Statistics in R: Task 12

Liuaza Etezova

```
library(corrplot)
library(car)
library(ggfortify)
```

Air Quality (multiple linear regression)

Preprocessing

```
air_quality <- read.csv('AirQualityUCI.csv', sep=';')

air_quality <- air_quality[-c(9358:nrow(air_quality)),
                           !(names(air_quality) %in% c('Date', 'Time', 'X','X.1'))]

columns <- c('CO.GT.', 'C6H6.GT.', 'T', 'RH', 'AH')
air_quality[columns] <- lapply(air_quality[columns], function(x) gsub(",", ".", x))
air_quality[columns] <- lapply(air_quality[columns], as.numeric)

air_quality[air_quality == -200] <- NA
```

```
str(air_quality)
```

```
## 'data.frame':    9357 obs. of  13 variables:
##  $ 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 NA ...
##  $ 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 NA ...
##  $ 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(air_quality)
```

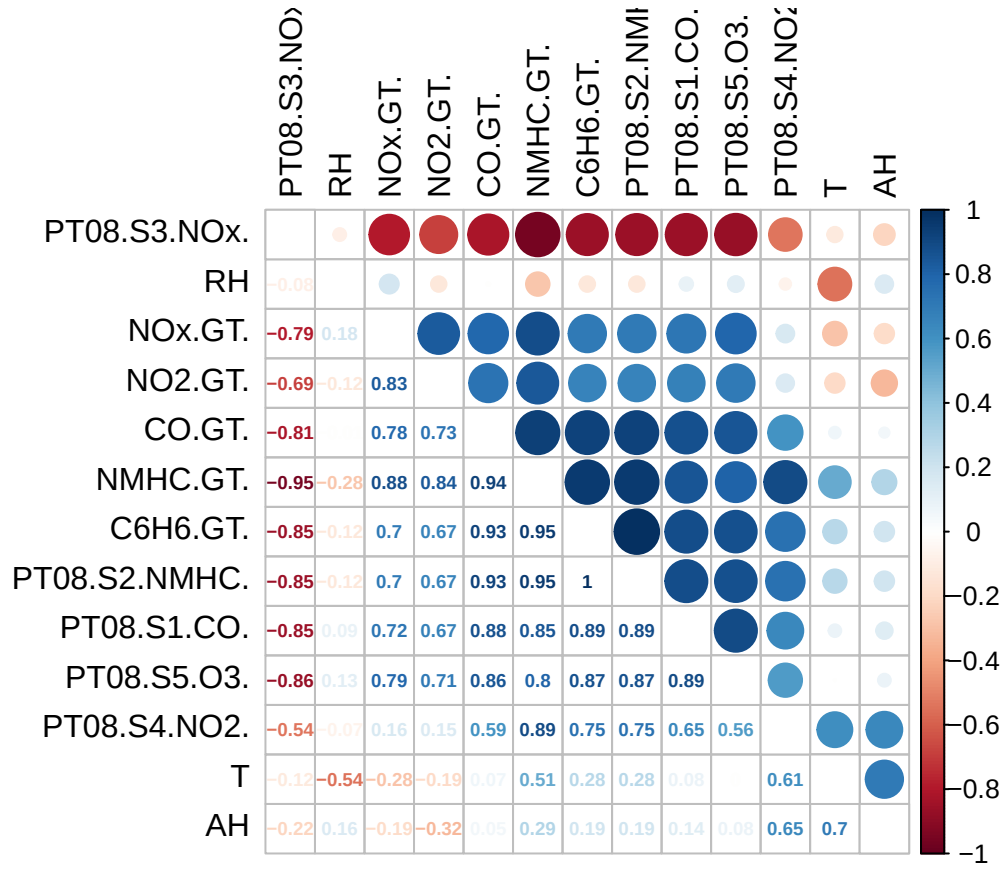
```
##      CO.GT.        PT08.S1.CO.      NMHC.GT.         C6H6.GT.
```

```
##  Min.   : 0.100   Min.   : 647   Min.   :   7.0   Min.   : 0.10
##  1st Qu.: 1.100   1st Qu.: 937   1st Qu.:  67.0   1st Qu.: 4.40
##  Median : 1.800   Median :1063   Median : 150.0   Median : 8.20
##  Mean   : 2.153   Mean   :1100   Mean   : 218.8   Mean   :10.08
##  3rd Qu.: 2.900   3rd Qu.:1231   3rd Qu.: 297.0   3rd Qu.:14.00
##  Max.   :11.900   Max.   :2040   Max.   :1189.0   Max.   :63.70
##  NA's   :1683     NA's   :366    NA's   :8443     NA's   :366
##  PT08.S2.NMHC.       NOx.GT.        PT08.S3.NOx.       NO2.GT.
##  Min.   : 383.0   Min.   :   2.0   Min.   : 322.0   Min.   :  2.0
##  1st Qu.: 734.5   1st Qu.:  98.0   1st Qu.: 658.0   1st Qu.: 78.0
##  Median : 909.0   Median : 180.0   Median : 806.0   Median :109.0
##  Mean   : 939.2   Mean   : 246.9   Mean   : 835.5   Mean   :113.1
##  3rd Qu.:1116.0   3rd Qu.: 326.0   3rd Qu.: 969.5   3rd Qu.:142.0
##  Max.   :2214.0   Max.   :1479.0   Max.   :2683.0   Max.   :340.0
##  NA's   :366      NA's   :1639     NA's   :366      NA's   :1642
##   PT08.S4.NO2.   PT08.S5.O3.           T               RH
##  Min.   : 551   Min.   : 221.0   Min.   :-1.90   Min.   : 9.20
##  1st Qu.:1227   1st Qu.: 731.5   1st Qu.:11.80   1st Qu.:35.80
##  Median :1463   Median : 963.0   Median :17.80   Median :49.60
##  Mean   :1456   Mean   :1022.9   Mean   :18.32   Mean   :49.23
##  3rd Qu.:1674   3rd Qu.:1273.5   3rd Qu.:24.40   3rd Qu.:62.50
##  Max.   :2775   Max.   :2523.0   Max.   :44.60   Max.   :88.70
##  NA's   :366    NA's   :366      NA's   :366     NA's   :366
##        AH
##  Min.   :0.1847
##  1st Qu.:0.7368
##  Median :0.9954
##  Mean   :1.0255
##  3rd Qu.:1.3137
##  Max.   :2.2310
##  NA's   :366
```

Exploring multicollinearity

```
cors <- cor(air_quality, use='pairwise.complete.obs', method='spearman')
```

```
corrplot.mixed(cors, order='hclust', lower='number', tl.pos='lt', tl.col='black',
               number.cex=0.55)
```




```
model_all <- lm(C6H6.GT. ~ ., data = air_quality)
summary(model_all)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ ., data = air_quality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6771 -0.3765 -0.0190  0.3200  3.4535
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.525e+01  6.323e-01 -39.943  < 2e-16 ***
## CO.GT.         1.050e+00  8.541e-02  12.292  < 2e-16 ***
## PT08.S1.CO.   -2.765e-03  3.917e-04  -7.058 3.63e-12 ***
## NMHC.GT.       2.351e-03  2.525e-04   9.310  < 2e-16 ***
## PT08.S2.NMHC.  2.166e-02  7.740e-04  27.977  < 2e-16 ***
## NOx.GT.       -2.297e-03  1.009e-03  -2.277   0.0231 *
## PT08.S3.NOx.   6.408e-03  2.710e-04  23.645  < 2e-16 ***
## NO2.GT.       -1.367e-02  1.780e-03  -7.680 4.56e-14 ***
## PT08.S4.NO2.   6.454e-03  5.341e-04  12.084  < 2e-16 ***
## PT08.S5.O3.    1.436e-03  1.701e-04   8.446  < 2e-16 ***
## T              1.329e-02  2.096e-02   0.634   0.5263
## RH            -1.221e-02  7.195e-03  -1.697   0.0900 .
```

```
## AH            -5.805e-01  4.901e-01  -1.184   0.2366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5966 on 814 degrees of freedom
##   (8530 observations deleted due to missingness)
## Multiple R-squared:  0.9936, Adjusted R-squared:  0.9935
## F-statistic: 1.058e+04 on 12 and 814 DF,  p-value: < 2.2e-16
```

```
vif(model_all)
```

```
##        CO.GT.   PT08.S1.CO.      NMHC.GT. PT08.S2.NMHC.       NOx.GT.
##     33.637317     20.826040      6.432465     98.706491     15.819663
##  PT08.S3.NOx.       NO2.GT.  PT08.S4.NO2.   PT08.S5.O3.             T
##     12.054763      7.298246     60.499344     10.747744     23.748039
##            RH            AH
##     28.003324     17.762660
```

```
model_vif_2p <- lm(C6H6.GT. ~ NO2.GT. + NMHC.GT., data = air_quality)
summary(model_vif_2p)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NO2.GT. + NMHC.GT., data = air_quality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1777 -1.3691 -0.2744  1.0660 16.5805
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.7364681  0.3135029  -11.92   <2e-16 ***
## NO2.GT.      0.0954497  0.0039068   24.43   <2e-16 ***
## NMHC.GT.     0.0214118  0.0006023   35.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.484 on 848 degrees of freedom
##   (8506 observations deleted due to missingness)
## Multiple R-squared:  0.8887, Adjusted R-squared:  0.8884
## F-statistic:  3385 on 2 and 848 DF,  p-value: < 2.2e-16
```

```
model_vif_3p <- lm(C6H6.GT. ~ NO2.GT. + NMHC.GT. + PT08.S5.O3., data = air_quality)
summary(model_vif_3p)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NO2.GT. + NMHC.GT. + PT08.S5.O3., data = air_quality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4204 -1.3398 -0.1235  1.1767 13.0933
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.7930833  0.2642149  -18.14   <2e-16 ***
## NO2.GT.      0.0420102  0.0041950   10.01   <2e-16 ***
## NMHC.GT.     0.0170583  0.0005431   31.41   <2e-16 ***
## PT08.S5.O3.  0.0070717  0.0003549   19.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.051 on 847 degrees of freedom
##   (8506 observations deleted due to missingness)
## Multiple R-squared:  0.9242, Adjusted R-squared:  0.9239
## F-statistic:  3443 on 3 and 847 DF,  p-value: < 2.2e-16
```

```
model_vif_4p <- lm(C6H6.GT. ~ NO2.GT. + NMHC.GT. + PT08.S5.O3. + NOx.GT., data = air_quality)
summary(model_vif_4p)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NO2.GT. + NMHC.GT. + PT08.S5.O3. + NOx.GT.,
##     data = air_quality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9716 -1.0756 -0.1798  0.9867 10.0380
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.1699236  0.2622303 -12.088  < 2e-16 ***
## NO2.GT.      0.0185914  0.0040961   4.539 6.48e-06 ***
## NMHC.GT.     0.0140348  0.0005301  26.476  < 2e-16 ***
## PT08.S5.O3.  0.0043141  0.0003712  11.622  < 2e-16 ***
## NOx.GT.      0.0300950  0.0020867  14.423  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.839 on 846 degrees of freedom
##   (8506 observations deleted due to missingness)
## Multiple R-squared:  0.9392, Adjusted R-squared:  0.9389
## F-statistic:  3265 on 4 and 846 DF,  p-value: < 2.2e-16
```

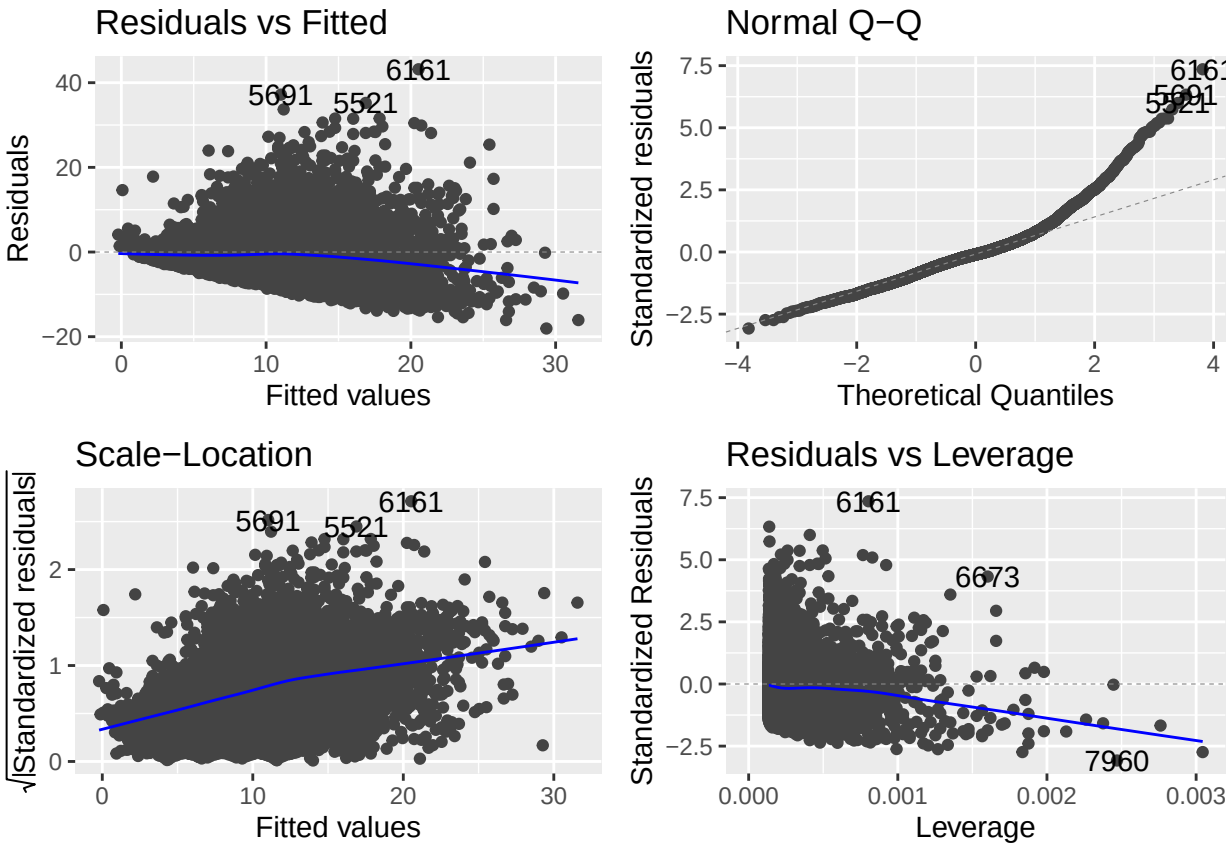
```
vif(model_vif_3p)
```

```
##     NO2.GT.    NMHC.GT. PT08.S5.O3.
##    3.653759    2.577353    4.067478
```

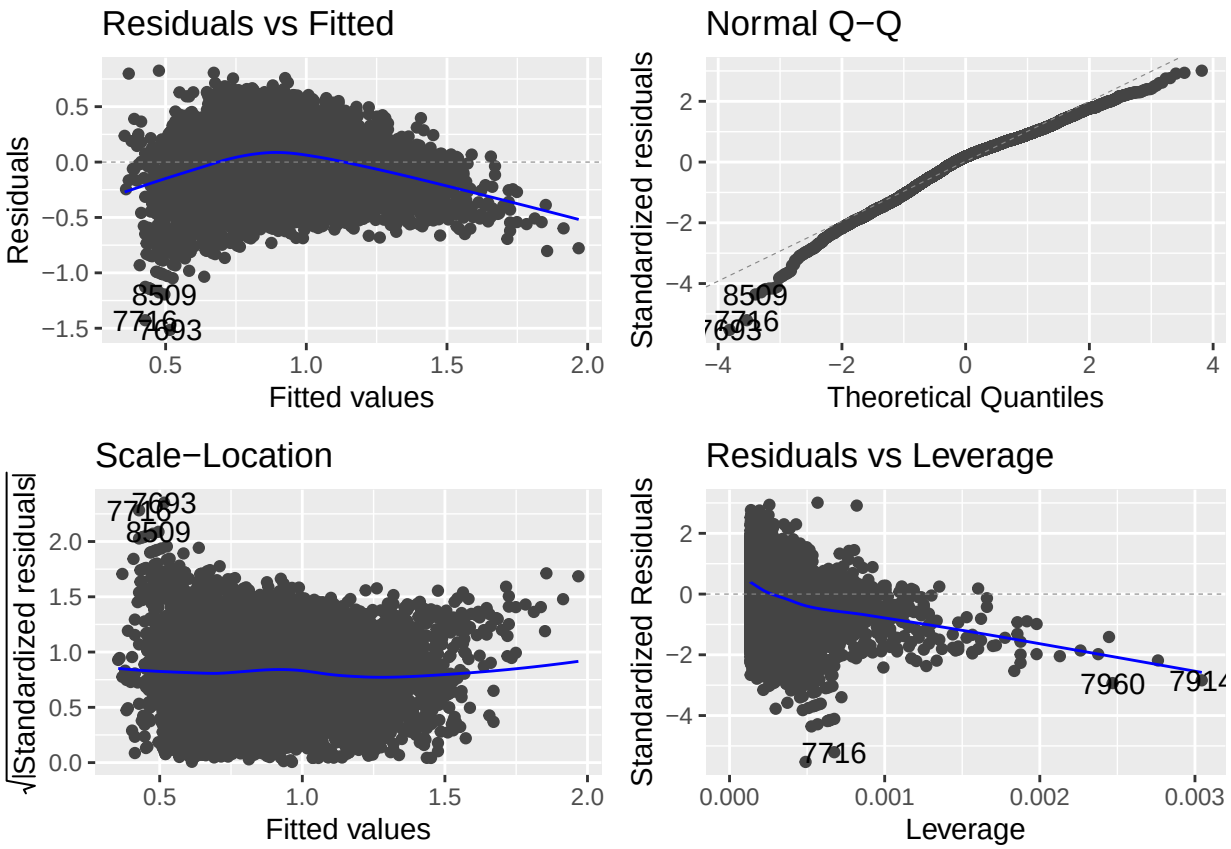
Checking residuals & Transforming selected predictors

```
model <- lm(C6H6.GT. ~ NO2.GT., data = air_quality)
autoplot(model)
```

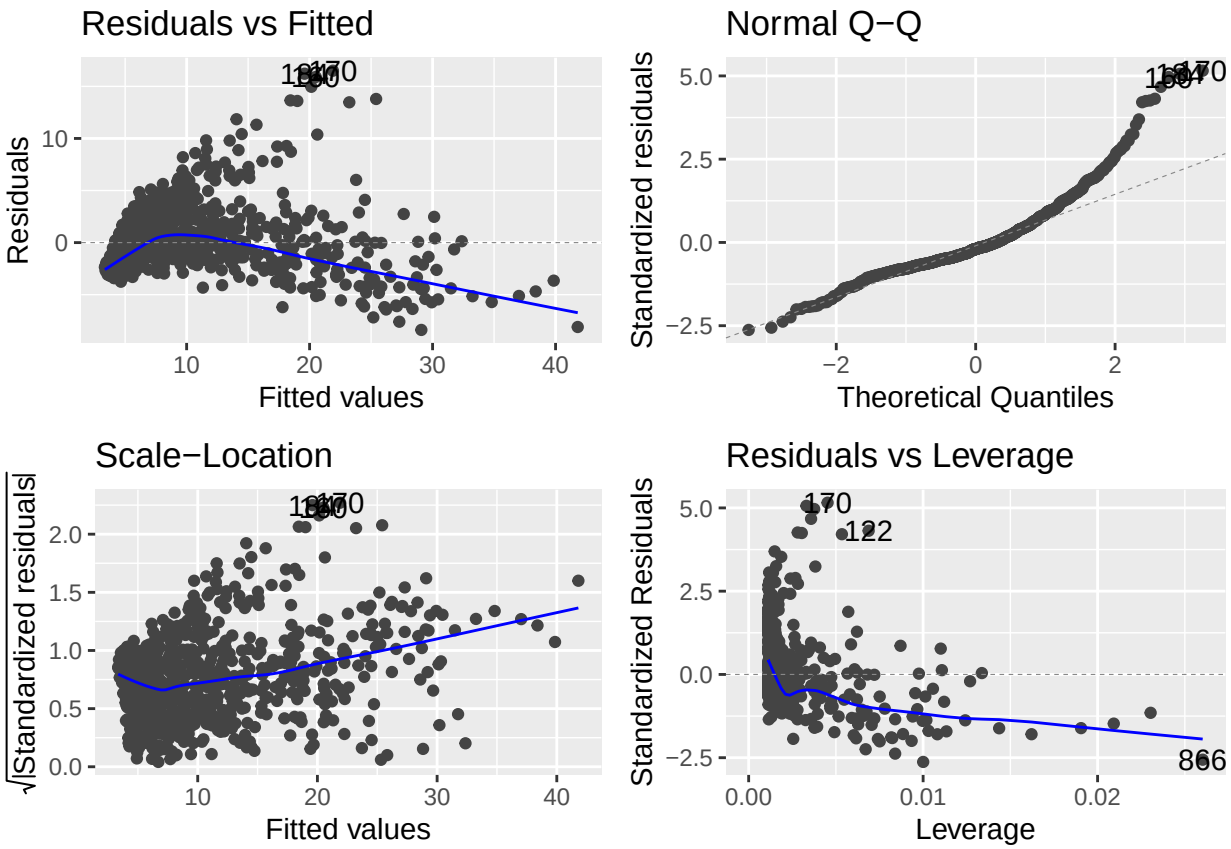
```
## Warning: `arrange_()` is deprecated as of dplyr 0.7.0.
## Please use `arrange()` instead.
## See vignette('programming') for more help
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```



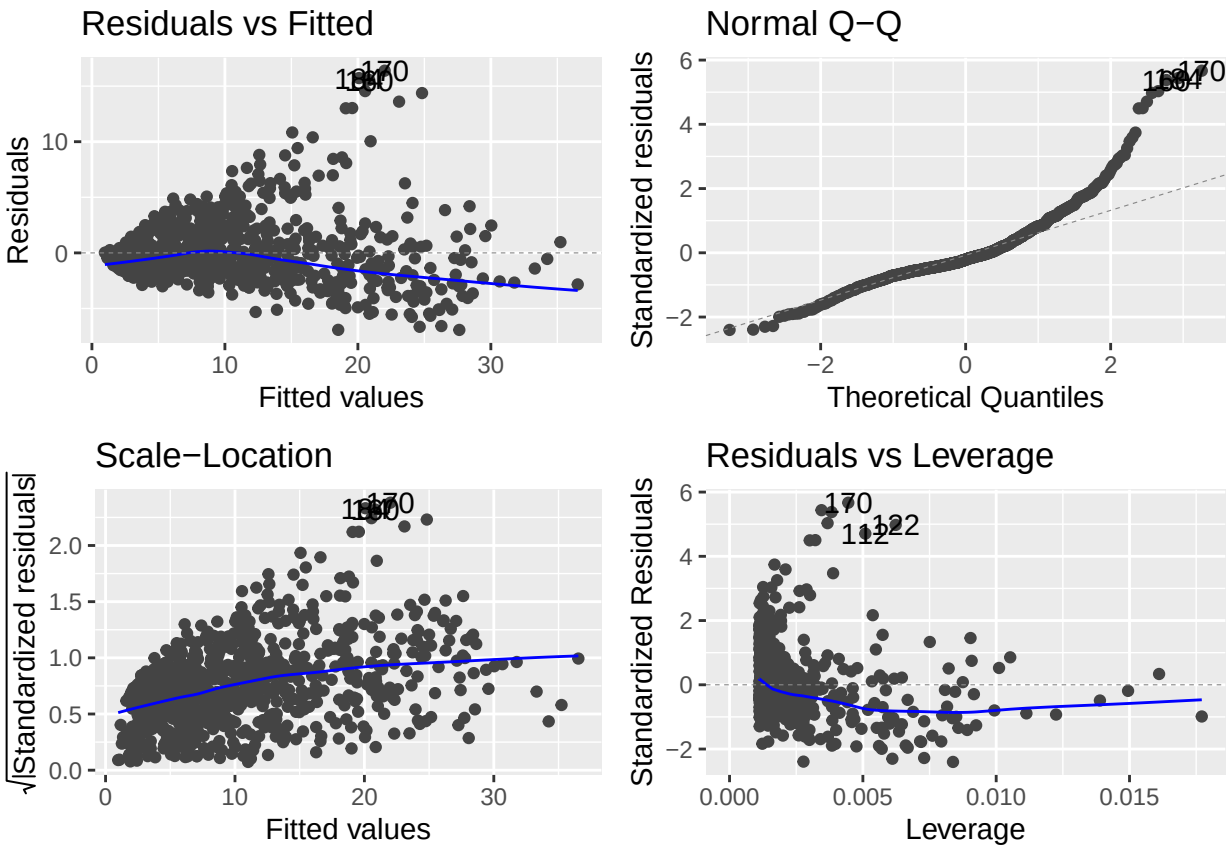

```
model <- lm(log10(C6H6.GT.) ~ NO2.GT., data = air_quality)
autoplot(model)
```



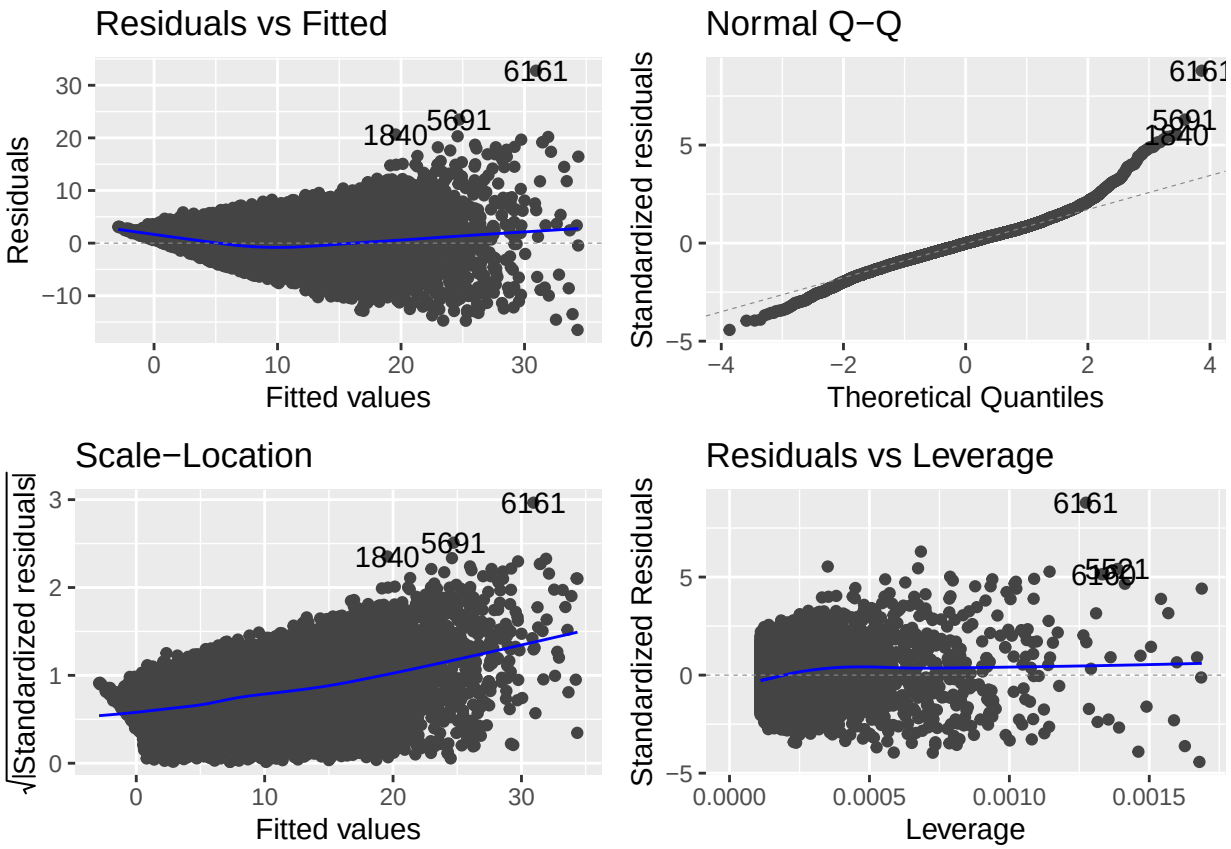

```
model <- lm(C6H6.GT. ~ NMHC.GT., data = air_quality)
autoplot(model)
```



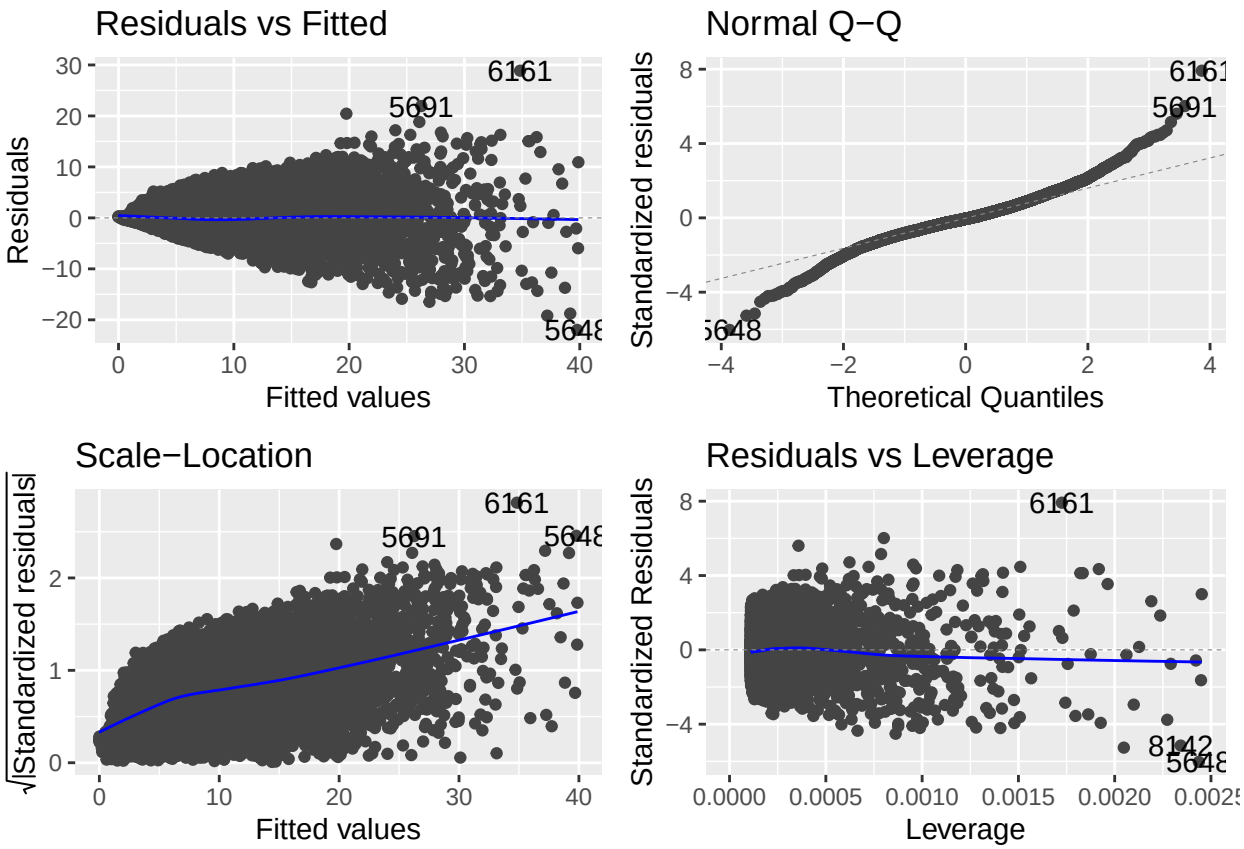

```
model <- lm(C6H6.GT. ~ I(NMHC.GT.^0.7), data = air_quality)
autoplot(model)
```




```
model <- lm(C6H6.GT. ~ PT08.S5.O3., data = air_quality)
autoplot(model)
```




```
model <- lm(C6H6.GT. ~ I(PT08.S5.O3.^1.5), data = air_quality)
autoplot(model)
```



The final linear model

```
final_model0 <- lm(C6H6.GT. ~ NO2.GT. + I(NMHC.GT.^0.7) + PT08.S5.O3., data = air_quality)
summary(final_model0)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NO2.GT. + I(NMHC.GT.^0.7) + PT08.S5.O3.,
##     data = air_quality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0671 -1.3143 -0.1256  1.0745 12.9074
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.6127958  0.2492907 -22.515  < 2e-16 ***
## NO2.GT.          0.0296702  0.0042675   6.953 7.16e-12 ***
## I(NMHC.GT.^0.7)  0.1448260  0.0045292  31.976  < 2e-16 ***
## PT08.S5.O3.      0.0070352  0.0003515  20.016  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.031 on 847 degrees of freedom
##   (8506 observations deleted due to missingness)
## Multiple R-squared:  0.9257, Adjusted R-squared:  0.9254
## F-statistic:  3516 on 3 and 847 DF,  p-value: < 2.2e-16
```

```
final_model1 <- lm(C6H6.GT. ~ NO2.GT. + I(NMHC.GT.^0.7) + I(PT08.S5.O3.^1.5), data = air_quality)
summary(final_model1)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ NO2.GT. + I(NMHC.GT.^0.7) + I(PT08.S5.O3.^1.5),
##     data = air_quality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6378 -1.1607 -0.0475  1.0062 11.8335
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -3.423e+00  2.380e-01 -14.386  < 2e-16 ***
## NO2.GT.             2.932e-02  3.881e-03   7.553  1.1e-13 ***
## I(NMHC.GT.^0.7)     1.368e-01  4.325e-03  31.634  < 2e-16 ***
## I(PT08.S5.O3.^1.5)  1.552e-04  6.529e-06  23.763  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.91 on 847 degrees of freedom
##   (8506 observations deleted due to missingness)
## Multiple R-squared:  0.9343, Adjusted R-squared:  0.9341
## F-statistic:  4016 on 3 and 847 DF,  p-value: < 2.2e-16
```

```
final_model2 <- lm(C6H6.GT. ~ I(NMHC.GT.^0.7) + I(PT08.S5.O3.^1.5), data = air_quality)
summary(final_model2)
```

```
##
## Call:
## lm(formula = C6H6.GT. ~ I(NMHC.GT.^0.7) + I(PT08.S5.O3.^1.5),
##     data = air_quality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3936 -1.2511 -0.0271  1.0102 10.8808
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.009e+00  1.325e-01  -15.16   <2e-16 ***
## I(NMHC.GT.^0.7)     1.502e-01  4.093e-03   36.70   <2e-16 ***
## I(PT08.S5.O3.^1.5)  1.808e-04  5.527e-06   32.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.95 on 884 degrees of freedom
##   (8470 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.9316, Adjusted R-squared:  0.9314
## F-statistic:  6019 on 2 and 884 DF,  p-value: < 2.2e-16
```