exercise5

Chutian Chen cc4515; Congcheng Yan cy2550; Mingrui Liu ml4404

4/13/2020

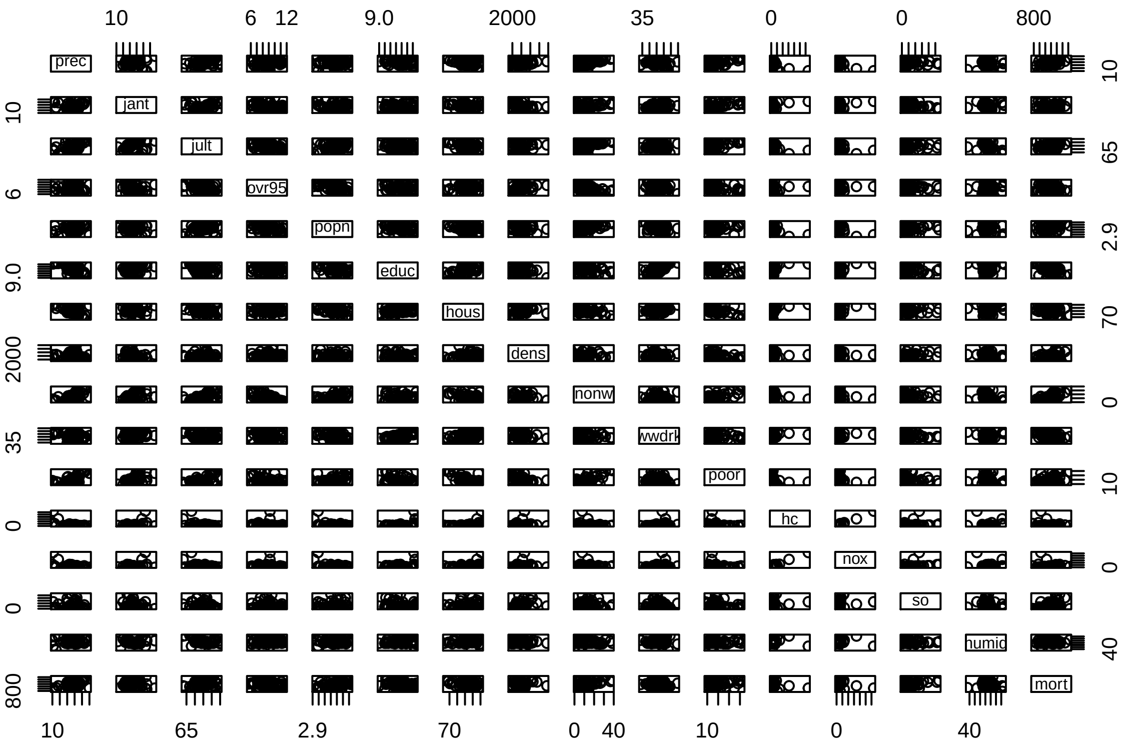
(1)

```
library(SMPracticals)
```

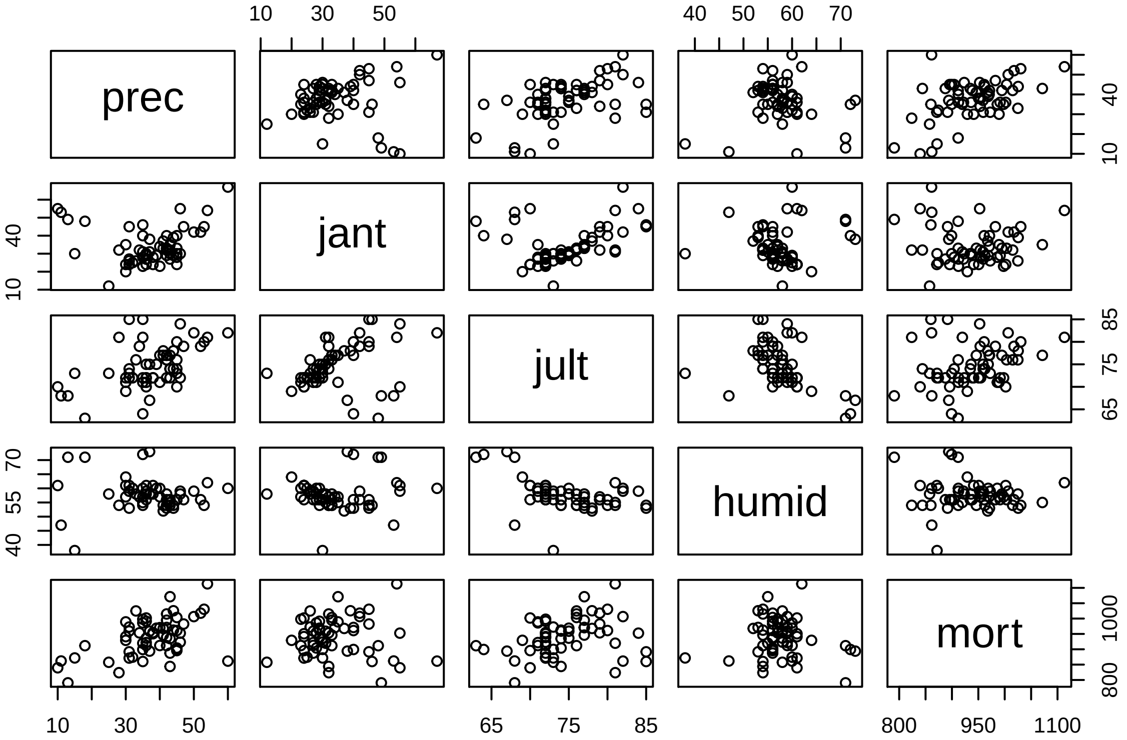
```
## Loading required package: ellipse

##
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':
##
##     pairs
```

```
data = pollution
```

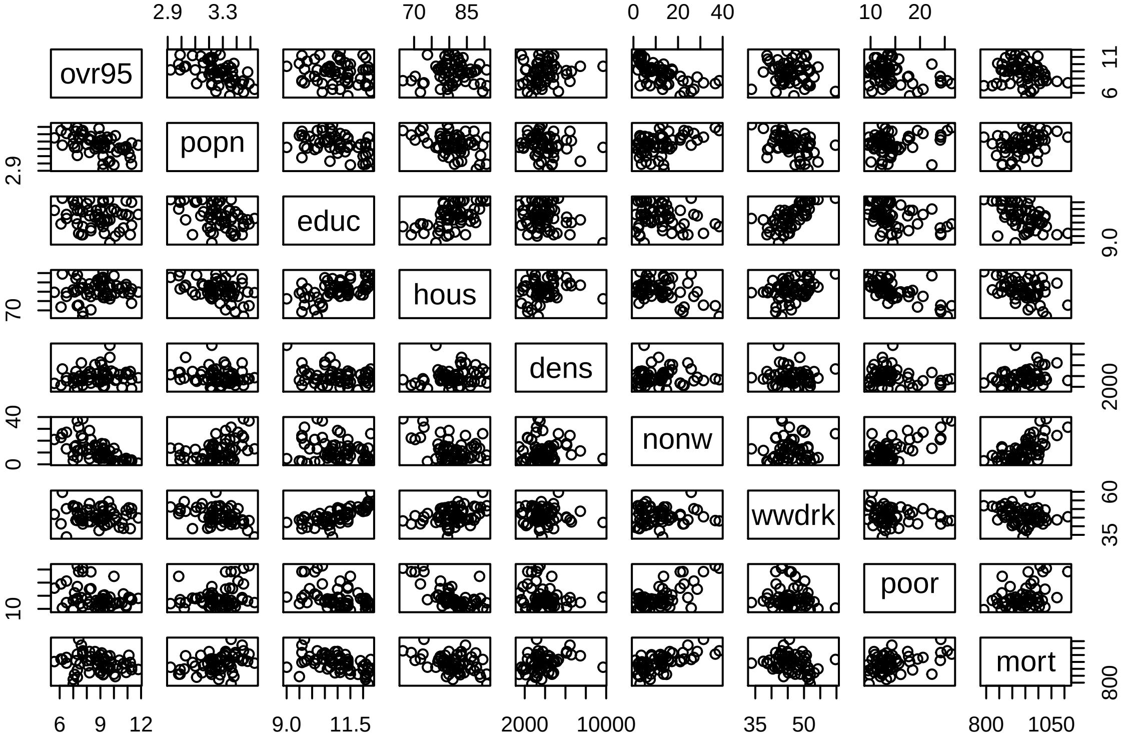
```
pairs(pollution)
```



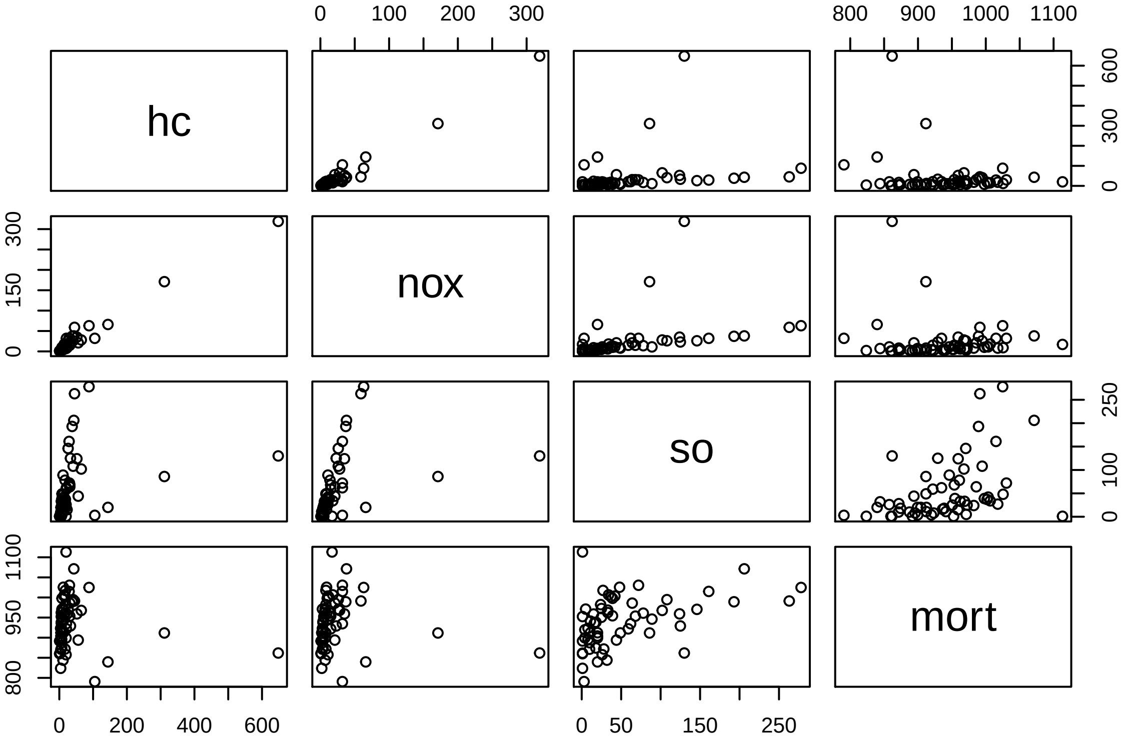
```
pairs(pollution[,c(1:3,15:16)]) # association of mortality with weather
```



```r
pairs(pollution[,c(4:11,16)])    # and social factors
```



```r
pairs(pollution[,c(12:14,16)])   # and pollution measures
```



There are some outliers in the dataset. We need to remove outliers or use log transformation in the dataset. For transformation on features and responses, it seems many features don't have linear relation with response. We can use Box-Cox transformation.

We can see clusters in the scatter plot of air pollution and mortality. It's hard for us to interpret the relation between them with linear model.

(2)

```
fit <- step(glm(mort~.-hc-nox-so,data=pollution))
```

```
## Start:  AIC=615.94
## mort ~ (prec + jant + jult + ovr95 + popn + educ + hous + dens +
##     nonw + wwdrk + poor + hc + nox + so + humid) - hc - nox -
##     so
##
##         Df Deviance    AIC
## - humid  1    63302 613.95
## - hous   1    63343 613.99
## - poor   1    63351 614.00
## - wwdrk  1    63365 614.01
## - ovr95  1    63707 614.34
## <none>        63288 615.94
## - dens   1    65434 615.94
## - popn   1    66050 616.50
## - jult   1    67033 617.39
## - educ   1    67999 618.25
## - prec   1    68175 618.40
## - jant   1    69624 619.66
## - nonw   1    96348 639.16
```

```
##
## Step:  AIC=613.95
## mort ~ prec + jant + jult + ovr95 + popn + educ + hous + dens +
##     nonw + wwdrk + poor
##
##         Df Deviance    AIC
## - hous   1    63351 612.00
## - poor   1    63360 612.01
## - wwdrk  1    63378 612.02
## - ovr95  1    63713 612.34
## <none>        63302 613.95
## - dens   1    65509 614.01
## - popn   1    66050 614.50
## - jult   1    67922 616.18
## - educ   1    68071 616.31
## - prec   1    68346 616.55
## - jant   1    69939 617.94
## - nonw   1    96365 637.17
##
## Step:  AIC=612
## mort ~ prec + jant + jult + ovr95 + popn + educ + dens + nonw +
##     wwdrk + poor
##
##         Df Deviance    AIC
## - poor   1    63368 610.01
## - wwdrk  1    63407 610.05
## - ovr95  1    63790 610.41
## <none>        63351 612.00
## - dens   1    65520 612.02
## - popn   1    66128 612.57
## - jult   1    68059 614.30
## - prec   1    68507 614.69
## - educ   1    68823 614.97
## - jant   1    73071 618.56
## - nonw   1    96499 635.25
##
## Step:  AIC=610.01
## mort ~ prec + jant + jult + ovr95 + popn + educ + dens + nonw +
##     wwdrk
##
##         Df Deviance    AIC
## - wwdrk  1    63420 608.06
## - ovr95  1    63947 608.56
## <none>        63368 610.01
## - dens   1    65988 610.45
## - popn   1    66284 610.71
## - prec   1    68707 612.87
## - educ   1    69060 613.18
## - jult   1    69164 613.27
## - jant   1    77841 620.36
## - nonw   1   109754 640.97
##
## Step:  AIC=608.06
## mort ~ prec + jant + jult + ovr95 + popn + educ + dens + nonw
```

```
##
##          Df Deviance    AIC
## - ovr95  1    64018 606.63
## <none>         63420 608.06
## - dens   1    65988 608.45
## - popn   1    66285 608.71
## - prec   1    68849 610.99
## - jult   1    69521 611.57
## - educ   1    73291 614.74
## - jant   1    77925 618.42
## - nonw   1   110819 639.55
##
## Step:  AIC=606.63
## mort ~ prec + jant + jult + popn + educ + dens + nonw
##
##        Df Deviance    AIC
## <none>        64018 606.63
## - popn  1    66596 607.00
## - dens  1    66953 607.32
## - prec  1    69428 609.49
## - jult  1    69614 609.65
## - educ  1    73806 613.16
## - jant  1    78989 617.24
## - nonw  1   129620 646.95
```

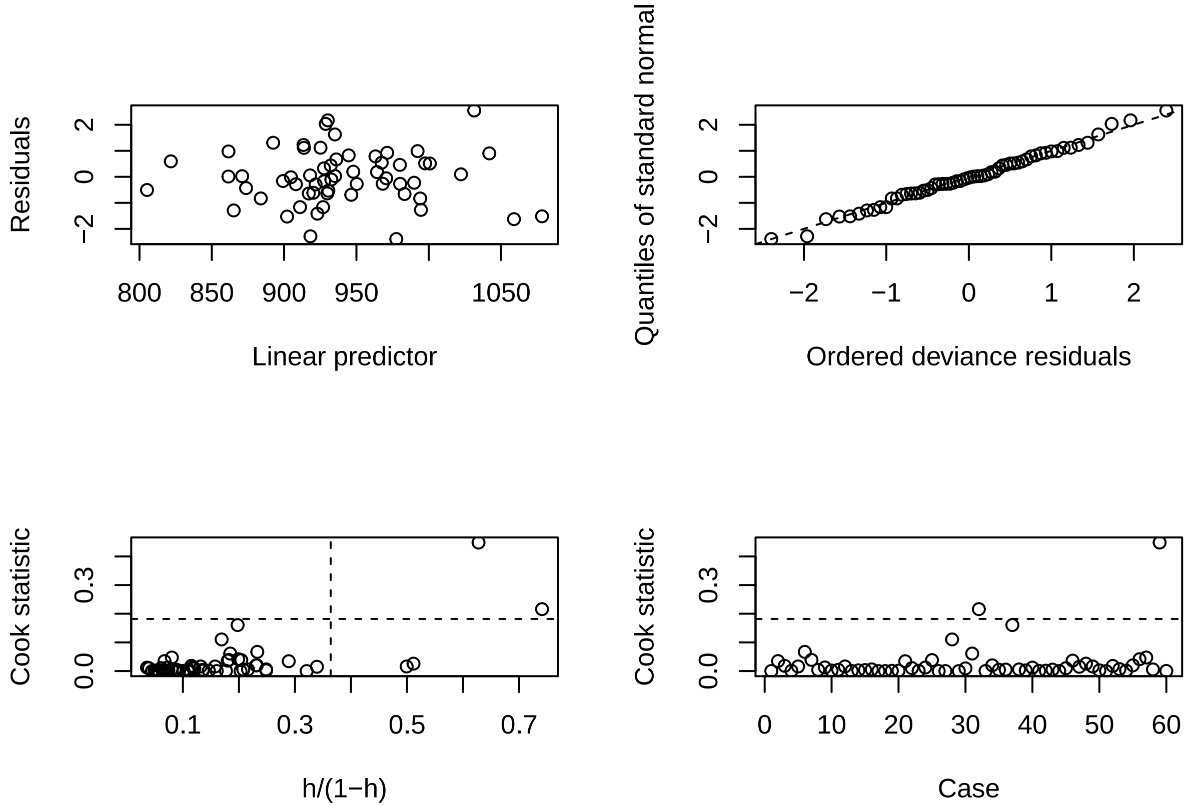
```
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'

## The following objects are masked from 'package:stats':
##
##     predict, predict.lm

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

```
boxcox(fit)
```

```
##
## Results of Box-Cox Transformation
## ---------------------------------
##
## Objective Name:                  PPCC
##
## Linear Model:                    fit
##
## Sample Size:                     60
##
##  lambda      PPCC
##    -2.0 0.9952407
##    -1.5 0.9959795
##    -1.0 0.9964582
##    -0.5 0.9962627
##     0.0 0.9958303
##     0.5 0.9953650
##     1.0 0.9947407
##     1.5 0.9936601
##     2.0 0.9923084
```

```
plot.glm.diag(fit) # model adequate?
```



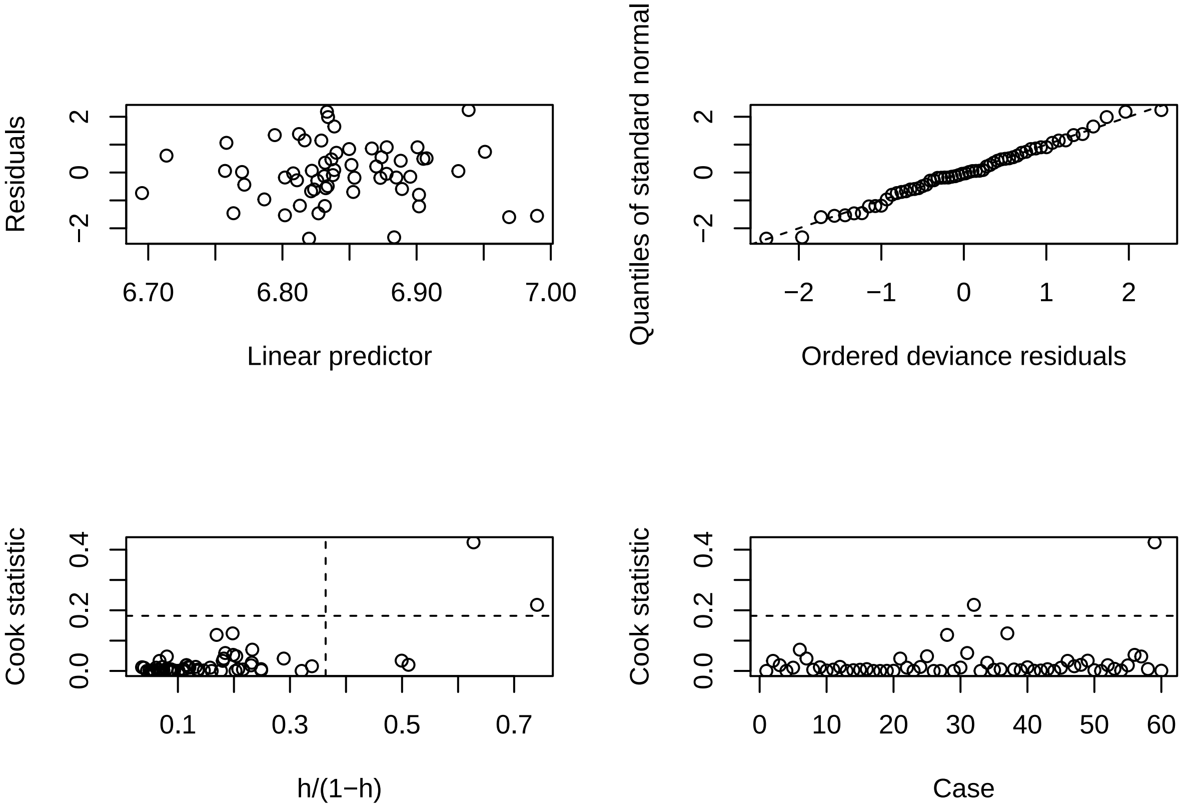
```
summary(fit)
```

```
##
## Call:
## glm(formula = mort ~ prec + jant + jult + popn + educ + dens +
##     nonw, data = pollution)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -74.148  -20.837   -1.231   19.548   81.714
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.526e+03  2.310e+02   6.608 2.09e-08 ***
## prec         1.274e+00  6.078e-01   2.096  0.04095 *
## jant        -2.125e+00  6.092e-01  -3.487  0.00100 **
## jult        -2.727e+00  1.279e+00  -2.132  0.03776 *
## popn        -7.025e+01  4.855e+01  -1.447  0.15388
## educ        -2.006e+01  7.116e+00  -2.820  0.00679 **
## dens         5.513e-03  3.571e-03   1.544  0.12867
## nonw         5.891e+00  8.070e-01   7.300 1.65e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1231.114)
##
##     Null deviance: 228308  on 59  degrees of freedom
## Residual deviance:  64018  on 52  degrees of freedom
## AIC: 606.63
##
## Number of Fisher Scoring iterations: 2
```

```
fit <- update(fit,log(mort)~.) # try log transform of response plot.glm.diag(fit) # model adequate?
summary(fit)
```

```
##
## 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
```

```
plot.glm.diag(fit) # model adequate?
```



We can see that both model (take log and not) are adequate. However, taking log on response can help us to deal with outliers. So we prefer taking log.

```
fit_all <- lm(log(mort)~.-hc-nox-so,data=pollution)
summary(fit_all)
```

```
##
## Call:
## lm(formula = log(mort) ~ . - hc - nox - so, data = pollution)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.078144 -0.018287 -0.002924  0.022202  0.078882
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.819e+00  4.700e-01  16.636  < 2e-16 ***
## prec         1.812e-03  8.728e-04   2.077   0.0433 *
## jant        -2.479e-03  1.057e-03  -2.345   0.0233 *
## jult        -3.484e-03  1.967e-03  -1.771   0.0830 .
## ovr95       -5.830e-03  8.746e-03  -0.667   0.5083
## popn        -1.258e-01  7.655e-02  -1.644   0.1068
## educ        -2.071e-02  1.186e-02  -1.746   0.0873 .
## hous        -5.675e-04  1.918e-03  -0.296   0.7686
## dens         5.722e-06  4.414e-06   1.296   0.2012
## nonw         6.099e-03  1.225e-03   4.978 9.06e-06 ***
## wwdrk       -7.432e-04  1.739e-03  -0.427   0.6711
## poor        -7.426e-04  3.482e-03  -0.213   0.8320
## humid       -2.347e-04  1.168e-03  -0.201   0.8416
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0388 on 47 degrees of freedom
## Multiple R-squared:  0.7271, Adjusted R-squared:  0.6575
## F-statistic: 10.44 on 12 and 47 DF,  p-value: 1.32e-09
```

```
fit_least <- lm(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw, data = pollution)
summary(fit_least)
```

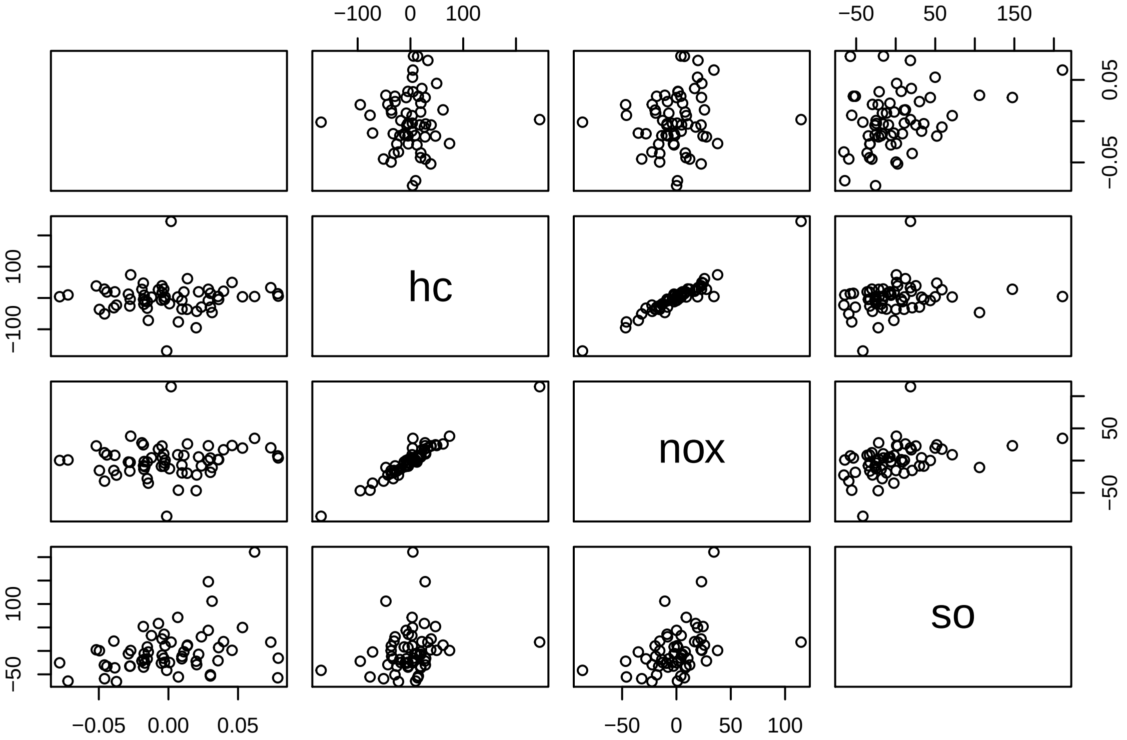
```
##
## Call:
## lm(formula = log(mort) ~ prec + jant + jult + popn + educ + dens +
##     nonw, data = pollution)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.081625 -0.021889 -0.001382  0.021198  0.078037
##
## Coefficients:
##               Estimate Std. Error t value 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
##
## Residual standard error: 0.03722 on 52 degrees of freedom
## Multiple R-squared:  0.7223, Adjusted R-squared:  0.6849
## F-statistic: 19.32 on 7 and 52 DF,  p-value: 1.913e-12
```

We can see that the step eliminates some insignificant features and imporves the adjusted R-square. So we use the features selected by step.

As we mentioned before, we think log transformation is good for the data. So we will choose log model with selected features.

(3)

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



The scatter plots all show a large cluster and it seems inappropriate to use linear regression on it.

There are outliers in all three pollution.

In all pollutions, SO has the strongest linear relation.

```r
fit_so <- lm(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + so, data = pollution)
summary(fit_so)
```

```
##
## Call:
## lm(formula = log(mort) ~ prec + jant + jult + popn + educ + dens +
##     nonw + so, data = pollution)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.082580 -0.020950 -0.002096  0.016160  0.088873
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.318e+00  2.461e-01  29.735  < 2e-16 ***
## prec         1.863e-03  6.429e-04   2.898  0.00553 **
## jant        -2.099e-03  6.336e-04  -3.313  0.00170 **
## jult        -2.484e-03  1.313e-03  -1.892  0.06415 .
## popn        -6.068e-02  5.016e-02  -1.210  0.23189
## educ        -1.622e-02  7.518e-03  -2.158  0.03567 *
## dens         3.051e-06  3.802e-06   0.803  0.42598
## nonw         5.517e-03  8.843e-04   6.239 8.63e-08 ***
## so           2.168e-04  9.094e-05   2.384  0.02088 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.03565 on 51 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.711
## F-statistic: 19.14 on 8 and 51 DF,  p-value: 6.641e-13
```

We can see the model is improved. R-squared increases.

Then we try to add the other two features.

```
fit_so_nox_hc <- lm(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + so + nox + hc, data = p
summary(fit_so_nox_hc)
```

```
##
## Call:
## lm(formula = log(mort) ~ prec + jant + jult + popn + educ + dens +
##     nonw + so + nox + hc, data = pollution)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.079690 -0.018963  0.001739  0.015901  0.082277
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.359e+00  2.593e-01  28.383  < 2e-16 ***
## prec         1.471e-03  7.406e-04   1.987  0.05254 .
## jant        -1.871e-03  6.982e-04  -2.680  0.00999 **
## jult        -2.740e-03  1.410e-03  -1.943  0.05778 .
## popn        -6.485e-02  5.145e-02  -1.261  0.21342
## educ        -1.629e-02  7.526e-03  -2.165  0.03527 *
## dens         3.668e-06  3.829e-06   0.958  0.34271
## nonw         5.527e-03  9.034e-04   6.118 1.54e-07 ***
## so           9.900e-05  1.342e-04   0.737  0.46434
## nox          1.242e-03  9.266e-04   1.340  0.18633
## hc          -6.469e-04  4.616e-04  -1.401  0.16743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03564 on 49 degrees of freedom
## Multiple R-squared:   0.76,  Adjusted R-squared:  0.711
## F-statistic: 15.52 on 10 and 49 DF,  p-value: 5.05e-12
```

Adding the two features doesn't improve the model. We try to adjust the model with taking log on nox and hc.

```
fit_so_nox_hc <- lm(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + so + log(nox) + log(hc
summary(fit_so_nox_hc)
```

```
##
## Call:
## lm(formula = log(mort) ~ prec + jant + jult + popn + educ + dens +
##     nonw + so + log(nox) + log(hc), data = pollution)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.079660 -0.021461  0.002049  0.017834  0.076347
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.322e+00  2.613e-01  28.018  < 2e-16 ***
## prec         1.916e-03  6.863e-04   2.792  0.00744 **
## jant        -2.110e-03  6.523e-04  -3.235  0.00218 **
## jult        -2.729e-03  1.780e-03  -1.533  0.13174
## popn        -6.292e-02  4.836e-02  -1.301  0.19935
## educ        -1.362e-02  7.331e-03  -1.857  0.06926 .
## dens         2.986e-06  3.761e-06   0.794  0.43115
## nonw         5.452e-03  9.928e-04   5.492 1.41e-06 ***
## so           8.186e-05  1.169e-04   0.700  0.48716
## log(nox)     3.625e-02  1.484e-02   2.443  0.01822 *
## log(hc)     -3.051e-02  1.565e-02  -1.949  0.05700 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03432 on 49 degrees of freedom
## Multiple R-squared:  0.7775, Adjusted R-squared:  0.7321
## F-statistic: 17.12 on 10 and 49 DF,  p-value: 8.653e-13
```

We can see that the model improves after taking log on nox and hc.

(4)

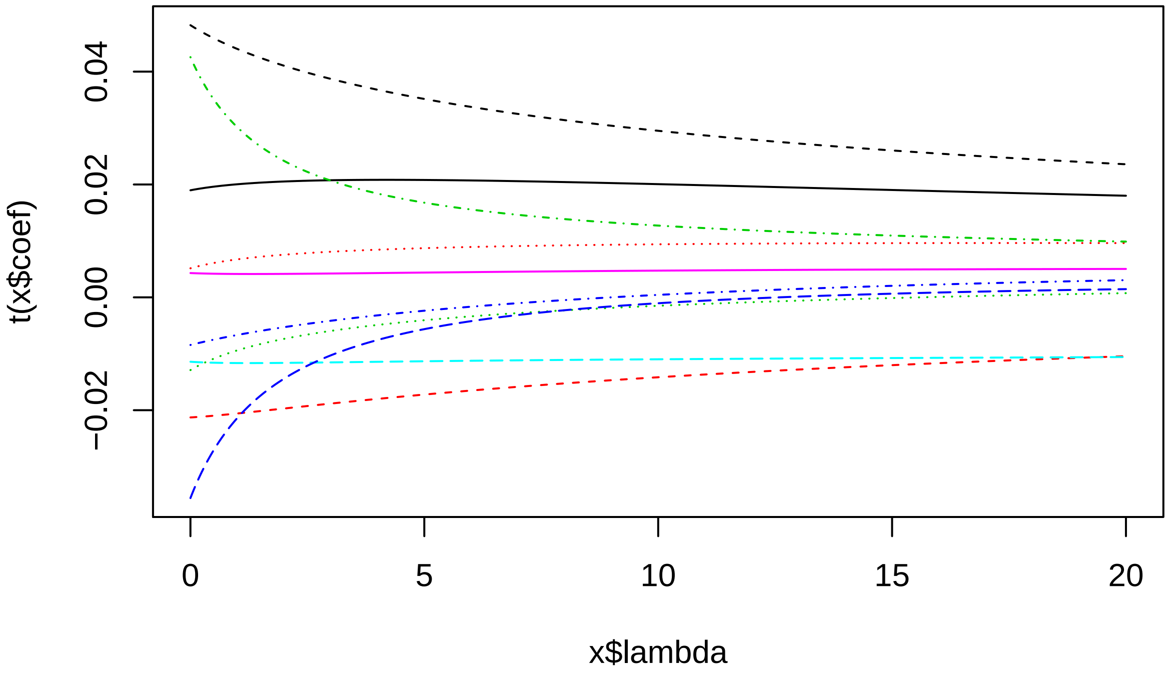
```
library(MASS)
```

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

```
## The following object is masked from 'package:EnvStats':
##
##     boxcox
```

```
## The following objects are masked from 'package:SMPracticals':
##
##     cement, forbes, leuk, shuttle
```

```
rfit <- lm.ridge(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + so + log(nox) + log(hc),da
plot(rfit)
```



```r
select(rfit)
```

```
## modified HKB estimator is 1.421146
## modified L-W estimator is 2.803817
## smallest value of GCV  at 1
```

These three estimators are three estimations of the ridge constants.

```r
coef(rfit)[which.min(rfit$GCV),]
```

```
##                        prec          jant          jult          popn
##  7.215998e+00  2.023964e-03 -2.041086e-03 -1.990832e-03 -4.974786e-02
##          educ          dens          nonw            so      log(nox)
## -1.389164e-02  2.873315e-06  4.977025e-03  1.070660e-04  2.568260e-02
##       log(hc)
## -1.836020e-02
```

(5)

```r
fit_lqs <- lqs(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + so + log(nox) + log(hc), da
fit_lqs
```

```
## Call:
## lqs.formula(formula = log(mort) ~ prec + jant + jult + popn +
##     educ + dens + nonw + so + log(nox) + log(hc), data = pollution)
##
## Coefficients:
## (Intercept)         prec         jant         jult         popn
##   8.875e+00   -5.598e-04   -3.169e-03   -4.664e-03   -3.150e-01
##        educ         dens         nonw           so     log(nox)
##  -5.672e-02    5.407e-06    1.089e-02   -3.651e-04    6.879e-02
##     log(hc)
##  -6.776e-02
```

```
##
## Scale estimates 0.02263 0.02341
```

```
fit_lqs <- lqs(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + so, data = pollution)
fit_lqs
```

```
## Call:
## lqs.formula(formula = log(mort) ~ prec + jant + jult + popn +
##     educ + dens + nonw + so, data = pollution)
##
## Coefficients:
## (Intercept)         prec         jant         jult         popn
##   6.494e+00    3.214e-03    1.197e-03   -2.280e-03    1.363e-01
##        educ         dens         nonw           so
##  -1.624e-02    2.114e-05    8.166e-04    5.853e-05
##
## Scale estimates 0.02134 0.02628
```

```
fit_lqs <- lqs(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + so + nox + hc, data = pollu
fit_lqs
```

```
## Call:
## lqs.formula(formula = log(mort) ~ prec + jant + jult + popn +
##     educ + dens + nonw + so + nox + hc, data = pollution)
##
## Coefficients:
## (Intercept)         prec         jant         jult         popn
##   8.643e+00   -1.169e-03   -2.556e-03   -5.433e-03   -2.377e-01
##        educ         dens         nonw           so          nox
##  -5.815e-02    9.130e-06    1.087e-02   -2.734e-04    7.098e-04
##          hc
##  -4.022e-04
##
## Scale estimates 0.02255 0.02437
```

Compared using scale estimates, the model with all pollutions and taking log on nox and hc is better. It's also the best model before.

```
fit_rlm <- rlm(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + so + log(nox) + log(hc), da
summary(fit_rlm)
```

```
##
## Call: rlm(formula = log(mort) ~ prec + jant + jult + popn + educ +
##     dens + nonw + so + log(nox) + log(hc), data = pollution)
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.085720 -0.020783  0.002227  0.016942  0.068745
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept)  7.3397  0.2583    28.4107
## prec         0.0018  0.0007     2.6426
## jant        -0.0020  0.0006    -3.0768
## jult        -0.0028  0.0018    -1.6171
## popn        -0.0537  0.0478    -1.1232
## educ        -0.0174  0.0072    -2.4048
## dens         0.0000  0.0000     1.2478
## nonw          0.0055  0.0010     5.5937
## so            0.0001  0.0001     0.5128
## log(nox)      0.0378  0.0147     2.5739
## log(hc)      -0.0325  0.0155    -2.0976
##
## Residual standard error: 0.0259 on 49 degrees of freedom
```

```
fit_rlm <- rlm(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + so, data = pollution)
summary(fit_rlm)
```

```
##
## Call: rlm(formula = log(mort) ~ prec + jant + jult + popn + educ +
##     dens + nonw + so, data = pollution)
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.087507 -0.020622  0.001456  0.017104  0.095952
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept)  7.2437  0.2581    28.0611
## prec         0.0018  0.0007     2.7238
## jant        -0.0019  0.0007    -2.8618
## jult        -0.0021  0.0014    -1.5545
## popn        -0.0461  0.0526    -0.8755
## educ        -0.0166  0.0079    -2.0989
## dens         0.0000  0.0000     1.0582
## nonw         0.0049  0.0009     5.2844
## so           0.0002  0.0001     2.4411
##
## Residual standard error: 0.02773 on 51 degrees of freedom
```

```
fit_rlm <- rlm(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + so + nox + hc, data = pollu
summary(fit_rlm)
```

```
##
## Call: rlm(formula = log(mort) ~ prec + jant + jult + popn + educ +
##     dens + nonw + so + nox + hc, data = pollution)
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.086027 -0.018349  0.002801  0.014112  0.081472
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept)  7.2841  0.2619    27.8127
## prec         0.0015  0.0007     1.9811
## jant        -0.0016  0.0007    -2.2373
## jult        -0.0023  0.0014    -1.6431
## popn        -0.0503  0.0520    -0.9671
## educ        -0.0176  0.0076    -2.3183
## dens         0.0000  0.0000     1.4495
## nonw         0.0050  0.0009     5.5276
## so           0.0001  0.0001     0.7600
## nox          0.0012  0.0009     1.2518
## hc          -0.0006  0.0005    -1.3114
##
```

```
## Residual standard error: 0.02351 on 49 degrees of freedom
```

Using robust M-estimation, the best model is using all pollutions and not taking log on it. It's different from that before.