Predicting Asthma Rates in California

Pranjal Srivastava

5/8/2022

##Data Cleaning

```
cal_envir <- read_delim("/Users/pranjalsrivastava/Desktop/previous semester/linear regression/Project/C
```

```
## New names:
## Rows: 7989 Columns: 9
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (9): ...1, Census Tract, Pesticides, Tox. Release, Pollution Burden Scor...
## 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.
## * `` -> `...1`
```

```
head(cal_envir)
```

```
## # A tibble: 6 x 9
##    ...1 `Census Tract` Pesticides Tox. Rel~1 Pollu~2 PM2.5 Traffic  Ozone Asthma
##   <dbl>          <dbl>      <dbl>      <dbl>   <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1     1     6019001100      1.00       4859.    9.64  13.9   1037. 0.0603  130.
## 2     2     6077000700     63.1         520.    8.97  11.9    856. 0.0459  106.
## 3     3     6037204920      0          3683.    9.48  12.3   2523. 0.0479   76.1
## 4     4     6019000700     44.6        1630.    8.28  13.5    691. 0.0603  139.
## 5     5     6019000200     16.6        1975.    8.16  13.8    910. 0.0603  139.
## 6     6     6037542402      0.407     12492.    9.33  12.3   1481. 0.0426   86.8
## # ... with abbreviated variable names 1: `Tox. Release`,
## #   2: `Pollution Burden Score`
```

```
cal_envir1 <- cal_envir %>% dplyr::select("Pesticides", "Tox. Release", "Pollution Burden Score", "PM2.5
cal_envir1 <- rename(cal_envir1, c(Tox_Release = "Tox. Release", Pollution_Burden_Score = "Pollution Bu
summary(cal_envir1)
```

```
##    Pesticides       Tox_Release      Pollution_Burden_Score     PM2.5
##  Min.   :    0.0   Min.   :    0.0   Min.   : 1.213         Min.   : 1.875
##  1st Qu.:    0.0   1st Qu.:  112.4   1st Qu.: 4.074         1st Qu.: 8.578
##  Median :    0.0   Median :  457.4   Median : 5.192         Median :10.125
##  Mean   :  269.9   Mean   : 1615.7   Mean   : 5.217         Mean   :10.157
##  3rd Qu.:    0.2   3rd Qu.: 1630.3   3rd Qu.: 6.321         3rd Qu.:11.939
##  Max.   :80811.1   Max.   :96985.6   Max.   :10.000         Max.   :16.395
##     Traffic            Ozone            Asthma
```

```
##  Min.   :   20.75   Min.   :0.02655   Min.   :  4.28
##  1st Qu.:  553.59   1st Qu.:0.04193   1st Qu.: 30.06
##  Median :  880.41   Median :0.04716   Median : 45.77
##  Mean   : 1116.40   Mean   :0.04865   Mean   : 52.00
##  3rd Qu.: 1383.99   3rd Qu.:0.05680   3rd Qu.: 65.83
##  Max.   :45752.00   Max.   :0.07313   Max.   :243.29
```

```
cal_envir2 <- na.omit(cal_envir1)
summary(cal_envir2)
```

```
##    Pesticides       Tox_Release      Pollution_Burden_Score     PM2.5
##  Min.   :    0.0   Min.   :    0.0   Min.   : 1.213         Min.   : 1.875
##  1st Qu.:    0.0   1st Qu.:  112.4   1st Qu.: 4.074         1st Qu.: 8.578
##  Median :    0.0   Median :  457.4   Median : 5.192         Median :10.125
##  Mean   :  269.9   Mean   : 1615.7   Mean   : 5.217         Mean   :10.157
##  3rd Qu.:    0.2   3rd Qu.: 1630.3   3rd Qu.: 6.321         3rd Qu.:11.939
##  Max.   :80811.1   Max.   :96985.6   Max.   :10.000         Max.   :16.395
##     Traffic             Ozone             Asthma
##  Min.   :   20.75   Min.   :0.02655   Min.   :  4.28
##  1st Qu.:  553.59   1st Qu.:0.04193   1st Qu.: 30.06
##  Median :  880.41   Median :0.04716   Median : 45.77
##  Mean   : 1116.40   Mean   :0.04865   Mean   : 52.00
##  3rd Qu.: 1383.99   3rd Qu.:0.05680   3rd Qu.: 65.83
##  Max.   :45752.00   Max.   :0.07313   Max.   :243.29
```

```
write.csv(cal_envir2, "Clean_Cal_Environtmentdata.csv")
```

##Univariate Analysis & Standard Deviation

```
#Print observations
nrow(cal_envir2)
```

```
## [1] 7989
```

```
#Univariate
summary(cal_envir2)
```

```
##    Pesticides       Tox_Release      Pollution_Burden_Score     PM2.5
##  Min.   :    0.0   Min.   :    0.0   Min.   : 1.213         Min.   : 1.875
##  1st Qu.:    0.0   1st Qu.:  112.4   1st Qu.: 4.074         1st Qu.: 8.578
##  Median :    0.0   Median :  457.4   Median : 5.192         Median :10.125
##  Mean   :  269.9   Mean   : 1615.7   Mean   : 5.217         Mean   :10.157
##  3rd Qu.:    0.2   3rd Qu.: 1630.3   3rd Qu.: 6.321         3rd Qu.:11.939
##  Max.   :80811.1   Max.   :96985.6   Max.   :10.000         Max.   :16.395
##     Traffic             Ozone             Asthma
##  Min.   :   20.75   Min.   :0.02655   Min.   :  4.28
##  1st Qu.:  553.59   1st Qu.:0.04193   1st Qu.: 30.06
##  Median :  880.41   Median :0.04716   Median : 45.77
##  Mean   : 1116.40   Mean   :0.04865   Mean   : 52.00
##  3rd Qu.: 1383.99   3rd Qu.:0.05680   3rd Qu.: 65.83
##  Max.   :45752.00   Max.   :0.07313   Max.   :243.29
```

```
sd(cal_envir2$Pesticides)
```

```
## [1] 2331.122
```

```
sd(cal_envir2$Tox_Release)
```

```
## [1] 3795.546
```

```
sd(cal_envir2$Pollution_Burden_Score)
```

```
## [1] 1.552934
```

```
sd(cal_envir2$PM2.5)
```

```
## [1] 2.165989
```

```
sd(cal_envir2$Traffic)
```

```
## [1] 988.4462
```

```
sd(cal_envir2$Ozone)
```

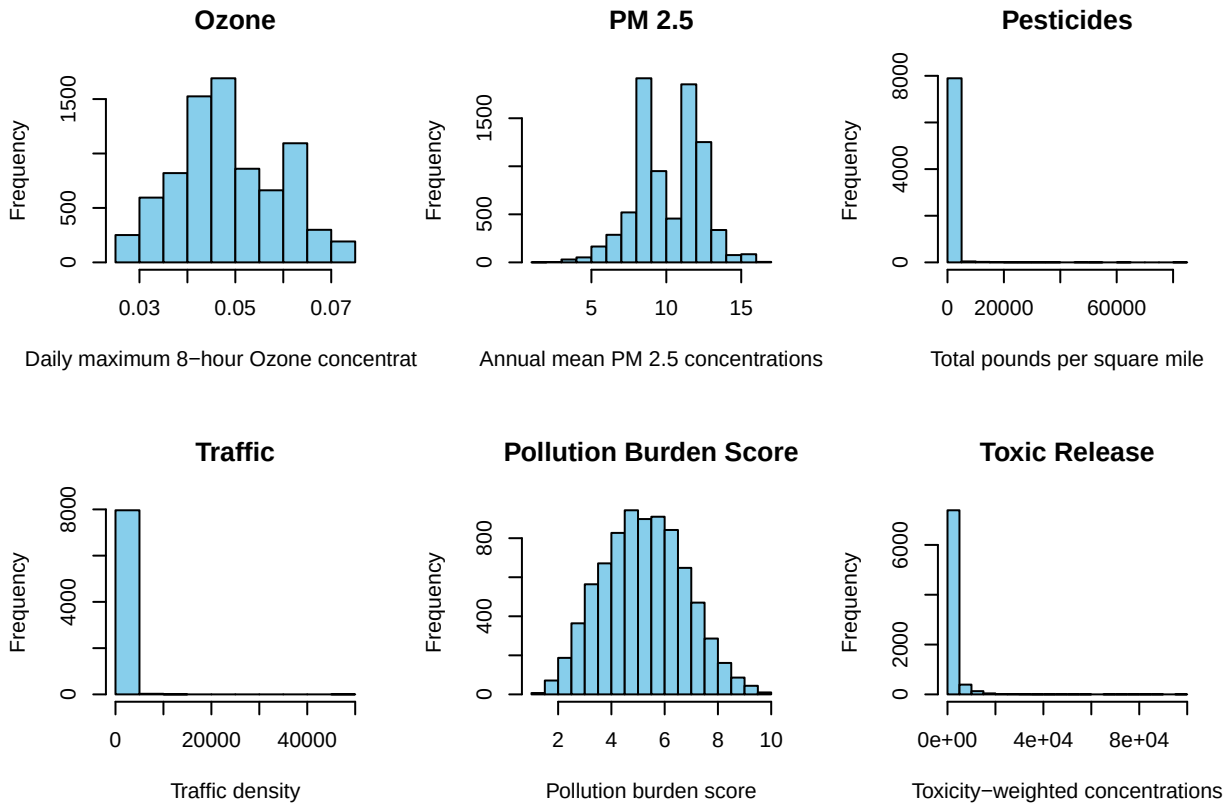
```
## [1] 0.01046916
```

```
sd(cal_envir2$Asthma)
```

```
## [1] 30.55984
```

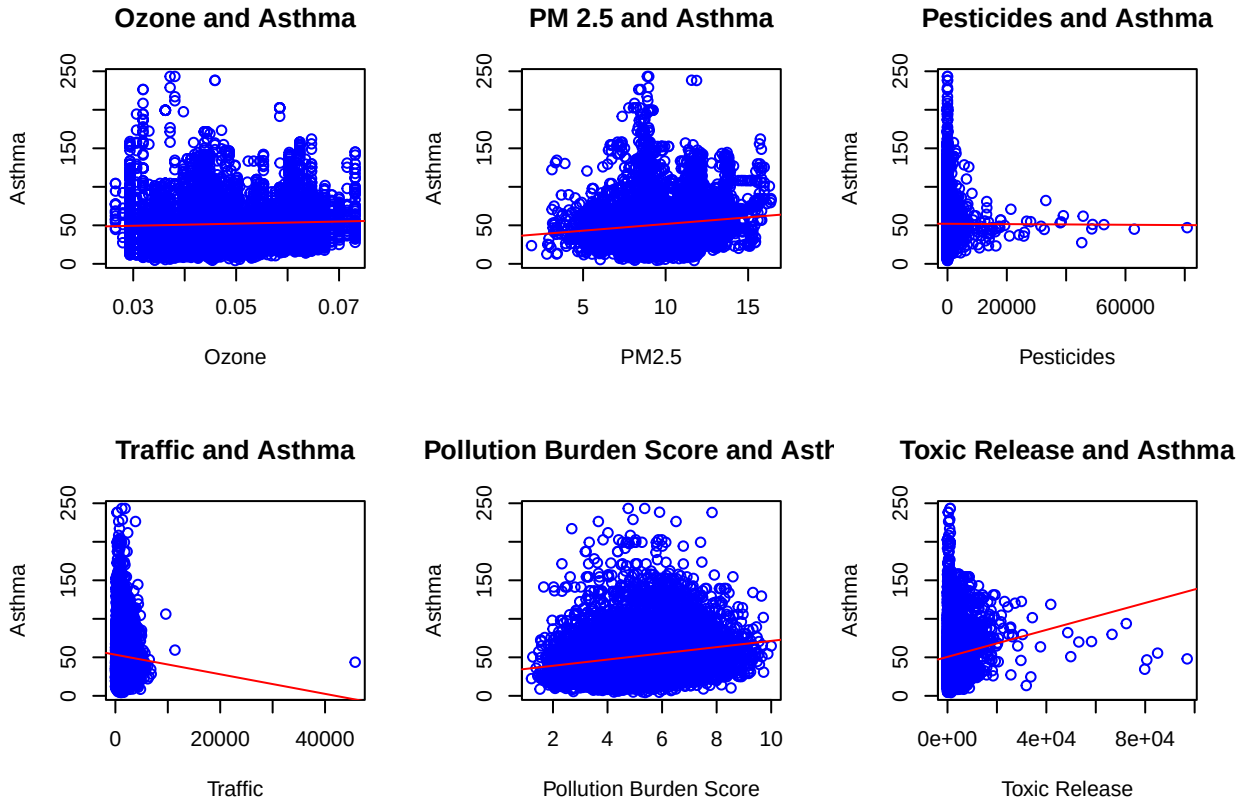
##Exploratory Plots

```
#Histograms
par(mfrow = c(2,3))
hist(cal_envir2$Ozone, main = "Ozone", xlab = "Daily maximum 8-hour Ozone concentration" , col = "skyblu
hist(cal_envir2$PM2.5, main = "PM 2.5" , xlab = "Annual mean PM 2.5 concentrations" , col = "skyblue")
hist(cal_envir2$Pesticides, main = "Pesticides" , xlab = "Total pounds per square mile" , col = "skyblue
hist(cal_envir2$Traffic, main = "Traffic" , xlab = "Traffic density" , col = "skyblue")
hist(cal_envir2$Pollution_Burden_Score, main = "Pollution Burden Score", xlab = "Pollution burden score
hist(cal_envir2$'Tox_Release', main = "Toxic Release" , xlab = "Toxicity-weighted concentrations", col =
```




```r
#Scatter plots
y <- cal_envir2$Asthma
par(mfrow= c(2,3))
plot(cal_envir2$Ozone,y,
     xlab = "Ozone", ylab = "Asthma",
     main = "Ozone and Asthma", col = "blue")
abline(lm(y ~ cal_envir2$Ozone, data = cal_envir2), col = "red")

plot(cal_envir2$PM2.5,y,
     xlab = "PM2.5", ylab = "Asthma",
     main = "PM 2.5 and Asthma", col = "blue")
abline(lm(y ~ cal_envir2$PM2.5, data = cal_envir2), col = "red")

plot(cal_envir2$Pesticides,y,
     xlab = "Pesticides", ylab = "Asthma",
     main = "Pesticides and Asthma", col = "blue")
abline(lm(y ~ cal_envir2$Pesticides, data = cal_envir2), col = "red")

plot(cal_envir2$Traffic,y,
     xlab = "Traffic", ylab = "Asthma",
     main = "Traffic and Asthma", col = "blue")
abline(lm(y ~ cal_envir2$Traffic, data = cal_envir2), col = "red")

plot(cal_envir2$Pollution_Burden_Score, y,
     xlab = "Pollution Burden Score", ylab = "Asthma",
     main = "Pollution Burden Score and Asthma", col = "blue")
```

```
abline(lm(y ~ cal_envir2$Pollution_Burden_Score, data = cal_envir2), col = "red")

plot(cal_envir2$Tox_Release, y,
     xlab = "Toxic Release", ylab = "Asthma",
     main = "Toxic Release and Asthma", col = "blue")
abline(lm(y ~ cal_envir2$Tox_Release, data = cal_envir2), col = "red")
```



##Fitting a Model

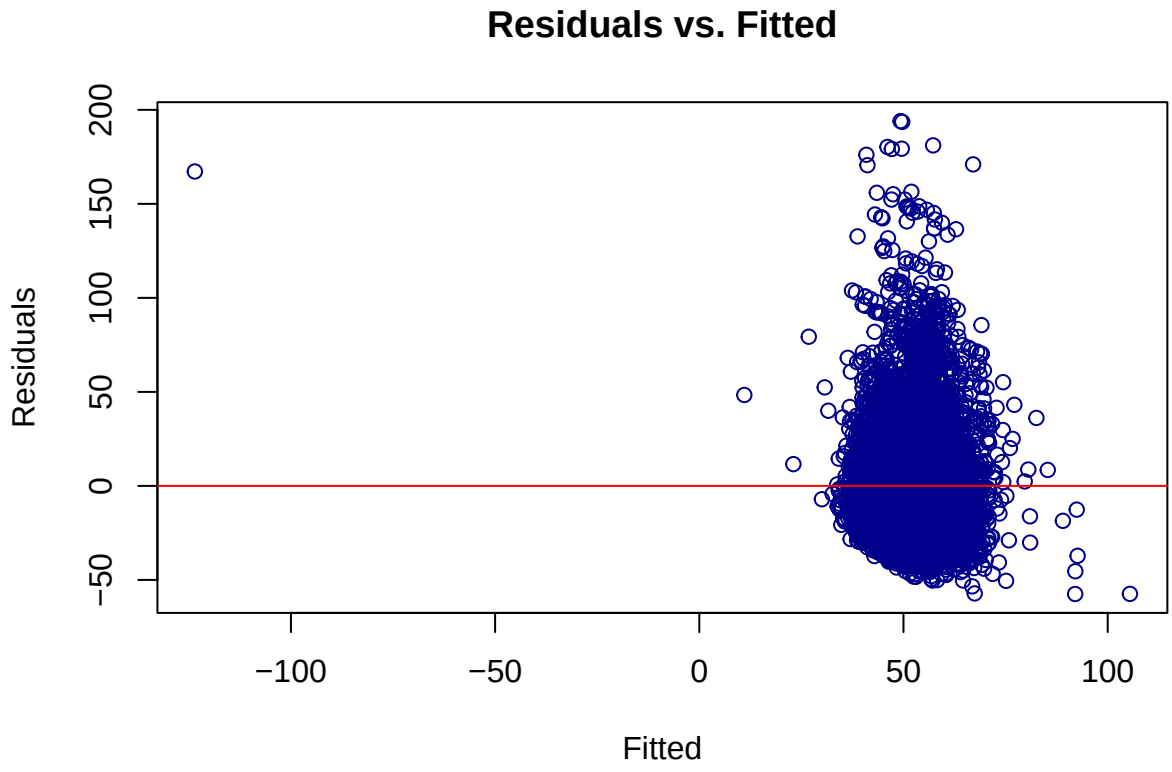
```
#fitting a model

model1 <- lm(Asthma ~ Ozone + PM2.5 + Pesticides + Traffic + Pollution_Burden_Score + Tox_Release, data
summary(model1)
```

```
##
## Call:
## lm(formula = Asthma ~ Ozone + PM2.5 + Pesticides + Traffic +
##     Pollution_Burden_Score + Tox_Release, data = cal_envir2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.482 -20.560  -6.274  12.968 194.008
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)             3.228e+01  1.906e+00  16.938  < 2e-16 ***
## Ozone                   2.726e+01  3.620e+01   0.753    0.452
## PM2.5                  -3.069e-01  2.158e-01  -1.422    0.155
## Pesticides             -2.338e-04  1.442e-04  -1.621    0.105
## Traffic                -4.007e-03  3.619e-04 -11.072  < 2e-16 ***
## Pollution_Burden_Score  4.872e+00  2.901e-01  16.797  < 2e-16 ***
## Tox_Release             3.919e-04  9.498e-05   4.126 3.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.65 on 7982 degrees of freedom
## Multiple R-squared:  0.05921,    Adjusted R-squared:  0.0585
## F-statistic: 83.73 on 6 and 7982 DF,  p-value: < 2.2e-16
```

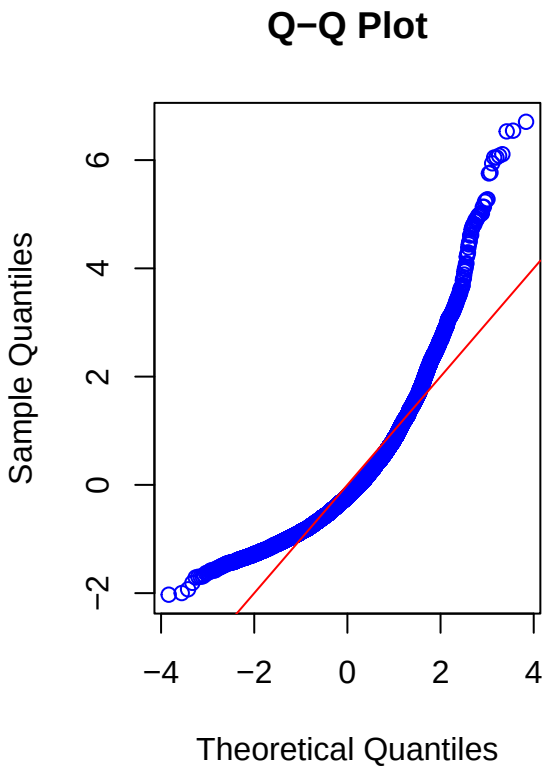
##Diagnostics

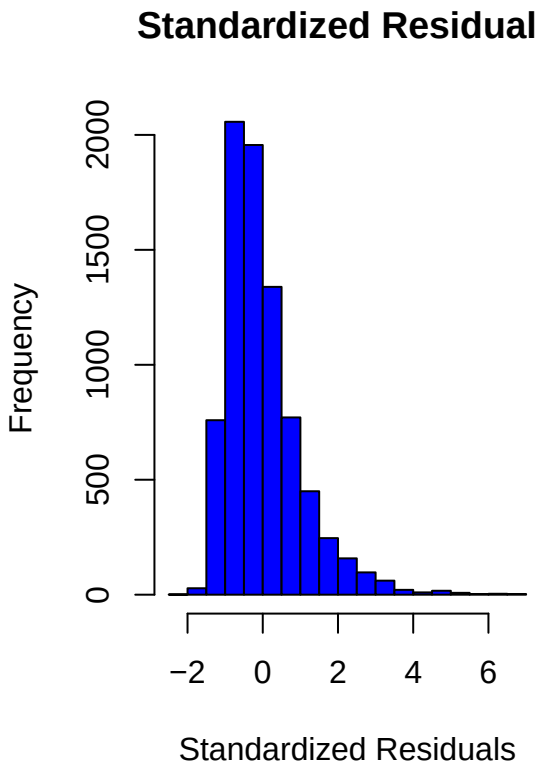
```
#Residuals vs. fitted plots
par (mfrow = c (1,1))
plot(fitted(model1), residuals(model1), main = "Residuals vs. Fitted" , xlab = "Fitted", ylab = "Residua
abline (h=0, col = "red")
```




```
#Quantile-Quantile plots with standardized residuals
par (mfrow = c(1,2))
qqnorm(rstandard(model1), main = "Q-Q Plot" , col = "blue")
abline(0,1, col = "red")
```

```
hist (rstandard(model1), main = "Standardized Residuals", xlab =
"Standardized Residuals" , col = "blue")
```






```
#Normality Shapiro-Wilk Test
#shapiro.test(residuals(model1)) # data too large for Shapiro so using Anderson-Darling normality test

# Anderson-Darling normality test
ad.test(residuals(model1))
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(model1)
## A = 201.15, p-value < 2.2e-16
```

```
# Interpretation for AD test: reject the hypothesis of normality when the p-value is less than or equal

# Outliers
range(rstudent(model1))
```

```
## [1] -2.028829  6.726322
```

```r
p <- 2 # one predictor + an intercept
n <- nrow (cal_envir2)
qt(1-.05 /(n*2),n-p-1)
```

```
## [1] 4.520488
```

```r
#correlation
cal_envir_modified <- cal_envir2[,-1]
cal_envir_corrmatrix <- cor(cal_envir2,cal_envir_modified)
cal_envir_corrmatrix <- round(cal_envir_corrmatrix,2)
cal_envir_corrmatrix
```

```
##                       Tox_Release Pollution_Burden_Score PM2.5 Traffic Ozone
## Pesticides                  -0.04                   0.04 -0.08   -0.04 -0.04
## Tox_Release                  1.00                   0.32  0.28    0.10 -0.07
## Pollution_Burden_Score       0.32                   1.00  0.61    0.35  0.18
## PM2.5                        0.28                   0.61  1.00    0.19  0.43
## Traffic                      0.10                   0.35  0.19    1.00 -0.03
## Ozone                       -0.07                   0.18  0.43   -0.03  1.00
## Asthma                       0.11                   0.21  0.12   -0.04  0.05
##                       Asthma
## Pesticides              0.00
## Tox_Release             0.11
## Pollution_Burden_Score  0.21
## PM2.5                   0.12
## Traffic                -0.04
## Ozone                   0.05
## Asthma                  1.00
```

```r
#eigenvalues
cal_envir_eigen <- model.matrix(model1)[,-1]
e <- eigen(t(cal_envir_eigen) %*% cal_envir_eigen)
examen_e <- sqrt(e$val[1] / e$val)
sum(examen_e > 30)
```

```
## [1] 3
```

```r
#variance inflation factors
require(MASS)
```

```
## Loading required package: MASS
```

```
## Warning: package 'MASS' was built under R version 4.1.2
```

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

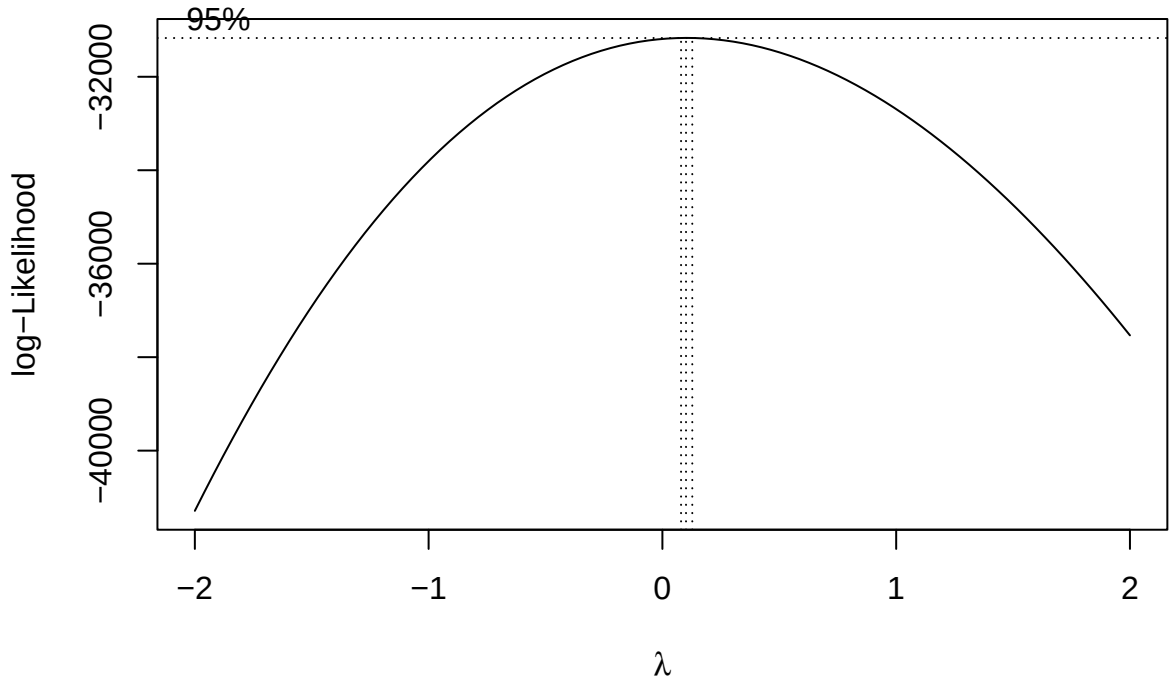
```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
require(faraway)
cal_envir_vif <- vif(model1)
cal_envir_vif
```

```
##                  Ozone                  PM2.5             Pesticides
##               1.305141               1.984789               1.026330
##                Traffic Pollution_Burden_Score            Tox_Release
##               1.162388               1.843443               1.180576
```

##Modified model Box-Cox Method

```
#boxcox
a <- boxcox(model1, plotit = TRUE)
```




```
#finding lambda
a$x[which.max(a$y)]
```

```
## [1] 0.1010101
```

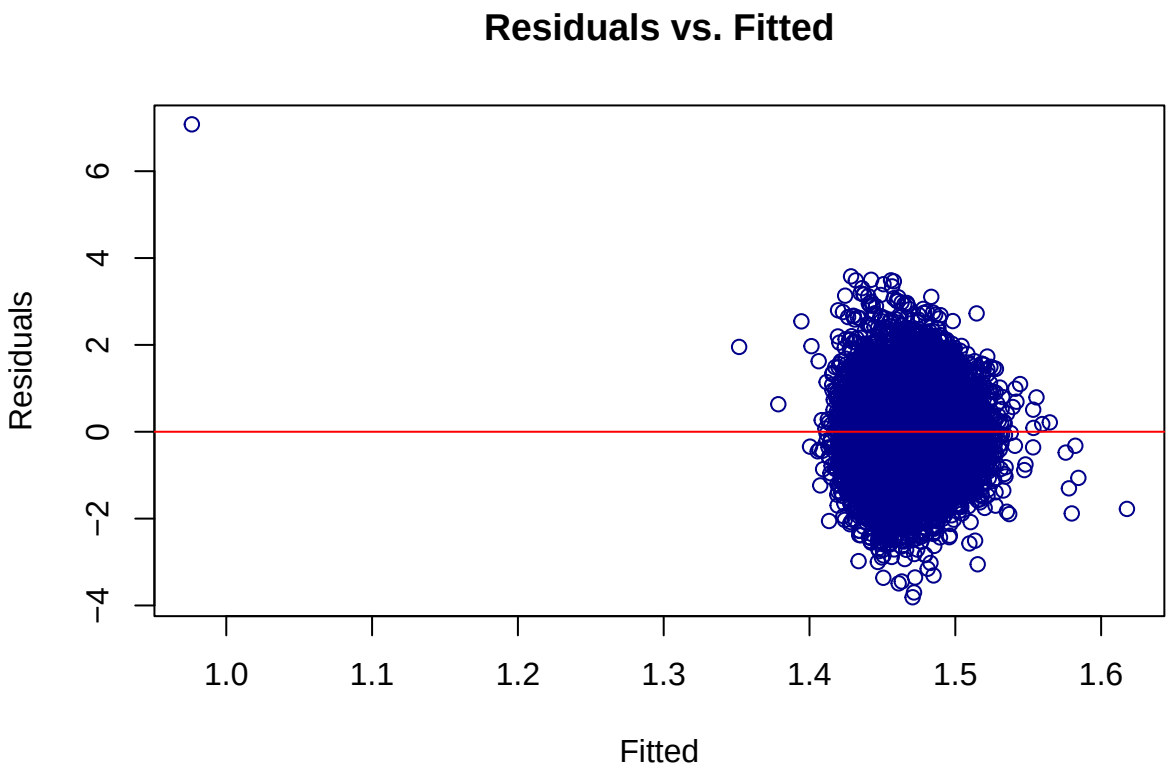
##Making a new model (Model 2)

```
#Model2 With y Transformed
model2 <- lm(Asthmaˆ(10/99) ~ Ozone + PM2.5 + Pesticides + Traffic + Pollution_Burden_Score + Tox_Releas
summary(model2)
```

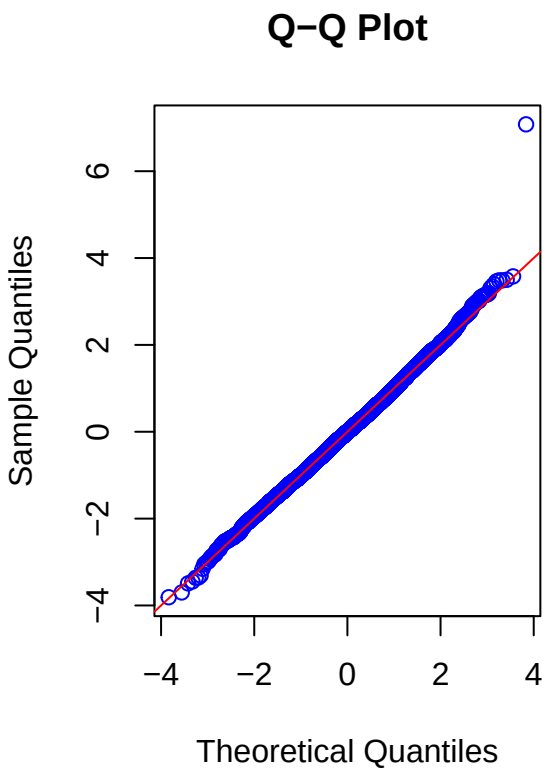
```
##
## Call:
## lm(formula = Asthma^(10/99) ~ Ozone + PM2.5 + Pesticides + Traffic +
##     Pollution_Burden_Score + Tox_Release, data = cal_envir2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31252 -0.05585 -0.00034  0.05303  0.48817
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.384e+00  5.272e-03 262.509  < 2e-16 ***
## Ozone                   4.997e-01  1.001e-01   4.990 6.17e-07 ***
## PM2.5                  -9.612e-04  5.969e-04  -1.610    0.107
## Pesticides             -4.919e-09  3.988e-07  -0.012    0.990
## Traffic                -1.122e-05  1.001e-06 -11.205  < 2e-16 ***
## Pollution_Burden_Score  1.557e-02  8.024e-04  19.410  < 2e-16 ***
## Tox_Release             1.061e-06  2.627e-07   4.039 5.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08202 on 7982 degrees of freedom
## Multiple R-squared:  0.08206,    Adjusted R-squared:  0.08137
## F-statistic: 118.9 on 6 and 7982 DF,  p-value: < 2.2e-16
```

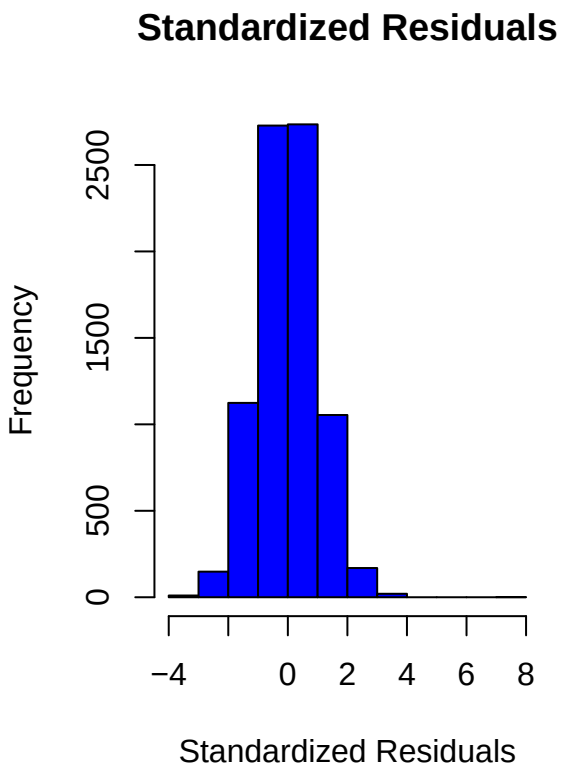
##Model 2 Diagnostics

```
#Residuals vs Fitted Plot
plot(fitted(model2), rstandard(model2),
     xlab = "Fitted",
     ylab = "Residuals", col = "darkblue" , main = "Residuals vs. Fitted")
abline(h=0, col = "red")
```



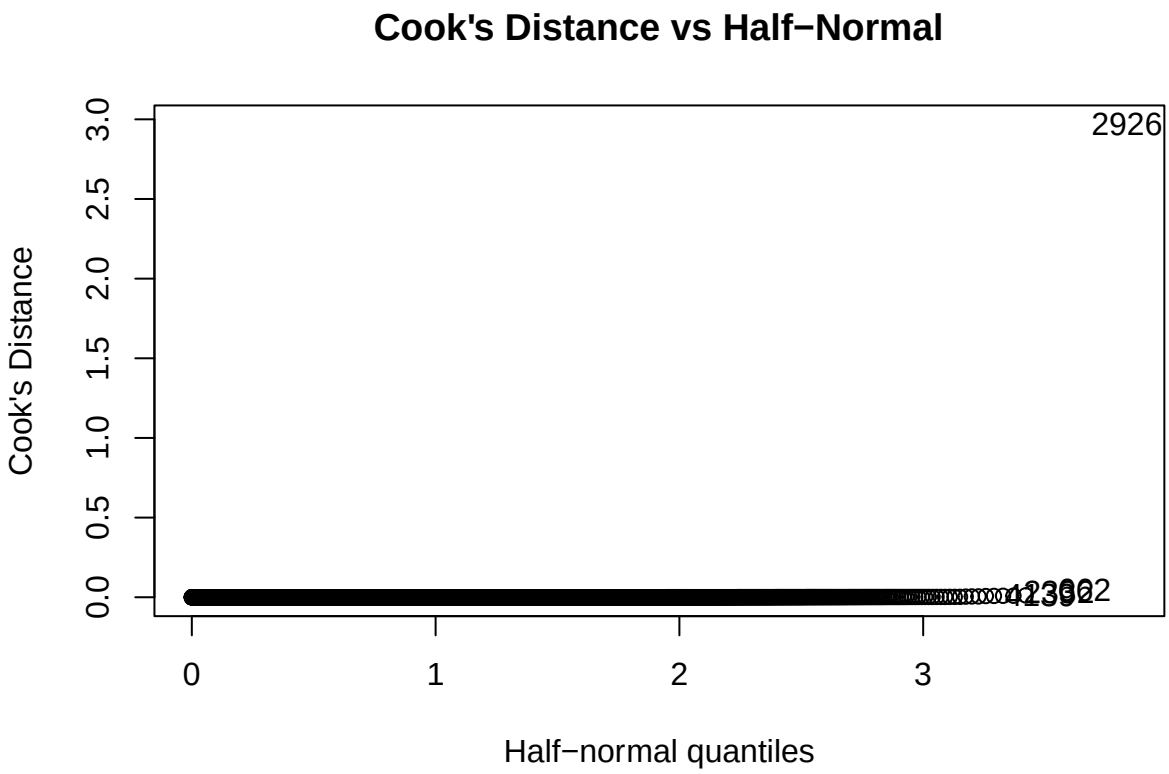

```
# Quantile-Quantile plots with standardized residuals
par (mfrow = c (1,2))
qqnorm(rstandard(model2),main = "Q-Q Plot", col = "blue")
abline(0,1, col = "red")
hist(rstandard(model2), main = "Standardized Residuals", xlab = "Standardized Residuals" , col = "blue"
```





##Cooks Distance to Remove Outliers

```
#Cooks distance
cal_envir2$cooked <- cooks.distance(model2)
sample_size <- nrow(cal_envir2)
cal_envir2$outlier <- ifelse((cal_envir2$cooked < 4/sample_size),"keep","delete")
cal_new <- cal_envir2[!(cal_envir2$outlier=="delete"),]

rname <- row.names(cal_envir2)


halfnorm(cal_envir2$cooked,4,labs = rname, ylab = "Cook's Distance" , main = "Cook's Distance vs Half-N
```



##Making a New Model (Model 3) with Outliers Removed

```
#Model 3 with new data, outliers removed
model3 <- lm(Asthma^(10/99) ~ Ozone + PM2.5 + Pesticides + Traffic + Pollution_Burden_Score + `Tox_Relea
summary(model3)
```

```
##
## Call:
## lm(formula = Asthma^(10/99) ~ Ozone + PM2.5 + Pesticides + Traffic +
##     Pollution_Burden_Score + Tox_Release, data = cal_new)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.310475 -0.053004  0.000806  0.051883  0.290543
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.376e+00  5.062e-03 271.932  < 2e-16 ***
## Ozone                    5.040e-01  9.816e-02   5.134  2.9e-07 ***
## PM2.5                   -6.992e-04  5.928e-04  -1.179 0.238255
## Pesticides              -3.469e-09  5.297e-07  -0.007 0.994775
## Traffic                 -1.667e-05  1.175e-06 -14.190  < 2e-16 ***
## Pollution_Burden_Score  1.705e-02  7.926e-04  21.510  < 2e-16 ***
## Tox_Release              1.210e-06  3.475e-07   3.482 0.000501 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

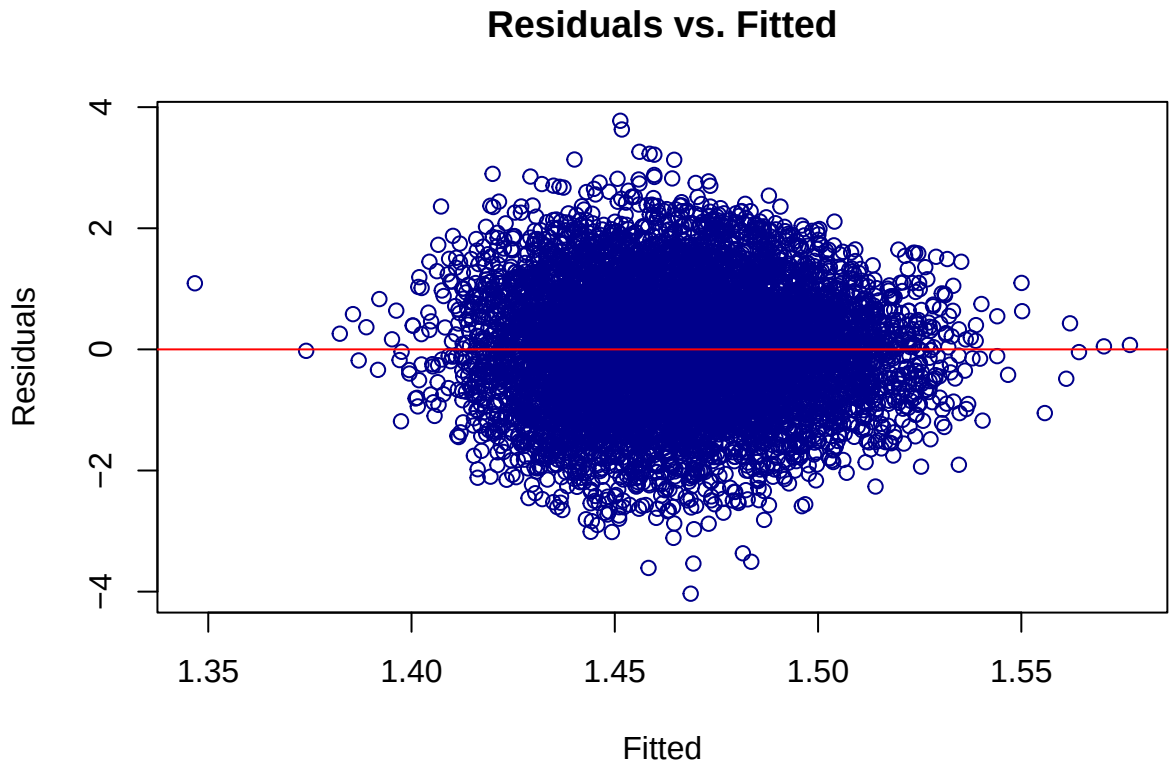
```
##
## Residual standard error: 0.07698 on 7750 degrees of freedom
## Multiple R-squared:  0.1037, Adjusted R-squared:  0.103
## F-statistic: 149.5 on 6 and 7750 DF,  p-value: < 2.2e-16
```

```
nrow(cal_new)
```

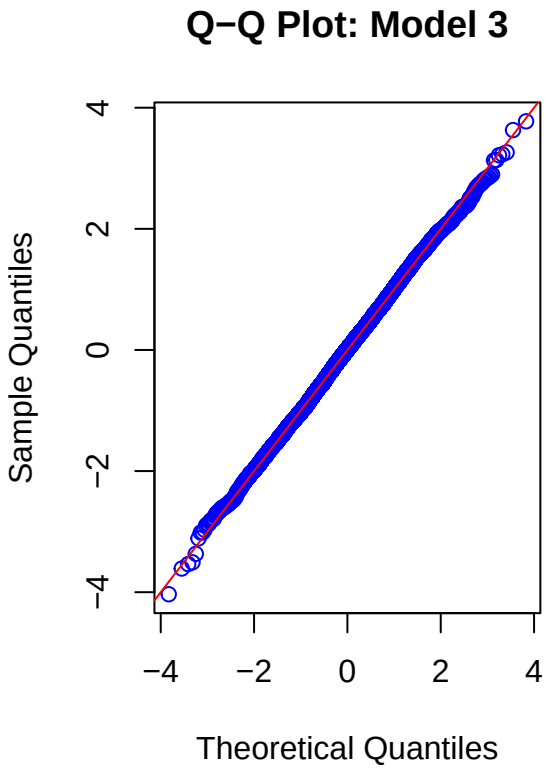
```
## [1] 7757
```

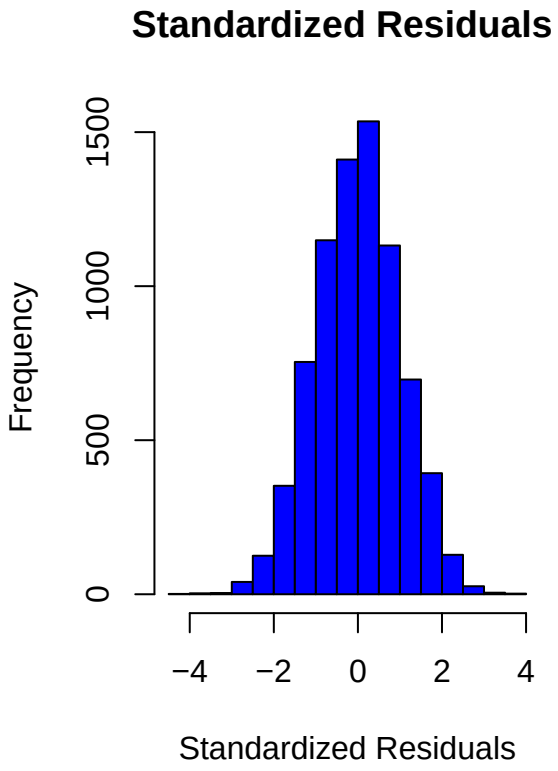
##Diagnostics for Model 3

```
#Residuals vs Fitted Plot
plot(fitted(model3), rstandard(model3),
     xlab = "Fitted",
     ylab = "Residuals", col = "darkblue" , main = "Residuals vs. Fitted")
abline(h=0, col = "red")
```



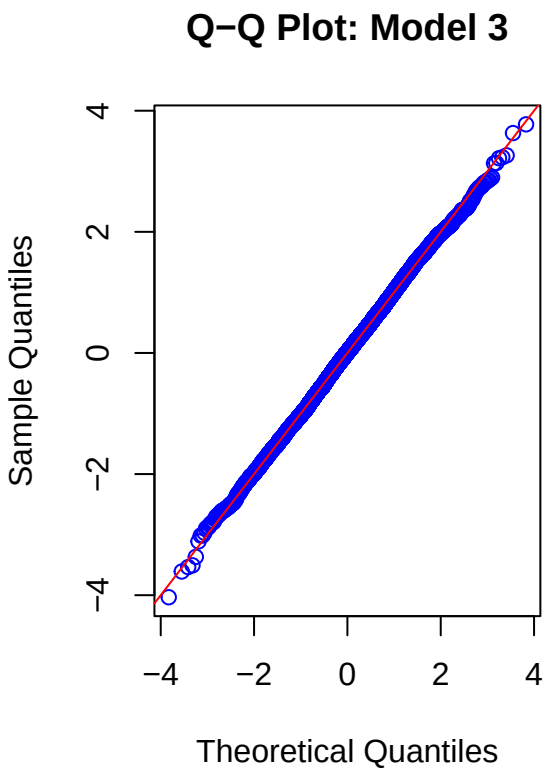

```
# Quantile-Quantile plots with standardized residuals
par (mfrow = c (1,2))
qqnorm(rstandard(model3),main = "Q-Q Plot: Model 3", col = "blue")
abline(0,1, col = "red")
hist(rstandard(model3), main = "Standardized Residuals ", xlab = "Standardized Residuals" , col = "blue
```

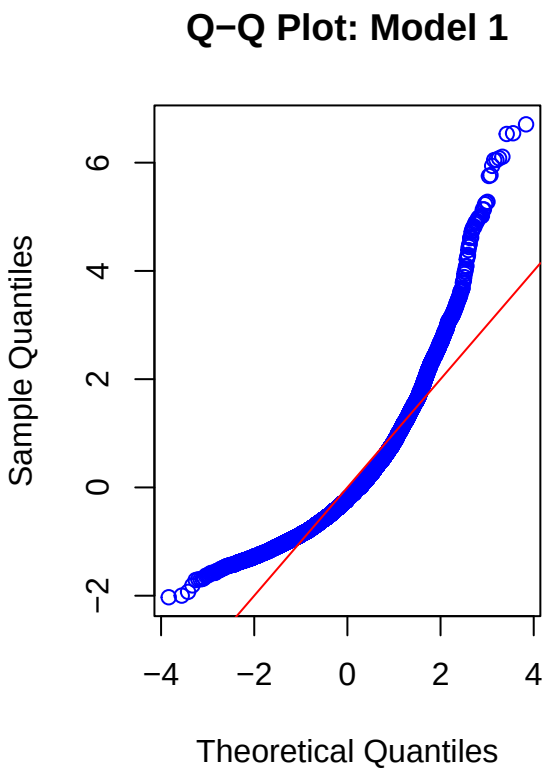




##Comparing Model1 and Model3

```
#QQplot M1 vs M3
par (mfrow = c (1,2))
qqnorm(rstandard(model3),main = "Q-Q Plot: Model 3", col = "blue")
abline(0,1, col = "red")
qqnorm(rstandard(model1),main = "Q-Q Plot: Model 1", col = "blue")
abline(0,1, col = "red")
```





##Model Selection Based on AIC

```
#AIC

b <- regsubsets(Asthma^(10/99) ~ Ozone + PM2.5 + Pesticides + Traffic + Pollution_Burden_Score + Tox_Rel
rs <- summary(b)
rs$which
```

```
##   (Intercept) Ozone PM2.5 Pesticides Traffic Pollution_Burden_Score Tox_Release
## 1        TRUE FALSE FALSE      FALSE   FALSE                    TRUE       FALSE
## 2        TRUE FALSE FALSE      FALSE    TRUE                    TRUE       FALSE
## 3        TRUE  TRUE FALSE      FALSE    TRUE                    TRUE       FALSE
## 4        TRUE  TRUE FALSE      FALSE    TRUE                    TRUE        TRUE
## 5        TRUE  TRUE  TRUE      FALSE    TRUE                    TRUE        TRUE
## 6        TRUE  TRUE  TRUE       TRUE    TRUE                    TRUE        TRUE
```

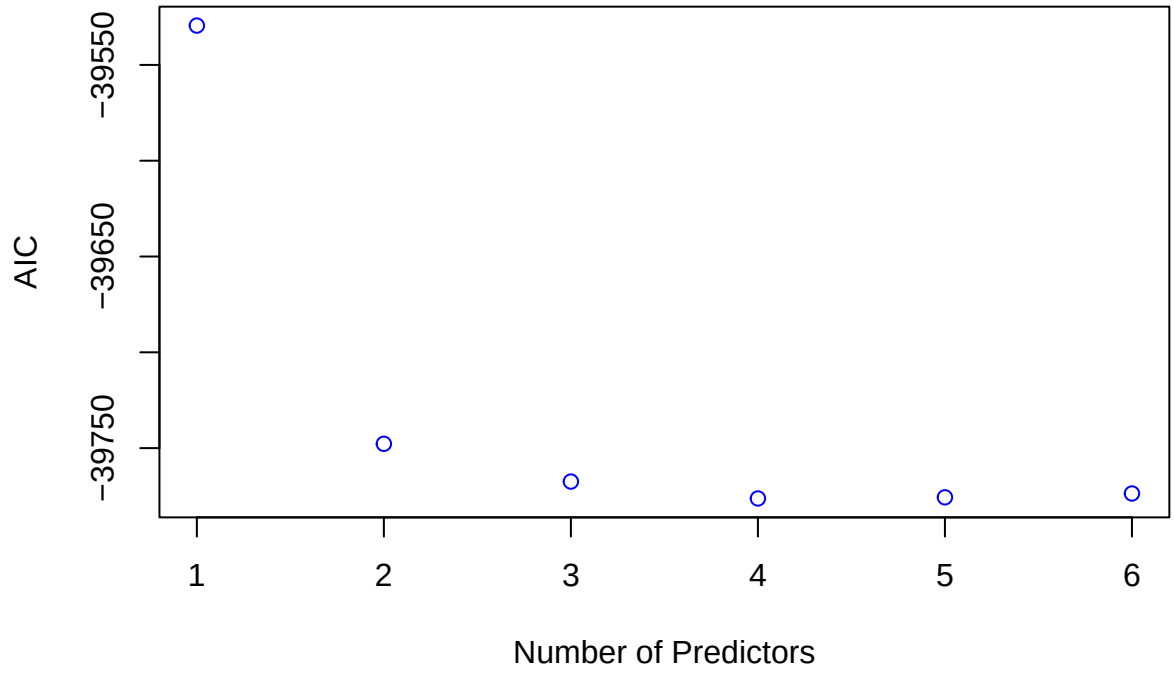
```
rs$rss
```

```
## [1] 47.45882 46.12999 46.00108 45.93700 45.92868 45.92868
```

```
n <- nrow(cal_new)
p <- 2:7
AIC <- n*log(rs$rss / n) + 2 * p
AIC
```

```
## [1] -39529.46 -39747.75 -39767.46 -39776.27 -39775.68 -39773.68
```

```
plot(AIC ~ I(p - 1), ylab = "AIC", xlab = "Number of Predictors" , col = "blue")
```




```
#Best model is model 4, with Ozone, Traffic, Pollution Burden Score, and Tox. Release as predictors
```

##Final Model (Model 4), removed predictors

```
#Final Model
model4 <- lm(Asthma^(10/99) ~ Ozone + Traffic + Pollution_Burden_Score + Tox_Release, data=cal_new)
summary(model4)
```

```
##
## Call:
## lm(formula = Asthma^(10/99) ~ Ozone + Traffic + Pollution_Burden_Score +
##     Tox_Release, data = cal_new)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.311433 -0.052913  0.000764  0.051872  0.290632
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.374e+00  4.744e-03 289.686  < 2e-16 ***
```

```
## Ozone                   4.515e-01  8.760e-02   5.155 2.61e-07 ***
## Traffic                -1.666e-05  1.169e-06 -14.252  < 2e-16 ***
## Pollution_Burden_Score  1.660e-02  6.850e-04  24.230  < 2e-16 ***
## Tox_Release             1.098e-06  3.337e-07   3.289  0.00101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07698 on 7752 degrees of freedom
## Multiple R-squared:  0.1036, Adjusted R-squared:  0.1031
## F-statistic: 223.9 on 4 and 7752 DF,  p-value: < 2.2e-16
```

```
confint(model4)
```

```
##                              2.5 %        97.5 %
## (Intercept)             1.365036e+00  1.383636e+00
## Ozone                   2.798243e-01  6.232721e-01
## Traffic                -1.895573e-05 -1.437178e-05
## Pollution_Burden_Score  1.525516e-02  1.794077e-02
## Tox_Release             4.433054e-07  1.751696e-06
```

```
model4$coefficients^(99/10)
```

```
##            (Intercept)                  Ozone                Traffic
##           2.328741e+01           3.815668e-04                    NaN
## Pollution_Burden_Score            Tox_Release
##           2.390800e-18           1.000015e-59
```