Multiple Linear Regression

DATA621 Blog 03

Zhi Ying Chen

27 November 2020

Generalized linear model (GLM) is a generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution like Gaussian distribution.

In R, using lm() is a special case of glm(). lm() fits models following the form Y = Xb + e, where e is Normal (0 , sˆ2).

glm() fits models following the form f(Y) = Xb + e. … i.e. if you don't specify the link function and error distribution, the parameters that glm() uses produce the same effect as running lm().

Load Packages

```
library(recommenderlab)
```

```
## Warning: package 'recommenderlab' was built under R version 3.5.3

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 3.5.3

## Loading required package: arules

## Warning: package 'arules' was built under R version 3.5.3

##
## Attaching package: 'arules'

## The following objects are masked from 'package:base':
##
##     abbreviate, write

## Loading required package: proxy

## Warning: package 'proxy' was built under R version 3.5.3

##
## Attaching package: 'proxy'

## The following object is masked from 'package:Matrix':
##
##     as.matrix

## The following objects are masked from 'package:stats':
##
##     as.dist, dist

## The following object is masked from 'package:base':
##
```

```
##     as.matrix

## Loading required package: registry
```

```r
library(ggplot2)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3

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

## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.4

## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'tidyr' was built under R version 3.5.3

## Warning: package 'readr' was built under R version 3.5.3

## Warning: package 'purrr' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

## Warning: package 'forcats' was built under R version 3.5.3

## -- Conflicts ----------------------------------------------------------------- tidyverse_conflicts()
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x dplyr::recode() masks arules::recode()
## x tidyr::unpack() masks Matrix::unpack()
```

```r
library(dplyr)
library(Metrics)
```

```
## Warning: package 'Metrics' was built under R version 3.5.3
```

```r
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 3.5.3

##
## Attaching package: 'kableExtra'

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

```r
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.5.3

##
## Attaching package: 'gridExtra'

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

```r
library(rmdformats)
```

```
## Warning: package 'rmdformats' was built under R version 3.5.3
```

```r
library(caTools)
library(formattable)
```

```
## Warning: package 'formattable' was built under R version 3.5.3
```

```
##
## Attaching package: 'formattable'
```

```
## The following object is masked from 'package:recommenderlab':
##
##     normalize
```

```r
library(mice)
```

```
## Warning: package 'mice' was built under R version 3.5.3
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```r
library(naniar)
library(reshape)
```

```
## Warning: package 'reshape' was built under R version 3.5.3
```

```
##
## Attaching package: 'reshape'
```

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

```
## The following objects are masked from 'package:tidyr':
##
##     expand, smiths
```

```
## The following object is masked from 'package:Matrix':
##
##     expand
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.5.3
```

```
## corrplot 0.84 loaded
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.5.3
```

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

```
## Warning: package 'lattice' was built under R version 3.5.3
```

```
##
## Attaching package: 'caret'
## The following objects are masked from 'package:Metrics':
##
##     precision, recall

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

## The following objects are masked from 'package:recommenderlab':
##
##     MAE, RMSE
```

```r
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.5.3
```

```r
library(scales)
```

```
##
## Attaching package: 'scales'

## The following objects are masked from 'package:formattable':
##
##     comma, percent, scientific

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

## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
library(gplots)
```

```
## Warning: package 'gplots' was built under R version 3.5.3

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.5.3

##
## Attaching package: 'MASS'

## The following object is masked from 'package:formattable':
##
##     area

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

```r
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 3.5.3
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following object is masked from 'package:Metrics':
##
##     auc

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

Read Data

```r
data(airquality)
str(airquality)
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

This provides insight telling us that airquality is a of class data.frame, the number of observation, the number of variables, and further details about each variable and the first 10 values in each column.

```r
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

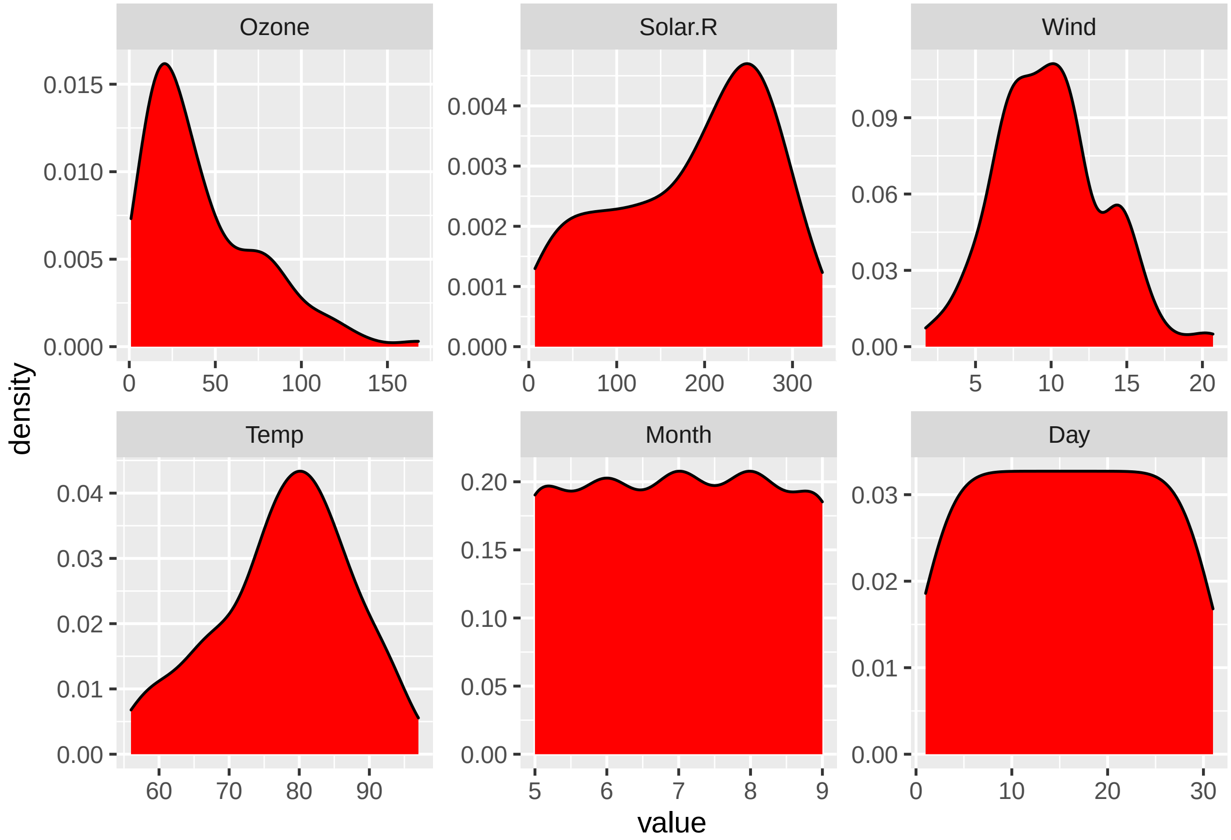
```r
par(mfrow = c(3, 3))

datasub = melt(airquality)
```

```
## Using  as id variables
```

```r
ggplot(datasub, aes(x= value)) +
    geom_density(fill='red') + facet_wrap(~variable, scales = 'free')
```

```
## Warning: Removed 44 rows containing non-finite values (stat_density).
```



DATA PREPARATION

```
set.seed(100)
ozone <- subset(na.omit(airquality),
        select = c("Ozone", "Solar.R", "Wind", "Temp"))
train <- ceiling(0.7 * nrow(ozone))
test <- nrow(ozone) - train
trainset <- sample(seq_len(nrow(ozone)), train)
testset <- setdiff(seq_len(nrow(ozone)), trainset)
```

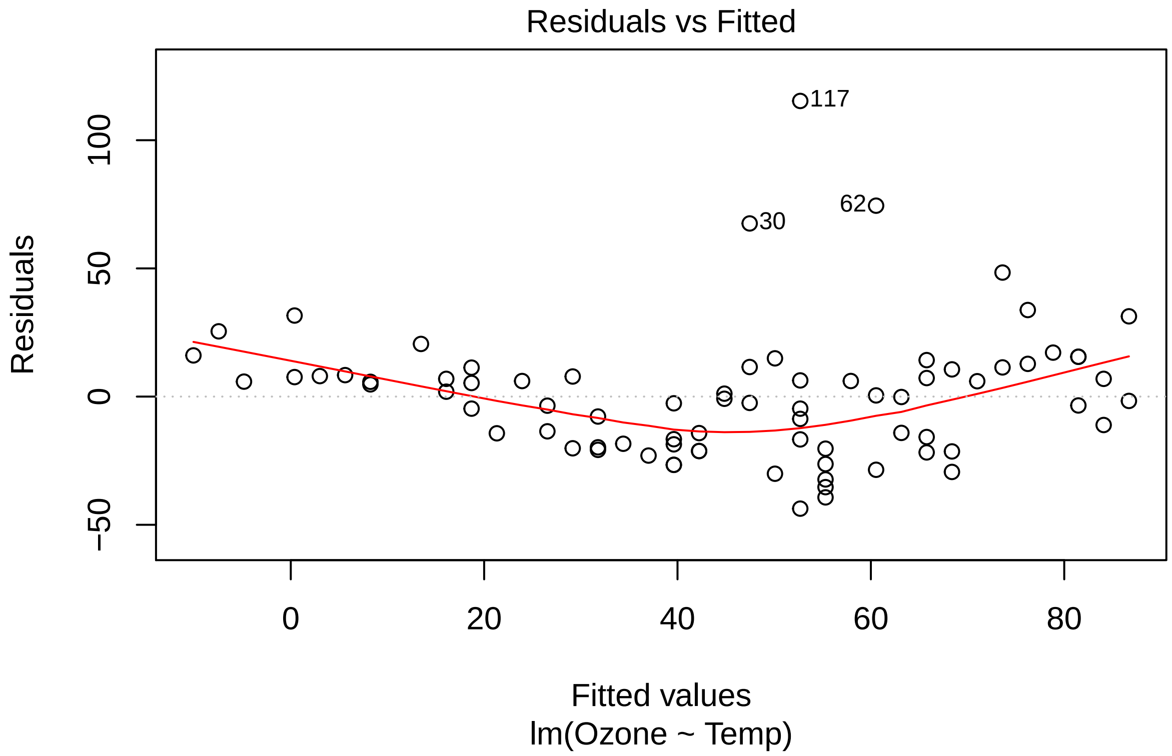
Build Models

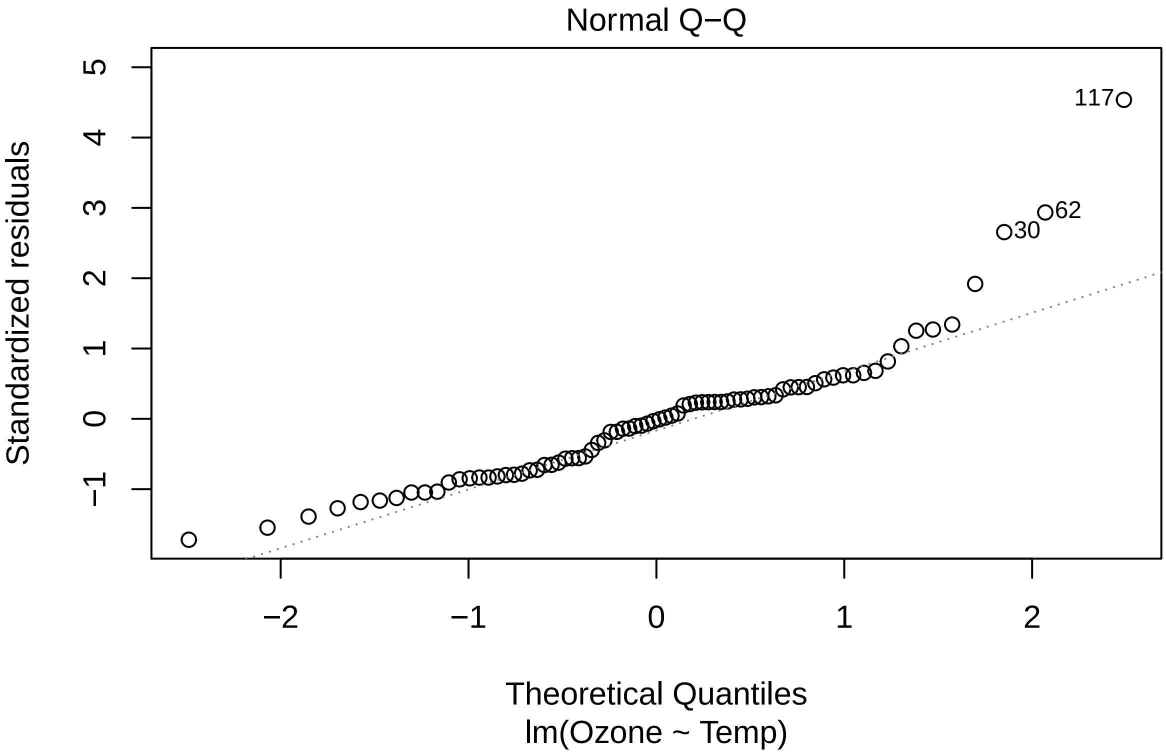
Model 1 - Standard linear model

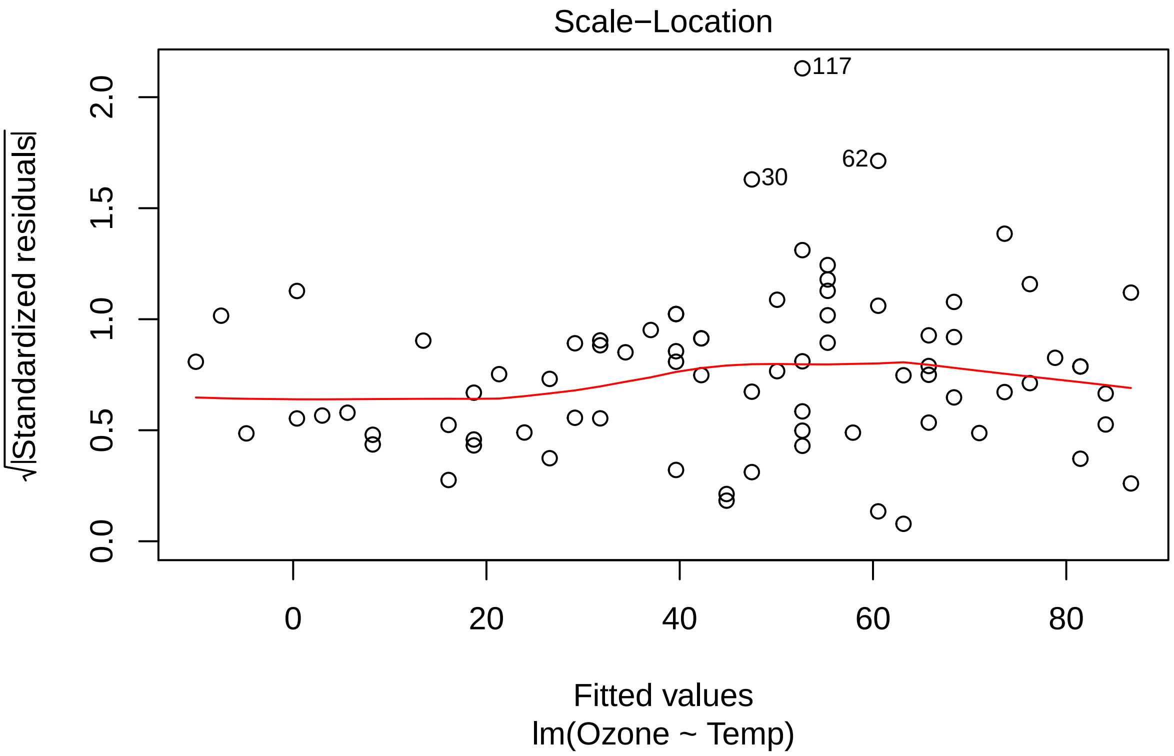
The simple function lm() creates a linear model of the data and will omit NAs if any automatically. For this example it suffices. Other options exists, or computations can ne one to impute the missing data, for example replacing each NA with the average (mean) of all values. The result of lm() is a slope and an intercept which describes a regression line. This can help show a trend, but it is also important to keep in mind that lm() is a simple model and that other regression methods exist.

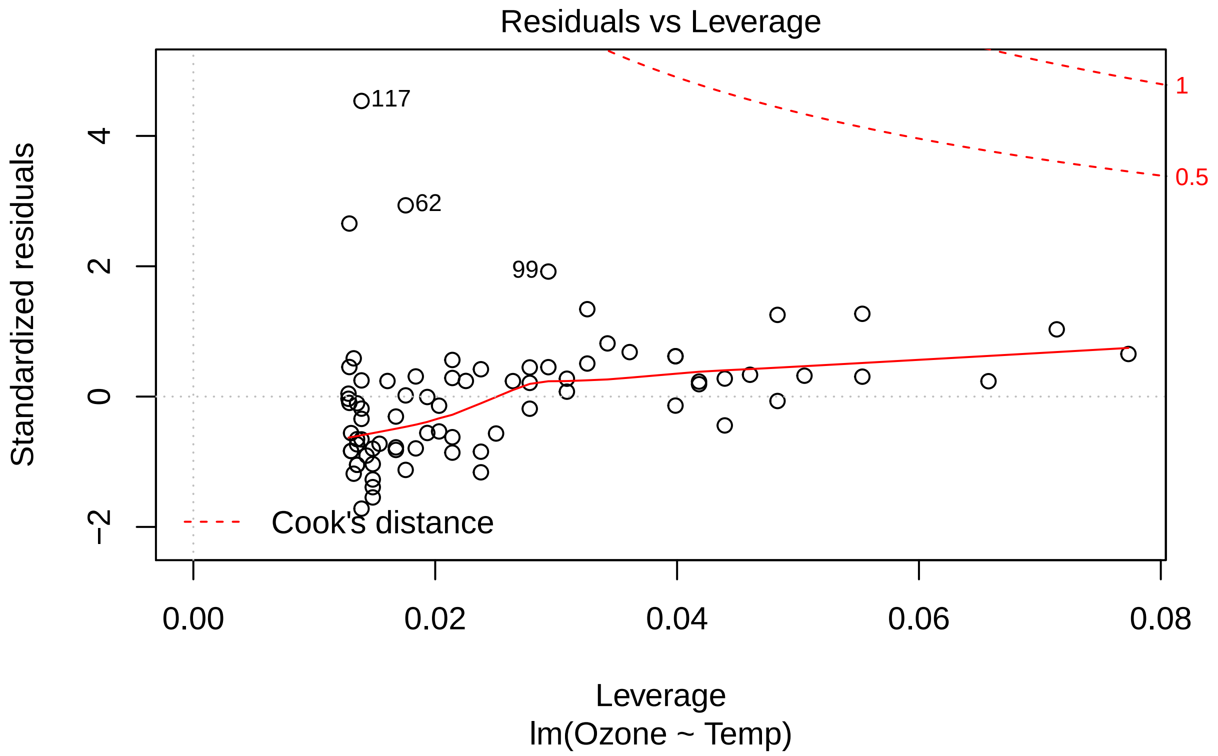
We can compute a simple regression line for the Ozone vs Temp by providing the values, as in a subset.

```
model1 <- lm(Ozone ~ Temp, data = ozone, subset = trainset)
plot(model1)
```


Residuals vs Fitted
Residuals
100
50
0
−50
0
20
40
60
80
Fitted values
lm(Ozone ~ Temp)
117
30
62


Normal Q−Q
Standardized residuals
5
4
3
2
1
0
−1
117
62
30
−2
−1
0
1
2
Theoretical Quantiles
lm(Ozone ~ Temp)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
20
40
60
80
117
62
30
Fitted values
lm(Ozone ~ Temp)



```r
summary(model1)
```

```
##
## Call:
## lm(formula = Ozone ~ Temp, data = ozone, subset = trainset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.696 -18.562  -0.503  10.054 115.304
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -159.134     24.119  -6.598 4.98e-09 ***
## Temp           2.615      0.306   8.546 9.91e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.6 on 76 degrees of freedom
## Multiple R-squared:   0.49,  Adjusted R-squared:  0.4833
## F-statistic: 73.03 on 1 and 76 DF,  p-value: 9.91e-13
```

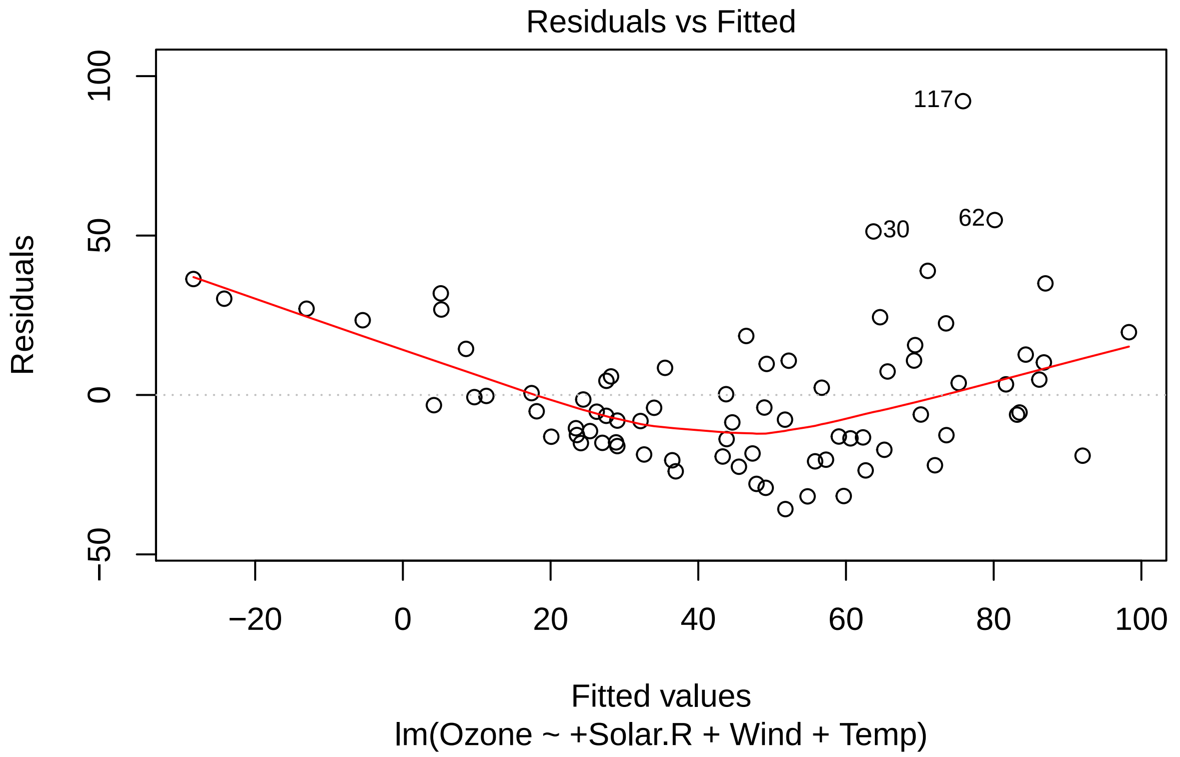
The model above is achieved by using the lm() function in R and the output is called using the summary() function on the model.The model above is also telling us that Adjusted R-squared is 0.4833.

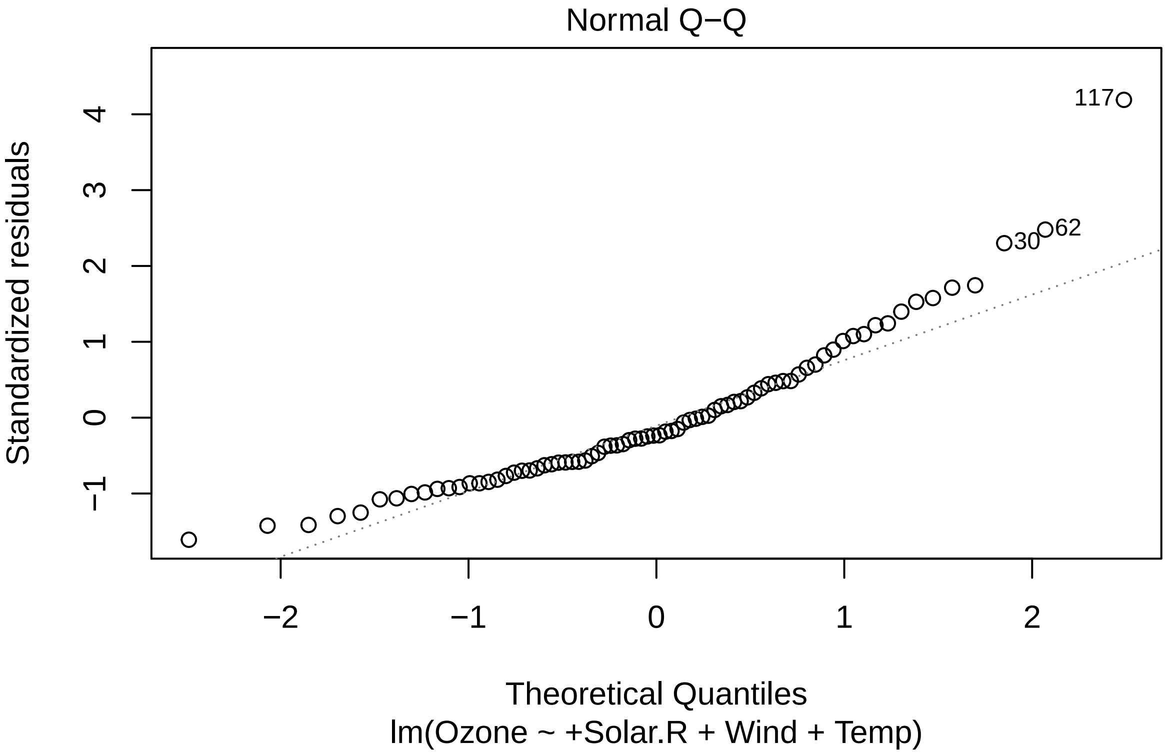
```r
coef(model1)
```

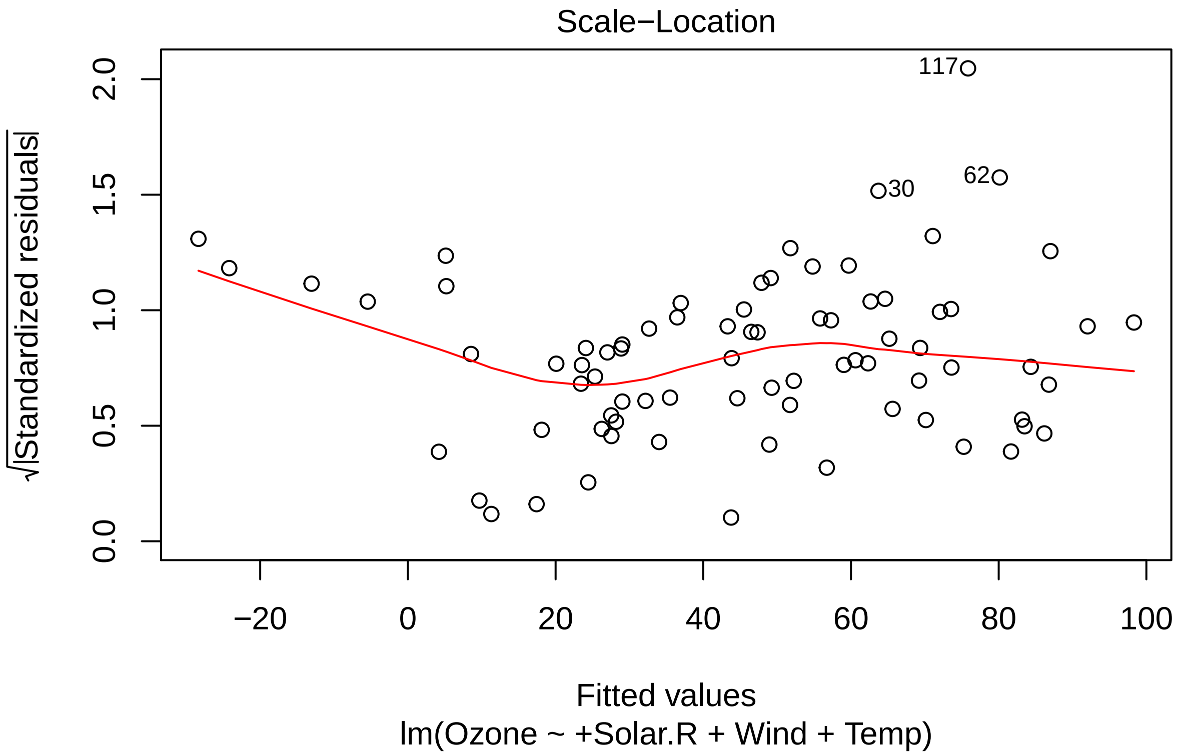
```
## (Intercept)        Temp
## -159.133578    2.615176
```

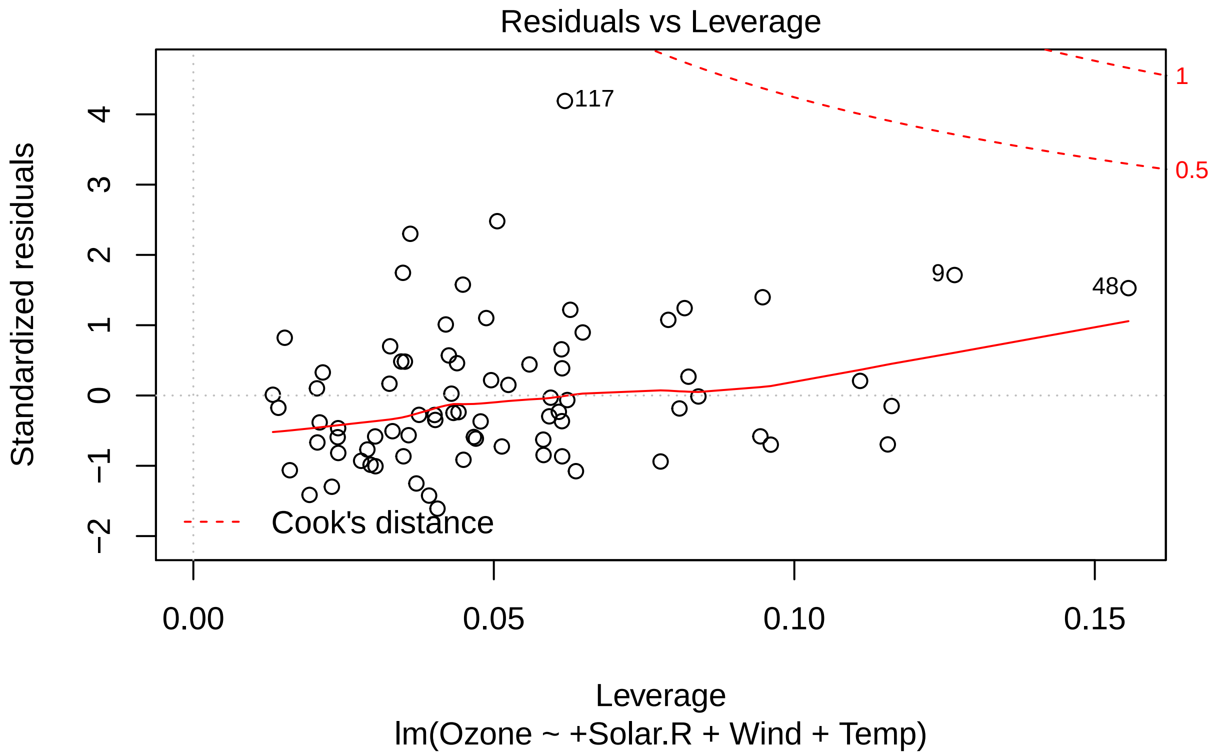
Model 2 - we model the relationship between Ozone, Solar.R, Wind and Temp.

```
model2 <- lm(Ozone ~ + Solar.R + Wind + Temp, data = ozone, subset = trainset)
plot(model2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Ozone ~ +Solar.R + Wind + Temp)
−1
0
1
2
3
4
−2
−1
0
1
2
117
30
62


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
−20
0
20
40
60
80
100
117
62
30
Fitted values
lm(Ozone ~ +Solar.R + Wind + Temp)



```r
summary(model2)
```

```
##
## Call:
## lm(formula = Ozone ~ +Solar.R + Wind + Temp, data = ozone, subset = trainset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.794 -14.977  -5.182  10.612  92.149
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -51.18946   32.57177  -1.572   0.1203
## Solar.R       0.07189    0.02977   2.415   0.0182 *
## Wind         -3.49641    0.82955  -4.215 6.98e-05 ***
## Temp          1.50391    0.35843   4.196 7.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.7 on 74 degrees of freedom
## Multiple R-squared:  0.6093, Adjusted R-squared:  0.5935
## F-statistic: 38.47 on 3 and 74 DF,  p-value: 4.304e-15
```

The model above is achieved by using the lm() function in R and the output is called using the summary() function on the model.The model above is also telling us that Adjusted R-squared is 0.5935.

```r
coef(model2)
```

```
##  (Intercept)      Solar.R         Wind         Temp 
## -51.18946407   0.07189438  -3.49641385   1.50391458
```