calscreening2018

Group 21; Group name: 21 and me; Members: Pluto Zhang, Linfeng Hu, Cynthia Ma

2022-10-26

```
library(rstudioapi)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(gam)
```

```
## Loading required package: splines
## Loading required package: foreach
##
## Attaching package: 'foreach'
##
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
##
## Loaded gam 1.20.2
```

```
library(splines)
library(splines2)
library(dplyr)
library(tidyr)
library(broom)
library(dslabs)
library(ggplot2)
library(ggthemes)
library(ggrepel)
library(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:dplyr':
```

```
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(nnet)
library(VGAM)
```

```
## Loading required package: stats4
##
## Attaching package: 'VGAM'
##
## The following object is masked from 'package:gam':
##
##     s
```

```
data_cal <- read_csv('calenviroscreen-3.0-results-june-2018-update.csv')
```

```
## Rows: 8035 Columns: 57
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): California County, Nearby City
## (to help approximate location only...
## dbl (53): Census Tract, Total Population, ZIP, Longitude, Latitude, CES 3.0 ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#data_cal
```

```
summary(data_cal)
```

```
##   Census Tract       Total Population California County       ZIP
##  Min.   :6.001e+09   Min.   :    0    Length:8035        Min.   :   32
##  1st Qu.:6.037e+09   1st Qu.: 3358    Class :character   1st Qu.:91602
##  Median :6.059e+09   Median : 4413    Mode  :character   Median :92691
##  Mean   :6.055e+09   Mean   : 4636                       Mean   :92837
##  3rd Qu.:6.073e+09   3rd Qu.: 5656                       3rd Qu.:94558
##  Max.   :6.115e+09   Max.   :37452                       Max.   :96161
##
##  Nearby City \n(to help approximate location only)   Longitude
##  Length:8035                                        Min.   :-124.3
##  Class :character                                   1st Qu.:-121.5
##  Mode  :character                                   Median :-118.4
##                                                     Mean   :-119.4
##                                                     3rd Qu.:-117.9
##                                                     Max.   :-114.3
##
##     Latitude     CES 3.0 Score   CES 3.0 Percentile CES 3.0 \nPercentile Range
```

```
##  Min.   :32.55   Min.   : 0.98   Min.   :  0.01     Length:8035
##  1st Qu.:33.92   1st Qu.:14.96   1st Qu.: 25.01     Class :character
##  Median :34.21   Median :25.06   Median : 50.01     Mode  :character
##  Mean   :35.50   Mean   :27.93   Mean   : 50.01
##  3rd Qu.:37.63   3rd Qu.:39.35   3rd Qu.: 75.00
##  Max.   :41.95   Max.   :94.09   Max.   :100.00
##                  NA's   :106     NA's   :106
##  SB 535 Disadvantaged Community     Ozone           Ozone Pctl
##  Length:8035                    Min.   :0.02600   Min.   :  0.24
##  Class :character               1st Qu.:0.04000   1st Qu.: 25.87
##  Mode  :character               Median :0.04600   Median : 53.02
##                                 Mean   :0.04743   Mean   : 53.30
##                                 3rd Qu.:0.05500   3rd Qu.: 77.87
##                                 Max.   :0.06800   Max.   :100.00
##
##      PM2.5          PM2.5 Pctl       Diesel PM       Diesel PM Pctl
##  Min.   : 1.651   Min.   :  0.01   Min.   :  0.021   Min.   :  0.01
##  1st Qu.: 8.698   1st Qu.: 30.70   1st Qu.:  8.812   1st Qu.: 25.01
##  Median :10.370   Median : 52.61   Median : 16.448   Median : 50.01
##  Mean   :10.378   Mean   : 53.59   Mean   : 19.196   Mean   : 50.02
##  3rd Qu.:12.050   3rd Qu.: 81.66   3rd Qu.: 24.646   3rd Qu.: 75.00
##  Max.   :19.600   Max.   :100.00   Max.   :253.731   Max.   :100.00
##  NA's   :19       NA's   :19
##  Drinking Water    Drinking Water Pctl   Pesticides       Pesticides Pctl
##  Min.   :   6.92   Min.   :  0.01      Min.   :    0.00   Min.   :  0.00
##  1st Qu.: 249.35   1st Qu.: 25.01      1st Qu.:    0.00   1st Qu.:  0.00
##  Median : 479.23   Median : 51.02      Median :    0.00   Median :  0.00
##  Mean   : 472.37   Mean   : 50.34      Mean   :  313.97   Mean   : 17.98
##  3rd Qu.: 664.07   3rd Qu.: 78.57      3rd Qu.:    0.37   3rd Qu.: 30.45
##  Max.   :1245.65   Max.   :100.00      Max.   :91316.19   Max.   :100.00
##  NA's   :18        NA's   :18
##   Tox. Release      Tox. Release Pctl    Traffic          Traffic Pctl
##  Min.   :     0.0   Min.   :  0.00    Min.   :   22.41   Min.   :  0.01
##  1st Qu.:    94.8   1st Qu.: 24.85    1st Qu.:  442.08   1st Qu.: 25.01
##  Median :   474.0   Median : 49.90    Median :  699.89   Median : 50.01
##  Mean   :  3182.7   Mean   : 49.90    Mean   :  943.04   Mean   : 50.01
##  3rd Qu.:  3474.2   3rd Qu.: 74.95    3rd Qu.: 1190.08   3rd Qu.: 75.00
##  Max.   :842751.3   Max.   :100.00    Max.   :45687.87   Max.   :100.00
##                                       NA's   :56         NA's   :56
##  Cleanup Sites    Cleanup Sites Pctl Groundwater Threats
##  Min.   :  0.00   Min.   :  0.00     Min.   :   0.0
##  1st Qu.:  0.00   1st Qu.:  0.00     1st Qu.:   0.2
##  Median :  2.00   Median : 27.29     Median :   5.6
##  Mean   :  8.37   Mean   : 34.45     Mean   :  15.7
##  3rd Qu.: 10.30   3rd Qu.: 63.41     3rd Qu.:  17.8
##  Max.   :323.75   Max.   :100.00     Max.   :1610.2
##
##  Groundwater Threats Pctl   Haz. Waste      Haz. Waste Pctl  Imp. Water Bodies
##  Min.   :  0.00           Min.   : 0.0000   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:  0.33           1st Qu.: 0.0000   1st Qu.:  0.00   1st Qu.: 0.000
##  Median : 33.60           Median : 0.0500   Median : 25.76   Median : 1.000
##  Mean   : 38.02           Mean   : 0.4534   Mean   : 34.71   Mean   : 3.279
##  3rd Qu.: 66.78           3rd Qu.: 0.2250   3rd Qu.: 63.00   3rd Qu.: 6.000
##  Max.   :100.00           Max.   :28.6950   Max.   :100.00   Max.   :34.000
```

```
##
##  Imp. Water Bodies Pctl  Solid Waste     Solid Waste Pctl Pollution Burden
##  Min.   :  0.00         Min.   : 0.000   Min.   :  0.00   Min.   : 8.37
##  1st Qu.:  0.00         1st Qu.: 0.000   1st Qu.:  0.00   1st Qu.:32.24
##  Median : 15.26         Median : 0.200   Median :  9.08   Median :41.80
##  Mean   : 30.68         Mean   : 2.233   Mean   : 27.33   Mean   :41.97
##  3rd Qu.: 63.17         3rd Qu.: 2.250   3rd Qu.: 52.84   3rd Qu.:51.02
##  Max.   :100.00         Max.   :97.800   Max.   :100.00   Max.   :81.19
##
##  Pollution Burden Score Pollution Burden Pctl     Asthma        Asthma Pctl
##  Min.   : 1.030         Min.   :  0.01        Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 3.970         1st Qu.: 25.01        1st Qu.: 29.86   1st Qu.: 24.88
##  Median : 5.150         Median : 50.01        Median : 45.27   Median : 49.93
##  Mean   : 5.169         Mean   : 50.01        Mean   : 51.98   Mean   : 49.93
##  3rd Qu.: 6.280         3rd Qu.: 75.00        3rd Qu.: 65.99   3rd Qu.: 74.96
##  Max.   :10.000         Max.   :100.00        Max.   :278.83   Max.   :100.00
##
##  Low Birth Weight Low Birth Weight Pctl Cardiovascular Disease
##  Min.   : 0.000   Min.   :  0.00        Min.   : 0.000
##  1st Qu.: 3.950   1st Qu.: 24.98        1st Qu.: 6.080
##  Median : 4.920   Median : 50.22        Median : 7.940
##  Mean   : 4.976   Mean   : 50.04        Mean   : 8.266
##  3rd Qu.: 5.930   3rd Qu.: 75.06        3rd Qu.:10.040
##  Max.   :14.890   Max.   :100.00        Max.   :21.260
##  NA's   :222      NA's   :222
##  Cardiovascular Disease Pctl   Education     Education Pctl
##  Min.   :  0.00              Min.   : 0.00   Min.   :  0.00
##  1st Qu.: 25.00              1st Qu.: 6.30   1st Qu.: 25.08
##  Median : 49.96              Median :14.00   Median : 50.00
##  Mean   : 49.98              Mean   :19.12   Mean   : 50.05
##  3rd Qu.: 75.07              3rd Qu.:28.70   3rd Qu.: 74.99
##  Max.   :100.00              Max.   :80.00   Max.   :100.00
##                              NA's   :96      NA's   :96
##  Linguistic Isolation Linguistic Isolation Pctl    Poverty
##  Min.   : 0.00        Min.   :  0.00            Min.   : 0.00
##  1st Qu.: 3.00        1st Qu.: 22.52            1st Qu.:19.20
##  Median : 7.40        Median : 48.34            Median :33.50
##  Mean   :10.42        Mean   : 48.36            Mean   :36.39
##  3rd Qu.:14.90        3rd Qu.: 74.23            3rd Qu.:51.50
##  Max.   :72.30        Max.   :100.00            Max.   :96.20
##  NA's   :242          NA's   :242               NA's   :79
##   Poverty Pctl     Unemployment    Unemployment Pctl Housing Burden
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00    Min.   : 1.00
##  1st Qu.: 25.10   1st Qu.:  6.60   1st Qu.: 25.46    1st Qu.:12.90
##  Median : 50.11   Median :  9.30   Median : 50.27    Median :18.00
##  Mean   : 50.07   Mean   : 10.21   Mean   : 50.32    Mean   :19.33
##  3rd Qu.: 75.02   3rd Qu.: 12.90   3rd Qu.: 75.52    3rd Qu.:24.40
##  Max.   :100.00   Max.   :100.00   Max.   :100.00    Max.   :67.20
##  NA's   :79       NA's   :155      NA's   :155       NA's   :157
##  Housing Burden Pctl   Pop. Char.    Pop. Char. Score Pop. Char. Pctl
##  Min.   :  0.03      Min.   : 2.53   Min.   : 0.260   Min.   :  0.01
##  1st Qu.: 25.51      1st Qu.:33.76   1st Qu.: 3.500   1st Qu.: 25.01
##  Median : 50.33      Median :49.96   Median : 5.180   Median : 50.01
##  Mean   : 50.18      Mean   :49.89   Mean   : 5.174   Mean   : 50.01
```

```
##  3rd Qu.: 75.01      3rd Qu.:66.45   3rd Qu.: 6.890   3rd Qu.: 75.00
##  Max.   :100.00      Max.   :96.43   Max.   :10.000   Max.   :100.00
##  NA's   :157         NA's   :106     NA's   :106      NA's   :106
```

```
#names(data_cal)
colnames(data_cal) <- gsub(" ", "", colnames(data_cal))
colnames(data_cal) <- gsub("\n", "", colnames(data_cal))
#names(data_cal)
```

---

columns for diseases: Asthma LowBirthWeight CardiovascularDisease

columns for socio-economic elements: ducation LinguisticIsolation Poverty Unemployment HousingBurden Pop.Char.

columns for air pollution elements: Ozone PM2.5 DieselPM DrinkingWater Pesticides Tox.Release Traffic CleanupSites GroundwaterThreats Haz.Waste Imp.WaterBodies SolidWaste PollutionBurden *****

```
air_pollutants_vec <- names(data_cal)[12:38]
air_pollutants_vec
```

```
##  [1] "Ozone"                 "OzonePctl"             "PM2.5"
##  [4] "PM2.5Pctl"             "DieselPM"              "DieselPMPctl"
##  [7] "DrinkingWater"         "DrinkingWaterPctl"     "Pesticides"
## [10] "PesticidesPctl"        "Tox.Release"           "Tox.ReleasePctl"
## [13] "Traffic"               "TrafficPctl"           "CleanupSites"
## [16] "CleanupSitesPctl"      "GroundwaterThreats"    "GroundwaterThreatsPctl"
## [19] "Haz.Waste"             "Haz.WastePctl"         "Imp.WaterBodies"
## [22] "Imp.WaterBodiesPctl"   "SolidWaste"            "SolidWastePctl"
## [25] "PollutionBurden"       "PollutionBurdenScore"  "PollutionBurdenPctl"
```

```
disease_vec <- names(data_cal)[39:44]
disease_vec
```

```
## [1] "Asthma"                  "AsthmaPctl"
## [3] "LowBirthWeight"          "LowBirthWeightPctl"
## [5] "CardiovascularDisease"   "CardiovascularDiseasePctl"
```

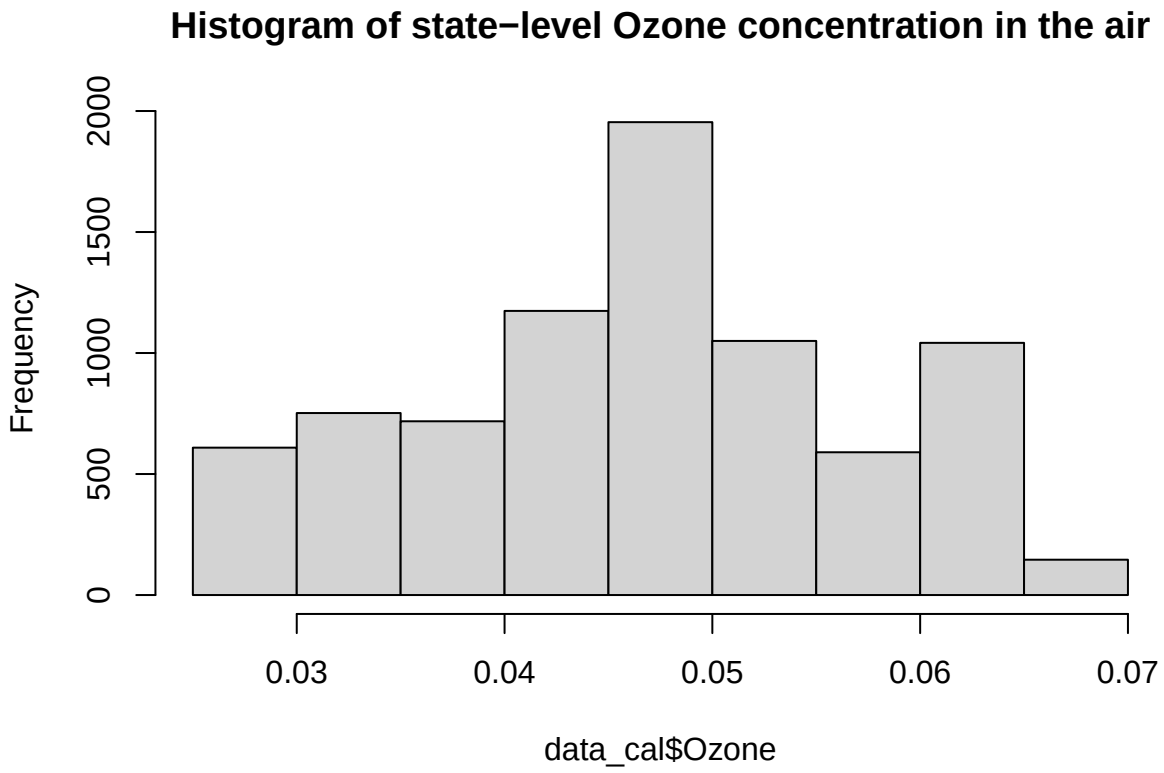
```
soecon_vec <- names(data_cal)[45:57]
soecon_vec
```

```
##  [1] "Education"              "EducationPctl"
##  [3] "LinguisticIsolation"    "LinguisticIsolationPctl"
##  [5] "Poverty"                "PovertyPctl"
##  [7] "Unemployment"           "UnemploymentPctl"
##  [9] "HousingBurden"          "HousingBurdenPctl"
## [11] "Pop.Char."              "Pop.Char.Score"
## [13] "Pop.Char.Pctl"
```

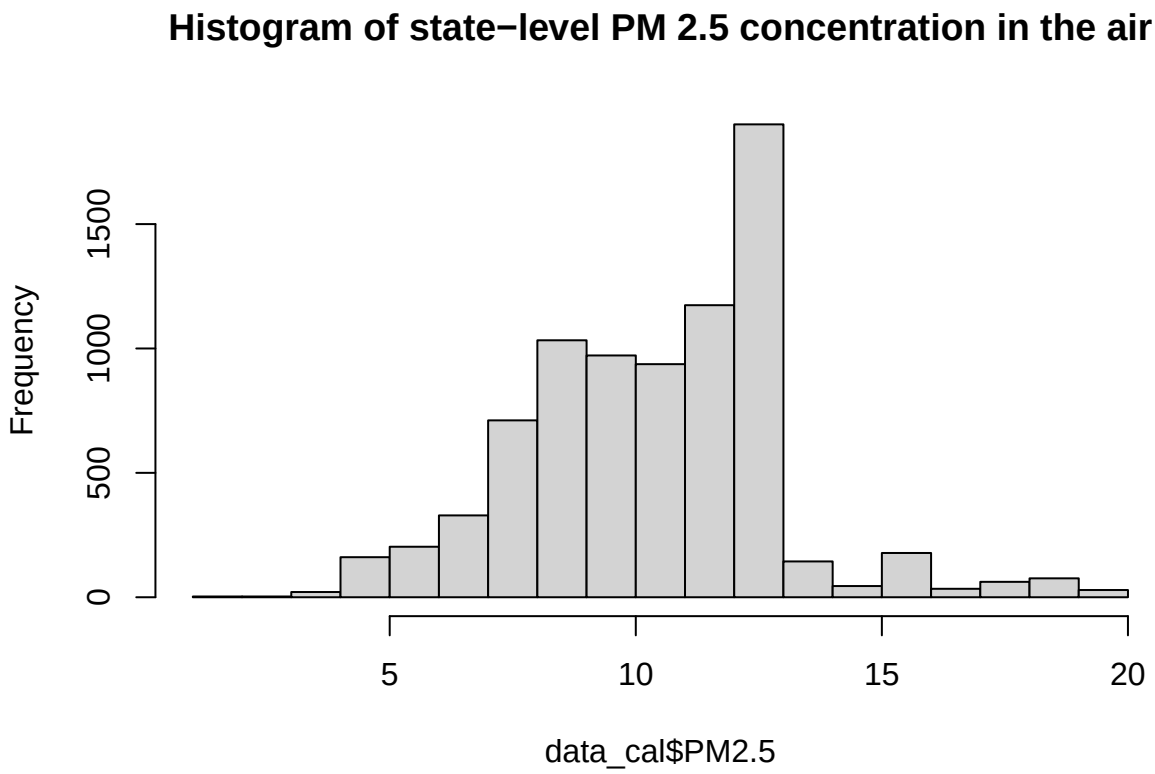
EDA

EDA for Pollution factors

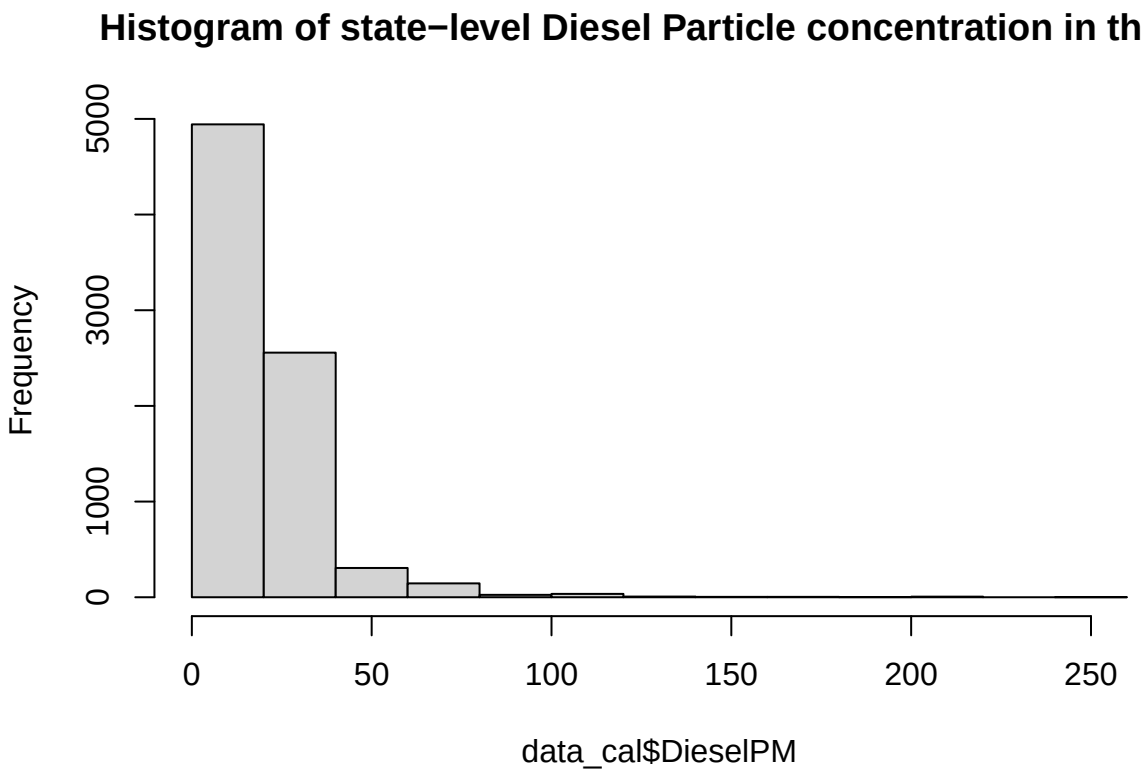
```
#histograms for the air pollution factors
hist(data_cal$Ozone, main='Histogram of state-level Ozone concentration in the air' )
```



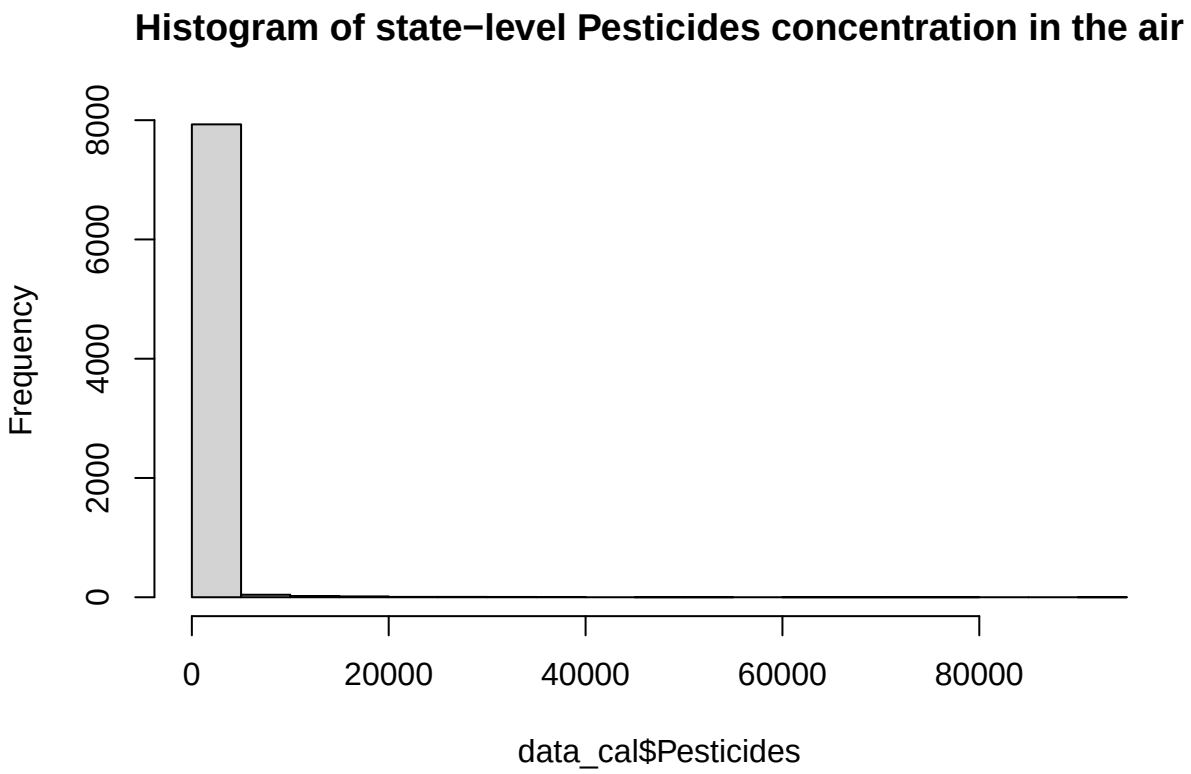

```
hist(data_cal$PM2.5, main='Histogram of state-level PM 2.5 concentration in the air' )
```



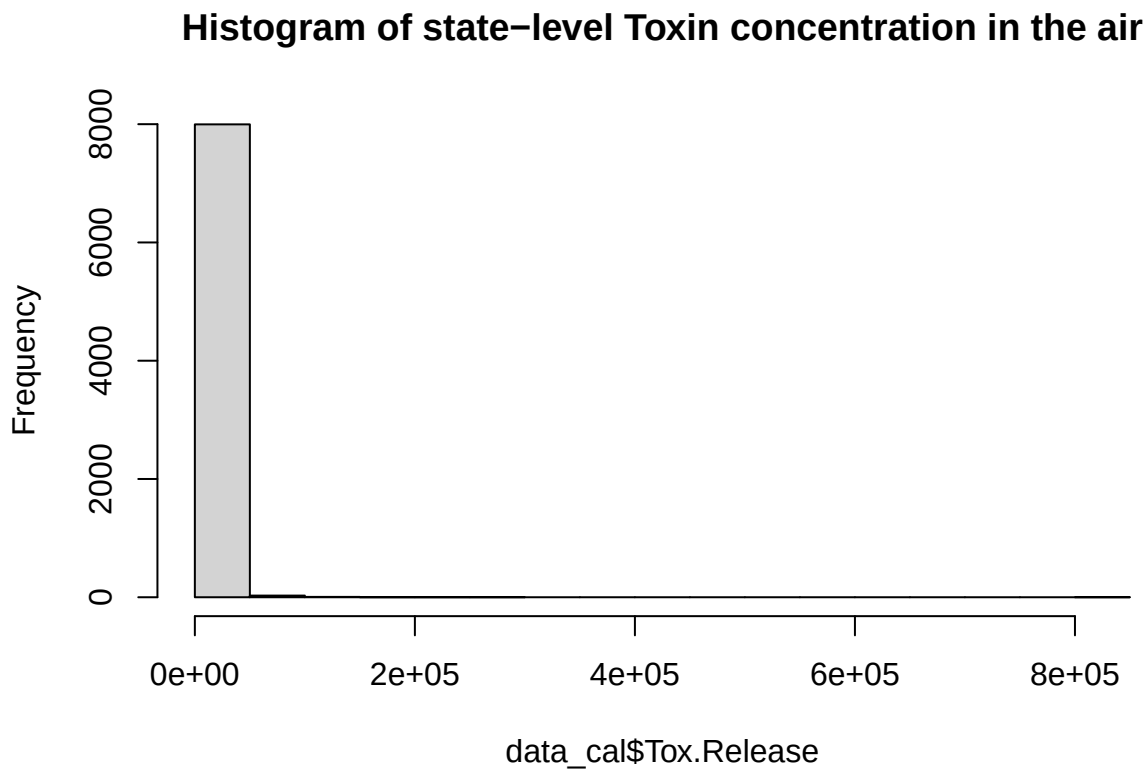

```
hist(data_cal$DieselPM, main='Histogram of state-level Diesel Particle concentration in the air' )
```



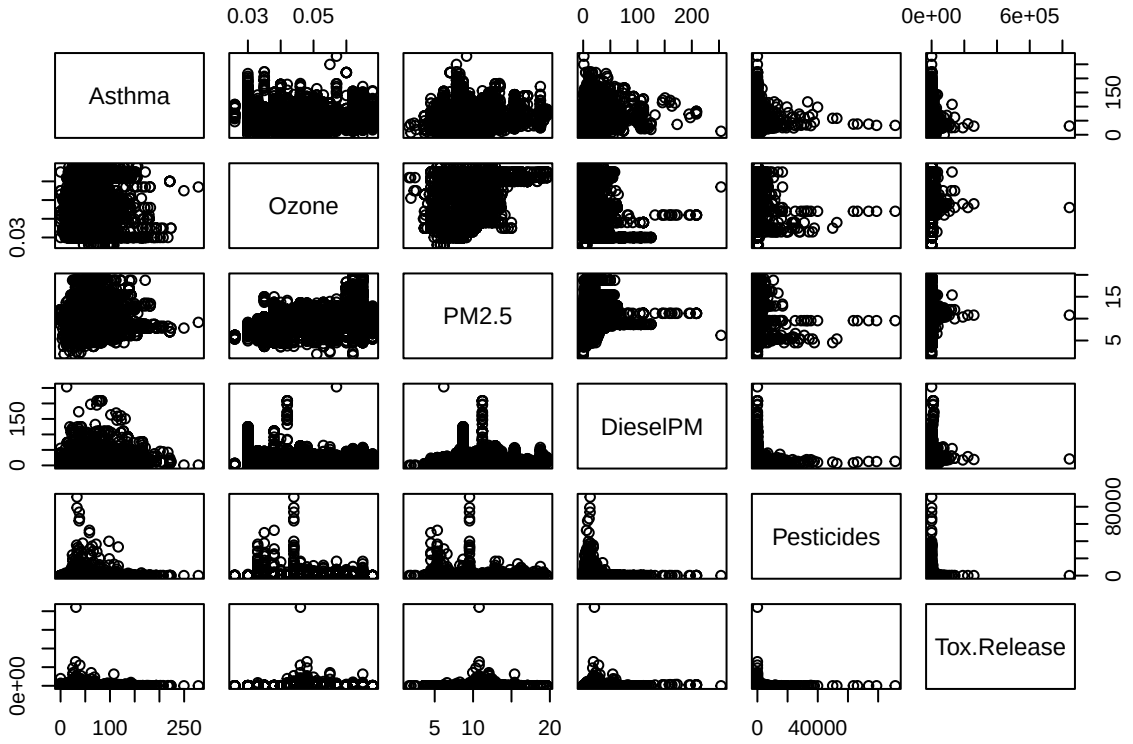

```
hist(data_cal$Pesticides, main='Histogram of state-level Pesticides concentration in the air' )
```




```
hist(data_cal$Tox.Release, main='Histogram of state-level Toxin concentration in the air' )
```

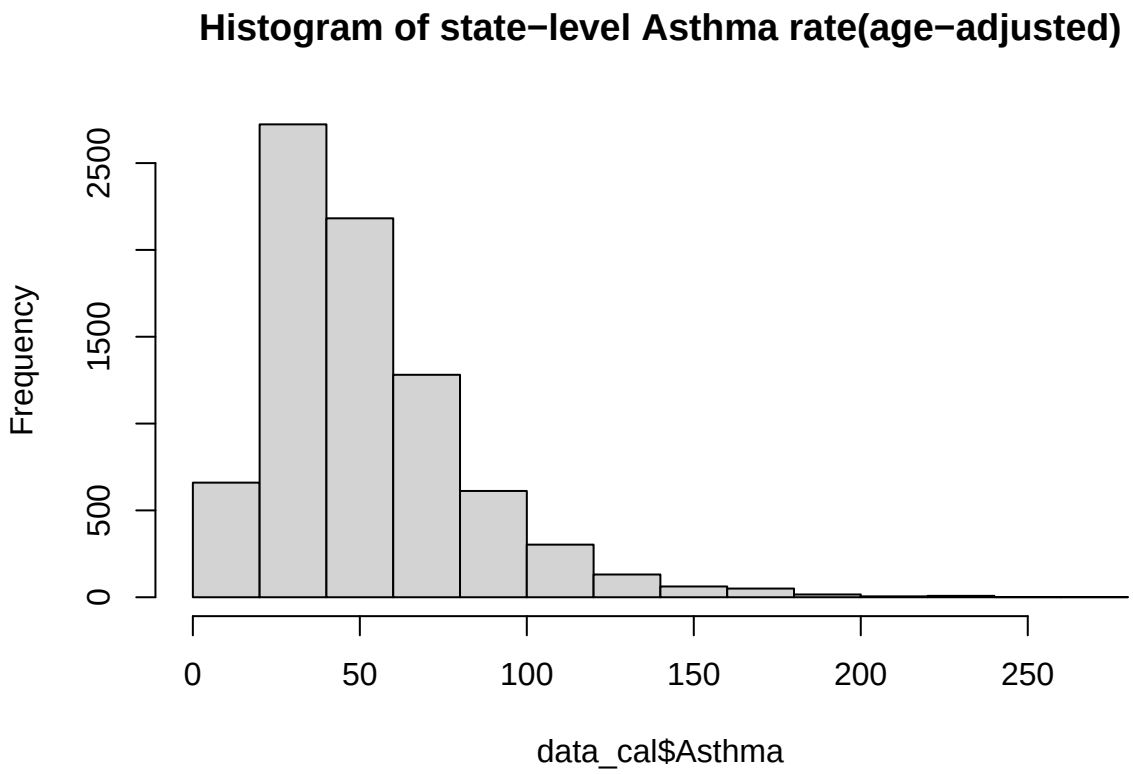



```
pairs(Asthma ~ Ozone + PM2.5 + DieselPM + Pesticides + Tox.Release , dat = data_cal)
```

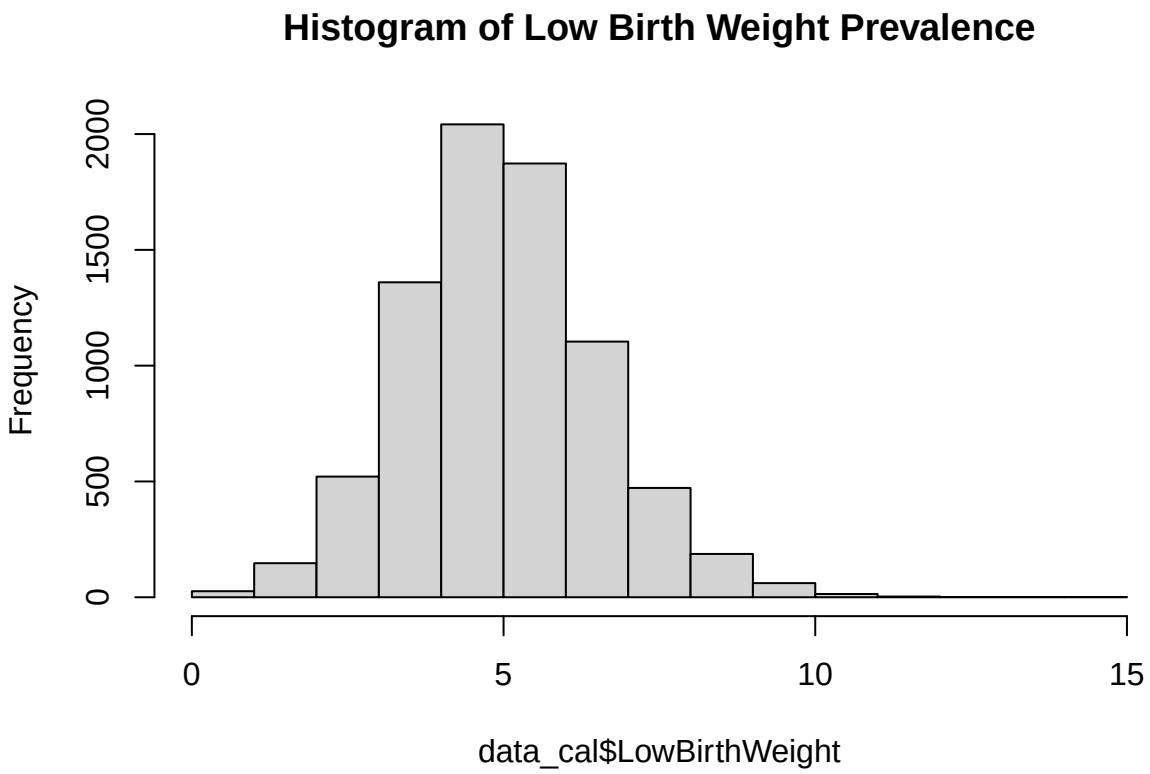


for other diseases

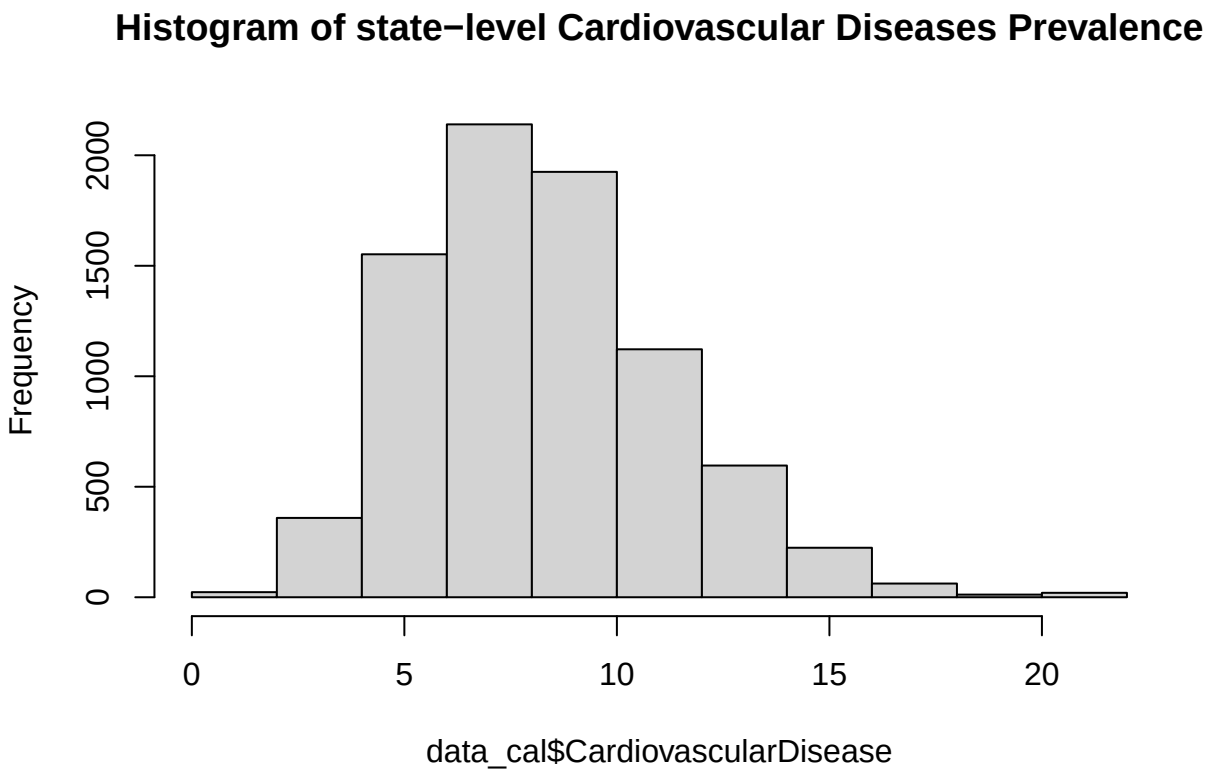
```
#histograms for the other diseases factors
hist(data_cal$Asthma, main='Histogram of state-level Asthma rate(age-adjusted)' )
```



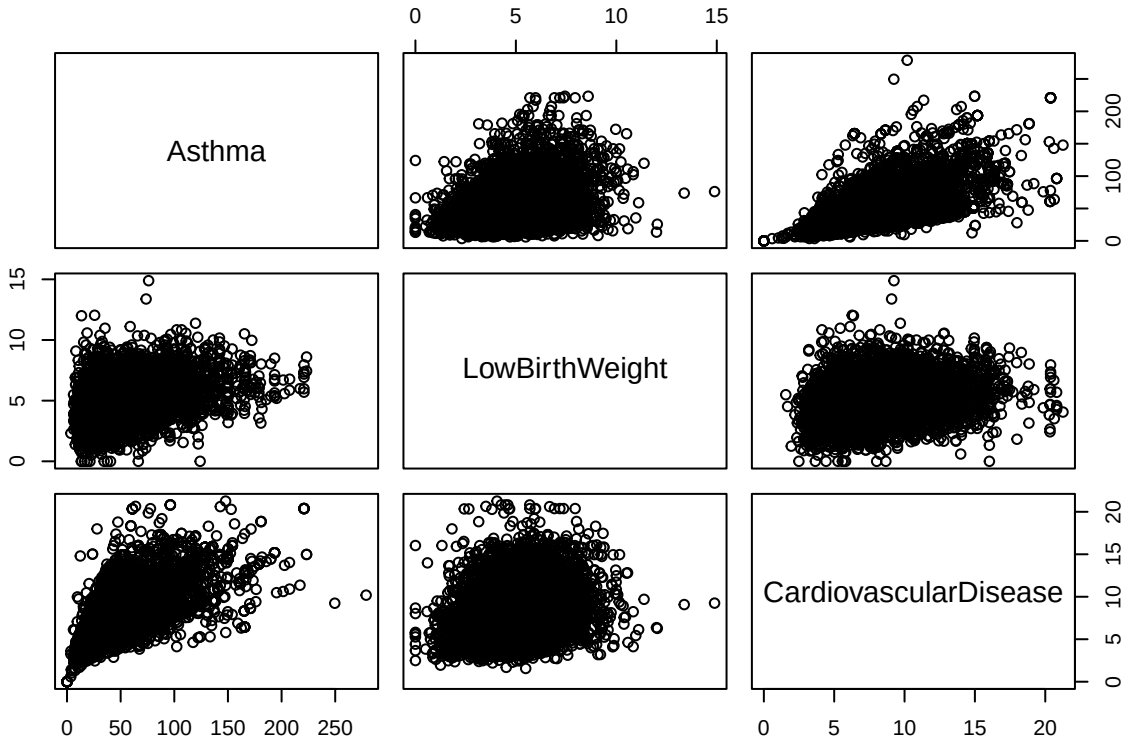

```
hist(data_cal$LowBirthWeight, main='Histogram of Low Birth Weight Prevalence' )
```




```
hist(data_cal$CardiovascularDisease, main='Histogram of state-level Cardiovascular Diseases Prevalence'
```

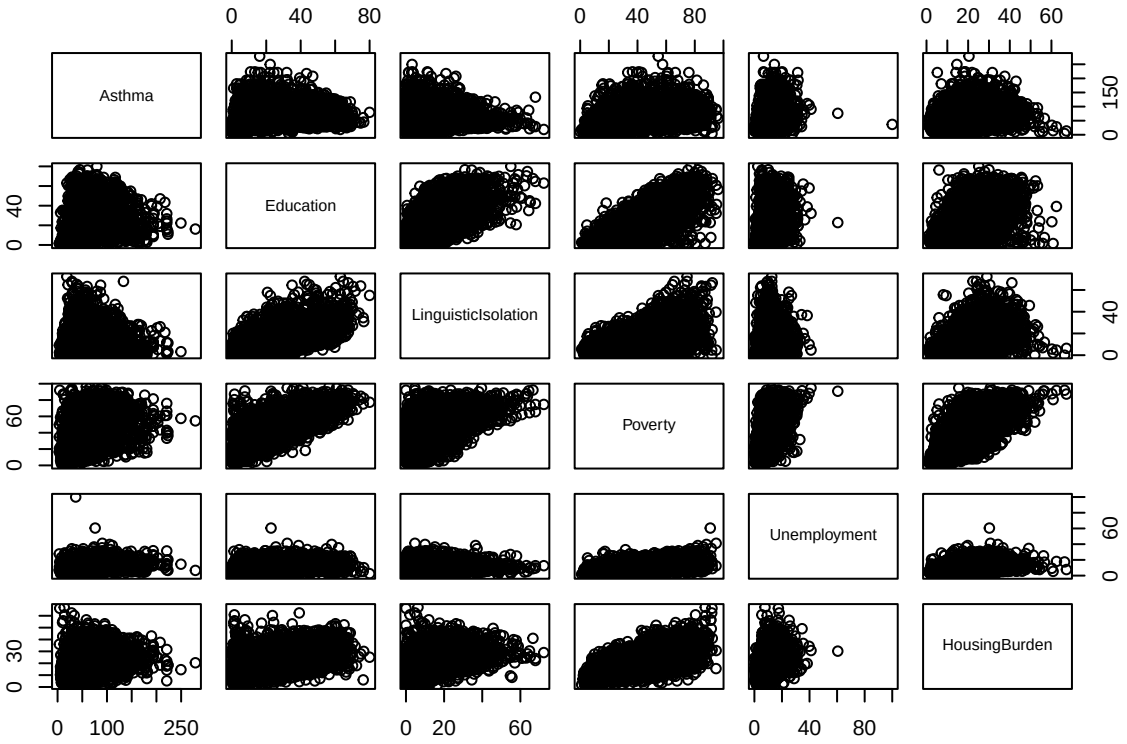



```
pairs(Asthma ~ LowBirthWeight + CardiovascularDisease , dat = data_cal)
```



#for socioeconomic factors

```
pairs(Asthma ~ Education + LinguisticIsolation + Poverty + Unemployment + HousingBurden, dat = data_cal
```




```
#pairs(Pop.Char.~ Education + LinguisticIsolation + Poverty + Unemployment + HousingBurden, dat = data_
```

```
#forward selection
require(broom)
data_cal1 <- na.omit(data_cal)
data_cal_old <- data_cal1
#forward selection procedure using AIC values
lm1 <- lm(Asthma ~ 1, data=data_cal1)
stepModel <- step(lm1, direction="forward",
scope=(~ Ozone + PM2.5 + DieselPM+ DrinkingWater+Pesticides+Tox.Release+Traffic +CleanupSites + Groundwa
SolidWaste + Education+LinguisticIsolation+Poverty+Unemployment+HousingBurden+ LowBirthWeight+Cardiovas
```

```
## Start:  AIC=51556.08
## Asthma ~ 1
##
##                          Df Sum of Sq     RSS   AIC
## + CardiovascularDisease   1   3081203 3854963 47119
## + Poverty                 1   1714938 5221228 49412
## + Unemployment            1   1390708 5545458 49867
## + Education               1   1087079 5849087 50270
## + LowBirthWeight          1    836269 6099897 50587
## + HousingBurden           1    808030 6128136 50622
## + DieselPM                1    248241 6687925 51283
## + LinguisticIsolation     1    217017 6719149 51318
## + CleanupSites            1    125110 6811056 51421
```

```
## + GroundwaterThreats     1    104527 6831639 51443
## + Haz.Waste              1     73133 6863033 51478
## + PM2.5                  1     56975 6879191 51496
## + Ozone                  1     30323 6905843 51525
## + SolidWaste             1     29046 6907120 51526
## + DrinkingWater          1     22675 6913491 51533
## + Tox.Release            1     12218 6923948 51545
## + Traffic                1      4747 6931419 51553
## + Imp.WaterBodies        1      3476 6932690 51554
## <none>                              6936166 51556
## + Pesticides             1        46 6936120 51558
##
## Step:  AIC=47119.19
## Asthma ~ CardiovascularDisease
##
##                       Df Sum of Sq     RSS   AIC
## + Poverty              1    386188 3468775 46323
## + Ozone                1    347611 3507352 46407
## + DieselPM             1    313537 3541425 46480
## + HousingBurden        1    296816 3558147 46516
## + LowBirthWeight       1    268075 3586888 46577
## + DrinkingWater        1    242008 3612955 46631
## + Unemployment         1    190404 3664558 46738
## + Education            1    164490 3690473 46792
## + GroundwaterThreats   1    107469 3747494 46908
## + CleanupSites         1     88893 3766070 46945
## + LinguisticIsolation  1     76369 3778594 46970
## + Imp.WaterBodies      1     51096 3803867 47020
## + Haz.Waste            1     38408 3816554 47046
## + Tox.Release          1      7980 3846982 47106
## + SolidWaste           1      4115 3850847 47113
## + Traffic              1      2197 3852766 47117
## <none>                             3854963 47119
## + PM2.5                1       872 3854090 47119
## + Pesticides           1        24 3854939 47121
##
## Step:  AIC=46323.48
## Asthma ~ CardiovascularDisease + Poverty
##
##                       Df Sum of Sq     RSS   AIC
## + Ozone                1    403395 3065380 45391
## + DrinkingWater        1    321933 3146842 45589
## + DieselPM             1    182534 3286241 45917
## + LowBirthWeight       1    137076 3331699 46021
## + GroundwaterThreats   1     80592 3388182 46148
## + Imp.WaterBodies      1     57006 3411769 46200
## + CleanupSites         1     50793 3417982 46214
## + Haz.Waste            1     26324 3442451 46268
## + PM2.5                1     24002 3444773 46273
## + LinguisticIsolation  1     23978 3444797 46273
## + HousingBurden        1     21374 3447400 46279
## + Education            1     20882 3447893 46280
## + Unemployment         1     19050 3449725 46284
## + Tox.Release          1     15958 3452817 46291
```

```
## <none>                              3468775 46323
## + Pesticides           1       250 3468525 46325
## + Traffic              1       149 3468625 46325
## + SolidWaste           1        11 3468764 46325
##
## Step:  AIC=45391.2
## Asthma ~ CardiovascularDisease + Poverty + Ozone
##
##                       Df Sum of Sq     RSS   AIC
## + LowBirthWeight       1    130838 2934542 45064
## + DieselPM             1     76607 2988772 45202
## + DrinkingWater        1     72582 2992797 45212
## + LinguisticIsolation  1     67364 2998016 45225
## + Unemployment         1     50829 3014551 45267
## + Education            1     35111 3030268 45306
## + GroundwaterThreats   1     26039 3039341 45329
## + CleanupSites         1     16413 3048967 45353
## + Tox.Release          1     12568 3052811 45362
## + PM2.5                1      9621 3055759 45369
## + Haz.Waste            1      9396 3055984 45370
## + Imp.WaterBodies      1      7241 3058139 45375
## + HousingBurden        1      3124 3062256 45385
## + Traffic              1      1720 3063659 45389
## + Pesticides           1      1313 3064067 45390
## <none>                              3065380 45391
## + SolidWaste           1       126 3065254 45393
##
## Step:  AIC=45063.57
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight
##
##                       Df Sum of Sq     RSS   AIC
## + LinguisticIsolation  1     74104 2860438 44872
## + DrinkingWater        1     71002 2863540 44880
## + DieselPM             1     53969 2880573 44925
## + Unemployment         1     39765 2894777 44962
## + Education            1     39680 2894862 44963
## + GroundwaterThreats   1     25146 2909396 45001
## + Tox.Release          1     12760 2921782 45033
## + CleanupSites         1     12550 2921993 45033
## + Imp.WaterBodies      1      8699 2925843 45043
## + Haz.Waste            1      7774 2926768 45046
## + Traffic              1      4332 2930210 45054
## + PM2.5                1      3209 2931333 45057
## <none>                              2934542 45064
## + Pesticides           1       629 2933913 45064
## + HousingBurden        1       468 2934074 45064
## + SolidWaste           1        79 2934463 45065
##
## Step:  AIC=44872.28
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight +
##     LinguisticIsolation
##
##                      Df Sum of Sq     RSS   AIC
## + DieselPM            1     81440 2778998 44656
```

```
## + DrinkingWater       1     55162 2805276 44727
## + Unemployment        1     22842 2837596 44814
## + GroundwaterThreats  1     20449 2839989 44820
## + CleanupSites        1     15441 2844997 44833
## + PM2.5               1     14690 2845748 44835
## + Haz.Waste           1      8842 2851596 44851
## + Tox.Release         1      8748 2851690 44851
## + Imp.WaterBodies     1      6711 2853727 44857
## + Education           1      4077 2856361 44864
## + HousingBurden       1      1878 2858560 44869
## + Traffic             1       946 2859492 44872
## <none>                            2860438 44872
## + Pesticides          1       446 2859992 44873
## + SolidWaste          1       396 2860042 44873
##
## Step:  AIC=44656
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight +
##     LinguisticIsolation + DieselPM
##
##                       Df Sum of Sq     RSS   AIC
## + DrinkingWater       1     49054 2729944 44523
## + Unemployment        1     28934 2750063 44579
## + GroundwaterThreats  1     17483 2761515 44610
## + Tox.Release         1     11318 2767679 44627
## + CleanupSites        1      9108 2769889 44633
## + Imp.WaterBodies     1      5782 2773216 44642
## + Traffic             1      5611 2773387 44643
## + Haz.Waste           1      4130 2774867 44647
## + Education           1      1933 2777064 44653
## + PM2.5               1      1780 2777218 44653
## <none>                            2778998 44656
## + HousingBurden       1       152 2778846 44658
## + Pesticides          1         4 2778994 44658
## + SolidWaste          1         2 2778996 44658
##
## Step:  AIC=44523.42
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight +
##     LinguisticIsolation + DieselPM + DrinkingWater
##
##                       Df Sum of Sq     RSS   AIC
## + Unemployment        1   30720.6 2699223 44440
## + GroundwaterThreats  1   19517.1 2710427 44471
## + Tox.Release         1   10853.3 2719090 44495
## + CleanupSites        1   10806.6 2719137 44495
## + PM2.5               1    5747.9 2724196 44509
## + Imp.WaterBodies     1    5442.8 2724501 44510
## + Haz.Waste           1    5311.2 2724632 44511
## + Traffic             1    4395.1 2725549 44513
## + SolidWaste          1     746.5 2729197 44523
## <none>                            2729944 44523
## + Pesticides          1     711.1 2729233 44523
## + Education           1     646.5 2729297 44524
## + HousingBurden       1     108.4 2729835 44525
##
```

```
## Step:  AIC=44439.9
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight +
##     LinguisticIsolation + DieselPM + DrinkingWater + Unemployment
##
##                       Df Sum of Sq     RSS   AIC
## + GroundwaterThreats  1   19444.3 2679779 44387
## + CleanupSites        1   11796.9 2687426 44409
## + Tox.Release         1    9601.2 2689622 44415
## + PM2.5               1    5365.0 2693858 44427
## + Haz.Waste           1    4946.8 2694276 44428
## + Imp.WaterBodies     1    3985.3 2695238 44431
## + Traffic             1    3727.9 2695495 44431
## + Pesticides          1     991.1 2698232 44439
## + SolidWaste          1     751.6 2698472 44440
## <none>                            2699223 44440
## + Education           1     633.6 2698590 44440
## + HousingBurden       1      77.2 2699146 44442
##
## Step:  AIC=44387.26
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight +
##     LinguisticIsolation + DieselPM + DrinkingWater + Unemployment +
##     GroundwaterThreats
##
##                    Df Sum of Sq     RSS   AIC
## + Tox.Release      1    9376.8 2670402 44363
## + PM2.5            1    5715.9 2674063 44373
## + Traffic          1    3733.4 2676045 44379
## + CleanupSites     1    3622.7 2676156 44379
## + Imp.WaterBodies  1    2237.4 2677541 44383
## + Haz.Waste        1    1296.5 2678482 44386
## + Pesticides       1     902.4 2678876 44387
## <none>                         2679779 44387
## + HousingBurden    1     318.9 2679460 44388
## + Education        1     310.6 2679468 44388
## + SolidWaste       1       4.0 2679775 44389
##
## Step:  AIC=44362.77
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight +
##     LinguisticIsolation + DieselPM + DrinkingWater + Unemployment +
##     GroundwaterThreats + Tox.Release
##
##                    Df Sum of Sq     RSS   AIC
## + PM2.5            1    6837.7 2663564 44345
## + CleanupSites     1    5096.3 2665306 44350
## + Traffic          1    2811.0 2667591 44357
## + Haz.Waste        1    1894.8 2668507 44359
## + Imp.WaterBodies  1    1854.5 2668547 44360
## + Pesticides       1     766.2 2669636 44363
## <none>                         2670402 44363
## + HousingBurden    1     406.8 2669995 44364
## + Education        1      82.2 2670320 44365
## + SolidWaste       1      10.0 2670392 44365
##
## Step:  AIC=44345.4
```

```
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight +
##     LinguisticIsolation + DieselPM + DrinkingWater + Unemployment +
##     GroundwaterThreats + Tox.Release + PM2.5
##
##                   Df Sum of Sq     RSS   AIC
## + CleanupSites     1    4645.1 2658919 44334
## + Traffic          1    3803.5 2659761 44337
## + Imp.WaterBodies  1    2449.8 2661114 44340
## + Haz.Waste        1    1724.1 2661840 44343
## + Pesticides       1    1047.8 2662517 44344
## <none>                         2663564 44345
## + HousingBurden    1     401.3 2663163 44346
## + Education        1     383.8 2663180 44346
## + SolidWaste       1      57.3 2663507 44347
##
## Step:  AIC=44334.21
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight +
##     LinguisticIsolation + DieselPM + DrinkingWater + Unemployment +
##     GroundwaterThreats + Tox.Release + PM2.5 + CleanupSites
##
##                   Df Sum of Sq     RSS   AIC
## + Traffic          1    4045.6 2654874 44325
## + Imp.WaterBodies  1    1971.0 2656948 44331
## + Pesticides       1    1088.6 2657831 44333
## <none>                         2658919 44334
## + Education        1     557.4 2658362 44335
## + HousingBurden    1     422.6 2658497 44335
## + Haz.Waste        1     343.5 2658576 44335
## + SolidWaste       1     222.8 2658696 44336
##
## Step:  AIC=44324.7
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight +
##     LinguisticIsolation + DieselPM + DrinkingWater + Unemployment +
##     GroundwaterThreats + Tox.Release + PM2.5 + CleanupSites +
##     Traffic
##
##                   Df Sum of Sq     RSS   AIC
## + Imp.WaterBodies  1   2131.71 2652742 44321
## + HousingBurden    1   1051.97 2653822 44324
## + Pesticides       1    932.76 2653941 44324
## <none>                         2654874 44325
## + Education        1    553.62 2654320 44325
## + Haz.Waste        1    362.63 2654511 44326
## + SolidWaste       1    224.83 2654649 44326
##
## Step:  AIC=44320.63
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight +
##     LinguisticIsolation + DieselPM + DrinkingWater + Unemployment +
##     GroundwaterThreats + Tox.Release + PM2.5 + CleanupSites +
##     Traffic + Imp.WaterBodies
##
##                  Df Sum of Sq     RSS   AIC
## + HousingBurden  1   1232.29 2651510 44319
## <none>                       2652742 44321
```

```
## + Education      1    674.48 2652067 44321
## + Pesticides     1    650.19 2652092 44321
## + SolidWaste     1    451.74 2652290 44321
## + Haz.Waste      1    250.72 2652491 44322
##
## Step:  AIC=44319.12
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight +
##     LinguisticIsolation + DieselPM + DrinkingWater + Unemployment +
##     GroundwaterThreats + Tox.Release + PM2.5 + CleanupSites +
##     Traffic + Imp.WaterBodies + HousingBurden
##
##               Df Sum of Sq     RSS   AIC
## + Pesticides  1    742.06 2650768 44319
## <none>                    2651510 44319
## + Education   1    532.65 2650977 44320
## + SolidWaste  1    381.52 2651128 44320
## + Haz.Waste   1    244.43 2651265 44320
##
## Step:  AIC=44319
## Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight +
##     LinguisticIsolation + DieselPM + DrinkingWater + Unemployment +
##     GroundwaterThreats + Tox.Release + PM2.5 + CleanupSites +
##     Traffic + Imp.WaterBodies + HousingBurden + Pesticides
##
##               Df Sum of Sq     RSS   AIC
## <none>                    2650768 44319
## + Education   1    659.52 2650108 44319
## + SolidWaste  1    395.82 2650372 44320
## + Haz.Waste   1    245.67 2650522 44320
```

The desired model: Asthma ~ CardiovascularDisease + Poverty + Ozone + LowBirthWeight + LinguisticIsolation + DieselPM + DrinkingWater + Unemployment + GroundwaterThreats + Tox.Release + PM2.5 + CleanupSites + PollutionBurden + Imp.WaterBodies + HousingBurden + Traffic + Pesticides + Haz.Waste

```
#stepwise selection
lm_stepwise <- lm(Asthma ~ Ozone + PM2.5 + DieselPM+ DrinkingWater+Pesticides+Tox.Release+Traffic +Clea
SolidWaste + PollutionBurden + Education+LinguisticIsolation+Poverty+Unemployment+HousingBurden+LowBirt
stepModel <- step(lm_stepwise, direction="both")
```

```
## Start:  AIC=44304.58
## Asthma ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Pesticides +
##     Tox.Release + Traffic + CleanupSites + GroundwaterThreats +
##     Haz.Waste + Imp.WaterBodies + SolidWaste + PollutionBurden +
##     Education + LinguisticIsolation + Poverty + Unemployment +
##     HousingBurden + LowBirthWeight + CardiovascularDisease
##
##                         Df Sum of Sq     RSS   AIC
## - SolidWaste             1         0 2642913 44303
## - Education              1       333 2643246 44304
## <none>                               2642913 44305
## - Haz.Waste              1       960 2643873 44305
## - Pesticides             1      1179 2644092 44306
## - Traffic                1      1320 2644233 44306
```

```
## - HousingBurden          1      1733 2644646 44308
## - Imp.WaterBodies        1      5162 2648075 44317
## - PollutionBurden        1      6426 2649339 44321
## - CleanupSites           1      6569 2649482 44321
## - Tox.Release            1      7636 2650549 44324
## - GroundwaterThreats     1     12219 2655132 44337
## - PM2.5                  1     14322 2657234 44343
## - Unemployment           1     26353 2669266 44378
## - DrinkingWater          1     30037 2672949 44388
## - LinguisticIsolation    1     37772 2680684 44410
## - DieselPM               1     61111 2704024 44475
## - Poverty                1     80129 2723041 44528
## - LowBirthWeight         1     92240 2735153 44562
## - Ozone                  1    110597 2753510 44612
## - CardiovascularDisease  1   1594762 4237674 47870
##
## Step:  AIC=44302.58
## Asthma ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Pesticides +
##     Tox.Release + Traffic + CleanupSites + GroundwaterThreats +
##     Haz.Waste + Imp.WaterBodies + PollutionBurden + Education +
##     LinguisticIsolation + Poverty + Unemployment + HousingBurden +
##     LowBirthWeight + CardiovascularDisease
##
##                         Df Sum of Sq     RSS   AIC
## - Education              1       333 2643246 44302
## <none>                               2642913 44303
## - Haz.Waste              1       996 2643909 44303
## - Pesticides             1      1179 2644092 44304
## - Traffic                1      1334 2644247 44304
## + SolidWaste             1         0 2642913 44305
## - HousingBurden          1      1741 2644654 44306
## - Imp.WaterBodies        1      5162 2648075 44315
## - CleanupSites           1      6701 2649614 44320
## - PollutionBurden        1      6922 2649835 44320
## - Tox.Release            1      7641 2650554 44322
## - GroundwaterThreats     1     12219 2655132 44335
## - PM2.5                  1     14892 2657805 44343
## - Unemployment           1     26358 2669270 44376
## - DrinkingWater          1     30043 2672956 44386
## - LinguisticIsolation    1     38020 2680932 44409
## - DieselPM               1     62624 2705537 44478
## - Poverty                1     80428 2723341 44527
## - LowBirthWeight         1     92245 2735158 44560
## - Ozone                  1    110699 2753611 44611
## - CardiovascularDisease  1   1597584 4240496 47874
##
## Step:  AIC=44301.53
## Asthma ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Pesticides +
##     Tox.Release + Traffic + CleanupSites + GroundwaterThreats +
##     Haz.Waste + Imp.WaterBodies + PollutionBurden + LinguisticIsolation +
##     Poverty + Unemployment + HousingBurden + LowBirthWeight +
##     CardiovascularDisease
##
##                         Df Sum of Sq     RSS   AIC
```

```
## <none>                              2643246 44302
## - Haz.Waste              1       980 2644226 44302
## + Education              1       333 2642913 44303
## - Pesticides             1      1082 2644328 44303
## - Traffic                1      1314 2644561 44303
## + SolidWaste             1         0 2643246 44304
## - HousingBurden          1      1894 2645140 44305
## - Imp.WaterBodies        1      5168 2648414 44314
## - CleanupSites           1      6677 2649923 44319
## - PollutionBurden        1      7276 2650522 44320
## - Tox.Release            1      7874 2651120 44322
## - GroundwaterThreats     1     12751 2655997 44336
## - PM2.5                  1     14661 2657907 44341
## - Unemployment           1     26322 2669568 44374
## - DrinkingWater          1     30069 2673315 44385
## - LinguisticIsolation    1     55381 2698627 44456
## - DieselPM               1     64585 2707831 44482
## - LowBirthWeight         1     92035 2735282 44558
## - Poverty                1    104727 2747973 44593
## - Ozone                  1    110505 2753752 44609
## - CardiovascularDisease  1   1641101 4284347 47949
```

The stepwise model selection method and forward selection method outputs match each other. Combined with our subject matter knowledge, we will proceed with modeling using the following covariates: Ozone + PM2.5 + DieselPM + DrinkingWater + Pesticides + Tox.Release + Traffic + CleanupSites + GroundwaterThreats + Haz.Waste + Imp.WaterBodies + PollutionBurden + LinguisticIsolation + Poverty + Unemployment + HousingBurden + LowBirthWeight + CardiovascularDisease

```
#check for missing data
anyNA(data_cal)
```

Missing Data

```
## [1] TRUE
```

```
colnames(data_cal)[colSums(is.na(data_cal)) > 0]
```

```
##  [1] "CES3.0Score"             "CES3.0Percentile"
##  [3] "CES3.0PercentileRange"   "PM2.5"
##  [5] "PM2.5Pctl"               "DrinkingWater"
##  [7] "DrinkingWaterPctl"       "Traffic"
##  [9] "TrafficPctl"             "LowBirthWeight"
## [11] "LowBirthWeightPctl"      "Education"
## [13] "EducationPctl"           "LinguisticIsolation"
## [15] "LinguisticIsolationPctl" "Poverty"
## [17] "PovertyPctl"             "Unemployment"
## [19] "UnemploymentPctl"        "HousingBurden"
## [21] "HousingBurdenPctl"       "Pop.Char."
## [23] "Pop.Char.Score"          "Pop.Char.Pctl"
```

```
sum(is.na(data_cal$PM2.5))
```

```
## [1] 19
```

covariates with missing data: CES 3.0 Score, PM2.5, DrinkingWater, Traffic, LowBirthWeight, Education, LinguisticIsolation, Poverty, Unemployment, HousingBurden, Population characteristics.

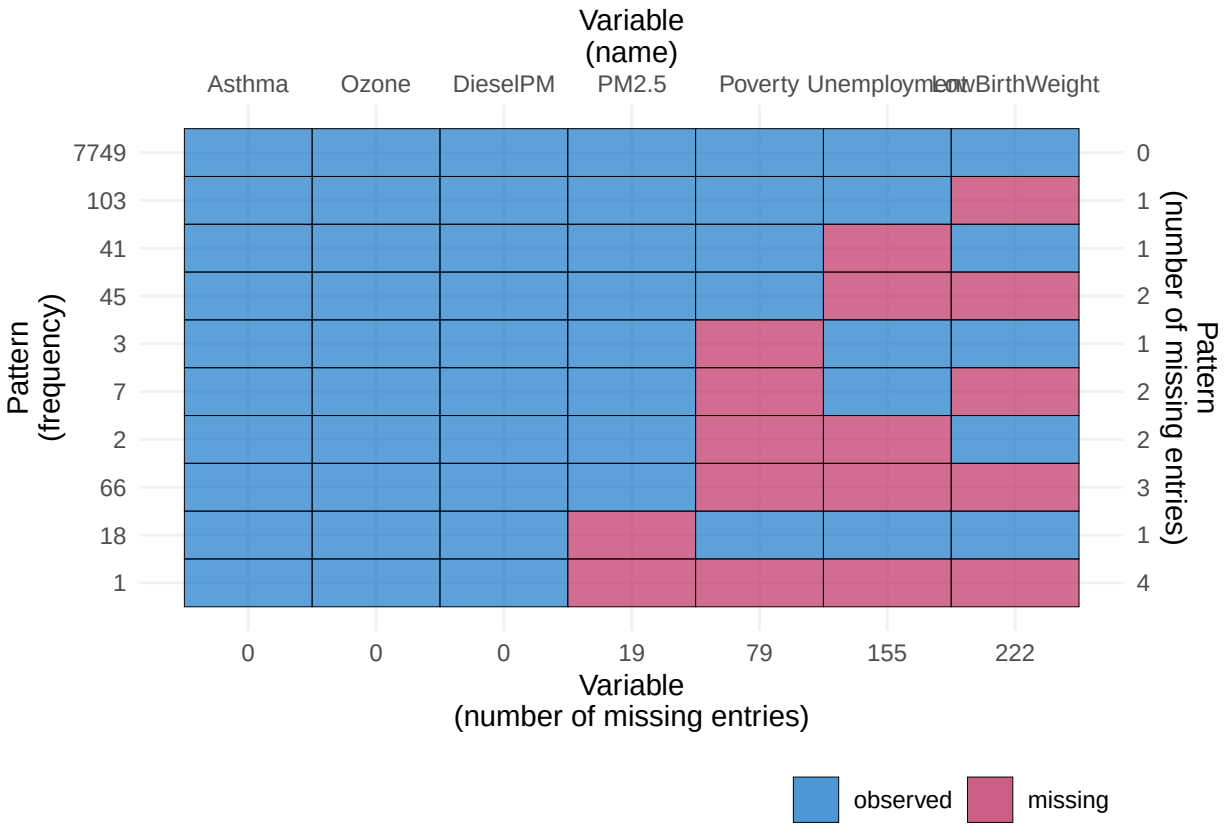
```
#look at rows with missing data
dat_NA <- data_cal[!complete.cases(data_cal), ]
rowSums(is.na(dat_NA))
```

```
##   [1]  2  4  2  2  2  6  2  2  2  2  6  2  2  2  2  6  4  4  2  2  6  2  2  2  2
##  [26]  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  4  2  6  4  2  2  2  2  2
##  [51]  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  4  2  6
##  [76]  2  2  2  2  2  4  2  2  2  2  4  2  4  2  2  2  2  4  2  2  4  2  2  2  2
## [101]  2  6  4  2  4  2  2  2  2  2  2  4  6  2  2  2  2  2  2  2  2  2  2  2  2  2
## [126]  2  2  4  2  2  2  2  2  2  6  2  2  2  2  2  2  4  2  6  2  2  2  2  2  4
## [151]  2  2  4  2  2  2  2  4  2  2  2  2  4  2  4  2  2  4  2  2  2  2  2  2  2
## [176]  2  2  6  2  2  6  2  2  2  4  2  6  2  2  2  2  2  2  2  2  4  6  2  2  4
## [201]  2  2  2  2  2  2  2  2  2  2  2  2  4  2  6  2  4  2  6  2  2  2  4  2  2
## [226]  2  2  2  2  2  2  2  2  2  2  2  2  2  4  4  2  2  2  2  2  2  2  2  2  2
## [251]  2  4  2  2  2  2  4  2  2  2  2  2  2  2  2  2  2  2  2  4  2  2  2  2  2
## [276]  4  6  4  2  2  2  2  2  2  4  2  4  2  2  2  6  2  2  2  2  2  6  2  2  2
## [301]  2  2  2  2  4  2  4  2  2  2  2  2  2  6  2  2  4  4  2  4  4  6  2  2  2
## [326]  2  2  2  2  2  2  4  2  2  2  2  2  4  6  2  4  4  2  2  2  2  2  2  4  2
## [351]  4  2  2  2  2  4  2  2  4  4  2  2  2  2  2  4  2  6  6  2  2  2 12 20 16
## [376] 14 18 16 18 20 18 14 16 20 14 20 18 20 20 14 16 18 20 16 14 20 18 18 14 18
## [401] 14 14 18 20 20 16 18 16 18 14 18 16 18 16 18 18 18 16 18 12 20 18 18 16 18
## [426] 16 16 16 18 16 16 18 14 16 14 16 18 18 16 14 16 18 18 16 14 16 16 14 16 16
## [451] 18 14 18 18 14 18 18 18 16 14 20 14 18 16 16 16 16 18 16 18 16 14 14 16 18
## [476] 14 22 22
```

```
#look at missing pattern
library(ggmice)
```

```
## Warning: package 'ggmice' was built under R version 4.2.2
```

```
dat <- data_cal[,c("Asthma", "Ozone", "PM2.5", "DieselPM","Poverty", "Unemployment",  "LowBirthWeight")]
plot_pattern(dat)
```




```
#Attempt Multivariate Imputation
library(mice)
```

```
## Warning: package 'mice' was built under R version 4.2.2
```

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

```
## The following objects are masked from 'package:ggmice':
##
##     bwplot, densityplot, stripplot, xyplot
```

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

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

```
tempData = mice(dat, m = 5, maxit = 10, seed = 210)
```

```
##
```

```
##  iter imp variable
##   1   1  PM2.5  Poverty  Unemployment  LowBirthWeight
##   1   2  PM2.5  Poverty  Unemployment  LowBirthWeight
##   1   3  PM2.5  Poverty  Unemployment  LowBirthWeight
##   1   4  PM2.5  Poverty  Unemployment  LowBirthWeight
##   1   5  PM2.5  Poverty  Unemployment  LowBirthWeight
##   2   1  PM2.5  Poverty  Unemployment  LowBirthWeight
##   2   2  PM2.5  Poverty  Unemployment  LowBirthWeight
##   2   3  PM2.5  Poverty  Unemployment  LowBirthWeight
##   2   4  PM2.5  Poverty  Unemployment  LowBirthWeight
##   2   5  PM2.5  Poverty  Unemployment  LowBirthWeight
##   3   1  PM2.5  Poverty  Unemployment  LowBirthWeight
##   3   2  PM2.5  Poverty  Unemployment  LowBirthWeight
##   3   3  PM2.5  Poverty  Unemployment  LowBirthWeight
##   3   4  PM2.5  Poverty  Unemployment  LowBirthWeight
##   3   5  PM2.5  Poverty  Unemployment  LowBirthWeight
##   4   1  PM2.5  Poverty  Unemployment  LowBirthWeight
##   4   2  PM2.5  Poverty  Unemployment  LowBirthWeight
##   4   3  PM2.5  Poverty  Unemployment  LowBirthWeight
##   4   4  PM2.5  Poverty  Unemployment  LowBirthWeight
##   4   5  PM2.5  Poverty  Unemployment  LowBirthWeight
##   5   1  PM2.5  Poverty  Unemployment  LowBirthWeight
##   5   2  PM2.5  Poverty  Unemployment  LowBirthWeight
##   5   3  PM2.5  Poverty  Unemployment  LowBirthWeight
##   5   4  PM2.5  Poverty  Unemployment  LowBirthWeight
##   5   5  PM2.5  Poverty  Unemployment  LowBirthWeight
##   6   1  PM2.5  Poverty  Unemployment  LowBirthWeight
##   6   2  PM2.5  Poverty  Unemployment  LowBirthWeight
##   6   3  PM2.5  Poverty  Unemployment  LowBirthWeight
##   6   4  PM2.5  Poverty  Unemployment  LowBirthWeight
##   6   5  PM2.5  Poverty  Unemployment  LowBirthWeight
##   7   1  PM2.5  Poverty  Unemployment  LowBirthWeight
##   7   2  PM2.5  Poverty  Unemployment  LowBirthWeight
##   7   3  PM2.5  Poverty  Unemployment  LowBirthWeight
##   7   4  PM2.5  Poverty  Unemployment  LowBirthWeight
##   7   5  PM2.5  Poverty  Unemployment  LowBirthWeight
##   8   1  PM2.5  Poverty  Unemployment  LowBirthWeight
##   8   2  PM2.5  Poverty  Unemployment  LowBirthWeight
##   8   3  PM2.5  Poverty  Unemployment  LowBirthWeight
##   8   4  PM2.5  Poverty  Unemployment  LowBirthWeight
##   8   5  PM2.5  Poverty  Unemployment  LowBirthWeight
##   9   1  PM2.5  Poverty  Unemployment  LowBirthWeight
##   9   2  PM2.5  Poverty  Unemployment  LowBirthWeight
##   9   3  PM2.5  Poverty  Unemployment  LowBirthWeight
##   9   4  PM2.5  Poverty  Unemployment  LowBirthWeight
##   9   5  PM2.5  Poverty  Unemployment  LowBirthWeight
##  10   1  PM2.5  Poverty  Unemployment  LowBirthWeight
##  10   2  PM2.5  Poverty  Unemployment  LowBirthWeight
##  10   3  PM2.5  Poverty  Unemployment  LowBirthWeight
##  10   4  PM2.5  Poverty  Unemployment  LowBirthWeight
##  10   5  PM2.5  Poverty  Unemployment  LowBirthWeight
```

```
summary(tempData)
```

```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##         Asthma          Ozone          PM2.5       DieselPM        Poverty
##             ""             ""          "pmm"             ""          "pmm"
##   Unemployment LowBirthWeight
##          "pmm"          "pmm"
## PredictorMatrix:
##               Asthma Ozone PM2.5 DieselPM Poverty Unemployment LowBirthWeight
## Asthma             0     1     1        1       1            1              1
## Ozone              1     0     1        1       1            1              1
## PM2.5              1     1     0        1       1            1              1
## DieselPM           1     1     1        0       1            1              1
## Poverty            1     1     1        1       0            1              1
## Unemployment       1     1     1        1       1            0              1
```

```
#complete(tempData,action=1)
```

```
lm_pureLinear = lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Pesticides + Tox.Release + T
summary(lm_pureLinear)
```

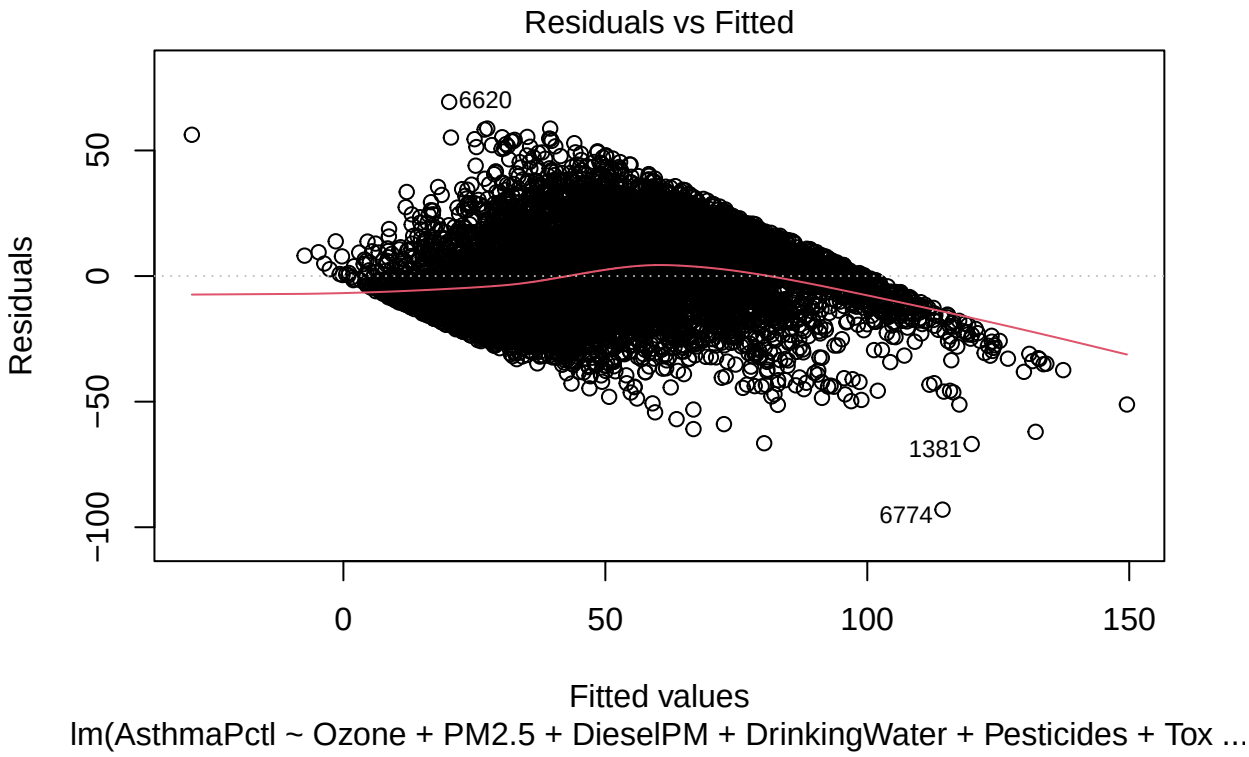
```
##
## Call:
## lm(formula = AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater +
##     Pesticides + Tox.Release + Traffic + CleanupSites + GroundwaterThreats +
##     Haz.Waste + Imp.WaterBodies + PollutionBurden + LinguisticIsolation +
##     Poverty + Unemployment + HousingBurden + LowBirthWeight +
##     CardiovascularDisease, data = data_cal1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.986 -12.050  -0.974  11.268  69.307
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -6.872e+00  1.353e+00  -5.078 3.90e-07 ***
## Ozone                 -5.165e+02  2.859e+01 -18.064  < 2e-16 ***
## PM2.5                  7.221e-01  1.082e-01   6.674 2.67e-11 ***
## DieselPM               1.812e-01  1.417e-02  12.792  < 2e-16 ***
## DrinkingWater         -7.974e-03  1.100e-03  -7.247 4.68e-13 ***
## Pesticides             1.983e-04  7.444e-05   2.664  0.00773 **
## Tox.Release           -8.799e-05  1.580e-05  -5.568 2.67e-08 ***
## Traffic               -6.894e-05  2.398e-04  -0.287  0.77378
## CleanupSites           9.481e-03  1.544e-02   0.614  0.53920
## GroundwaterThreats     2.756e-02  6.349e-03   4.341 1.44e-05 ***
## Haz.Waste              7.928e-02  1.631e-01   0.486  0.62686
## Imp.WaterBodies       -2.371e-02  5.037e-02  -0.471  0.63790
## PollutionBurden       -2.082e-02  2.982e-02  -0.698  0.48510
## LinguisticIsolation   -2.902e-01  2.776e-02 -10.455  < 2e-16 ***
## Poverty                4.424e-01  1.938e-02  22.824  < 2e-16 ***
## Unemployment           3.143e-01  5.398e-02   5.823 6.03e-09 ***
## HousingBurden          3.054e-02  3.436e-02   0.889  0.37419
## LowBirthWeight         1.784e+00  1.378e-01  12.939  < 2e-16 ***
```

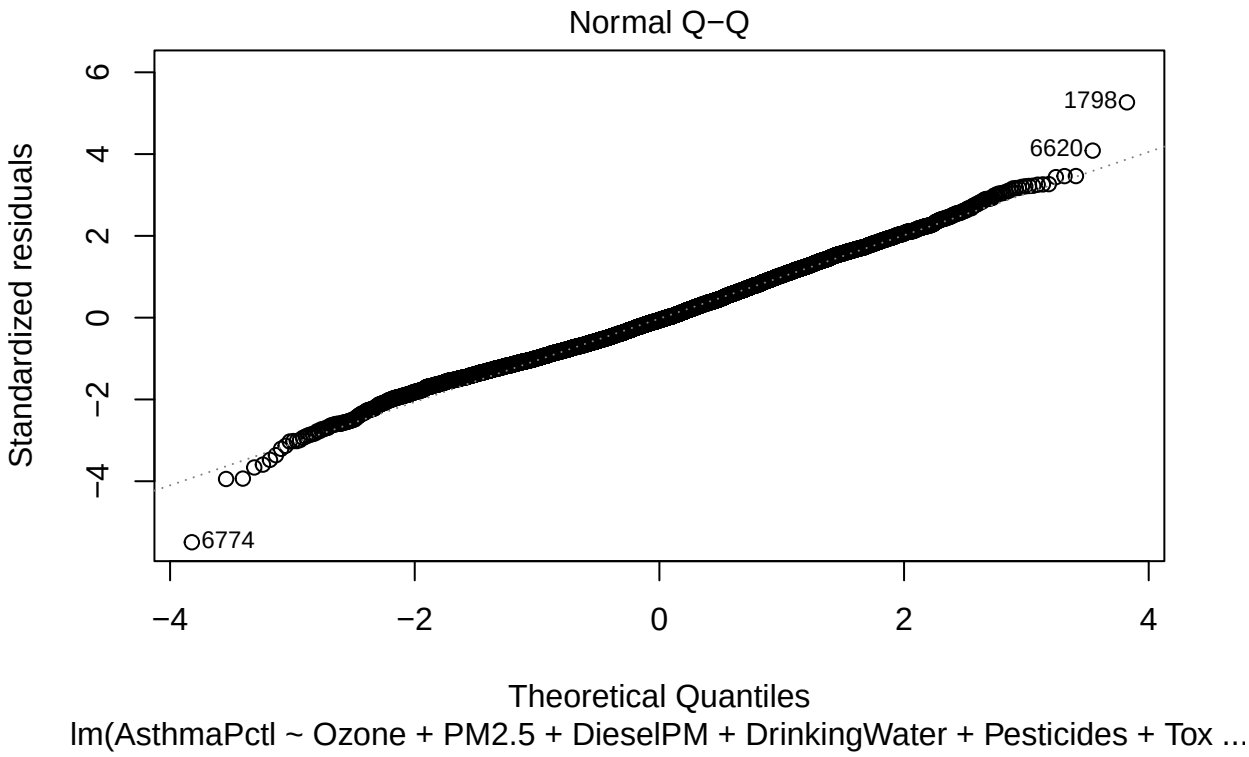
```
## CardiovascularDisease  5.968e+00  8.250e-02  72.344  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.98 on 7538 degrees of freedom
## Multiple R-squared:  0.6499, Adjusted R-squared:  0.649
## F-statistic: 777.3 on 18 and 7538 DF,  p-value: < 2.2e-16
```

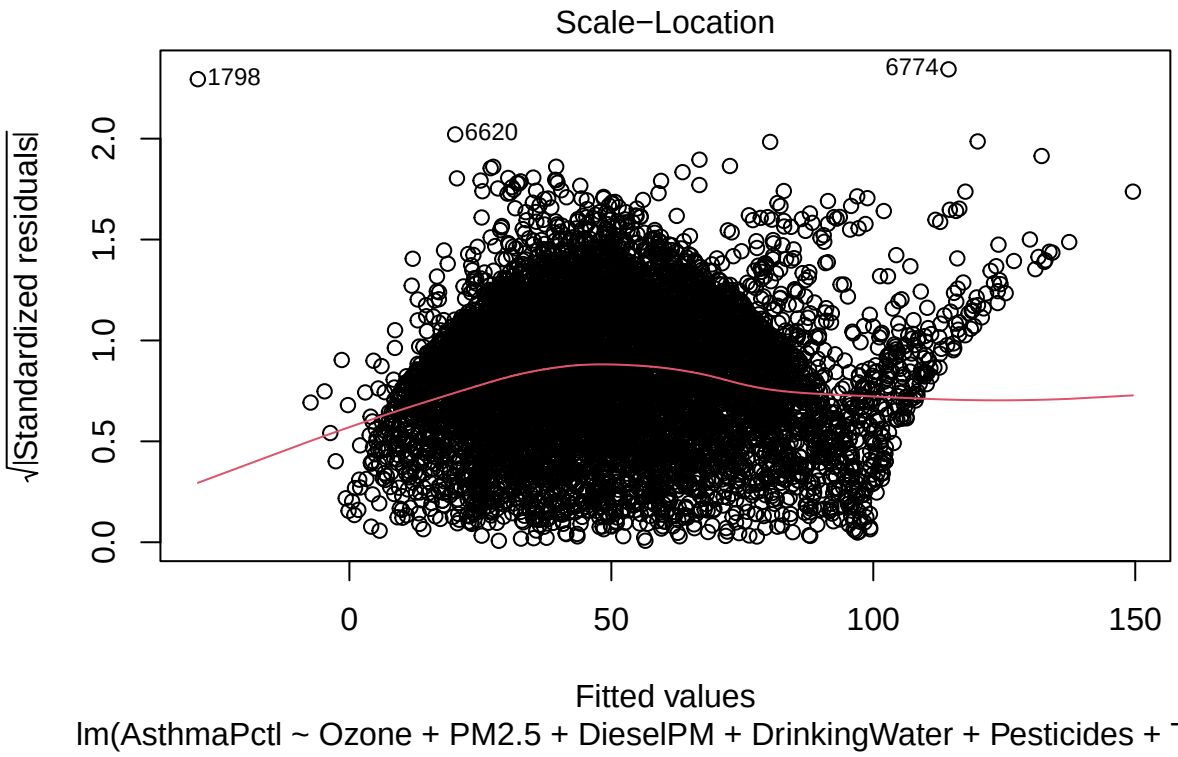
```
confint(lm_pureLinear)
```

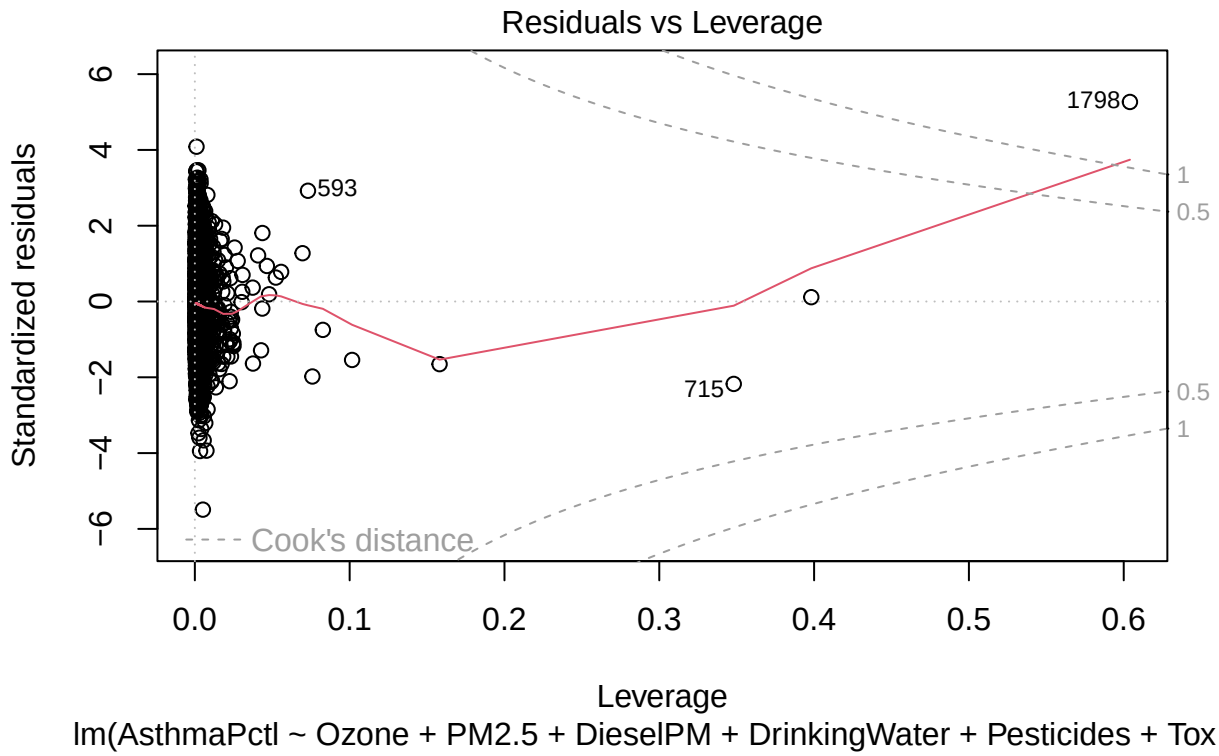
```
##                              2.5 %        97.5 %
## (Intercept)          -9.525282e+00 -4.219647e+00
## Ozone                -5.725459e+02 -4.604463e+02
## PM2.5                 5.099830e-01  9.341929e-01
## DieselPM              1.534544e-01  2.089956e-01
## DrinkingWater        -1.013082e-02 -5.817111e-03
## Pesticides            5.240025e-05  3.442423e-04
## Tox.Release          -1.189753e-04 -5.701363e-05
## Traffic              -5.391043e-04  4.012211e-04
## CleanupSites         -2.078607e-02  3.974830e-02
## GroundwaterThreats    1.511663e-02  4.000940e-02
## Haz.Waste            -2.403698e-01  3.989204e-01
## Imp.WaterBodies      -1.224416e-01  7.503004e-02
## PollutionBurden      -7.928410e-02  3.764107e-02
## LinguisticIsolation  -3.446608e-01 -2.358249e-01
## Poverty               4.044166e-01  4.804108e-01
## Unemployment          2.084894e-01  4.201184e-01
## HousingBurden        -3.682229e-02  9.789950e-02
## LowBirthWeight        1.513459e+00  2.053901e+00
## CardiovascularDisease  5.806570e+00  6.130009e+00
```

```
#Evaluation for this simple linear model
plot(lm_pureLinear)
```


Residuals vs Fitted
6620
1381
6774
Residuals
50
0
−50
−100
0
50
100
150
Fitted values
lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Pesticides + Tox ...


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Pesticides + Tox ...
1798
6620
6774
−4
−2
0
2
4
6


Scale−Location
1798
6620
6774
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
50
100
150
Fitted values
lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Pesticides + Tox ...



This simple linear model is unsatisfactory. We can see from the plot of Residuals vs. Fitted that there's fanning trend, no equal variance above and below the line. The QQ-plot also shows non-normality with clear deviation from the diagonal line on both ends of the fitted curve.

- pesticides, traffic, hazard waste are not statistically significant. We may consider removing these covariates.
- Housing burden as a lower significance level (higher p-value) compared to other covariates may consider removing it from linear model as well.

```
#remove coefficients with lower significance level, include only if coefficients are coded "***" signif
lm_rmSig = lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Tox.Release + CleanupSites + Grou
summary(lm_rmSig)
```

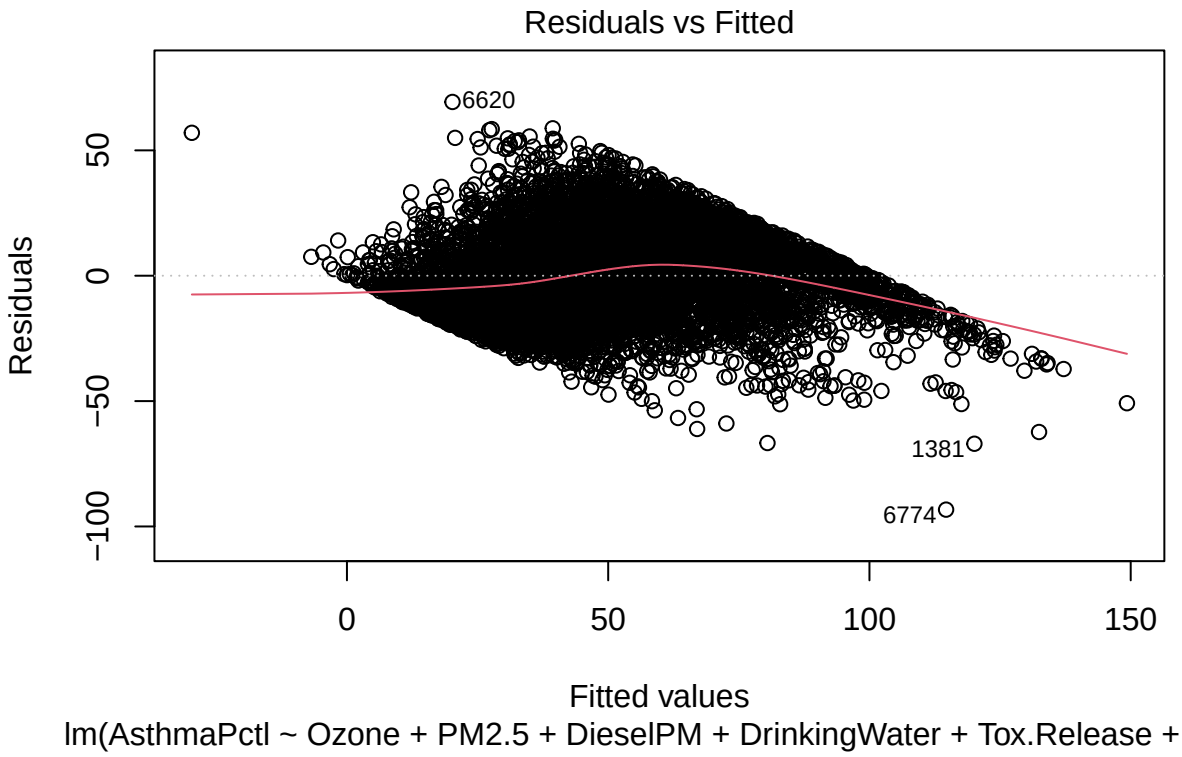
```
##
## Call:
## lm(formula = AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater +
##     Tox.Release + CleanupSites + GroundwaterThreats + Imp.WaterBodies +
##     PollutionBurden + LinguisticIsolation + Poverty + Unemployment +
##     LowBirthWeight + CardiovascularDisease, data = data_cal1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -93.29 -12.05  -1.00  11.26  69.30
##
## Coefficients:
```

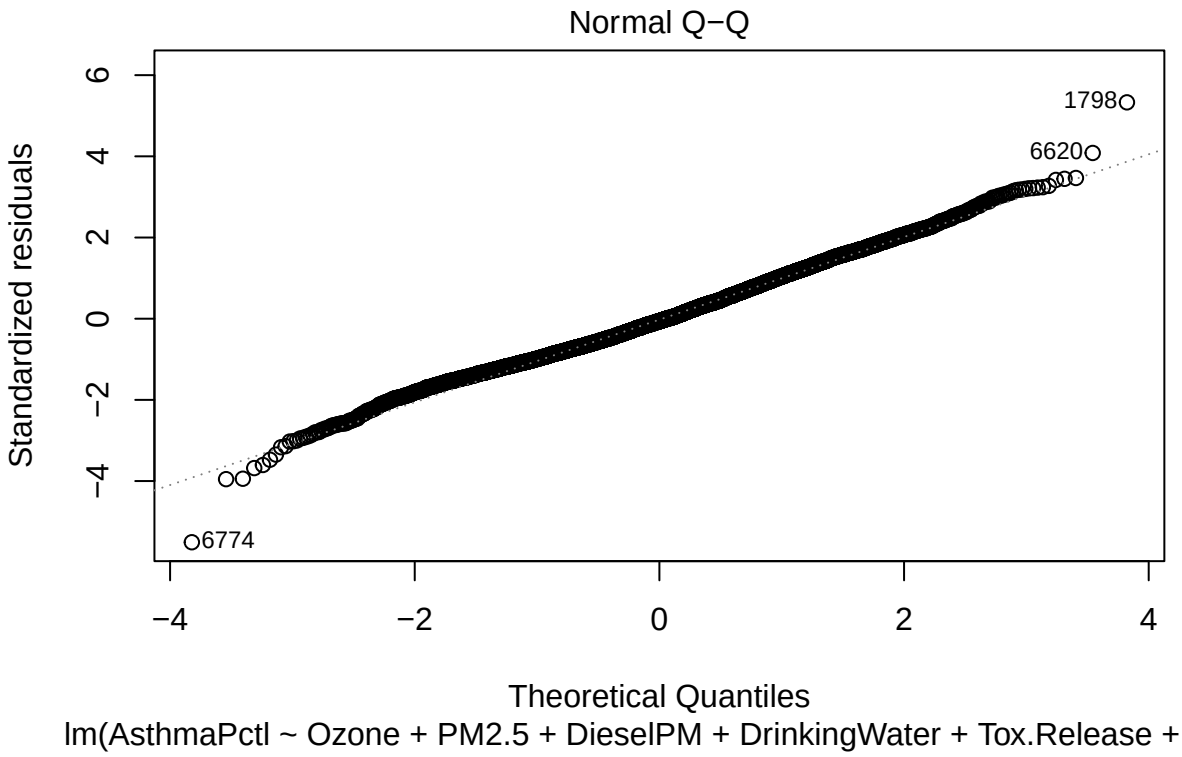
```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -6.377e+00  1.291e+00  -4.941 7.93e-07 ***
## Ozone                 -5.231e+02  2.834e+01 -18.455  < 2e-16 ***
## PM2.5                  6.966e-01  1.071e-01   6.506 8.21e-11 ***
## DieselPM               1.794e-01  1.411e-02  12.721  < 2e-16 ***
## DrinkingWater         -7.770e-03  1.082e-03  -7.178 7.73e-13 ***
## Tox.Release           -8.890e-05  1.580e-05  -5.626 1.91e-08 ***
## CleanupSites           9.993e-03  1.477e-02   0.677    0.499
## GroundwaterThreats     2.742e-02  6.299e-03   4.353 1.36e-05 ***
## Imp.WaterBodies       -1.465e-02  4.977e-02  -0.294    0.768
## PollutionBurden       -1.486e-02  2.732e-02  -0.544    0.586
## LinguisticIsolation   -2.886e-01  2.773e-02 -10.409  < 2e-16 ***
## Poverty                4.530e-01  1.640e-02  27.621  < 2e-16 ***
## Unemployment           3.116e-01  5.392e-02   5.780 7.79e-09 ***
## LowBirthWeight         1.787e+00  1.374e-01  13.003  < 2e-16 ***
## CardiovascularDisease  5.965e+00  8.196e-02  72.773  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.98 on 7542 degrees of freedom
## Multiple R-squared:  0.6495, Adjusted R-squared:  0.6488
## F-statistic: 998.2 on 14 and 7542 DF,  p-value: < 2.2e-16
```

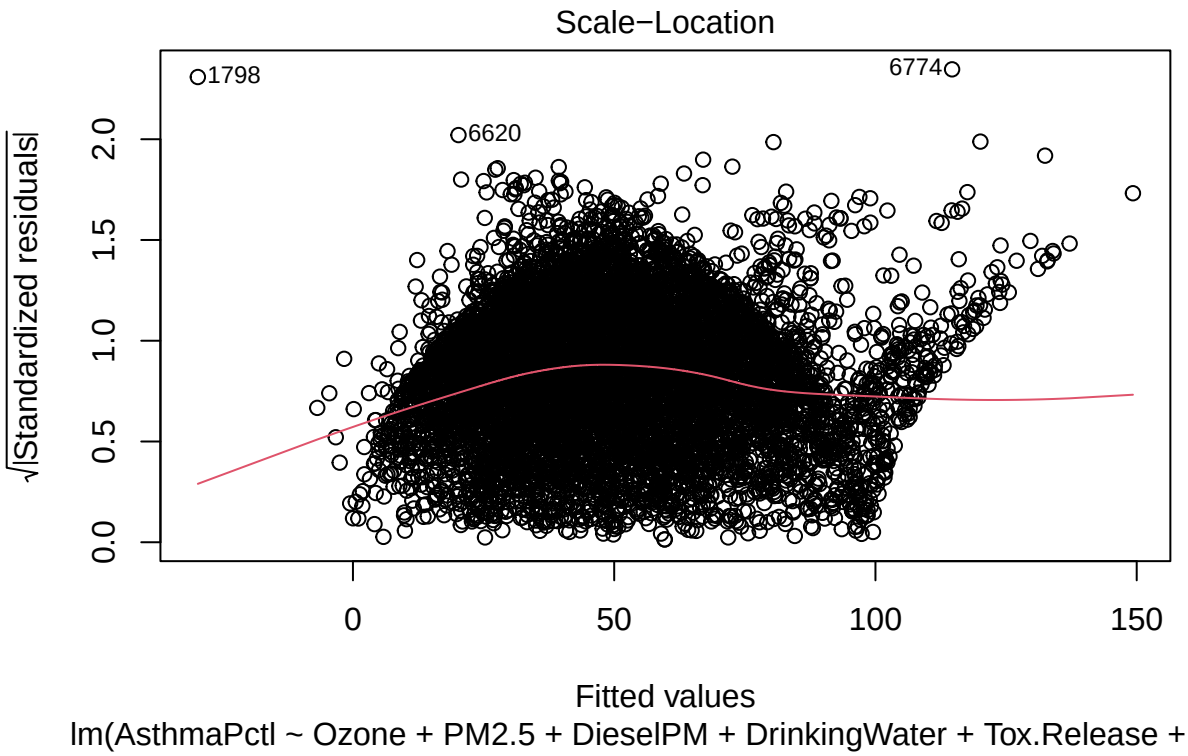
```
confint(lm_rmSig)
```

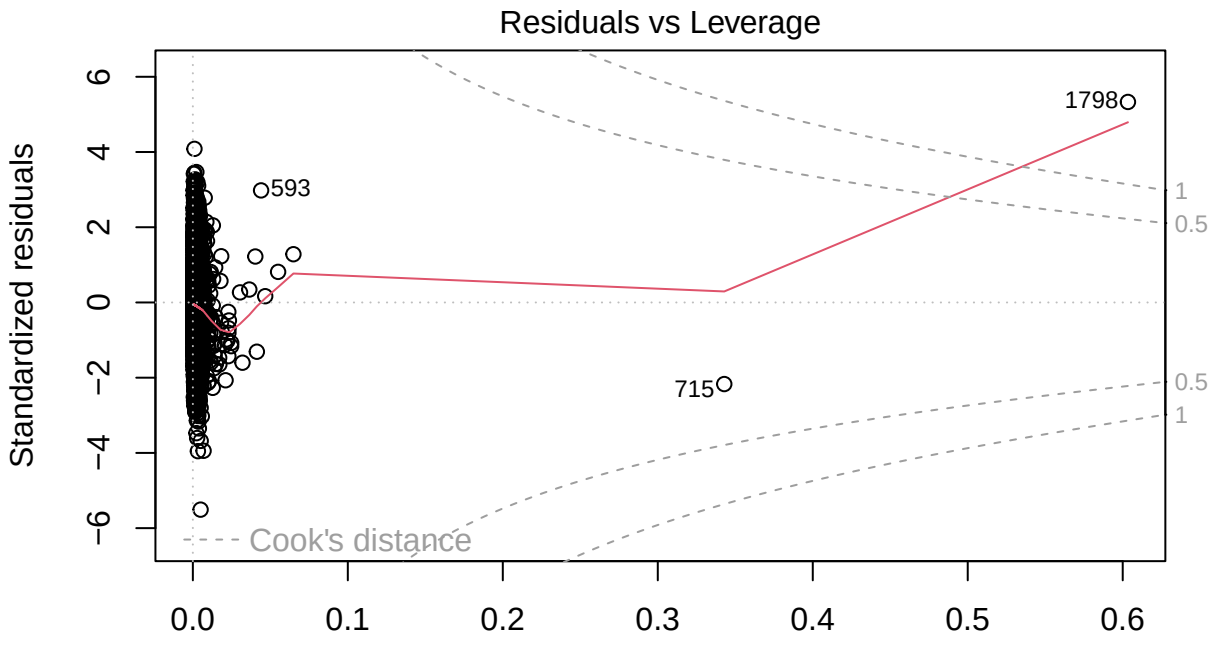
```
##                             2.5 %        97.5 %
## (Intercept)           -8.907489e+00 -3.847369e+00
## Ozone                 -5.786509e+02 -4.675247e+02
## PM2.5                  4.867100e-01  9.064792e-01
## DieselPM               1.517946e-01  2.070983e-01
## DrinkingWater         -9.891879e-03 -5.648174e-03
## Tox.Release           -1.198756e-04 -5.792624e-05
## CleanupSites          -1.895892e-02  3.894586e-02
## GroundwaterThreats     1.507043e-02  3.976518e-02
## Imp.WaterBodies       -1.122051e-01  8.290628e-02
## PollutionBurden       -6.841046e-02  3.869044e-02
## LinguisticIsolation   -3.429982e-01 -2.342776e-01
## Poverty                4.208759e-01  4.851802e-01
## Unemployment           2.059361e-01  4.173301e-01
## LowBirthWeight         1.517740e+00  2.056570e+00
## CardiovascularDisease  5.803848e+00  6.125179e+00
```

```
#Evaluation for this simple linear model
plot(lm_rmSig)
```


Residuals vs Fitted
6620
1381
6774
Residuals
50
0
−50
−100
0
50
100
150
Fitted values
lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Tox.Release + Cl ..


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Tox.Release + Cl ..
1798
6620
6774
6
4
2
0
−2
−4
−4
−2
0
2
4


Scale−Location
1798
6620
6774
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
50
100
150
Fitted values
lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Tox.Release + Cl ..




```
#leave coefficients with significant level <2e-16
lm_highSig = lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater + LinguisticIsolation + Poverty +
summary(lm_highSig)
```

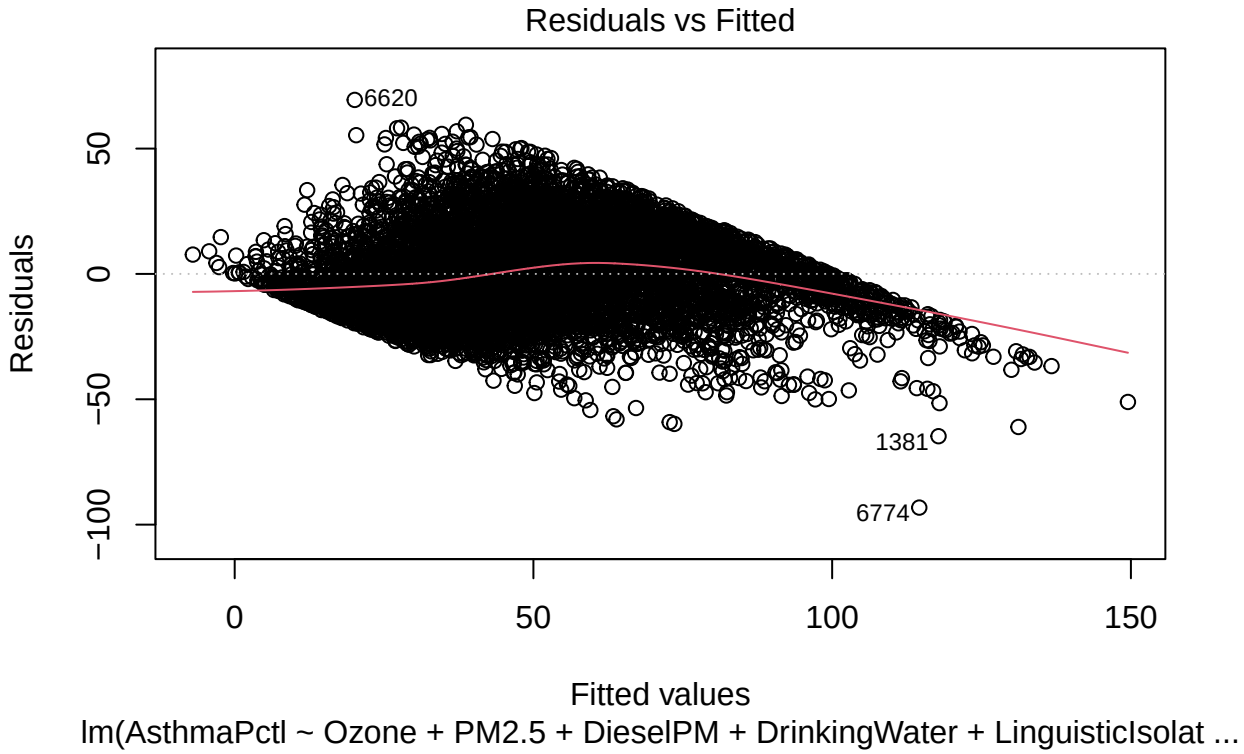
```
##
## Call:
## lm(formula = AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater +
##     LinguisticIsolation + Poverty + Unemployment + LowBirthWeight +
##     CardiovascularDisease, data = data_cal1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.203 -12.186  -1.056  11.235  69.408
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -5.359e+00  1.207e+00  -4.439 9.18e-06 ***
## Ozone                 -5.458e+02  2.768e+01 -19.716  < 2e-16 ***
## PM2.5                  6.282e-01  9.335e-02   6.729 1.83e-11 ***
## DieselPM               1.788e-01  1.357e-02  13.179  < 2e-16 ***
## DrinkingWater         -7.854e-03  9.719e-04  -8.081 7.44e-16 ***
## LinguisticIsolation   -3.049e-01  2.752e-02 -11.079  < 2e-16 ***
## Poverty                4.584e-01  1.638e-02  27.988  < 2e-16 ***
## Unemployment           3.225e-01  5.388e-02   5.985 2.27e-09 ***
## LowBirthWeight         1.801e+00  1.378e-01  13.068  < 2e-16 ***
## CardiovascularDisease  5.984e+00  8.199e-02  72.983  < 2e-16 ***
```

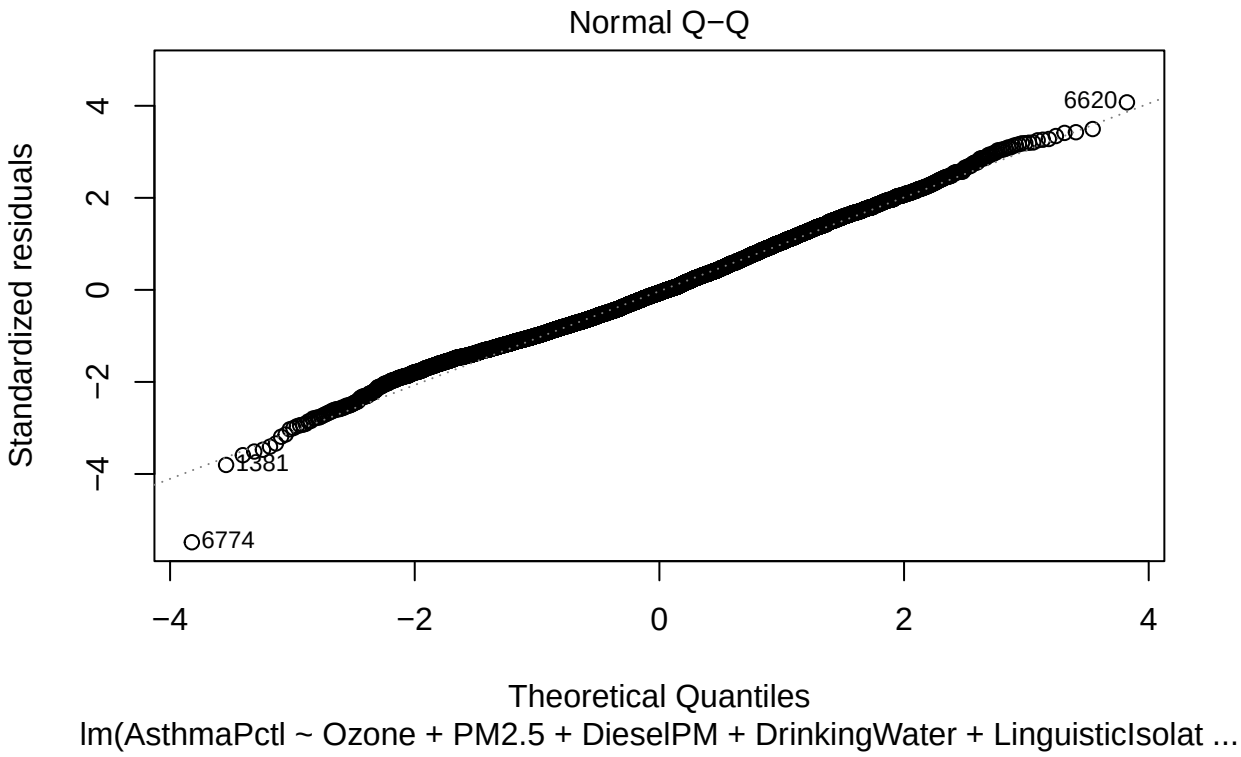
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.04 on 7547 degrees of freedom
## Multiple R-squared:  0.6468, Adjusted R-squared:  0.6464
## F-statistic:  1536 on 9 and 7547 DF,  p-value: < 2.2e-16
```

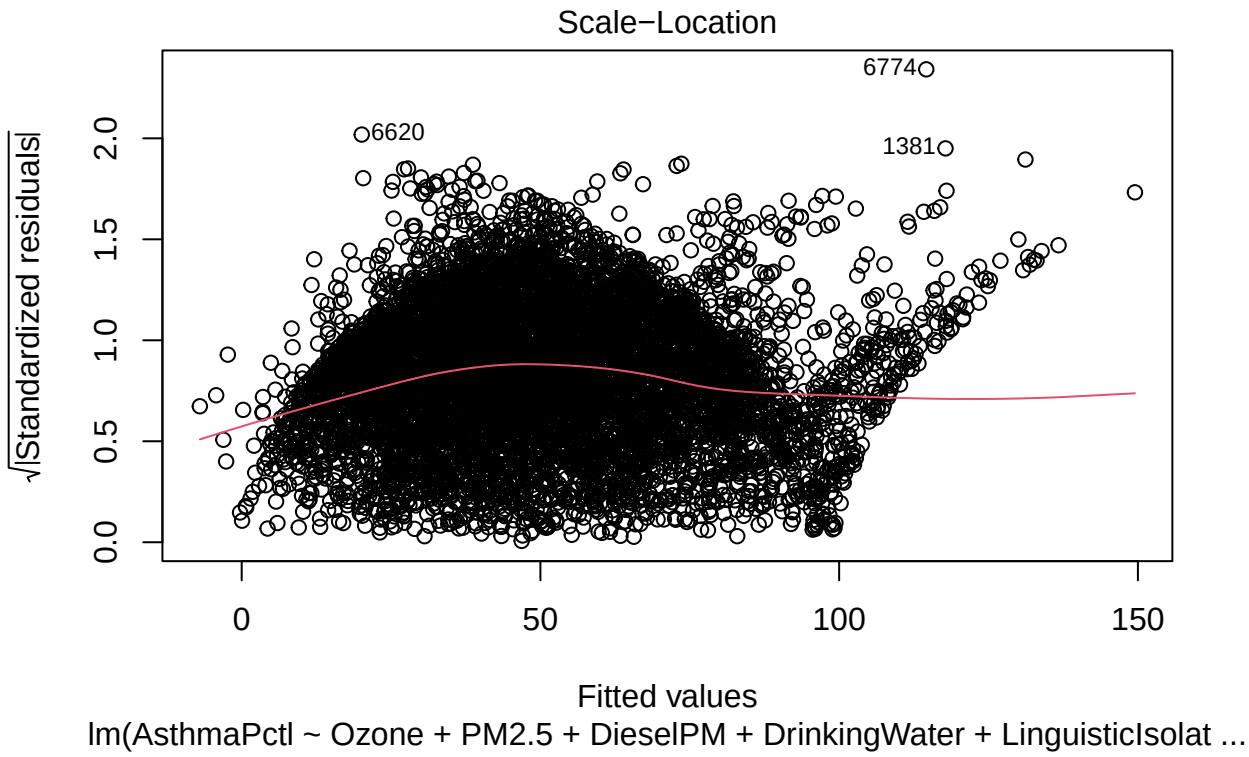
```
confint(lm_highSig)
```

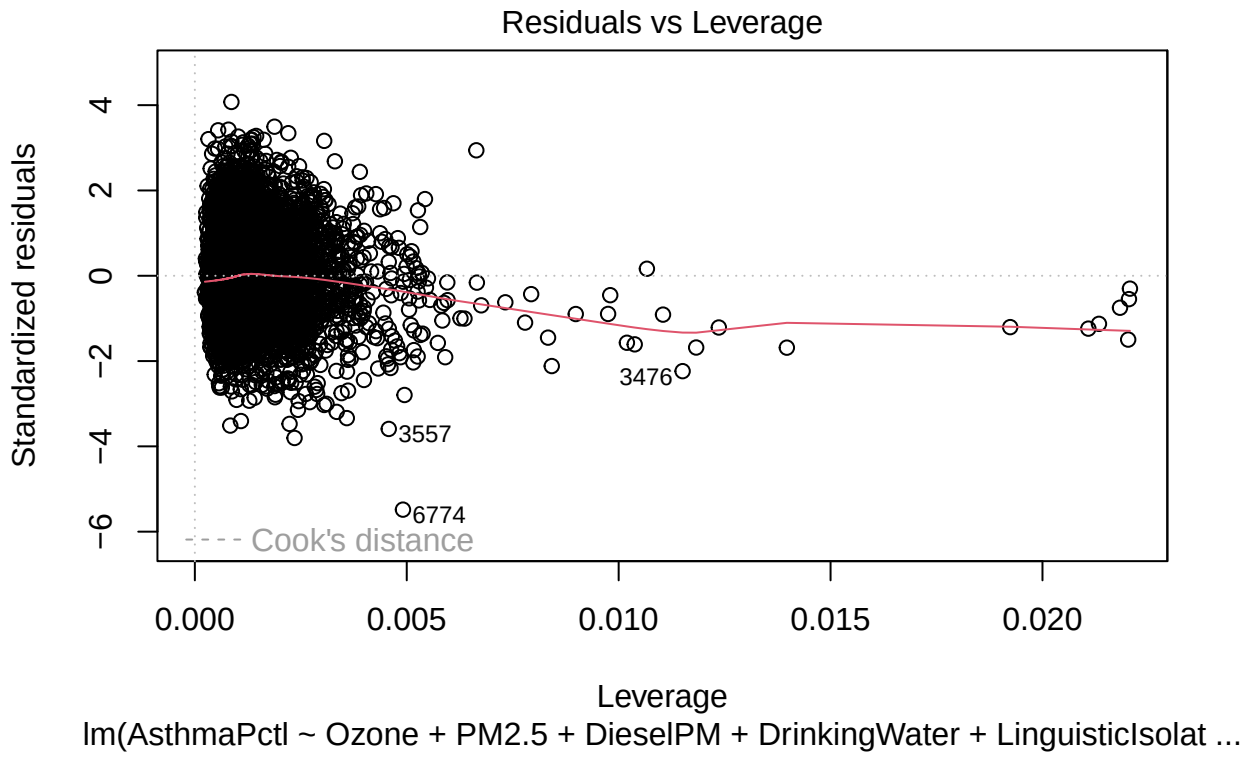
```
##                              2.5 %        97.5 %
## (Intercept)           -7.725208e+00 -2.992127e+00
## Ozone                 -6.000380e+02 -4.915096e+02
## PM2.5                  4.451869e-01  8.111557e-01
## DieselPM               1.522358e-01  2.054368e-01
## DrinkingWater         -9.759135e-03 -5.948642e-03
## LinguisticIsolation   -3.588963e-01 -2.509872e-01
## Poverty                4.262653e-01  4.904740e-01
## Unemployment           2.168346e-01  4.280698e-01
## LowBirthWeight         1.530568e+00  2.070772e+00
## CardiovascularDisease  5.823507e+00  6.144971e+00
```

```
#Evaluation for this simple linear model
plot(lm_highSig)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater + LinguisticIsolat ...
6620
1381
6774
−4
−2
0
2
4


Scale−Location
6774
6620
1381
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
50
100
150
Fitted values
lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + DrinkingWater + LinguisticIsolat ...




```
#vector of covariates of interest
vec_cov <- c("Ozone", "PM2.5", "DieselPM", "LinguisticIsolation", "Poverty", "Unemployment", "LowBirthW

#model with only covariates of interest
data_cal <- data_cal[,c("CensusTract", "TotalPopulation", "CaliforniaCounty", "ZIP", "NearbyCity(tohelp
data_cal1 <- na.omit(data_cal)

#check proportion of missing values
sum(!complete.cases(data_cal))/nrow(data_cal)
```

Modeling

```
## [1] 0.04965775
```

```
library(glmnet)
```

Linear, additive, or other models (LASSO, ridge)

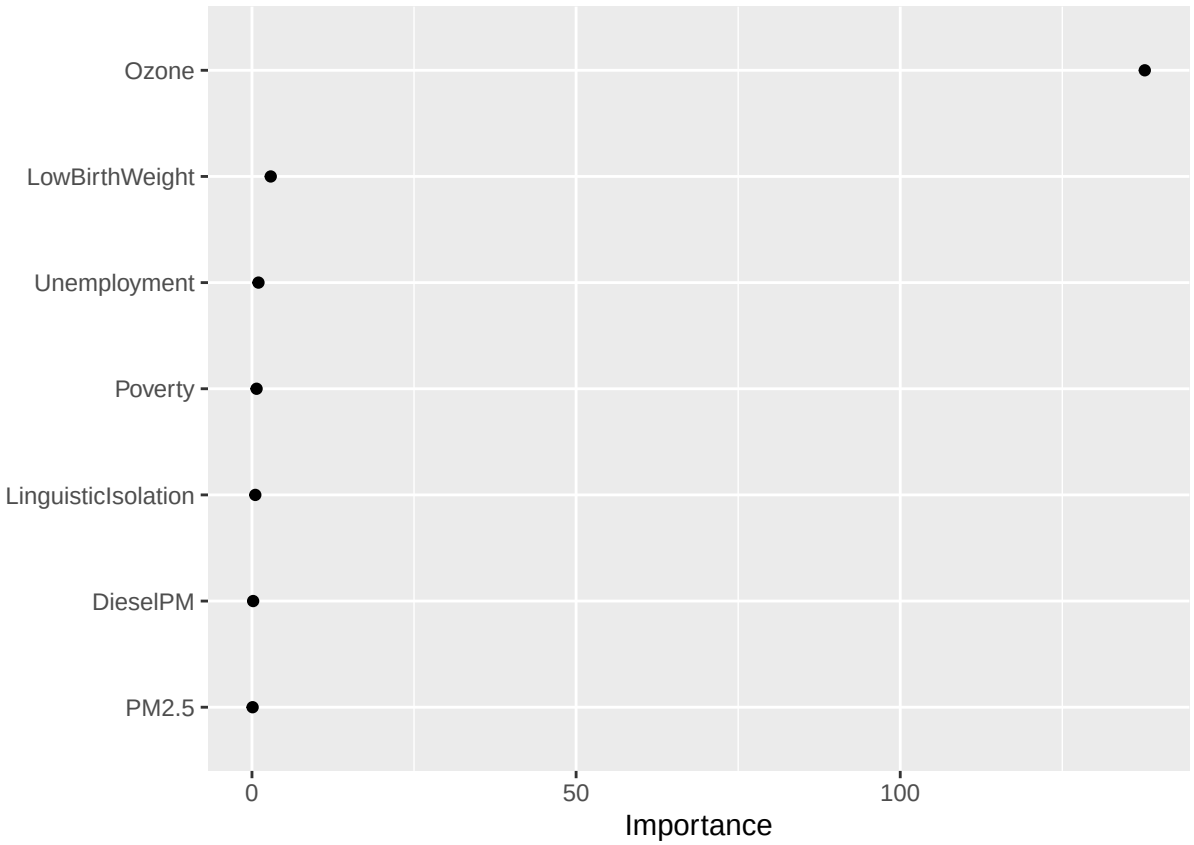
```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-4
```

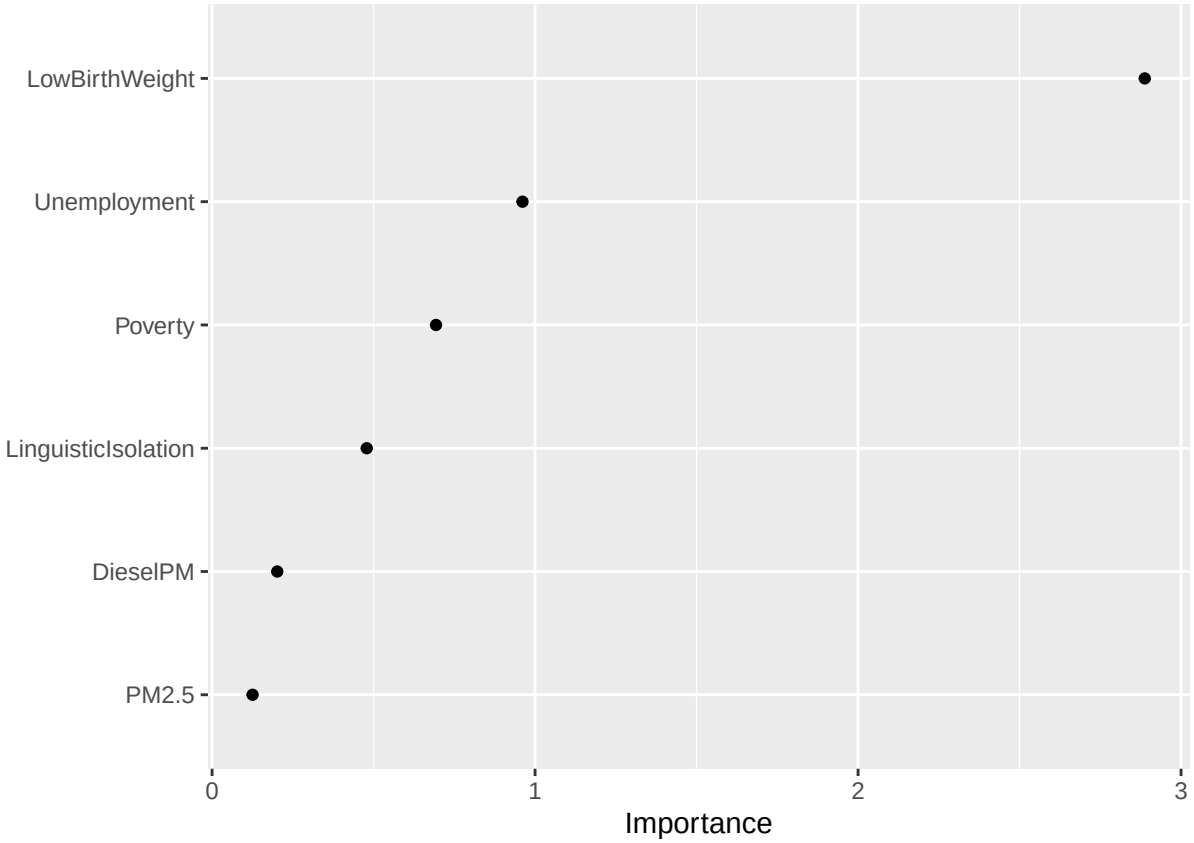
```
library(vip)
```

```
##
## Attaching package: 'vip'

## The following object is masked from 'package:utils':
##
##     vi
```

```
#define outcome variable
y <- data_cal1[,"AsthmaPctl"] |> as.matrix()
#define matrix of predictor variables
x <- data_cal1[, c("Ozone", "PM2.5", "DieselPM", "LinguisticIsolation",  "Poverty", "Unemployment", "Lo
elasticnet.mod = glmnet(x,y,alpha=0.5,family="gaussian")
vip(elasticnet.mod, num_features=10, geom = "point")
```



```r
#now we look at the Variable importance (vip) for factors excluding Ozone
x_noOzone <- data_cal1[, c("PM2.5", "DieselPM", "LinguisticIsolation",  "Poverty", "Unemployment", "Low
elasticnet.mod.noOzone = glmnet(x_noOzone,y,alpha=0.5,family="gaussian")
vip(elasticnet.mod.noOzone, num_features=10, geom = "point")
```

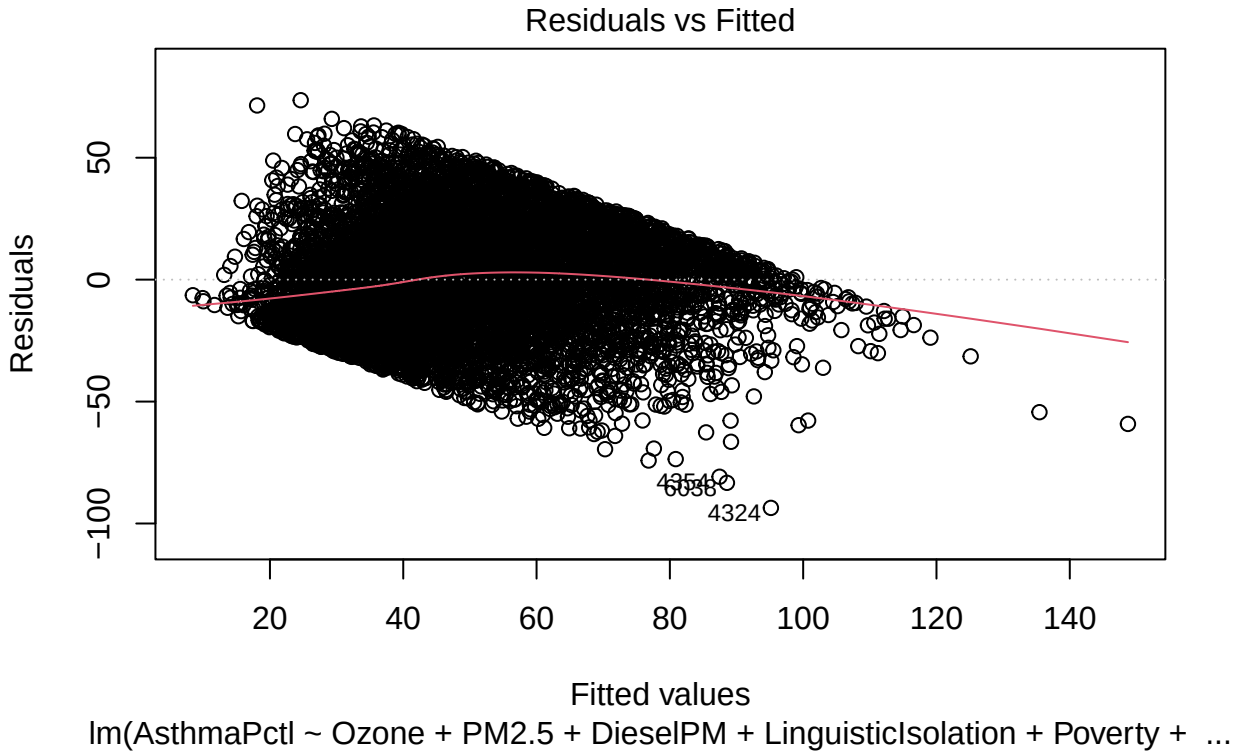


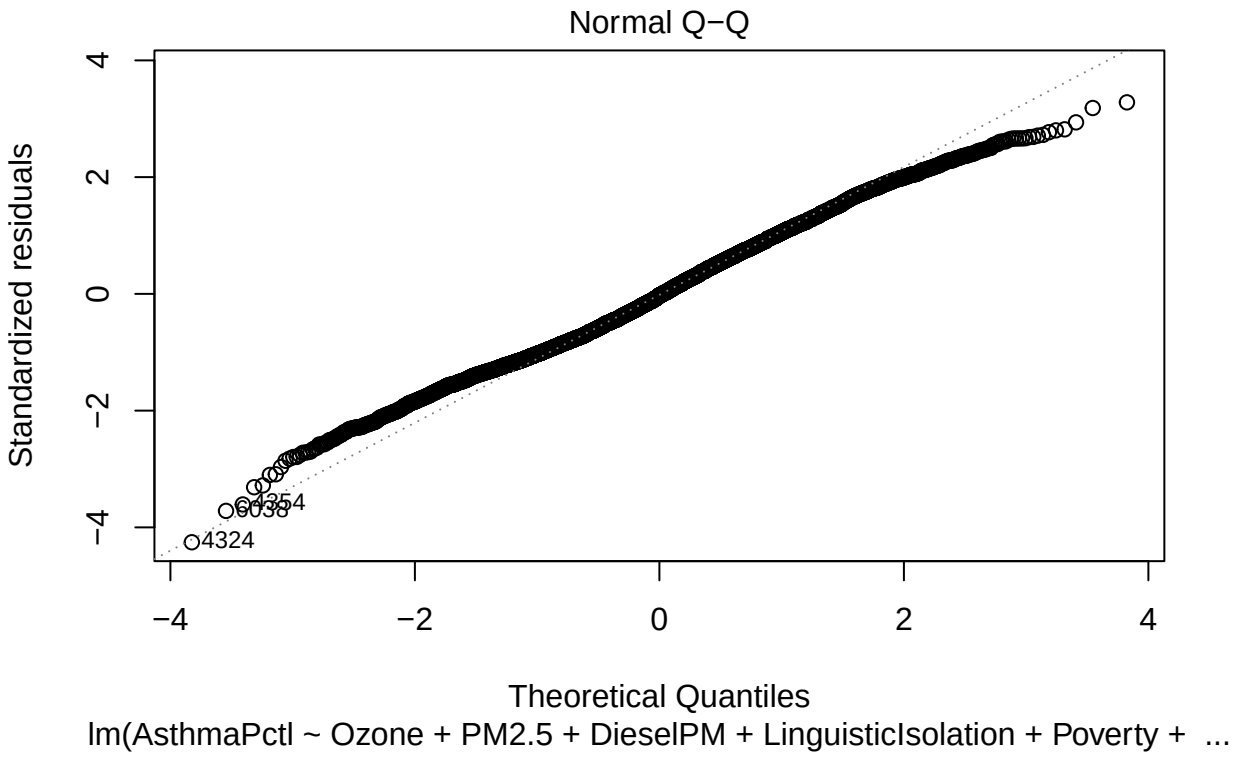

```r
#linear model with all linear terms for 7 covariates of interest
lm_7linear = lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + Unemployment + L
summary(lm_7linear)
```

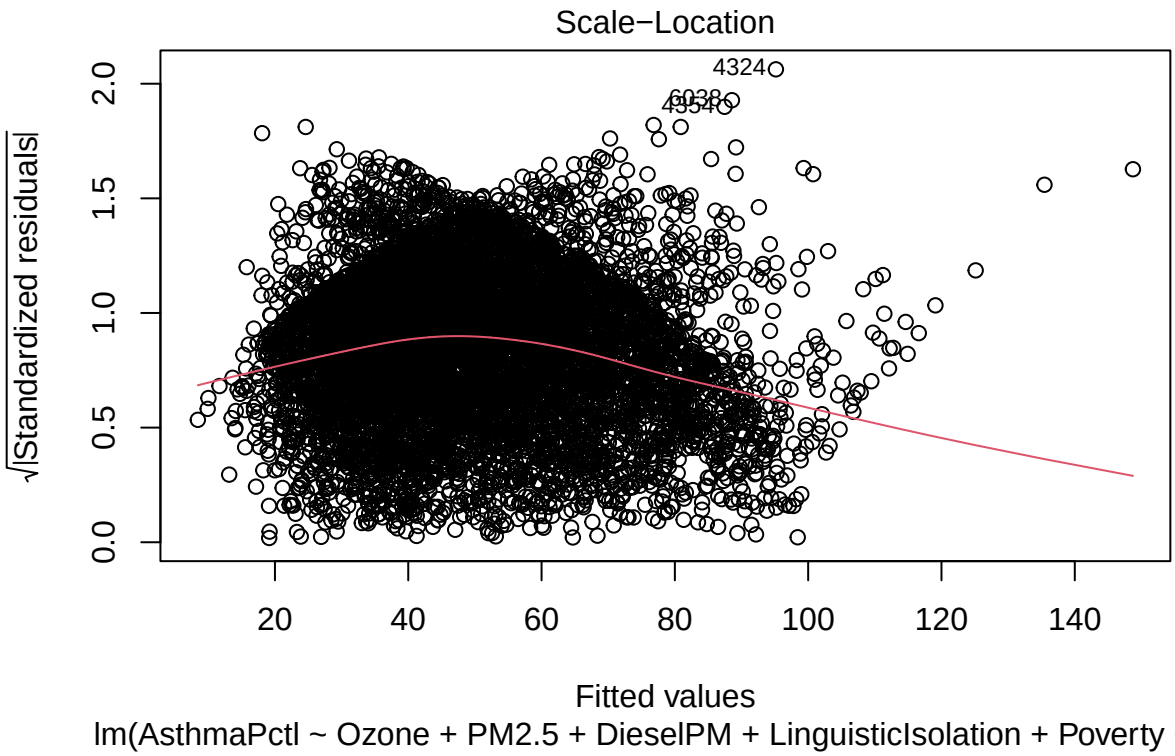
```
##
## Call:
## lm(formula = AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
##     Poverty + Unemployment + LowBirthWeight, data = data_cal1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.601 -16.946  -0.704  16.207  73.572
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            7.52398    1.54370   4.874 1.12e-06 ***
## Ozone               -145.85212   31.32216  -4.657 3.27e-06 ***
## PM2.5                  0.12053    0.12045   1.001    0.317
## DieselPM               0.17899    0.01741  10.279  < 2e-16 ***
## LinguisticIsolation   -0.52567    0.03562 -14.760  < 2e-16 ***
```

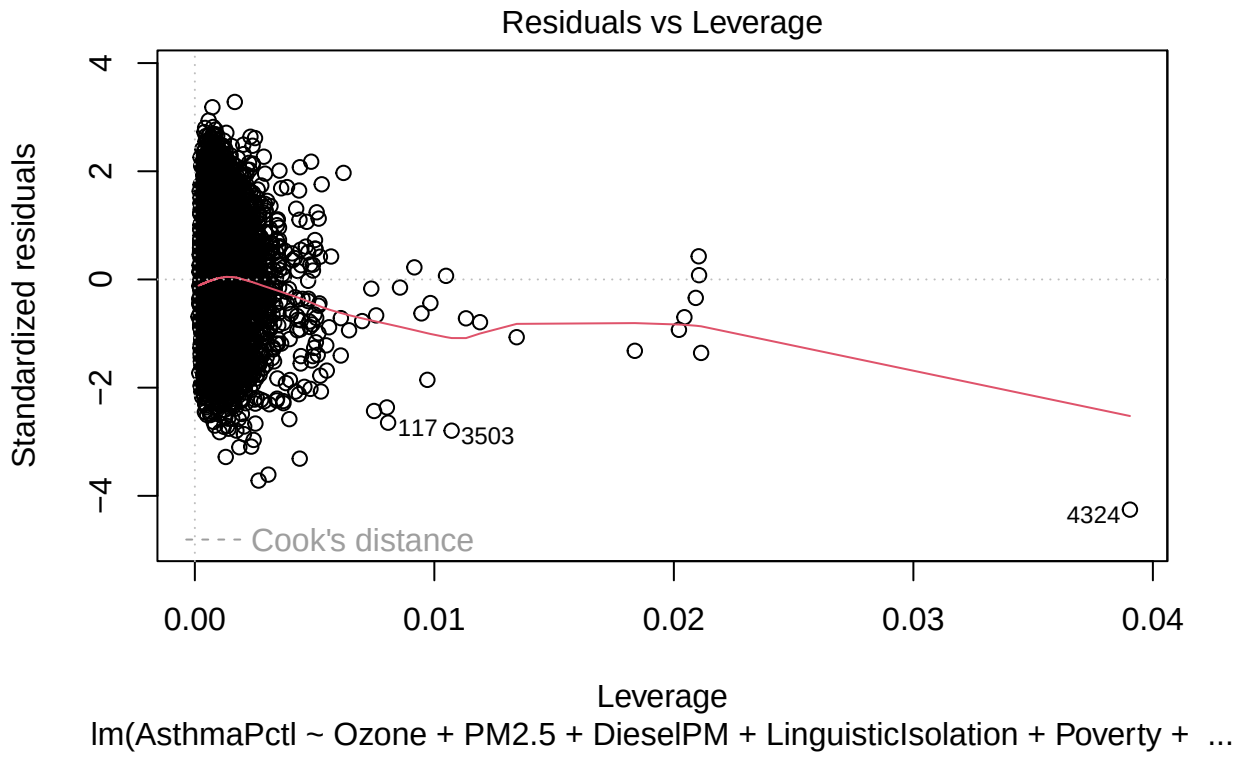
```
## Poverty                0.71665    0.02090  34.281  < 2e-16 ***
## Unemployment           1.00085    0.06935  14.433  < 2e-16 ***
## LowBirthWeight         2.90441    0.17917  16.210  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.44 on 7628 degrees of freedom
## Multiple R-squared:  0.3883, Adjusted R-squared:  0.3877
## F-statistic: 691.7 on 7 and 7628 DF,  p-value: < 2.2e-16
```

```
plot(lm_7linear)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + ...
4324
6038
4354
−4
−2
0
2
4


Scale−Location
4324
6038
4354
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
20
40
60
80
100
120
140
Fitted values
lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + ...




```
library(splines2)
library(foreign)
library(gam)
library(Hmisc)
```

```
## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following objects are masked from 'package:ggmice':
##
##     bwplot, densityplot, stripplot, xyplot

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

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

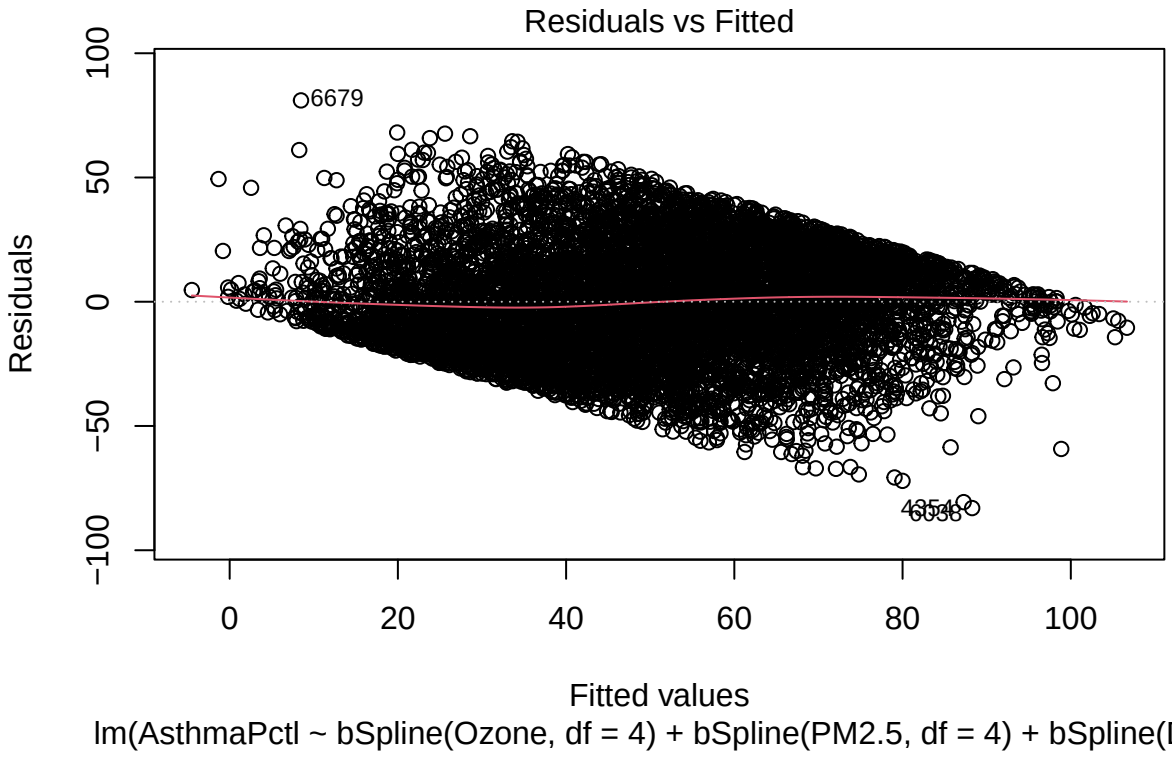
```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

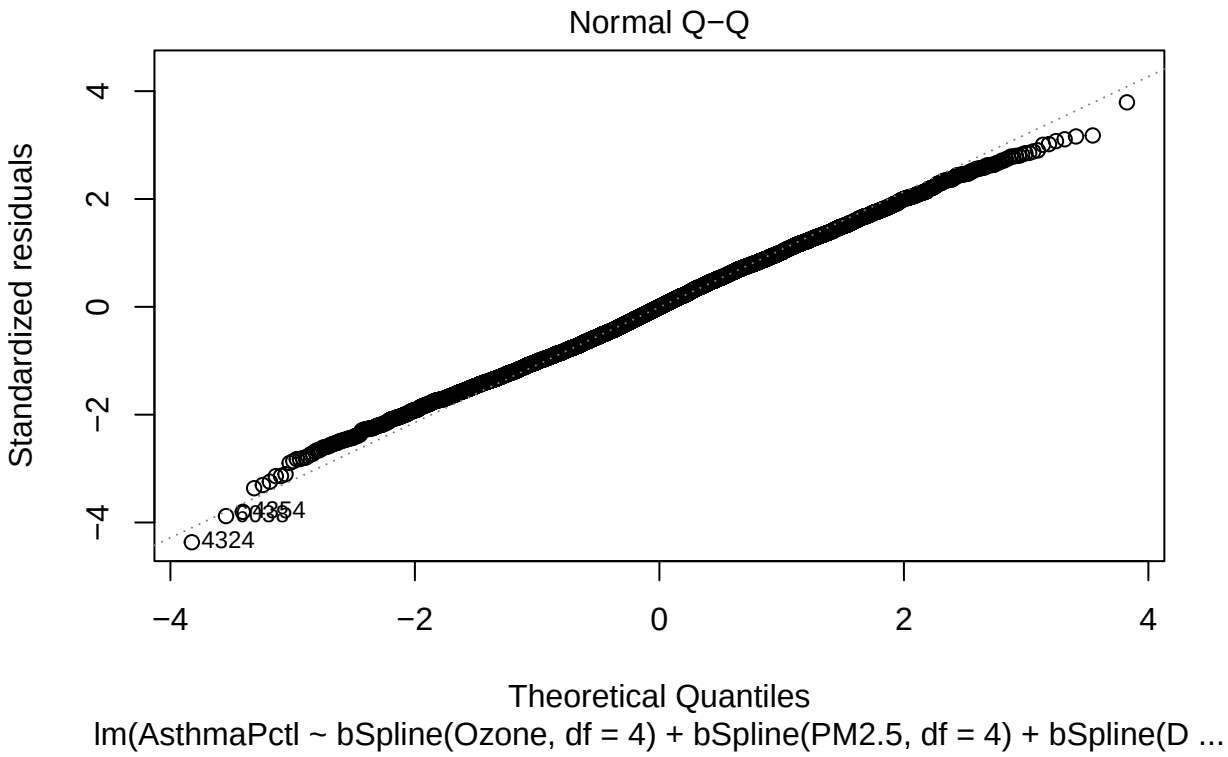
```
#add spline term to every term with significance level <2e-16 & PM2.5
model_spline=lm(AsthmaPctl ~ bSpline(Ozone,df=4) + bSpline(PM2.5, df=4) + bSpline(DieselPM, df=4) + bSp
summary(model_spline)
```

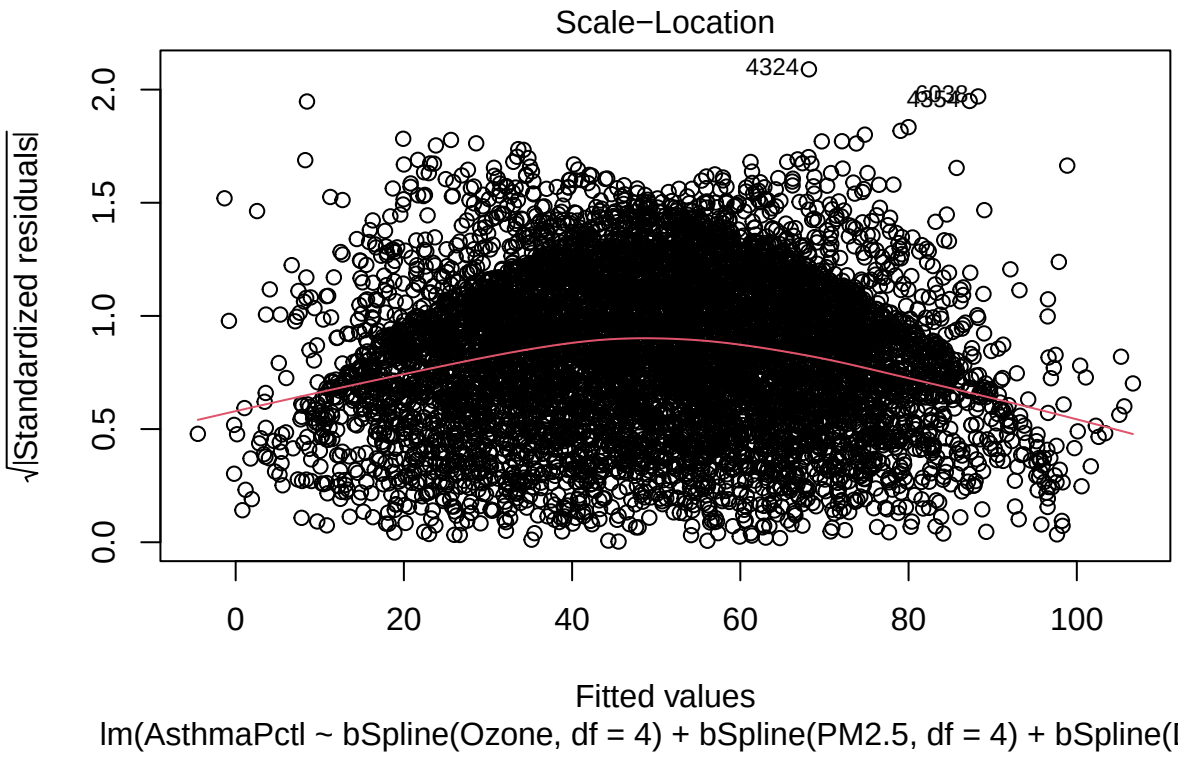
```
##
## Call:
## lm(formula = AsthmaPctl ~ bSpline(Ozone, df = 4) + bSpline(PM2.5,
##     df = 4) + bSpline(DieselPM, df = 4) + bSpline(LinguisticIsolation,
##     df = 4) + bSpline(Poverty, df = 4) + bSpline(Unemployment,
##     df = 4) + bSpline(LowBirthWeight, df = 4), data = data_cal1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.033 -15.578  -0.033  15.320  81.020
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           -44.697      9.155  -4.882 1.07e-06 ***
## bSpline(Ozone, df = 4)1                 1.749      4.142   0.422 0.672886
## bSpline(Ozone, df = 4)2               -19.500      2.451  -7.955 2.05e-15 ***
## bSpline(Ozone, df = 4)3                -2.228      3.602  -0.618 0.536319
## bSpline(Ozone, df = 4)4                -9.197      2.789  -3.298 0.000978 ***
## bSpline(PM2.5, df = 4)1                71.624      8.747   8.188 3.08e-16 ***
## bSpline(PM2.5, df = 4)2                 4.830      5.094   0.948 0.343020
## bSpline(PM2.5, df = 4)3                66.448      8.174   8.130 4.99e-16 ***
## bSpline(PM2.5, df = 4)4                41.338      6.476   6.383 1.83e-10 ***
## bSpline(DieselPM, df = 4)1             15.511      1.600   9.693  < 2e-16 ***
## bSpline(DieselPM, df = 4)2             27.422      4.019   6.824 9.55e-12 ***
## bSpline(DieselPM, df = 4)3             30.199     12.554   2.406 0.016169 *
## bSpline(DieselPM, df = 4)4             19.339     15.014   1.288 0.197781
## bSpline(LinguisticIsolation, df = 4)1  -2.672      1.378  -1.939 0.052520 .
## bSpline(LinguisticIsolation, df = 4)2  -5.140      2.859  -1.798 0.072186 .
## bSpline(LinguisticIsolation, df = 4)3 -25.295      6.090  -4.154 3.31e-05 ***
## bSpline(LinguisticIsolation, df = 4)4 -47.047      8.731  -5.388 7.32e-08 ***
## bSpline(Poverty, df = 4)1              12.827      3.283   3.908 9.40e-05 ***
## bSpline(Poverty, df = 4)2              49.696      2.436  20.404  < 2e-16 ***
## bSpline(Poverty, df = 4)3              48.849      4.326  11.293  < 2e-16 ***
## bSpline(Poverty, df = 4)4              58.928      4.391  13.419  < 2e-16 ***
## bSpline(Unemployment, df = 4)1         -3.187      4.430  -0.719 0.471890
## bSpline(Unemployment, df = 4)2         25.293      3.472   7.285 3.53e-13 ***
## bSpline(Unemployment, df = 4)3         16.250      7.830   2.075 0.037984 *
## bSpline(Unemployment, df = 4)4         13.685     12.235   1.118 0.263410
## bSpline(LowBirthWeight, df = 4)1        1.564      6.071   0.258 0.796735
## bSpline(LowBirthWeight, df = 4)2       13.033      4.196   3.106 0.001904 **
## bSpline(LowBirthWeight, df = 4)3       28.988      7.702   3.764 0.000169 ***
## bSpline(LowBirthWeight, df = 4)4       16.656     13.594   1.225 0.220522
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.45 on 7607 degrees of freedom
```

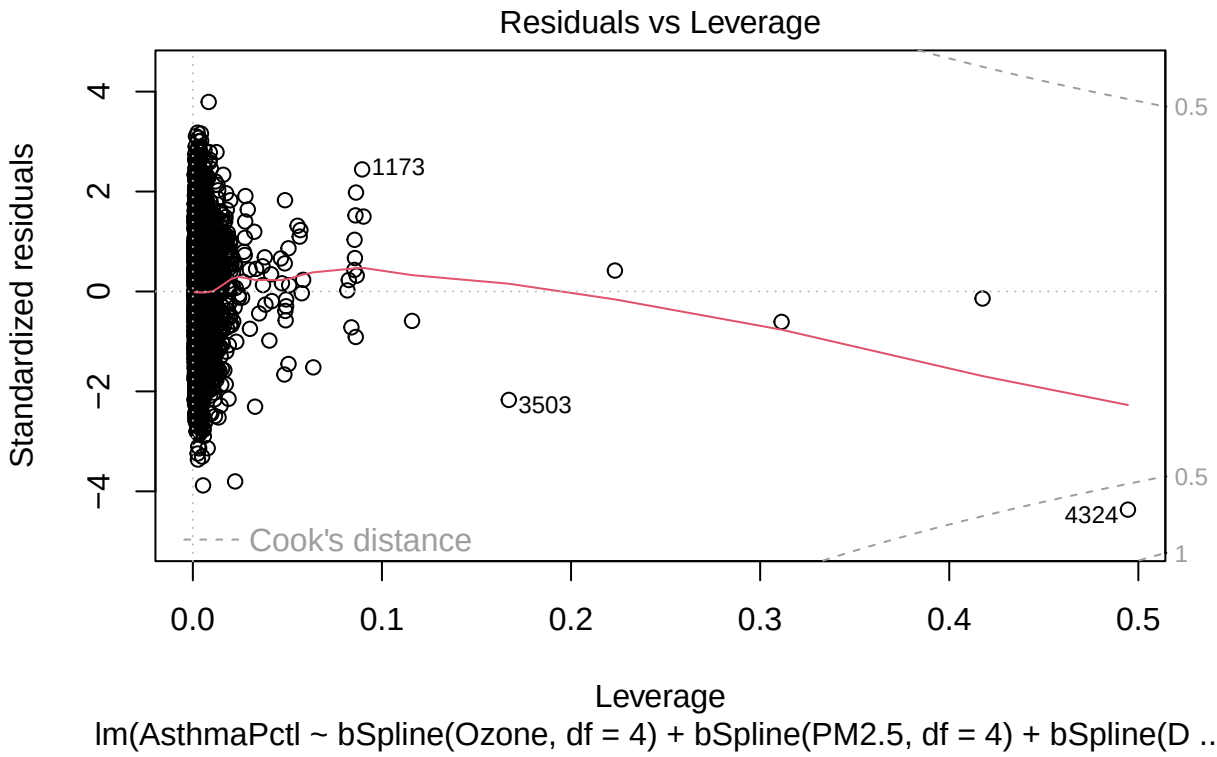
```
## Multiple R-squared:  0.4426, Adjusted R-squared:  0.4405
## F-statistic: 215.7 on 28 and 7607 DF,  p-value: < 2.2e-16
```

```
#model evaluation
plot(model_spline)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AsthmaPctl ~ bSpline(Ozone, df = 4) + bSpline(PM2.5, df = 4) + bSpline(D ...
4324
−4
−2
0
2
4


Scale−Location
4324
6038
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
20
40
60
80
100
Fitted values
lm(AsthmaPctl ~ bSpline(Ozone, df = 4) + bSpline(PM2.5, df = 4) + bSpline(D ...




```
#Ridge
library(dplyr)
library(MASS)
```

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

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

```
fit = lm.ridge(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + Unemployment + L
#view summary of model
summary(fit)
```

```
##        Length Class  Mode
## coef   2807   -none- numeric
## scales    7   -none- numeric
## Inter     1   -none- numeric
## lambda  401   -none- numeric
## ym        1   -none- numeric
## xm        7   -none- numeric
## GCV     401   -none- numeric
## kHKB      1   -none- numeric
## kLW       1   -none- numeric
```

More flexible modeling:

```
#interaction terms
lm_interAir = lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + Unemployment +
lm_interDiesel = lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + Unemployment
lm_interLing = lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + Unemployment +
lm_interPov = lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + Unemployment +
lm_interEmp = lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + Unemployment +
lm_interWgt = lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + Unemployment +
lm_interSoc = lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + Unemployment +

#quadratic terms
lm_quadOzone = lm(AsthmaPctl ~ Ozone + I(Ozone)^2+PM2.5 + DieselPM + LinguisticIsolation + Poverty + Un
lm_quadWgt = lm(AsthmaPctl ~ Ozone + I(LowBirthWeight)^2 + PM2.5 + DieselPM + LinguisticIsolation + Pov
```

Model comparison:

```
mods_linear <- list(lm_7linear, model_spline, lm_interWgt, lm_interEmp, lm_interPov, lm_interLing, lm_i
mod_names <- c("7Linear", "spline", "interWgt", "interEmp", "interPov", "interLing", "interDiesel", "in

library(AICcmodavg)
```

```
## Warning: package 'AICcmodavg' was built under R version 4.2.2
```

```
##
## Attaching package: 'AICcmodavg'
```

```
## The following object is masked from 'package:VGAM':
##
##     AICc
```

```
aictab(cand.set = mods_linear, modnames = mod_names)
```

```
## Warning in aictab.AIClm(cand.set = mods_linear, modnames = mod_names):
## Check model structure carefully as some models may be redundant
```

```
##
## Model selection based on AICc:
##
##              K     AICc Delta_AICc AICcWt Cum.Wt        LL
## spline      30 68522.95       0.00      1      1 -34231.35
## interAir    10 69061.24     538.29      0      1 -34520.61
## internSoc   10 69084.59     561.64      0      1 -34532.28
## interEmp    10 69154.66     631.71      0      1 -34567.32
## interLing   10 69179.26     656.31      0      1 -34579.62
## interWgt    10 69185.02     662.07      0      1 -34582.49
## interPov    10 69189.31     666.36      0      1 -34584.64
## 7Linear      9 69190.24     667.29      0      1 -34586.11
## quadOzon     9 69190.24     667.29      0      1 -34586.11
## quadWgt      9 69190.24     667.29      0      1 -34586.11
## interDiesel 10 69191.43     668.47      0      1 -34585.70
```

```
r_square <- c(summary(lm_7linear)$adj.r.squared, summary(model_spline)$adj.r.squared, summary(lm_interWg
cbind(mod_names,r_square)
```

```
##       mod_names     r_square
##  [1,] "7Linear"     "0.387716396235505"
##  [2,] "spline"      "0.440504028459097"
##  [3,] "interWgt"    "0.388215257887555"
##  [4,] "interEmp"    "0.390642416255224"
##  [5,] "interPov"    "0.387871548512763"
##  [6,] "interLing"   "0.388676145598628"
##  [7,] "interDiesel" "0.387701699931608"
##  [8,] "internSoc"   "0.396208720178323"
##  [9,] "interAir"    "0.398052138846566"
## [10,] "quadOzon"    "0.387716396235505"
## [11,] "quadWgt"     "0.387716396235505"
```

```
f_score <- c(summary(lm_7linear)$fstatistic[1], summary(model_spline)$fstatistic[1], summary(lm_interWg
cbind(mod_names, f_score)
```

```
##       mod_names     f_score
## value "7Linear"     "691.673095898192"
## value "spline"      "215.686101953418"
## value "interWgt"    "606.609965797965"
## value "interEmp"    "612.823608926694"
## value "interPov"    "605.734036152844"
## value "interLing"   "607.786065658977"
## value "interDiesel" "605.30154686188"
## value "internSoc"   "627.262269690055"
## value "interAir"    "632.102848814442"
## value "quadOzon"    "691.673095898192"
## value "quadWgt"     "691.673095898192"
```

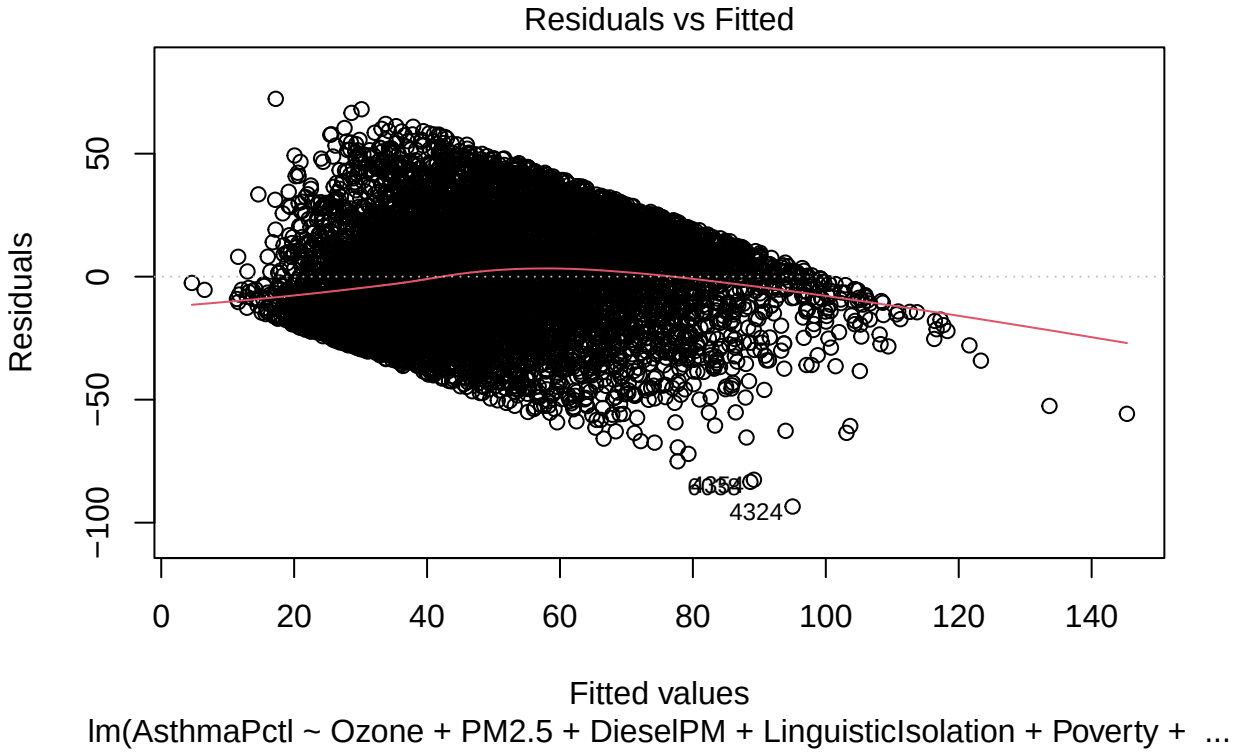
Further interpretation of lm_interAir

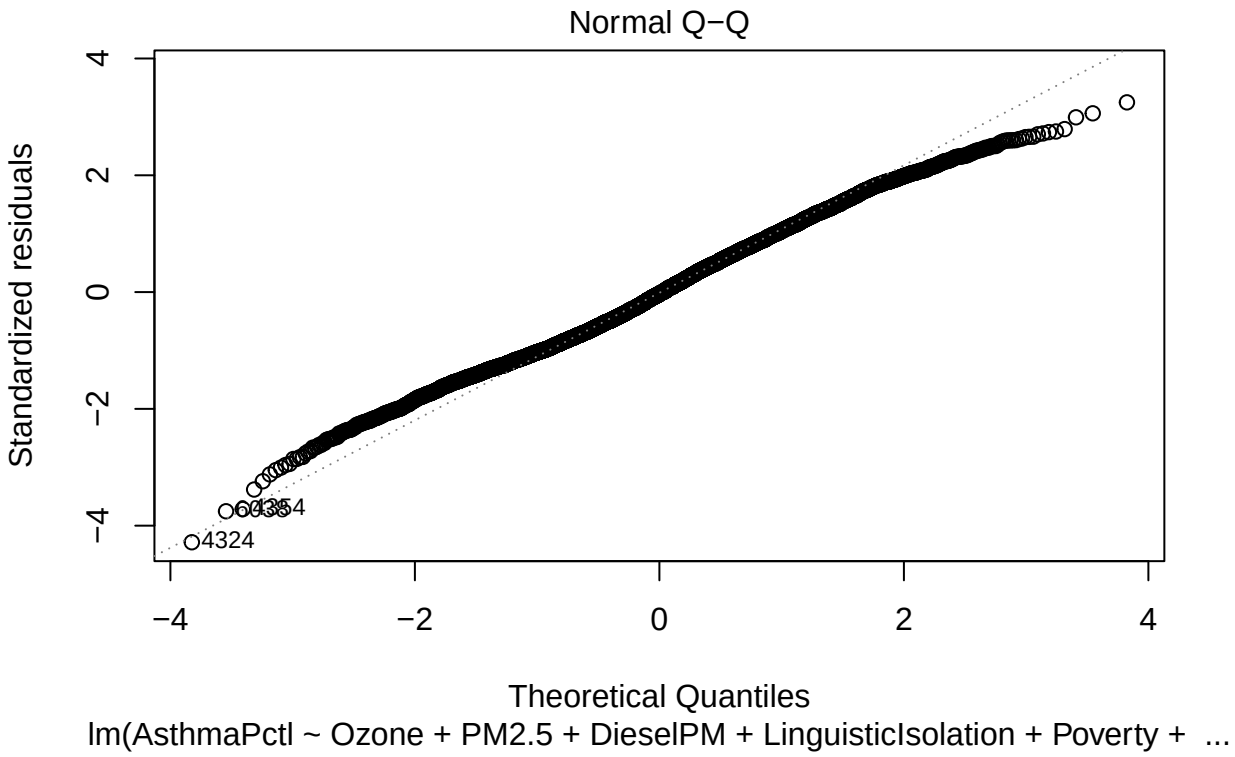
```
summary(lm_interAir)
```

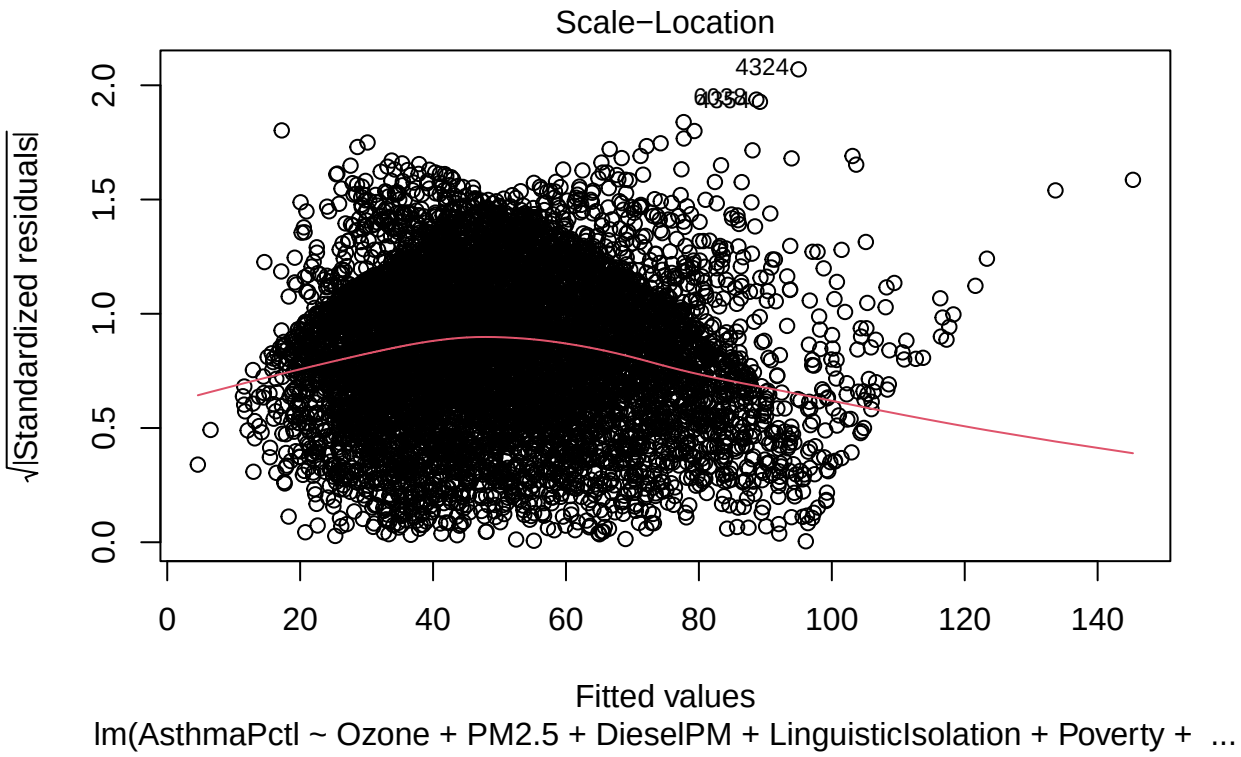
```
##
## Call:
## lm(formula = AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
##     Poverty + Unemployment + LowBirthWeight + Ozone * PM2.5,
##     data = data_cal1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -93.44 -16.65  -0.65  16.10  72.27
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.911e+01  4.744e+00   12.46   <2e-16 ***
## Ozone               -1.130e+03  9.110e+01  -12.40   <2e-16 ***
## PM2.5               -5.472e+00  5.012e-01  -10.92   <2e-16 ***
```

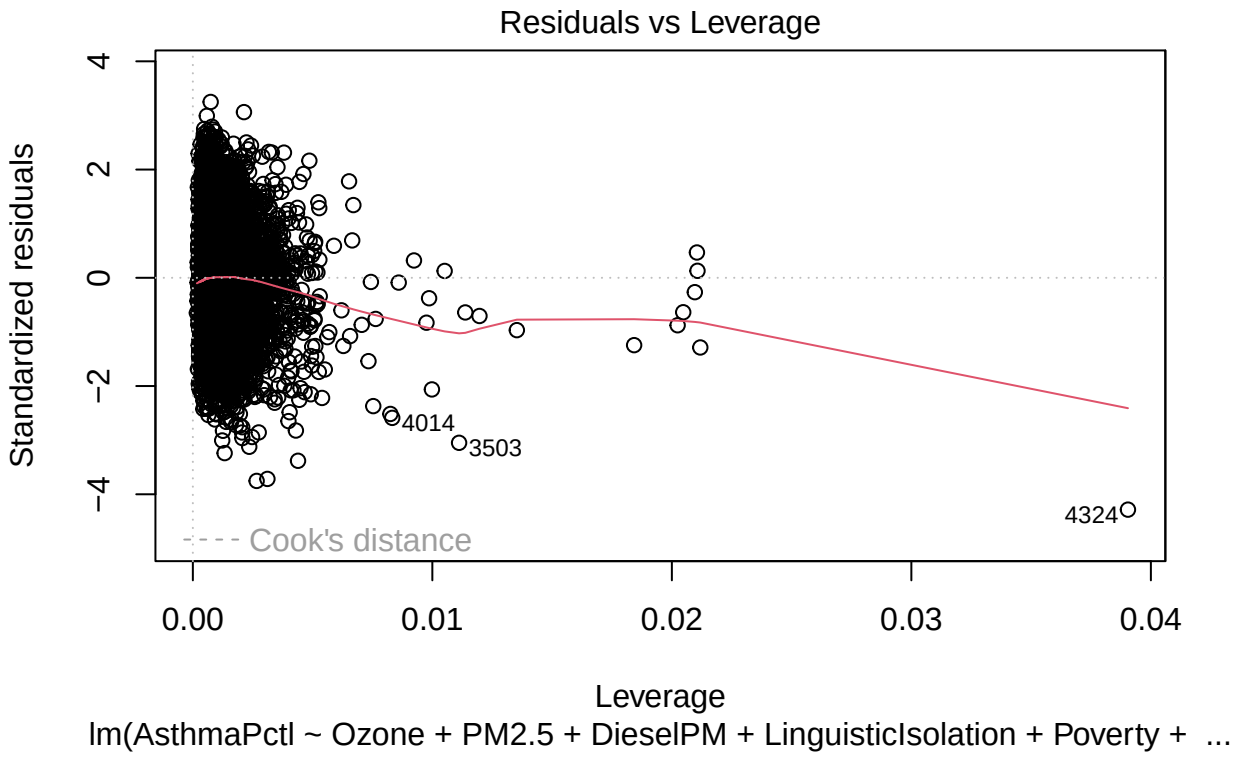
```
## DieselPM             1.872e-01  1.728e-02   10.83   <2e-16 ***
## LinguisticIsolation -4.636e-01  3.572e-02  -12.98   <2e-16 ***
## Poverty              6.887e-01  2.087e-02   33.00   <2e-16 ***
## Unemployment         1.025e+00  6.879e-02   14.90   <2e-16 ***
## LowBirthWeight       3.057e+00  1.782e-01   17.16   <2e-16 ***
## Ozone:PM2.5          1.041e+02  9.059e+00   11.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.25 on 7627 degrees of freedom
## Multiple R-squared:  0.3987, Adjusted R-squared:  0.3981
## F-statistic: 632.1 on 8 and 7627 DF,  p-value: < 2.2e-16
```

```
plot(lm_interAir)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + ...
4324
−4
−2
0
2
4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
20
40
60
80
100
120
140
4324
Fitted values
lm(AsthmaPctl ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + ...




```
library(dplyr)
library(tidyverse)
#round the Asthma rate to make sure that it's integers
dataint <- data_cal_old |> mutate(Asthma = round(Asthma))

mod.poisson.full <- glm(Asthma ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Pesticides +
    Tox.Release + Traffic + CleanupSites + GroundwaterThreats +
    Haz.Waste + Imp.WaterBodies + PollutionBurden + LinguisticIsolation +
    Poverty + Unemployment + HousingBurden + LowBirthWeight +
    CardiovascularDisease, data=dataint, family=quasipoisson)
summary(mod.poisson.full)
```

Poisson

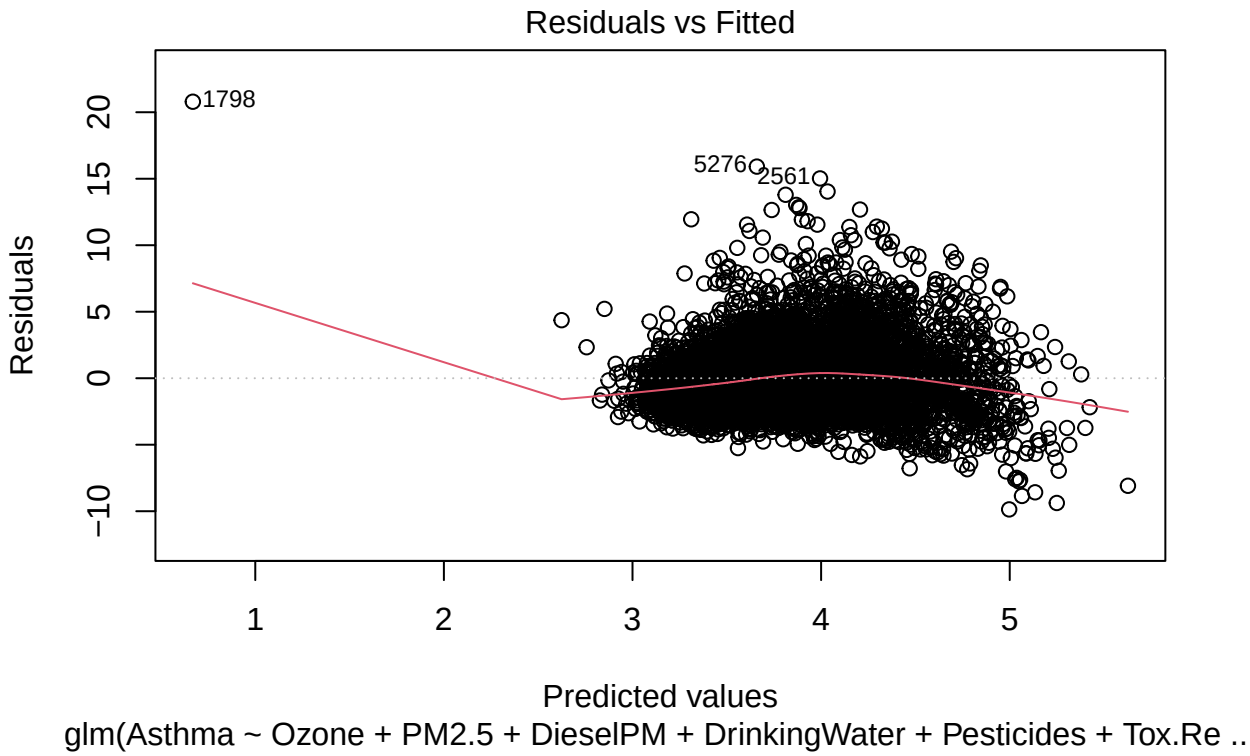
```
##
## Call:
## glm(formula = Asthma ~ Ozone + PM2.5 + DieselPM + DrinkingWater +
##     Pesticides + Tox.Release + Traffic + CleanupSites + GroundwaterThreats +
##     Haz.Waste + Imp.WaterBodies + PollutionBurden + LinguisticIsolation +
##     Poverty + Unemployment + HousingBurden + LowBirthWeight +
##     CardiovascularDisease, family = quasipoisson, data = dataint)
##
## Deviance Residuals:
```

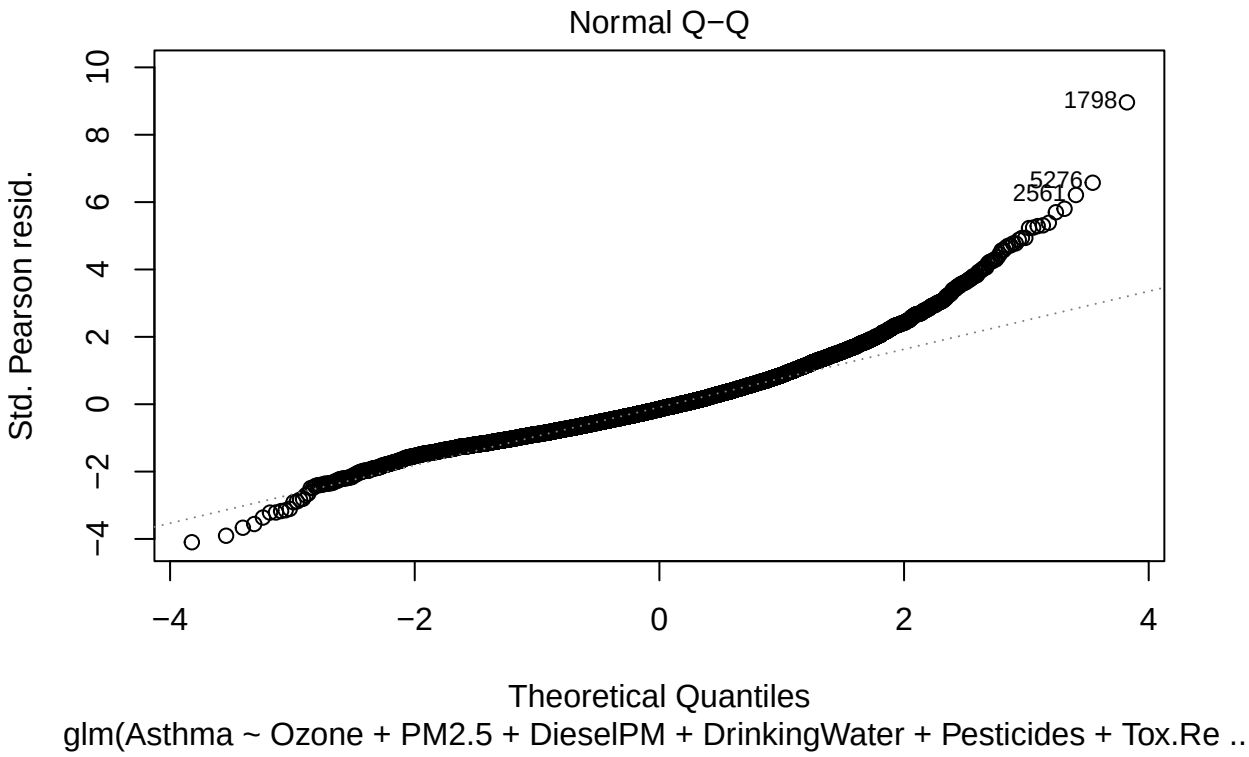
```
##      Min        1Q    Median        3Q       Max
## -12.1141   -1.6975   -0.3578    1.1534   12.3134
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.891e+00  2.712e-02 106.608  < 2e-16 ***
## Ozone                 -1.044e+01  5.631e-01 -18.541  < 2e-16 ***
## PM2.5                  1.616e-02  2.075e-03   7.787 7.77e-15 ***
## DieselPM               3.223e-03  2.344e-04  13.752  < 2e-16 ***
## DrinkingWater         -2.153e-04  2.224e-05  -9.681  < 2e-16 ***
## Pesticides             4.029e-06  1.445e-06   2.787  0.00533 **
## Tox.Release           -3.771e-06  5.944e-07  -6.345 2.36e-10 ***
## Traffic               -4.761e-06  4.998e-06  -0.952  0.34093
## CleanupSites           1.123e-03  2.644e-04   4.247 2.20e-05 ***
## GroundwaterThreats     4.874e-04  8.905e-05   5.473 4.57e-08 ***
## Haz.Waste              6.068e-03  2.851e-03   2.129  0.03331 *
## Imp.WaterBodies        1.661e-03  9.840e-04   1.688  0.09154 .
## PollutionBurden       -1.430e-03  5.913e-04  -2.418  0.01563 *
## LinguisticIsolation   -5.679e-03  5.306e-04 -10.703  < 2e-16 ***
## Poverty                6.813e-03  3.714e-04  18.345  < 2e-16 ***
## Unemployment           7.333e-03  9.791e-04   7.489 7.73e-14 ***
## HousingBurden          1.296e-03  6.450e-04   2.009  0.04458 *
## LowBirthWeight         4.529e-02  2.676e-03  16.926  < 2e-16 ***
## CardiovascularDisease  1.060e-01  1.485e-03  71.393  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 5.866488)
##
##     Null deviance: 118332  on 7556  degrees of freedom
## Residual deviance:  41860  on 7538  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

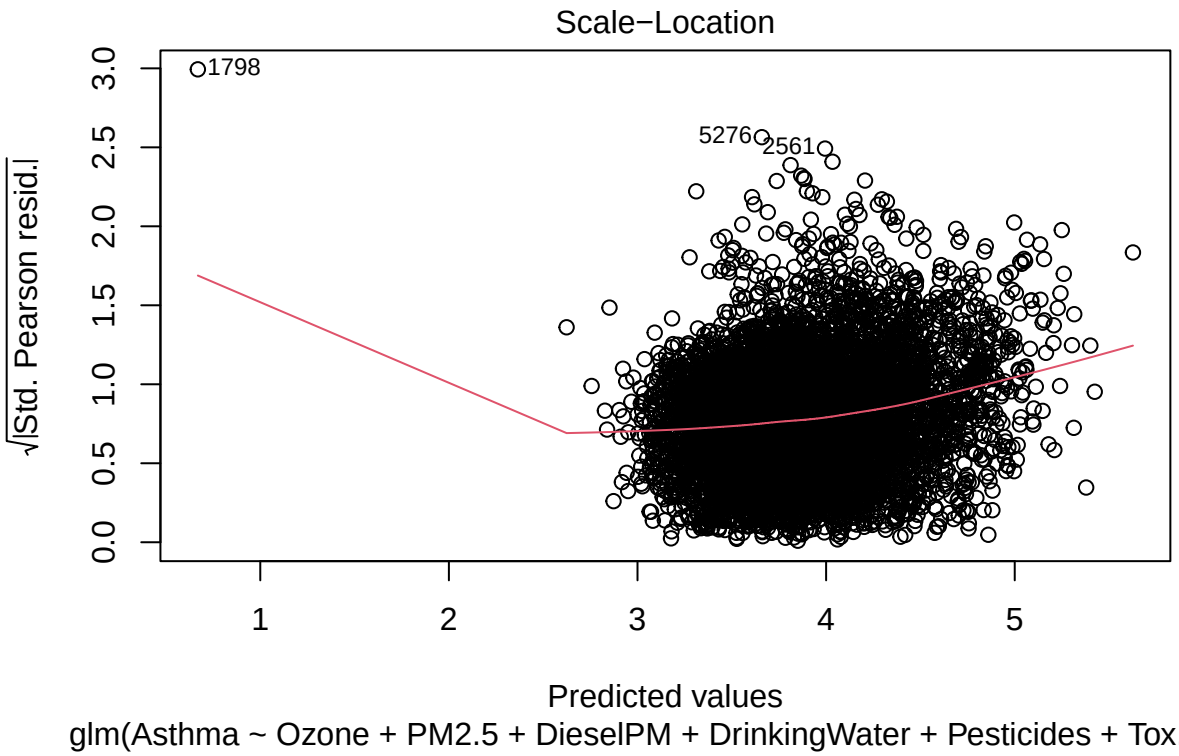
```
sort(coef(mod.poisson.full) , decreasing = TRUE)
```

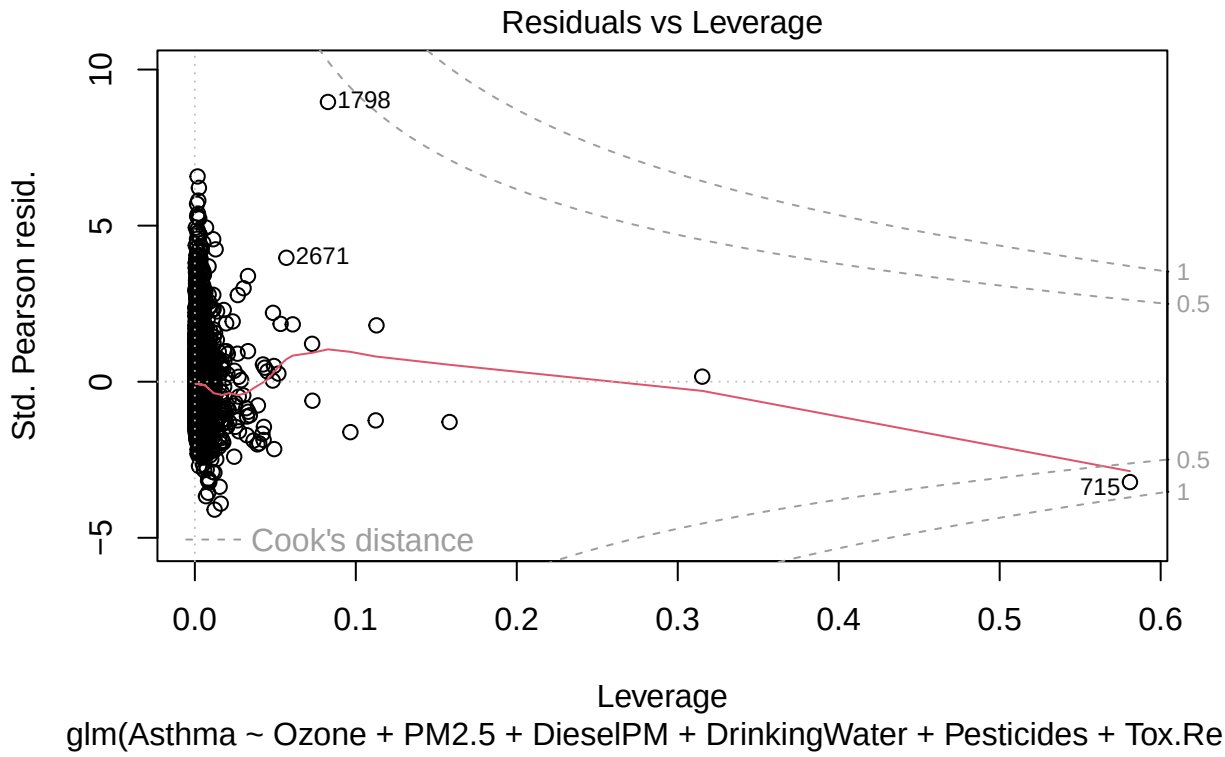
```
##           (Intercept) CardiovascularDisease        LowBirthWeight
##          2.890944e+00          1.060299e-01          4.529040e-02
##                 PM2.5          Unemployment               Poverty
##          1.615568e-02          7.332720e-03          6.813452e-03
##             Haz.Waste              DieselPM       Imp.WaterBodies
##          6.068494e-03          3.223275e-03          1.660507e-03
##         HousingBurden          CleanupSites    GroundwaterThreats
##          1.295736e-03          1.122964e-03          4.873904e-04
##            Pesticides           Tox.Release               Traffic
##          4.028706e-06         -3.771252e-06         -4.760469e-06
##         DrinkingWater       PollutionBurden   LinguisticIsolation
##         -2.153015e-04         -1.429799e-03         -5.679291e-03
##                 Ozone
##         -1.044095e+01
```

```
plot(mod.poisson.full)
```




Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm(Asthma ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Pesticides + Tox.Re ..
1798
5276
2561
−4
−2
0
2
4
6
8
10


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1798
5276
2561
1
2
3
4
5
Predicted values
glm(Asthma ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Pesticides + Tox.Re ..



From the summary statistics of the full model, the Cardiovascular Disease, Low Birth Weight and PM 2.5 are the three major contributors to Asthma rate, using Poisson models. While Ozone is the factor that is negatively correlated with Asthma rate, as is shown in the linear model.

Since pesticides, traffic, Haz.waste, Imp.WaterBodies, PollutionBurden and Housing burden variables have both relatively lower p-values and small Since drinking water, cleanup sites and Ground Water threats have relatively very low coefficients, we exclude their effects for the next round of analysis.

Thus, what we keep for the next round of analysis are 8 variables: PM2.5, Ozone, DieselPM, Unemployment, Poverty, Linguistic Isolation, Low Birth Weight and Cardiovascular Diseases.

checking for dispersion:

```
deviance(mod.poisson.full)/mod.poisson.full$df.residual
```

```
## [1] 5.553204
```

Since the quotient is greater than 1, There exists some form of overdispersion in the data.

First, we look at the model with the air pollution factors: We want to approximate the effects of Pmm 2.5 and Asthma

```
mod.p_base <- glm(Asthma ~ PM2.5, data=dataint, family=quasipoisson)
summary(mod.p_base)
```

```
##
## Call:
```

```
## glm(formula = Asthma ~ PM2.5, family = quasipoisson, data = dataint)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.6372  -3.2860  -0.9569   1.7947  17.9512
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.748848   0.028041  133.69  < 2e-16 ***
## PM2.5       0.020259   0.002577    7.86 4.37e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 17.52806)
##
##     Null deviance: 118332  on 7556  degrees of freedom
## Residual deviance: 117254  on 7555  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(mod.p_base)[2])
```

```
##    PM2.5
## 1.020465
```

```
deviance(mod.p_base)/mod.p_base$df.residual
```

```
## [1] 15.52003
```

The model with Pm 2.5 alone has very great overdispersion.

From the summary statistics: we can see that with every 1 more unit increase in PM 2.5 concentration is estimated to be associated with, on average, 2% increase in the incidence rate of asthma among individuals that live in the counties with the current level of PM2.5 concentration.

```
mod.p_pmd <- glm(Asthma ~ PM2.5 + DieselPM, data=dataint, family=poisson())
summary(mod.p_pmd)
```

```
##
## Call:
## glm(formula = Asthma ~ PM2.5 + DieselPM, family = poisson(),
##     data = dataint)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.9550  -3.1749  -0.8681   1.7670  18.5001
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.715e+00  6.802e-03  546.13   <2e-16 ***
## PM2.5       1.343e-02  6.322e-04   21.24   <2e-16 ***
```

```
## DieselPM    5.203e-03  8.197e-05   63.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 118332  on 7556  degrees of freedom
## Residual deviance: 113709  on 7554  degrees of freedom
## AIC: 156411
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(mod.p_pmd)[2])
```

```
##    PM2.5
## 1.013516
```

```
exp(coef(mod.p_pmd)[3])
```

```
## DieselPM
## 1.005216
```

From the summary statistics: we can see that with every 1 more unit increase in PM 2.5 concentration is estimated to be associated with, on average, 1.3% increase in the incidence rate of asthma among individuals that live in the counties with the current level of PM2.5 concentration.

Diesel is likely a confonder for the effects of PM 2.5, because dieselPM contributes to a portion of PM 2.5, while the total PM2.5 amount doesn't directly lead to DieselPM. The changes in coefficient of PM 2.5: (0.02-0.013)/0.013 = 53.8% > 10% Thus, diesel PM satisfies the confounding effect.

Then we test the effect modifying of DieselPM: The interaction terms has a p-value below the significant threshold. SO we can conclude that it is an effect modifier of PM 2.5.

```
mod.p_pmdinter <- glm(Asthma ~ PM2.5 * DieselPM, data=dataint, family=poisson())
summary(mod.p_pmdinter)
```

```
##
## Call:
## glm(formula = Asthma ~ PM2.5 * DieselPM, family = poisson(),
##     data = dataint)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.6138  -3.1880  -0.8655   1.7323  18.2446
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     3.850e+00  9.749e-03 394.923  < 2e-16 ***
## PM2.5          -2.559e-04  9.501e-04  -0.269    0.788
## DieselPM       -2.972e-03  4.346e-04  -6.839 7.96e-12 ***
## PM2.5:DieselPM  7.843e-04  4.077e-05  19.235  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 118332  on 7556  degrees of freedom
## Residual deviance: 113341  on 7553  degrees of freedom
## AIC: 156045
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(mod.p_pmdinter)[2])
```

```
##     PM2.5
## 0.9997441
```

```
exp(coef(mod.p_pmdinter)[3])
```

```
##  DieselPM
## 0.9970319
```

Checking for Pm 2.5 + Ozone as the predictor

```
mod.p_ozone <- glm(Asthma ~ PM2.5 + Ozone, data=dataint, family=poisson())
summary(mod.p_ozone)
```

```
##
## Call:
## glm(formula = Asthma ~ PM2.5 + Ozone, family = poisson(), data = dataint)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.5015  -3.2701  -0.9574   1.7453  18.0991
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.6920347  0.0084332  437.80   <2e-16 ***
## PM2.5       0.0169882  0.0006795   25.00   <2e-16 ***
## Ozone       1.9153022  0.1721921   11.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 118332  on 7556  degrees of freedom
## Residual deviance: 117131  on 7554  degrees of freedom
## AIC: 159833
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(mod.p_ozone)[2])
```

```
##    PM2.5
## 1.017133
```

```
exp(coef(mod.p_ozone)[3])
```

```
##   Ozone
## 6.78899
```

From the summary statistics: we can see that with every 1 more unit increase in ozone concentration is estimated to be associated with, on average, 600% increase in the incidence rate of asthma among individuals that live in the counties with the current level of ozone concentration, holding PM 2.5 constant

Something to note: Ozone showed a very significant positive association with Asthma rate when modeled with Asthma rate alone, but showed negative association in the full model, which is open for later discussion and examinations.

Testing whether we can use low birth weight as predictors for Asthma rate:

```
#coef(mod.p_ozone)
mod.p_diseases <- glm(Asthma ~ LowBirthWeight, data=dataint, family=quasipoisson)
summary(mod.p_diseases)
```

```
##
## Call:
## glm(formula = Asthma ~ LowBirthWeight, family = quasipoisson,
##     data = dataint)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -10.9131   -2.8769   -0.7546    1.7829   16.5595
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.31314    0.02126  155.83   <2e-16 ***
## LowBirthWeight  0.12631    0.00385   32.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 14.84471)
##
##     Null deviance: 118332  on 7556  degrees of freedom
## Residual deviance: 102709  on 7555  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coef(mod.p_diseases)[2])
```

```
## LowBirthWeight
##       1.134639
```

```
#exp(coef(mod.p_diseases)[3])

summary(mod.p_diseases)$dispersion
```

```
## [1] 14.84471
```

With every 1 unit increase in percent of population with Low Birth Weight, is estimated to be associated with, on average,13% increase in the incidence rate of asthma among individuals.

Testing whether socioeconomic factors can be prediction factors for Asthma.

```
mod.p_soecon <- glm(Asthma ~ Poverty + Unemployment + LinguisticIsolation, data=dataint, family=poisson
summary(mod.p_soecon)
```

```
##
## Call:
## glm(formula = Asthma ~ Poverty + Unemployment + LinguisticIsolation,
##     family = poisson(), data = dataint)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.8954  -2.4703  -0.7264   1.4618  17.8889
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          3.3347631  0.0039923  835.31   <2e-16 ***
## Poverty              0.0132957  0.0001208  110.03   <2e-16 ***
## Unemployment         0.0184299  0.0003835   48.05   <2e-16 ***
## LinguisticIsolation -0.0085289  0.0001998  -42.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 118332  on 7556  degrees of freedom
## Residual deviance:  80997  on 7553  degrees of freedom
## AIC: 123701
##
## Number of Fisher Scoring iterations: 4
```

poverty and unemployment both demonstrated positive relationship with Asthma, while linguistic isolation demonstrated negative relationship with Asthma.

```
mod.poisson.sim <- glm(Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
    Poverty + Unemployment+ LowBirthWeight, data=dataint, family=quasipoisson)
summary(mod.poisson.sim)
```

```
##
## Call:
## glm(formula = Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
##     Poverty + Unemployment + LowBirthWeight, family = quasipoisson,
##     data = dataint)
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -14.5085   -2.3304   -0.6375    1.4568   16.4379
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.2644372  0.0313205 104.227  < 2e-16 ***
## Ozone               -4.5921903  0.6258301  -7.338  2.4e-13 ***
## PM2.5               -0.0053669  0.0023483  -2.285   0.0223 *
## DieselPM             0.0031105  0.0003051  10.195  < 2e-16 ***
## LinguisticIsolation -0.0105184  0.0006889 -15.269  < 2e-16 ***
## Poverty              0.0122395  0.0004103  29.828  < 2e-16 ***
## Unemployment         0.0192106  0.0012864  14.934  < 2e-16 ***
## LowBirthWeight       0.0657500  0.0035665  18.435  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 10.80237)
##
##     Null deviance: 118332  on 7556  degrees of freedom
## Residual deviance:  74138  on 7549  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```
summary(mod.poisson.sim)$dispersion
```

```
## [1] 10.80237
```

```
summary(mod.poisson.full)$dispersion
```

```
## [1] 5.866488
```

```
anova(mod.poisson.full,mod.poisson.sim, test ='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: Asthma ~ Ozone + PM2.5 + DieselPM + DrinkingWater + Pesticides +
##     Tox.Release + Traffic + CleanupSites + GroundwaterThreats +
##     Haz.Waste + Imp.WaterBodies + PollutionBurden + LinguisticIsolation +
##     Poverty + Unemployment + HousingBurden + LowBirthWeight +
##     CardiovascularDisease
## Model 2: Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty +
##     Unemployment + LowBirthWeight
##   Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
## 1      7538      41860
## 2      7549      74138 -11   -32278 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Next, incoporating the quadratic terms of PM2.5, Diesel PM, Ozone into the model:

Hypothesis testing: comparing the linear simple model with the 8 variables, to the model that has the quadratic terms of the three environmental factors PM 2.5, Ozone and DieselPM.

Hypothesis H0: this new model is better in terms of predicting the Asthma compared to the simple model

```
mod.poisson.qua <- glm(Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
    Poverty + Unemployment+ LowBirthWeight  + I(Ozone^2) + I(PM2.5^2) + I(DieselPM ^2), data=dataint, fa
summary(mod.poisson.qua)
```

```
##
## Call:
## glm(formula = Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
##     Poverty + Unemployment + LowBirthWeight + I(Ozone^2) + I(PM2.5^2) +
##     I(DieselPM^2), family = quasipoisson, data = dataint)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13.3062   -2.2865   -0.5859    1.4387   16.5410
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.599e+00  1.049e-01  43.844  < 2e-16 ***
## Ozone               -5.469e+01  4.697e+00 -11.644  < 2e-16 ***
## PM2.5               -4.376e-02  1.081e-02  -4.050 5.17e-05 ***
## DieselPM             5.732e-03  6.796e-04   8.435  < 2e-16 ***
## LinguisticIsolation -9.644e-03  6.921e-04 -13.933  < 2e-16 ***
## Poverty              1.189e-02  4.093e-04  29.055  < 2e-16 ***
## Unemployment         1.866e-02  1.265e-03  14.755  < 2e-16 ***
## LowBirthWeight       6.398e-02  3.511e-03  18.224  < 2e-16 ***
## I(Ozone^2)           5.163e+02  4.750e+01  10.870  < 2e-16 ***
## I(PM2.5^2)           1.809e-03  4.728e-04   3.827 0.000131 ***
## I(DieselPM^2)       -2.707e-05  5.048e-06  -5.363 8.44e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 10.38138)
##
##     Null deviance: 118332  on 7556  degrees of freedom
## Residual deviance:  71707  on 7546  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```
anova(mod.poisson.sim,mod.poisson.qua, test ='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty +
##     Unemployment + LowBirthWeight
## Model 2: Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty +
##     Unemployment + LowBirthWeight + I(Ozone^2) + I(PM2.5^2) +
##     I(DieselPM^2)
```

```
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      7549      74138
## 2      7546      71707  3   2430.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is lower than zero, it's sufficient for us to reject the null hypothesis. Thus, the model with quadratic terms is better in predicting Asthma rate, as compared to the simple model.

Examining the effects of interactionterms between PM 2.5 and DieselPM:

```
mod.poisson.qua1 <- glm(Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
    Poverty + Unemployment+ LowBirthWeight  + I(Ozone^2) + I(PM2.5^2) + I(DieselPM ^2) + PM2.5*DieselPM
summary(mod.poisson.qua1)
```

```
##
## Call:
## glm(formula = Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
##     Poverty + Unemployment + LowBirthWeight + I(Ozone^2) + I(PM2.5^2) +
##     I(DieselPM^2) + PM2.5 * DieselPM, family = quasipoisson,
##     data = dataint)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -12.8222   -2.2952   -0.5766    1.4518   16.5638
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.450e+00  1.149e-01  38.722  < 2e-16 ***
## Ozone               -4.897e+01  5.033e+00  -9.729  < 2e-16 ***
## PM2.5               -5.064e-02  1.099e-02  -4.606 4.18e-06 ***
## DieselPM             1.088e-02  1.744e-03   6.237 4.70e-10 ***
## LinguisticIsolation -9.760e-03  6.923e-04 -14.099  < 2e-16 ***
## Poverty              1.200e-02  4.100e-04  29.254  < 2e-16 ***
## Unemployment         1.858e-02  1.264e-03  14.704  < 2e-16 ***
## LowBirthWeight       6.432e-02  3.508e-03  18.335  < 2e-16 ***
## I(Ozone^2)           4.628e+02  5.041e+01   9.180  < 2e-16 ***
## I(PM2.5^2)           2.422e-03  5.080e-04   4.768 1.89e-06 ***
## I(DieselPM^2)       -2.794e-05  5.116e-06  -5.462 4.87e-08 ***
## PM2.5:DieselPM      -4.686e-04  1.464e-04  -3.202  0.00137 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 10.35752)
##
##     Null deviance: 118332  on 7556  degrees of freedom
## Residual deviance:  71601  on 7545  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

The interaction term of PM2.5 and DieselPM does not have a p-value as significant as the other terms, butthe chisq test betweeb the model with the interaction vs not has a p-value under the significance leve,

which indicates that it's not sufficient to reject the null hypothesis that the interaction term is significant in terms of predicting the Asthma rate .

```
anova(mod.poisson.qua,mod.poisson.qua1, test ='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty +
##     Unemployment + LowBirthWeight + I(Ozone^2) + I(PM2.5^2) +
##     I(DieselPM^2)
## Model 2: Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty +
##     Unemployment + LowBirthWeight + I(Ozone^2) + I(PM2.5^2) +
##     I(DieselPM^2) + PM2.5 * DieselPM
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      7546      71707
## 2      7545      71601  1   106.14 0.001369 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check for overdispersion:

```
summary(mod.poisson.qua)$dispersion
```

```
## [1] 10.38138
```

```
summary(mod.poisson.qua1)$dispersion
```

```
## [1] 10.35752
```

```
summary(mod.poisson.sim)$dispersion
```

```
## [1] 10.80237
```

The quadratic model has smaller overdispersion quotient than the linear model, which indicates that it has better goodness of fit.

Next: to remedy for the overdispersion effects, we consider incorporating the negative binomial model:

```
nbin1 <- MASS::glm.nb(Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
    Poverty + Unemployment+ LowBirthWeight +  I(Ozone^2) + I(PM2.5^2) + I(DieselPM ^2) + PM2.5 * Diesel
summary(nbin1)
```

```
##
## Call:
## MASS::glm.nb(formula = Asthma ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
##     Poverty + Unemployment + LowBirthWeight + I(Ozone^2) + I(PM2.5^2) +
##     I(DieselPM^2) + PM2.5 * DieselPM, data = dataint, init.theta = 6.255171645,
##     link = log)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -3.8815  -0.8145  -0.1982   0.4649   4.4267
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          4.376e+00  1.123e-01  38.952  < 2e-16 ***
## Ozone               -4.491e+01  4.909e+00  -9.148  < 2e-16 ***
## PM2.5               -7.043e-02  1.109e-02  -6.349 2.17e-10 ***
## DieselPM             1.306e-02  1.816e-03   7.194 6.31e-13 ***
## LinguisticIsolation -1.027e-02  6.927e-04 -14.823  < 2e-16 ***
## Poverty              1.279e-02  4.062e-04  31.498  < 2e-16 ***
## Unemployment         2.048e-02  1.314e-03  15.589  < 2e-16 ***
## LowBirthWeight       6.239e-02  3.428e-03  18.199  < 2e-16 ***
## I(Ozone^2)           4.318e+02  4.957e+01   8.710  < 2e-16 ***
## I(PM2.5^2)           3.406e-03  5.227e-04   6.515 7.26e-11 ***
## I(DieselPM^2)       -2.924e-05  5.086e-06  -5.750 8.93e-09 ***
## PM2.5:DieselPM      -6.140e-04  1.574e-04  -3.900 9.61e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(6.2552) family taken to be 1)
##
##     Null deviance: 12760  on 7556  degrees of freedom
## Residual deviance:  7714  on 7545  degrees of freedom
## AIC: 66647
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  6.255
##           Std. Err.:  0.113
##
##  2 x log-likelihood:  -66620.539
```

```
summary(nbin1)$dispersion
```

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

The negative binomial model has dispersion quotient of 1, and all the terms are statistically significant with p-values much smaller than the 0.05 threshold. Thus, it's a desirable model at this step.

Multinomial Regression Testing:

```
#multinomial regression
data_cal <- data_cal %>% mutate(asthma_cat = case_when(
  AsthmaPctl <=25 ~ 1,
  AsthmaPctl <=50 ~ 2,
  AsthmaPctl <=75 ~ 3,
  AsthmaPctl <=100 ~ 4
))

summ.MNfit <- function(fit, digits=3){
  s <- summary(fit)
```

```
  for(i in 2:length(fit$lev)) {
##
cat("\nLevel", fit$lev[i], "vs. Level", fit$lev[1], "\n")
##
betaHat <- s$coefficients[(i-1),]
se <- s$standard.errors[(i-1),]
zStat <- betaHat / se
pval <- 2 * pnorm(abs(zStat), lower.tail=FALSE)
##
RRR <- exp(betaHat)
RRR.lower <- exp(betaHat - qnorm(0.975)*se)
RRR.upper <- exp(betaHat + qnorm(0.975)*se)
##
results <- cbind(betaHat, se, pval, RRR, RRR.lower, RRR.upper)
print(round(results, digits=digits))
}
}

mod.multi <- multinom(asthma_cat ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
    Poverty + Unemployment + LowBirthWeight, data=data_cal)
```

```
## # weights:  36 (24 variable)
## initial  value 10585.743742
## iter  10 value 9679.913746
## iter  20 value 9232.293865
## iter  30 value 8794.734380
## iter  40 value 8794.526687
## final  value 8794.524398
## converged
```

```
summ.MNfit(mod.multi)
```

```
##
## Level 2 vs. Level 1
##                      betaHat    se  pval         RRR RRR.lower    RRR.upper
## (Intercept)          -2.895 0.205 0.000       0.055     0.037 8.300000e-02
## Ozone                14.609 2.612 0.000 2211868.286 13235.962 3.696264e+08
## PM2.5                 0.010 0.016 0.543       1.010     0.978 1.043000e+00
## DieselPM              0.008 0.003 0.005       1.008     1.003 1.014000e+00
## LinguisticIsolation  -0.012 0.006 0.039       0.988     0.977 9.990000e-01
## Poverty               0.046 0.003 0.000       1.047     1.040 1.054000e+00
## Unemployment          0.069 0.011 0.000       1.072     1.048 1.096000e+00
## LowBirthWeight        0.081 0.025 0.001       1.085     1.033 1.139000e+00
##
## Level 3 vs. Level 1
##                      betaHat    se  pval    RRR RRR.lower RRR.upper
## (Intercept)          -4.418 0.219 0.000  0.012     0.008     0.019
## Ozone                 3.966 2.326 0.088 52.790     0.553  5041.086
## PM2.5                 0.029 0.017 0.086  1.029     0.996     1.063
## DieselPM              0.015 0.003 0.000  1.015     1.009     1.020
## LinguisticIsolation  -0.027 0.006 0.000  0.974     0.962     0.985
## Poverty               0.073 0.004 0.000  1.076     1.068     1.083
## Unemployment          0.105 0.012 0.000  1.111     1.086     1.136
```

```
## LowBirthWeight        0.194 0.026 0.000  1.214     1.153     1.279
##
## Level 4 vs. Level 1
##                     betaHat    se pval   RRR RRR.lower RRR.upper
## (Intercept)          -6.218 0.243    0 0.002     0.001     0.003
## Ozone               -23.287 2.323    0 0.000     0.000     0.000
## PM2.5                 0.063 0.018    0 1.065     1.028     1.103
## DieselPM              0.027 0.003    0 1.028     1.022     1.034
## LinguisticIsolation  -0.080 0.006    0 0.923     0.912     0.935
## Poverty               0.105 0.004    0 1.111     1.102     1.119
## Unemployment          0.163 0.012    0 1.177     1.149     1.205
## LowBirthWeight        0.377 0.029    0 1.458     1.378     1.543
```

```
mod.ord <- vglm(asthma_cat ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
    Poverty + Unemployment + LowBirthWeight, cumulative(parallel=TRUE, reverse=TRUE),data=data_cal)

summary(mod.ord)
```

```
##
## Call:
## vglm(formula = asthma_cat ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation +
##     Poverty + Unemployment + LowBirthWeight, family = cumulative(parallel = TRUE,
##     reverse = TRUE), data = data_cal)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept):1        -1.946341   0.134196 -14.504  < 2e-16 ***
## (Intercept):2        -3.444484   0.137676 -25.019  < 2e-16 ***
## (Intercept):3        -4.992315   0.144187 -34.624  < 2e-16 ***
## Ozone               -10.947596   2.676715  -4.090 4.31e-05 ***
## PM2.5                 0.016818   0.010333   1.628    0.104
## DieselPM              0.017787   0.001687  10.542  < 2e-16 ***
## LinguisticIsolation  -0.048035   0.003098 -15.504  < 2e-16 ***
## Poverty               0.057567   0.001885  30.540  < 2e-16 ***
## Unemployment          0.080727   0.006102  13.230  < 2e-16 ***
## LowBirthWeight        0.223158   0.015426  14.466  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y>=2]), logitlink(P[Y>=3]),
## logitlink(P[Y>=4])
##
## Residual deviance: 17797.28 on 22898 degrees of freedom
##
## Log-likelihood: -8898.638 on 22898 degrees of freedom
##
## Number of Fisher scoring iterations: 6
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##               Ozone               PM2.5            DieselPM LinguisticIsolation
##        1.760027e-05        1.016960e+00        1.017946e+00        9.531006e-01
```

```
##             Poverty        Unemployment      LowBirthWeight
##        1.059256e+00        1.084075e+00        1.250018e+00
```

```
mod.ord.npo <- vglm(asthma_cat ~ Ozone + PM2.5 + DieselPM + LinguisticIsolation + Poverty + Unemploymen
```

```
## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y = y, extra = extra): It seems
## that the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1
```

```
## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.
```

```
## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1

## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.

## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1
```

```
## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.
```

```
## Warning in Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals =
## residuals, : fitted values close to 0 or 1
```

```
## Warning in slot(family, "validparams")(eta, y, extra = extra): It seems that
## the nonparallelism assumption has resulted in intersecting linear/additive
## predictors. Try propodds() or fitting a partial nonproportional odds model or
## choosing some other link function, etc.
```

```
## Warning in vglm.fitter(x = x, y = y, w = w, offset = offset, Xm2 = Xm2, :
## iterations terminated because half-step sizes are very small
```

```
## Warning in vglm.fitter(x = x, y = y, w = w, offset = offset, Xm2 = Xm2, : some
## quantities such as z, residuals, SEs may be inaccurate due to convergence at a
## half-step
```

```
## Warning in log(prob): NaNs produced
```

```
pchisq(deviance(mod.ord)-deviance(mod.ord.npo), df=df.residual(mod.ord)-df.residual(mod.ord.npo),lower.
```

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

```
#proportional odds does not hold --> use multinomial
```