HW2.2_Mary_Futey

Mary Futey

4/26/2020

Air Quality dataset

- Convert date and time to factor, check for NAs, switch comma to decimal

```
airq <- read.csv("/Users/maryfutey/desktop/AirQualityUCI/AirQualityUCI.csv",
                 header = TRUE,
                 dec=",")
#deleted column  NMHC.GT. as there were many "-200" values
airq <- airq[, -5]

summary(airq)
```

```
##          Date            Time          CO.GT.         PT08.S1.CO.
##  01/01/2005:  24   00.00.00: 390   Min.   :-200.00   Min.   :-200
##  01/02/2005:  24   01.00.00: 390   1st Qu.:   0.60   1st Qu.: 921
##  01/03/2005:  24   02.00.00: 390   Median :   1.50   Median :1053
##  01/04/2004:  24   03.00.00: 390   Mean   : -34.21   Mean   :1049
##  01/04/2005:  24   04.00.00: 390   3rd Qu.:   2.60   3rd Qu.:1221
##  01/05/2004:  24   05.00.00: 390   Max.   :  11.90   Max.   :2040
##  (Other)   :9213   (Other) :7017
##     C6H6.GT.        PT08.S2.NMHC.       NOx.GT.        PT08.S3.NOx.
##  Min.   :-200.000   Min.   :-200.0   Min.   :-200.0   Min.   :-200
##  1st Qu.:   4.000   1st Qu.: 711.0   1st Qu.:  50.0   1st Qu.: 637
##  Median :   7.900   Median : 895.0   Median : 141.0   Median : 794
##  Mean   :   1.866   Mean   : 894.6   Mean   : 168.6   Mean   : 795
##  3rd Qu.:  13.600   3rd Qu.:1105.0   3rd Qu.: 284.0   3rd Qu.: 960
##  Max.   :  63.700   Max.   :2214.0   Max.   :1479.0   Max.   :2683
##
##     NO2.GT.         PT08.S4.NO2.   PT08.S5.O3.           T
##  Min.   :-200.00   Min.   :-200   Min.   :-200.0   Min.   :-200.000
##  1st Qu.:  53.00   1st Qu.:1185   1st Qu.: 700.0   1st Qu.:  10.900
##  Median :  96.00   Median :1446   Median : 942.0   Median :  17.200
##  Mean   :  58.15   Mean   :1391   Mean   : 975.1   Mean   :   9.778
##  3rd Qu.: 133.00   3rd Qu.:1662   3rd Qu.:1255.0   3rd Qu.:  24.100
##  Max.   : 340.00   Max.   :2775   Max.   :2523.0   Max.   :  44.600
##
##        RH                 AH
##  Min.   :-200.00   Min.   :-200.0000
##  1st Qu.:  34.10   1st Qu.:   0.6923
##  Median :  48.60   Median :   0.9768
##  Mean   :  39.49   Mean   :  -6.8376
##  3rd Qu.:  61.90   3rd Qu.:   1.2962
##  Max.   :  88.70   Max.   :   2.2310
```

```
##
```

```
airq_long <- gather(airq, 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)
```

- Clean and remove "-200" values as they are not possible / erroneous

```
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
```

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

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

- Need to normalize data

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

norm <- function(x) {
    (x - min(x)) / (max(x) - min(x))
  }

airq_norm <- as.data.frame(lapply(airq_wide[3:14], norm))
airq_norm$Date <- airq_wide$Date
airq_norm$Time <- airq_wide$Time

airq_norm_long <- gather(airq_norm, key = "measurement", value = "value", -c(Date, Time))
```

#Explore multicollinearity

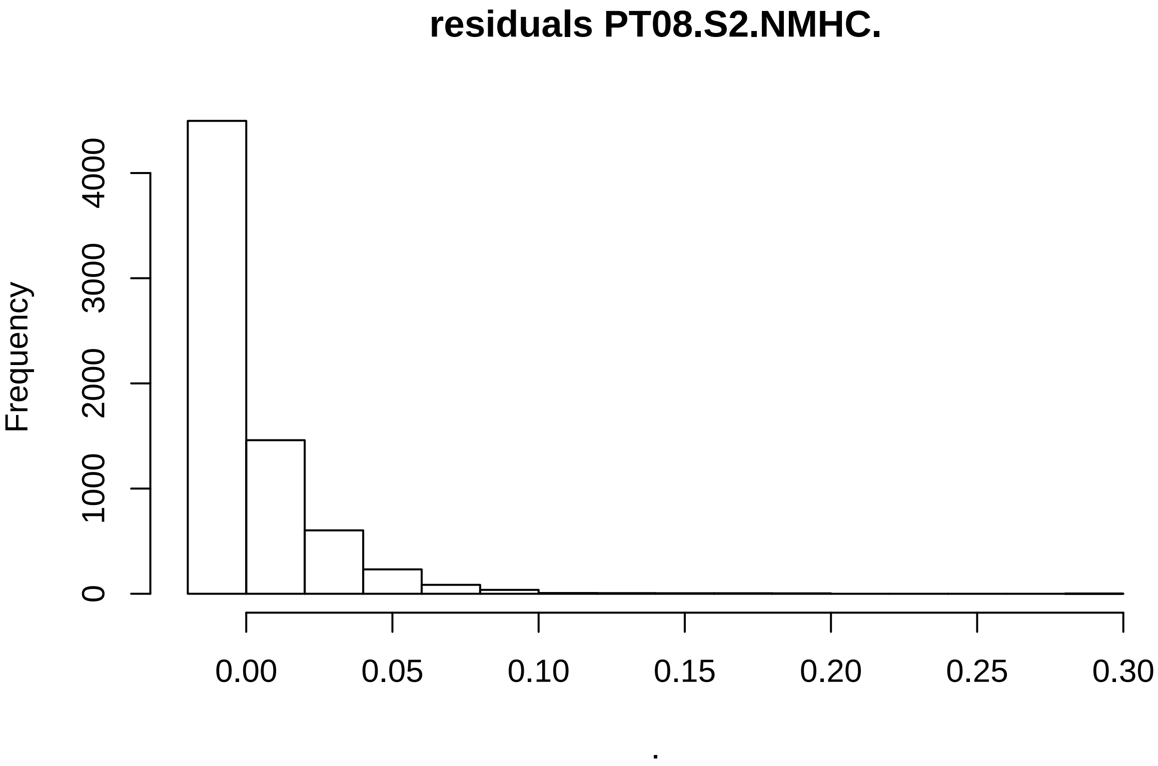
*Choose good predictors

*Check residuals

*Do you need some non-linear transformations for some of predictors? Or maybe response? +log(x) +sqrt(x) +(x)^2

** selected two predictor with high R^2, but non-linear tp transform: PT08.S2.NMHC.

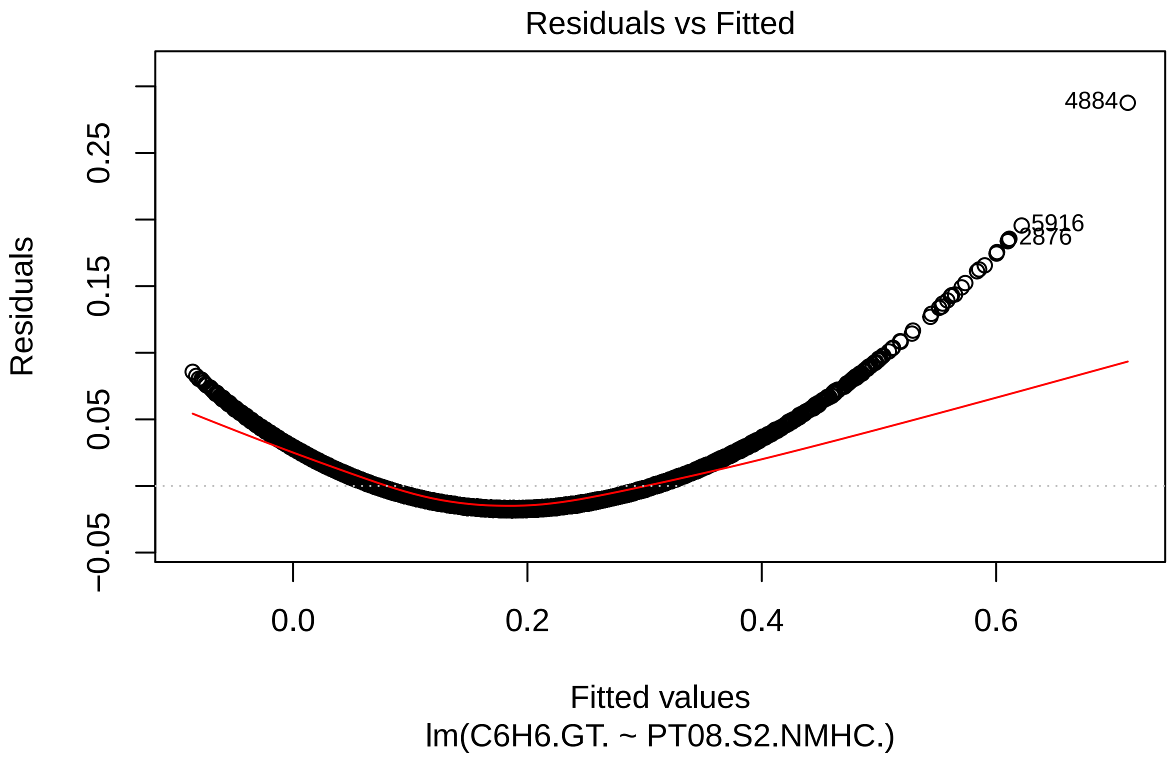
```
# PT08.S2.NMHC.: residuals look better after quadratic transformtion,
#however residuals vs fitted plot cubic
# R^2 of 0.999, how so high?
lm_3 <- lm(C6H6.GT. ~ PT08.S2.NMHC., airq_norm)
residuals(lm_3) %>% hist(main = "residuals PT08.S2.NMHC.")
```



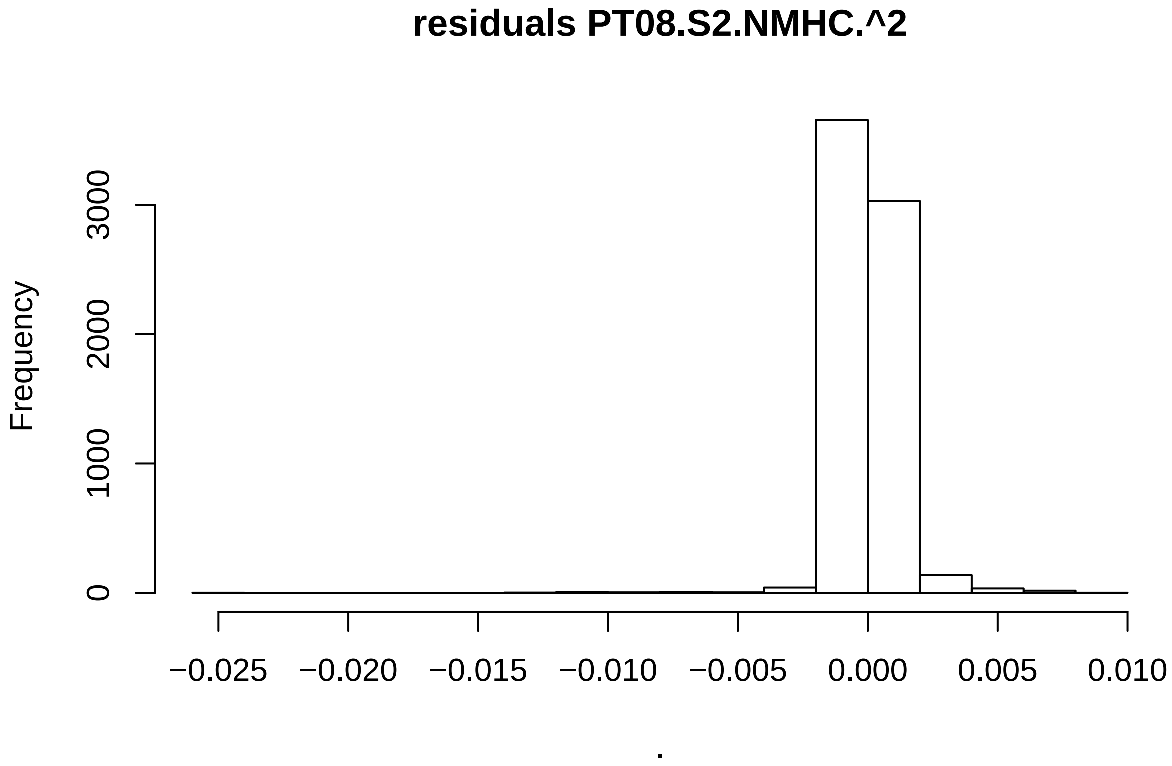
```r
summary(lm_3)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S2.NMHC., data = airq_norm)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.018385 -0.015140 -0.007871  0.007973  0.287650
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.0856861  0.0006204  -138.1   <2e-16 ***
## PT08.S2.NMHC.  0.7980363  0.0018053   442.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02177 on 6939 degrees of freedom
## Multiple R-squared:  0.9657, Adjusted R-squared:  0.9657
## F-statistic: 1.954e+05 on 1 and 6939 DF,  p-value: < 2.2e-16
```

```r
plot(lm_3, which = 1)
```



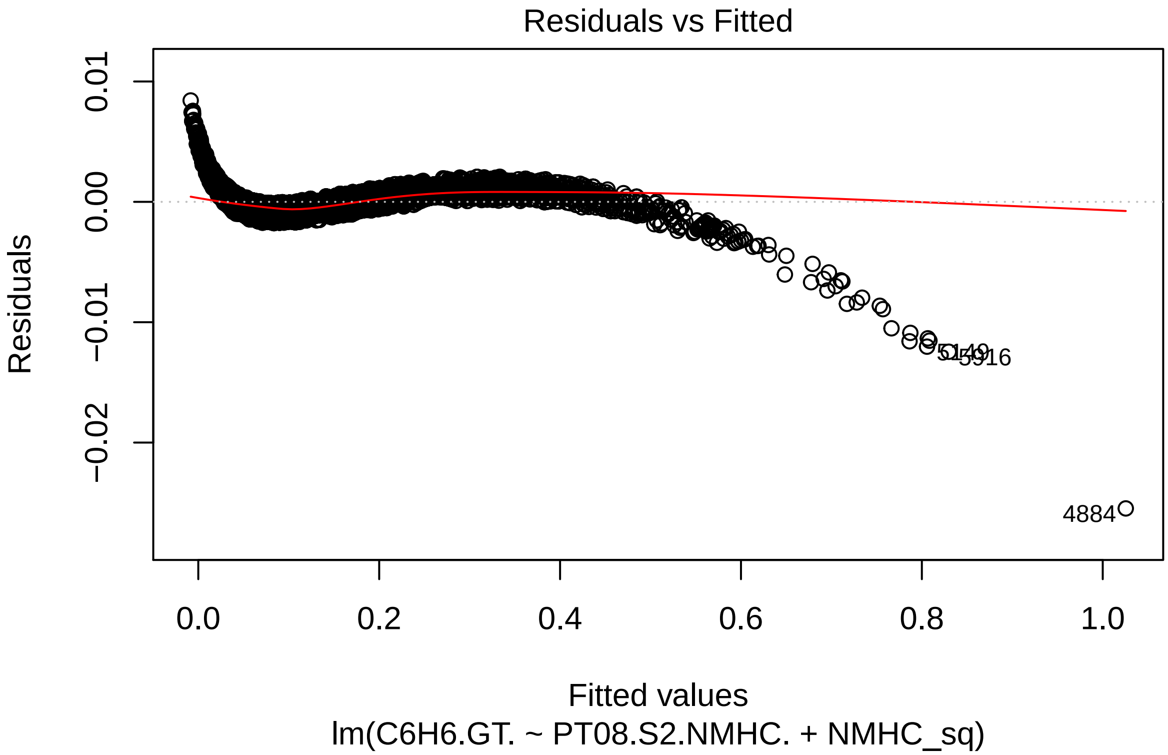
```
airq_norm$NMHC_sq <- (airq_norm$PT08.S2.NMHC.)^2

lm_4 <- lm(C6H6.GT. ~ PT08.S2.NMHC. + NMHC_sq, airq_norm)
residuals(lm_4) %>% hist(main = "residuals PT08.S2.NMHC.^2")
```



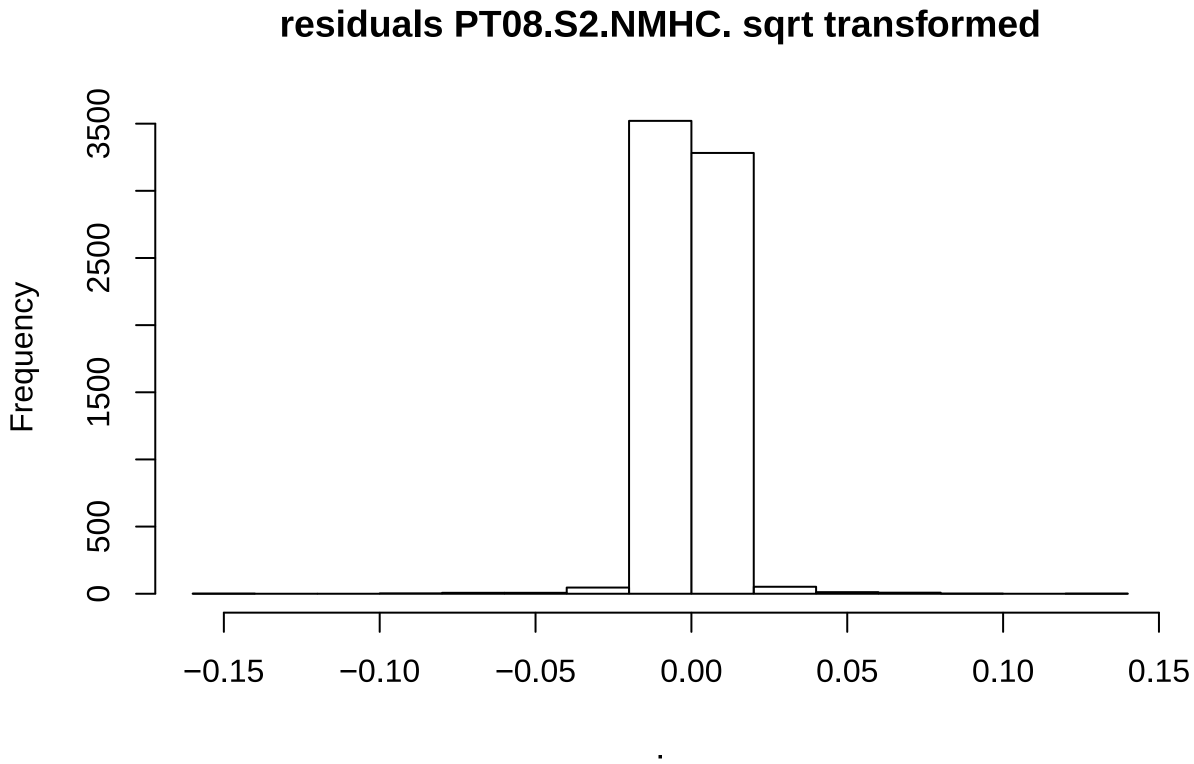
```
summary(lm_4)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S2.NMHC. + NMHC_sq, data = airq_norm)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0254697 -0.0007130 -0.0000817  0.0006329  0.0084216
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.422e-03  6.223e-05  -135.3   <2e-16 ***
## PT08.S2.NMHC.  2.550e-01  3.773e-04   675.8   <2e-16 ***
## NMHC_sq        7.789e-01  5.217e-04  1493.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001213 on 6938 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 3.26e+07 on 2 and 6938 DF,  p-value: < 2.2e-16
```

```
plot(lm_4, which = 1)
```



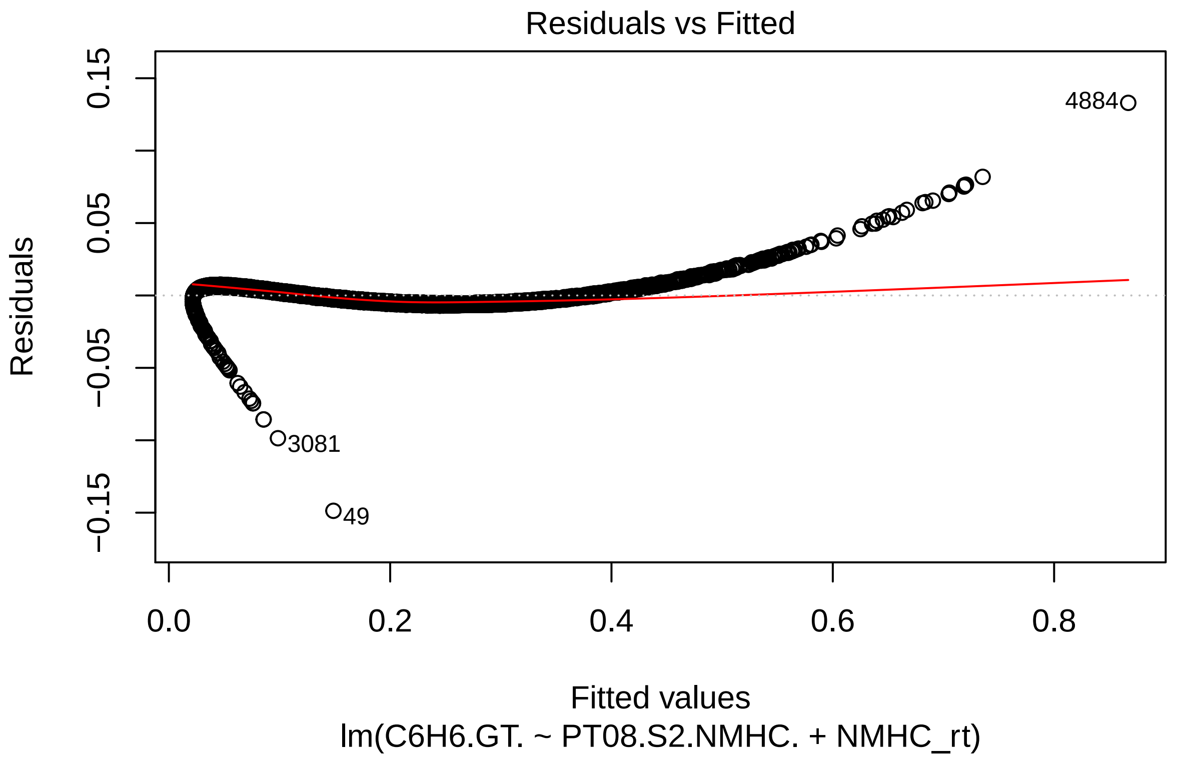
```
airq_norm$NMHC_rt <- sqrt(airq_norm$PT08.S2.NMHC.)

lm_2 <- lm(C6H6.GT. ~ PT08.S2.NMHC. + NMHC_rt, airq_norm)
residuals(lm_2) %>% hist(main = "residuals PT08.S2.NMHC. sqrt transformed")
```



```r
summary(lm_2)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S2.NMHC. + NMHC_rt, data = airq_norm)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.148700 -0.004664 -0.000351  0.004615  0.132983
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.148700   0.001115   133.3   <2e-16 ***
## PT08.S2.NMHC.  1.628181   0.003924   414.9   <2e-16 ***
## NMHC_rt       -0.909864   0.004241  -214.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00788 on 6938 degrees of freedom
## Multiple R-squared:  0.9955, Adjusted R-squared:  0.9955
## F-statistic: 7.688e+05 on 2 and 6938 DF,  p-value: < 2.2e-16
```

```r
plot(lm_2, which = 1)
```



Make a coorelation matrix to select predictors to use in multiple linear regression

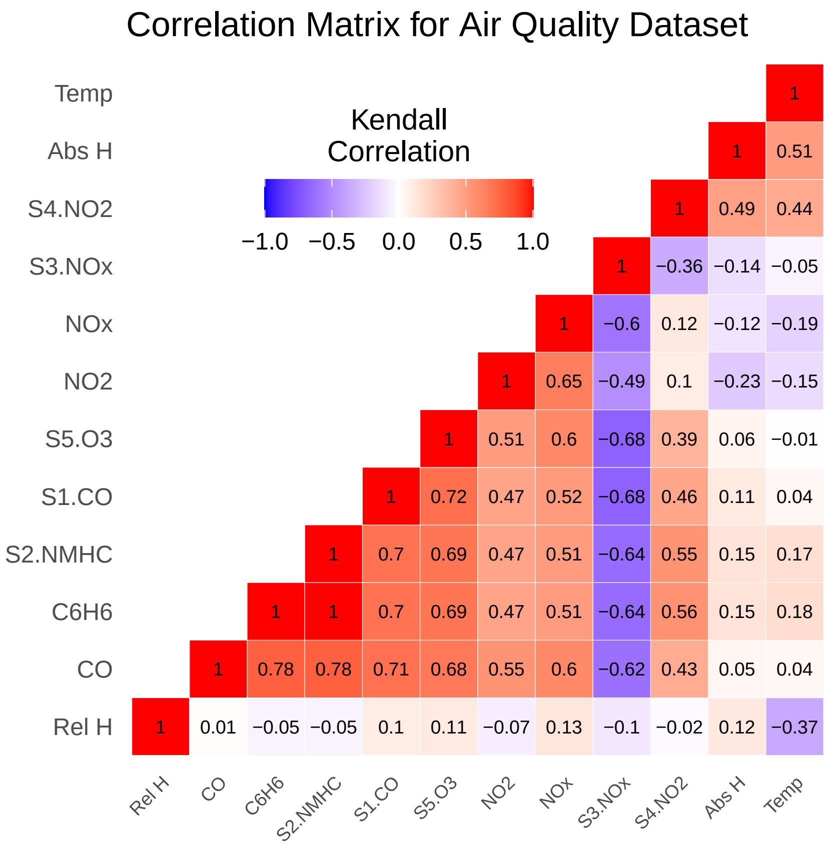
- due to deletion of NHMC column, matrix is differnet from the previous homework

```
airq_sub <- airq_norm[, 1:12]
colnames(airq_sub) <- c(
  "Abs H",
  "C6H6",
  "CO",
  "NO2",
  "NOx",
  "S1.CO",
  "S2.NMHC",
  "S3.NOx",
  "S4.NO2",
  "S5.O3",
  "Rel H",
  "Temp"
)

airq_cor <- round(cor(airq_sub, method = "kendall"),2)

#reorder, create upper triangle
reorder_airq_cor <- function(airq_cor){
  # Use correlation between variables as distance
  dd <- as.dist((1-airq_cor)/2)
  hc <- hclust(dd)
  airq_cor <-airq_cor[hc$order, hc$order]
}
```

```r
get_upper_tri <- function(airq_cor){
  airq_cor[lower.tri(airq_cor)]<- NA
  return(airq_cor)
}

airq_cor <- reorder_airq_cor(airq_cor)
upper_tri <- get_upper_tri(airq_cor)
melt_uppertri <- melt(upper_tri, na.rm = TRUE)
```

```
## Warning in melt(upper_tri, na.rm = TRUE): The melt generic in data.table has
## been passed a matrix and will attempt to redirect to the relevant reshape2
## method; please note that reshape2 is deprecated, and this redirection is now
## deprecated as well. To continue using melt methods from reshape2 while both
## libraries are attached, e.g. melt.list, you can prepend the namespace like
## reshape2::melt(upper_tri). In the next version, this warning will become an
## error.
```

```r
p1 <- ggplot(melt_uppertri,
             aes(Var2, Var1,
                 fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Kendall\nCorrelation") +
  ggtitle("Correlation Matrix for Air Quality Dataset") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 7, hjust = 1)) +
  coord_fixed() +
  geom_text(aes(Var2,
                Var1,
                label = value),
            color = "black", size = 2.6) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))

p1
```

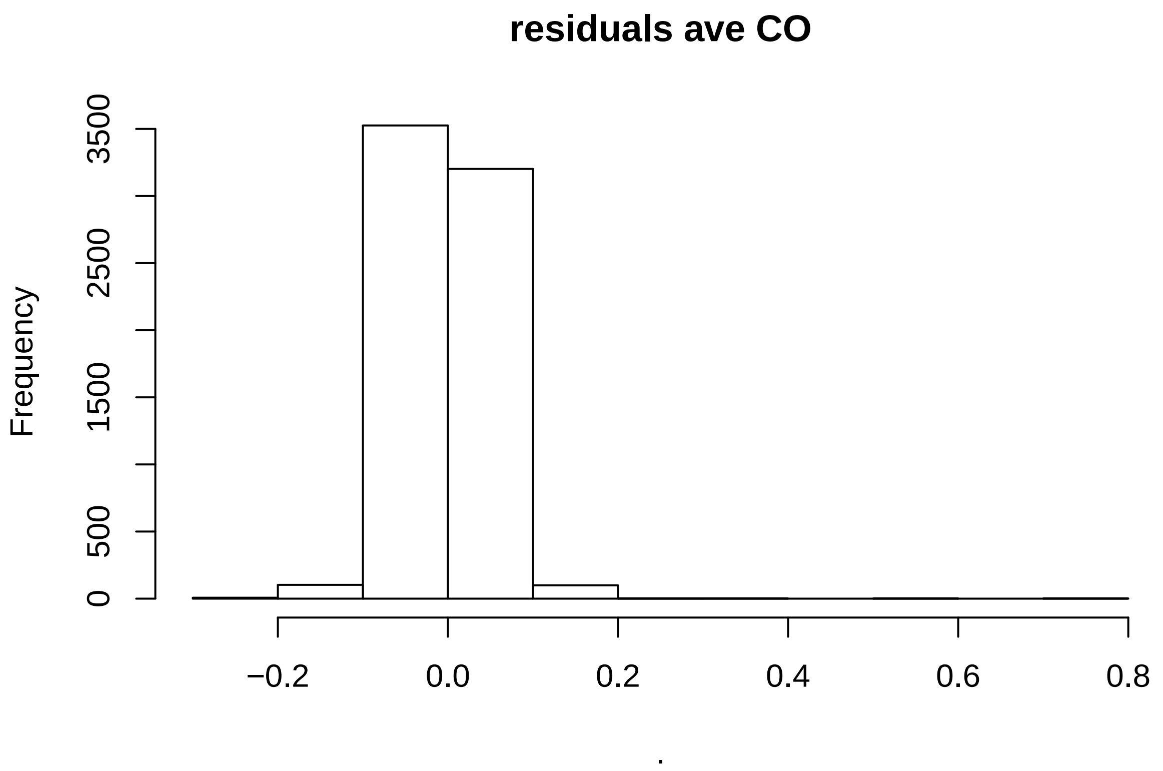


	Rel H	CO	C6H6	S2.NMHC	S1.CO	S5.O3	NO2	NOx	S3.NOx	S4.NO2	Abs H	Temp
Temp												1
Abs H											1	0.51
S4.NO2										1	0.49	0.44
S3.NOx									1	−0.36	−0.14	−0.05
NOx								1	−0.6	0.12	−0.12	−0.19
NO2							1	0.65	−0.49	0.1	−0.23	−0.15
S5.O3						1	0.51	0.6	−0.68	0.39	0.06	−0.01
S1.CO					1	0.72	0.47	0.52	−0.68	0.46	0.11	0.04
S2.NMHC				1	0.7	0.69	0.47	0.51	−0.64	0.55	0.15	0.17
C6H6			1	1	0.7	0.69	0.47	0.51	−0.64	0.56	0.15	0.18
CO		1	0.78	0.78	0.71	0.68	0.55	0.6	−0.62	0.43	0.05	0.04
Rel H	1	0.01	−0.05	−0.05	0.1	0.11	−0.07	0.13	−0.1	−0.02	0.12	−0.37

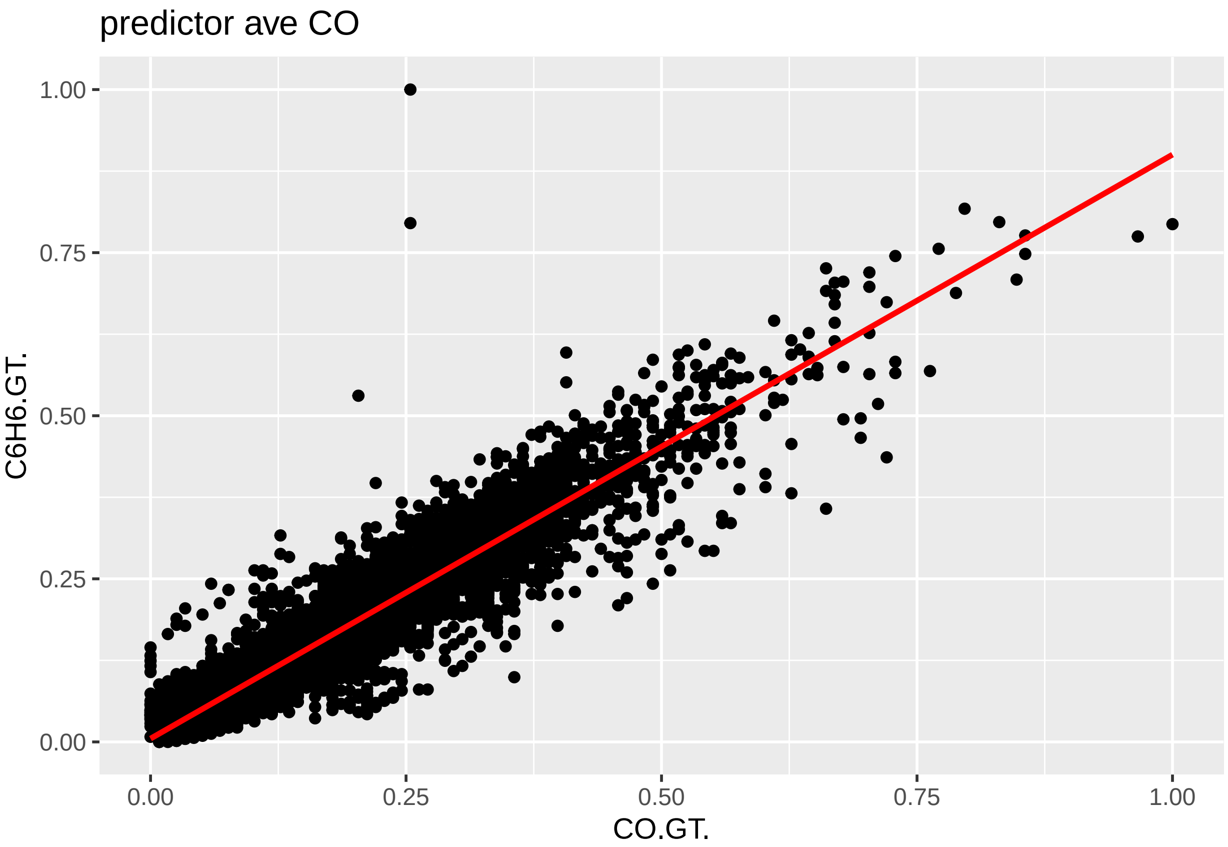
```r
#single linear regression with good predictor for baseline comparison

lm_5 <- lm(C6H6.GT. ~ CO.GT., airq_norm)
summary(lm_5)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ CO.GT., data = airq_norm)
##
## 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 ***
## 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
residuals(lm_5) %>% hist(main = "residuals ave CO")
```



```
ggplot(lm_5,
       aes(x = CO.GT., y = C6H6.GT.)) +
  geom_point() +
  geom_line(aes(y = .fitted), color = "red", size = 1) +
  ggtitle("predictor ave CO")
```

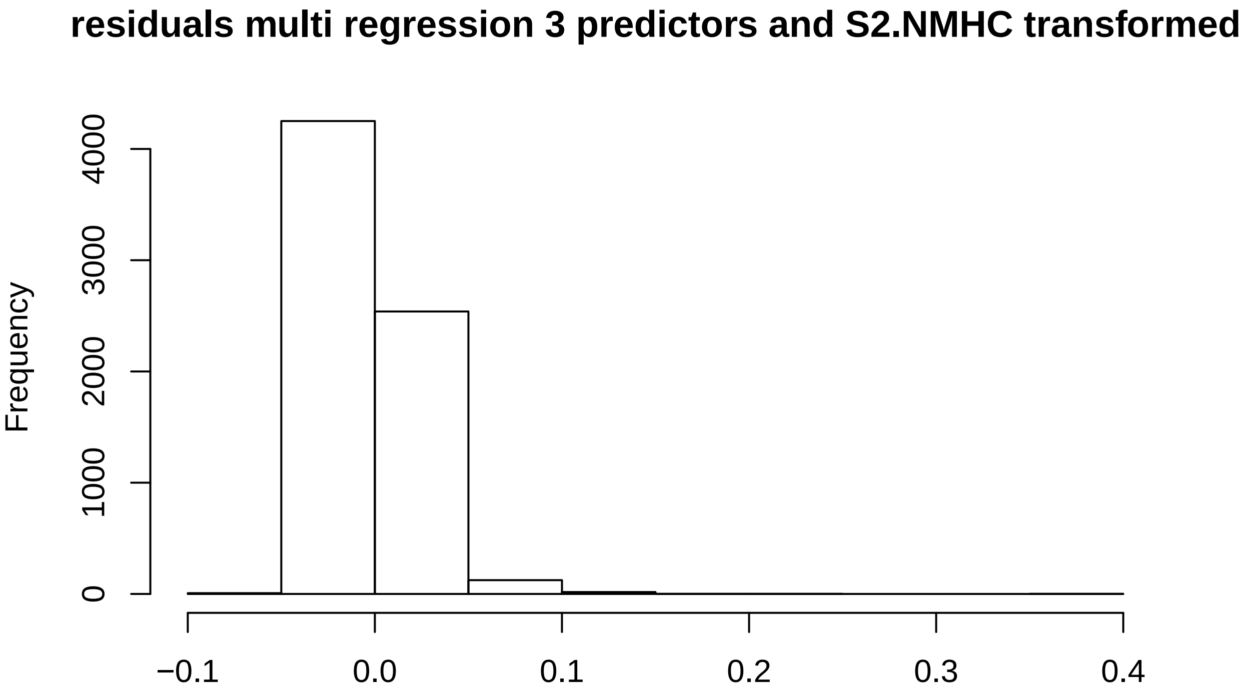


selected the following predictors: CO.GT., PT08.S1.CO. and PT08.S2.NMHC. for building a model

```
#CO: corr 0.88, good residuals, high R^2, sigif, linear
#PT08.S1.CO., corr 0.83, residuals okay, high R^2, sigif, linear
#PT08.S2.NMHC., sym residuals, high R^2, sigif, linear

#very high R^2, but residuals vs fitted plot not linear

lm_6 <- lm(C6H6.GT. ~ CO.GT. + PT08.S1.CO. + PT08.S2.NMHC., airq_norm)
summary(lm_6)
```

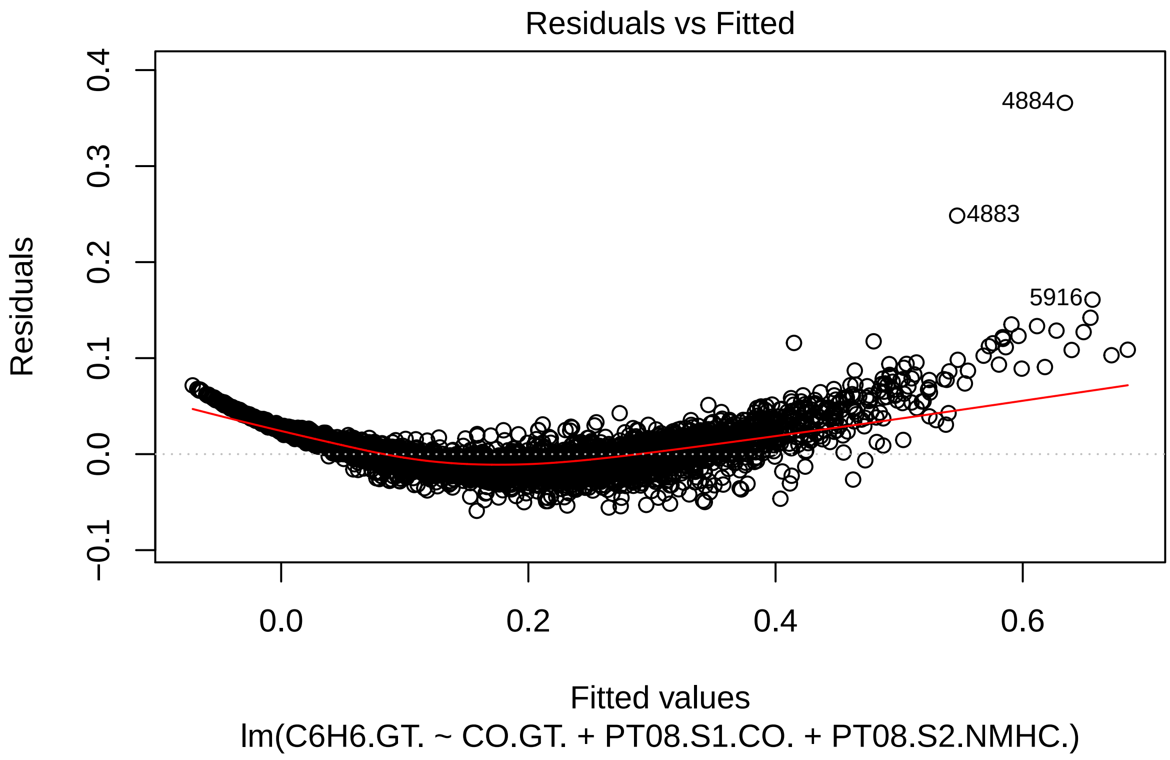
```
##
## Call:
## lm(formula = C6H6.GT. ~ CO.GT. + PT08.S1.CO. + PT08.S2.NMHC.,
##     data = airq_norm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05902 -0.01174 -0.00468  0.00846  0.36586
##
## Coefficients:
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   -0.072055   0.000688 -104.737  < 2e-16 ***
## CO.GT.         0.197796   0.005113   38.683  < 2e-16 ***
## PT08.S1.CO.   -0.024579   0.003475   -7.073 1.66e-12 ***
## PT08.S2.NMHC.  0.669088   0.004471  149.644  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01969 on 6937 degrees of freedom
## Multiple R-squared:  0.972,  Adjusted R-squared:  0.972
## F-statistic: 8.016e+04 on 3 and 6937 DF,  p-value: < 2.2e-16
```

```r
residuals(lm_6) %>% hist(main = "residuals multi regression 3 predictors and S2.NMHC transformed")
```



.

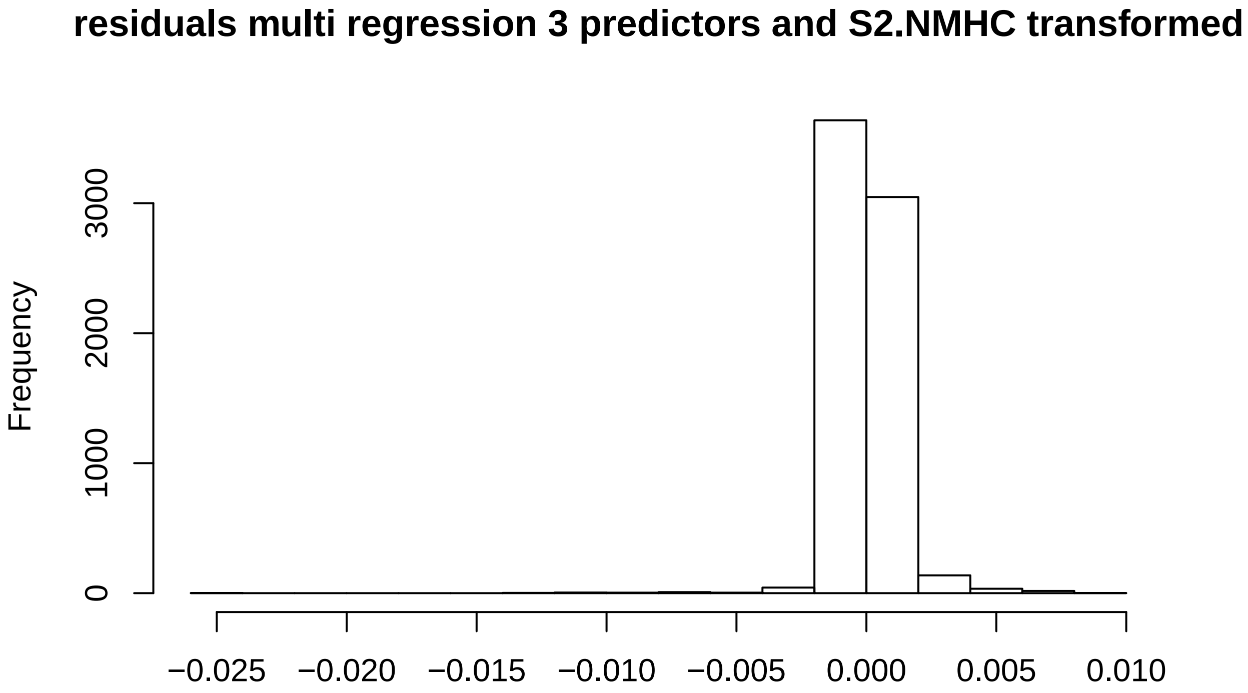
```r
plot(lm_6, which = 1)
```



```
#CO: corr 0.88, good residuals, high R^2, sigif, linear
#PT08.S1.CO., corr 0.83, residuals okay, high R^2, sigif, linear
#PT08.S2.NMHC., sym residuals, high R^2, sigif, linear

#inclusion of transformed PT08.S2.NMHC appears to make residuals vs fitted plot cubic?

lm_6 <- lm(C6H6.GT. ~ CO.GT. + PT08.S1.CO. + PT08.S2.NMHC. + NMHC_sq, airq_norm)
summary(lm_6)
```

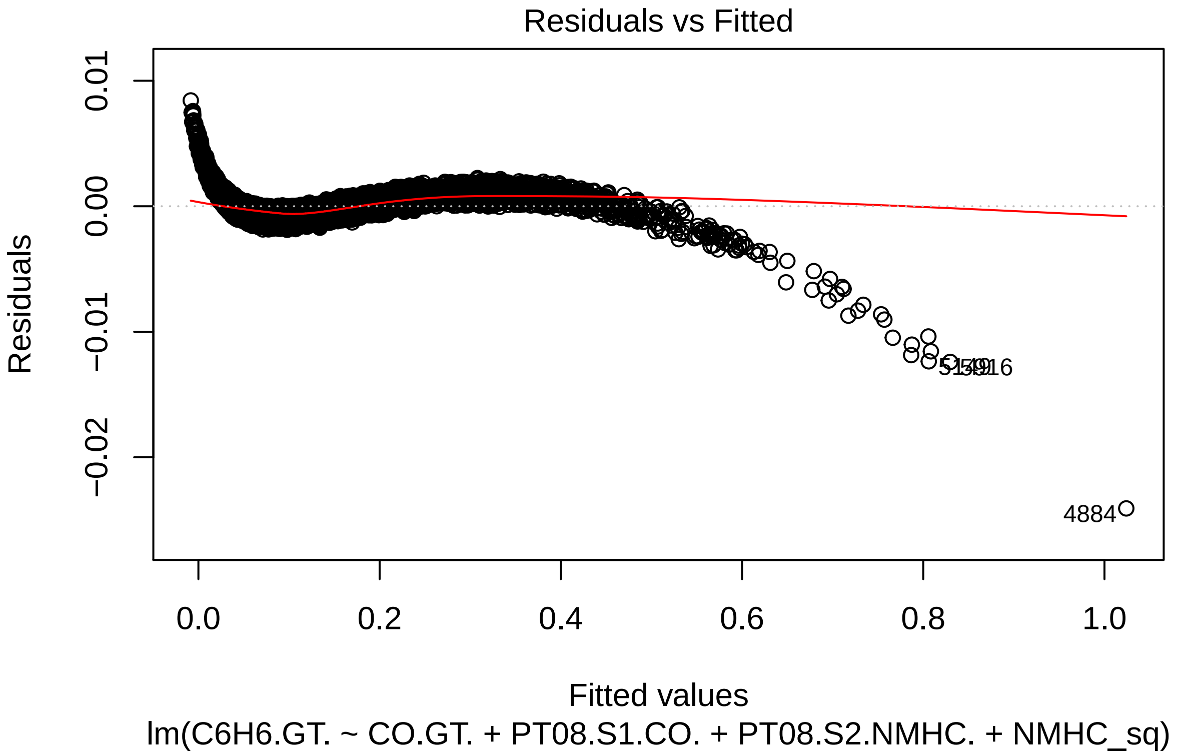
```
##
## Call:
## lm(formula = C6H6.GT. ~ CO.GT. + PT08.S1.CO. + PT08.S2.NMHC. +
##     NMHC_sq, data = airq_norm)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0240720 -0.0007106 -0.0000823  0.0006286  0.0084289
##
## Coefficients:
##                 Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   -8.471e-03  6.321e-05 -134.011  < 2e-16 ***
## CO.GT.         1.430e-03  3.461e-04    4.133 3.63e-05 ***
## PT08.S1.CO.    5.964e-04  2.143e-04    2.783   0.0054 **
## PT08.S2.NMHC.  2.542e-01  4.117e-04  617.324  < 2e-16 ***
## NMHC_sq        7.777e-01  5.748e-04 1353.018  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00121 on 6936 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 1.638e+07 on 4 and 6936 DF,  p-value: < 2.2e-16
```

```
residuals(lm_6) %>% hist(main = "residuals multi regression 3 predictors and S2.NMHC transformed")
```



.

```
plot(lm_6, which = 1)
```

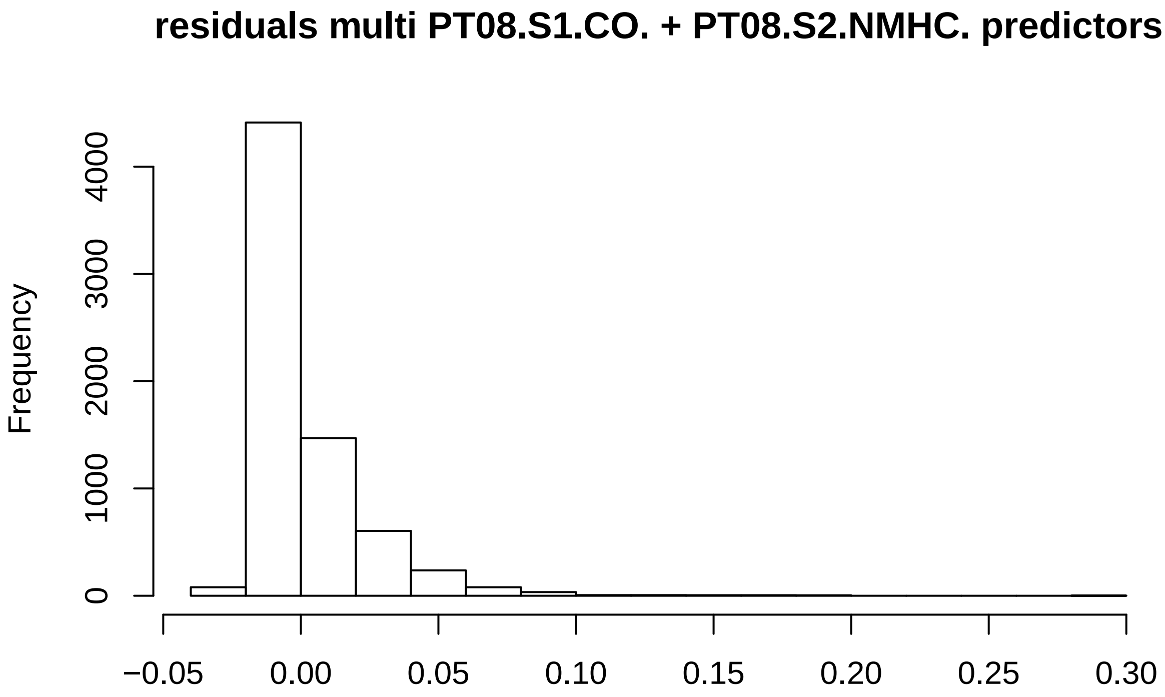


Model with the following predictors: PT08.S1.CO. and PT08.S2.NMHC.

```
# residuals vs fitted plot is not linear
lm_7 <- lm(C6H6.GT. ~ PT08.S1.CO. + PT08.S2.NMHC., airq_norm)
summary(lm_7)
```

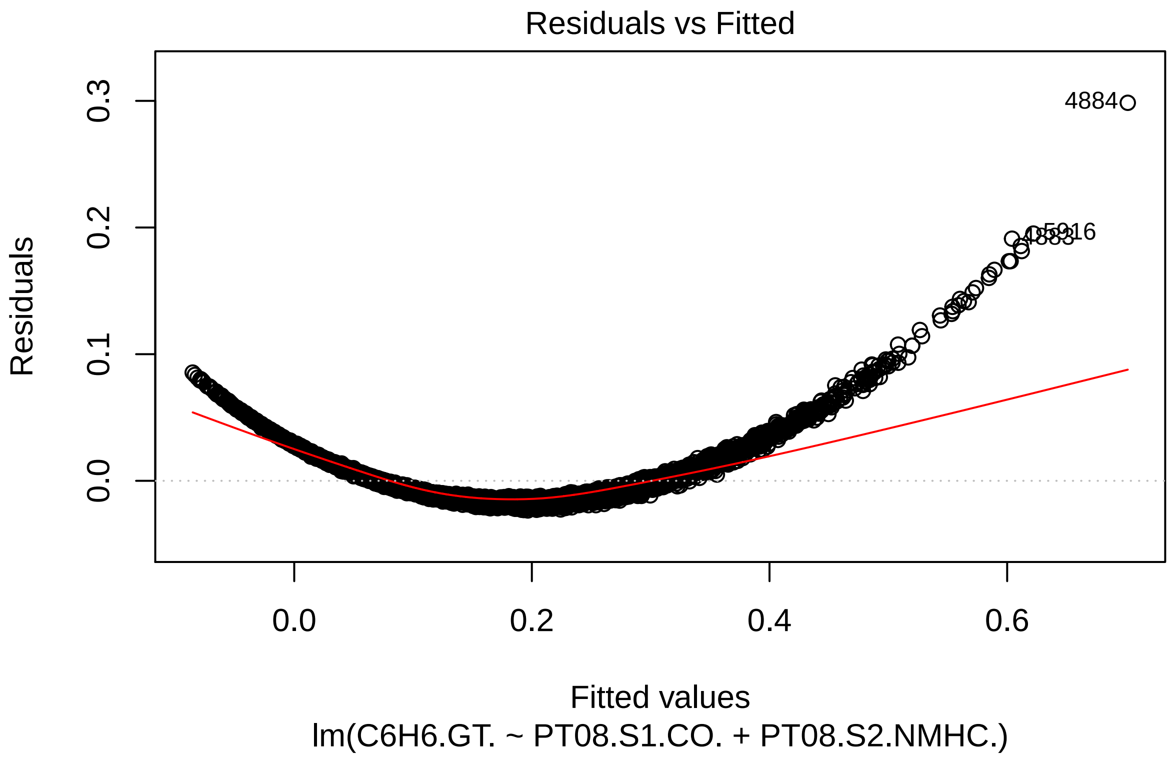
```
##
## Call:
## lm(formula = C6H6.GT. ~ PT08.S1.CO. + PT08.S2.NMHC., data = airq_norm)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.023401 -0.014539 -0.007786  0.008009  0.298478
##
## Coefficients:
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   -0.0866162  0.0006349 -136.429  < 2e-16 ***
## PT08.S1.CO.    0.0232981  0.0035798    6.508 8.14e-11 ***
## PT08.S2.NMHC.  0.7756448  0.0038829  199.760  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02171 on 6938 degrees of freedom
## Multiple R-squared:  0.9659, Adjusted R-squared:  0.9659
## F-statistic: 9.831e+04 on 2 and 6938 DF,  p-value: < 2.2e-16
```

```
residuals(lm_7) %>% hist(main = "residuals multi PT08.S1.CO. + PT08.S2.NMHC. predictors")
```



.

```
plot(lm_7, which = 1)
```



```
# variance inflation factor:
#how much the variance of a coefficient is inflated due to multicollinearity
#vif <5, which is good, however based on the plots and residuals,
# am not selecting this model as the "final" model
vif(lm_7)
```

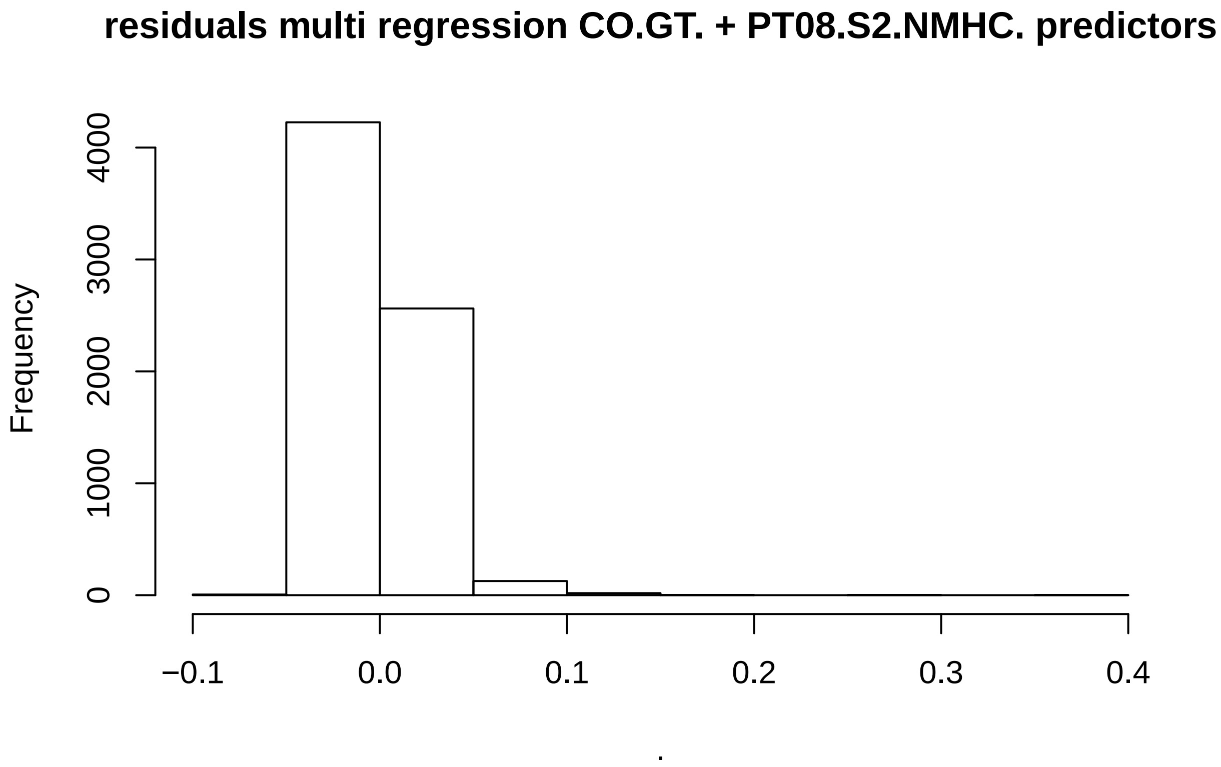
```
##   PT08.S1.CO. PT08.S2.NMHC.
##      4.653696      4.653696
```

Model with CO.GT.and PT08.S2.NMHC.

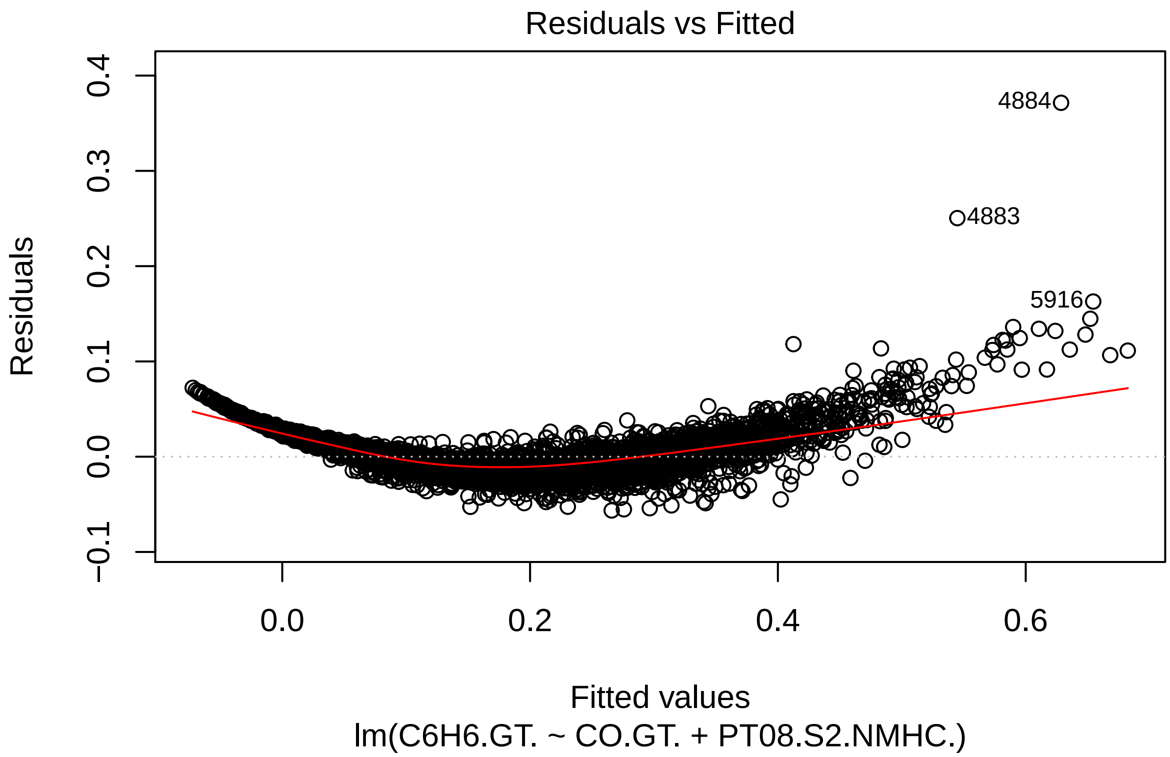
```
# seems to be quadratic, also built a model with PT08.S2.NMHC.^2, however did not improve the model

lm_8 <- lm(C6H6.GT. ~ CO.GT. + PT08.S2.NMHC., airq_norm)
summary(lm_8)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ CO.GT. + PT08.S2.NMHC., data = airq_norm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05643 -0.01185 -0.00467  0.00841  0.37145
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.0738605  0.0006411 -115.20   <2e-16 ***
## CO.GT.         0.1849137  0.0047948   38.56   <2e-16 ***
## PT08.S2.NMHC.  0.6554025  0.0040451  162.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01976 on 6938 degrees of freedom
## Multiple R-squared:  0.9718, Adjusted R-squared:  0.9718
## F-statistic: 1.194e+05 on 2 and 6938 DF,  p-value: < 2.2e-16
```

```
residuals(lm_8) %>% hist(main = "residuals multi regression CO.GT. + PT08.S2.NMHC. predictors")
```



```
plot(lm_8, which = 1)
```



```
# vif >5: indicates multicollinearity
vif(lm_8)
```

```
##        CO.GT. PT08.S2.NMHC.
##      6.096192      6.096192
```

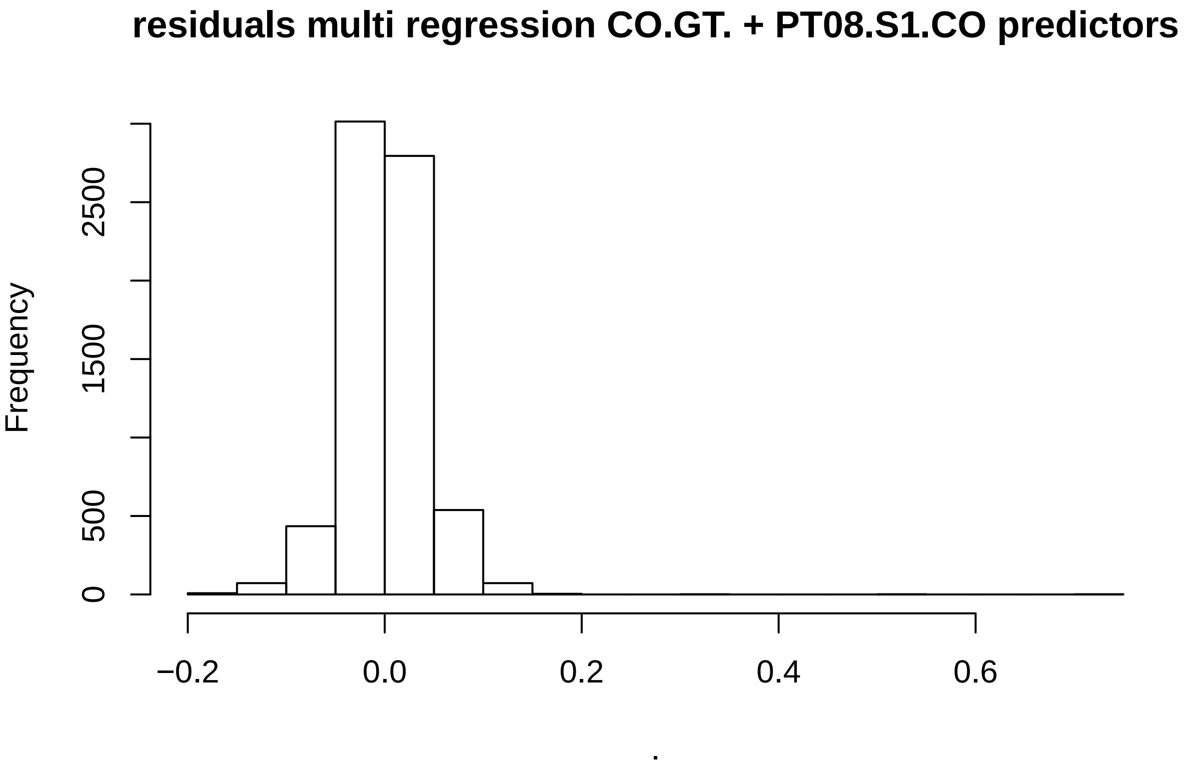
Final model with the following predictors: CO.GT.and PT08.S1.CO.

```
# predictors significant, best residuals with a few outliers on plot,
# high R^2 (larger than single regression with CO) and low vif

lm_9 <- lm(C6H6.GT. ~ CO.GT. + PT08.S1.CO., airq_norm)
summary(lm_9)
```

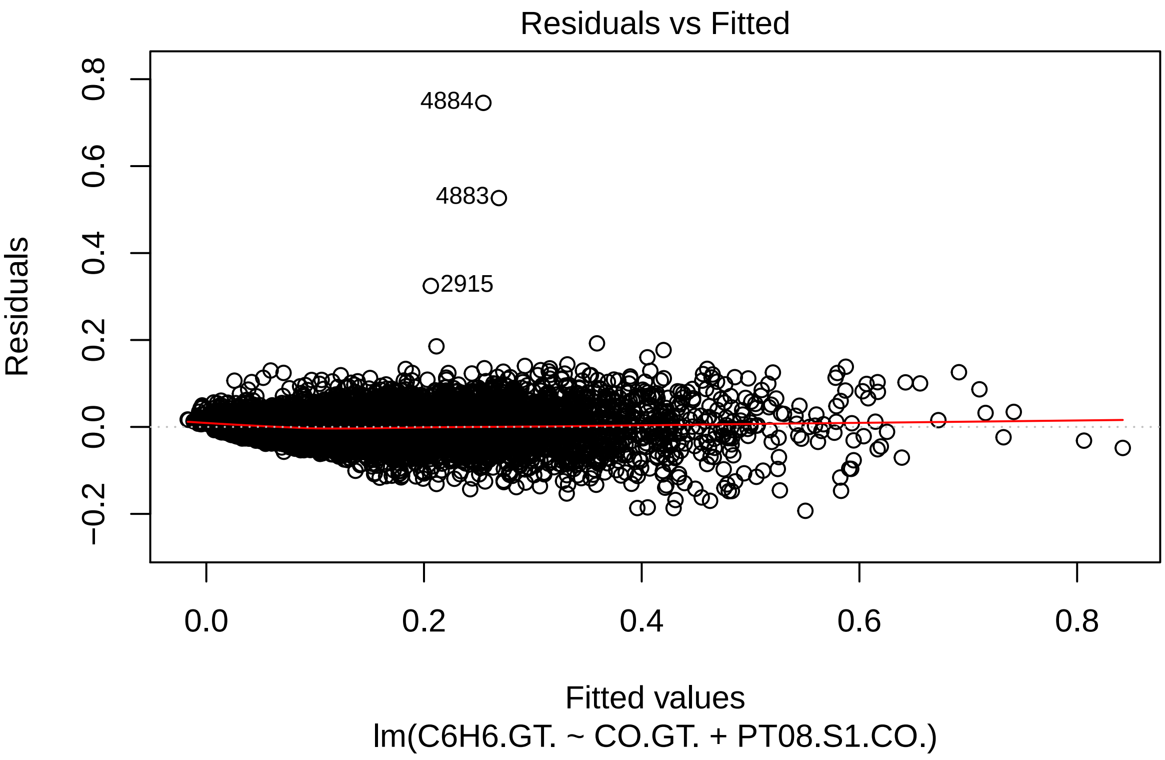
```
##
## Call:
## lm(formula = C6H6.GT. ~ CO.GT. + PT08.S1.CO., data = airq_norm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19287 -0.02482 -0.00104  0.02346  0.74546
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.023085   0.001244  -18.55   <2e-16 ***
## CO.GT.       0.669212   0.008281   80.81   <2e-16 ***
## PT08.S1.CO.  0.200433   0.006441   31.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.04048 on 6938 degrees of freedom
## Multiple R-squared:  0.8815, Adjusted R-squared:  0.8814
## F-statistic: 2.58e+04 on 2 and 6938 DF,  p-value: < 2.2e-16
```

```r
residuals(lm_9) %>% hist(main = "residuals multi regression CO.GT. + PT08.S1.CO predictors")
```



```r
plot(lm_9, which = 1)
```



```
# vif < 5:
#indicates that predictors are not redunant (not providing overlapping data to inform response)
vif(lm_9)
```

```
##      CO.GT. PT08.S1.CO.
##    4.331877    4.331877
```