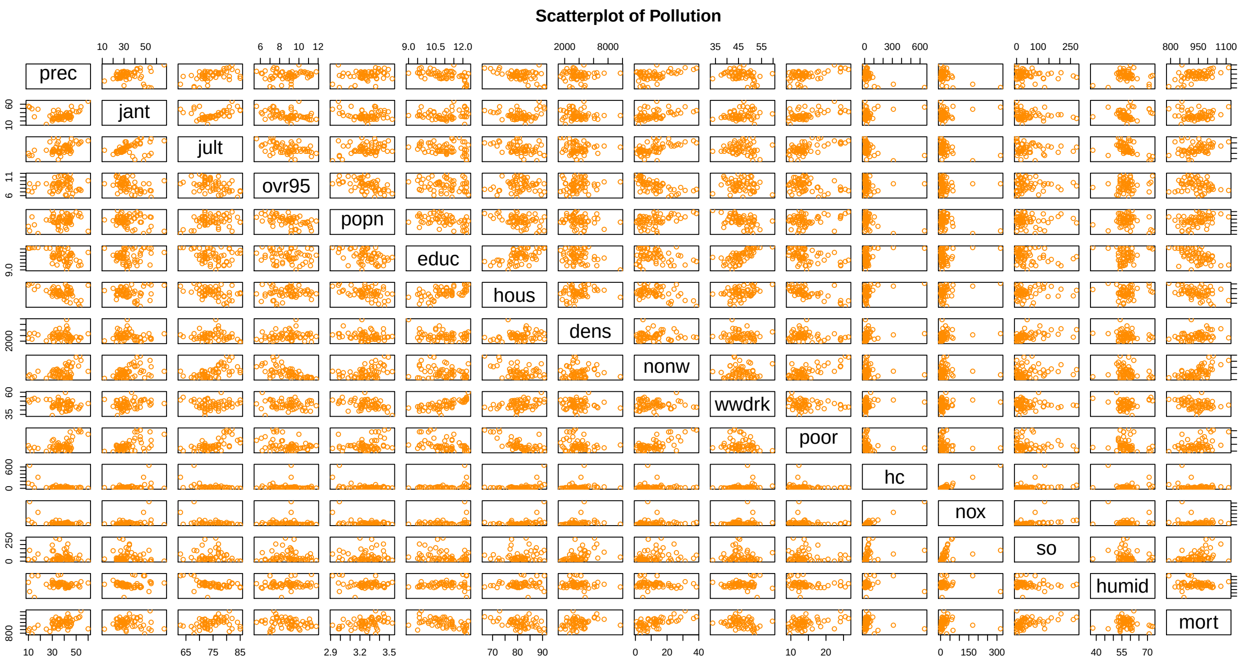
HW3-Exercise 5

Zining Fan(zf2234), Mutian Wang(mw3386), Siyuan Wang(sw3418)

1.

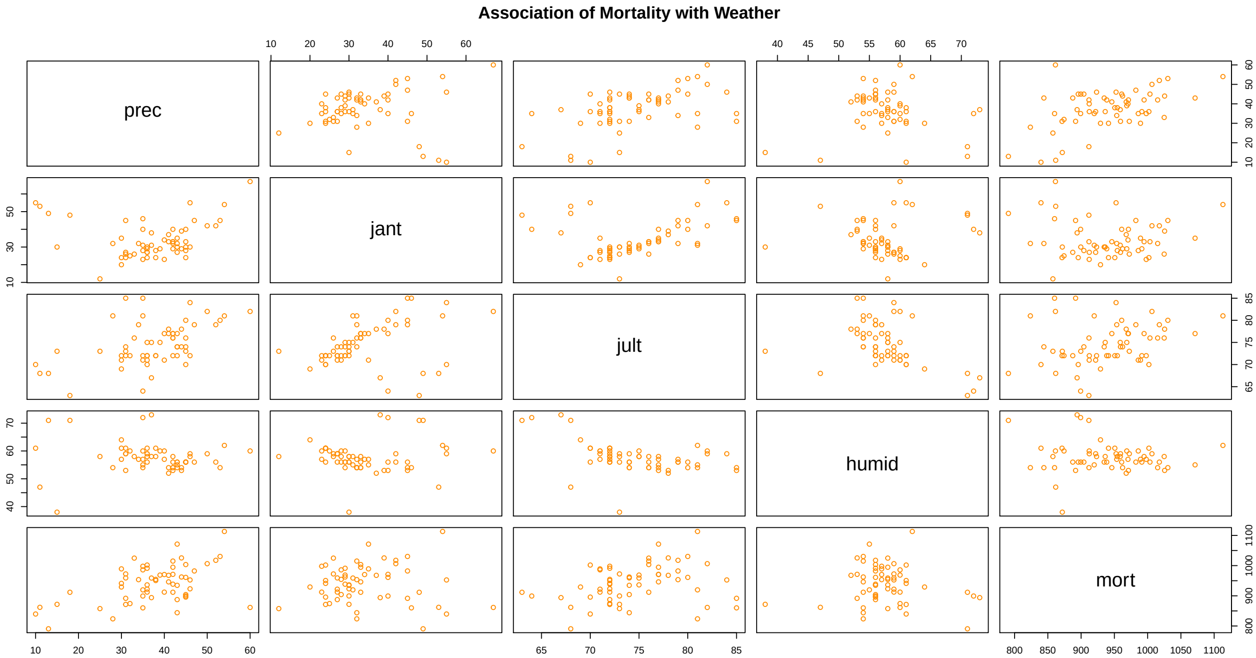
1)

The scatterplot of data pollution with 16 variables.



2)

The scatterplot which shows the association of mortality with weather.

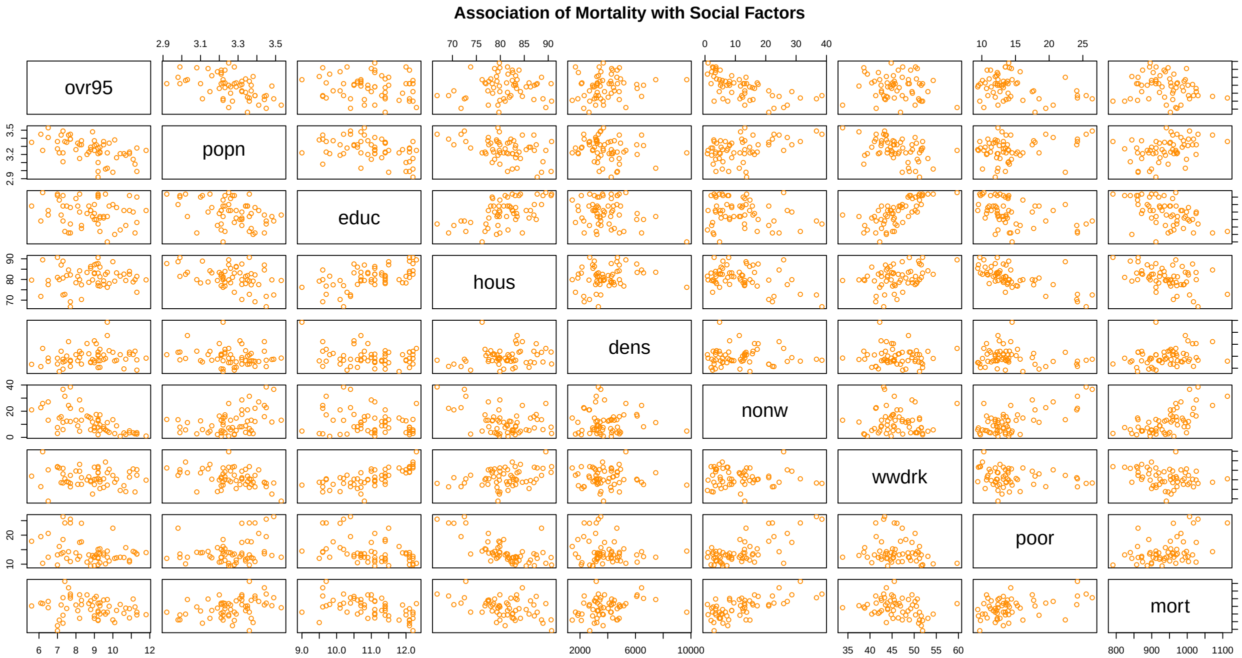


In the plot, we include 4 varibales about weather.
1) prec: Average annual precipitation in inches
2) jant: Average January temperature in degrees F
3) jult: Average July temperature in degrees F
4) humid: Annual average percentage relative humidity at 1pm

From the plot above, we could find that:
1) The temperature in January and July are highly positive correlated.
2) The mortality is not strongly related to the weather.
3) The data points in the plots at the bottom seems random.
4) There are outliers in the data.

3)

The scatterplot which shows the association of mortality social factors.

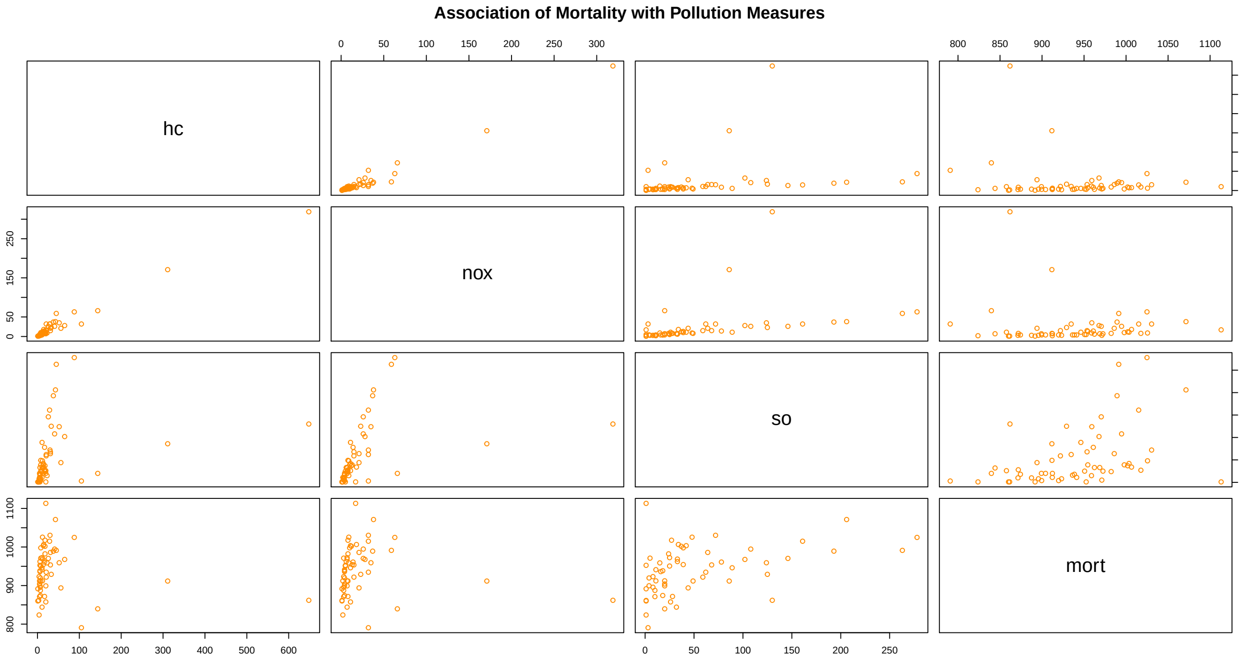


In the plot, we include 8 varibales about social factors.
1) ovr95: Percentage of 1960 SMSA population aged 65 or older
2) popn: Average household size
3) educ: Median school years completed by those over 22
4) hous: percentage of housing units which are sound and with all facilities
5) dens: Population per square mile in urbanized areas, 1960
6) nonw: Percentage non-white population in urbanized areas, 1960
7) wwdrk: Percentage employed in white collar occupations
8) poor: Percentage of families with income < 3000 dollars

From the plot above, we could find that:
1) The mortality is not strongly related to the social factors.
2) There are some outliers within the data.

4)

The scatterplot which shows the association of mortality with pollution measures.



In the plot, we include 3 varibales about pullution measures.
1) hc: Relative hydrocarbon pollution potential
2) nox: Relative nitric oxides pollution potential
3) so: Relative sulphur dioxide pollution potential

From the plot above, we could find that:
1) We might transform the variable mortality to move the data points from bottom to the center of the scatter plot.
2) The sulphur dioxide pollution is positively related with mortality.
3) We might transform variables hydrocarbon and nitric oxides to move the datapoints from corner to center.

2.

1)

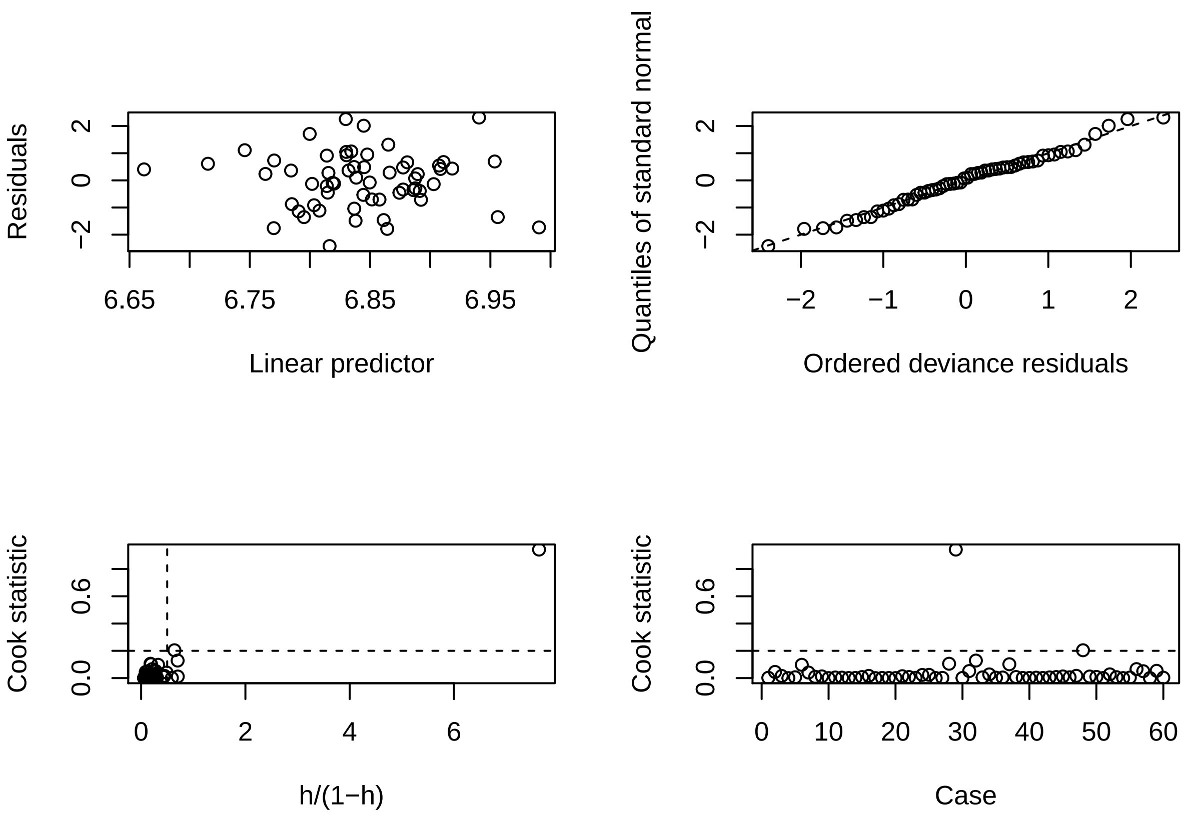
This model includes all the variables,

```r
summary(fit1)
```

```
##
## Call:
## glm(formula = log(mort) ~ prec + jant + jult + ovr95 + popn +
##     educ + nonw + hc + nox, data = pollution)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.078122  -0.023591   0.002894   0.017840   0.075813
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.9491203  0.3513514  22.624  < 2e-16 ***
## prec         0.0021323  0.0008821   2.417 0.019332 *
## jant        -0.0026029  0.0007329  -3.552 0.000846 ***
## jult        -0.0035889  0.0014719  -2.438 0.018353 *
## ovr95       -0.0124618  0.0075220  -1.657 0.103842
```

```
## popn        -0.1571025  0.0629750  -2.495 0.015956 *
## educ        -0.0248652  0.0074400  -3.342 0.001580 **
## nonw         0.0049152  0.0010208   4.815  1.4e-05 ***
## hc          -0.0009916  0.0003363  -2.948 0.004847 **
## nox          0.0020210  0.0006421   3.147 0.002777 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.001228892)
##
##     Null deviance: 0.259349  on 59  degrees of freedom
## Residual deviance: 0.061445  on 50  degrees of freedom
## AIC: -220.77
##
## Number of Fisher Scoring iterations: 2
```

```r
plot.glm.diag(fit1)
```



From the results above, we could see that:
1) Not all the variables are needed.
2) The most important variables are: jant, nonw, educ, hc, nox.
3) This model seems satisfies the random residuals assumption.

2)

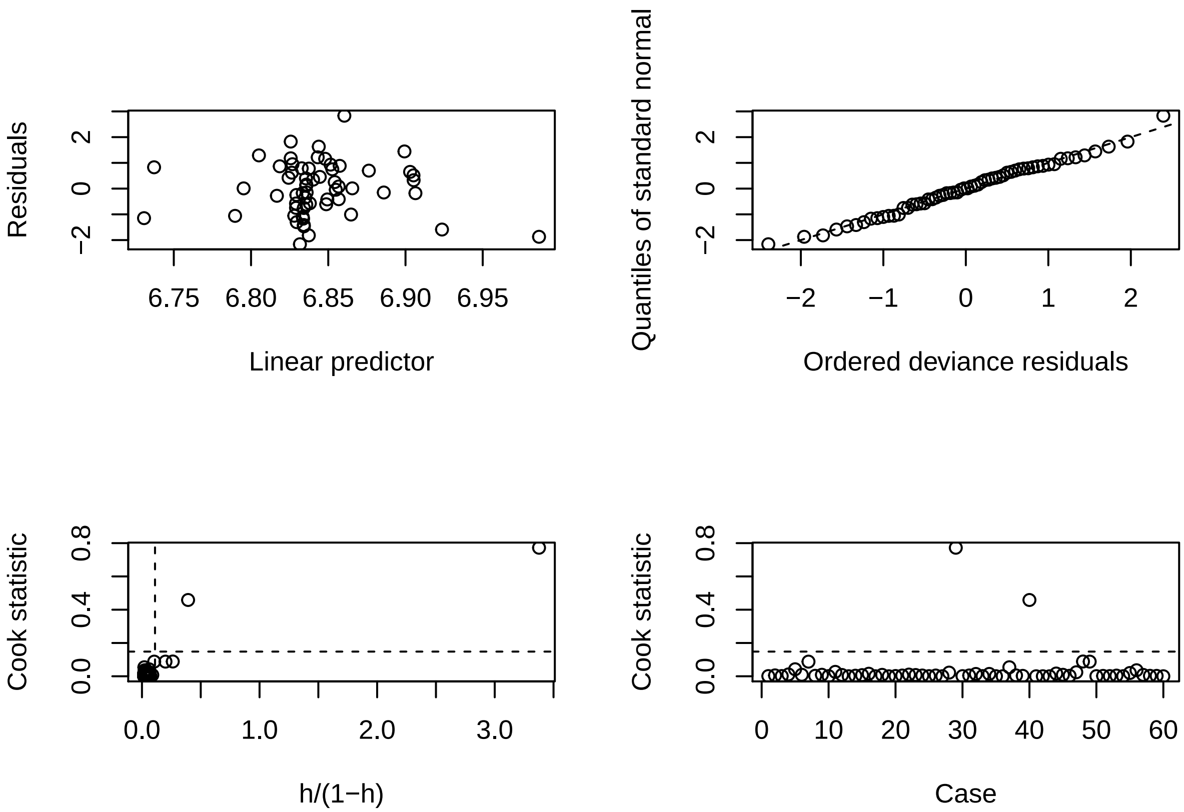
This model only include variables about pollution measures: hc,nox,so.

```r
summary(fit2)
```

```
##
```

```
## Call:
## glm(formula = log(mort) ~ hc + nox, data = pollution)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.117765  -0.035629   0.000526   0.038709   0.154534
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.8320999  0.0084081 812.561  < 2e-16 ***
## hc          -0.0022966  0.0004358  -5.270 2.17e-06 ***
## nox          0.0043678  0.0008651   5.049 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.003038591)
##
##     Null deviance: 0.25935  on 59  degrees of freedom
## Residual deviance: 0.17320  on 57  degrees of freedom
## AIC: -172.59
##
## Number of Fisher Scoring iterations: 2
```

```r
plot.glm.diag(fit2)
```



From the results above, we could see that: This model is not apropriate for this problem.

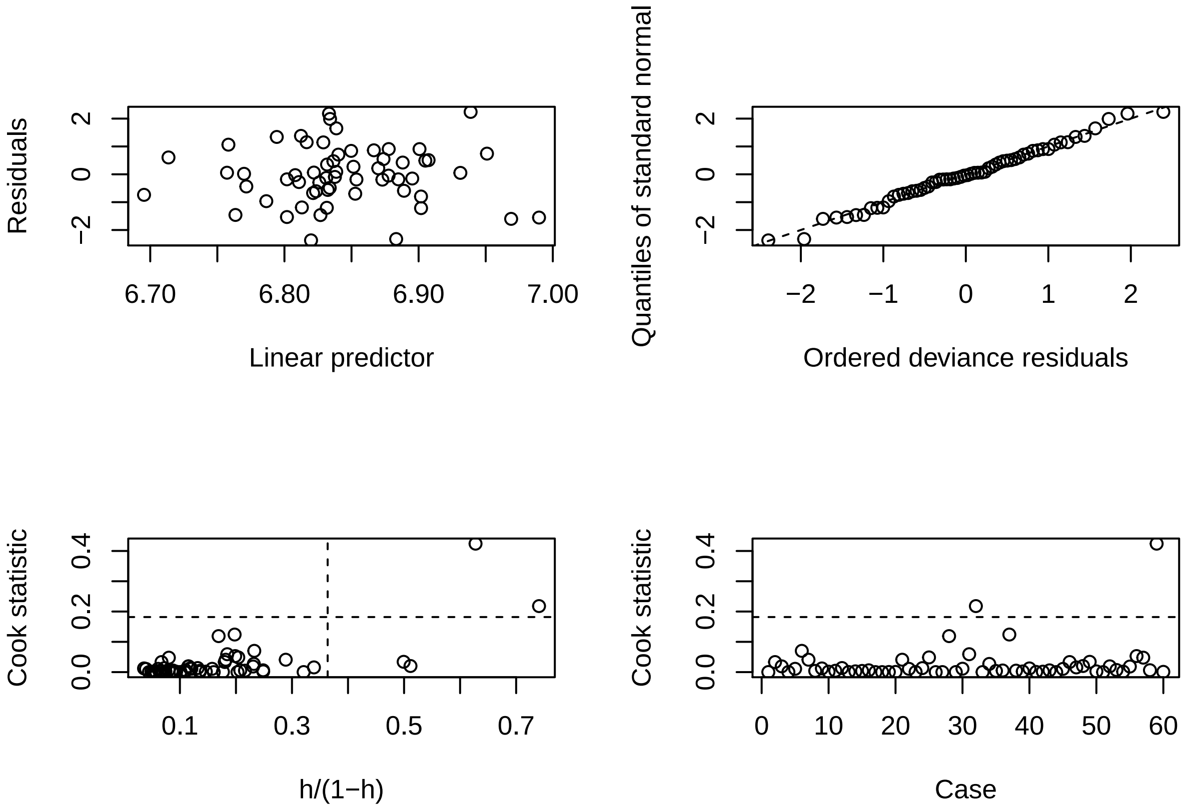
3)

This model includes all the variables except those about pollution measures,

```r
summary(fit3)
```

```
##
## Call:
## glm(formula = log(mort) ~ prec + jant + jult + popn + educ +
##     dens + nonw, data = pollution)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.081625  -0.021889  -0.001382   0.021198   0.078037
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.495e+00  2.450e-01  30.588  < 2e-16 ***
## prec         1.436e-03  6.446e-04   2.227 0.030290 *
## jant        -2.423e-03  6.462e-04  -3.749 0.000447 ***
## jult        -2.928e-03  1.357e-03  -2.158 0.035561 *
## popn        -8.240e-02  5.150e-02  -1.600 0.115617
## educ        -2.115e-02  7.548e-03  -2.802 0.007125 **
## dens         5.767e-06  3.788e-06   1.523 0.133906
## nonw         6.307e-03  8.560e-04   7.368 1.28e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.001385035)
##
##     Null deviance: 0.259349  on 59  degrees of freedom
## Residual deviance: 0.072022  on 52  degrees of freedom
## AIC: -215.24
##
## Number of Fisher Scoring iterations: 2
```

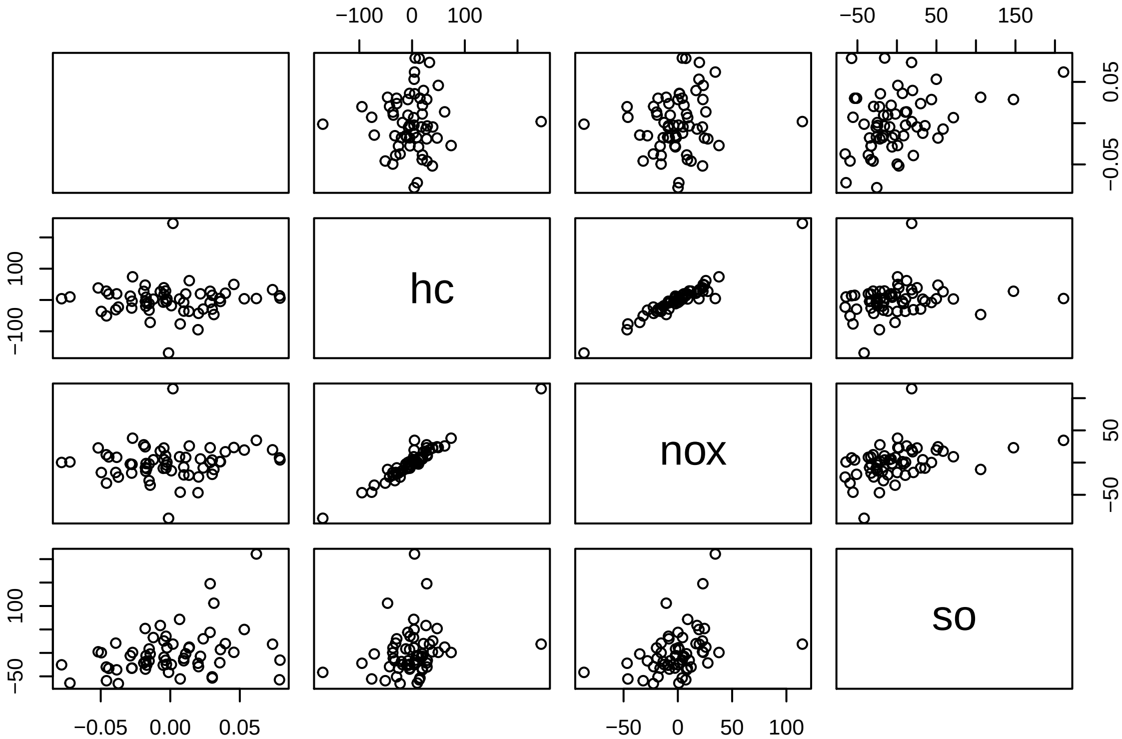
```r
plot.glm.diag(fit3)
```



From the results above, we could see that:
1) This model seems satisfies the random residuals assumption.
2) This model is reasonable for this problem.

3.

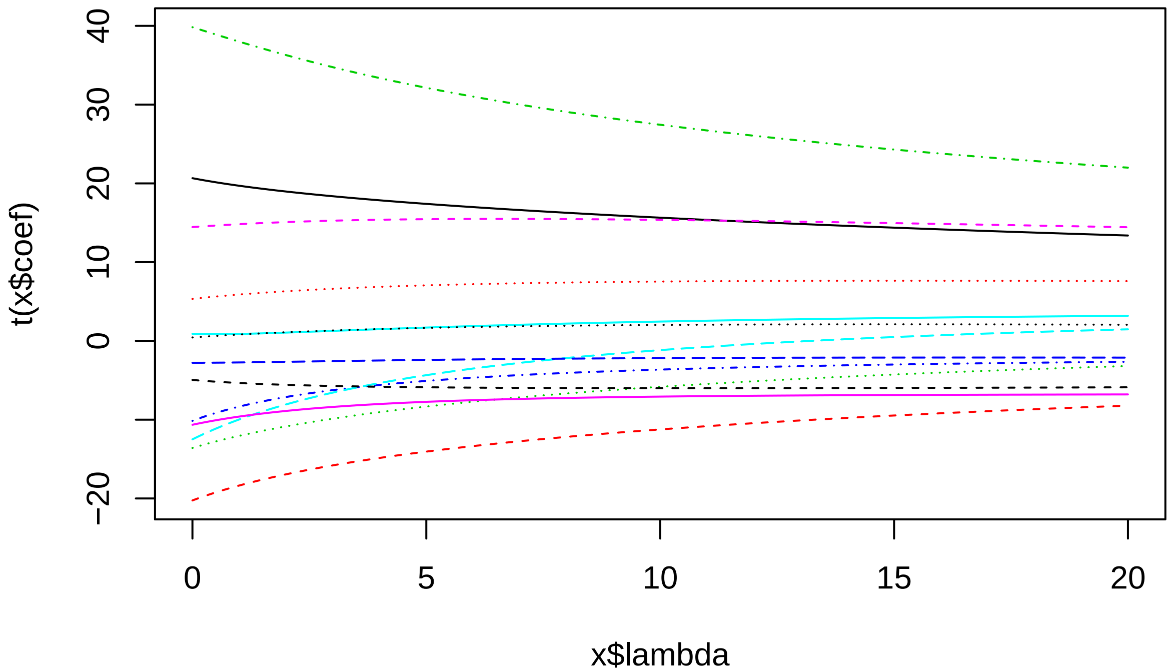
```
pairs(resid(lm(cbind(log(mort),hc,nox,so)~.,data=pollution)))
```



From the plot above, we could see that:
1) Log mort is not positively correlated with the pollution variables.
2) Variable hc and noc are highly correlated.
3) There are outliers.

4.

```
rfit <- lm.ridge(mort~.-hc-nox,data=pollution,lambda=seq(0,20,0.01))
plot(rfit)
```




```r
select(rfit)
```

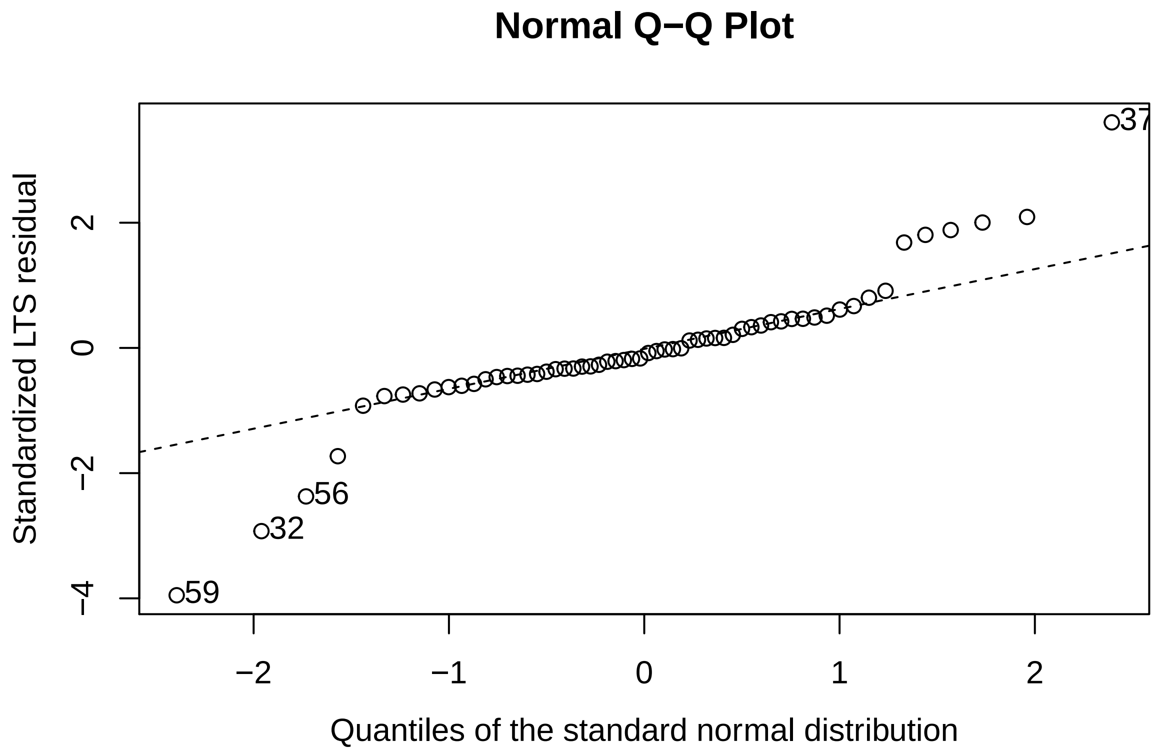
```
## modified HKB estimator is 4.116757
## modified L-W estimator is 4.659869
## smallest value of GCV  at 6.27
```

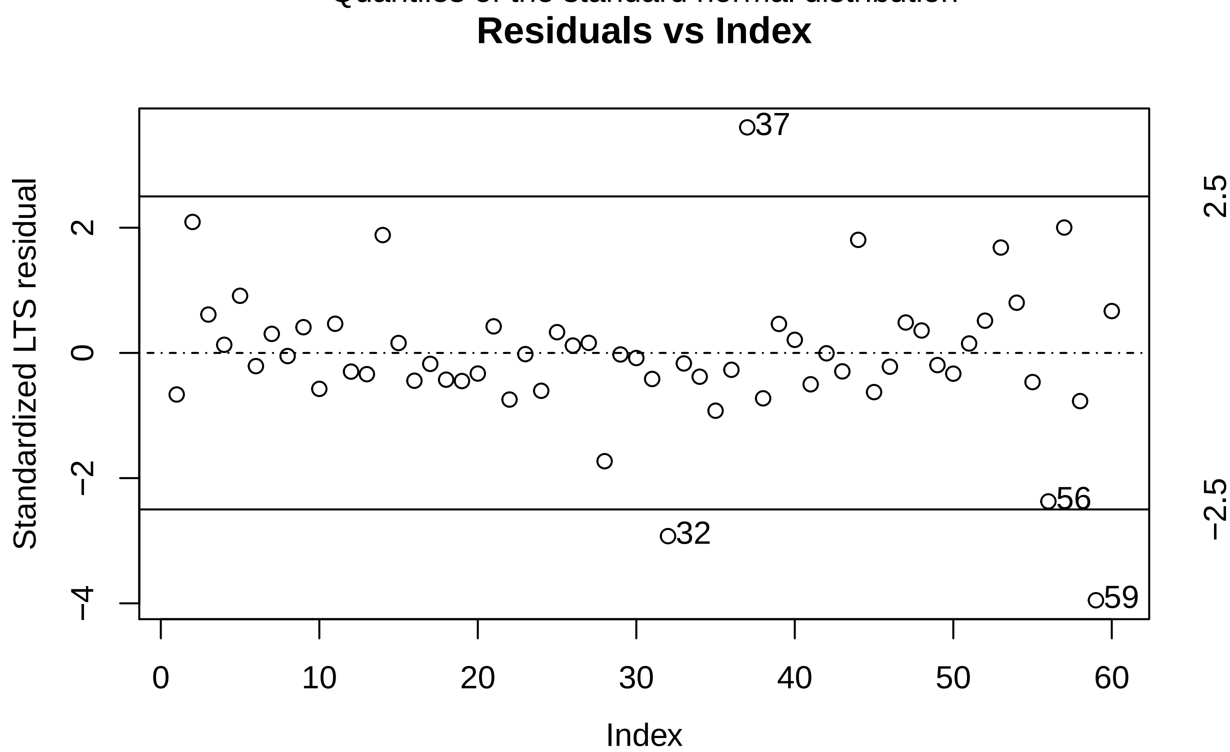
From the plot above, we could see that:
1) Some coefficients are about 0 no matter what value λ is, which means theses variables are not needed.
2) All the variables go to 0 when λ increases, which makes sense for ridge regression.

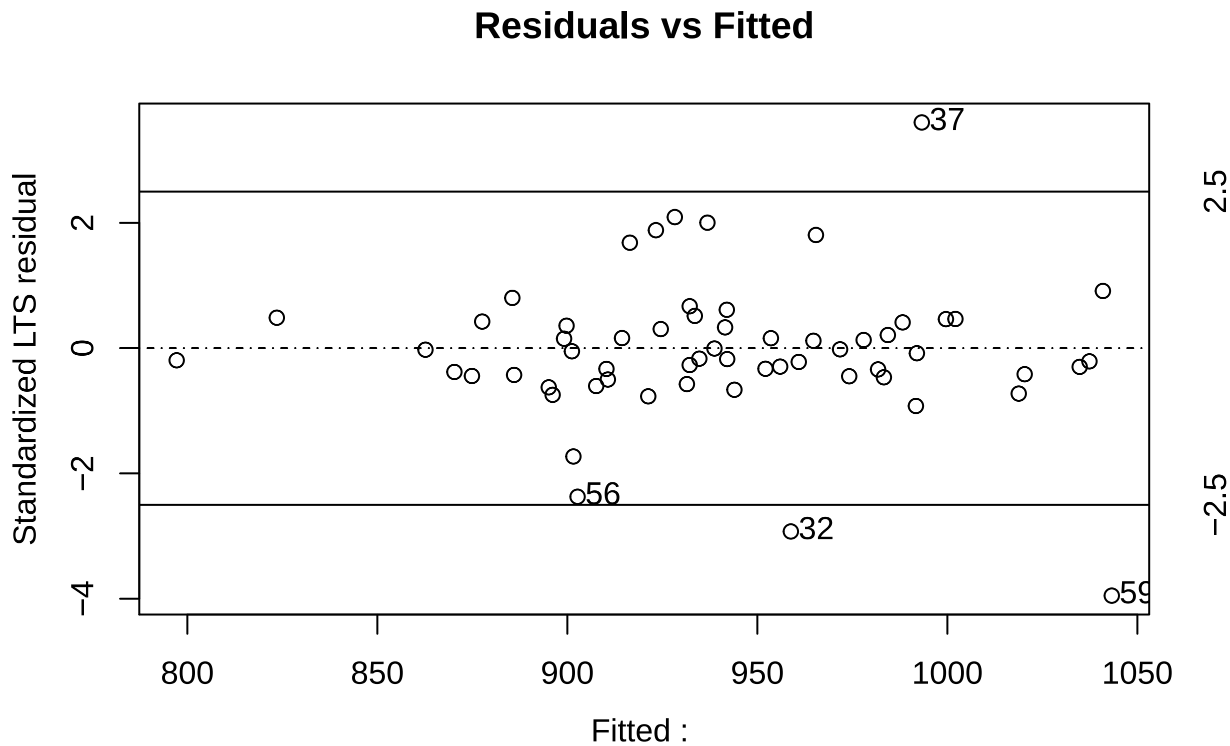
5.

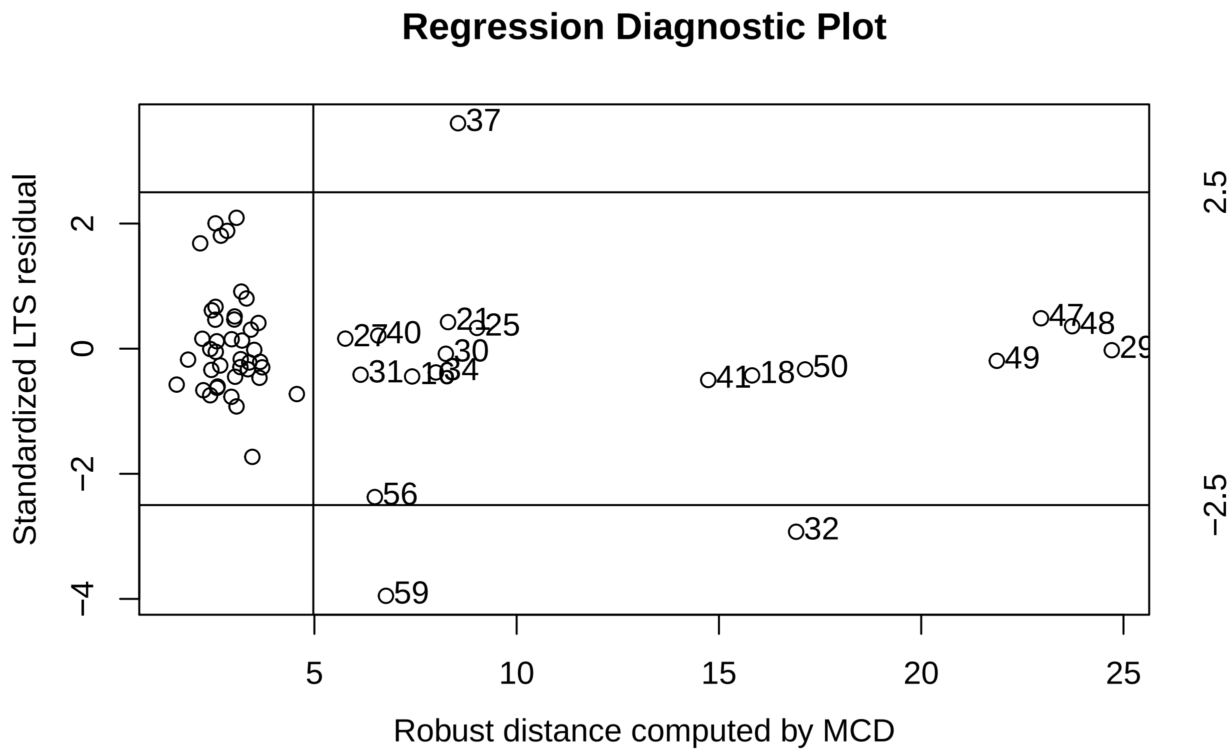
1) least trimmed squares regression

```r
tfit <- ltsReg(mort~.-hc-nox,data=pollution,lambda=seq(0,20,0.01))
plot(tfit)
```


Normal Q−Q Plot
Standardized LTS residual
Quantiles of the standard normal distribution
37
56
32
59


Residuals vs Index
Standardized LTS residual
Index
2.5
−2.5
37
56
32
59






```r
summary(tfit)
```

```
##
## Call:
## ltsReg.formula(formula = mort ~ . - hc - nox, data = pollution,
##     lambda = seq(0, 20, 0.01))
##
```

```
## Residuals (from reweighted LS):
##      Min       1Q   Median       3Q      Max
## -57.5390 -12.9518  -0.7345  12.3532  69.5978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## Intercept  1.527e+03  3.759e+02   4.064 0.000207 ***
## prec       2.959e+00  7.241e-01   4.087 0.000193 ***
## jant      -1.904e+00  8.000e-01  -2.380 0.021918 *
## jult      -1.777e+00  1.499e+00  -1.186 0.242362
## ovr95     -1.667e+01  6.695e+00  -2.490 0.016812 *
## popn      -1.010e+02  5.678e+01  -1.778 0.082629 .
## educ       6.383e-01  9.993e+00   0.064 0.949371
## hous      -1.547e+00  1.658e+00  -0.933 0.356078
## dens       1.626e-02  4.189e-03   3.882 0.000360 ***
## nonw       6.569e-01  1.129e+00   0.582 0.563718
## wwdrk     -1.305e+00  1.323e+00  -0.986 0.329784
## poor       3.866e+00  2.899e+00   1.334 0.189524
## so         2.719e-01  8.096e-02   3.359 0.001674 **
## humid      2.462e-01  8.852e-01   0.278 0.782316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.71 on 42 degrees of freedom
## Multiple R-Squared: 0.8042,  Adjusted R-squared: 0.7436
## F-statistic: 13.27 on 13 and 42 DF,  p-value: 6.116e-11
```

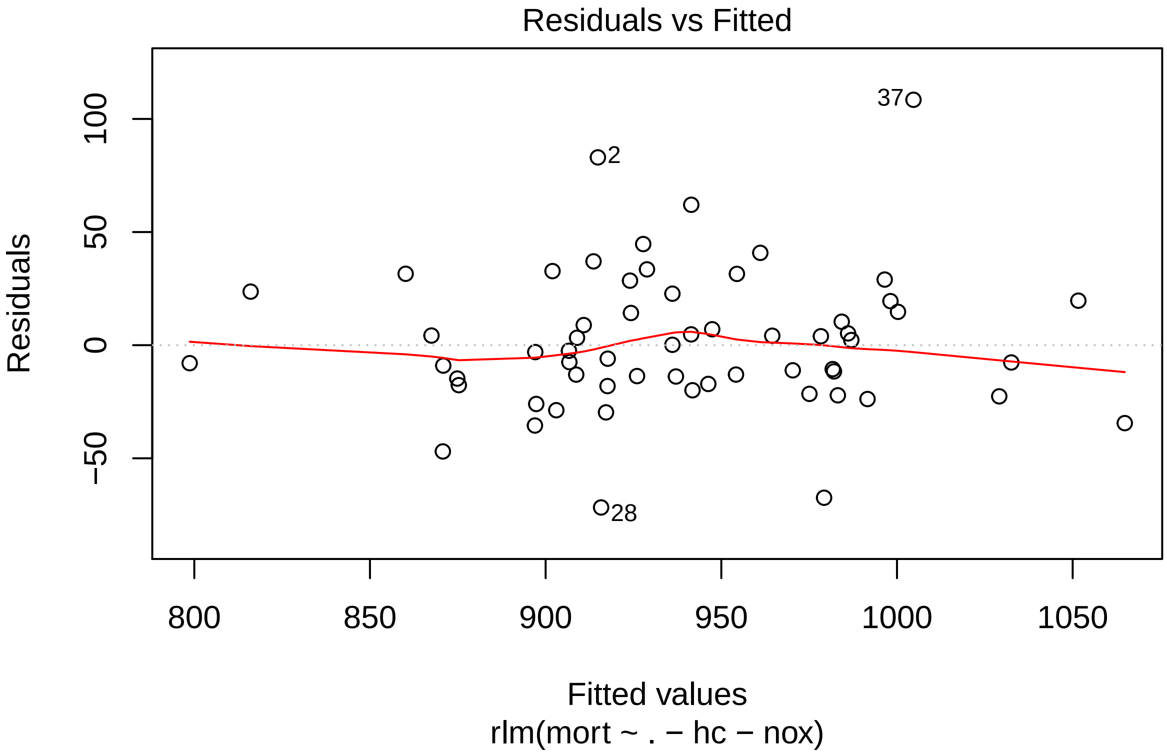
From the results above, we could see that:
1) The most important variables are prec, jant, dens and so, which is different from the results above.
2) This model is very plausible.

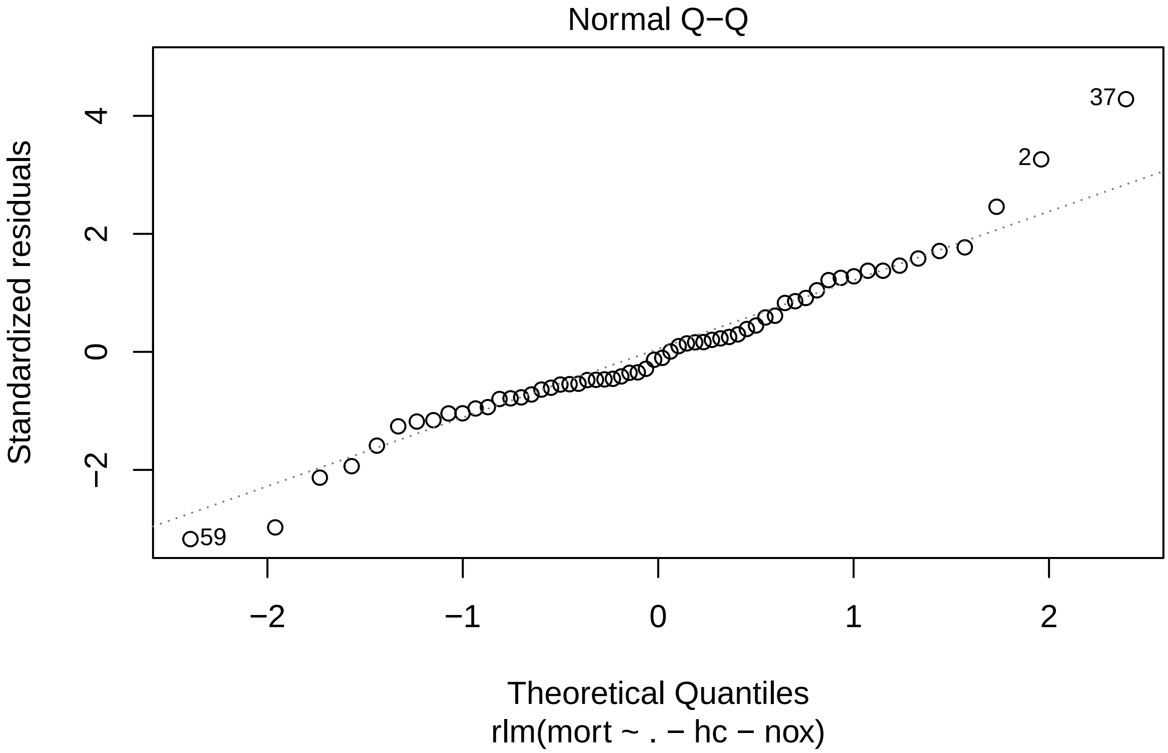
2) robust M-estimation

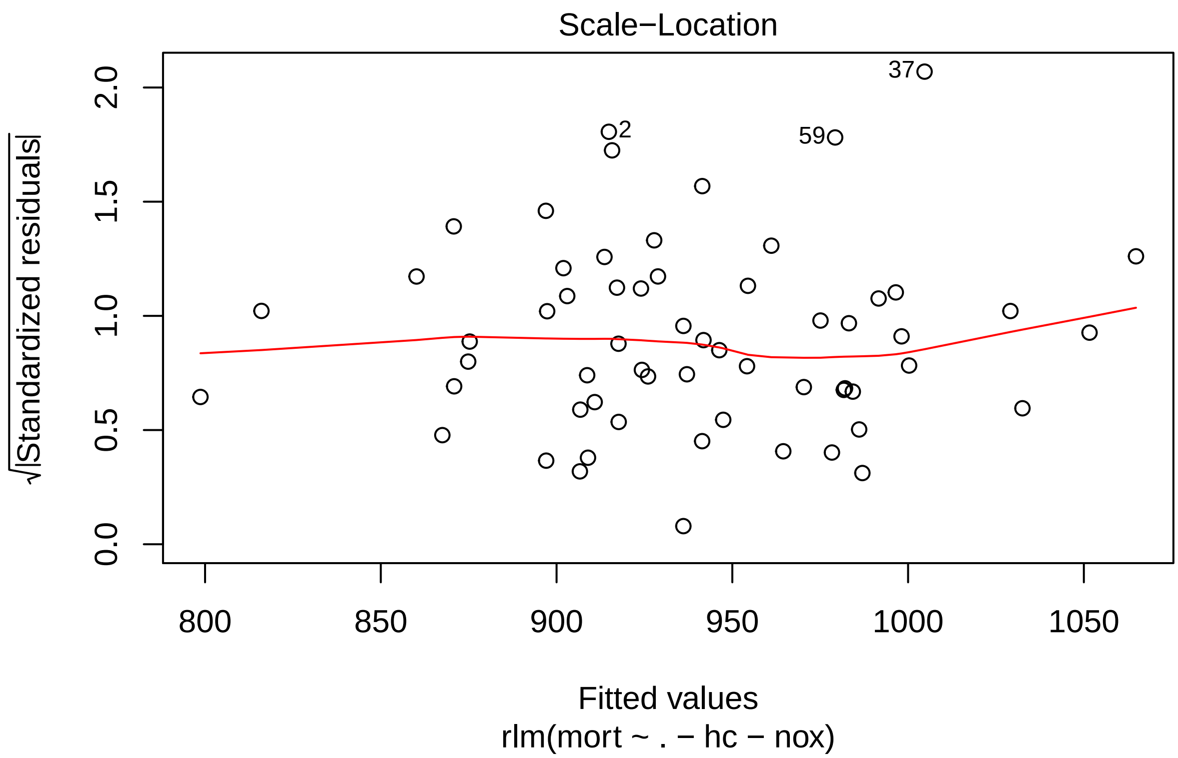
```
mfit <- rlm(mort~.-hc-nox,data=pollution,lambda=seq(0,20,0.01))
```

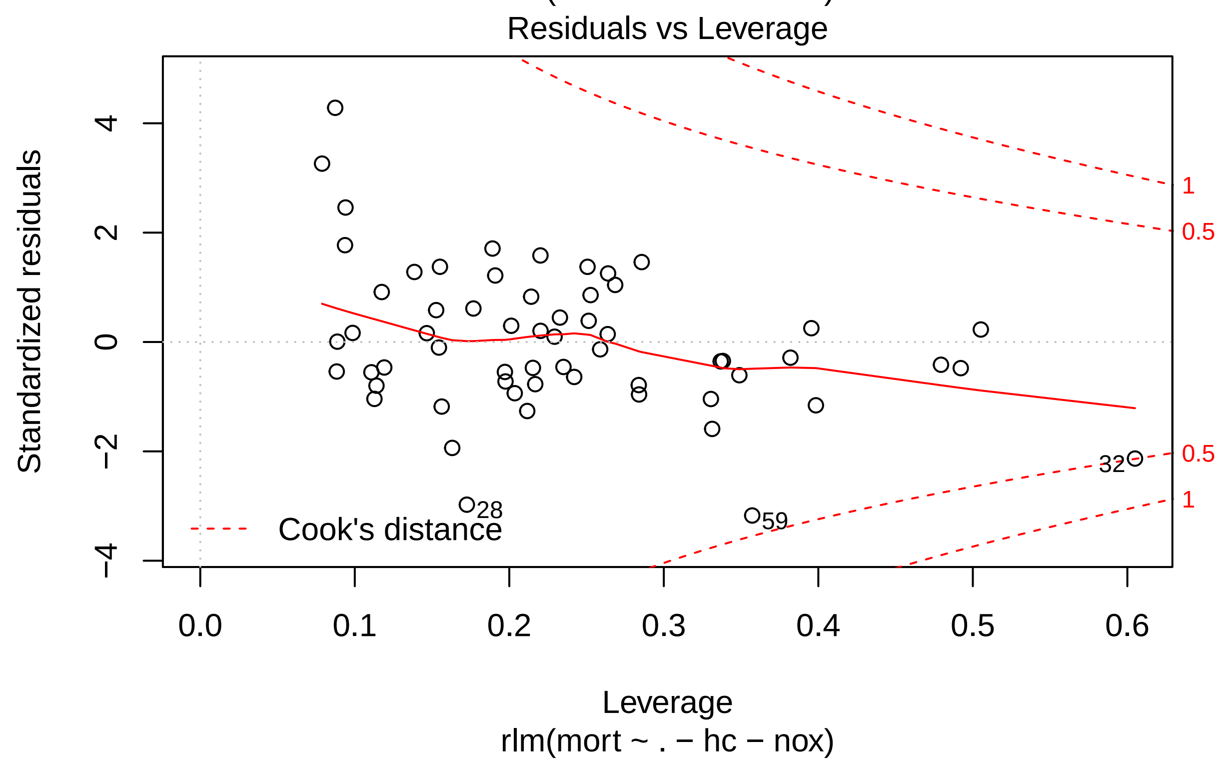
```
## Warning in rlm.default(x, y, weights, method = method, wt.method =
## wt.method, : some of ... do not match
```

```
plot(mfit)
```


Residuals vs Fitted
Residuals
100
50
0
−50
37
2
28
800
850
900
950
1000
1050
Fitted values
rlm(mort ~ . − hc − nox)


Normal Q−Q
Standardized residuals
4
2
0
−2
37
2
59
−2
−1
0
1
2
Theoretical Quantiles
rlm(mort ~ . − hc − nox)






```r
summary(mfit)
```

```
##
## Call: rlm(formula = mort ~ . - hc - nox, data = pollution, lambda = seq(0,
##     20, 0.01))
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -71.738 -17.276  -2.769  19.524 108.462
##
## Coefficients:
##             Value     Std. Error t value
## (Intercept) 1672.3291  379.2052     4.4101
## prec           2.0616    0.7252     2.8427
## jant          -1.6291    0.8507    -1.9150
## jult          -2.6195    1.5811    -1.6568
## ovr95         -7.2603    7.0738    -1.0264
## popn         -78.8129   61.5684    -1.2801
## educ         -10.7272    9.9884    -1.0740
## hous          -1.5532    1.5554    -0.9986
## dens           0.0055    0.0036     1.5353
## nonw           4.0195    1.0820     3.7147
## wwdrk         -1.0775    1.3982    -0.7706
## poor          -0.9276    2.8156    -0.3294
## so             0.2459    0.0837     2.9372
## humid         -0.2518    0.9404    -0.2677
##
## Residual standard error: 26.51 on 46 degrees of freedom
```

From the results of this model, we coold see that variables hc and nox does not have much significance.