333 Final Report

Ruohe Zhou & Zening Duan
rzhou73@wisc.edu
zduan26@wisc.edu

Introduction

Scientists, policy makers, and the publics have a long-standing criticism that pollution may lead to higher mortality in the U.S. This report aims to uncover the statistical relationships between morality rate and a set of explanatory variables including pollution indicators, climate and socioeconomic variables, via selecting and applying the best linear regression models on a second-hand dataset.

Based on the dataset, we then tested the possible statistical relationship among variables of our interest and compared the performance of different multivariable linear regression models. The response variable is *mortality* and the potential explanatory variables are the two pollution variables and a set of climate and socioeconomic variables.

The dataset has following ingredients, see Table 1.

Table 1: Dataset description

Variables	Description	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
City	60 cities surveyed	—	—	—	—	—	—
Mort	Total adjusted mortality from all cases, in deaths per 100,000 population	790.7	898.4	943.7	940.4	983.2	1113.1
Precip	Mean annual precipitation (in inches)	10.00	32.75	38.00	37.37	43.25	60.00
Educ*	Median number of school years completed for person of age 25 years or older	9.00	10.40	11.05	10.97	11.50	12.30
NonWhite	Percentage of 1960 population that is nonwhite	0.80	4.95	10.40	11.87	15.65	38.50
NOX**	Relative pollution potential of oxides of nitrogen	1.00	4.00	9.00	22.65	23.75	319.00
SO2**	Relative pollution potential of sulfur dioxide, SO2	1.00	11.00	30.00	53.77	69.00	278.00

* Notes: The education statistics of two cities (Lancaster and York) do not reflect the actual situation given the existing information in the material. To avoid introducing bias into our final results, these two cities along with their data have been removed accordingly.
** NOX and SO2 are the two pollution variables of our interest.

Method

Step 1: Drew a matrix plot on data *pollution* to get an overview of the linear relationships between arbitrary pair of variables. Given the plot below, we found that: (1) Explanatory variables like *Precip*, *Education*, *NonWhite*, and *SO2* have a linear relationship with the response variable *Mortality*. (2) According to the matrix plot (Figure 1) and summary table (Table 1), explanatory variables *NOX* and *SO2* have very large maximum values but small mean, suggesting that the distributions of the two variables are skewed and have a nonlinear relationship with mortality.

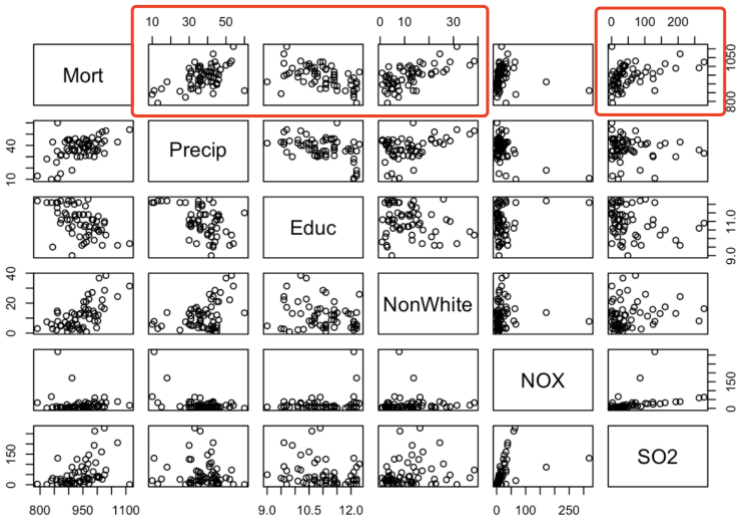


Figure 1 Matrix plot of the original dataset

Step 2: In order to transforming a highly skewed variable into a more normalized dataset (see Figure 2), we applied logarithmic transformation on two explanatory variables, *NOX* and *SO2*. Then we found a basic model:

$$Mort \sim Precip + Educ + NonWhite + NOX_{log} + SO2_{log}$$

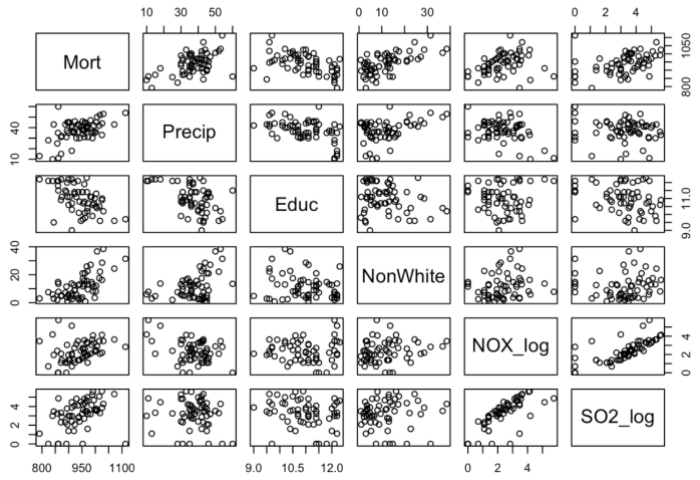


Figure 2 Matrix plot of the basic model

Step 3: Collinearity detection. We checked the variance inflation Factor (VIF), Eigenvalue, and condition Index of the basic model. Statistics of Eigenvalue and condition index suggested two groups of collinearity problem as the two groups have condition index larger than 15 while their eigenvalues are very close to 0. Although the VIF statistics of the model did not provide additional evidence since the collinearity problem in this case is not very severe.

Step 4: We used the Principal Component Regression (PCR) approach to find a better model to take care of the collinearity problem. We decided to keep two principal components (PCs) because under such condition, the model has the largest adjusted R-square value (adj R^2) and the lowest mean square prediction error (MSEP). Another reason is that the two indictors did not change significantly beyond 2 PCs, so it is unnecessary to choose more than 2 PCs (see Appendix A). We found our first model by using PCR:

$$Mort = 1078.97 + 17.71 * Precip - 21.62 * Educ + 19.12 * NonWhite + 9.57 * NOX_{log} + 14.76 * SO2_{log}$$

Step 5: We used three variable selection procedures, Forward Selection (FS), Backward Elimination (BE), and Stepwise Method (SM), to select other potential models. We found that the BE and SM procedure recommended a same model which has a relative lower Akaike Information Criterion score (AIC) comparing with the one suggested by FS; then we have our potential model:

$$Mort \sim Precip + Educ + NonWhite + SO2_{log}$$

Step 6: We used "myleaps" function to create a table including all possible models along with their performances under different model evaluation criteria, for instance, two of them commonly adopted are adj R^2 and Bayesian Information Criterion score (BIC). We have our second model:

$$Mort \sim Precip + Educ + NonWhite + SO2_{log}$$

The model above has the highest R2_adj (0.66) and the lowest BIC (-48.557) among all the options. This model is the same one as suggested by BE and SM method in Step 5.

Step 7: Assumption testing and influential points detection. We drew the residual plot and the Quantile-Quantile(Q-Q) plot of the two models (see the code). We found a funnel shape in the residual plot and the line in Q-Q plot did not fit the diagonal line well, which means the two models do not meet the assumptions of constant variance (i.e., homosecdasity) and normality. We found that the residual plot of *Precip* has an obvious funnel shape, so we decided to transform *Precip* into *Precip^2*. We also drew a Cook's plot (see the code) and found point 60 is very different from other points since it has a very large mortality rate. We dropped point 60 because its leverage is high, and our model failed to explain the high mortality well if including this point.

Step 8: Remove low-quality data points. We also dropped two points (Lancaster city and York city) because people in these two cities prefer to teach children at home, so the education data from the two cities cannot indicate the social climate that other cities have.

Step 9: Collinearity detection. We ran VIF, Eigenvalue, and condition index of the revised model and found that there is still one set of collinearity problem because it has an eigenvalue very close to 0 and a condition index larger than 15. Thus, we applied PCR approach again to address this problem and chose two as the number of PCs because it has the largest R^2 and lowest MSEP, furthermore, two indicator values did not change too much beyond 2. We found a new model (third model):

$$Mort = 1104.70 - 15.18 * Precip^2 - 19.52 * Educ + 18.69 * NonWhite + 9.57 * NOX_{log} + 15.85 * SO2_{log}$$

Step 10: Repeated step 5, we used FS, BE, and SM approaches again to select variables and find good models. Again, both BE and SM suggested the same model (fourth model) with a lowest AIC:

$$Mort \sim Precip^2 + Educ + NonWhite + SO2_{log}$$

Step 11: Similarly, we used *myleaps* function to build a table of all possible models. The table (see code) suggested two possible options by giving a highest R square adjusted and a lowest

BIC value. The last step, we went through a cross-validation process to finally compare the three models suggested by different model selection criteria. We picked the models as they have relatively lower MSEP score (see Result part).

Result

The two best model we found are:

Model 1 (MSEP = 1001):

$$Mort = 1104.70 - 15.18 * Precip^2 - 19.52 * Educ + 18.69 * NonWhite + 9.57 * NOX_{log} + 15.85 * SO2_{log}$$

Model 2 (MSEP = 991):

$$Mort \sim Precip^2 + Educ + NonWhite + SO2_{log}$$

Model	Response	Predictors	R^2%	Adj R^2%	BIC	Constant Variance	Normality	Influential Points	Needed action
1	Original	Nox_log & SO2_log	69	65.9	-45.37	Does not hold	no	60	Remove 60, 4*, 20*, precip_square
2	Original	Nox_log & SO2_log	68	66	-48.56	Does not hold	no	60	Remove 60, 4, 20, Precip_square
3	Original	Nox_log SO2_log Precip_sq	77	74	-58.51	holds	yes	50	Considering remove 50
4	Original	Nox_log SO2_log Precip_sq	75	73	-59.13	holds	yes	50	Considering remove 50

*4: Lankaster *20: York

We remove these two points because they do not reflect the actual situation given the existing information in the material.

Conclusion and Discussion

This report aims to uncover the statistically relationship with pollution variables and mortality rate in 60 U.S. cities the 60s. The method section illustrates the way how we initially examined and strategically transformed the raw data, made comparisons between different models suggested by a set of widely-used model/variable selection criteria, for instance, principal component regression, forward selection procedure, backward elimination procedure, stepwise method, and etc. Two fundamental assumptions of linear regression model, normality and homoscedastic, were inspected accordingly.

Given all these efforts, we finally have two models:

Model 1:

$$Mort = 1104.70 - 15.18 * Precip^2 - 19.52 * Educ + 18.69 * NonWhite + 9.57 * NOX_{log} + 15.85 * SO2_{log}$$

Model 2:

$$Mort \sim Precip + Educ + NonWhite + SO2_{log}$$

i.e.,

$$Mort = 1054.00 + 0.02 * Precip^2 - 20.670 * Educ + 2.39 * NonWhite + 18.23 * SO2_{log}$$

Given the two models we currently have, the answer for our research question becomes obvious: there is a strong evidence that mortality is associated with pollution variables. Specifically, if we adopt Model 1, then all the two pollution variables are believed to have positive influences on mortality rate in the U.S. cities; in other words, about one unit increase in the **log** of relatively pollution potential of oxides of nitrogen (*NOX*), will lead to 9.57 units increase in mortality rate. Similarly, about 15.85 units increase in mortality rate increases by one unit increase in the **log** of relative pollution potential of sulfur dioxide (*SO2*); However, if we adopt Model 2, only one pollution variable, *SO2*, is believed to have positive relation with mortality. About 18.23 units increase in mortality rate increases by one unit increase in the **log** of relative pollution potential of sulfur dioxide (*SO2*)

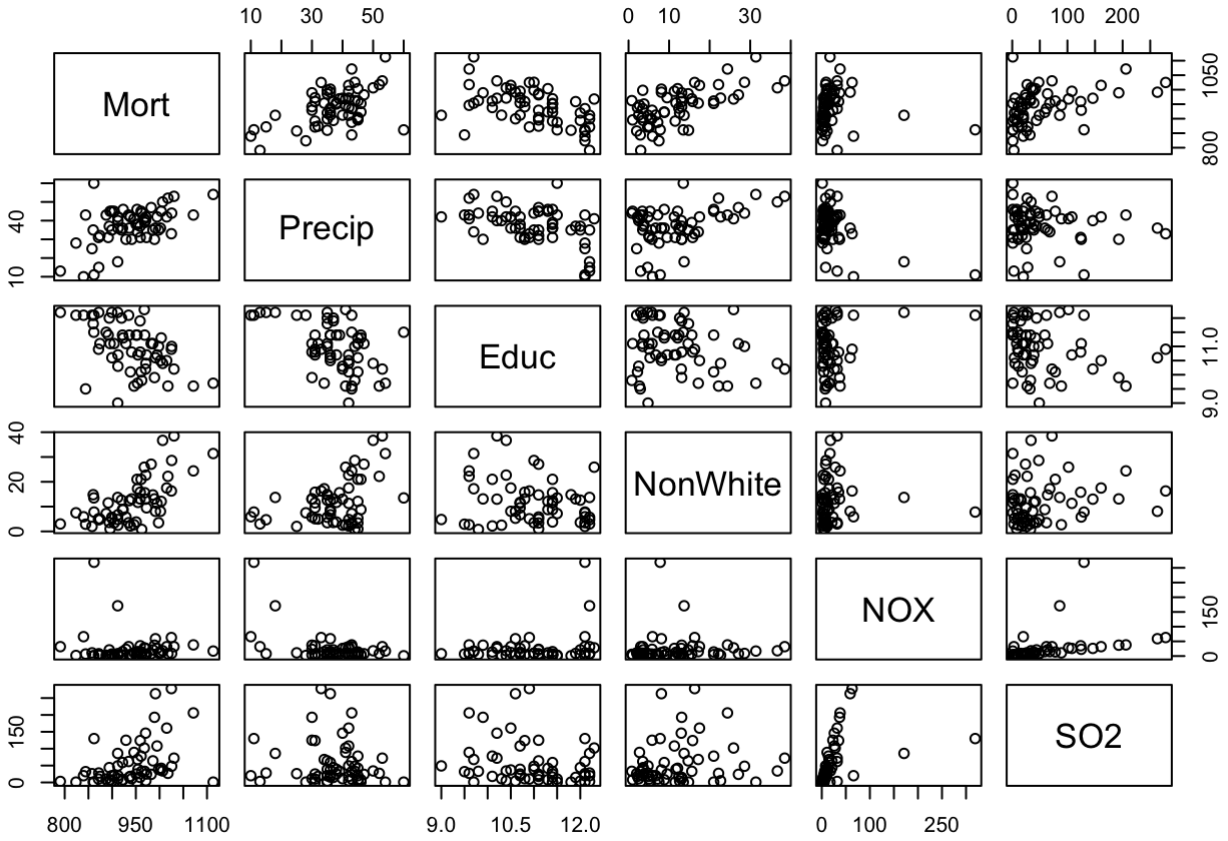
Model selection is a never-end endeavor, and every model has its own limitations. Although we have addressed most problems we identified in the original models; for example, collinearity problem, the existence of influential points, and violation of linear regression models, we realize that the two models we finally adopted still have influential points left. We stopped at the second round of model selection & examine since we did not want to remove too many observations and make it difficult to interpret.

Appendix (complete code and output)

Untitled

```
knitr::opts_chunk$set(echo = TRUE)
setwd("/Users/zhaoyanpeng/Downloads/New\ Folder\ With\ Items\ 2")
pollution <- read.csv("pollution.csv")
```

```
### Matrix Plot
pairs(pollution[,2:7])
```

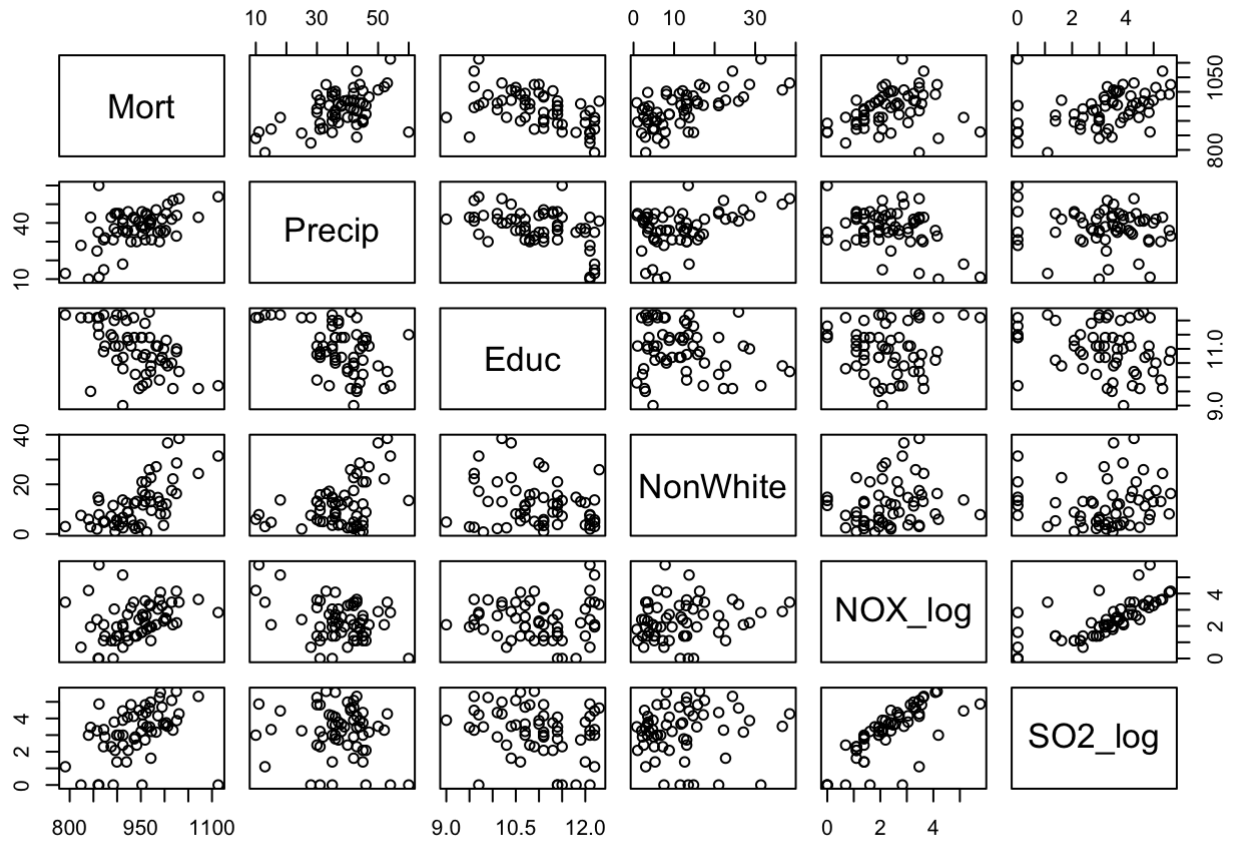



```
summary(pollution)
```

```
##              City         Mort            Precip           Educ
##  Akron, OH     : 1   Min.   : 790.7   Min.   :10.00   Min.   : 9.00
##  Albany, NY    : 1   1st Qu.: 898.4   1st Qu.:32.75   1st Qu.:10.40
##  Allentown, PA : 1   Median : 943.7   Median :38.00   Median :11.05
##  Atlanta, GA   : 1   Mean   : 940.4   Mean   :37.37   Mean   :10.97
##  Baltimore, MD : 1   3rd Qu.: 983.2   3rd Qu.:43.25   3rd Qu.:11.50
##  Birmingham, AL: 1   Max.   :1113.1   Max.   :60.00   Max.   :12.30
##  (Other)       :54
##     NonWhite          NOX              SO2
##  Min.   : 0.80   Min.   :  1.00   Min.   :  1.00
##  1st Qu.: 4.95   1st Qu.:  4.00   1st Qu.: 11.00
##  Median :10.40   Median :  9.00   Median : 30.00
##  Mean   :11.87   Mean   : 22.65   Mean   : 53.77
##  3rd Qu.:15.65   3rd Qu.: 23.75   3rd Qu.: 69.00
##  Max.   :38.50   Max.   :319.00   Max.   :278.00
##
```

Comment: (1)Precip, Education, NonWhite, and SO2 have a linear relationship with Mortality. (2)According to the matrix plot and summary table, NOX and SO2 have very large maximum values but small mean values, suggesting that the distribution of the two variables is skewed and there is a nonlinear relationship between mortality.

```
###Apply Log Transformation
pollution$NOX_log <- log(pollution$NOX)
pollution$SO2_log <- log(pollution$SO2)
pairs(pollution[c(2:5, 8, 9)])
```




```
#Find VIF, eigenvalue, and Condition index
model1 <- lm(Mort ~., pollution[,c(2:5, 8, 9)])
summary(model1)
```

```
##
## Call:
## lm(formula = Mort ~ ., data = pollution[, c(2:5, 8, 9)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.222  -19.547    0.239   20.084   95.386
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 940.6584    94.0551  10.001 6.81e-14 ***
## Precip        1.9467     0.7007   2.778   0.0075 **
## Educ        -14.6645     6.9379  -2.114   0.0392 *
## NonWhite      3.0289     0.6685   4.531 3.29e-05 ***
## NOX_log       6.7164     7.3990   0.908   0.3680
## SO2_log      11.3578     5.2955   2.145   0.0365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.3 on 54 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6594
## F-statistic: 23.85 on 5 and 54 DF,  p-value: 1.418e-12
```

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_vif_tol(model1)
```

```
##   Variables Tolerance      VIF
## 1    Precip 0.4562987 2.191547
## 2      Educ 0.6493911 1.539904
## 3  NonWhite 0.6279259 1.592545
## 4   NOX_log 0.2908238 3.438508
## 5   SO2_log 0.3551147 2.815991
```

```
ols_coll_diag(model1)
```

```
## Tolerance and Variance Inflation Factor
## ---------------------------------------
##   Variables Tolerance      VIF
## 1    Precip 0.4562987 2.191547
## 2      Educ 0.6493911 1.539904
## 3  NonWhite 0.6279259 1.592545
## 4   NOX_log 0.2908238 3.438508
## 5   SO2_log 0.3551147 2.815991
##
##
## Eigenvalue and Condition Index
## ------------------------------
##    Eigenvalue Condition Index    intercept       Precip         Educ
## 1 5.366301613        1.000000 8.327776e-05 0.0009590522 1.247298e-04
## 2 0.323445180        4.073214 2.937158e-05 0.0039169078 7.955257e-05
## 3 0.224853295        4.885263 1.376898e-03 0.0263500757 2.301258e-03
## 4 0.063571607        9.187682 1.796009e-03 0.0821492361 9.454325e-03
## 5 0.020327291       16.247921 5.653159e-03 0.5866007350 2.499935e-02
## 6 0.001501015       59.792268 9.910613e-01 0.3000239932 9.630408e-01
##      NonWhite      NOX_log     SO2_log
## 1 0.005847166 1.808200e-03 0.001955392
## 2 0.438601860 2.520439e-02 0.034409536
## 3 0.194029830 6.864535e-02 0.023722597
## 4 0.001415999 1.736497e-01 0.429326387
## 5 0.357499358 7.306554e-01 0.425130444
## 6 0.002605787 3.692961e-05 0.085455645
```

Eigenvalue and condition index suggest there are two sets of colinearity because the two condition indexes are larger than 15 and the eigenvalues are very close to 0.

```
###Use PCR
library(pls)
```

```
## Warning: package 'pls' was built under R version 3.6.2
```

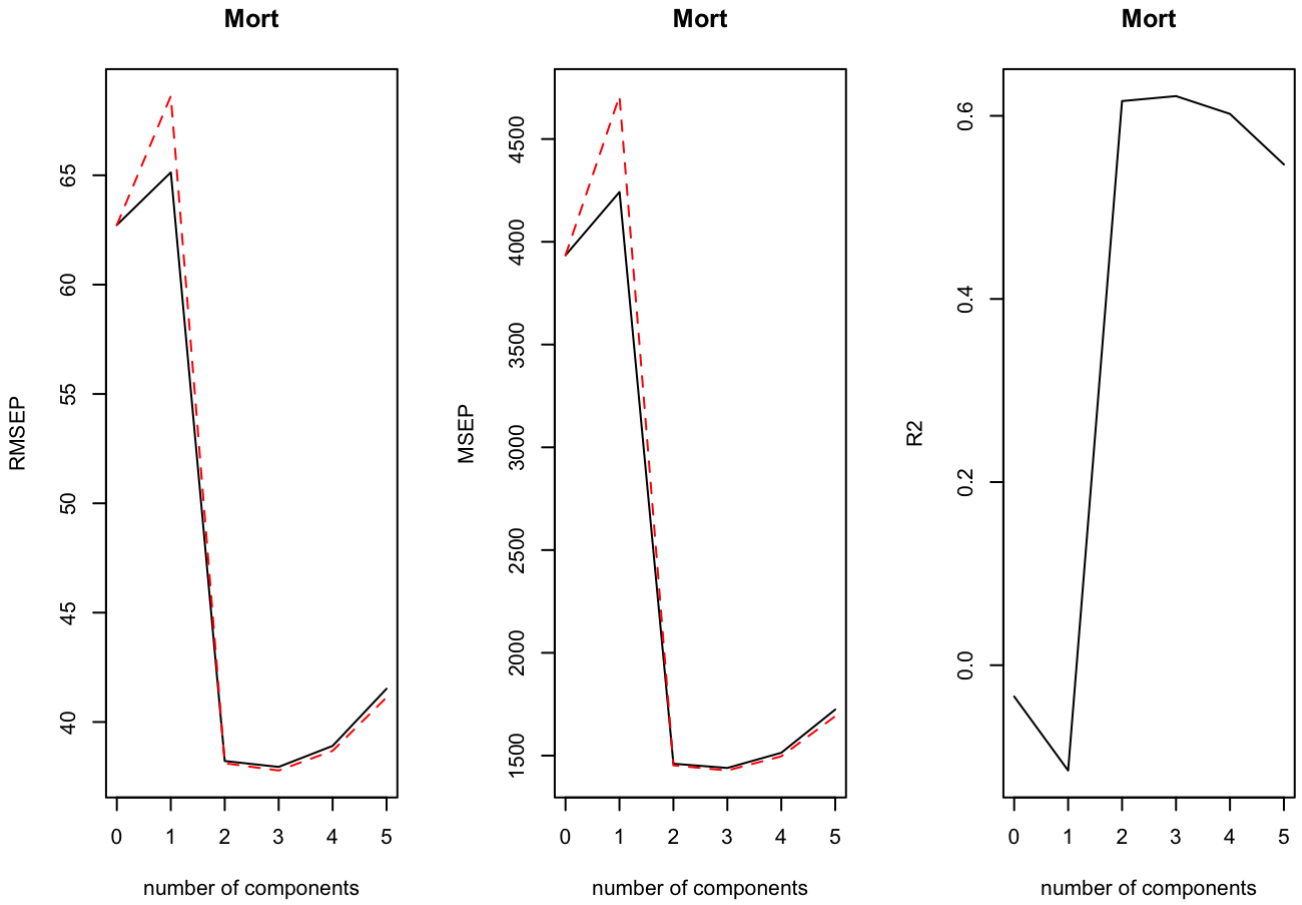
```
##
## Attaching package: 'pls'
```

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

```
pcr_model <- pcr(Mort~., data=pollution[,c(2:5, 8, 9)], scale = T, center=TRUE, validati
on = "CV")
summary(pcr_model)
```

```
## Data:    X dimension: 60 5
##  Y dimension: 60 1
## Fit method: svdpc
## Number of components considered: 5
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
## CV           62.73    65.13    38.22    37.95    38.91    41.52
## adjCV        62.73    68.62    38.11    37.78    38.68    41.13
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps
## X     37.5309    72.64    89.88    96.91   100.00
## Mort   0.6449    66.11    68.47    68.74    68.83
```

```
### Determine how many PCs to keep in the model
par(mfrow=c(1,3))
# Plot the root mean squared error
validationplot(pcr_model)
# Plot the cross validation MSE
validationplot(pcr_model, val.type="MSEP")
# Plot the R2
validationplot(pcr_model, val.type = "R2")
```



Decide to keep 2 PCs because it has large R^2 and low MSEP, and these two values haven't changed much after 2 PCs, so it is unneccessary to choose a larger PCs.

```
# First possible model found using Pcr
library(pls)
possible_model1 <- pcr(Mort~., data = pollution[, c(2:5, 8, 9)], scale = T,
                   center=TRUE, ncomp = 2)
coef(possible_model1, ncomp = 2, intercept = TRUE)
```

```
## , , 2 comps
##
##                    Mort
## (Intercept) 1078.965866
## Precip        17.714628
## Educ         -21.617556
## NonWhite      19.119834
## NOX_log        9.573169
## SO2_log       14.757224
```

```
#Selecting the second possible model by using forward, backward, and stepwise
n=dim(pollution)[1]
## Get the full and the null models
fit0 <- lm(Mort ~ 1, data=pollution[, c(2:5, 8, 9)])
## Null model
fitall <- lm(Mort~., data=pollution[, c(2:5, 8, 9)])
# Foward
m1=step(fit0, fitall, direction = "forward", k = log(n))
```

```
## Start:  AIC=498.73
## Mort ~ 1
```

```
summary(m1)
```

```
##
## Call:
## lm(formula = Mort ~ 1, data = pollution[, c(2:5, 8, 9)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -149.627  -41.986    3.323   42.849  172.703
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   940.36       8.03   117.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.2 on 59 degrees of freedom
```

```
m2=step(fitall, direction = "backward", k = log(n))
```

```
## Start:  AIC=449.27
## Mort ~ Precip + Educ + NonWhite + NOX_log + SO2_log
##
##            Df Sum of Sq   RSS    AIC
## - NOX_log   1    1085.8 72245 446.08
## <none>                  71159 449.27
## - Educ      1    5887.3 77046 449.94
## - SO2_log   1    6061.9 77221 450.08
## - Precip    1   10171.6 81331 453.19
## - NonWhite  1   27050.8 98210 464.50
##
## Step:  AIC=446.08
## Mort ~ Precip + Educ + NonWhite + SO2_log
##
##            Df Sum of Sq    RSS    AIC
## <none>                  72245 446.08
## - Educ      1      5405  77650 446.31
## - Precip    1      9400  81645 449.32
## - SO2_log   1     25118  97363 459.89
## - NonWhite  1     42342 114587 469.66
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + SO2_log, data = pollution[,
##     c(2:5, 8, 9)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.364  -22.544   -0.716   19.216  110.767
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 944.8448    93.7916  10.074 4.28e-14 ***
## Precip        1.6410     0.6134   2.675  0.00982 **
## Educ        -13.9637     6.8838  -2.028  0.04736 *
## NonWhite      3.3212     0.5850   5.678 5.31e-07 ***
## SO2_log      15.0134     3.4333   4.373 5.50e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.24 on 55 degrees of freedom
## Multiple R-squared:  0.6835, Adjusted R-squared:  0.6605
## F-statistic:  29.7 on 4 and 55 DF,  p-value: 3.599e-13
```

```
m3=step(fitall, direction = "both", k = log(n))
```

```
## Start:  AIC=449.27
## Mort ~ Precip + Educ + NonWhite + NOX_log + SO2_log
##
##            Df Sum of Sq   RSS    AIC
## - NOX_log   1    1085.8 72245 446.08
## <none>                  71159 449.27
## - Educ      1    5887.3 77046 449.94
## - SO2_log   1    6061.9 77221 450.08
## - Precip    1   10171.6 81331 453.19
## - NonWhite  1   27050.8 98210 464.50
##
## Step:  AIC=446.08
## Mort ~ Precip + Educ + NonWhite + SO2_log
##
##            Df Sum of Sq    RSS    AIC
## <none>                  72245 446.08
## - Educ      1      5405  77650 446.31
## + NOX_log   1      1086  71159 449.27
## - Precip    1      9400  81645 449.32
## - SO2_log   1     25118  97363 459.89
## - NonWhite  1     42342 114587 469.66
```

```
summary(m3)
```

```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + SO2_log, data = pollution[,
##     c(2:5, 8, 9)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.364  -22.544   -0.716   19.216  110.767
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 944.8448    93.7916  10.074 4.28e-14 ***
## Precip        1.6410     0.6134   2.675  0.00982 **
## Educ        -13.9637     6.8838  -2.028  0.04736 *
## NonWhite      3.3212     0.5850   5.678 5.31e-07 ***
## SO2_log      15.0134     3.4333   4.373 5.50e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.24 on 55 degrees of freedom
## Multiple R-squared:  0.6835, Adjusted R-squared:  0.6605
## F-statistic:  29.7 on 4 and 55 DF,  p-value: 3.599e-13
```

###We choose our second model based on the result of backward selection and stepwise selection.

```
library(leaps)
myleaps <- regsubsets(Mort~., data = pollution[, c(2:5, 8, 9)], nbest = 6)
(myleaps.summary <- summary(myleaps))
```

```
## Subset selection object
## Call: regsubsets.formula(Mort ~ ., data = pollution[, c(2:5, 8, 9)],
##     nbest = 6)
## 5 Variables  (and intercept)
##          Forced in Forced out
## Precip       FALSE      FALSE
## Educ         FALSE      FALSE
## NonWhite     FALSE      FALSE
## NOX_log      FALSE      FALSE
## SO2_log      FALSE      FALSE
## 6 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          Precip Educ NonWhite NOX_log SO2_log
## 1  ( 1 ) " "    " "  "*"      " "     " "
## 1  ( 2 ) " "    "*"  " "      " "     " "
## 1  ( 3 ) "*"    " "  " "      " "     " "
## 1  ( 4 ) " "    " "  " "      " "     "*"
## 1  ( 5 ) " "    " "  " "      "*"     " "
## 2  ( 1 ) " "    "*"  "*"      " "     " "
## 2  ( 2 ) " "    " "  "*"      " "     "*"
## 2  ( 3 ) "*"    " "  " "      "*"     " "
## 2  ( 4 ) "*"    " "  "*"      " "     " "
## 2  ( 5 ) "*"    " "  " "      " "     "*"
## 2  ( 6 ) " "    " "  "*"      "*"     " "
## 3  ( 1 ) "*"    " "  "*"      " "     "*"
## 3  ( 2 ) " "    "*"  "*"      " "     "*"
## 3  ( 3 ) "*"    " "  "*"      "*"     " "
## 3  ( 4 ) " "    "*"  "*"      "*"     " "
## 3  ( 5 ) "*"    "*"  "*"      " "     " "
## 3  ( 6 ) " "    " "  "*"      "*"     "*"
## 4  ( 1 ) "*"    "*"  "*"      " "     "*"
## 4  ( 2 ) "*"    " "  "*"      "*"     "*"
## 4  ( 3 ) "*"    "*"  "*"      "*"     " "
## 4  ( 4 ) " "    "*"  "*"      "*"     "*"
## 4  ( 5 ) "*"    "*"  " "      "*"     "*"
## 5  ( 1 ) "*"    "*"  "*"      "*"     "*"
```

```
bettertable <- cbind(myleaps.summary$which,
                     myleaps.summary$rsq, myleaps.summary$rss,
                     myleaps.summary$adjr2, myleaps.summary$cp,myleaps.summary$bic)
dimnames(bettertable)[[2]] <- c(dimnames(myleaps.summary$which)[[2]],
                                "R2", "sse", "R2_ADJ", "CP", "BIC")
show(bettertable)
```

```
##   (Intercept) Precip Educ NonWhite NOX_log SO2_log         R2       sse
## 1           1      0    0        1       0       0 0.41439274 133679.73
## 1           1      0    1        0       0       0 0.26110447 168671.67
## 1           1      1    0        0       0       0 0.25958245 169019.11
## 1           1      0    0        0       0       1 0.16251215 191177.87
## 1           1      0    0        0       1       0 0.08526371 208811.79
## 2           1      0    1        1       0       0 0.56267518  99830.50
## 2           1      0    0        1       0       1 0.55121781 102445.93
## 2           1      1    0        0       1       0 0.52574813 108260.03
## 2           1      1    0        1       0       0 0.48589205 117358.19
## 2           1      1    0        0       0       1 0.47890149 118953.97
## 2           1      0    0        1       1       0 0.44433807 126843.95
## 3           1      1    0        1       0       1 0.65984087  77649.96
## 3           1      0    1        1       0       1 0.64234156  81644.62
## 3           1      1    0        1       1       0 0.61118317  88757.32
## 3           1      0    1        1       1       0 0.60128460  91016.92
## 3           1      1    1        1       0       0 0.57348310  97363.31
## 3           1      0    0        1       1       1 0.57297556  97479.17
## 4           1      1    1        1       0       1 0.68351829  72244.99
## 4           1      1    0        1       1       1 0.66248471  77046.44
## 4           1      1    1        1       1       0 0.66171986  77221.03
## 4           1      0    1        1       1       1 0.64371658  81330.74
## 4           1      1    1        0       1       1 0.56977432  98209.94
## 5           1      1    1        1       1       1 0.68827500  71159.15
##       R2_ADJ          CP        BIC
## 1 0.40429607  45.444517 -23.917666
## 1 0.24836489  71.998586  -9.967235
## 1 0.24681663  72.262244  -9.843771
## 1 0.14807270  89.077695  -2.452222
## 1 0.06949239 102.459412   2.841522
## 2 0.54733045  21.757608 -37.341710
## 2 0.53547107  23.742364 -35.790023
## 2 0.50910771  28.154467 -32.477970
## 2 0.46785318  35.058719 -27.636288
## 2 0.46061733  36.269692 -26.825937
## 2 0.42484116  42.257101 -22.972678
## 3 0.64161806   6.925633 -48.323127
## 3 0.62318129   9.957032 -45.313231
## 3 0.59035370  15.354588 -40.301437
## 3 0.57992485  17.069313 -38.793066
## 3 0.55063398  21.885356 -34.748819
## 3 0.55009925  21.973278 -34.677463
## 4 0.66050144   4.824003 -48.557667
## 4 0.63793815   8.467642 -44.696946
## 4 0.63711767   8.600136 -44.561133
## 4 0.61780506  11.718838 -41.450001
## 4 0.53848518  24.527827 -30.135000
## 5 0.65941158   6.000000 -45.371966
```

According to model selection table above, the model: Mort ~ Precip +Educ + NonWhite + SO2_log has the highest R2_adj (0.66) and the lowest BIC (-48.557). This model is the same one as the model selected by using backward and stepwise selection method.

#(Ze)cross-validation could be used to select the best model from two models

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.2
```

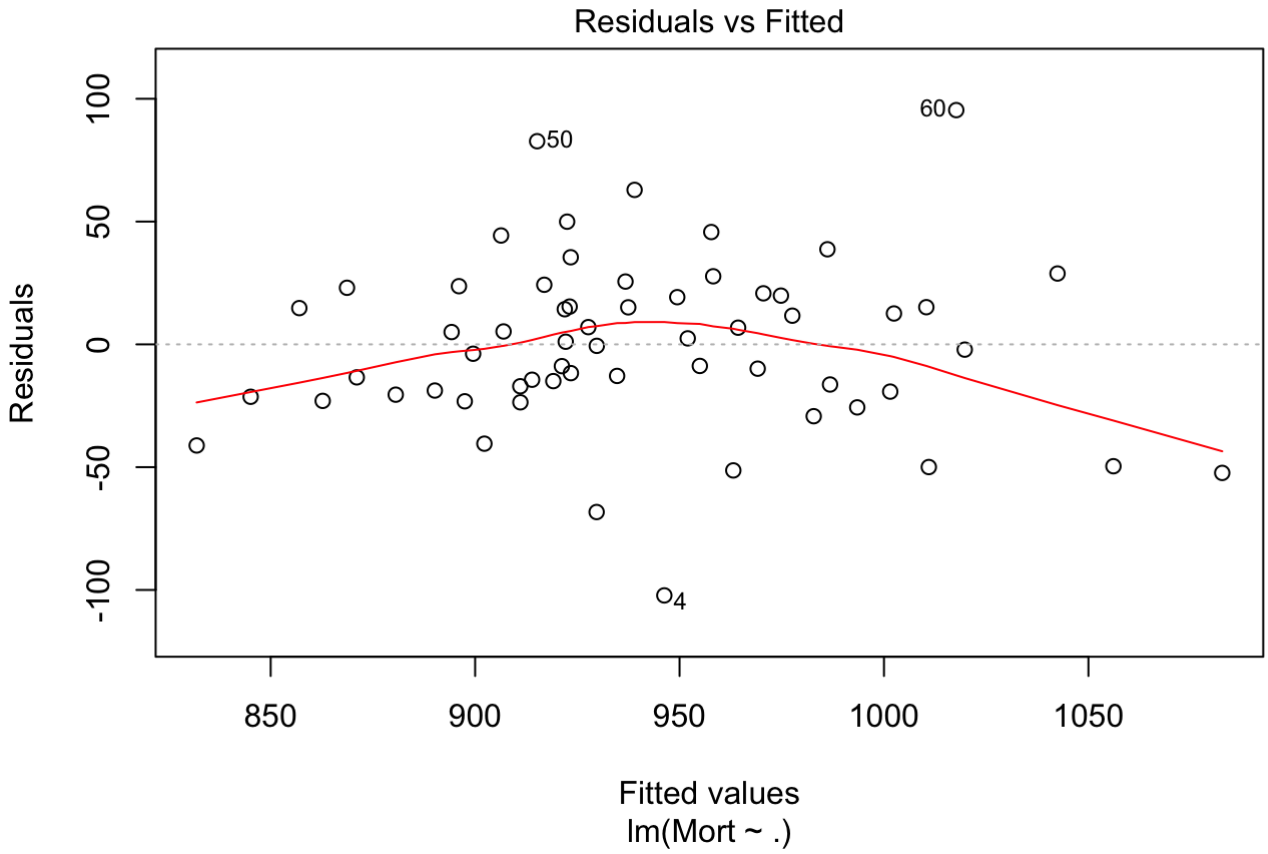
```
## Loading required package: carData
```

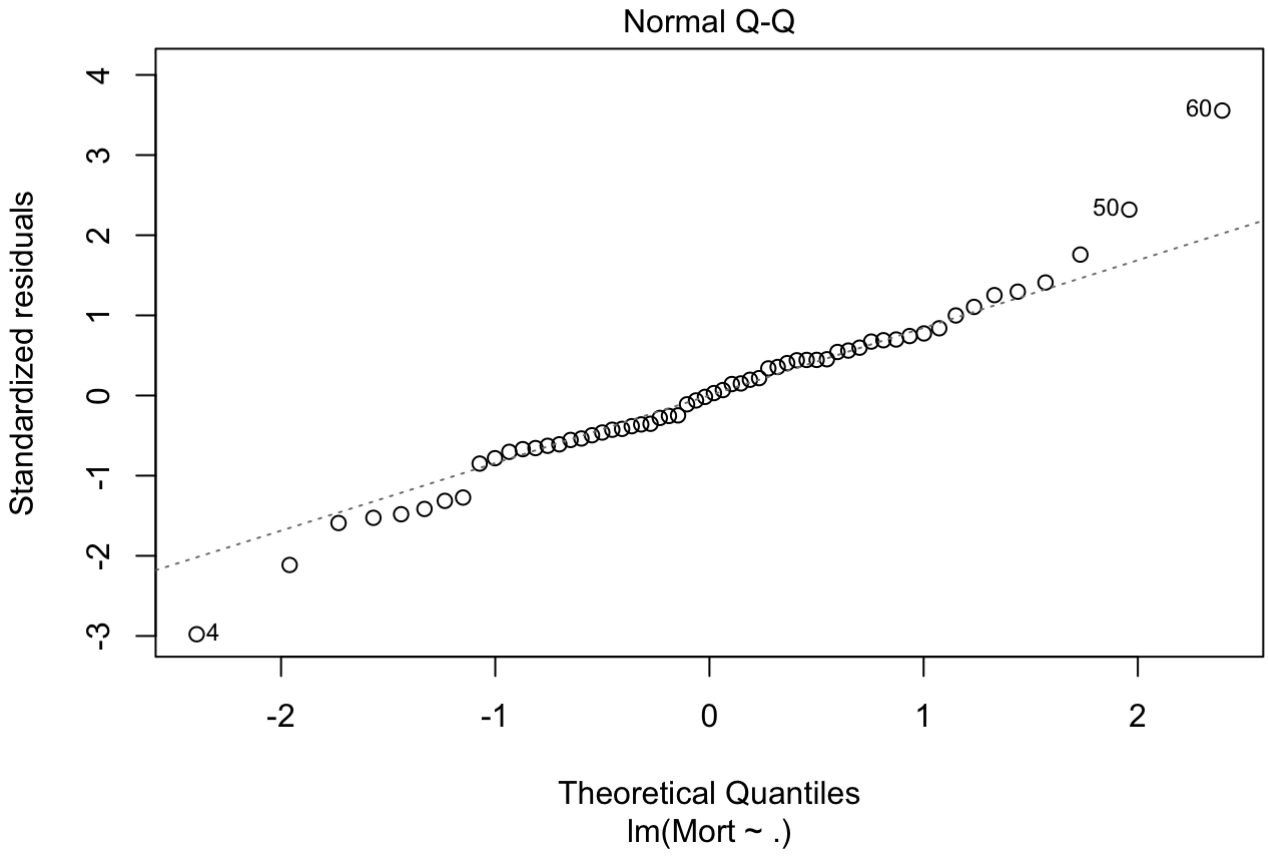
```
## Warning: package 'carData' was built under R version 3.6.2
```

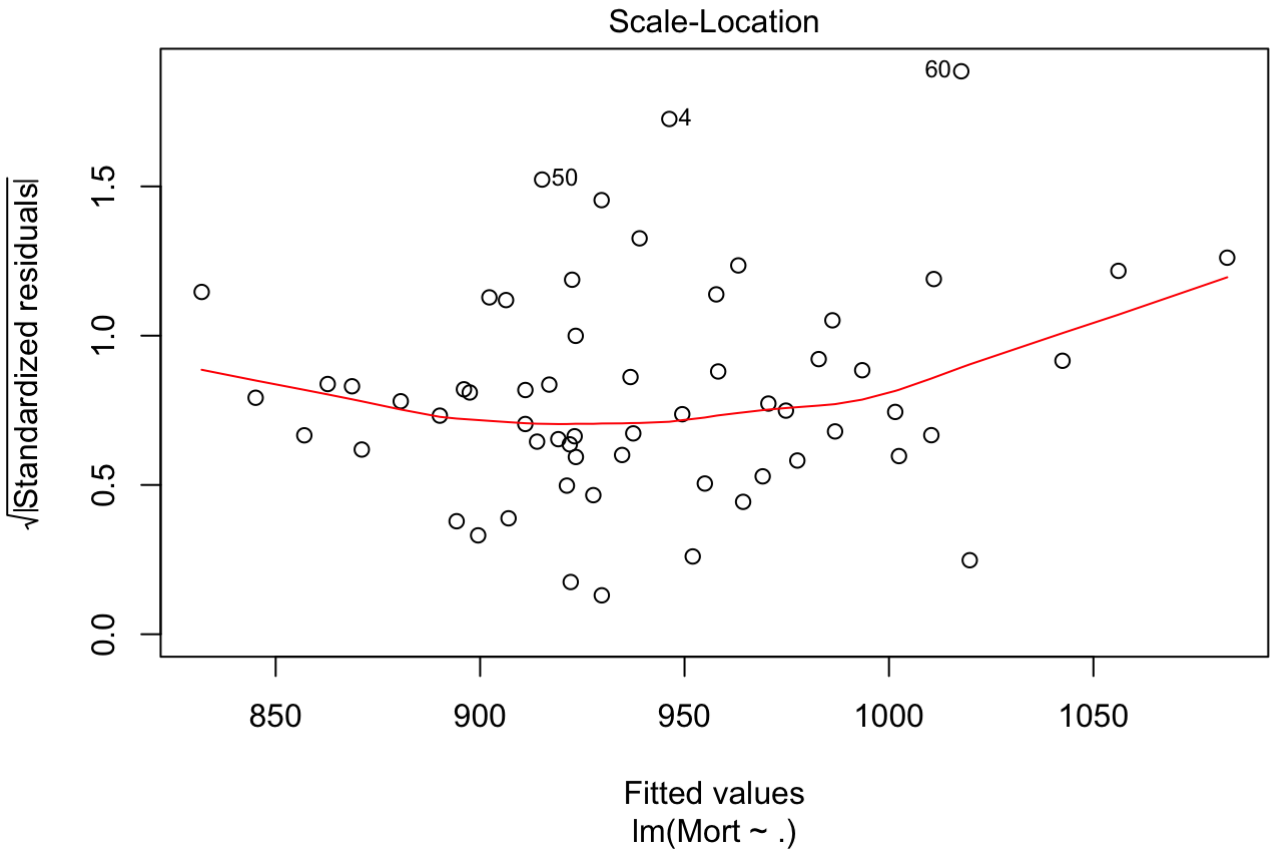
```
library(ggplot2)
```

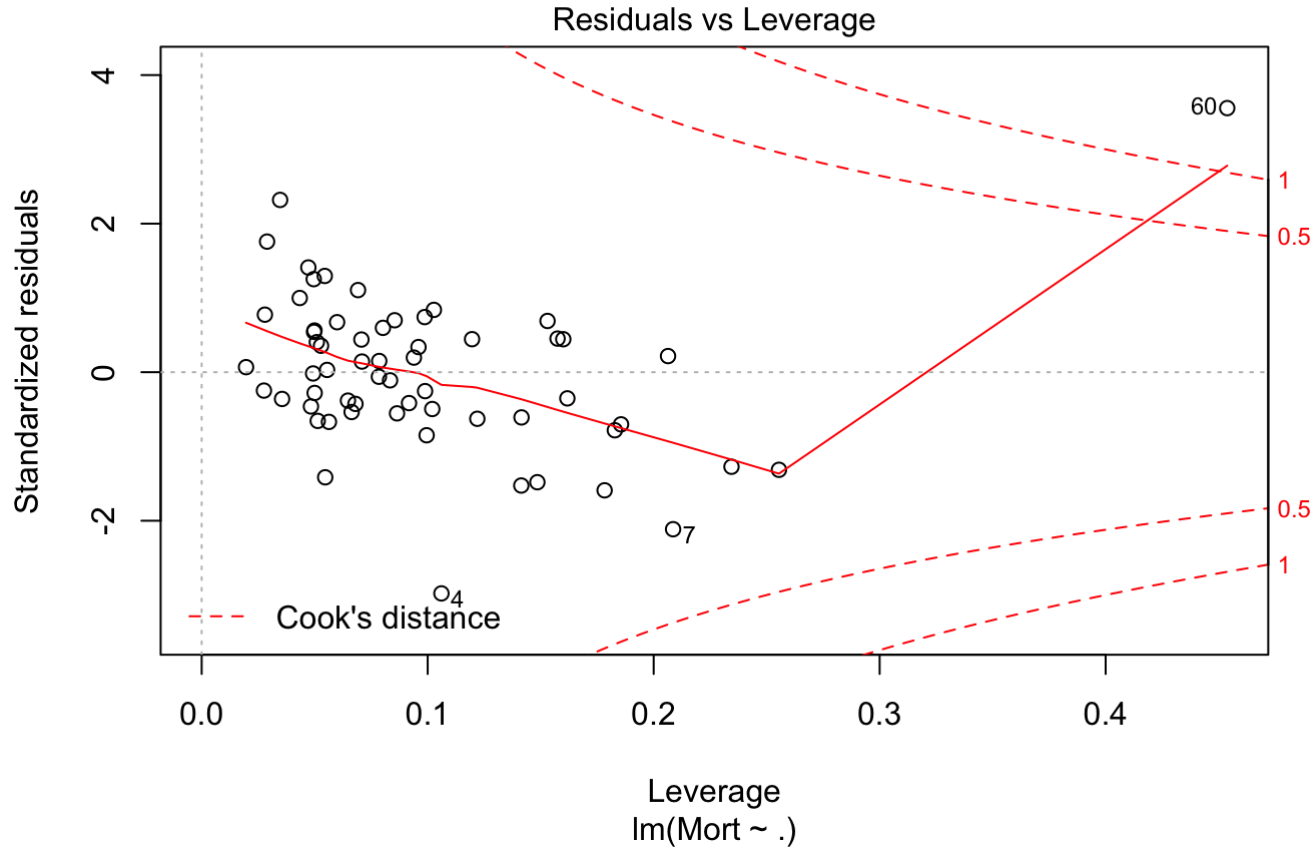
```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
#Check Assumption
possiblemodel1 <- lm(Mort ~., data = pollution[, c(2:5, 8, 9)])
possiblemodel2 <- lm(Mort ~ Precip + Educ + NonWhite + SO2_log, data = pollution[, c(2:5
, 8, 9)])
plot(possiblemodel1)
```

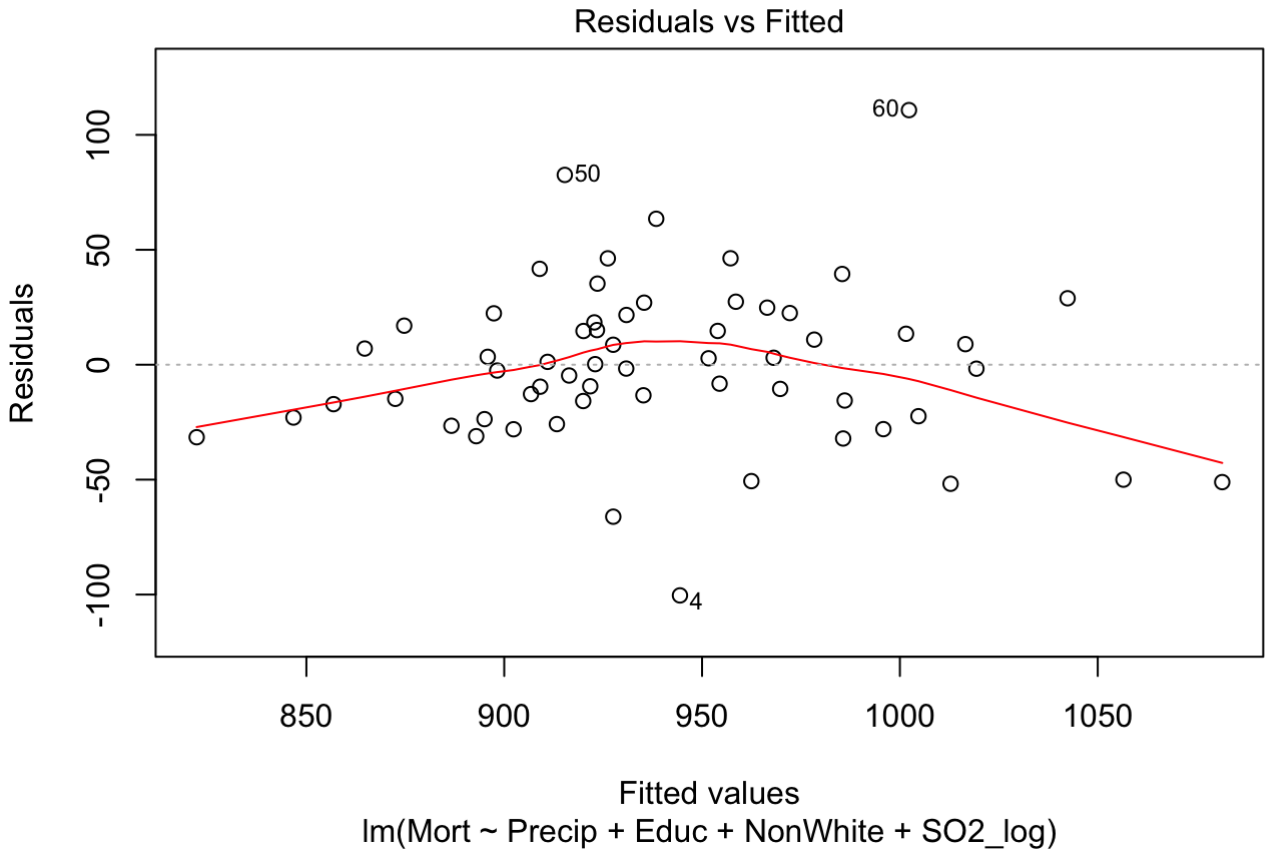

Residuals vs Fitted
Residuals
Fitted values
lm(Mort ~ .)
60
50
4
100
50
0
-50
-100
850
900
950
1000
1050

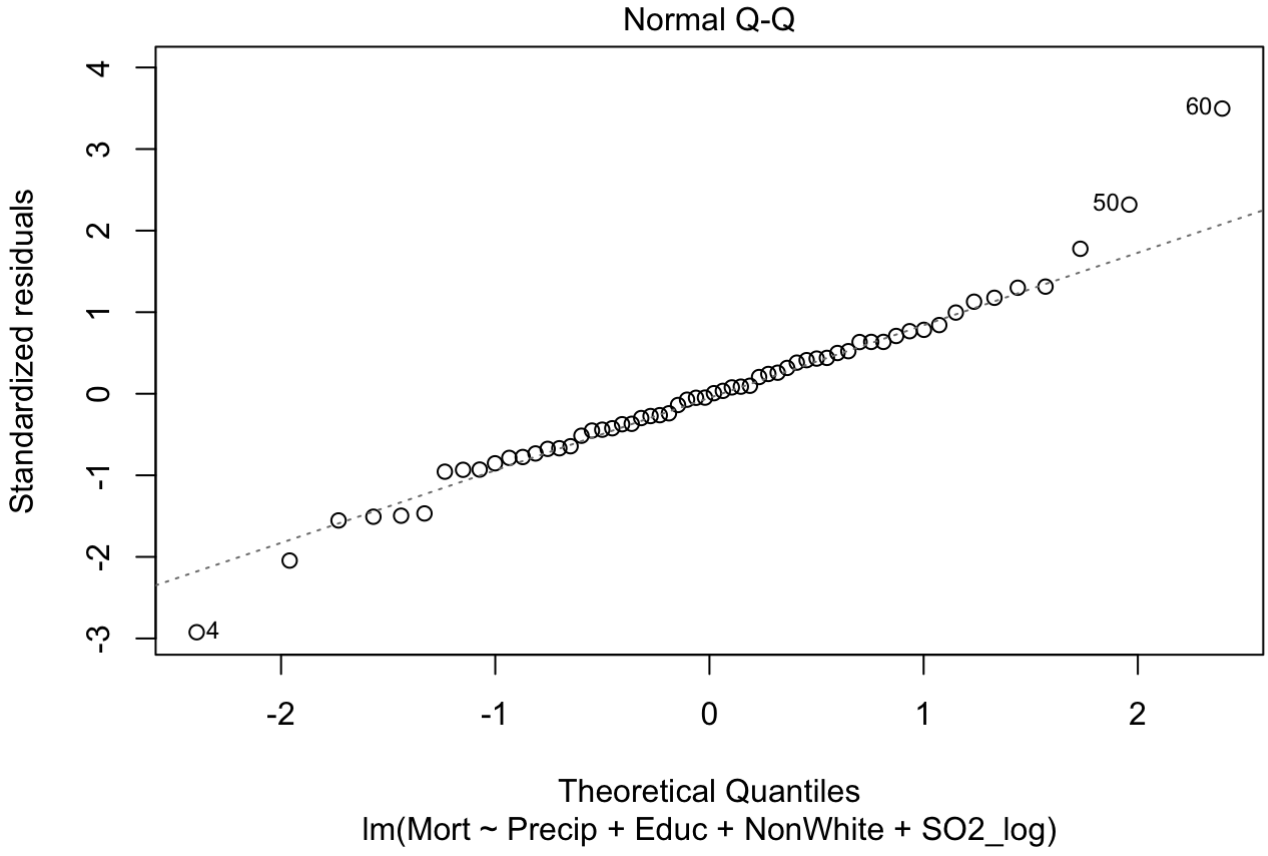

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Mort ~ .)
60
50
4
4
3
2
1
0
-1
-2
-3
-2
-1
0
1
2

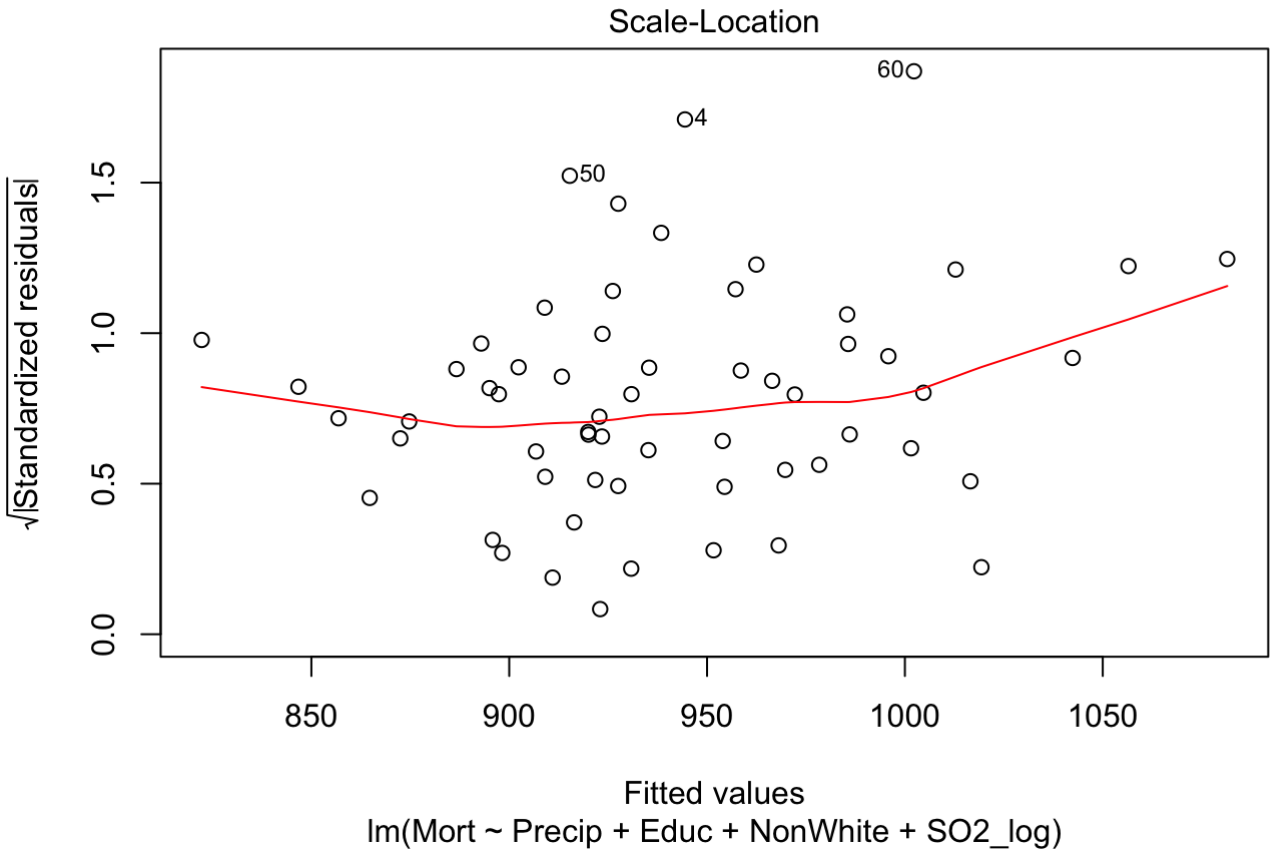

Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
850
900
950
1000
1050
60
4
50
Fitted values
lm(Mort ~ .)

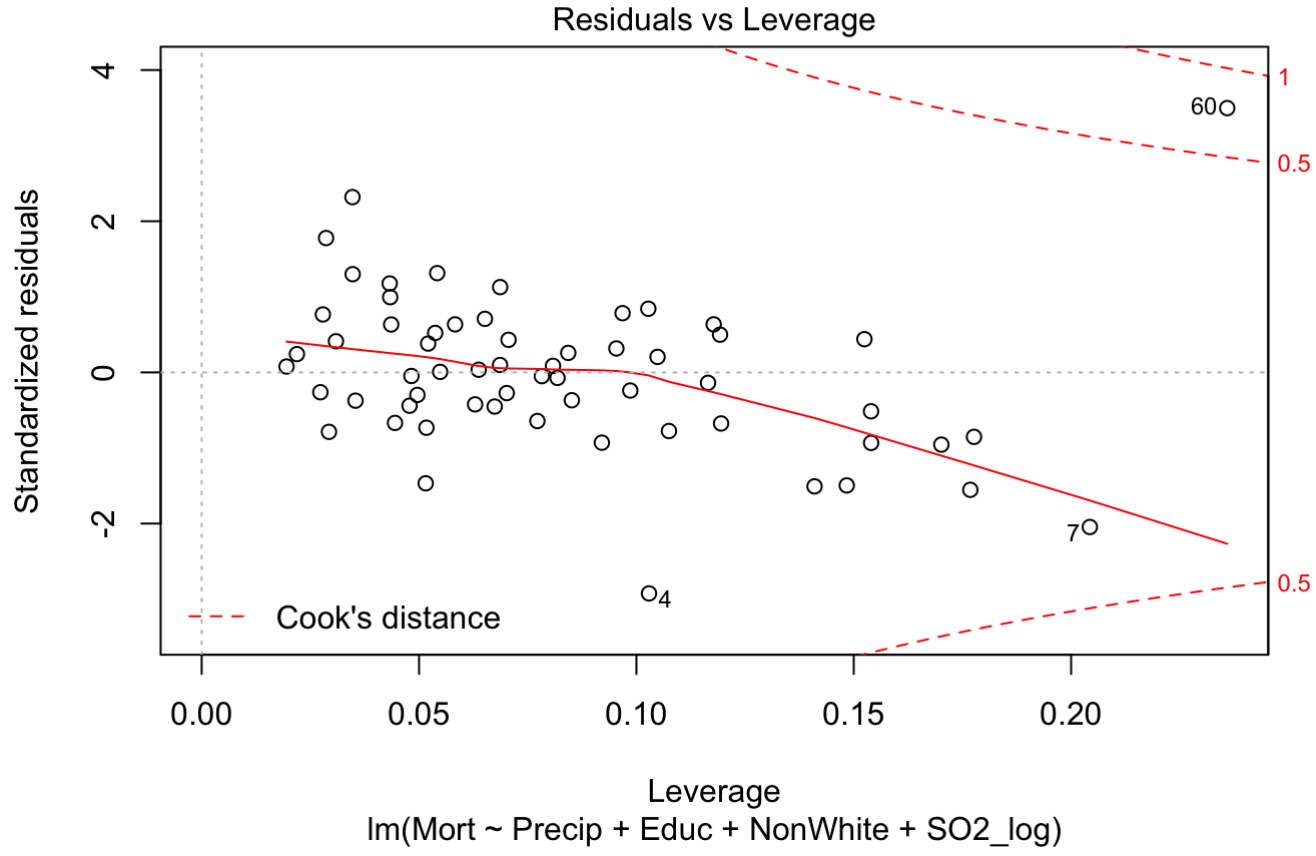

Residuals vs Leverage
Standardized residuals
-2
0
2
4
0.0
0.1
0.2
0.3
0.4
60
7
4
1
0.5
0.5
1
Cook's distance
Leverage
lm(Mort ~ .)

```
plot(possiblemodel2)
```

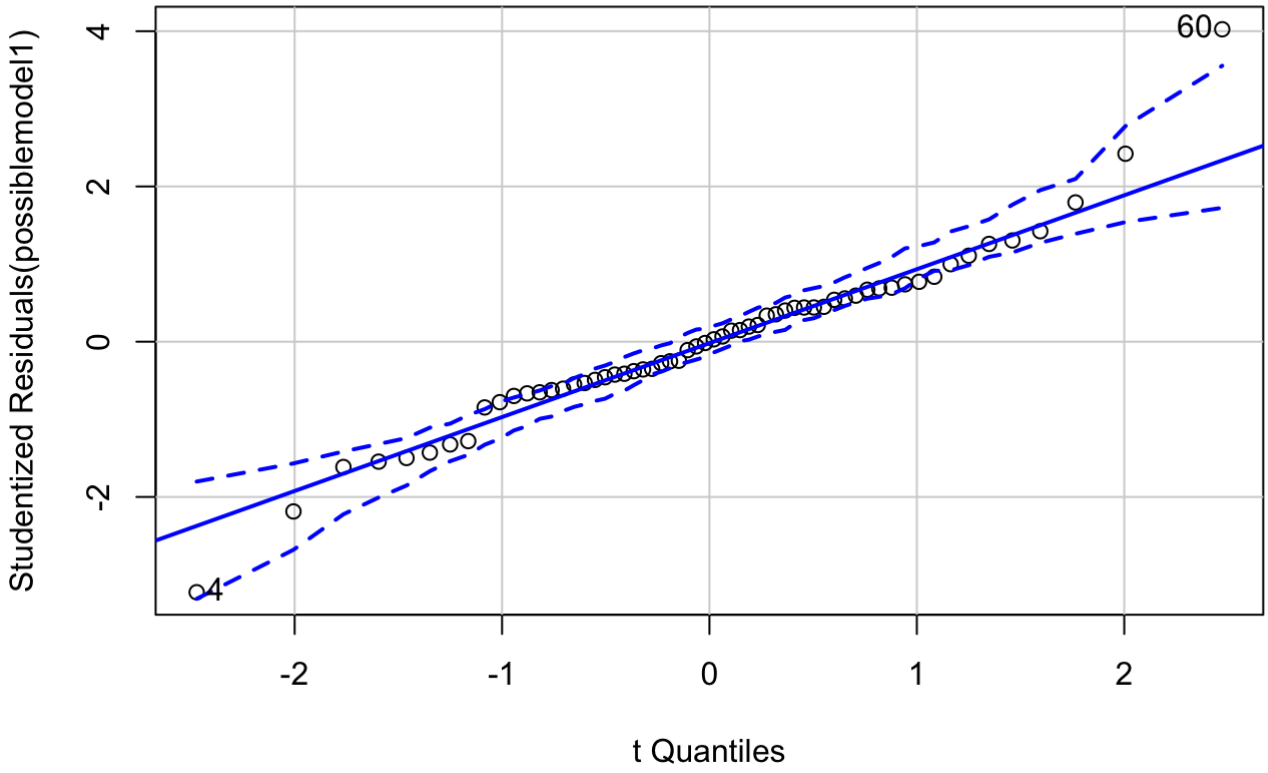

Residuals vs Fitted
Residuals
Fitted values
lm(Mort ~ Precip + Educ + NonWhite + SO2_log)
60
50
4
-100
-50
0
50
100
850
900
950
1000
1050


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Mort ~ Precip + Educ + NonWhite + SO2_log)
60
50
4
-3
-2
-1
0
1
2
3
4


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
850
900
950
1000
1050
60
4
50
Fitted values
lm(Mort ~ Precip + Educ + NonWhite + SO2_log)

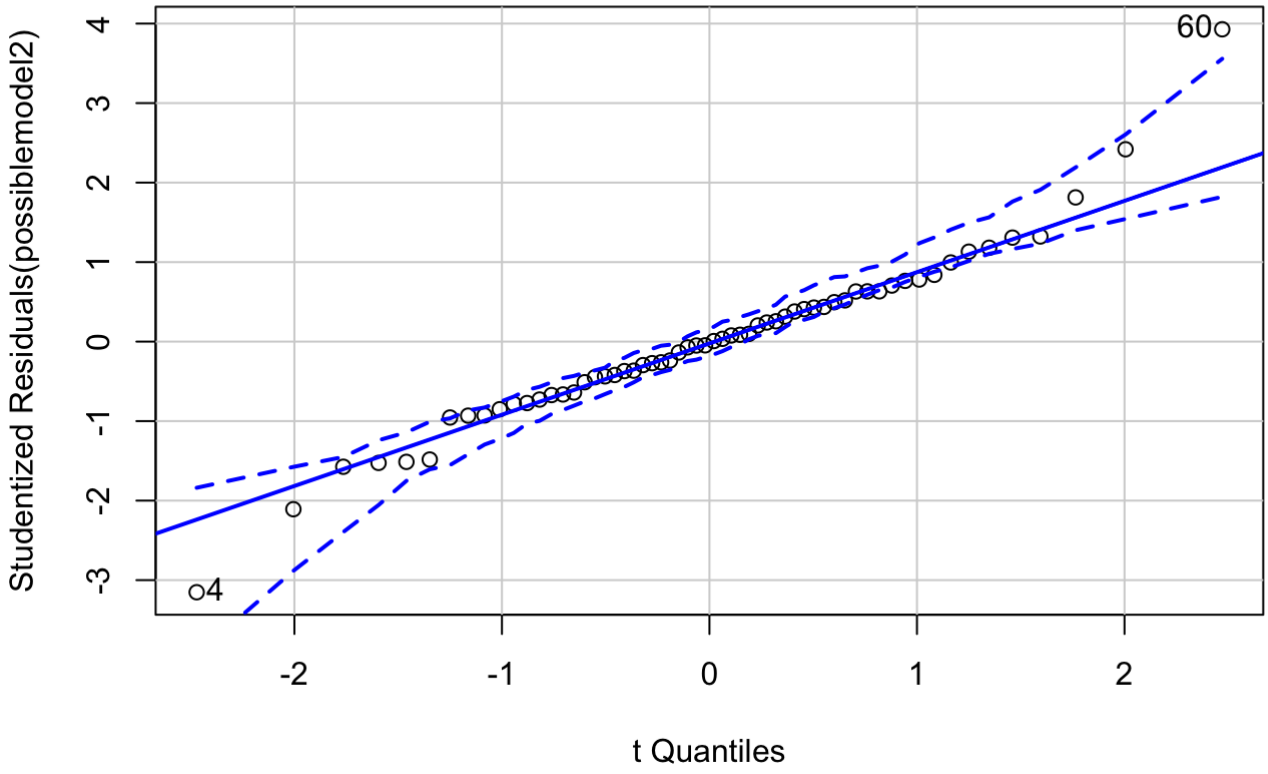

Residuals vs Leverage
Standardized residuals
-2
0
2
4
0.00
0.05
0.10
0.15
0.20
60
7
4
1
0.5
0.5
Cook's distance
Leverage
lm(Mort ~ Precip + Educ + NonWhite + SO2_log)

```
qqPlot(possiblemodel1)
```



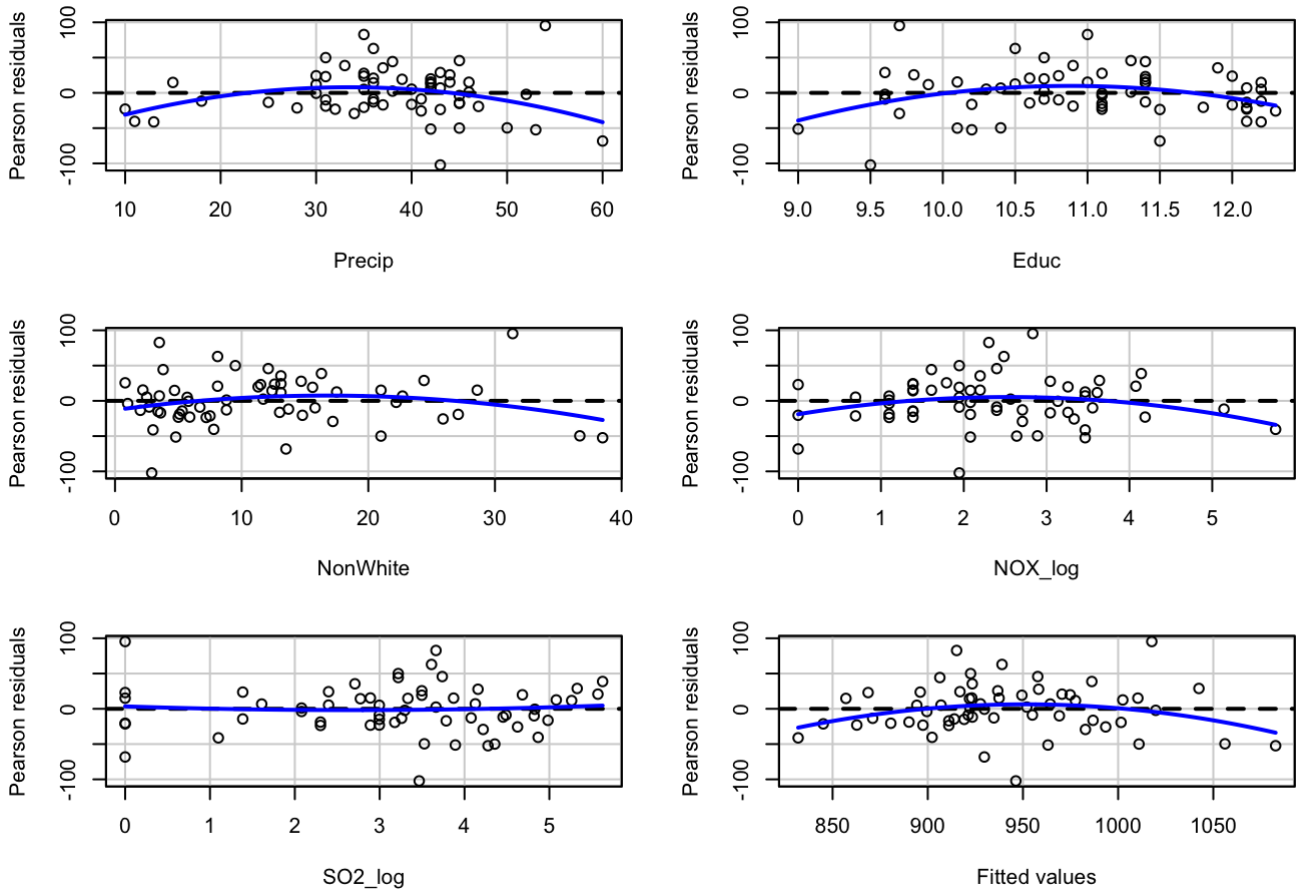

```
## [1]  4 60
```

```
qqPlot(possiblemodel2)
```



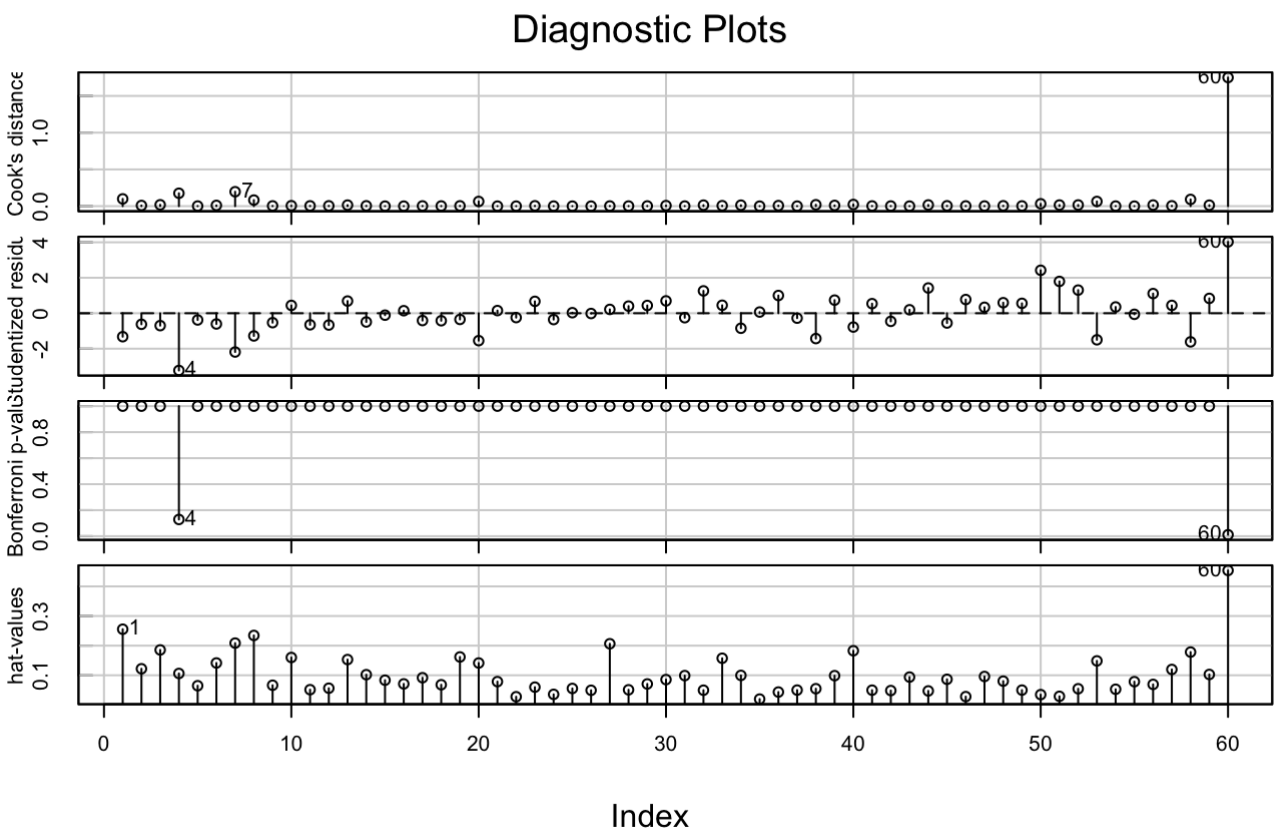

```
## [1]  4 60
```

```
residualPlots(possiblemodel1)
```

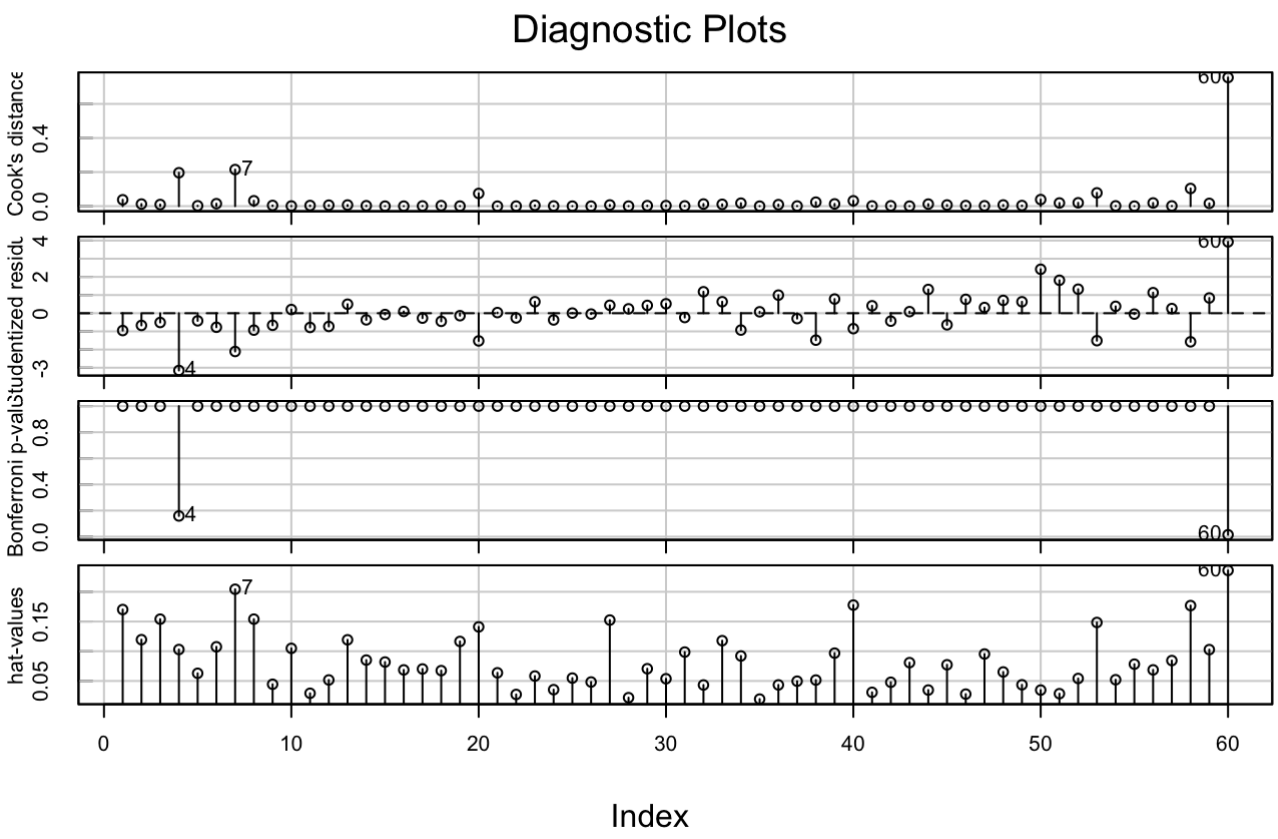



```
##            Test stat Pr(>|Test stat|)
## Precip       -3.1965         0.002346 **
## Educ         -2.5515         0.013648 *
## NonWhite     -1.6035         0.114759
## NOX_log      -1.8892         0.064340 .
## SO2_log       0.4547         0.651183
## Tukey test   -2.2359         0.025358 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Find influential point(s) to drop
influenceIndexPlot(possiblemodel1)
```




```
influenceIndexPlot(possiblemodel2)
```

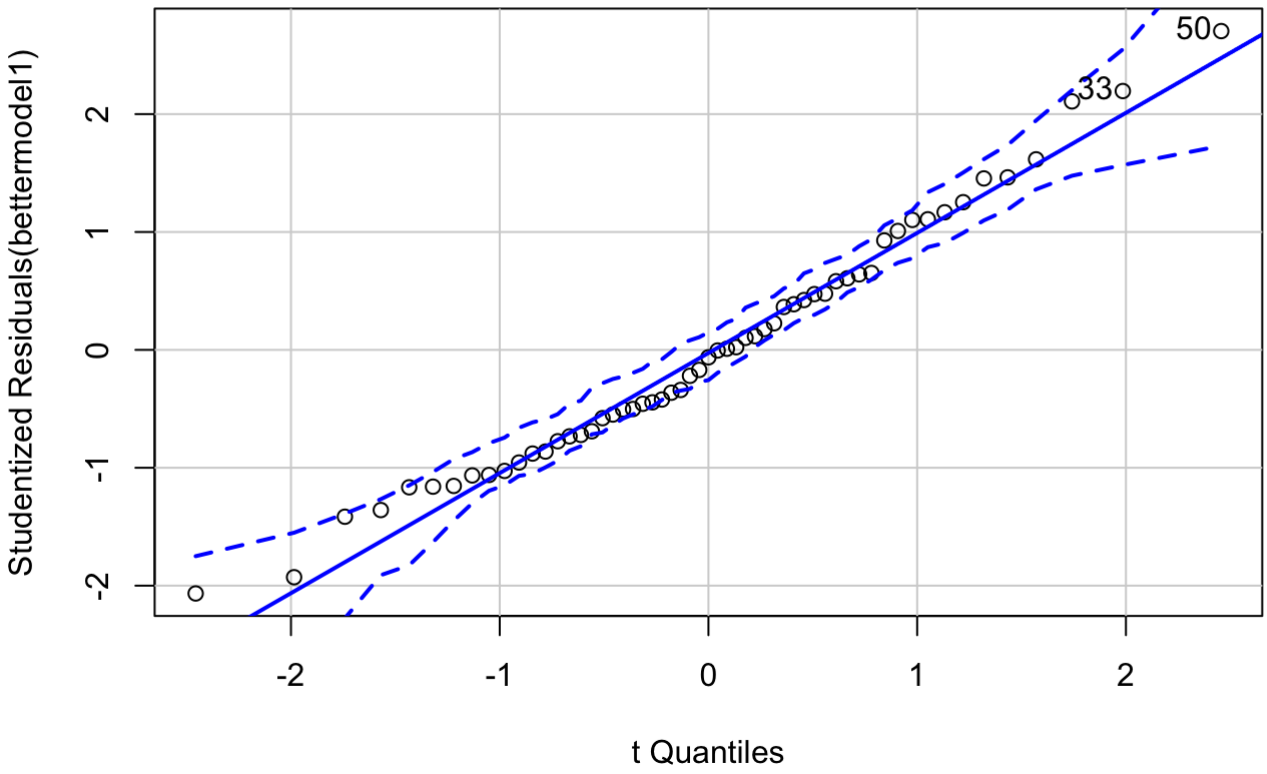



```
#Points Removal
pollution <- pollution[-c(4,20,60),]
pollution
```

```
##                 City    Mort Precip Educ NonWhite NOX SO2   NOX_log  SO2_log
## 1       San Jose, CA  790.73     13 12.2      3.0  32   3 3.4657359 1.098612
## 2        Wichita, KS  823.76     28 12.1      7.5   2   1 0.6931472 0.000000
## 3      San Diego, CA  839.71     10 12.1      5.9  66  20 4.1896547 2.995732
## 5    Minneapolis, MN  857.62     25 12.1      2.0  11  26 2.3978953 3.258097
## 6         Dallas, TX  860.10     35 11.8     14.8   1   1 0.0000000 0.000000
## 7          Miami, FL  861.44     60 11.5     13.5   1   1 0.0000000 0.000000
## 8    Los Angeles, CA  861.83     11 12.1      7.8 319 130 5.7651911 4.867534
## 9   Grand Rapids, MI  871.34     31 10.9      5.1   3  10 1.0986123 2.302585
## 10        Denver, CO  871.77     15 12.2      4.7   8  28 2.0794415 3.332205
## 11     Rochester, NY  874.28     32 11.1      5.0   4  18 1.3862944 2.890372
## 12      Hartford, CT  887.47     43 11.5      7.2   3  10 1.0986123 2.302585
## 13    Fort Worth, TX  891.71     31 11.4     11.5   1   1 0.0000000 0.000000
## 14      Portland, OR  893.99     37 12.0      3.6  21  44 3.0445224 3.784190
## 15     Worcester, MA  895.70     45 11.1      1.0   3   8 1.0986123 2.079442
## 16       Seattle, WA  899.26     35 12.2      5.7   7  20 1.9459101 2.995732
## 17    Bridgeport, CT  899.53     45 10.6      5.3   4   4 1.3862944 1.386294
## 18   Springfield, MA  904.16     45 11.1      3.4   4  20 1.3862944 2.995732
## 19 San Francisco, CA  911.70     18 12.2     13.7 171  86 5.1416636 4.454347
## 21         Utica, NY  912.20     40 10.3      2.5   2  11 0.6931472 2.397895
## 22        Canton, OH  912.35     36 10.7      6.7   7  20 1.9459101 2.995732
## 23   Kansas City, MO  919.73     35 12.0     12.6   4   4 1.3862944 1.386294
## 24         Akron, OH  921.87     36 11.4      8.8  15  59 2.7080502 4.077537
## 25     New Haven, CT  923.23     46 11.3      8.8   3   8 1.0986123 2.079442
## 26     Milwaukee, WI  929.15     30 11.1      5.8  23 125 3.1354942 4.828314
## 27        Boston, MA  934.70     43 12.1      3.5  32  62 3.4657359 4.127134
## 28        Dayton, OH  936.23     36 11.4     12.4   4  16 1.3862944 2.772589
## 29    Providence, RI  938.50     42 10.1      2.2   4  18 1.3862944 2.890372
## 30         Flint, MI  941.18     30 10.8     13.1   4  11 1.3862944 2.397895
## 31       Reading, PA  946.18     41  9.6      2.7  11  89 2.3978953 4.488636
## 32      Syracuse, NY  950.67     38 11.4      3.8   5  25 1.6094379 3.218876
## 33       Houston, TX  952.53     46 11.4     21.0   5   1 1.6094379 0.000000
## 34   Saint Louis, MO  953.56     34  9.7     17.2  15  68 2.7080502 4.219508
## 35    Youngstown, OH  954.44     38 10.7     11.7  13  39 2.5649494 3.663562
## 36      Columbus, OH  958.84     37 11.9     13.1   9  15 2.1972246 2.708050
## 37       Detroit, MI  959.22     31 10.8     15.8  35 124 3.5553481 4.820282
## 38     Nashville, TN  961.01     45 10.1     21.0  14  78 2.6390573 4.356709
## 39     Allentown, PA  962.35     44  9.8      0.8   6  33 1.7917595 3.496508
## 40    Washington, DC  967.80     41 12.3     25.9  28 102 3.3322045 4.624973
## 41  Indianapolis, IN  968.66     39 11.4     15.6   7  33 1.9459101 3.496508
## 42    Cincinnati, OH  970.47     40 10.2     13.0  26 146 3.2580965 4.983607
## 43   Greeensboro, NC  971.12     42 10.4     22.7   3   5 1.0986123 1.609438
## 44         Toledo, OH  972.46     31 10.7      9.5   7  25 1.9459101 3.218876
## 45       Atlanta, GA  982.29     47 11.1     27.1   8  24 2.0794415 3.178054
## 46     Cleveland, OH  985.95     35 11.1     14.7  21  64 3.0445224 4.158883
## 47    Louisville, KY  989.27     30  9.9     13.1  37 193 3.6109179 5.262690
## 48    Pittsburgh, PA  991.29     36 10.6      8.1  59 263 4.0775374 5.572154
## 49      New York, NY  994.65     42 10.7     11.3  26 108 3.2580965 4.682131
## 50        Albany, NY  997.88     35 11.0      3.5  10  39 2.3025851 3.663562
## 51       Buffalo, NY 1001.90     36 10.5      8.1  12  37 2.4849066 3.610918
## 52    Wilmington, DE 1003.50     45 11.3     12.1  11  42 2.3978953 3.737670
## 53       Memphis, TN 1006.49     50 10.4     36.7  18  34 2.8903718 3.526361
## 54  Philadelphia, PA 1015.02     42 10.5     17.5  32 161 3.4657359 5.081404
```

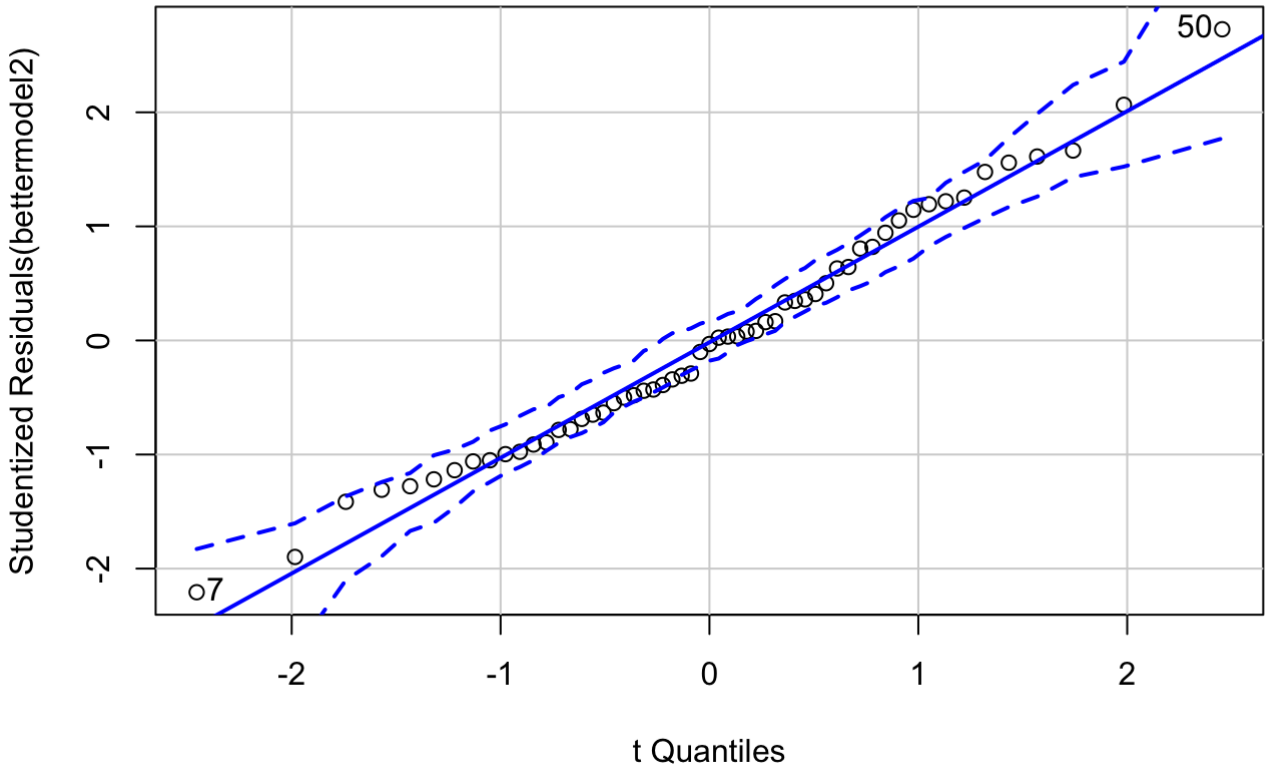
```
## 55   Chattanooga, TN 1017.61     52  9.6     22.2   8  27 2.0794415 3.295837
## 56        Chicago, IL 1024.89     33 10.9     16.3  63 278 4.1431347 5.627621
## 57       Richmond, VA 1025.50     44 11.0     28.6   9  48 2.1972246 3.871201
## 58     Birmingham, AL 1030.38     53 10.2     38.5  32  72 3.4657359 4.276666
## 59      Baltimore, MD 1071.29     43  9.6     24.4  38 206 3.6375862 5.327876
```

```
bettermodel1 <- lm (Mort~., pollution[, c(2:5, 8, 9)])
bettermodel2 <- lm(Mort ~ Precip + Educ + NonWhite + SO2_log, data = pollution[, c(2:5,
8, 9)])
qqPlot(bettermodel1)
```



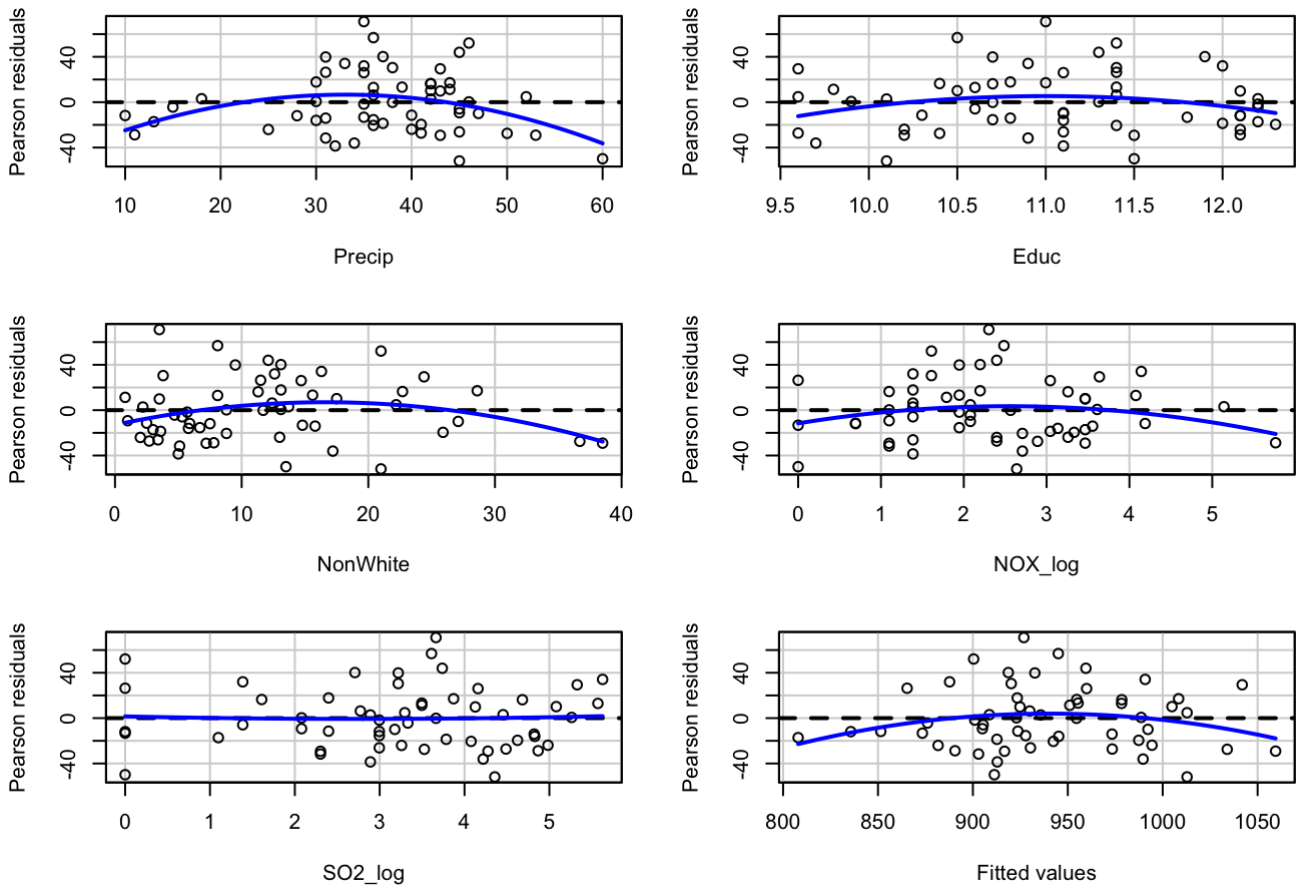

```
## 33 50
## 31 48
```

```
qqPlot(bettermodel2)
```



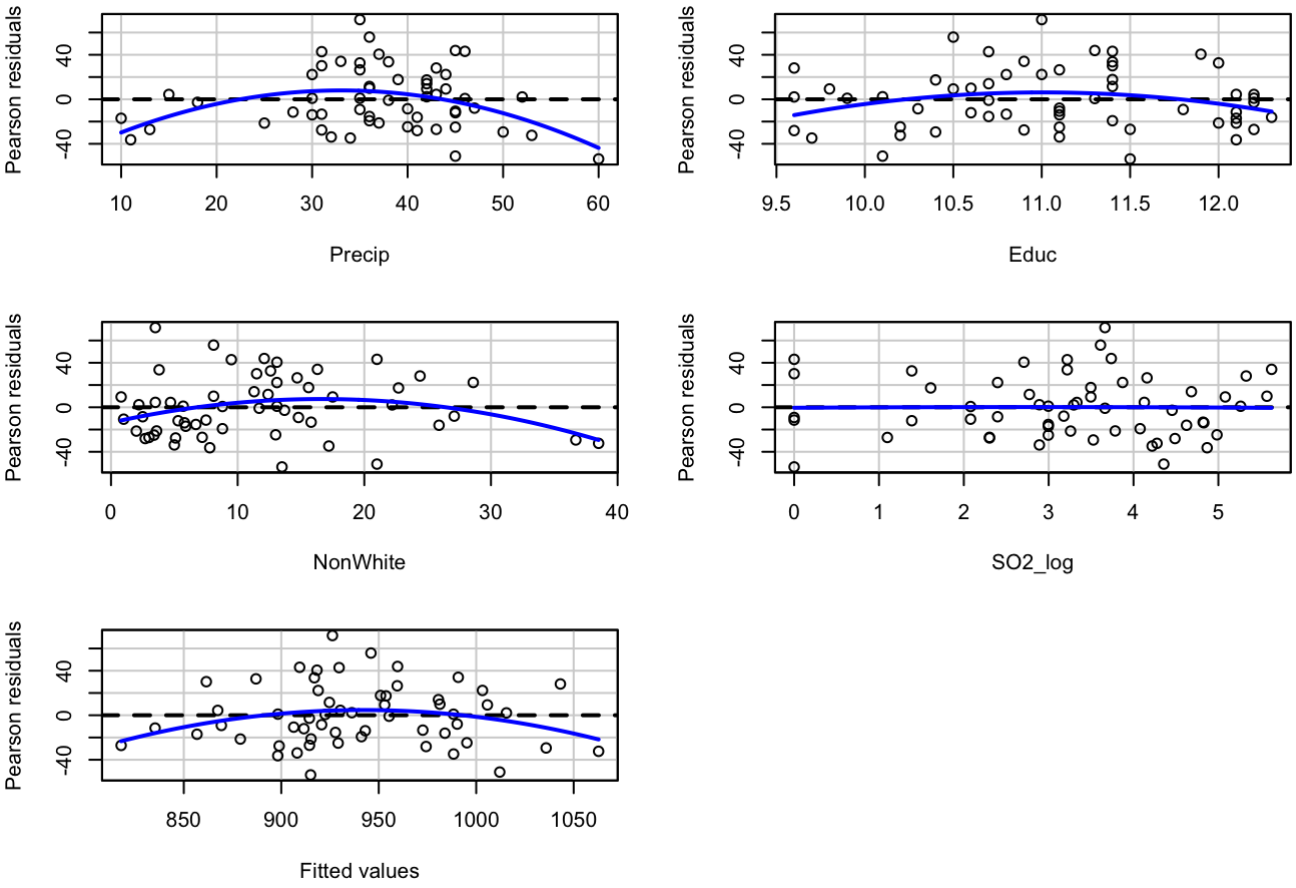

```
##  7 50
##  6 48
```

```
residualPlots(bettermodel1)
```



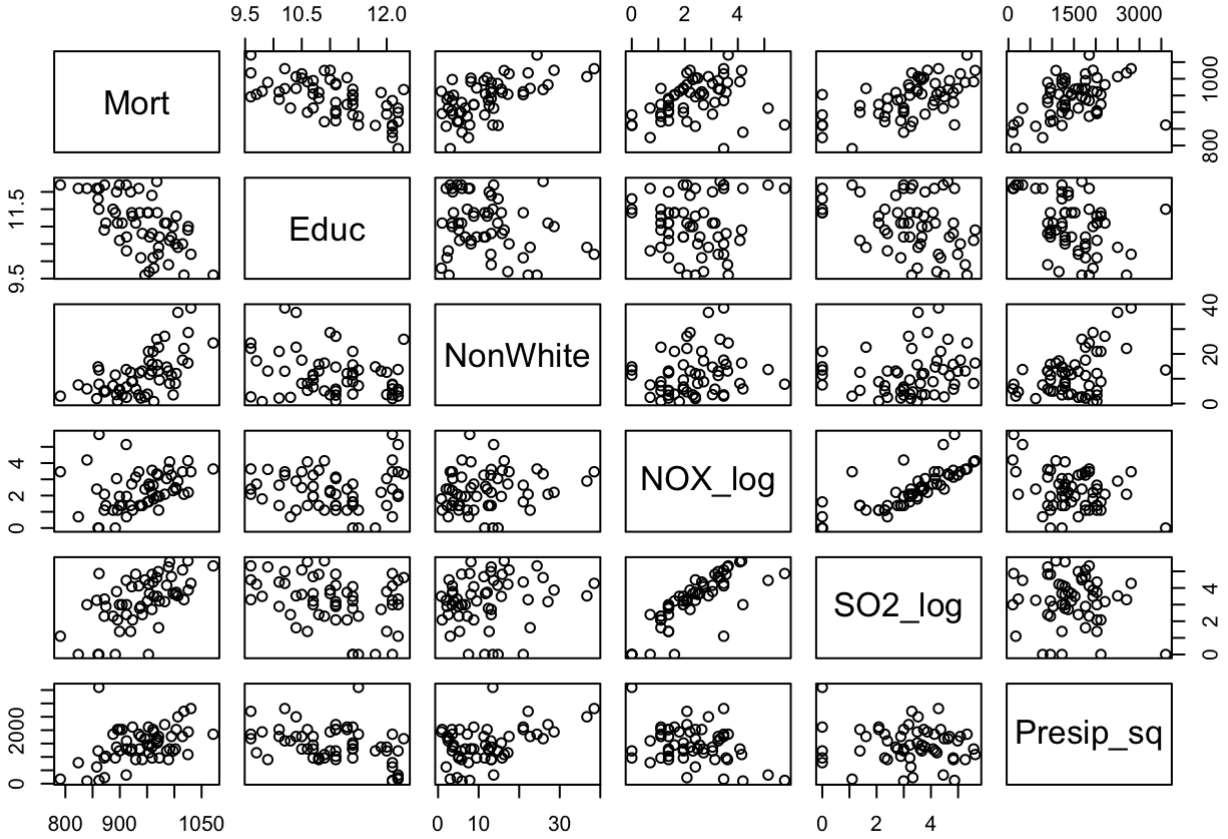

```
##            Test stat Pr(>|Test stat|)
## Precip       -3.2415         0.002119 **
## Educ         -1.6114         0.113392
## NonWhite     -1.9775         0.053512 .
## NOX_log      -1.4967         0.140754
## SO2_log       0.2188         0.827681
## Tukey test   -1.9889         0.046709 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residualPlots(bettermodel2)
```




```
##            Test stat Pr(>|Test stat|)
## Precip       -3.3147         0.001694 **
## Educ         -1.8115         0.075956 .
## NonWhite     -2.0925         0.041388 *
## SO2_log      -0.0603         0.952116
## Tukey test   -2.1475         0.031752 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##check vif, eigenvalue, and codition index to find collinearity problem for the first p
ossible model after points removal and Presip transforming into Precip_sq.
pollution$Presip_sq <- (pollution$Precip)^2
pairs(pollution[,c(2,4,5,8:10)])
```




```
bettermodel_1<- lm(Mort ~., pollution[,c(2,4,5,8:10)])
summary(bettermodel_1)
```

```
##
## Call:
## lm(formula = Mort ~ ., data = pollution[, c(2, 4, 5, 8:10)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.585 -19.416  -0.308  16.481  72.027
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.013e+03  8.098e+01  12.507  < 2e-16 ***
## Educ        -1.649e+01  6.564e+00  -2.512   0.0152 *
## NonWhite     2.742e+00  5.550e-01   4.940 8.80e-06 ***
## NOX_log     -1.170e+01  6.584e+00  -1.778   0.0814 .
## SO2_log      2.597e+01  5.302e+00   4.897 1.02e-05 ***
## Presip_sq    1.358e-02  7.934e-03   1.712   0.0930 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.4 on 51 degrees of freedom
## Multiple R-squared:  0.7659, Adjusted R-squared:  0.743
## F-statistic: 33.37 on 5 and 51 DF,  p-value: 5.891e-15
```

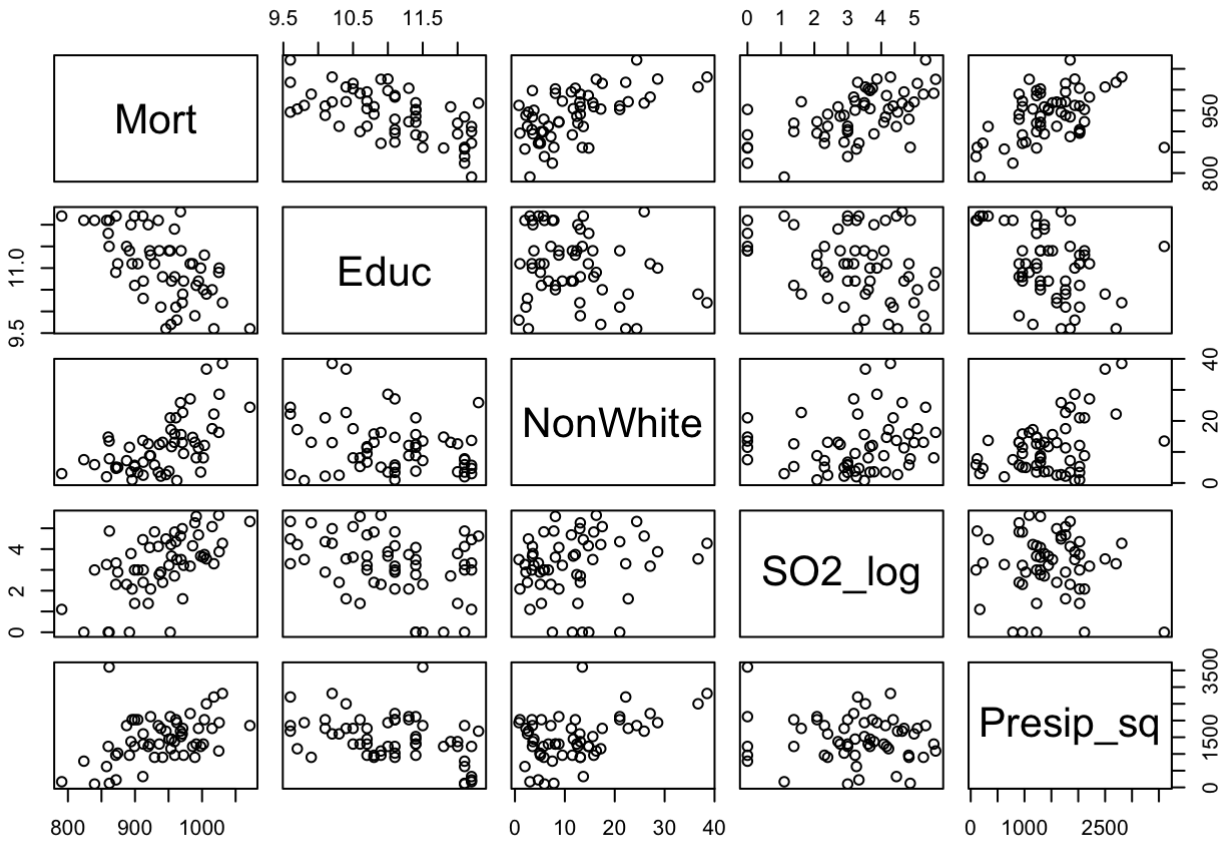
```
ols_vif_tol(bettermodel_1)
```

```
##   Variables Tolerance      VIF
## 1      Educ 0.5886459 1.698814
## 2  NonWhite 0.6707774 1.490808
## 3   NOX_log 0.2422798 4.127460
## 4   SO2_log 0.2531349 3.950463
## 5 Presip_sq 0.5379741 1.858826
```

```
ols_coll_diag(bettermodel_1)
```

```
## Tolerance and Variance Inflation Factor
## ----------------------------------------
##   Variables Tolerance      VIF
## 1      Educ 0.5886459 1.698814
## 2  NonWhite 0.6707774 1.490808
## 3   NOX_log 0.2422798 4.127460
## 4   SO2_log 0.2531349 3.950463
## 5 Presip_sq 0.5379741 1.858826
##
##
## Eigenvalue and Condition Index
## ------------------------------
##    Eigenvalue Condition Index    intercept         Educ    NonWhite
## 1 5.302393764        1.000000 7.894202e-05 9.641781e-05 0.006398777
## 2 0.354901648        3.865289 5.451622e-06 2.231605e-05 0.198575585
## 3 0.229590671        4.805724 1.270923e-03 1.801586e-03 0.478007145
## 4 0.083543085        7.966742 3.095112e-03 7.802246e-03 0.153816492
## 5 0.028296512       13.688931 2.701854e-03 6.908682e-04 0.151707369
## 6 0.001274319       64.505512 9.928477e-01 9.895866e-01 0.011494633
##        NOX_log     SO2_log   Presip_sq
## 1 0.0015725754 0.001375716 0.002825171
## 2 0.0395474843 0.019637170 0.066954248
## 3 0.0226992567 0.004315645 0.052888528
## 4 0.0005352204 0.139427012 0.371924219
## 5 0.8227126548 0.548635558 0.406042011
## 6 0.1129328084 0.286608898 0.099365823
```

```
##check vif, eigenvalue, and codition index to find collinearity problem for the second
 possible model after points removal and Presip transforming into Precip_sq.
pollution$Presip_sq <- (pollution$Precip)^2
pairs(pollution[,c(2,4,5,9:10)])
```




```
bettermodel_2<- lm(Mort ~ Presip_sq + Educ + NonWhite + SO2_log, pollution[,c(2,4,5,9:10
)])
summary(bettermodel_2)
```

```
##
## Call:
## lm(formula = Mort ~ Presip_sq + Educ + NonWhite + SO2_log, data = pollution[,
##     c(2, 4, 5, 9:10)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.310 -20.688  -5.607  19.092  73.633
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.054e+03  7.923e+01  13.300  < 2e-16 ***
## Presip_sq    1.885e-02  7.509e-03   2.511  0.01520 *
## Educ        -2.070e+01  6.247e+00  -3.314  0.00168 **
## NonWhite     2.385e+00  5.281e-01   4.516 3.64e-05 ***
## SO2_log      1.823e+01  3.088e+00   5.902 2.75e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30 on 52 degrees of freedom
## Multiple R-squared:  0.7514, Adjusted R-squared:  0.7323
## F-statistic: 39.29 on 4 and 52 DF,  p-value: 3.937e-15
```

```
ols_vif_tol(bettermodel_2)
```

```
##   Variables Tolerance      VIF
## 1 Presip_sq 0.6254263 1.598909
## 2      Educ 0.6769463 1.477222
## 3  NonWhite 0.7716287 1.295960
## 4   SO2_log 0.7771829 1.286698
```

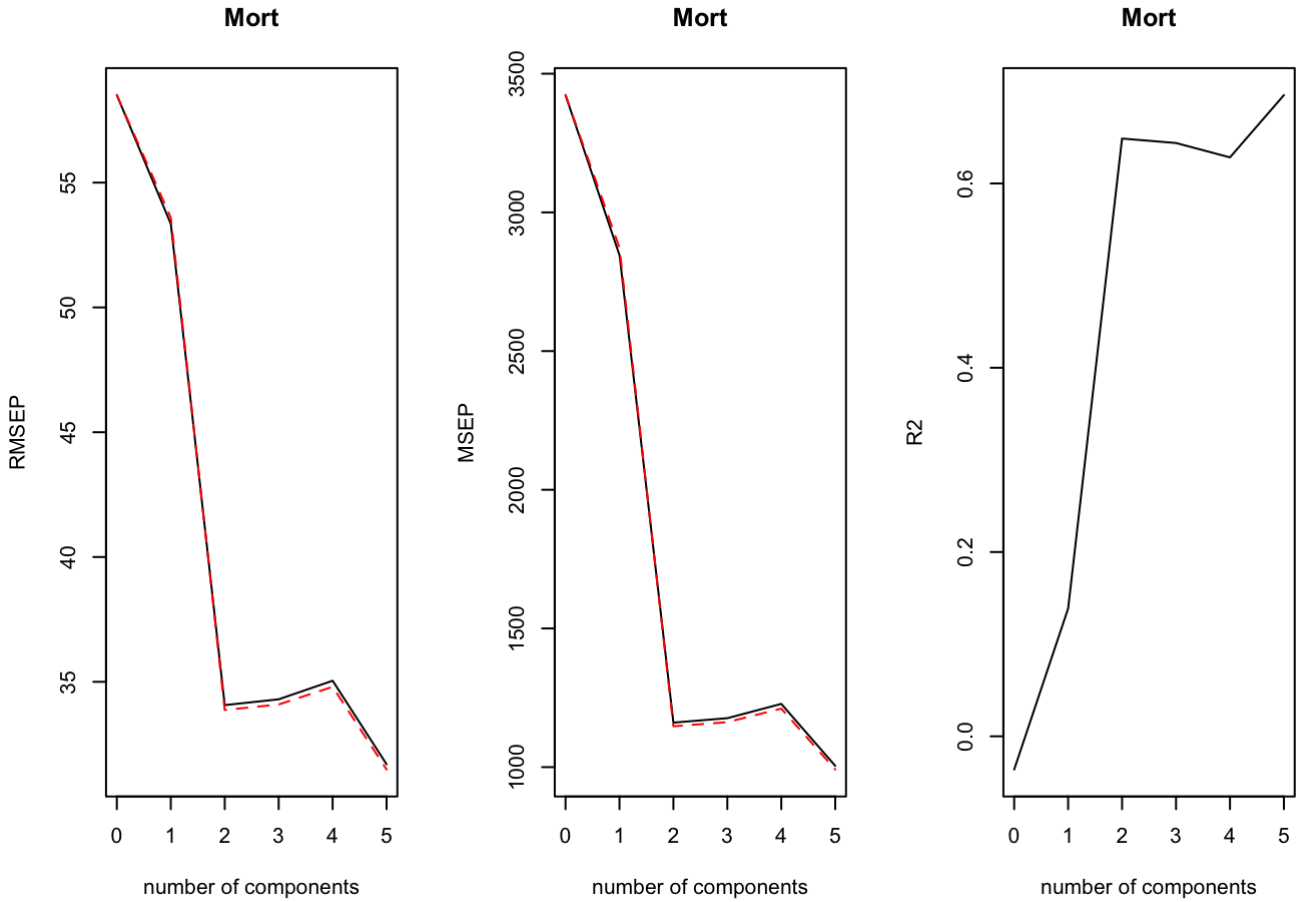
```
ols_coll_diag(bettermodel_2)
```

```
## Tolerance and Variance Inflation Factor
## ------------------------------------
##   Variables Tolerance      VIF
## 1 Presip_sq 0.6254263 1.598909
## 2      Educ 0.6769463 1.477222
## 3  NonWhite 0.7716287 1.295960
## 4   SO2_log 0.7771829 1.286698
##
##
## Eigenvalue and Condition Index
## ------------------------------
##    Eigenvalue Condition Index    intercept   Presip_sq         Educ
## 1 4.463401312        1.000000 0.0001216637 0.004860998 0.0001567461
## 2 0.281541278        3.981639 0.0006883095 0.019380809 0.0011117663
## 3 0.170218842        5.120698 0.0004028114 0.245559417 0.0006550279
## 4 0.083409693        7.315175 0.0034590381 0.473144042 0.0090845624
## 5 0.001428876       55.890207 0.9953281773 0.257054735 0.9889918973
##       NonWhite     SO2_log
## 1 0.0106104939 0.005722843
## 2 0.6016758737 0.065892079
## 3 0.1959893486 0.331351224
## 4 0.1914335831 0.366342804
## 5 0.0002907008 0.230691049
```

```
#final model 1
###Model selection using pcr to fix collinearity problem
pcr_model1 <- pcr(Mort~., data=pollution[,c(2,4,5,8:10)], scale = T, center=TRUE, valida
tion = "CV")
summary(pcr_model1)
```

```
## Data:    X dimension: 57 5
##  Y dimension: 57 1
## Fit method: svdpc
## Number of components considered: 5
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
## CV           58.51    53.35    34.07    34.30    35.04    31.70
## adjCV        58.51    53.60    33.87    34.09    34.80    31.49
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps
## X       38.90    74.12    90.16    97.53   100.00
## Mort    41.83    72.24    72.28    72.29    76.59
```

```
### Determine how many PCs to keep in the model
par(mfrow=c(1,3))
# Plot the root mean squared error
validationplot(pcr_model1)
# Plot the cross validation MSE
validationplot(pcr_model1, val.type="MSEP")
# Plot the R2
validationplot(pcr_model1, val.type = "R2")
```




```
bestmodel1 <- pcr(Mort~., data = pollution[, c(2,4,5,8:10)], scale = T,
                  center=TRUE, ncomp = 2)
coef(bestmodel1, ncomp = 2, intercept = TRUE)
```

```
## , , 2 comps
##
##                    Mort
## (Intercept) 1104.69859
## Educ         -19.52407
## NonWhite      18.68695
## NOX_log        9.56547
## SO2_log       15.85861
## Presip_sq     15.18117
```

```
#Final model 2
#Selecting model by using forward, backward, and stepwise
n=dim(pollution)[1]
## Get the full and the null models
fitnull <- lm(Mort ~ 1, data=pollution[, c(2,4,5,9:10)])
## Null model
fitall1 <- lm(Mort~., data=pollution[, c(2,4,5,9:10)])
# Foward
forwardmodel1=step(fitnull, fitall1, direction = "forward", k = log(n))
```

```
## Start:  AIC=465.91
## Mort ~ 1
```

```
summary(forwardmodel1)
```

```
##
## Call:
## lm(formula = Mort ~ 1, data = pollution[, c(2, 4, 5, 9:10)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -148.787  -40.257    6.663   42.773  131.773
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  939.517      7.681   122.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.99 on 56 degrees of freedom
```

```
# Backword
backwardmodel2=step(fitall1, direction = "backward", k = log(n))
```

```
## Start:  AIC=402.74
## Mort ~ Educ + NonWhite + SO2_log + Presip_sq
##
##             Df Sum of Sq   RSS    AIC
## <none>                   46814 402.74
## - Presip_sq  1    5674.5 52489 405.21
## - Educ       1    9885.5 56700 409.61
## - NonWhite   1   18364.3 65179 417.56
## - SO2_log    1   31359.2 78174 427.92
```

```
summary(backwardmodel2)
```

```
##
## Call:
## lm(formula = Mort ~ Educ + NonWhite + SO2_log + Presip_sq, data = pollution[,
##     c(2, 4, 5, 9:10)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.310 -20.688  -5.607  19.092  73.633
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.054e+03  7.923e+01  13.300  < 2e-16 ***
## Educ        -2.070e+01  6.247e+00  -3.314  0.00168 **
## NonWhite     2.385e+00  5.281e-01   4.516 3.64e-05 ***
## SO2_log      1.823e+01  3.088e+00   5.902 2.75e-07 ***
## Presip_sq    1.885e-02  7.509e-03   2.511  0.01520 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30 on 52 degrees of freedom
## Multiple R-squared:  0.7514, Adjusted R-squared:  0.7323
## F-statistic: 39.29 on 4 and 52 DF,  p-value: 3.937e-15
```

```
# Step
stepmodel3=step(fitall1, direction = "both", k = log(n))
```

```
## Start:  AIC=402.74
## Mort ~ Educ + NonWhite + SO2_log + Presip_sq
##
##             Df Sum of Sq   RSS    AIC
## <none>                   46814 402.74
## - Presip_sq  1    5674.5 52489 405.21
## - Educ       1    9885.5 56700 409.61
## - NonWhite   1   18364.3 65179 417.56
## - SO2_log    1   31359.2 78174 427.92
```

```
summary(stepmodel3)
```

```
##
## Call:
## lm(formula = Mort ~ Educ + NonWhite + SO2_log + Presip_sq, data = pollution[,
##     c(2, 4, 5, 9:10)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.310 -20.688  -5.607  19.092  73.633
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.054e+03  7.923e+01  13.300  < 2e-16 ***
## Educ        -2.070e+01  6.247e+00  -3.314  0.00168 **
## NonWhite     2.385e+00  5.281e-01   4.516 3.64e-05 ***
## SO2_log      1.823e+01  3.088e+00   5.902 2.75e-07 ***
## Presip_sq    1.885e-02  7.509e-03   2.511  0.01520 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30 on 52 degrees of freedom
## Multiple R-squared:  0.7514, Adjusted R-squared:  0.7323
## F-statistic: 39.29 on 4 and 52 DF,  p-value: 3.937e-15
```

###Mort ~ Educ + NonWhite + SO2_log + Presip_sq

```
myleaps1 <- regsubsets(Mort~., data = pollution[, c(2,4,5, 8:10)], nbest = 6)
(myleaps1.summary <- summary(myleaps1))
```

```
## Subset selection object
## Call: regsubsets.formula(Mort ~ ., data = pollution[, c(2, 4, 5, 8:10)],
##     nbest = 6)
## 5 Variables  (and intercept)
##           Forced in Forced out
## Educ          FALSE      FALSE
## NonWhite      FALSE      FALSE
## NOX_log       FALSE      FALSE
## SO2_log       FALSE      FALSE
## Presip_sq     FALSE      FALSE
## 6 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          Educ NonWhite NOX_log SO2_log Presip_sq
## 1  ( 1 ) "*"  " "      " "     " "     " "
## 1  ( 2 ) " "  "*"      " "     " "     " "
## 1  ( 3 ) " "  " "      " "     "*"     " "
## 1  ( 4 ) " "  " "      " "     " "     "*"
## 1  ( 5 ) " "  " "      "*"     " "     " "
## 2  ( 1 ) " "  "*"      " "     "*"     " "
## 2  ( 2 ) " "  " "      " "     "*"     "*"
## 2  ( 3 ) "*"  "*"      " "     " "     " "
## 2  ( 4 ) "*"  " "      " "     "*"     " "
## 2  ( 5 ) "*"  " "      "*"     " "     " "
## 2  ( 6 ) " "  " "      "*"     "*"     " "
## 3  ( 1 ) "*"  "*"      " "     "*"     " "
## 3  ( 2 ) " "  "*"      "*"     "*"     " "
## 3  ( 3 ) " "  "*"      " "     "*"     "*"
## 3  ( 4 ) "*"  " "      " "     "*"     "*"
## 3  ( 5 ) "*"  "*"      "*"     " "     " "
## 3  ( 6 ) " "  " "      "*"     "*"     "*"
## 4  ( 1 ) "*"  "*"      "*"     "*"     " "
## 4  ( 2 ) "*"  "*"      " "     "*"     "*"
## 4  ( 3 ) " "  "*"      "*"     "*"     "*"
## 4  ( 4 ) "*"  "*"      "*"     " "     "*"
## 4  ( 5 ) "*"  " "      "*"     "*"     "*"
## 5  ( 1 ) "*"  "*"      "*"     "*"     "*"
```

```
bettertable1 <- cbind(myleaps1.summary$which,
                    myleaps1.summary$rsq, myleaps1.summary$rss,
                    myleaps1.summary$adjr2, myleaps1.summary$cp,myleaps1.summary$bic)
dimnames(bettertable1)[[2]] <- c(dimnames(myleaps1.summary$which)[[2]],
                                 "R2", "sse", "R2_ADJ", "CP", "BIC")
show(bettertable1)
```

```
##   (Intercept) Educ NonWhite NOX_log SO2_log Presip_sq         R2       sse
## 1           1    1        0       0       0         0 0.37893702 116958.23
## 1           1    0        1       0       0         0 0.35121059 122179.66
## 1           1    0        0       0       1         0 0.34652008 123062.97
## 1           1    0        0       0       0         1 0.19461570 151669.52
## 1           1    0        0       1       0         0 0.08354933 172585.47
## 2           1    0        1       0       1         0 0.60357087  74655.31
## 2           1    0        0       0       1         1 0.60297811  74766.94
## 2           1    1        1       0       0         0 0.58364156  78408.39
## 2           1    1        0       0       1         0 0.54325217  86014.49
## 2           1    1        0       1       0         0 0.46701832 100370.81
## 2           1    0        0       1       1         0 0.42539189 108209.88
## 3           1    1        1       0       1         0 0.72127725  52488.91
## 3           1    0        1       1       1         0 0.71001843  54609.17
## 3           1    0        1       0       1         1 0.69891663  56699.85
## 3           1    1        0       0       1         1 0.65389299  65178.68
## 3           1    1        1       1       0         0 0.63066683  69552.61
## 3           1    0        0       1       1         1 0.60888017  73655.47
## 4           1    1        1       1       1         0 0.75246622  46615.42
## 4           1    1        1       0       1         1 0.75140963  46814.40
## 4           1    0        1       1       1         1 0.73696077  49535.40
## 4           1    1        1       1       0         1 0.65583460  64813.03
## 4           1    1        0       1       1         1 0.65389337  65178.61
## 5           1    1        1       1       1         1 0.76591309  44083.11
##       R2_ADJ          CP        BIC
## 1 0.36764497  82.309626 -19.064297
## 1 0.33941442  88.350323 -16.574782
## 1 0.33463862  89.372235 -16.164175
## 1 0.17997234 122.467309  -4.250733
## 1 0.06688659 146.665095   3.113022
## 2 0.58888831  35.369144 -40.610551
## 2 0.58827360  35.498286 -40.525387
## 2 0.56822088  39.711098 -37.814746
## 2 0.52633558  48.510647 -32.537405
## 2 0.44727826  65.119547 -23.739136
## 2 0.40411010  74.188608 -19.452666
## 3 0.70550049  11.724714 -56.647444
## 3 0.69360438  14.177648 -54.390255
## 3 0.68187417  16.596373 -52.248775
## 3 0.63430203  26.405573 -44.305210
## 3 0.60976118  31.465806 -40.602995
## 3 0.58674131  36.212418 -37.336049
## 4 0.73342516   6.929640 -59.368613
## 4 0.73228730   7.159837 -59.125828
## 4 0.71672698  10.307780 -55.905512
## 4 0.62936034  27.982559 -40.582820
## 4 0.62726978  28.405492 -40.262220
## 5 0.74296340   6.000000 -58.509274
```

###Mort ~ Educ + NonWhite + SO2_log + Presip_sq

```
#Cross-validation
library(DAAG)
```

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

```
##
## Attaching package: 'DAAG'
```

```
## The following object is masked from 'package:car':
##
##     vif
```

```
mod1="Mort~ Educ +NonWhite + NOX_log +SO2_log"
mod2="Mort~ Educ +NonWhite + NOX_log +SO2_log+ Presip_sq"
mod3="Mort~Educ +NonWhite + SO2_log+ Presip_sq"
mod1cv=CVlm(data = pollution, form.lm = formula(mod1), m=8, seed = 7,plotit=F) #ms = 991
```

```
## Analysis of Variance Table
##
## Response: Mort
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Educ       1  71361   71361   79.60 4.6e-12 ***
## NonWhite   1  38550   38550   43.00 2.5e-08 ***
## NOX_log    1   8856    8856    9.88  0.0028 **
## SO2_log    1  22937   22937   25.59 5.6e-06 ***
## Residuals 52  46615     896
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## fold 1
## Observations in test set: 7
##                10    12    13    24    36    41   57
## Predicted   901.1 910.4 881.2 939.9 915.1 958.1 1014
## cvpred      903.1 911.1 879.4 941.1 914.7 958.1 1012
## Mort        871.8 887.5 891.7 921.9 958.8 968.7 1026
## CV residual -31.3 -23.6  12.3 -19.3  44.2  10.6   14
##
## Sum of squares = 4321    Mean square = 617    n = 7
##
## fold 2
## Observations in test set: 8
##                15    33     38    43     52   54     56   59
## Predicted   892.15 886.1 1012.9 962.6  948.2 1000  992.8 1044
## cvpred      891.51 868.8 1009.1 953.0  947.2  998  991.0 1038
## Mort        895.70 952.5  961.0 971.1 1003.5 1015 1024.9 1071
## CV residual   4.19  83.7  -48.1  18.2   56.3   17   33.9   33
##
## Sum of squares = 15365    Mean square = 1921    n = 8
##
## fold 3
## Observations in test set: 7
##               1     7   9     16      47     51     58
## Predicted   813 885.7 915 897.33 1000.45  946.0 1051.8
## cvpred      822 892.1 916 897.48  999.24  946.6 1055.9
## Mort        791 861.4 871 899.26  989.27 1001.9 1030.4
## CV residual -31 -30.7 -45   1.78   -9.97   55.3  -25.5
##
## Sum of squares = 7744    Mean square = 1106    n = 7
##
## fold 4
## Observations in test set: 7
##                 2    11    22     28     30    40    48
## Predicted   843.8 922.4 929.5 937.15 940.84 982.3 971.8
## cvpred      849.1 925.9 931.0 941.48 943.57 986.1 970.2
## Mort        823.8 874.3 912.4 936.23 941.18 967.8 991.3
## CV residual -25.3 -51.6 -18.6  -5.25  -2.39 -18.3  21.1
##
## Sum of squares = 4461    Mean square = 637    n = 7
##
```

```
## fold 5
## Observations in test set: 7
##                 5     17    18    27    32  37    50
## Predicted   887.3 892.25 920.1 898.7 918.1 981 925.9
## cvpred      882.3 890.24 915.7 893.9 913.5 979 921.5
## Mort        857.6 899.53 904.2 934.7 950.7 959 997.9
## CV residual -24.6   9.29 -11.5  40.8  37.2 -20  76.4
##
## Sum of squares = 10107    Mean square = 1444    n = 7
##
## fold 6
## Observations in test set: 7
##              25    26     29     34    42    44    46
## Predicted   913 949.6 932.72 1003.6 992.4 944.6 961.5
## cvpred      912 951.0 935.21 1008.0 996.1 945.8 962.7
## Mort        923 929.1 938.50  953.6 970.5 972.5 986.0
## CV residual  11 -21.9   3.29  -54.4 -25.6  26.6  23.3
##
## Sum of squares = 5483    Mean square = 783    n = 7
##
## fold 7
## Observations in test set: 7
##                 14      19    23    39     45    49     55
## Predicted   898.33 911.673 888.6 944.0 990.09 969.1 1006.5
## cvpred      897.73 911.591 887.9 940.2 989.83 967.5 1003.7
## Mort        893.99 911.700 919.7 962.4 982.29 994.6 1017.6
## CV residual  -3.74   0.109  31.8  22.1  -7.54  27.1   13.9
##
## Sum of squares = 2498    Mean square = 357    n = 7
##
## fold 8
## Observations in test set: 7
##                 3     6     8    21  31     35   53
## Predicted   864.2 884.1 895.9 927.5 971 953.81 1031
## cvpred      879.0 884.3 912.9 926.0 973 957.50 1036
## Mort        839.7 860.1 861.8 912.2 946 954.44 1006
## CV residual -39.3 -24.2 -51.1 -13.8 -27  -3.06  -29
##
## Sum of squares = 6513    Mean square = 930    n = 7
##
## Overall (Sum over all 7 folds)
##  ms
## 991
```

```
mod2cv=CVlm(data = pollution, form.lm = formula(mod2), m=8, seed = 7,plotit=F) #ms = 100
1
```

```
## Analysis of Variance Table
##
## Response: Mort
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Educ       1  71361   71361   82.56 3.0e-12 ***
## NonWhite   1  38550   38550   44.60 1.8e-08 ***
## NOX_log    1   8856    8856   10.25  0.0024 **
## SO2_log    1  22937   22937   26.54 4.2e-06 ***
## Presip_sq  1   2532    2532    2.93  0.0930 .
## Residuals 51  44083     864
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## fold 1
## Observations in test set: 7
##                10    12    13    24    36    41     57
## Predicted   889.8 915.0 869.4 940.8 915.7 956.3 1011.0
## cvpred      888.6 914.8 864.4 941.3 914.0 954.6 1007.0
## Mort        871.8 887.5 891.7 921.9 958.8 968.7 1025.5
## CV residual -16.9 -27.3  27.3 -19.4  44.8  14.1   18.5
##
## Sum of squares = 4706    Mean square = 672    n = 7
##
## fold 2
## Observations in test set: 8
##                 15    33     38    43   52     54     56     59
## Predicted   901.17 892.3 1013.6 956.5  956 1003.0  990.2 1042.3
## cvpred      899.04 874.3 1009.5 948.4  954  999.8  988.8 1036.7
## Mort        895.70 952.5  961.0 971.1 1004 1015.0 1024.9 1071.3
## CV residual  -3.34  78.3  -48.5  22.7   50   15.2   36.1   34.6
##
## Sum of squares = 14225    Mean square = 1778    n = 8
##
## fold 3
## Observations in test set: 7
##                 1     7     9     16     47     51     58
## Predicted   810.1 909.1 907.1 898.94 992.11  944.2 1058.8
## cvpred      816.6 939.2 902.5 901.31 981.36  942.0 1071.7
## Mort        790.7 861.4 871.3 899.26 989.27 1001.9 1030.4
## CV residual -25.8 -77.7 -31.2  -2.05   7.91   59.9  -41.3
##
## Sum of squares = 13041    Mean square = 1863    n = 7
##
## fold 4
## Observations in test set: 7
##                 2    11  22       28     30    40  48
## Predicted   836.4 916.2 927 932.2134 928.92 984.9 975
## cvpred      840.9 919.6 928 936.3291 932.27 988.7 973
## Mort        823.8 874.3 912 936.2300 941.18 967.8 991
## CV residual -17.1 -45.3 -16  -0.0991   8.91 -20.9  18
##
## Sum of squares = 3445    Mean square = 492    n = 7
```

```
##
## fold 5
## Observations in test set: 7
##                 5     17    18  27  32    37    50
## Predicted   883.8 899.84 928.2 915 920 974.7 925.9
## cvpred      880.7 896.92 923.0 908 916 974.7 922.4
## Mort        857.6 899.53 904.2 935 951 959.2 997.9
## CV residual -23.1   2.61 -18.8  27  35 -15.5  75.5
##
## Sum of squares = 8791    Mean square = 1256    n = 7
##
## fold 6
## Observations in test set: 7
##                 25    26     29    34    42    44  46
## Predicted   920.50 946.6 935.09 993.6 993.3 936.3 959
## cvpred      919.07 947.8 936.84 997.6 996.0 937.7 960
## Mort        923.23 929.1 938.50 953.6 970.5 972.5 986
## CV residual   4.16 -18.6   1.66 -44.1 -25.6  34.8  26
##
## Sum of squares = 4844    Mean square = 692    n = 7
##
## fold 7
## Observations in test set: 7
##                14  19    23    39     45    49      55
## Predicted   906.0 909 885.9 949.5 992.28 974.8 1013.35
## cvpred      905.4 909 884.9 947.3 992.03 973.9 1011.63
## Mort        894.0 912 919.7 962.4 982.29 994.6 1017.61
## CV residual -11.4   3  34.8  15.1  -9.74  20.8    5.98
##
## Sum of squares = 2138    Mean square = 305    n = 7
##
## fold 8
## Observations in test set: 7
##                3     6     8    21    31     35     53
## Predicted   860 875.5 895.3 925.7 973.2 953.18 1033.6
## cvpred      874 875.3 911.6 924.4 975.6 956.88 1038.1
## Mort        840 860.1 861.8 912.2 946.2 954.44 1006.5
## CV residual -34 -15.2 -49.8 -12.2 -29.4  -2.44  -31.6
##
## Sum of squares = 5881    Mean square = 840    n = 7
##
## Overall (Sum over all 7 folds)
##   ms
## 1001
```

```
mod3cv=CVlm(data = pollution, form.lm = formula(mod3), m=8, seed = 7,plotit=F) #ms = 1009
```

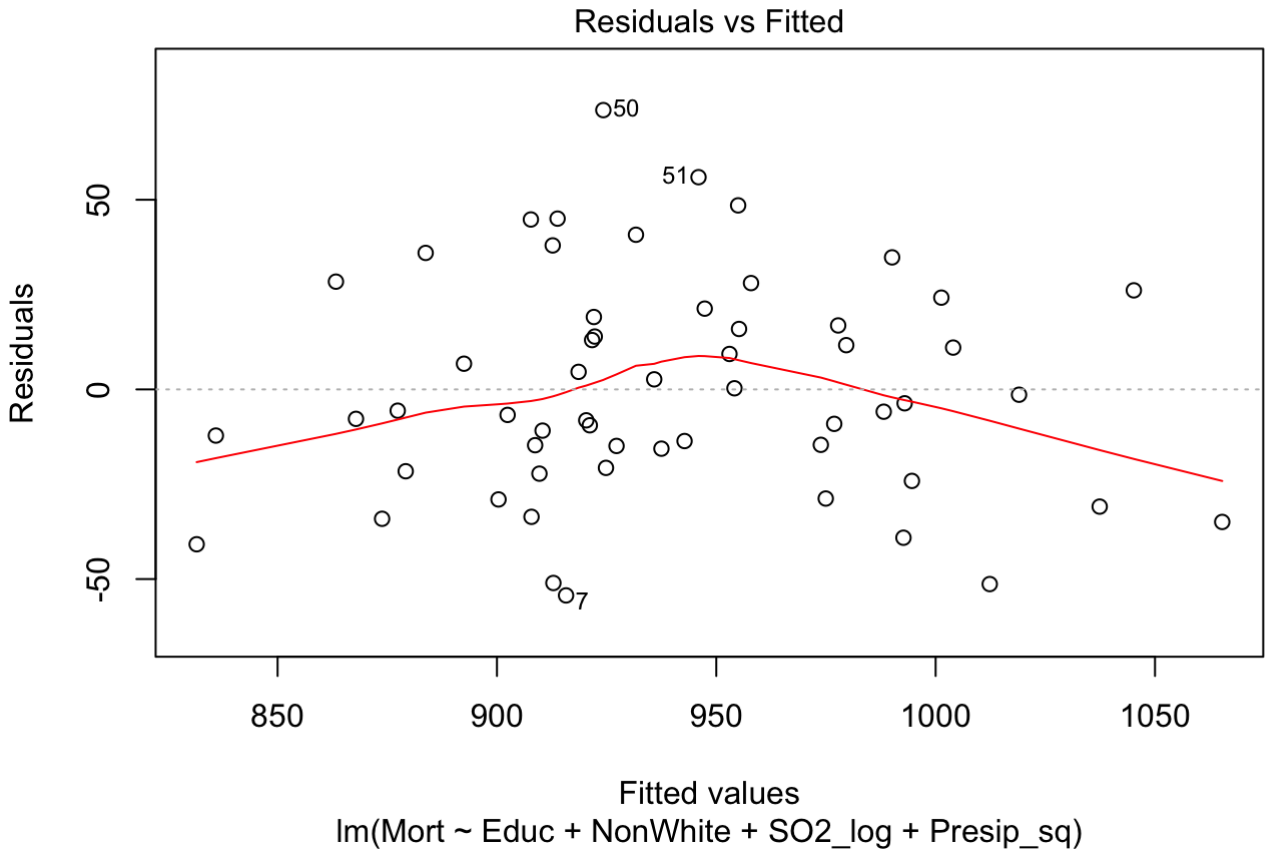
```
## Analysis of Variance Table
##
## Response: Mort
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Educ       1  71361   71361    79.3 4.9e-12 ***
## NonWhite   1  38550   38550    42.8 2.6e-08 ***
## SO2_log    1  25919   25919    28.8 1.9e-06 ***
## Presip_sq  1   5675    5675     6.3   0.015 *
## Residuals 52  46814     900
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## fold 1
## Observations in test set: 7
##                 10    12    13    24    36    41     57
## Predicted   877.38 909.7 863.3 937.5 913.8 947.4 1001.3
## cvpred      874.81 908.8 856.8 937.1 910.4 944.7  996.2
## Mort        871.77 887.5 891.7 921.9 958.8 968.7 1025.5
## CV residual  -3.04 -21.3  34.9 -15.3  48.4  23.9   29.3
##
## Sum of squares = 5689    Mean square = 813    n = 7
##
## fold 2
## Observations in test set: 8
##                 15    33     38    43     52     54     56     59
## Predicted   902.42 907.7 1012.3 955.2  955.0 1004.0  990.1 1045.2
## cvpred      901.53 900.8 1008.3 949.4  952.0 1001.0  987.7 1041.1
## Mort        895.70 952.5  961.0 971.1 1003.5 1015.0 1024.9 1071.3
## CV residual  -5.83  51.8  -47.2  21.7   51.5   14.1   37.2   30.2
##
## Sum of squares = 10566    Mean square = 1321    n = 8
##
## fold 3
## Observations in test set: 7
##                 1     7     9     16     47     51     58
## Predicted   831.6 915.7 900.4 892.48 992.93  945.9 1065.3
## cvpred      831.2 944.7 898.7 897.67 981.44  943.2 1076.3
## Mort        790.7 861.4 871.3 899.26 989.27 1001.9 1030.4
## CV residual -40.4 -83.2 -27.4   1.59   7.83   58.7  -45.9
##
## Sum of squares = 14936    Mean square = 2134    n = 7
##
## fold 4
## Observations in test set: 7
##                 2    11    22    28    30     40    48
## Predicted   835.9 907.9 927.3 922.3 922.1 976.88 979.6
## cvpred      838.3 908.8 927.9 923.2 922.8 977.05 979.2
## Mort        823.8 874.3 912.4 936.2 941.2 967.80 991.3
## CV residual -14.6 -34.5 -15.5  13.1  18.4  -9.25  12.1
##
## Sum of squares = 2385    Mean square = 341    n = 7
##
```

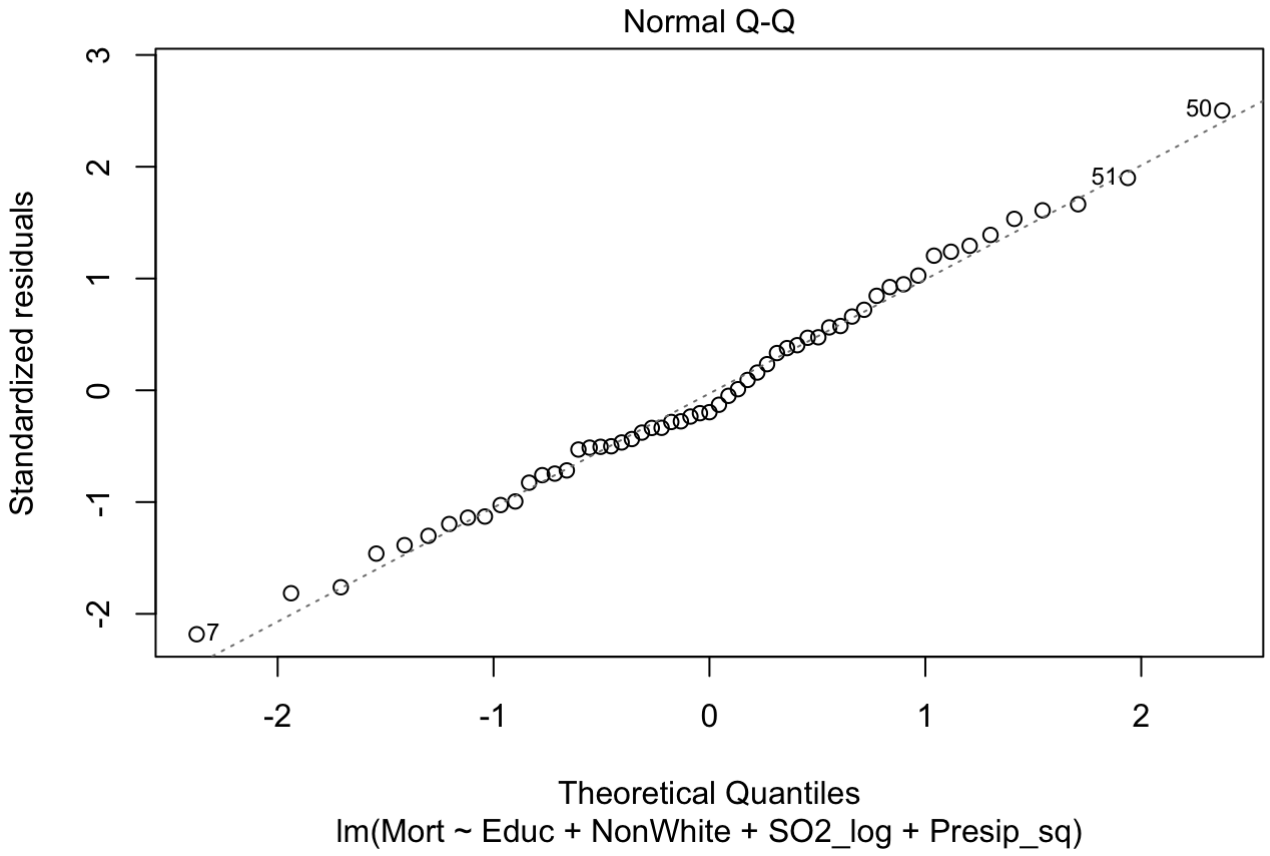
```
## fold 5
## Observations in test set: 7
##                5     17    18    27    32    37    50
## Predicted   879 910.40 924.8 921.7 912.7 973.8 924.2
## cvpred      876 908.27 920.1 915.1 908.7 973.4 920.7
## Mort        858 899.53 904.2 934.7 950.7 959.2 997.9
## CV residual -18  -8.74 -15.9  19.6  41.9 -14.1  77.1
##
## Sum of squares = 8950    Mean square = 1279    n = 7
##
## fold 6
## Observations in test set: 7
##                  25    26     29    34    42    44    46
## Predicted   918.60 942.8 935.85 992.7 994.6 931.7 957.9
## cvpred      917.24 943.7 937.29 996.3 997.0 932.9 958.6
## Mort        923.23 929.1 938.50 953.6 970.5 972.5 986.0
## CV residual   5.99 -14.6   1.21 -42.7 -26.5  39.5  27.3
##
## Sum of squares = 5090    Mean square = 727    n = 7
##
## fold 7
## Observations in test set: 7
##               14    19  23  39    45    49       55
## Predicted   909 921.2 884 953 988.2 977.8 1019.009
## cvpred      909 922.3 883 951 987.9 977.8 1017.493
## Mort        894 911.7 920 962 982.3 994.6 1017.610
## CV residual -15 -10.6  37  11  -5.6  16.8    0.117
##
## Sum of squares = 2142    Mean square = 306    n = 7
##
## fold 8
## Observations in test set: 7
##                 3     6     8     21    31     35     53
## Predicted   873.8 867.9 912.9 920.35 975.0 954.14 1037.4
## cvpred      882.6 871.4 922.7 920.97 976.0 957.59 1039.7
## Mort        839.7 860.1 861.8 912.20 946.2 954.44 1006.5
## CV residual -42.9 -11.3 -60.9  -8.77 -29.8  -3.15  -33.2
##
## Sum of squares = 7752    Mean square = 1107    n = 7
##
## Overall (Sum over all 7 folds)
##   ms
## 1009
```

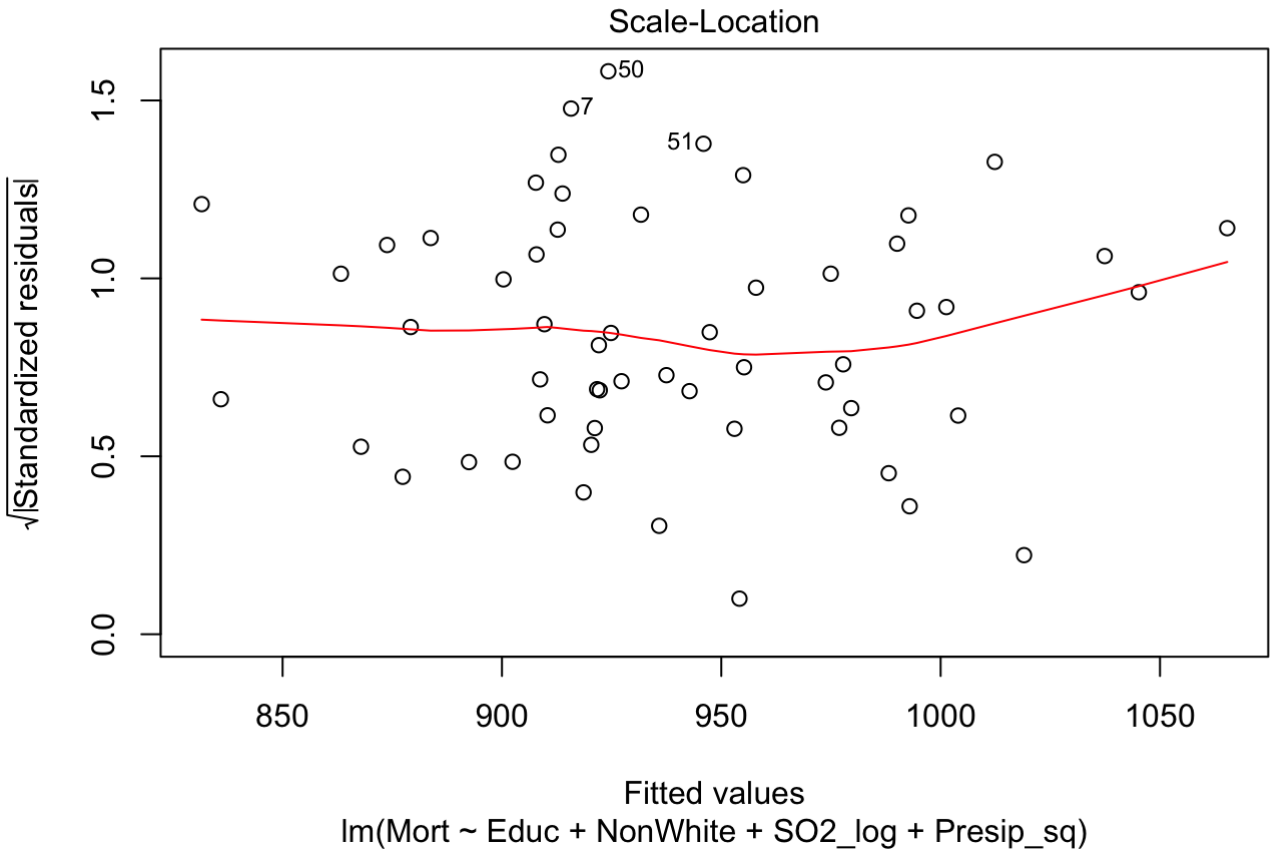
```
#So we choose the model1 that is being selected by backward selection and step-wise sele
ction and model2 which is the same as the model selected through PCR since they both hav
e small msep value.
```

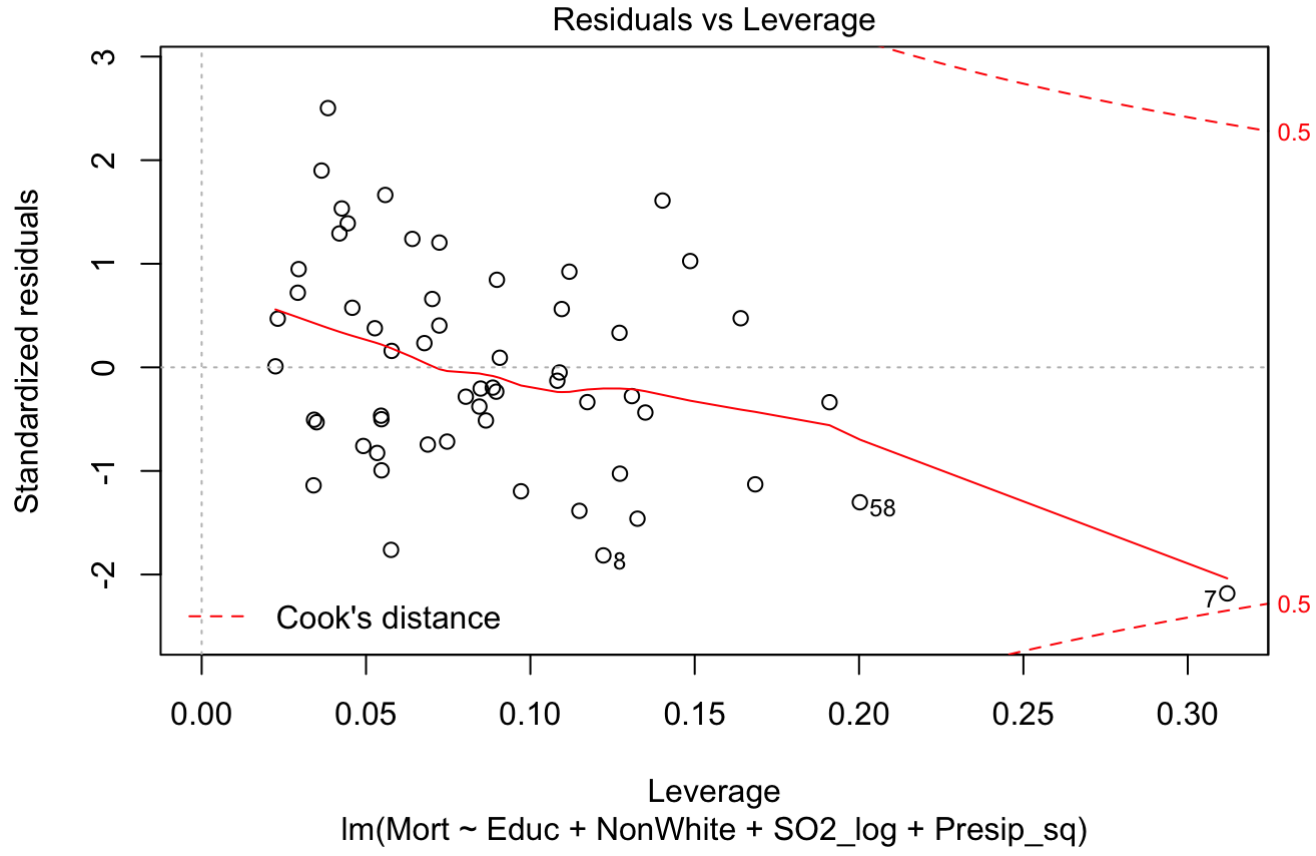
Educ + NonWhite + SO2_log + Presip_sq

```
finalmodel1= lm(Mort~ Educ +NonWhite +SO2_log + Presip_sq, data = pollution)
finalmodel2=lm(Mort~ Educ +NonWhite + NOX_log +SO2_log+ Presip_sq, data = pollution)
plot(finalmodel1)
```

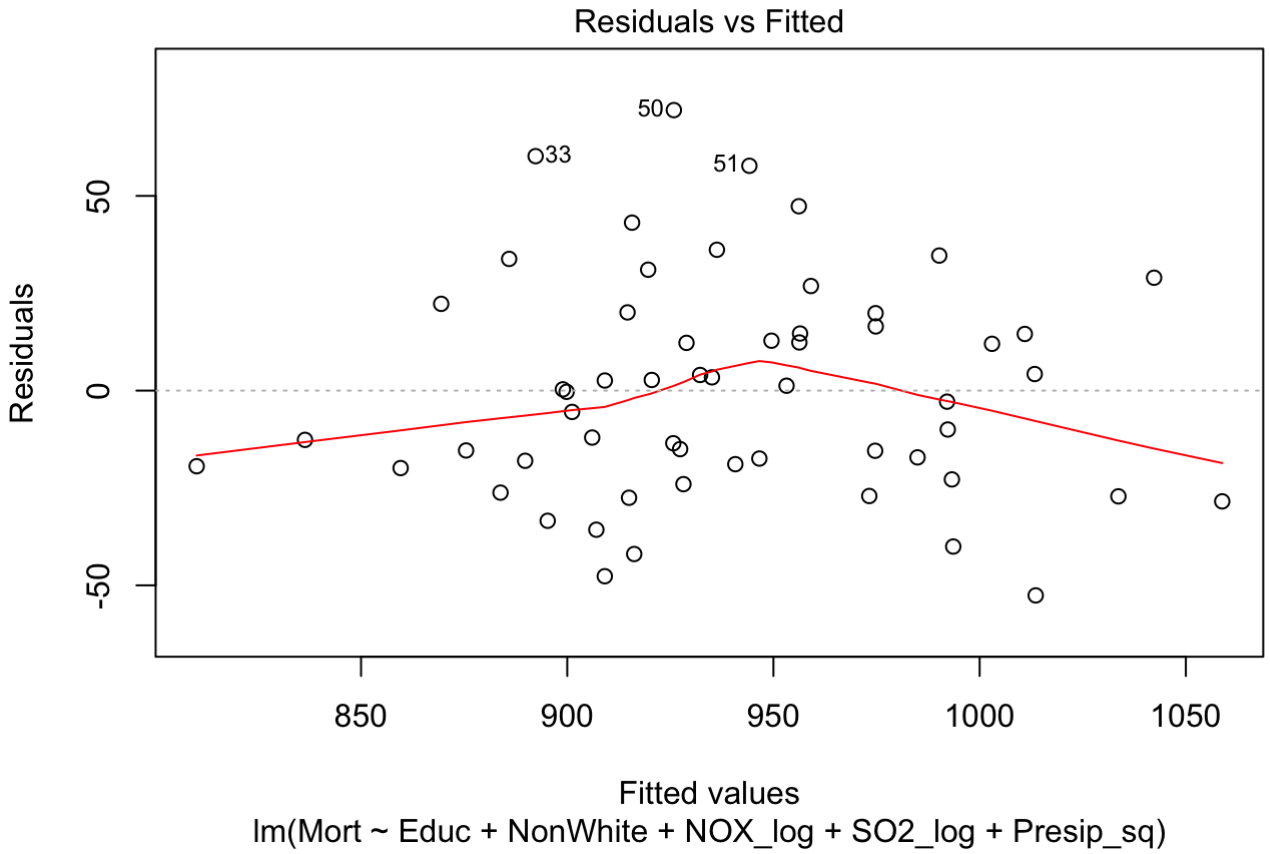

Residuals vs Fitted
Residuals
Fitted values
lm(Mort ~ Educ + NonWhite + SO2_log + Presip_sq)
50
51
7
850
900
950
1000
1050
-50
0
50

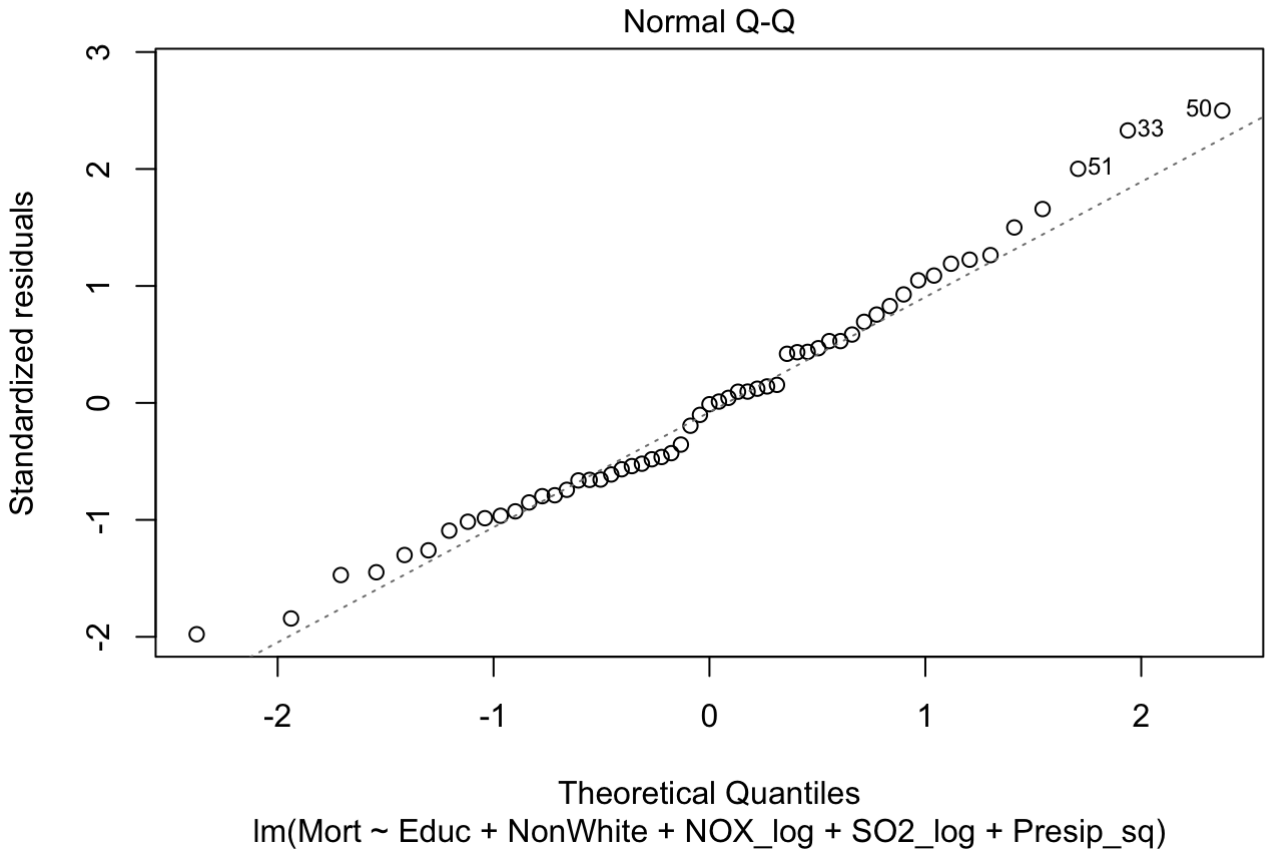

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Mort ~ Educ + NonWhite + SO2_log + Presip_sq)
50
51
7
-2
-1
0
1
2
3

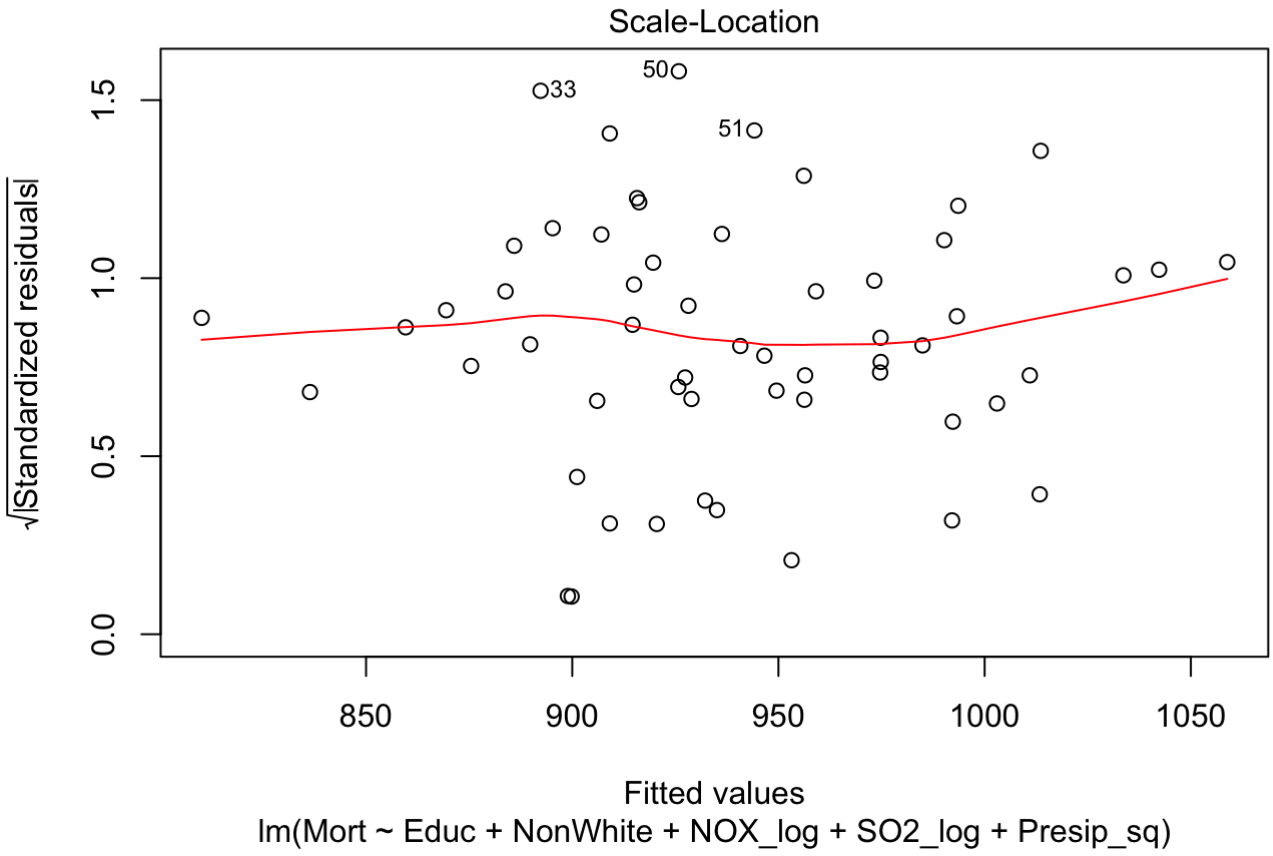

Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
850
900
950
1000
1050
50
7
51
Fitted values
lm(Mort ~ Educ + NonWhite + SO2_log + Presip_sq)

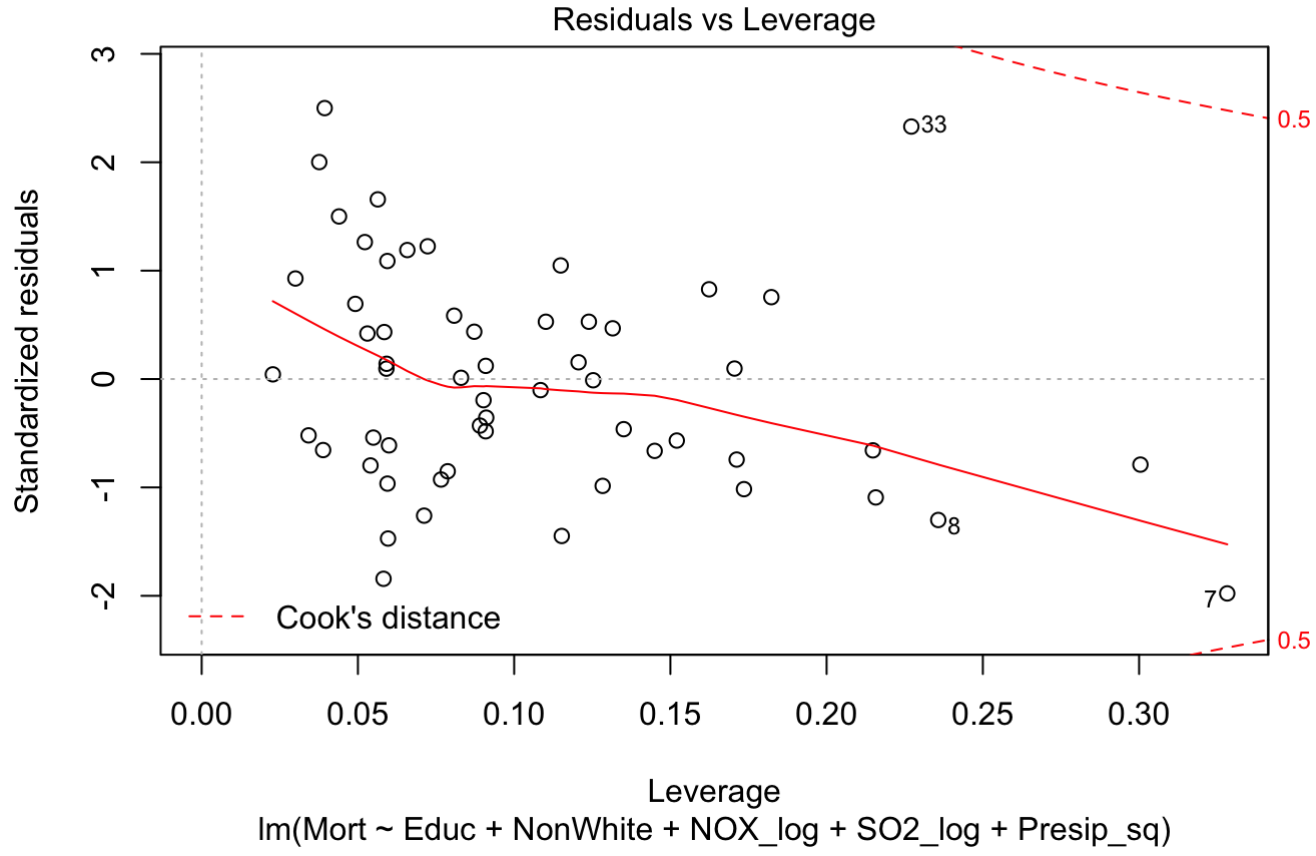

Residuals vs Leverage
Standardized residuals
-2
-1
0
1
2
3
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.5
0.5
58
8
7
Cook's distance
Leverage
lm(Mort ~ Educ + NonWhite + SO2_log + Presip_sq)

```
plot(finalmodel2)
```

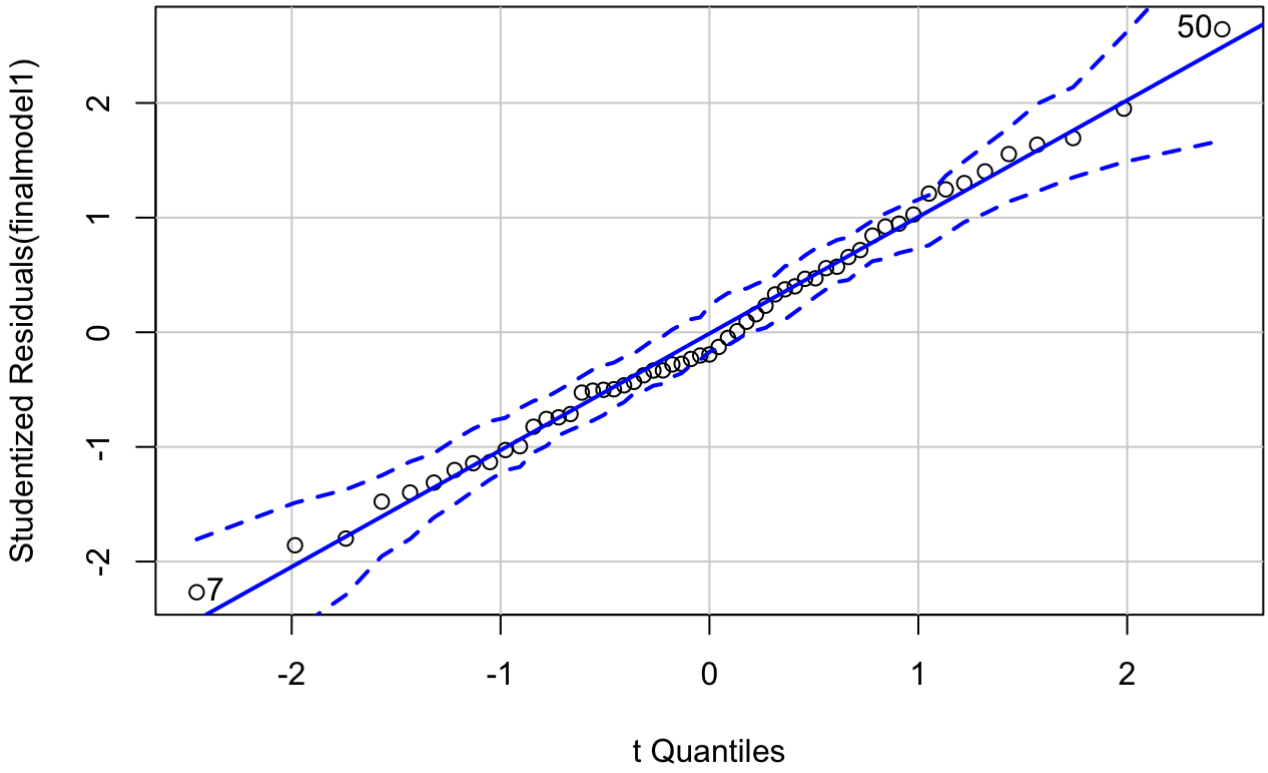

Residuals vs Fitted
50
33
51
Residuals
50
0
-50
850
900
950
1000
1050
Fitted values
lm(Mort ~ Educ + NonWhite + NOX_log + SO2_log + Presip_sq)


Normal Q-Q
50
33
51
Standardized residuals
3
2
1
0
-1
-2
-2
-1
0
1
2
Theoretical Quantiles
lm(Mort ~ Educ + NonWhite + NOX_log + SO2_log + Presip_sq)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
33
50
51
850
900
950
1000
1050
Fitted values
lm(Mort ~ Educ + NonWhite + NOX_log + SO2_log + Presip_sq)

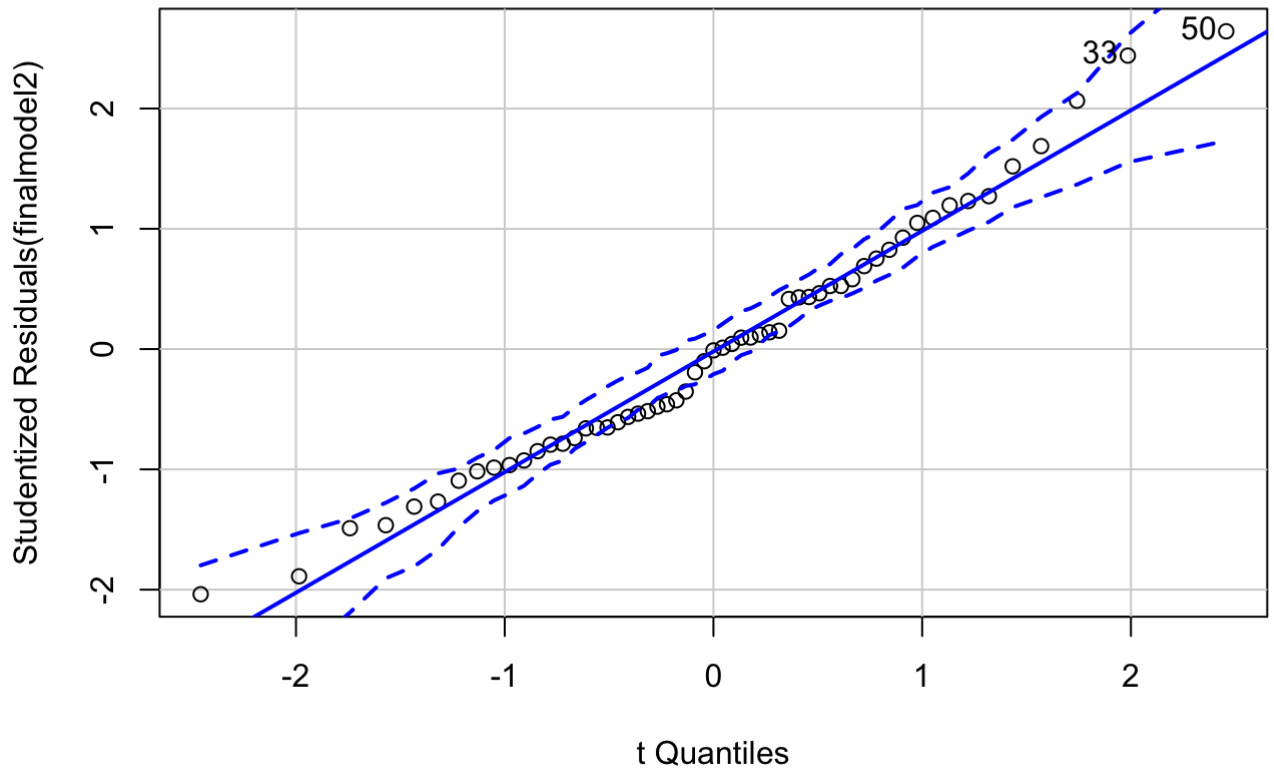

Residuals vs Leverage
Standardized residuals
-2
-1
0
1
2
3
33
8
7
0.5
0.5
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(Mort ~ Educ + NonWhite + NOX_log + SO2_log + Presip_sq)

```
qqPlot(finalmodel1)
```



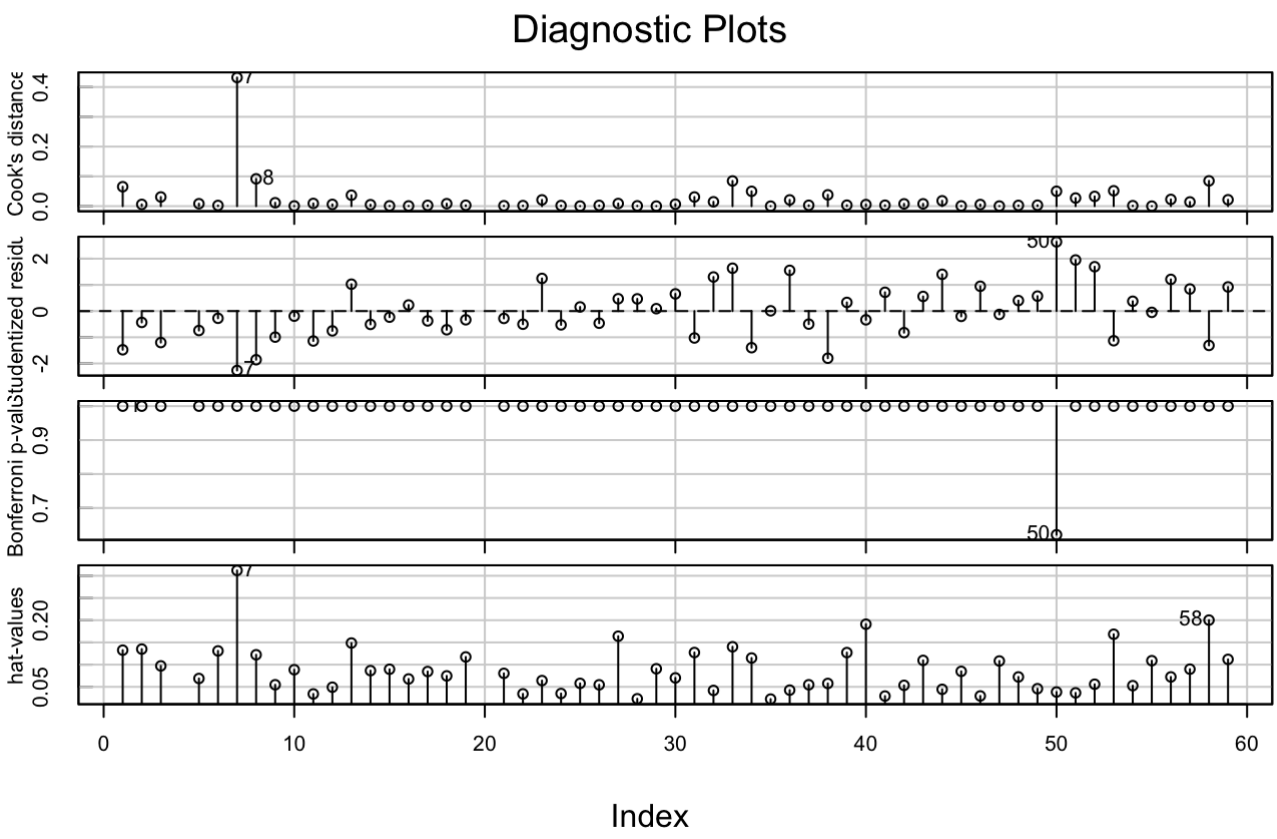

```
##  7 50
##  6 48
```

```
qqPlot(finalmodel2)
```

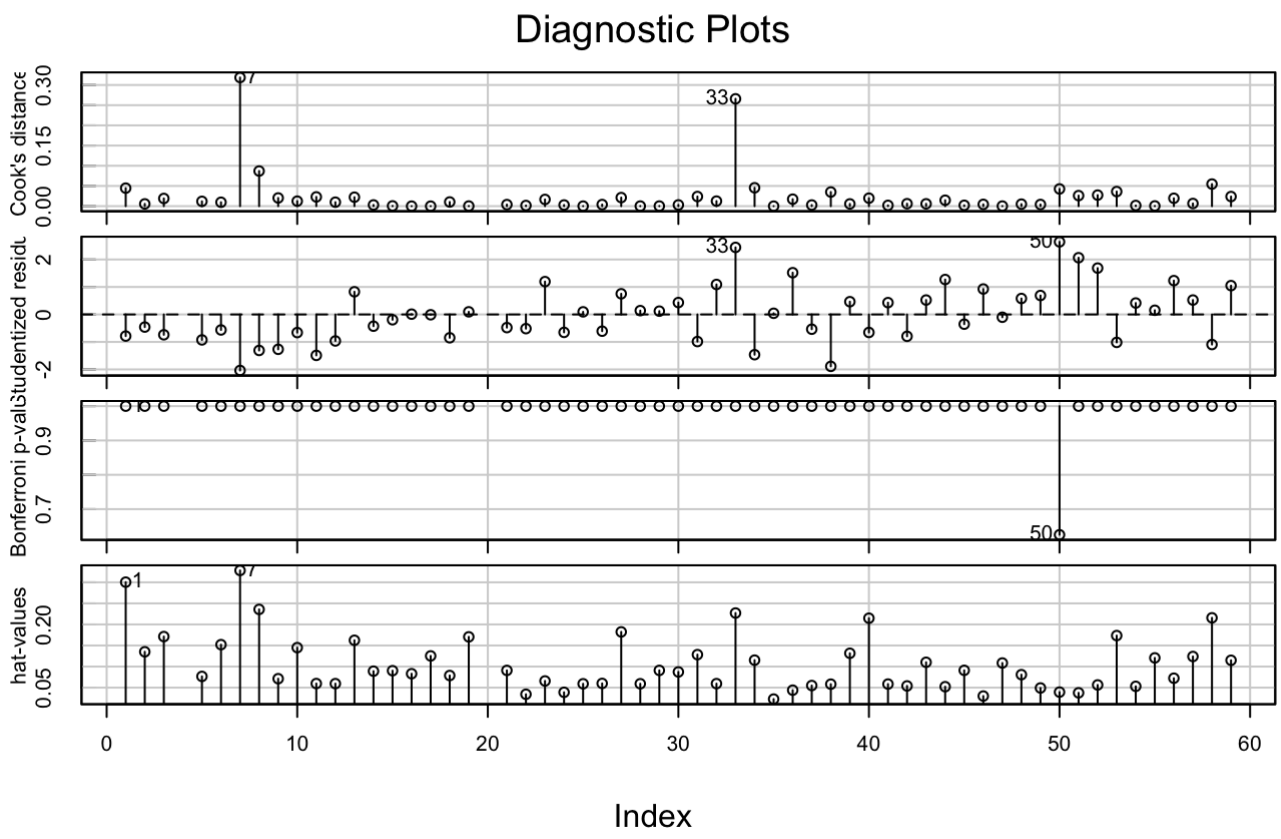



```
## 33 50
## 31 48
```

```
#Find influential point(s) to drop
influenceIndexPlot(finalmodel1)
```




```
influenceIndexPlot(finalmodel2)
```




```
###Final two models
#(1)lm(formula = 1104.70 -19.53*Educ +18.69*NonWhite + 9.57*Nox_Log +16.86*SO2_log +15.1
8*Presip_sq)
#(2)lm(formula = Mort ~ Educ + NonWhite + SO2_log + Presip_sq)
```