hw_2.2 Elizaveta Grigoreva_v2

```
library(data.table)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(ggpubr)
library(tidyr)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:ggpubr':
##
##     mutate
```

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

```
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

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

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

Load and clean dataset

```
aq <- read.csv2("/Users/Lisa/Downloads/Telegram Desktop/AirQualityUCI/AirQualityUCI.csv", header=T)
str(aq)
```

```
## 'data.frame':    9471 obs. of  17 variables:
##  $ Date         : Factor w/ 392 levels "","01/01/2005",..: 116 116 116 116 116 116 129 129 129 129 .
##  $ Time         : Factor w/ 25 levels "","00.00.00",..: 20 21 22 23 24 25 2 3 4 5 ...
##  $ CO.GT.       : num  2.6 2 2.2 2.2 1.6 1.2 1.2 1 0.9 0.6 ...
##  $ PT08.S1.CO.  : int  1360 1292 1402 1376 1272 1197 1185 1136 1094 1010 ...
##  $ NMHC.GT.     : int  150 112 88 80 51 38 31 31 24 19 ...
##  $ C6H6.GT.     : num  11.9 9.4 9 9.2 6.5 4.7 3.6 3.3 2.3 1.7 ...
##  $ PT08.S2.NMHC.: int  1046 955 939 948 836 750 690 672 609 561 ...
##  $ NOx.GT.      : int  166 103 131 172 131 89 62 62 45 -200 ...
##  $ PT08.S3.NOx. : int  1056 1174 1140 1092 1205 1337 1462 1453 1579 1705 ...
##  $ NO2.GT.      : int  113 92 114 122 116 96 77 76 60 -200 ...
##  $ PT08.S4.NO2. : int  1692 1559 1555 1584 1490 1393 1333 1333 1276 1235 ...
##  $ PT08.S5.O3.  : int  1268 972 1074 1203 1110 949 733 730 620 501 ...
##  $ T            : num  13.6 13.3 11.9 11 11.2 11.2 11.3 10.7 10.7 10.3 ...
##  $ RH           : num  48.9 47.7 54 60 59.6 59.2 56.8 60 59.7 60.2 ...
##  $ AH           : num  0.758 0.726 0.75 0.787 0.789 ...
##  $ X            : logi  NA NA NA NA NA NA ...
##  $ X.1          : logi  NA NA NA NA NA NA ...
```

```
#Drop strange variables 16,17
aq <- aq[,-c(16, 17)]
str(aq)
```

```
## 'data.frame':    9471 obs. of  15 variables:
##  $ Date         : Factor w/ 392 levels "","01/01/2005",..: 116 116 116 116 116 116 129 129 129 129 .
##  $ Time         : Factor w/ 25 levels "","00.00.00",..: 20 21 22 23 24 25 2 3 4 5 ...
##  $ CO.GT.       : num  2.6 2 2.2 2.2 1.6 1.2 1.2 1 0.9 0.6 ...
##  $ PT08.S1.CO.  : int  1360 1292 1402 1376 1272 1197 1185 1136 1094 1010 ...
```

```
##  $ NMHC.GT.      : int  150 112 88 80 51 38 31 31 24 19 ...
##  $ C6H6.GT.      : num  11.9 9.4 9 9.2 6.5 4.7 3.6 3.3 2.3 1.7 ...
##  $ PT08.S2.NMHC.: int  1046 955 939 948 836 750 690 672 609 561 ...
##  $ NOx.GT.       : int  166 103 131 172 131 89 62 62 45 -200 ...
##  $ PT08.S3.NOx. : int  1056 1174 1140 1092 1205 1337 1462 1453 1579 1705 ...
##  $ NO2.GT.       : int  113 92 114 122 116 96 77 76 60 -200 ...
##  $ PT08.S4.NO2. : int  1692 1559 1555 1584 1490 1393 1333 1333 1276 1235 ...
##  $ PT08.S5.O3.  : int  1268 972 1074 1203 1110 949 733 730 620 501 ...
##  $ T             : num  13.6 13.3 11.9 11 11.2 11.2 11.3 10.7 10.7 10.3 ...
##  $ RH            : num  48.9 47.7 54 60 59.6 59.2 56.8 60 59.7 60.2 ...
##  $ AH            : num  0.758 0.726 0.75 0.787 0.789 ...
```

```
summary(aq)
```

```
##          Date             Time          CO.GT.          PT08.S1.CO.
##             : 114   00.00.00: 390   Min.   :-200.00   Min.   :-200
##  01/01/2005:  24   01.00.00: 390   1st Qu.:   0.60   1st Qu.: 921
##  01/02/2005:  24   02.00.00: 390   Median :   1.50   Median :1053
##  01/03/2005:  24   03.00.00: 390   Mean   : -34.21   Mean   :1049
##  01/04/2004:  24   04.00.00: 390   3rd Qu.:   2.60   3rd Qu.:1221
##  01/04/2005:  24   05.00.00: 390   Max.   :  11.90   Max.   :2040
##  (Other)   :9237   (Other) :7131   NA's   :114       NA's   :114
##     NMHC.GT.         C6H6.GT.        PT08.S2.NMHC.       NOx.GT.
##  Min.   :-200.0   Min.   :-200.000   Min.   :-200.0   Min.   :-200.0
##  1st Qu.:-200.0   1st Qu.:   4.000   1st Qu.: 711.0   1st Qu.:  50.0
##  Median :-200.0   Median :   7.900   Median : 895.0   Median : 141.0
##  Mean   :-159.1   Mean   :   1.866   Mean   : 894.6   Mean   : 168.6
##  3rd Qu.:-200.0   3rd Qu.:  13.600   3rd Qu.:1105.0   3rd Qu.: 284.0
##  Max.   :1189.0   Max.   :  63.700   Max.   :2214.0   Max.   :1479.0
##  NA's   :114      NA's   :114        NA's   :114      NA's   :114
##   PT08.S3.NOx.      NO2.GT.         PT08.S4.NO2.   PT08.S5.O3.
##  Min.   :-200   Min.   :-200.00   Min.   :-200   Min.   :-200.0
##  1st Qu.: 637   1st Qu.:  53.00   1st Qu.:1185   1st Qu.: 700.0
##  Median : 794   Median :  96.00   Median :1446   Median : 942.0
##  Mean   : 795   Mean   :  58.15   Mean   :1391   Mean   : 975.1
##  3rd Qu.: 960   3rd Qu.: 133.00   3rd Qu.:1662   3rd Qu.:1255.0
##  Max.   :2683   Max.   : 340.00   Max.   :2775   Max.   :2523.0
##  NA's   :114    NA's   :114       NA's   :114    NA's   :114
##        T                RH                AH
##  Min.   :-200.000   Min.   :-200.00   Min.   :-200.0000
##  1st Qu.:  10.900   1st Qu.:  34.10   1st Qu.:   0.6923
##  Median :  17.200   Median :  48.60   Median :   0.9768
##  Mean   :   9.778   Mean   :  39.49   Mean   :  -6.8376
##  3rd Qu.:  24.100   3rd Qu.:  61.90   3rd Qu.:   1.2962
##  Max.   :  44.600   Max.   :  88.70   Max.   :   2.2310
##  NA's   :114        NA's   :114       NA's   :114
```

```
#We can see in columns contain  a lot of '-200' that =NA, we can drop column with a lot of NA's
aq <- aq[, -5]
```

```
airq_long <- gather(aq, key="measurement", value="value", -c(Date,Time))
airq_long$Date <- as.factor(airq_long$Date)
airq_long$Time <- as.factor(airq_long$Time)
airq_long$measurement <- as.factor(airq_long$measurement)
```

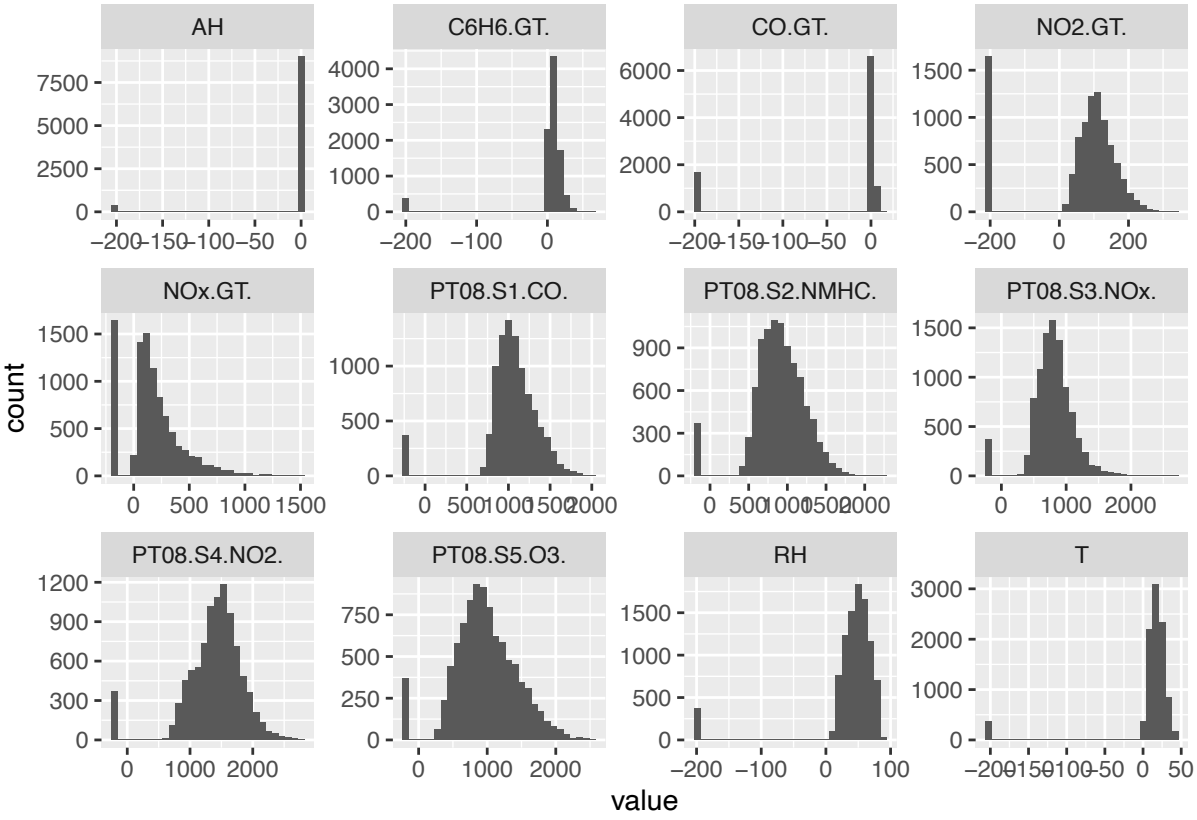
```r
colSums(is.na(airq_long))
```

```
##         Date         Time measurement        value
##            0            0            0         1368
```

```r
ggplot(airq_long, aes(value)) +
  geom_histogram() +
  facet_wrap(~measurement, scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1368 rows containing non-finite values (stat_bin).
```



Normalization

```r
#We can clean -200 variables
airq_fil <- airq_long %>%
  filter(value != -200)
summary(airq_fil)
```

```
##          Date              Time              measurement        value
##  02/04/2005:   288   09.00.00: 4396   AH           : 8991   Min.   :  -1.9
##  03/04/2005:   288   10.00.00: 4392   C6H6.GT.     : 8991   1st Qu.:  13.2
##  15/03/2005:   288   12.00.00: 4389   PT08.S1.CO.  : 8991   Median : 135.0
##  16/03/2005:   288   13.00.00: 4385   PT08.S2.NMHC.: 8991   Mean   : 496.4
##  18/03/2005:   288   11.00.00: 4378   PT08.S3.NOx. : 8991   3rd Qu.: 948.0
##  19/03/2005:   288   05.00.00: 4376   PT08.S4.NO2. : 8991   Max.   :2775.0
##  (Other)   :102298   (Other) :77710   (Other)      :50080
```

```r
colSums(is.na(airq_fil))
```

```
##        Date        Time measurement       value
##           0           0           0           0
```

```r
airq_transformed <- spread(airq_fil, key = "measurement" ,value = "value")
airq_transformed<- na.omit(airq_transformed)
summary(airq_transformed)
```

```
##          Date             Time             AH              C6H6.GT.
##  02/04/2005:  24   10.00.00: 312   Min.   :0.1847   Min.   : 0.20
##  03/04/2005:  24   20.00.00: 310   1st Qu.:0.6941   1st Qu.: 4.90
##  15/03/2005:  24   09.00.00: 309   Median :0.9539   Median : 8.80
##  16/03/2005:  24   12.00.00: 309   Mean   :0.9856   Mean   :10.55
##  18/03/2005:  24   18.00.00: 309   3rd Qu.:1.2516   3rd Qu.:14.60
##  19/03/2005:  24   21.00.00: 309   Max.   :2.1806   Max.   :63.70
##  (Other)   :6797   (Other) :5083
##      CO.GT.           NO2.GT.          NOx.GT.         PT08.S1.CO.
##  Min.   : 0.100   Min.   :  2.0   Min.   :   2.0   Min.   : 647
##  1st Qu.: 1.100   1st Qu.: 79.0   1st Qu.: 103.0   1st Qu.: 956
##  Median : 1.900   Median :110.0   Median : 186.0   Median :1085
##  Mean   : 2.182   Mean   :113.9   Mean   : 250.7   Mean   :1120
##  3rd Qu.: 2.900   3rd Qu.:142.0   3rd Qu.: 335.0   3rd Qu.:1254
##  Max.   :11.900   Max.   :333.0   Max.   :1479.0   Max.   :2040
##
##  PT08.S2.NMHC.     PT08.S3.NOx.     PT08.S4.NO2.   PT08.S5.O3.
##  Min.   : 390.0   Min.   : 322.0   Min.   : 551   Min.   : 221
##  1st Qu.: 760.0   1st Qu.: 642.0   1st Qu.:1207   1st Qu.: 760
##  Median : 931.0   Median : 786.0   Median :1457   Median :1006
##  Mean   : 958.5   Mean   : 816.9   Mean   :1453   Mean   :1058
##  3rd Qu.:1135.0   3rd Qu.: 947.0   3rd Qu.:1683   3rd Qu.:1322
##  Max.   :2214.0   Max.   :2683.0   Max.   :2775   Max.   :2523
##
##        RH              T
##  Min.   : 9.20   Min.   :-1.90
##  1st Qu.:35.30   1st Qu.:11.20
##  Median :49.20   Median :16.80
##  Mean   :48.88   Mean   :17.76
##  3rd Qu.:62.20   3rd Qu.:23.70
##  Max.   :88.70   Max.   :44.60
##
```

```r
str(airq_transformed)
```

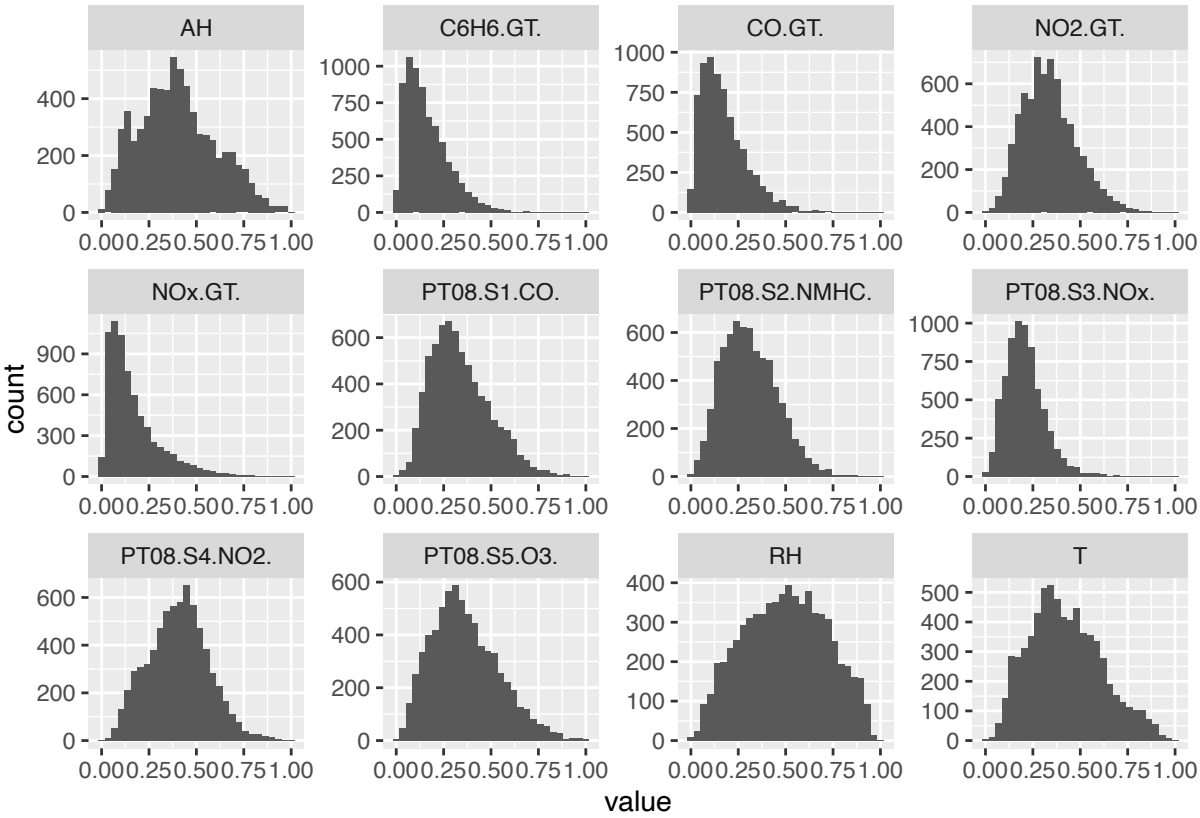
```
## 'data.frame':    6941 obs. of  14 variables:
##  $ Date         : Factor w/ 392 levels "","01/01/2005",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Time         : Factor w/ 25 levels "","00.00.00",..: 3 4 6 7 8 9 10 11 12 13 ...
##  $ AH           : num  0.456 0.469 0.465 0.476 0.464 ...
##  $ C6H6.GT.     : num  8.8 7.5 5.6 4.8 5.3 4.5 3 3 4.7 5.4 ...
##  $ CO.GT.       : num  1.6 2.5 1.9 1.4 1.5 1.4 1.1 1 1.2 1.7 ...
##  $ NO2.GT.      : num  106 129 126 106 99 97 94 86 97 113 ...
##  $ NOx.GT.      : num  215 300 253 181 171 168 169 145 190 225 ...
##  $ PT08.S1.CO.  : num  1275 1173 1054 1004 1001 ...
##  $ PT08.S2.NMHC.: num  930 878 791 753 777 736 653 649 748 782 ...
##  $ PT08.S3.NOx. : num  649 738 830 879 859 888 973 996 878 846 ...
```

```
##  $ PT08.S4.NO2. : num  1024 1002 967 942 954 ...
##  $ PT08.S5.O3.  : num  1617 1355 1131 1036 1009 ...
##  $ RH           : num  50.7 50 55.3 57.1 58.3 60.7 63.2 59 57.2 48.6 ...
##  $ T            : num  5.3 5.9 4.3 4.2 3.5 3 2.6 3.9 4.7 6.8 ...
##  - attr(*, "na.action")= 'omit' Named int  1 4 28 52 76 87 88 89 101 124 ...
##   ..- attr(*, "names")= chr  "1" "4" "28" "52" ...
```

```
#Function
normalize <- function(x) {
return ((x - min(x)) / (max(x) - min(x)))
}
air_norm <- as.data.frame(lapply(airq_transformed[3:14], normalize))
air_norm$Date <- airq_transformed$Date
air_norm$Time <- airq_transformed$Time
air_norm_long <- gather(air_norm, key = "measurement", value = "value", -c(Date, Time))
```

```
#Let's check hisogram
ggplot(air_norm_long, aes(value)) +
  geom_histogram() +
  facet_wrap(~ measurement, scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Explore multicollinearity, choose good predictors. I choose predictors based on the previous hw (that had linear deoendency)

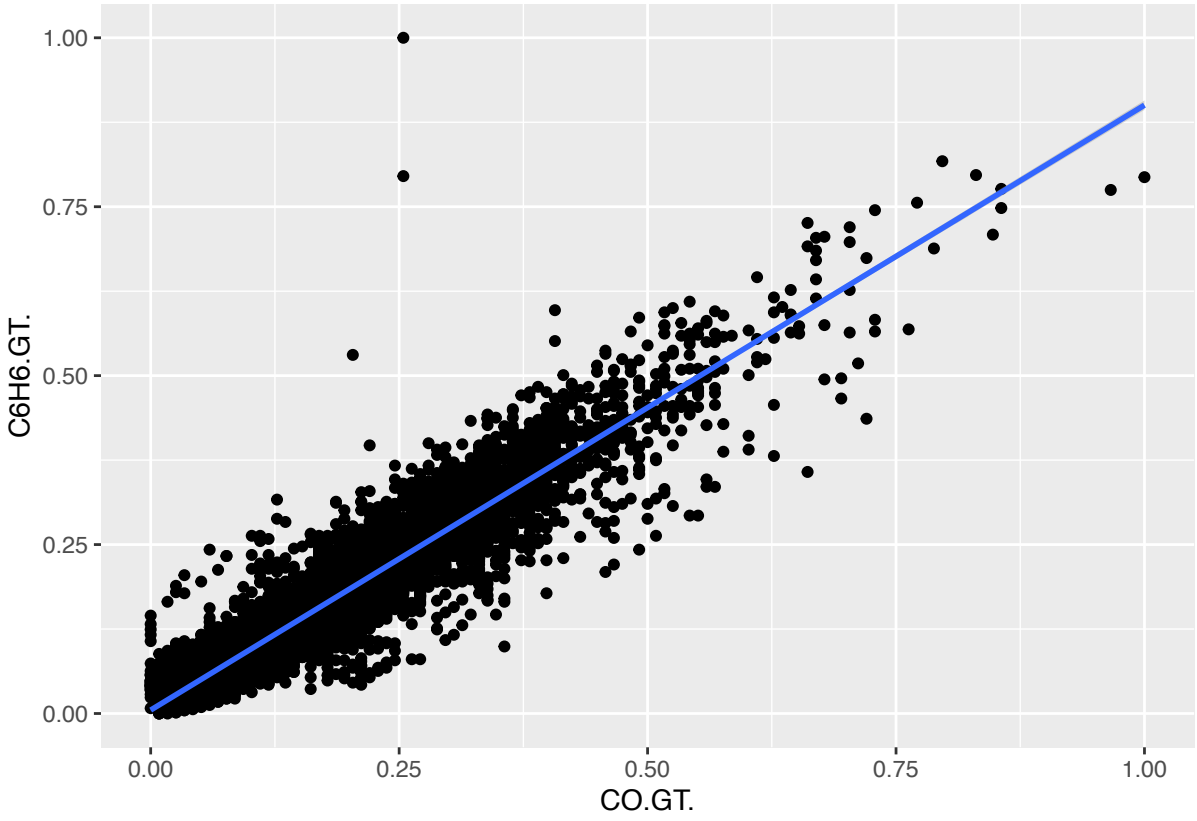
```
#Response C6H6.GT ~ CO.GT
summary(air_norm)
```

```
##        AH            C6H6.GT.          CO.GT.          NO2.GT.
```

```
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.2552   1st Qu.:0.07402   1st Qu.:0.08475   1st Qu.:0.2326
##  Median :0.3854   Median :0.13543   Median :0.15254   Median :0.3263
##  Mean   :0.4013   Mean   :0.16306   Mean   :0.17648   Mean   :0.3380
##  3rd Qu.:0.5345   3rd Qu.:0.22677   3rd Qu.:0.23729   3rd Qu.:0.4230
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##
##     NOx.GT.          PT08.S1.CO.     PT08.S2.NMHC.     PT08.S3.NOx.
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.06838   1st Qu.:0.2218   1st Qu.:0.2029   1st Qu.:0.1355
##  Median :0.12458   Median :0.3144   Median :0.2966   Median :0.1965
##  Mean   :0.16836   Mean   :0.3395   Mean   :0.3117   Mean   :0.2096
##  3rd Qu.:0.22546   3rd Qu.:0.4358   3rd Qu.:0.4084   3rd Qu.:0.2647
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##   PT08.S4.NO2.     PT08.S5.O3.           RH                T
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2950   1st Qu.:0.2341   1st Qu.:0.3283   1st Qu.:0.2817
##  Median :0.4074   Median :0.3410   Median :0.5031   Median :0.4022
##  Mean   :0.4054   Mean   :0.3635   Mean   :0.4991   Mean   :0.4227
##  3rd Qu.:0.5090   3rd Qu.:0.4783   3rd Qu.:0.6667   3rd Qu.:0.5505
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##          Date             Time
##  02/04/2005:  24   10.00.00: 312
##  03/04/2005:  24   20.00.00: 310
##  15/03/2005:  24   09.00.00: 309
##  16/03/2005:  24   12.00.00: 309
##  18/03/2005:  24   18.00.00: 309
##  19/03/2005:  24   21.00.00: 309
##  (Other)   :6797   (Other) :5083
```

```
air_norm %>%
  ggplot(aes(x= CO.GT., y= C6H6.GT.)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



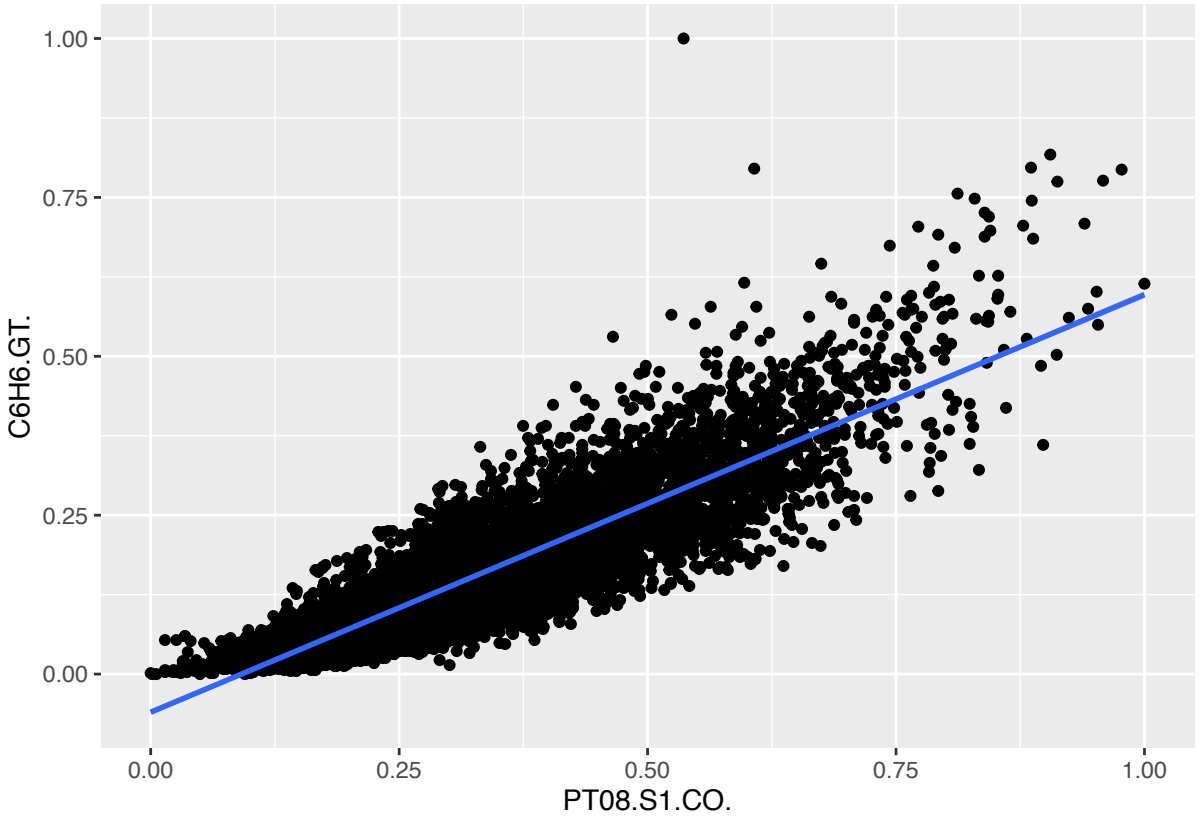

```
#R^2 is  significant
air_norm  %>%
  lm(data= .,CO.GT. ~ C6H6.GT.)%>%
  summary()
```

```
##
## Call:
## lm(formula = CO.GT. ~ C6H6.GT., data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73085 -0.02546 -0.00156  0.02100  0.29670
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0189361  0.0009214   20.55   <2e-16 ***
## C6H6.GT.    0.9661559  0.0045837  210.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04489 on 6939 degrees of freedom
## Multiple R-squared:  0.8649, Adjusted R-squared:  0.8649
## F-statistic: 4.443e+04 on 1 and 6939 DF,  p-value: < 2.2e-16
```

```
#Another predictor with linear response
air_norm %>%
  ggplot(aes(x= PT08.S1.CO., y= C6H6.GT.)) +
  geom_point() +
```

```
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

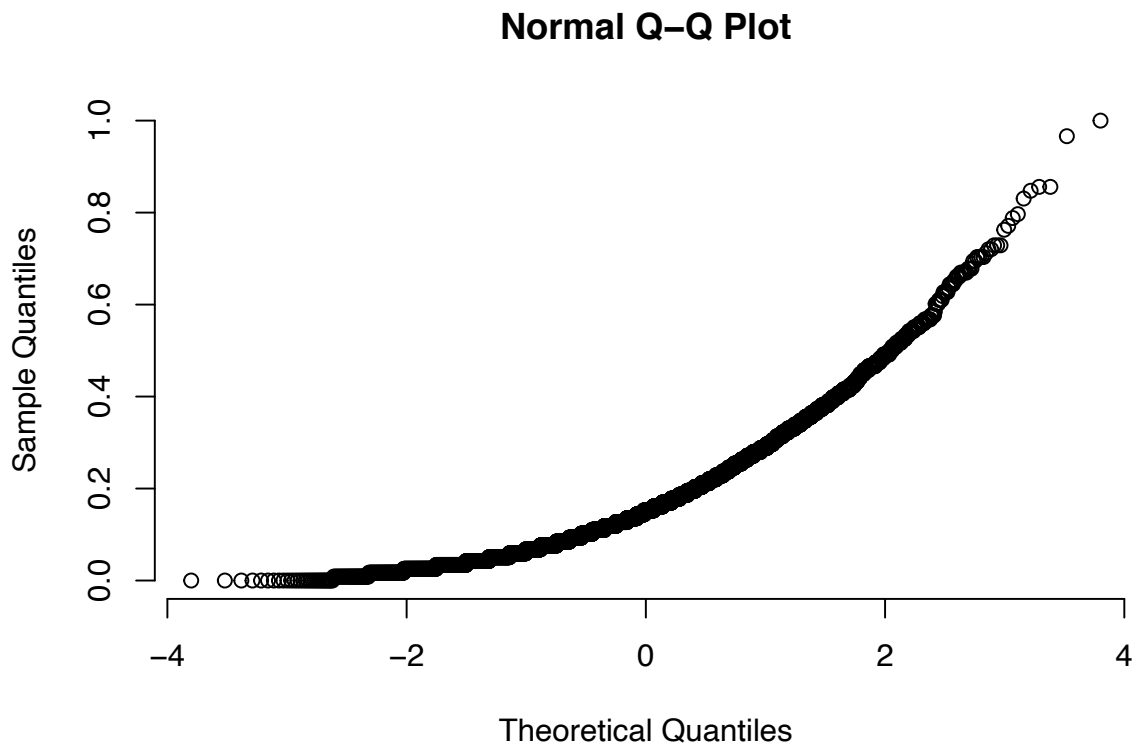



```
air_norm  %>%
  lm(data= .,C6H6.GT. ~ PT08.S1.CO.)%>%
  summary()
```

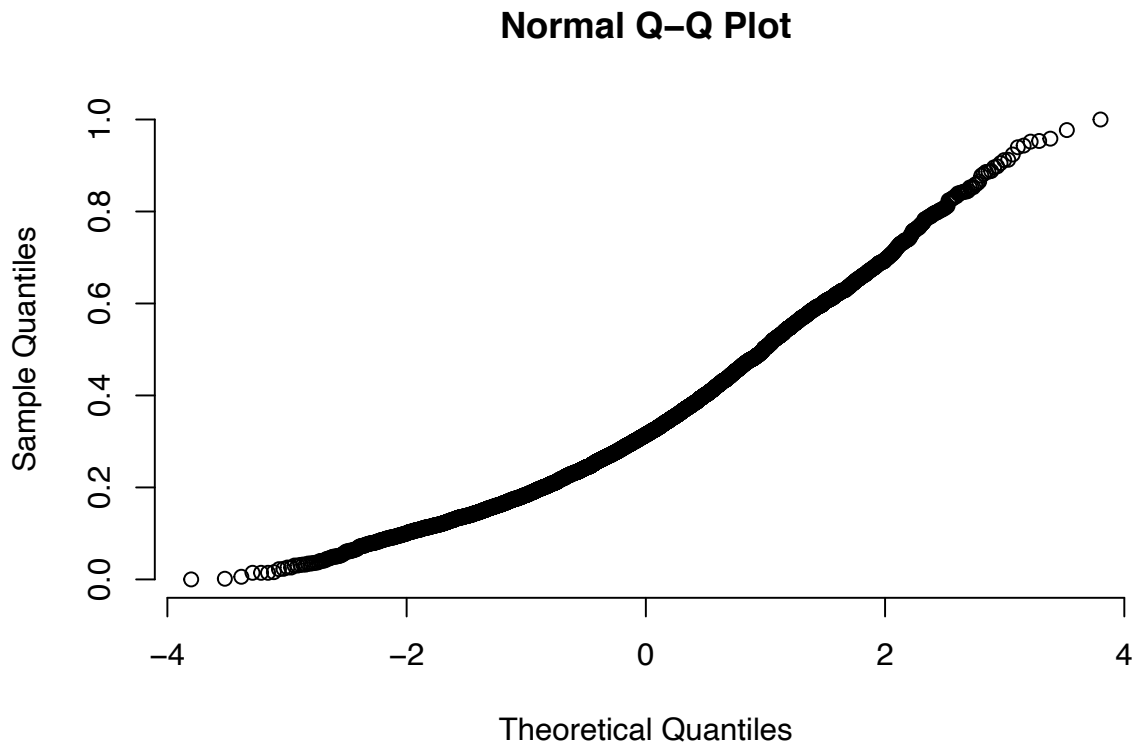
```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S1.CO., data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18826 -0.03559 -0.00288  0.03079  0.70768
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.059959   0.001613  -37.18   <2e-16 ***
## PT08.S1.CO.  0.656927   0.004312  152.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0564 on 6939 degrees of freedom
## Multiple R-squared:  0.7699, Adjusted R-squared:  0.7699
## F-statistic: 2.322e+04 on 1 and 6939 DF,  p-value: < 2.2e-16
```

We can check quantiles for this data

```
qqnorm(air_norm$CO.GT., pch = 1, frame = FALSE)
```




```
qqnorm(air_norm$PT08.S1.CO., pch = 1, frame = FALSE)
```

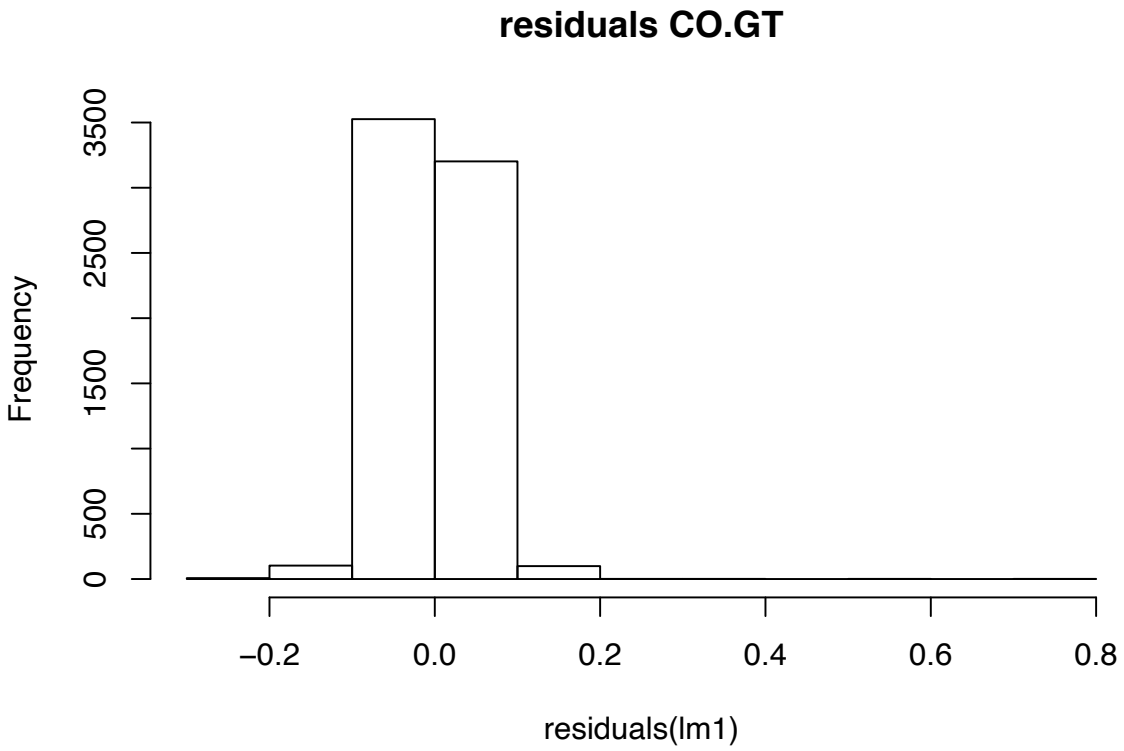


Check residuals for this predictors PT08.S1.CO and CO.GT

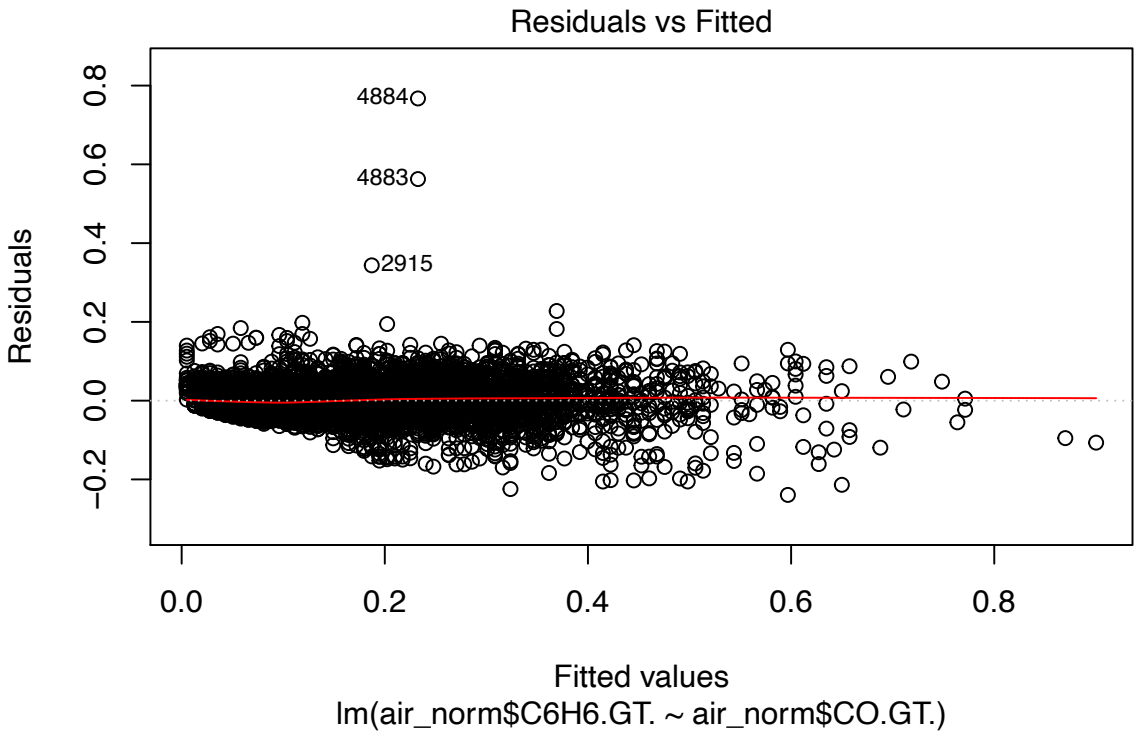
```r
lm1 <- lm(air_norm$C6H6.GT. ~ air_norm$CO.GT.)
summary(lm1)
```

```
##
## Call:
## lm(formula = air_norm$C6H6.GT. ~ air_norm$CO.GT.)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23935 -0.02496 -0.00249  0.02357  0.76733
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.0050753  0.0009115   5.568 2.67e-08 ***
## air_norm$CO.GT. 0.8952134  0.0042471 210.782  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04321 on 6939 degrees of freedom
## Multiple R-squared:  0.8649, Adjusted R-squared:  0.8649
## F-statistic: 4.443e+04 on 1 and 6939 DF,  p-value: < 2.2e-16
```

```r
#Check residuals
hist(residuals(lm1),main ="residuals CO.GT")
```



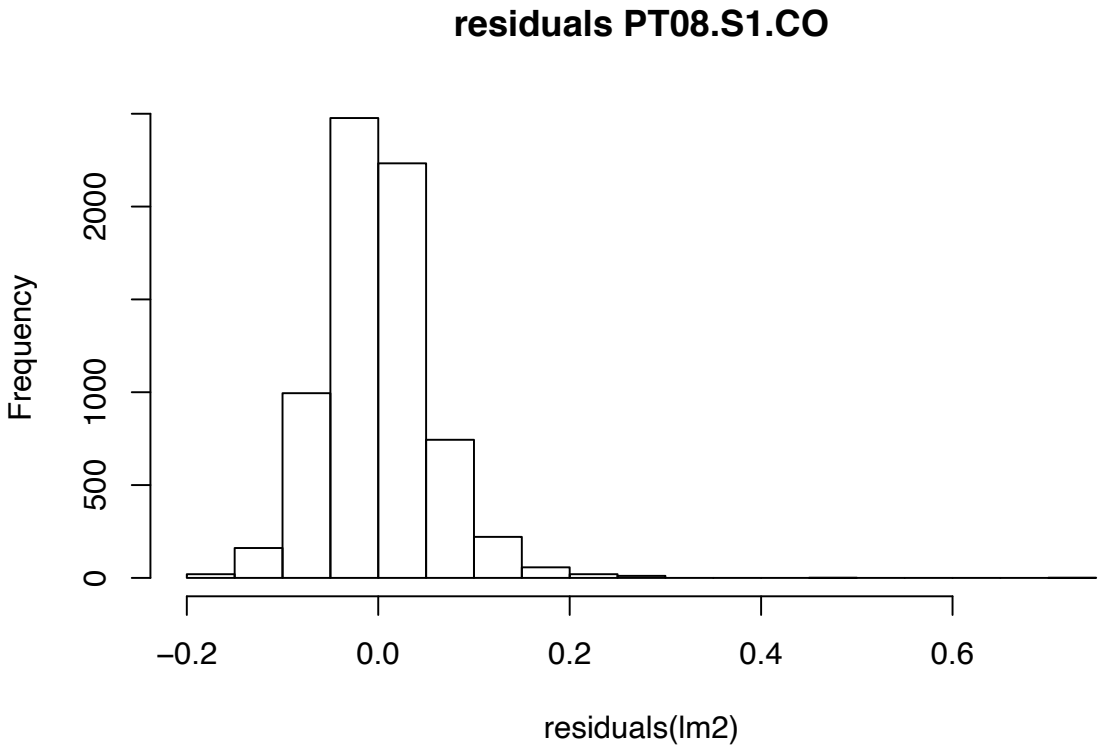

```r
#We can see some outliers in Residuals plot
plot(lm1,which=1)
```



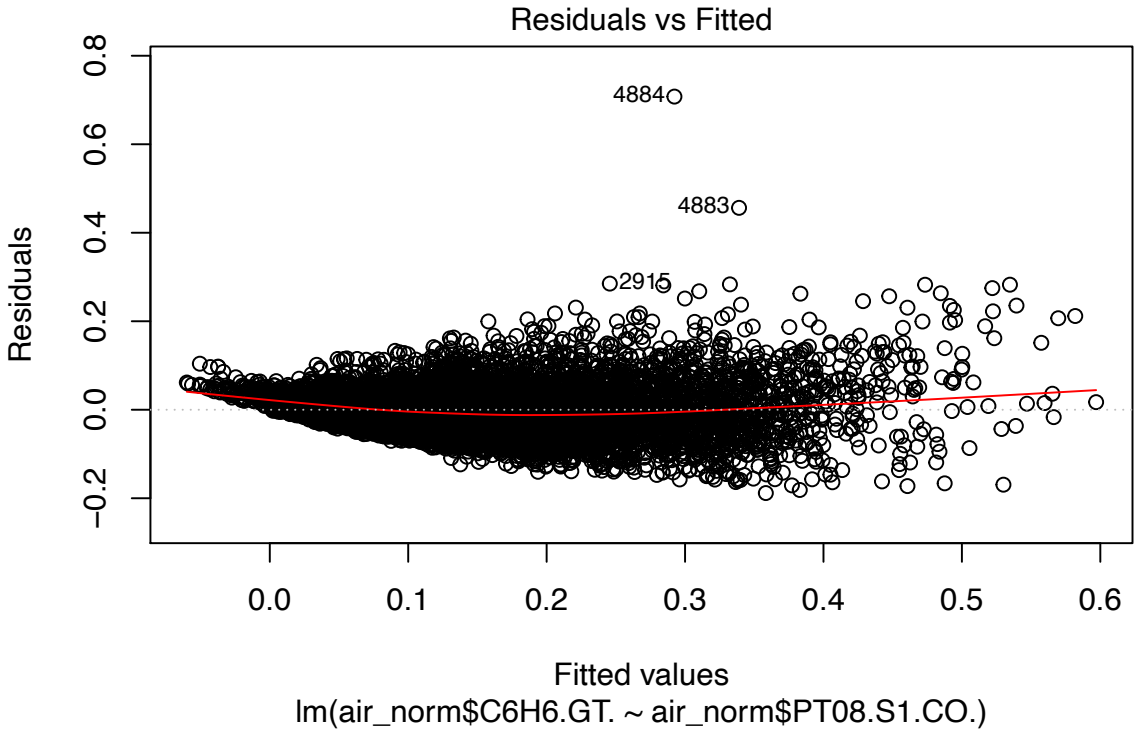

```
lm2 <- lm(air_norm$C6H6.GT. ~ air_norm$PT08.S1.CO.)
summary(lm2)
```

```
##
## Call:
## lm(formula = air_norm$C6H6.GT. ~ air_norm$PT08.S1.CO.)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18826 -0.03559 -0.00288  0.03079  0.70768
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.059959   0.001613  -37.18   <2e-16 ***
## air_norm$PT08.S1.CO.  0.656927   0.004312  152.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0564 on 6939 degrees of freedom
## Multiple R-squared:  0.7699, Adjusted R-squared:  0.7699
## F-statistic: 2.322e+04 on 1 and 6939 DF,  p-value: < 2.2e-16
```

```
hist(residuals(lm2),main ="residuals PT08.S1.CO")
```

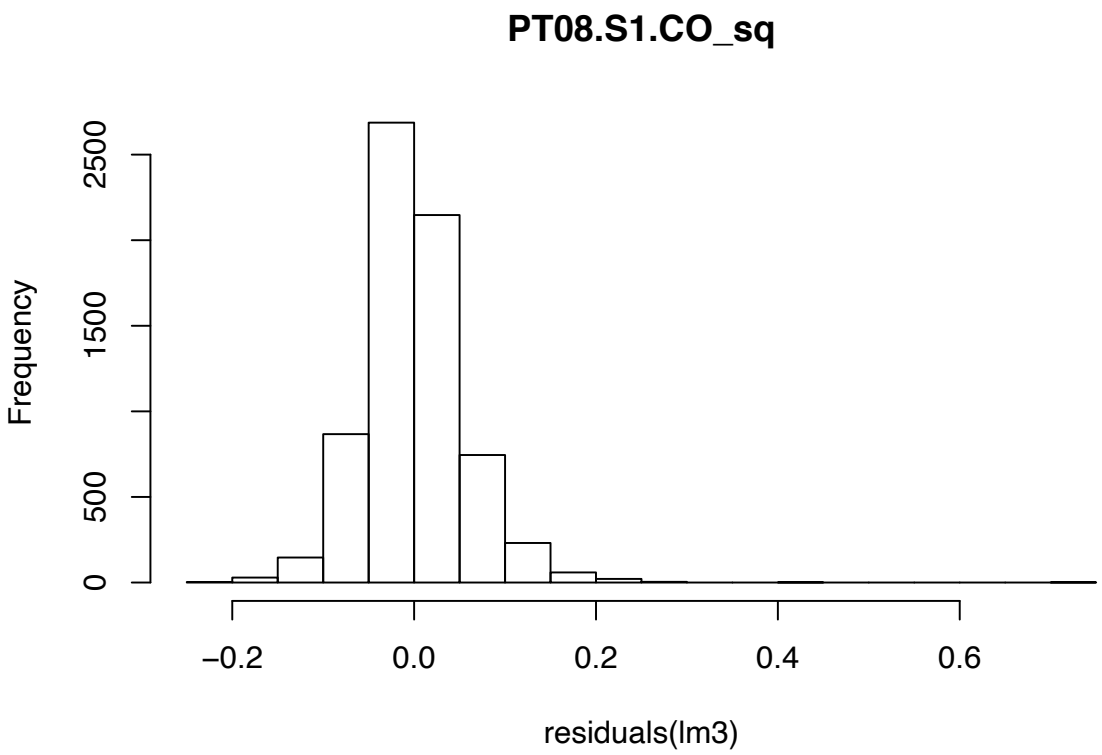



```
plot(lm2,which=1)
```

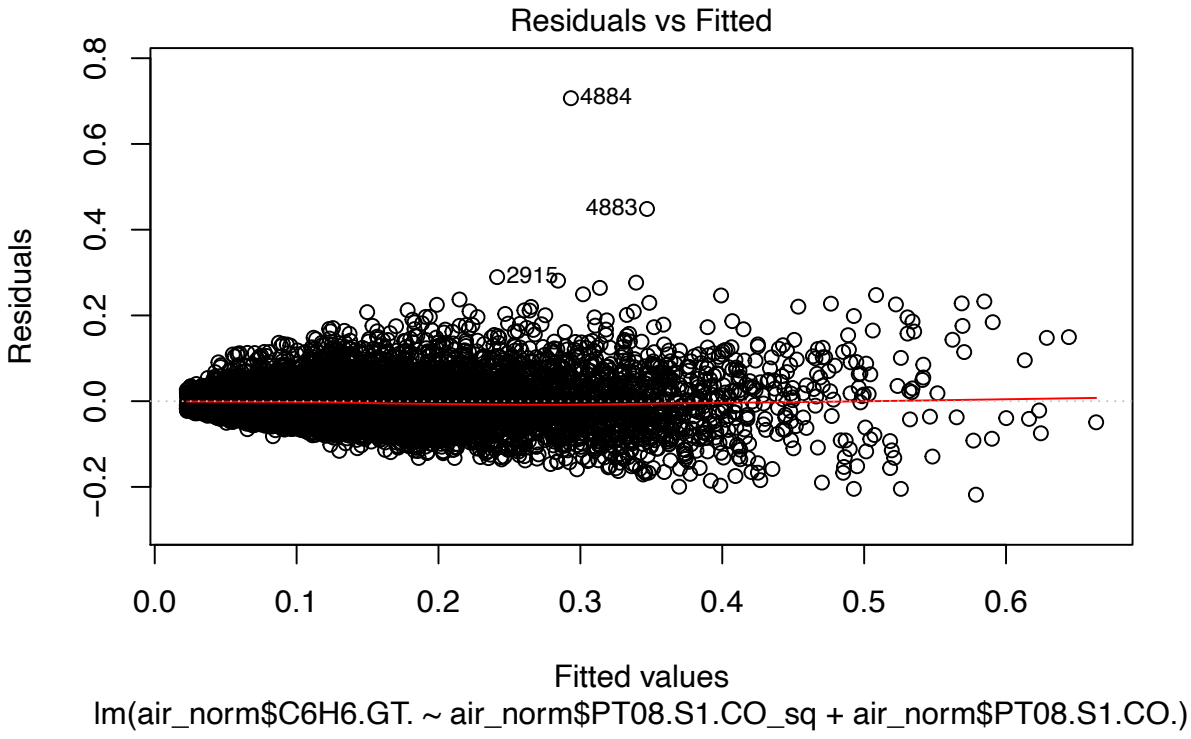


Data is normaly distributed but we have some outliers. We can try to transform our predictors using x^2, log or sqrt and see what will improve plot PT08.S1.CO sqrt

```r
air_norm$PT08.S1.CO_sq <- sqrt(air_norm$PT08.S1.CO.)
lm3 <- lm(air_norm$C6H6.GT. ~ air_norm$PT08.S1.CO_sq + air_norm$PT08.S1.CO.)
hist(residuals(lm3),main ="PT08.S1.CO_sq")
```




```r
plot(lm3,which=1)
```



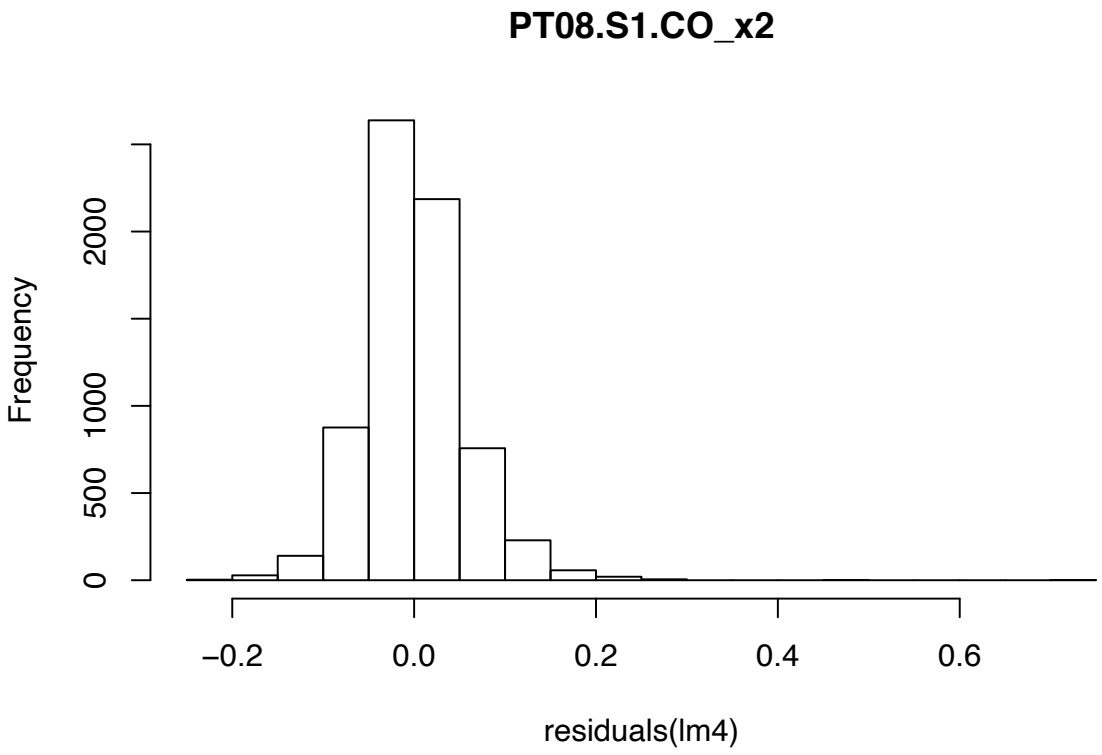
Check vif (vif >5: indicates multicollinearity)

```
vif(lm3)
```

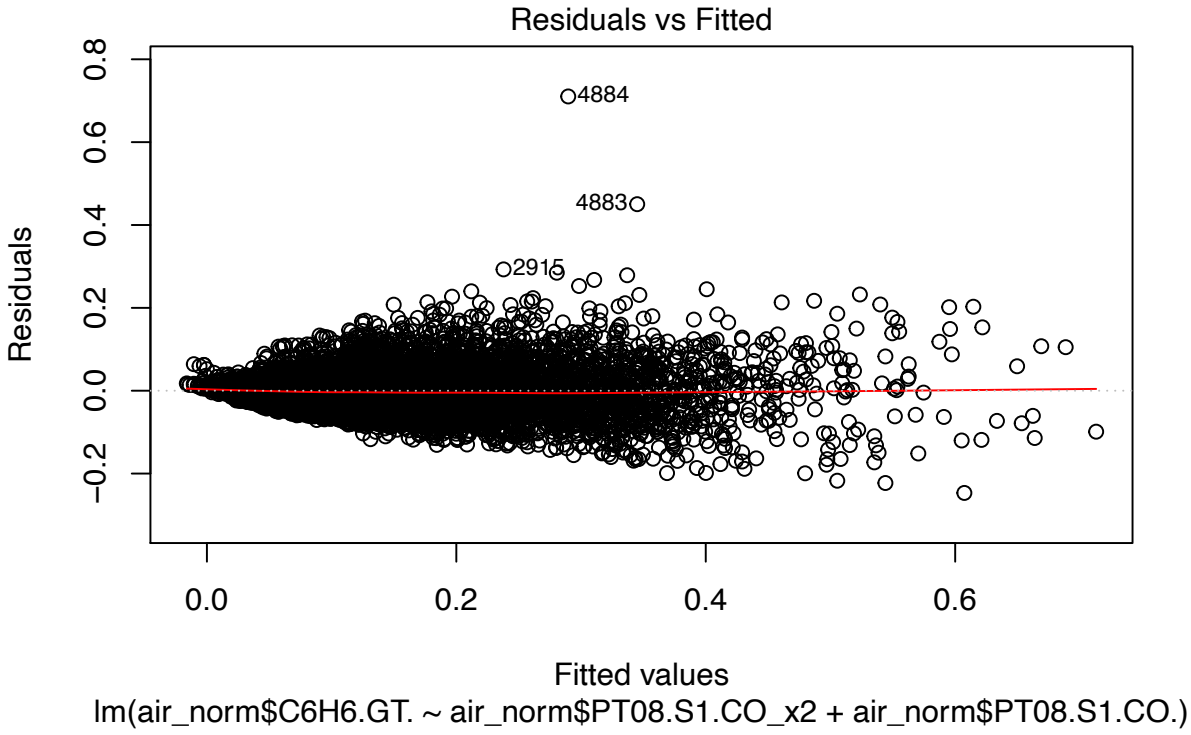
```
## air_norm$PT08.S1.CO_sq   air_norm$PT08.S1.CO.
##               40.63274                40.63274
```

PT08.S1.CO x^2

```
air_norm$PT08.S1.CO_x2 <- (air_norm$PT08.S1.CO.)^2
lm4 <- lm(air_norm$C6H6.GT. ~ air_norm$PT08.S1.CO_x2  + air_norm$PT08.S1.CO.)
hist(residuals(lm4),main ="PT08.S1.CO_x2")
```




```
plot(lm4,which=1)
```



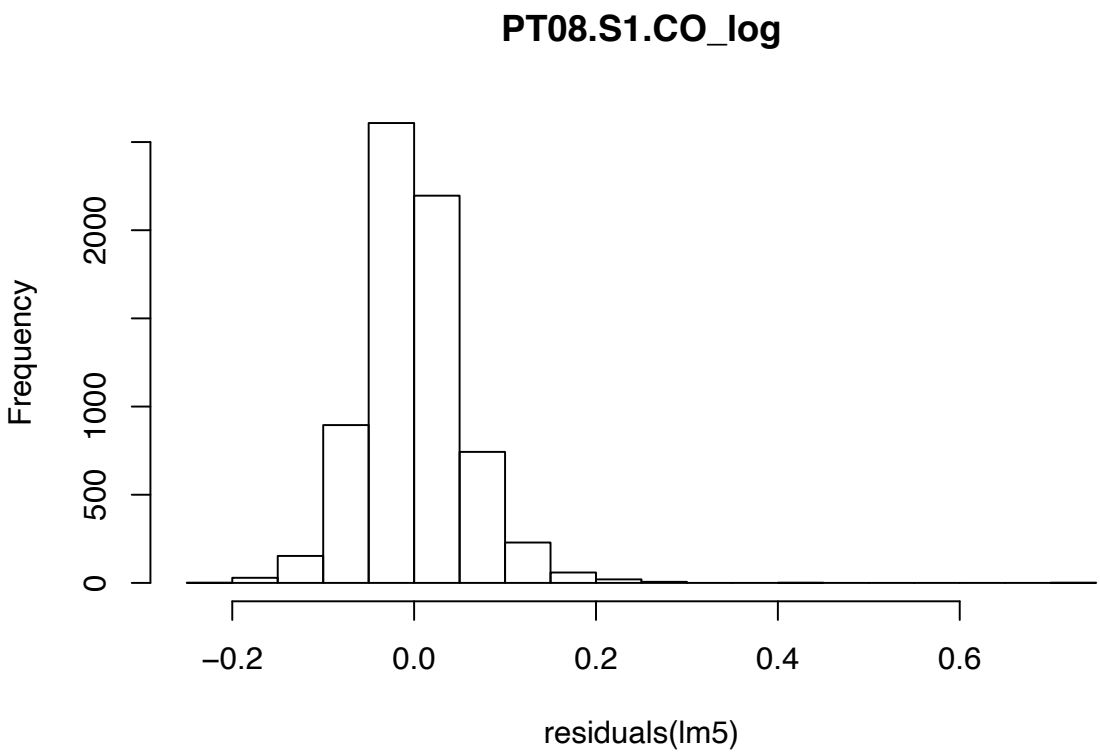
Check vif (vif >5: indicates multicollinearity) We can use x^2 transformation

```
vif(lm4)
```

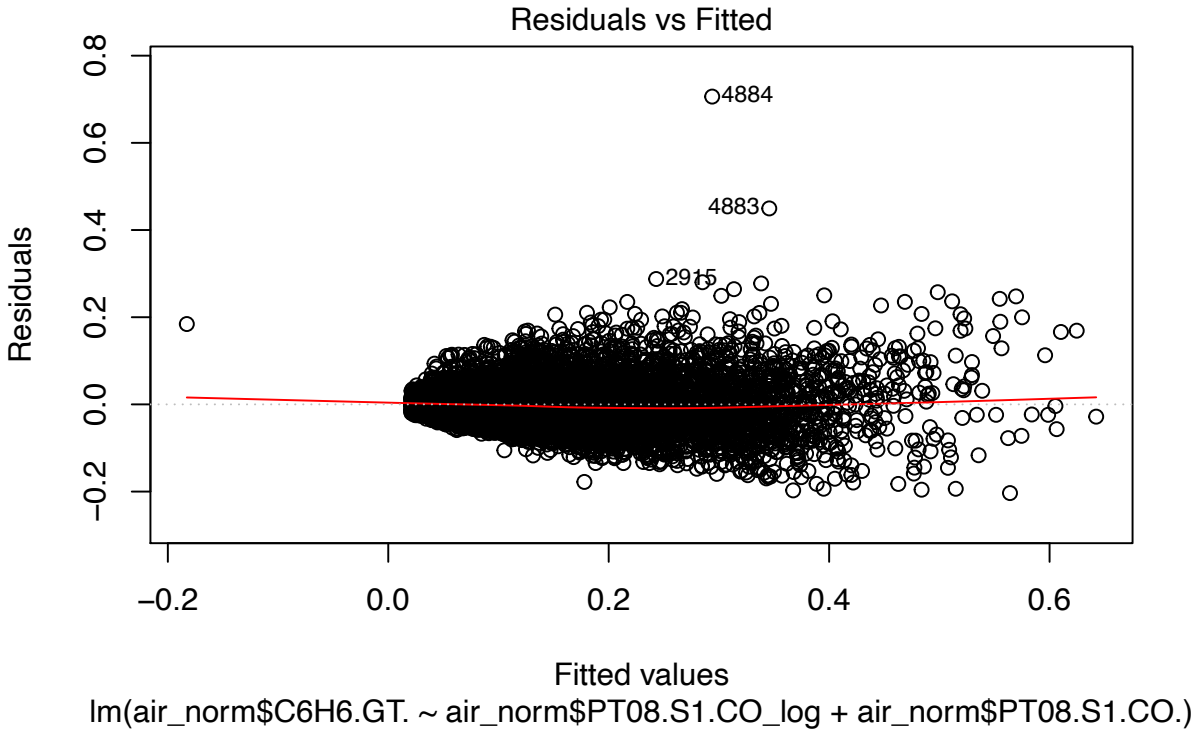
```
## air_norm$PT08.S1.CO_x2   air_norm$PT08.S1.CO.
##               15.30628                15.30628
```

PT08.S1.CO log(x)

```
air_norm$PT08.S1.CO_log <- log(air_norm$PT08.S1.CO.)
air_norm$PT08.S1.CO_log[!is.finite(air_norm$PT08.S1.CO_log)] <- 0
lm5 <- lm(air_norm$C6H6.GT. ~ air_norm$PT08.S1.CO_log  + air_norm$PT08.S1.CO.)
hist(residuals(lm5),main ="PT08.S1.CO_log")
```




```
plot(lm5,which=1)
```



Look on vif

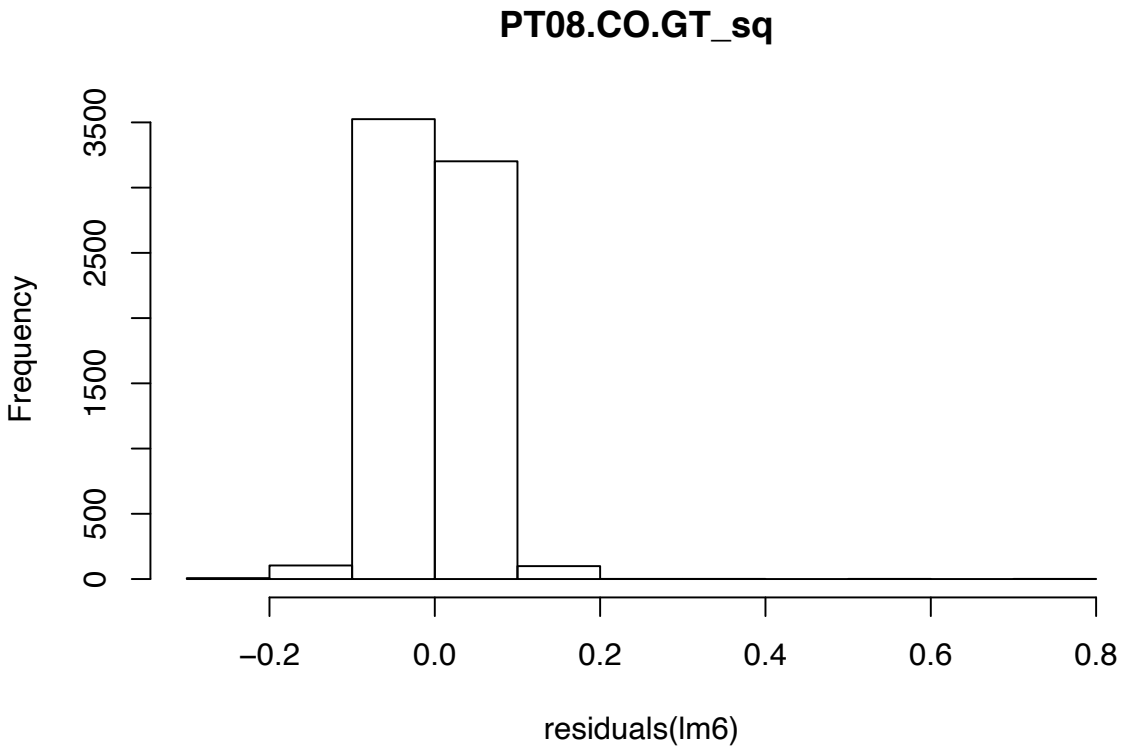
```
vif(lm5)
```

```
## air_norm$PT08.S1.CO_log     air_norm$PT08.S1.CO.
```

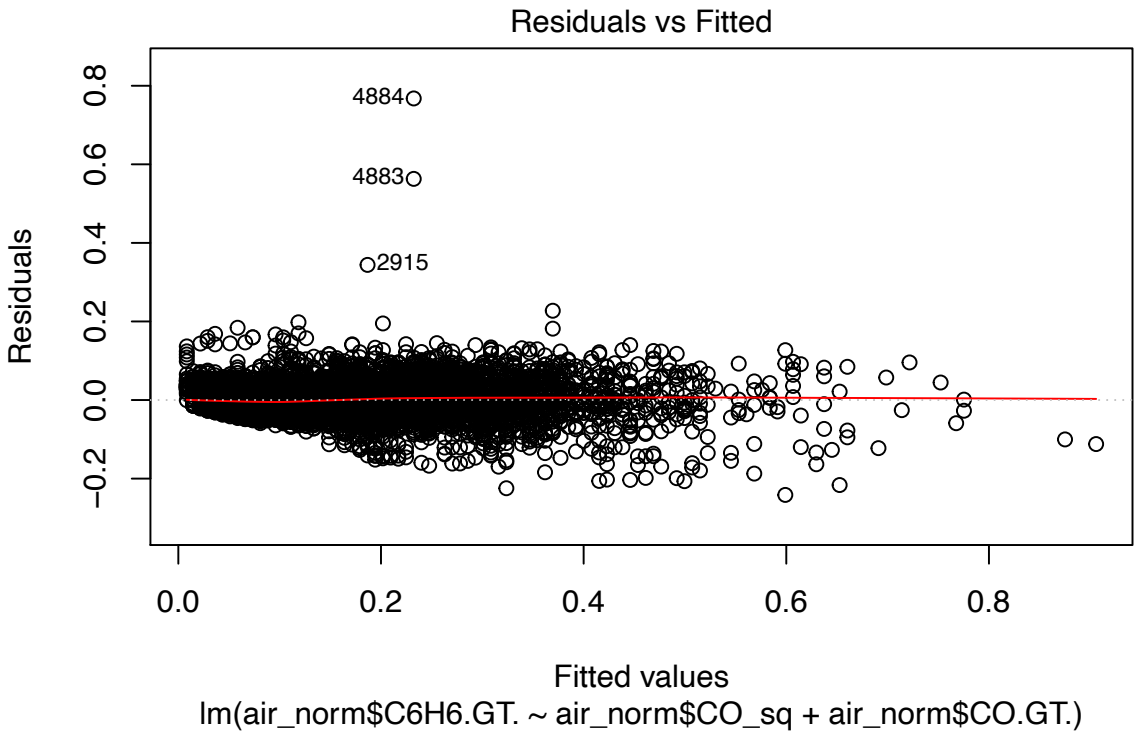
```
##                 8.073383                 8.073383
```

We can choose x^2 transformed predictor based on vif Let's look on the second one

```
air_norm$CO_sq <- sqrt(air_norm$CO.GT.)
lm6 <- lm(air_norm$C6H6.GT. ~ air_norm$CO_sq + air_norm$CO.GT.)
hist(residuals(lm6),main ="PT08.CO.GT_sq ")
```




```
plot(lm6,which=1)
```



Chieck

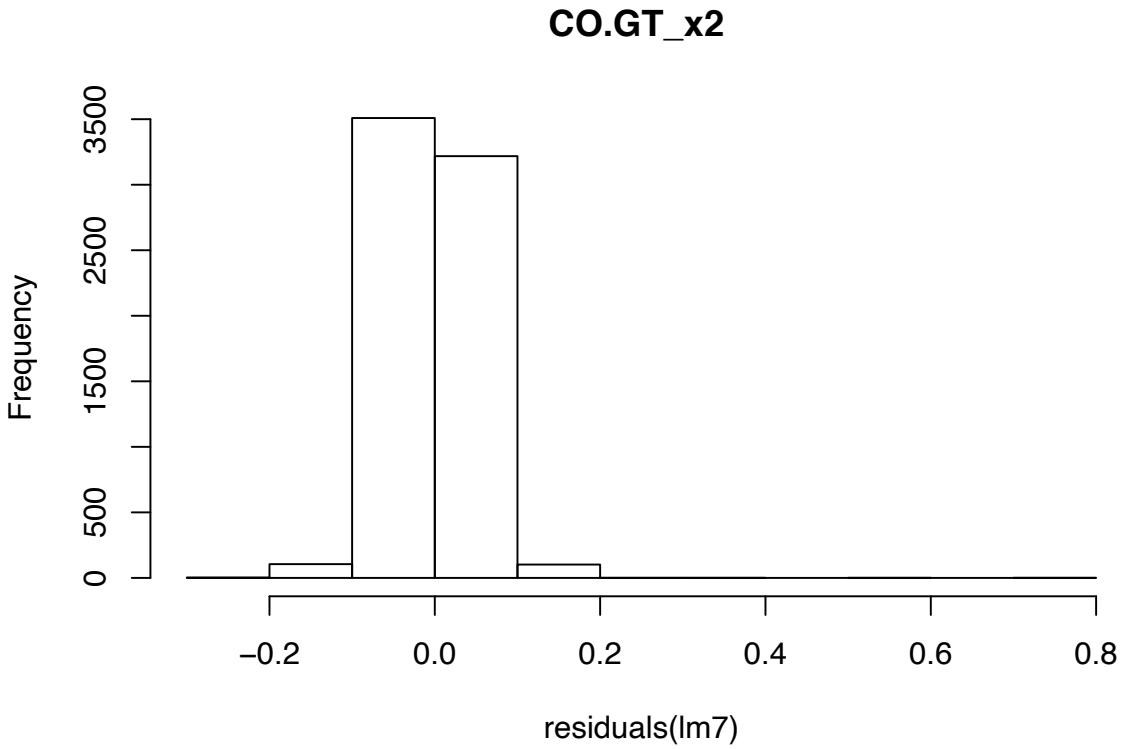
vif

```
vif(lm6)
```

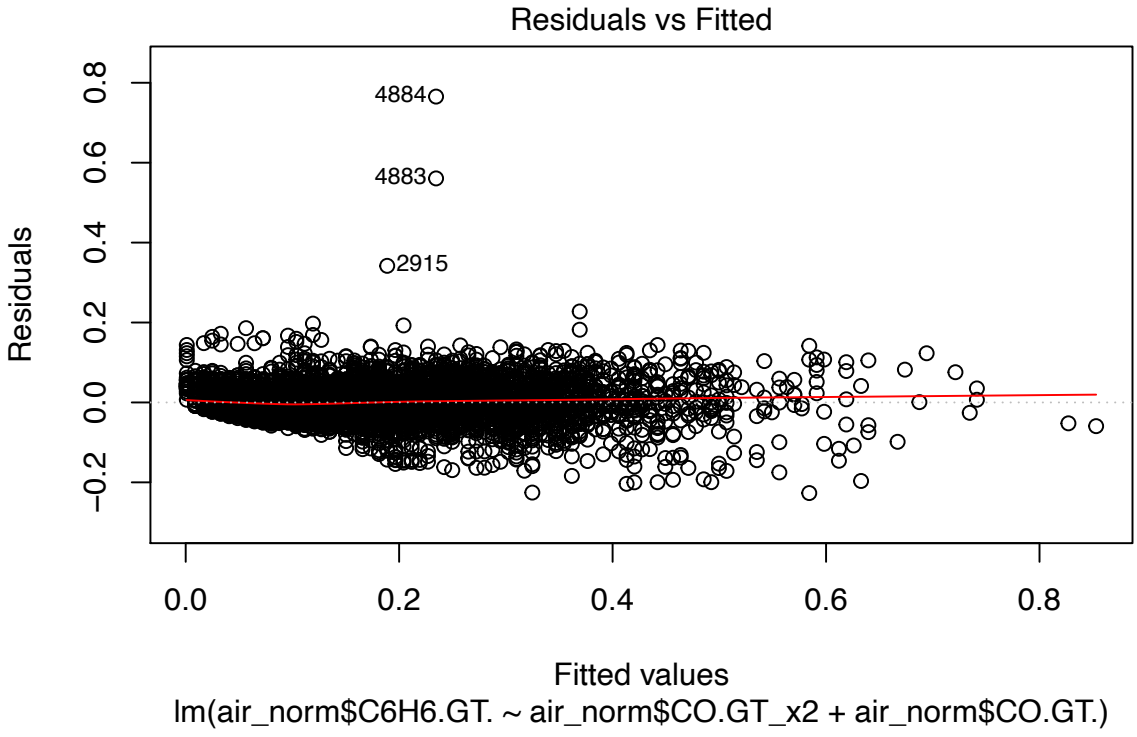
```
##  air_norm$CO_sq air_norm$CO.GT.
##        18.51628        18.51628
```

PT08.S1.CO x^2 (we cab use it)

```
air_norm$CO.GT_x2 <- (air_norm$CO.GT.)^2
lm7 <- lm(air_norm$C6H6.GT. ~ air_norm$CO.GT_x2 + air_norm$CO.GT.)
hist(residuals(lm7),main ="CO.GT_x2 ")
```




```
plot(lm7,which=1)
```



Chieck

vif

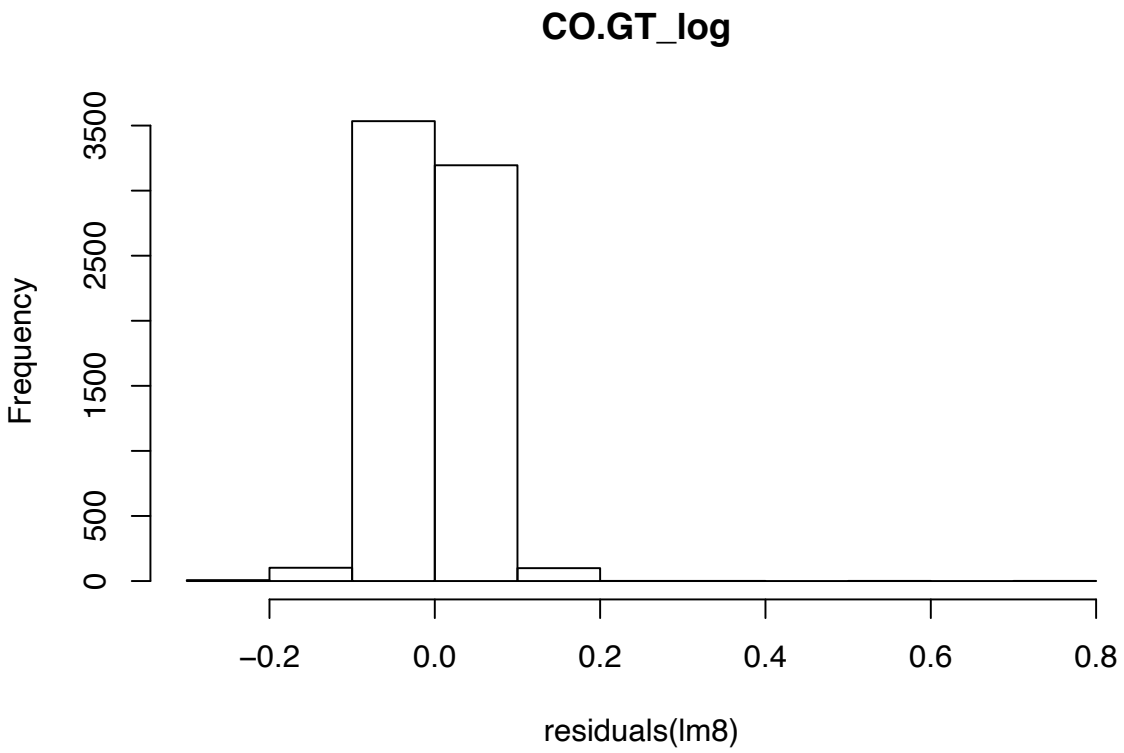
```
vif(lm7)
```

```
## air_norm$CO.GT_x2   air_norm$CO.GT.
```

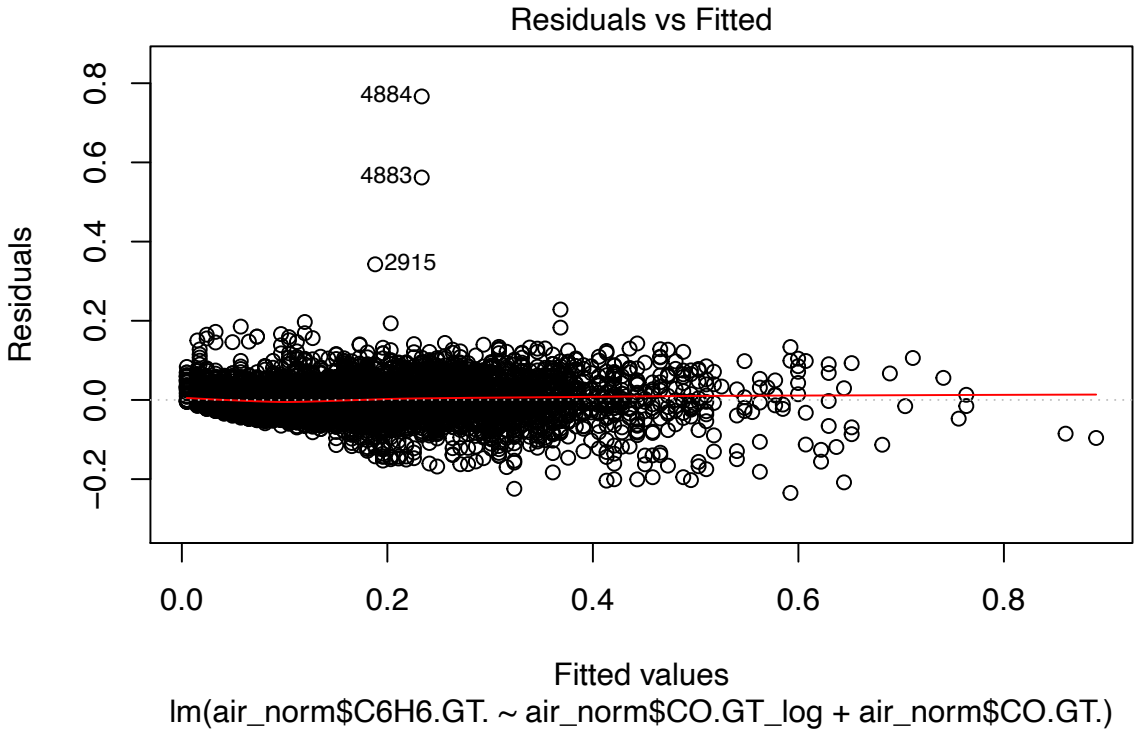
```
##          7.329655          7.329655
```

PT08.S1.CO log(x)

```
air_norm$CO.GT_log <- log(air_norm$CO.GT.)
air_norm$CO.GT_log[!is.finite(air_norm$CO.GT_log)] <- 0
lm8 <- lm(air_norm$C6H6.GT. ~ air_norm$CO.GT_log + air_norm$CO.GT.)
hist(residuals(lm8),main ="CO.GT_log ")
```




```
plot(lm8,which=1)
```



Chieck

vif

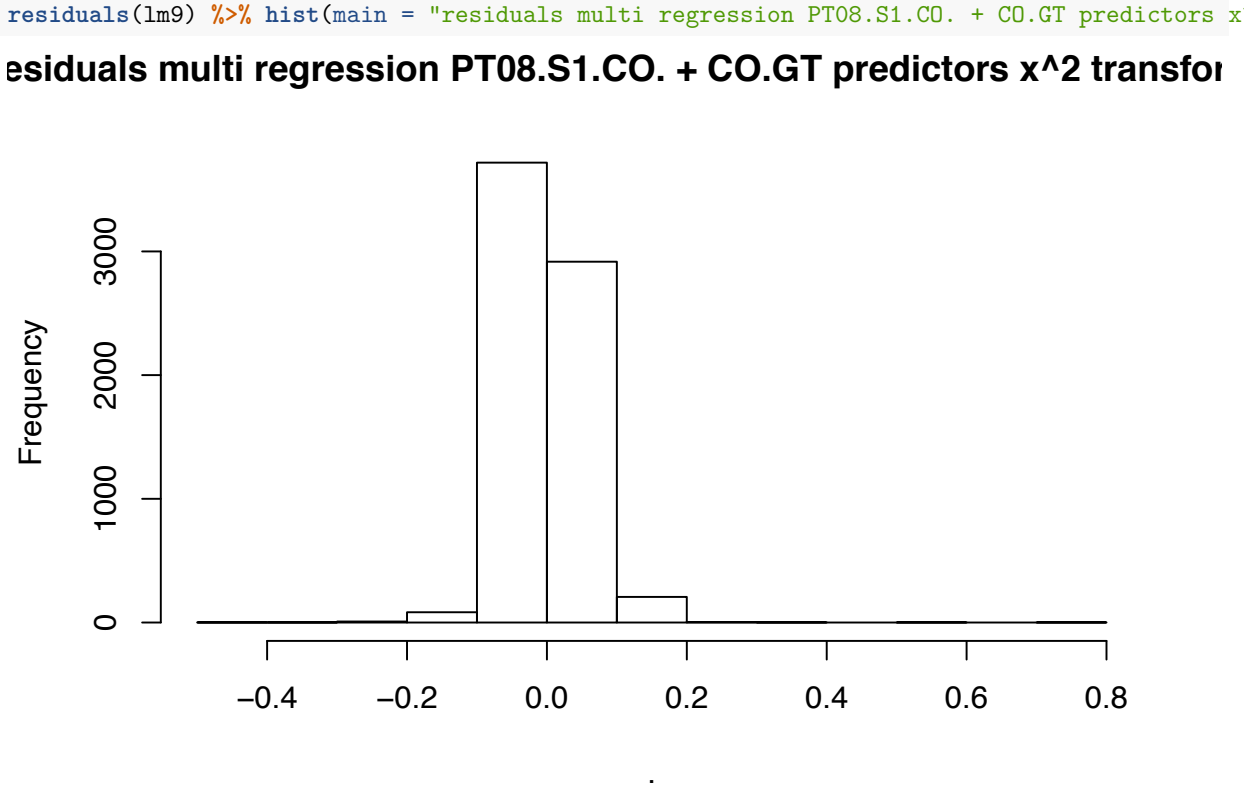
```
vif(lm8)
```

```
## air_norm$CO.GT_log    air_norm$CO.GT.
##            3.86905            3.86905
```

```
#Let's take predictors with r^2
lm9 <- lm(air_norm$C6H6.GT. ~ air_norm$PT08.S1.CO_x2 + air_norm$CO.GT_x2)
summary(lm9)
```

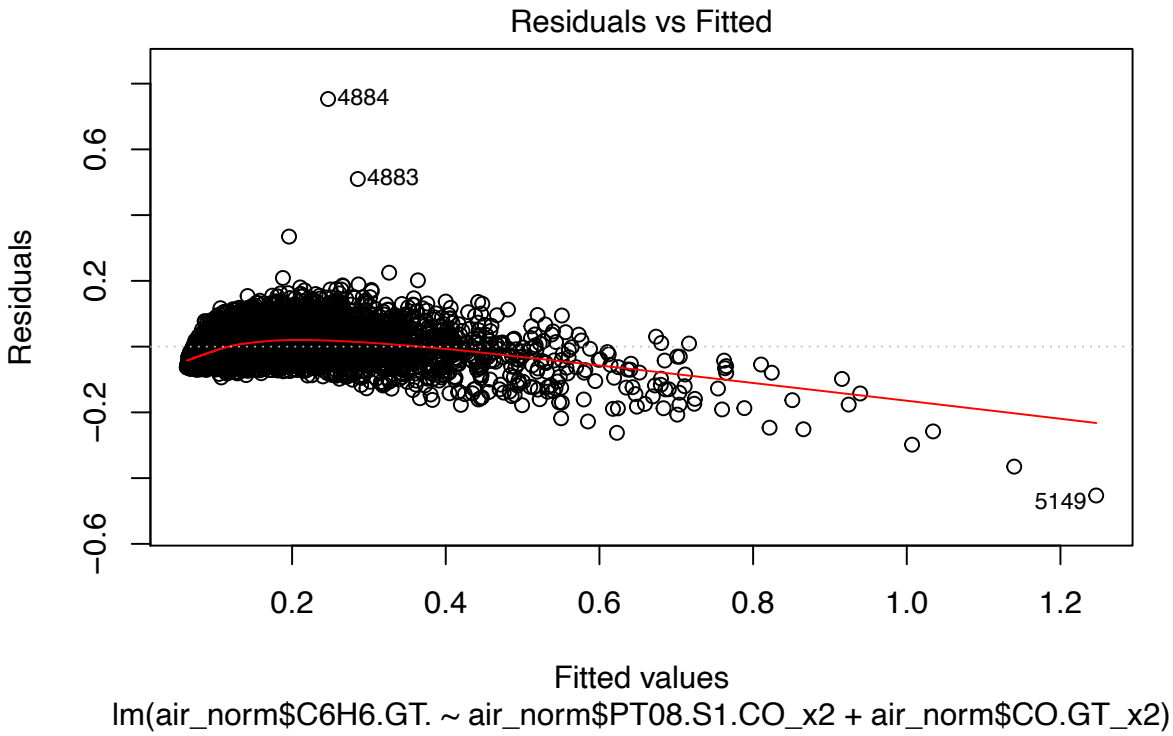
```
##
## Call:
## lm(formula = air_norm$C6H6.GT. ~ air_norm$PT08.S1.CO_x2 + air_norm$CO.GT_x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45282 -0.03458 -0.00584  0.02961  0.75316
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.0629180  0.0009415   66.83   <2e-16 ***
## air_norm$PT08.S1.CO_x2 0.4755961  0.0087834   54.15   <2e-16 ***
## air_norm$CO.GT_x2      0.7296029  0.0165848   43.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05077 on 6938 degrees of freedom
## Multiple R-squared:  0.8136, Adjusted R-squared:  0.8135
## F-statistic: 1.514e+04 on 2 and 6938 DF,  p-value: < 2.2e-16
```

```
residuals(lm9) %>% hist(main = "residuals multi regression PT08.S1.CO. + CO.GT predictors x^2 transfort
```



.

```
plot(lm9, which = 1)
```



vif < 5: Indicates that predictors are not redunant (not providing overlapping data to inform response)

```r
vif(lm9)
```

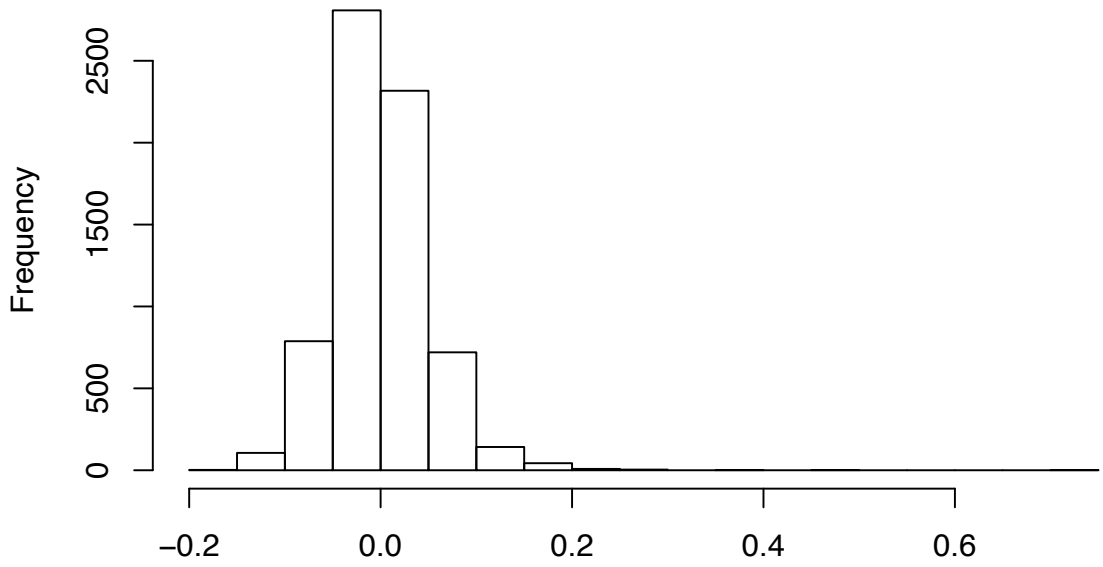
```
## air_norm$PT08.S1.CO_x2      air_norm$CO.GT_x2
##                3.41241                3.41241
```

```r
lm10 <- lm(air_norm$C6H6.GT. ~ air_norm$PT08.S1.CO. + air_norm$PT08.S4.NO2.)
summary(lm10)
```

```
##
## Call:
## lm(formula = air_norm$C6H6.GT. ~ air_norm$PT08.S1.CO. + air_norm$PT08.S4.NO2.)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15790 -0.03102 -0.00365  0.02645  0.72038
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -0.099796   0.001668  -59.81   <2e-16 ***
## air_norm$PT08.S1.CO.    0.499710   0.005141   97.20   <2e-16 ***
## air_norm$PT08.S4.NO2.   0.229913   0.005082   45.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04956 on 6938 degrees of freedom
## Multiple R-squared:  0.8223, Adjusted R-squared:  0.8223
## F-statistic: 1.605e+04 on 2 and 6938 DF,  p-value: < 2.2e-16
```

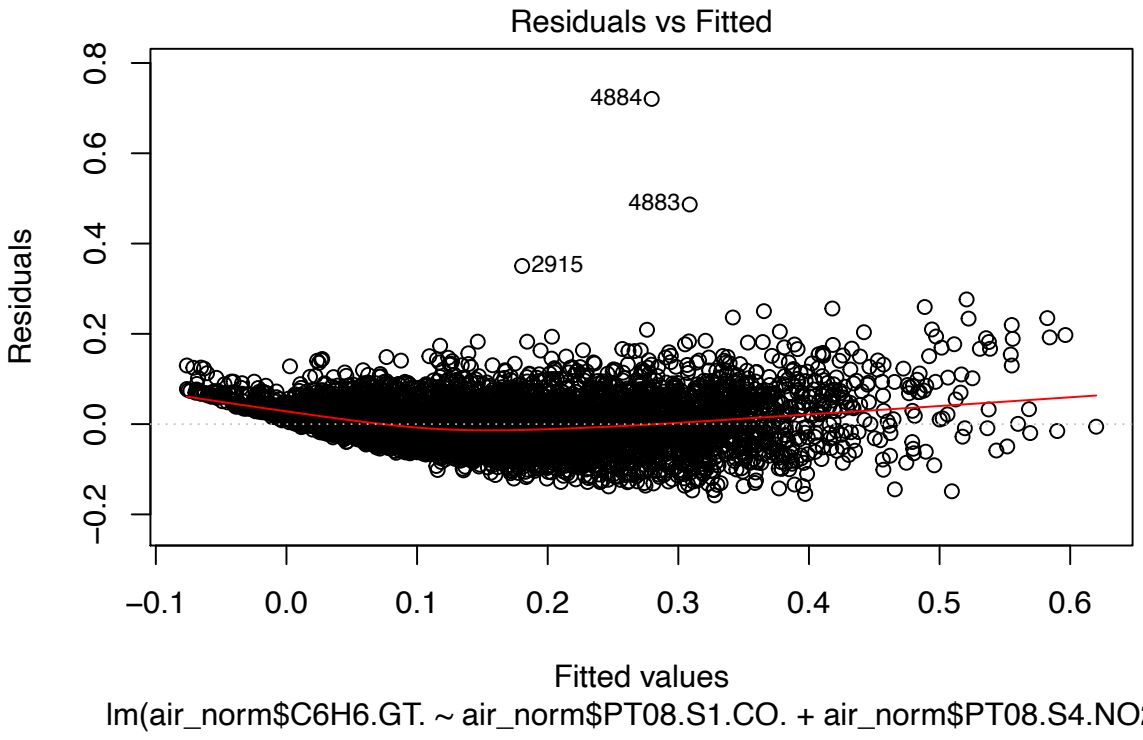
```r
residuals(lm10) %>% hist(main = "residuals multi regression PT08.S1.CO. + PT08.S4.NO2 predictors not tra
```

duals multi regression PT08.S1.CO. + PT08.S4.NO2 predictors not tran



.

```r
plot(lm10, which = 1)
```



We have good model with few number of outliers, high R^2 and low vif