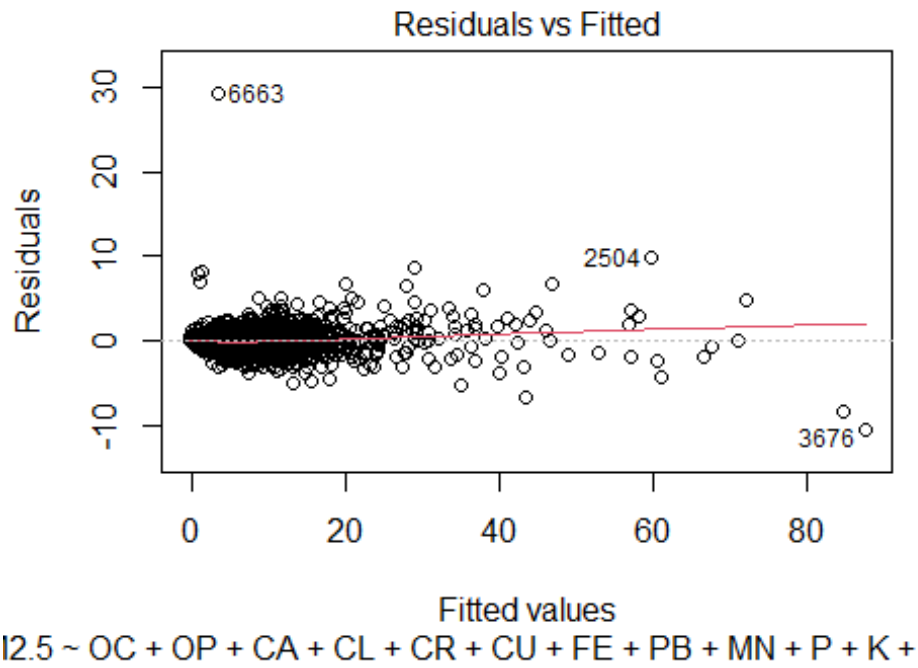


M2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K + CA + CU + PB + P

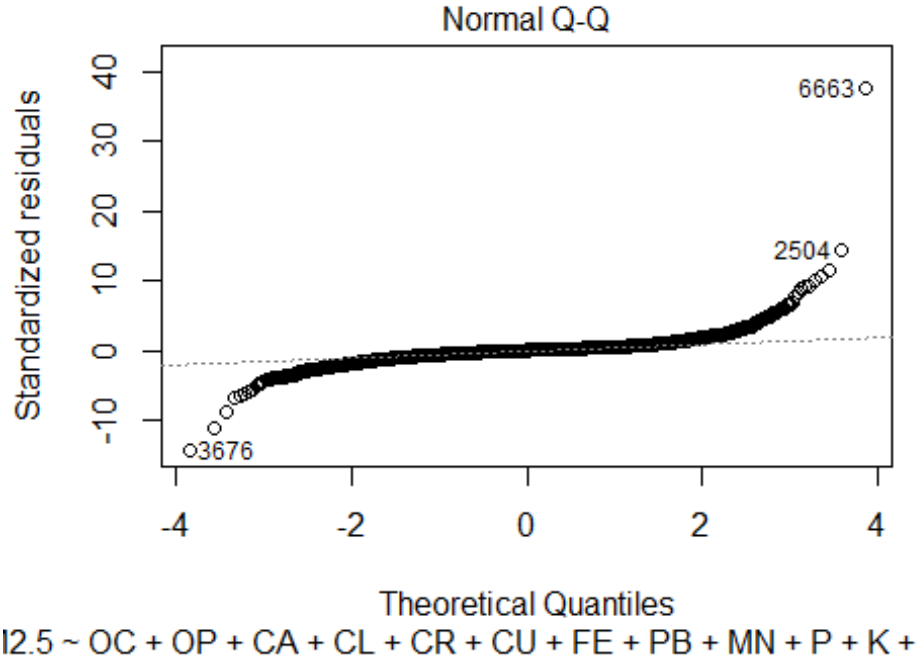


M2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K + CA + CU + PB + P

```
plot(mod2, which = c(1,2))
```

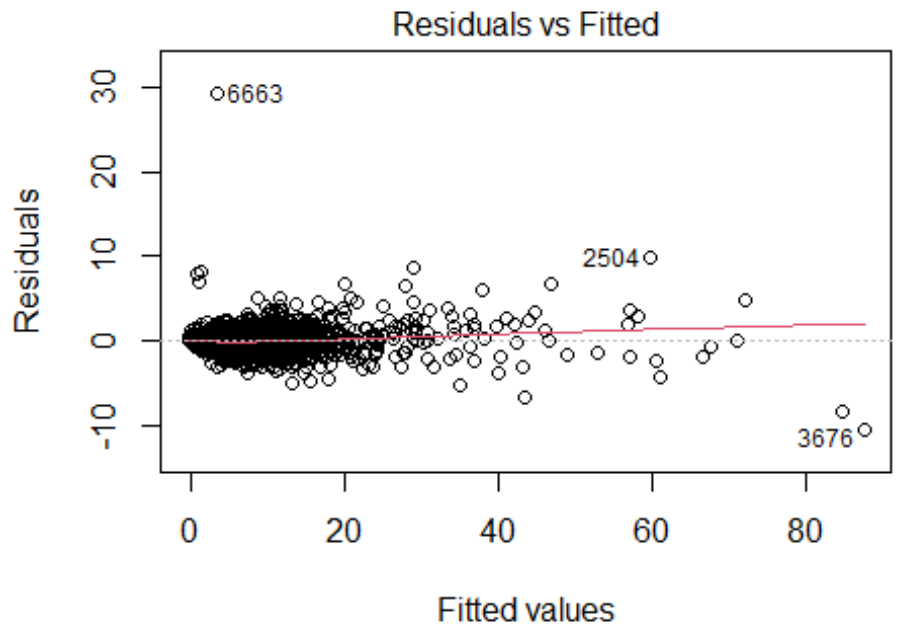


I2.5 ~ OC + OP + CA + CL + CR + CU + FE + PB + MN + P + K + RB

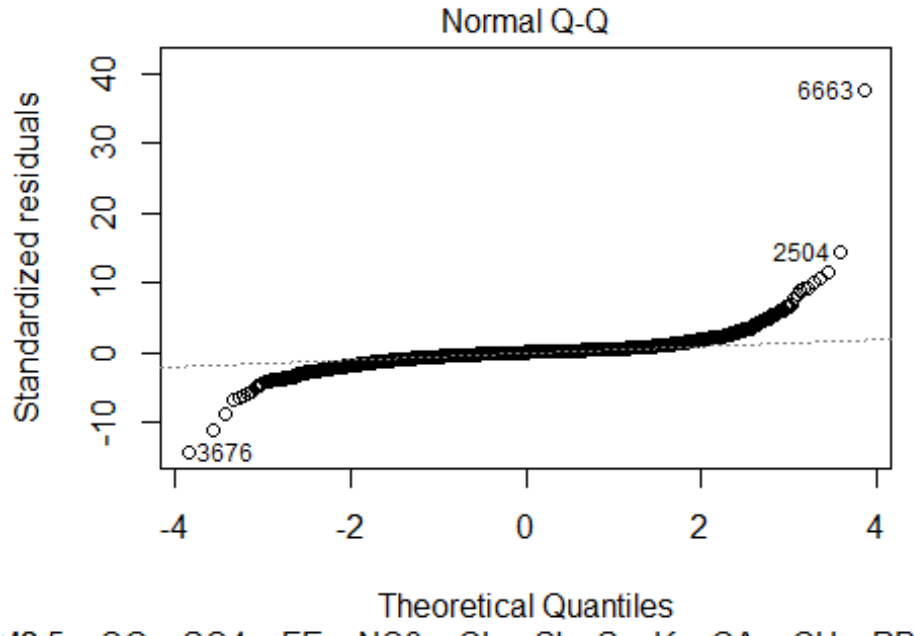


I2.5 ~ OC + OP + CA + CL + CR + CU + FE + PB + MN + P + K + RB

```
plot(mod3, which = c(1,2))
```

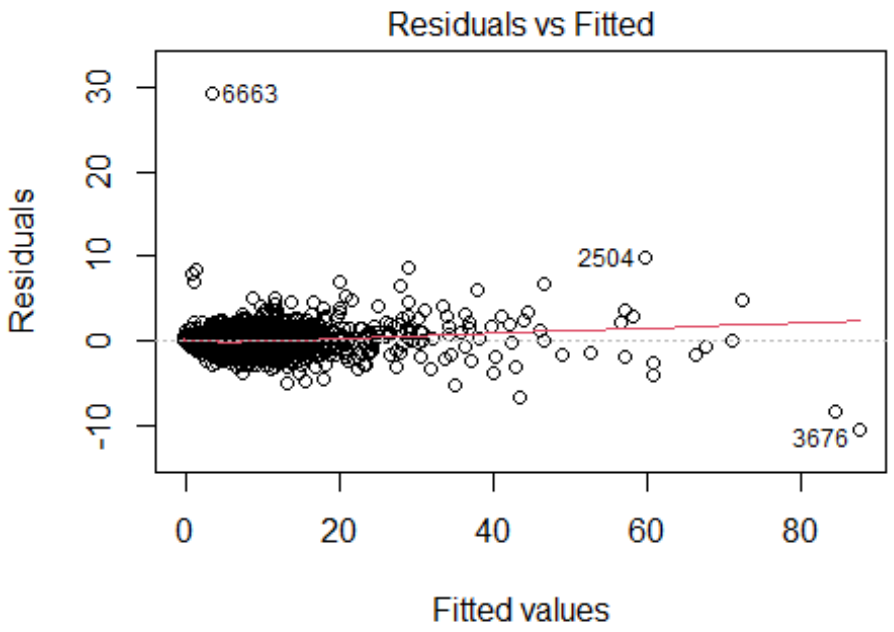


M2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K + CA + CU + PB + P

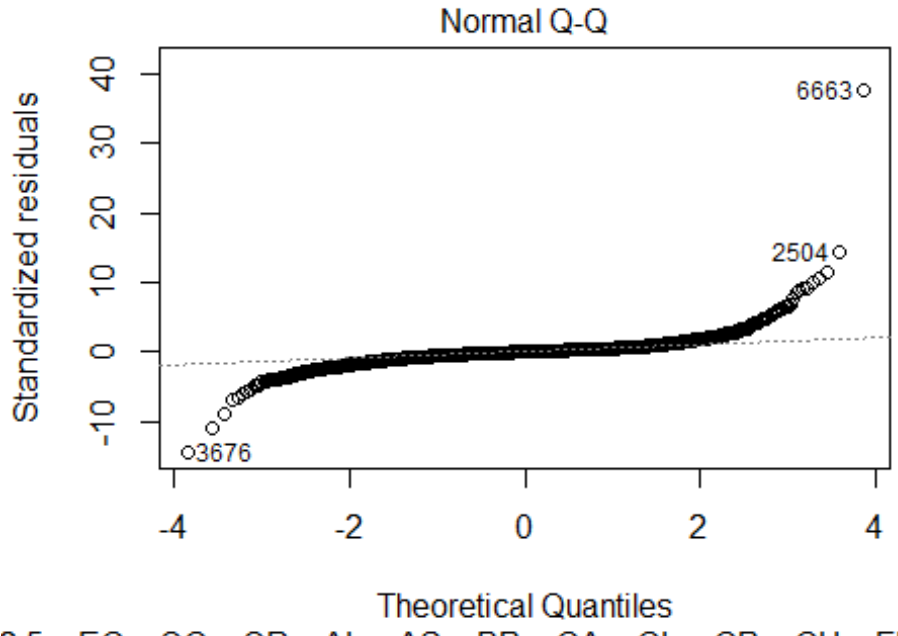


M2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K + CA + CU + PB + P

```
plot(mod4, which = c(1,2))
```

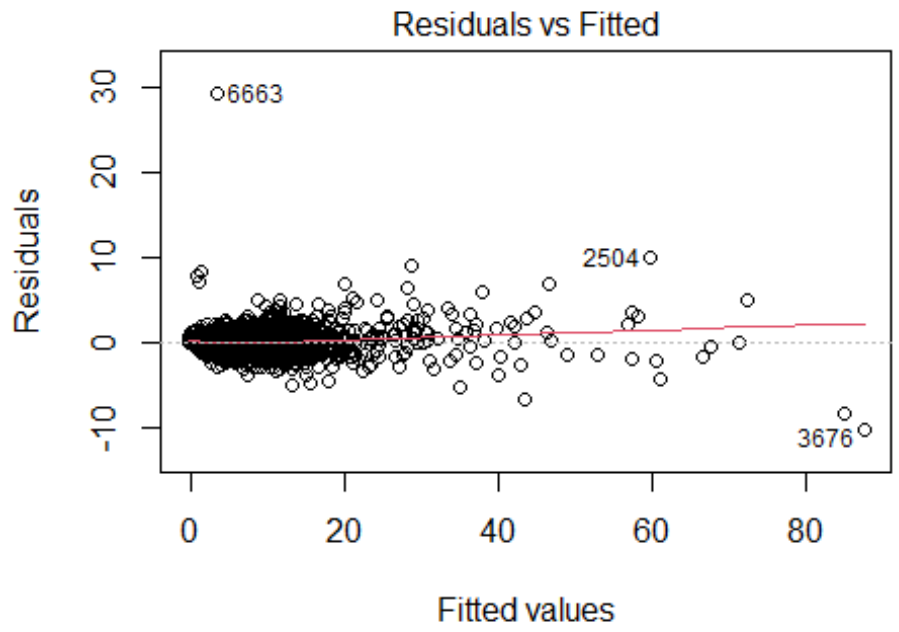


2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL + CR + CU + FE + P

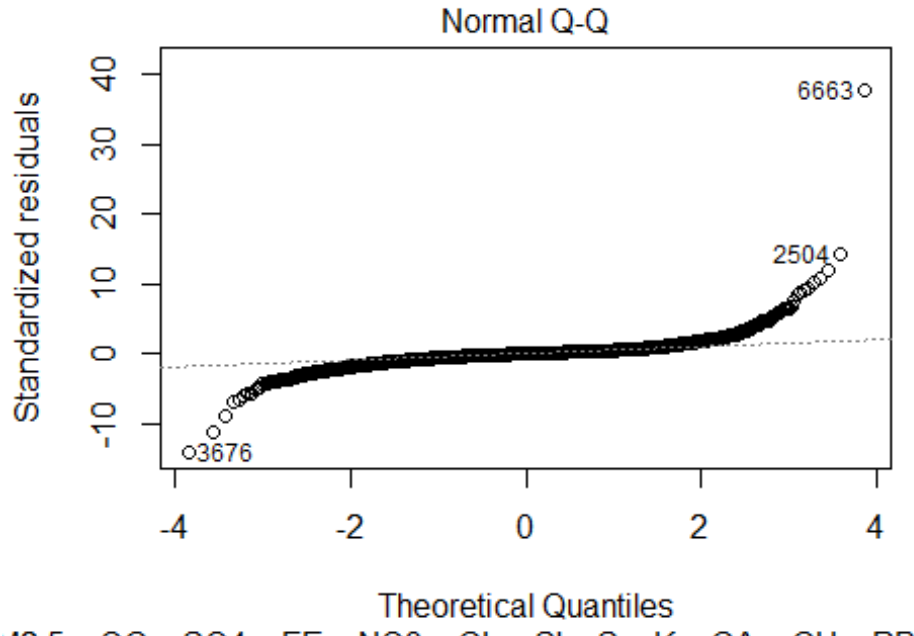


2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL + CR + CU + FE + P

```r
plot(mod5, which = c(1,2))
```

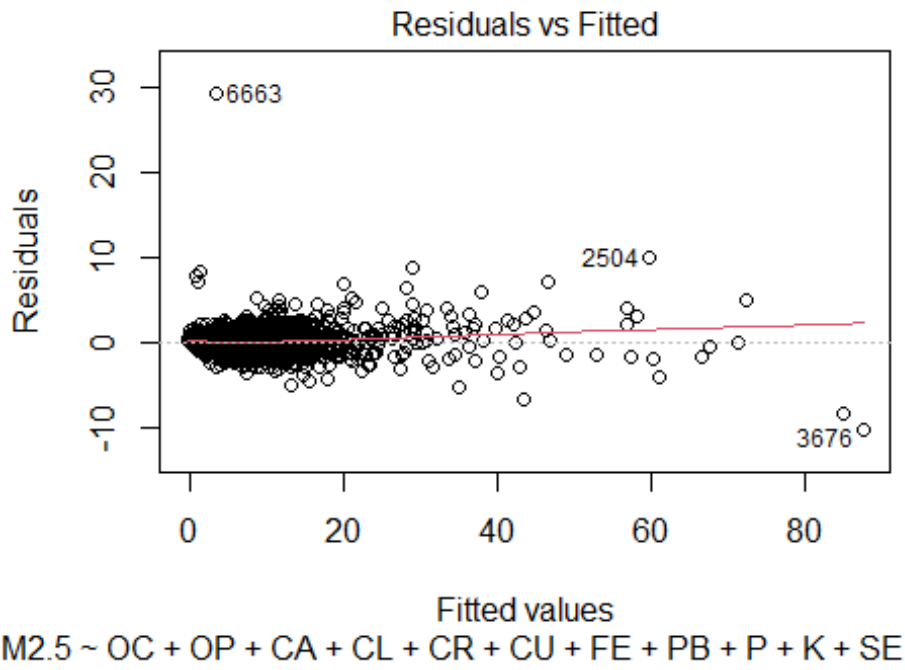


M2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K + CA + CU + PB + P

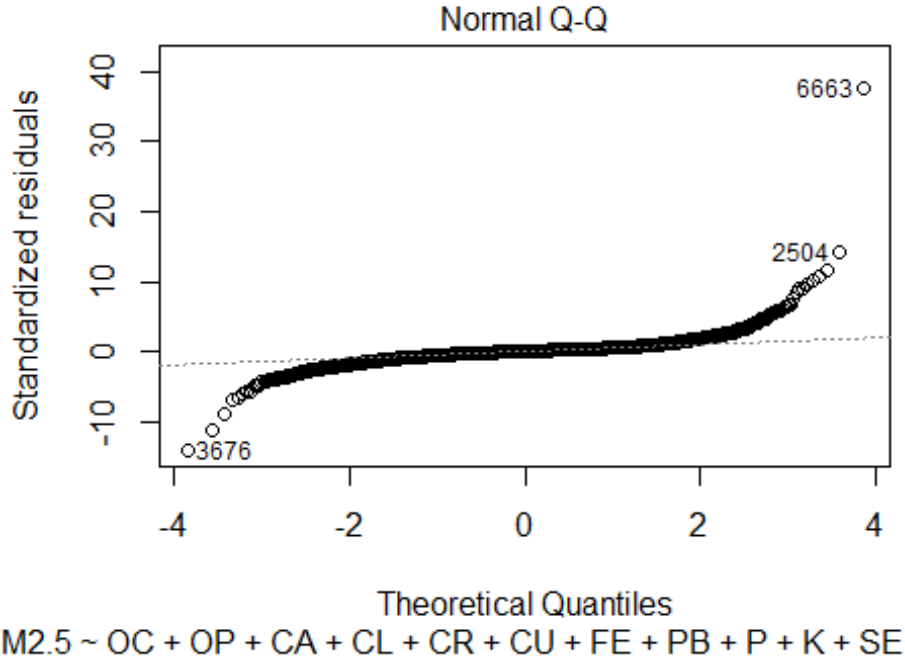


M2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K + CA + CU + PB + P

```
plot(mod6, which = c(1,2))
```

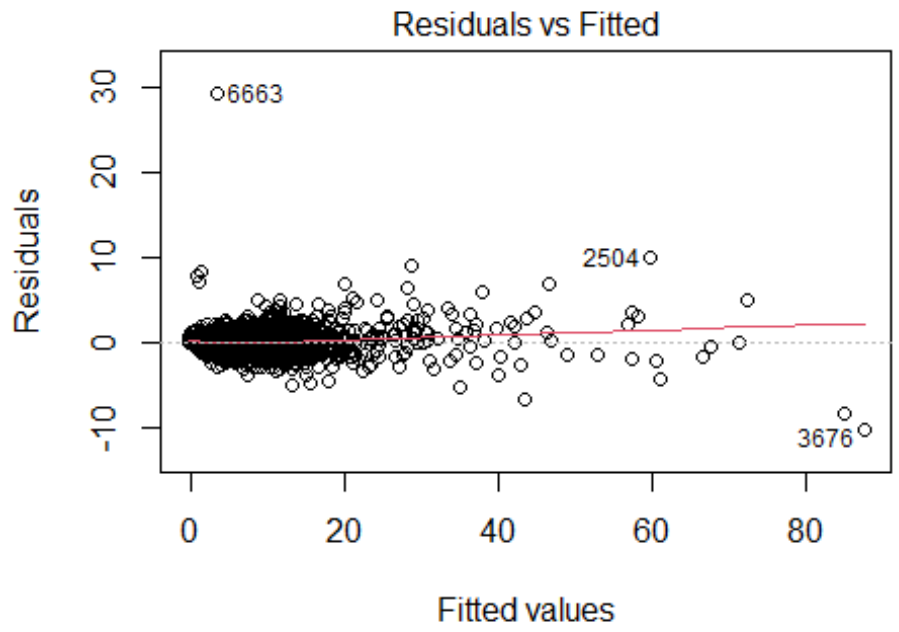


M2.5 ~ OC + OP + CA + CL + CR + CU + FE + PB + P + K + SE + SI

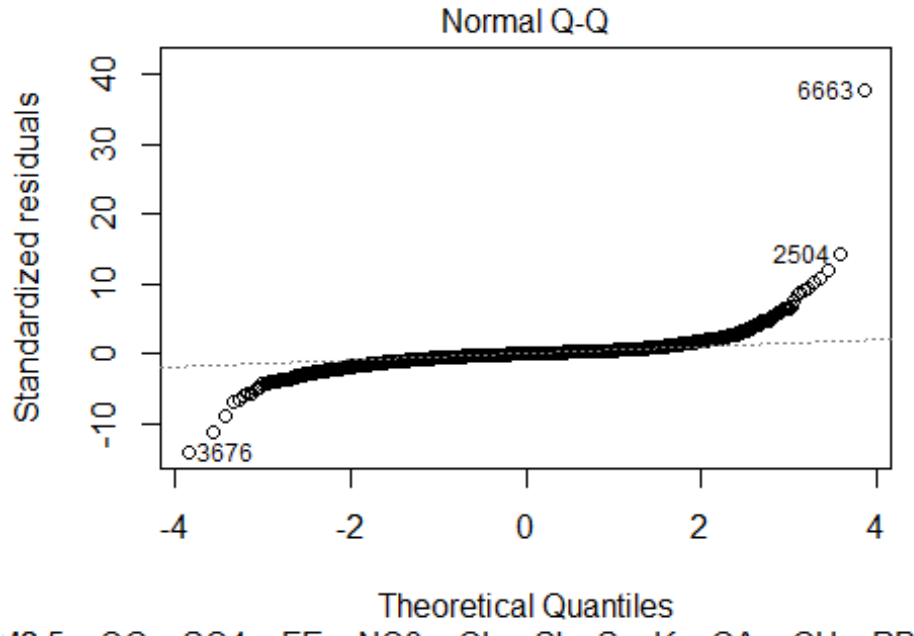


M2.5 ~ OC + OP + CA + CL + CR + CU + FE + PB + P + K + SE + SI

```
plot(mod7, which = c(1,2))
```

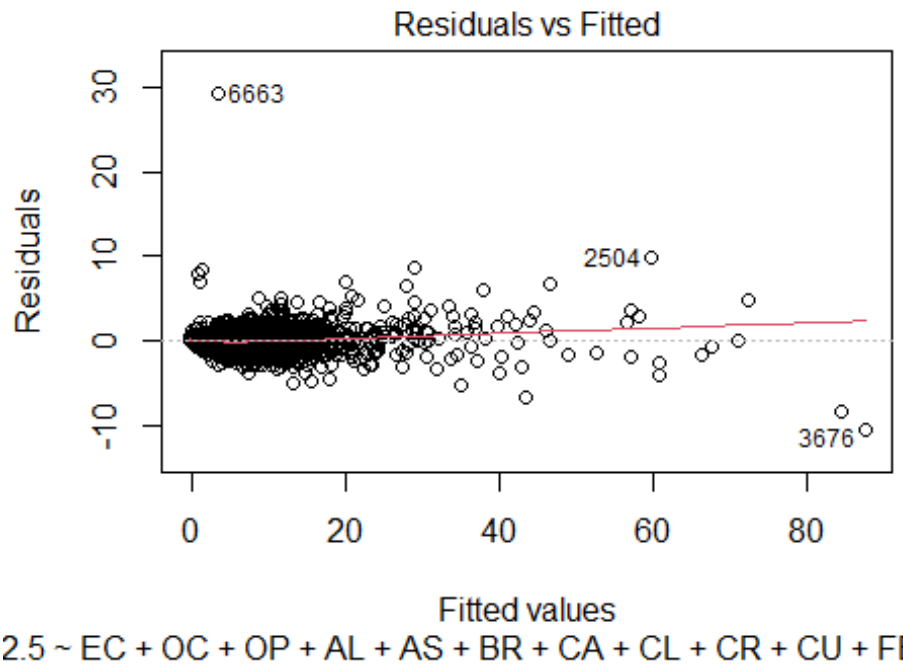


M2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K + CA + CU + PB + P

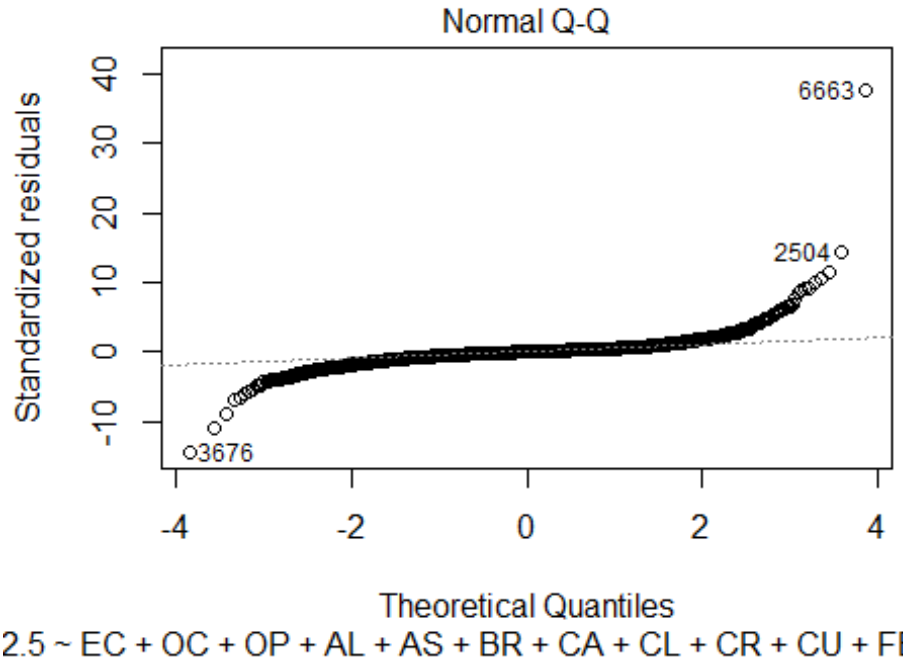


M2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K + CA + CU + PB + P

```
plot(mod8, which = c(1,2))
```



2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL + CR + CU + FE + P



2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL + CR + CU + FE + P

```r
#model1
prediction = mod1 %>% predict(US_DATA_LRG_test)
data.frame( R2 = R2(prediction, US_DATA_LRG_test$PM2.5),
```

```
            RMSE = RMSE(prediction, US_DATA_LRG_test$PM2.5),
            MAE = MAE(prediction, US_DATA_LRG_test$PM2.5))

##          R2      RMSE      MAE
## 1 0.9780208 0.7720496 0.432098

#model2
prediction = mod2 %>% predict(US_DATA_LRG_test)
data.frame( R2 = R2(prediction, US_DATA_LRG_test$PM2.5),
            RMSE = RMSE(prediction, US_DATA_LRG_test$PM2.5),
            MAE = MAE(prediction, US_DATA_LRG_test$PM2.5))

##          R2      RMSE      MAE
## 1 0.9780208 0.7720496 0.432098

#model3
prediction = mod3 %>% predict(US_DATA_LRG_test)
data.frame( R2 = R2(prediction, US_DATA_LRG_test$PM2.5),
            RMSE = RMSE(prediction, US_DATA_LRG_test$PM2.5),
            MAE = MAE(prediction, US_DATA_LRG_test$PM2.5))

##          R2      RMSE      MAE
## 1 0.9780208 0.7720496 0.432098

#model4
prediction = mod4 %>% predict(US_DATA_LRG_test)
data.frame( R2 = R2(prediction, US_DATA_LRG_test$PM2.5),
            RMSE = RMSE(prediction, US_DATA_LRG_test$PM2.5),
            MAE = MAE(prediction, US_DATA_LRG_test$PM2.5))

##          R2      RMSE      MAE
## 1 0.9779738 0.7728304 0.432242

#model5
prediction = mod5 %>% predict(US_DATA_LRG_test)
data.frame( R2 = R2(prediction, US_DATA_LRG_test$PM2.5),
            RMSE = RMSE(prediction, US_DATA_LRG_test$PM2.5),
            MAE = MAE(prediction, US_DATA_LRG_test$PM2.5))

##          R2      RMSE       MAE
## 1 0.9780446 0.7717444 0.4316845

#model6
prediction = mod6 %>% predict(US_DATA_LRG_test)
data.frame( R2 = R2(prediction, US_DATA_LRG_test$PM2.5),
            RMSE = RMSE(prediction, US_DATA_LRG_test$PM2.5),
            MAE = MAE(prediction, US_DATA_LRG_test$PM2.5))

##          R2      RMSE      MAE
## 1 0.9780007 0.7724435 0.431843
```

```r
#model7
prediction = mod7 %>% predict(US_DATA_LRG_test)
data.frame( R2 = R2(prediction, US_DATA_LRG_test$PM2.5),
            RMSE = RMSE(prediction, US_DATA_LRG_test$PM2.5),
            MAE = MAE(prediction, US_DATA_LRG_test$PM2.5))

##          R2      RMSE       MAE
## 1 0.9780446 0.7717444 0.4316845

#model8
prediction = mod8 %>% predict(US_DATA_LRG_test)
data.frame( R2 = R2(prediction, US_DATA_LRG_test$PM2.5),
            RMSE = RMSE(prediction, US_DATA_LRG_test$PM2.5),
            MAE = MAE(prediction, US_DATA_LRG_test$PM2.5))

##          R2      RMSE      MAE
## 1 0.9779738 0.7728304 0.432242
```

#8 models were produced based on 8 different processes. They have similar adjusted coefficient of determination and their assumptions are valid. When testing their predictive ability, all of them have high R2 value and low Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) value.

```r
#consistency of regression coefficient
valid1 = lm(PM2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K + CA + CU + PB + P
+ OP + TI + SE + V + CR + SR + MN + RB, data = US_DATA_LRG)
valid2 = lm(PM2.5 ~ OC + OP + CA + CL + CR + CU + FE + PB + MN + P + K + RB +
SE + SI + SR + S + TI + V + NO3 + SO4, data = US_DATA_LRG)
valid3 = lm(PM2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K + CA + CU + PB + P
+ OP + TI + SE + V + CR + SR + MN + RB, data = US_DATA_LRG)
valid4 = lm(PM2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL + CR + CU + FE + PB
+ MG + MN + NI + N2 + P + K + RB + SE + SI + NA. + SR + S + TI + V + ZN + ZR
+ NO3 + SO4, data = US_DATA_LRG)
valid5 = lm(PM2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K + CA + CU + PB + P
+ OP + TI + SE, data = US_DATA_LRG)
valid6 = lm(PM2.5 ~ OC + OP + CA + CL + CR + CU + FE + PB + P + K + SE + SI +
S + TI + V + NO3 + SO4, data = US_DATA_LRG)
valid7 = lm(PM2.5 ~ OC + SO4 + FE + NO3 + CL + SI + S + K + CA + CU + PB + P
+ OP + TI + SE, data = US_DATA_LRG)
valid8 = lm(PM2.5 ~ EC + OC + OP + AL + AS + BR + CA + CL + CR + CU + FE + PB
+ MG + MN + NI + N2 + P + K + RB + SE + SI + NA. + SR + S + TI + V + ZN + ZR
+ NO3 + SO4, data = US_DATA_LRG)
cbind(coef(summary(mod1))[,1], coef(summary(valid1))[,1])

##                    [,1]         [,2]
## (Intercept)   -0.2045776   -0.2045776
## OC             1.9231454    1.9231454
## SO4            0.3992574    0.3992574
## FE             3.7193158    3.7193158
## NO3            1.2209776    1.2209776
## CL             3.5368317    3.5368317
```

```
## SI             2.7651292    2.7651292
## S              3.9182363    3.9182363
## K              2.9583179    2.9583179
## CA             2.0211397    2.0211397
## CU           -26.0709656  -26.0709656
## PB            26.1187091   26.1187091
## P             45.1030623   45.1030623
## OP             0.2382850    0.2382850
## TI            15.0222095   15.0222095
## SE           146.7717588  146.7717588
## V             38.3276088   38.3276088
## CR          -154.9148578 -154.9148578
## SR           -15.1059510  -15.1059510
## MN           -22.3443123  -22.3443123
## RB            63.6468034   63.6468034
```

```
cbind(coef(summary(mod2))[,1], coef(summary(valid2))[,1])
```

```
##                    [,1]         [,2]
## (Intercept)   -0.2045776   -0.2045776
## OC             1.9231454    1.9231454
## OP             0.2382850    0.2382850
## CA             2.0211397    2.0211397
## CL             3.5368317    3.5368317
## CR          -154.9148578 -154.9148578
## CU           -26.0709656  -26.0709656
## FE             3.7193158    3.7193158
## PB            26.1187091   26.1187091
## MN           -22.3443123  -22.3443123
## P             45.1030623   45.1030623
## K              2.9583179    2.9583179
## RB            63.6468034   63.6468034
## SE           146.7717588  146.7717588
## SI             2.7651292    2.7651292
## SR           -15.1059510  -15.1059510
## S              3.9182363    3.9182363
## TI            15.0222095   15.0222095
## V             38.3276088   38.3276088
## NO3            1.2209776    1.2209776
## SO4            0.3992574    0.3992574
```

```
cbind(coef(summary(mod3))[,1], coef(summary(valid3))[,1])
```

```
##                    [,1]         [,2]
## (Intercept)   -0.2045776   -0.2045776
## OC             1.9231454    1.9231454
## SO4            0.3992574    0.3992574
## FE             3.7193158    3.7193158
## NO3            1.2209776    1.2209776
## CL             3.5368317    3.5368317
## SI             2.7651292    2.7651292
```

```
## S              3.9182363    3.9182363
## K              2.9583179    2.9583179
## CA             2.0211397    2.0211397
## CU           -26.0709656  -26.0709656
## PB            26.1187091   26.1187091
## P             45.1030623   45.1030623
## OP             0.2382850    0.2382850
## TI            15.0222095   15.0222095
## SE           146.7717588  146.7717588
## V             38.3276088   38.3276088
## CR          -154.9148578 -154.9148578
## SR           -15.1059510  -15.1059510
## MN           -22.3443123  -22.3443123
## RB            63.6468034   63.6468034

cbind(coef(summary(mod4))[,1], coef(summary(valid4))[,1])

##                      [,1]          [,2]
## (Intercept)   -0.21256462   -0.21256462
## EC            -0.11101415   -0.11101415
## OC             1.93475697    1.93475697
## OP             0.22447699    0.22447699
## AL            -0.58136572   -0.58136572
## AS            16.62335672   16.62335672
## BR             5.48740356    5.48740356
## CA             1.91323246    1.91323246
## CL             3.45763092    3.45763092
## CR          -148.03755914 -148.03755914
## CU           -26.39870378  -26.39870378
## FE             3.65516332    3.65516332
## PB            24.95061278   24.95061278
## MG            -0.03643026   -0.03643026
## MN           -20.57501447  -20.57501447
## NI            49.78919714   49.78919714
## N2             0.04224962    0.04224962
## P             44.97507876   44.97507876
## K              2.96531018    2.96531018
## RB            62.93135116   62.93135116
## SE           144.02217877  144.02217877
## SI             2.99261731    2.99261731
## NA.            0.11403627    0.11403627
## SR           -14.43223775  -14.43223775
## S              3.92477661    3.92477661
## TI            17.30420427   17.30420427
## V             25.86339643   25.86339643
## ZN            -0.55865933   -0.55865933
## ZR             9.27635580    9.27635580
## NO3            1.22060219    1.22060219
## SO4            0.39604787    0.39604787
```

```r
cbind(coef(summary(mod5))[,1], coef(summary(valid5))[,1])
```

```
##                    [,1]        [,2]
## (Intercept)  -0.2067957  -0.2067957
## OC            1.9301502   1.9301502
## SO4           0.4268946   0.4268946
## FE            2.2614638   2.2614638
## NO3           1.2238739   1.2238739
## CL            3.5397484   3.5397484
## SI            2.9902615   2.9902615
## S             3.8677503   3.8677503
## K             2.4362540   2.4362540
## CA            1.9113693   1.9113693
## CU          -29.5284818 -29.5284818
## PB           24.5011139  24.5011139
## P            47.1500977  47.1500977
## OP            0.2289468   0.2289468
## TI           18.4562834  18.4562834
## SE          141.3512728 141.3512728
```

```r
cbind(coef(summary(mod6))[,1], coef(summary(valid6))[,1])
```

```
##                     [,1]         [,2]
## (Intercept)   -0.2036429   -0.2036429
## OC             1.9290377    1.9290377
## OP             0.2314089    0.2314089
## CA             1.8704291    1.8704291
## CL             3.5338299    3.5338299
## CR          -170.0702343 -170.0702343
## CU           -25.8685233  -25.8685233
## FE             3.0005846    3.0005846
## PB            25.3082208   25.3082208
## P             45.4777712   45.4777712
## K              2.5654955    2.5654955
## SE           141.9082334  141.9082334
## SI             2.8900618    2.8900618
## S              3.8931129    3.8931129
## TI            15.8010448   15.8010448
## V             37.0164352   37.0164352
## NO3            1.2261903    1.2261903
## SO4            0.4097229    0.4097229
```

```r
cbind(coef(summary(mod7))[,1], coef(summary(valid7))[,1])
```

```
##                    [,1]        [,2]
## (Intercept)  -0.2067957  -0.2067957
## OC            1.9301502   1.9301502
## SO4           0.4268946   0.4268946
## FE            2.2614638   2.2614638
## NO3           1.2238739   1.2238739
## CL            3.5397484   3.5397484
```

```
## SI            2.9902615   2.9902615
## S             3.8677503   3.8677503
## K             2.4362540   2.4362540
## CA            1.9113693   1.9113693
## CU          -29.5284818 -29.5284818
## PB           24.5011139  24.5011139
## P            47.1500977  47.1500977
## OP            0.2289468   0.2289468
## TI           18.4562834  18.4562834
## SE          141.3512728 141.3512728
```

```
cbind(coef(summary(mod8))[,1], coef(summary(valid8))[,1])
```

```
##                      [,1]          [,2]
## (Intercept)   -0.21256462   -0.21256462
## EC            -0.11101415   -0.11101415
## OC             1.93475697    1.93475697
## OP             0.22447699    0.22447699
## AL            -0.58136572   -0.58136572
## AS            16.62335672   16.62335672
## BR             5.48740356    5.48740356
## CA             1.91323246    1.91323246
## CL             3.45763092    3.45763092
## CR          -148.03755914 -148.03755914
## CU           -26.39870378  -26.39870378
## FE             3.65516332    3.65516332
## PB            24.95061278   24.95061278
## MG            -0.03643026   -0.03643026
## MN           -20.57501447  -20.57501447
## NI            49.78919714   49.78919714
## N2             0.04224962    0.04224962
## P             44.97507876   44.97507876
## K              2.96531018    2.96531018
## RB            62.93135116   62.93135116
## SE           144.02217877  144.02217877
## SI             2.99261731    2.99261731
## NA.            0.11403627    0.11403627
## SR           -14.43223775  -14.43223775
## S              3.92477661    3.92477661
## TI            17.30420427   17.30420427
## V             25.86339643   25.86339643
## ZN            -0.55865933   -0.55865933
## ZR             9.27635580    9.27635580
## NO3            1.22060219    1.22060219
## SO4            0.39604787    0.39604787
```

#The regression coefficients are consistency between training data and testing data in all of these models.

```
#Complexity of models
length(coef(summary(mod1))[,1])
```

```
## [1] 21
```

```
length(coef(summary(mod2))[,1])
```

```
## [1] 21
```

```
length(coef(summary(mod3))[,1])
```

```
## [1] 21
```

```
length(coef(summary(mod4))[,1])
```

```
## [1] 31
```

```
length(coef(summary(mod5))[,1])
```

```
## [1] 16
```

```
length(coef(summary(mod6))[,1])
```

```
## [1] 18
```

```
length(coef(summary(mod7))[,1])
```

```
## [1] 16
```

```
length(coef(summary(mod8))[,1])
```

```
## [1] 31
```