STAT5703 HW3 Ex1

Chao Huang (ch3474), Wancheng Chen (wc2687), Chengchao Jin (cj2628)

Exercise 1

```r
library(SMPracticals)
```

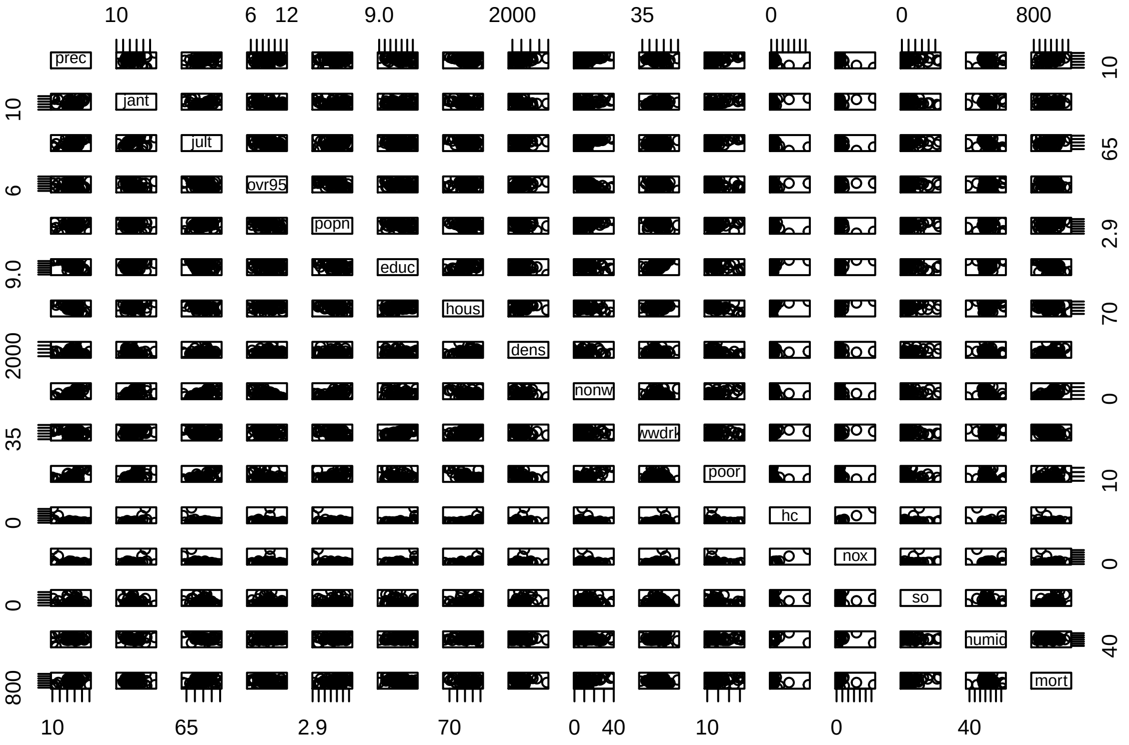
```
## Loading required package: ellipse

##
## Attaching package: 'ellipse'

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

```r
data <- pollution
```

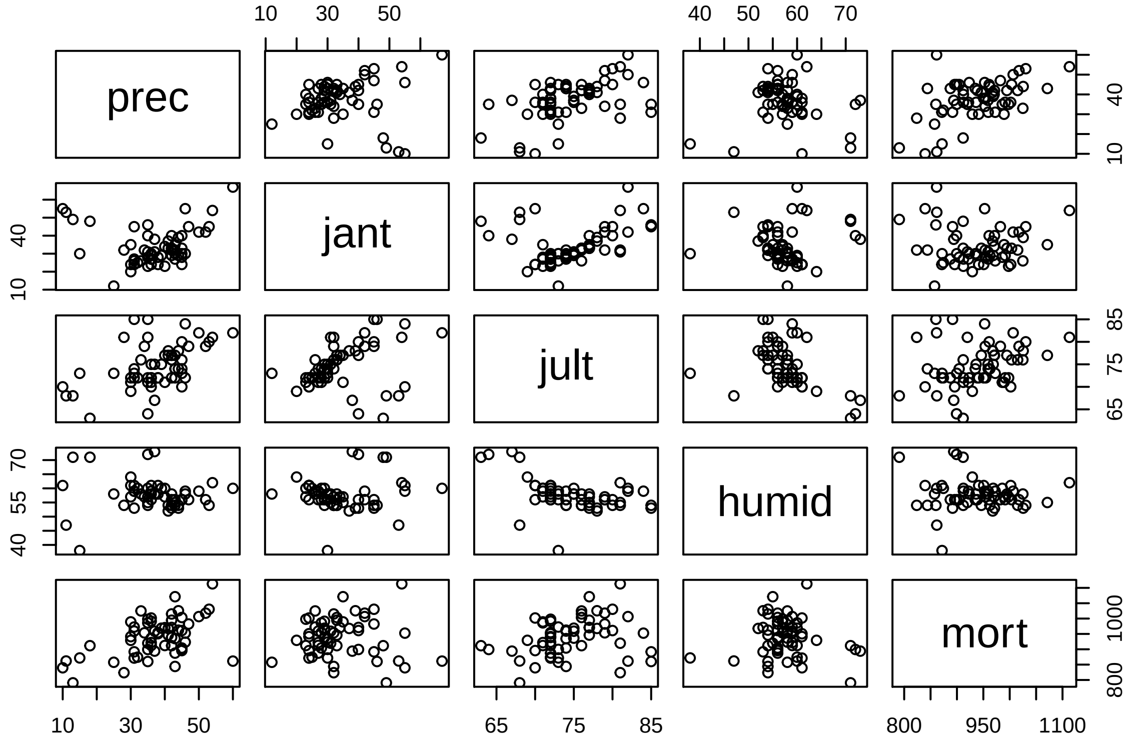
Question 1. Initial Examination

Examine these plots carefully, and comment. Are there outliers? Should covariates and/or the response be transformed? What difficulties might arise in accounting for the effect of air pollution on mortality?

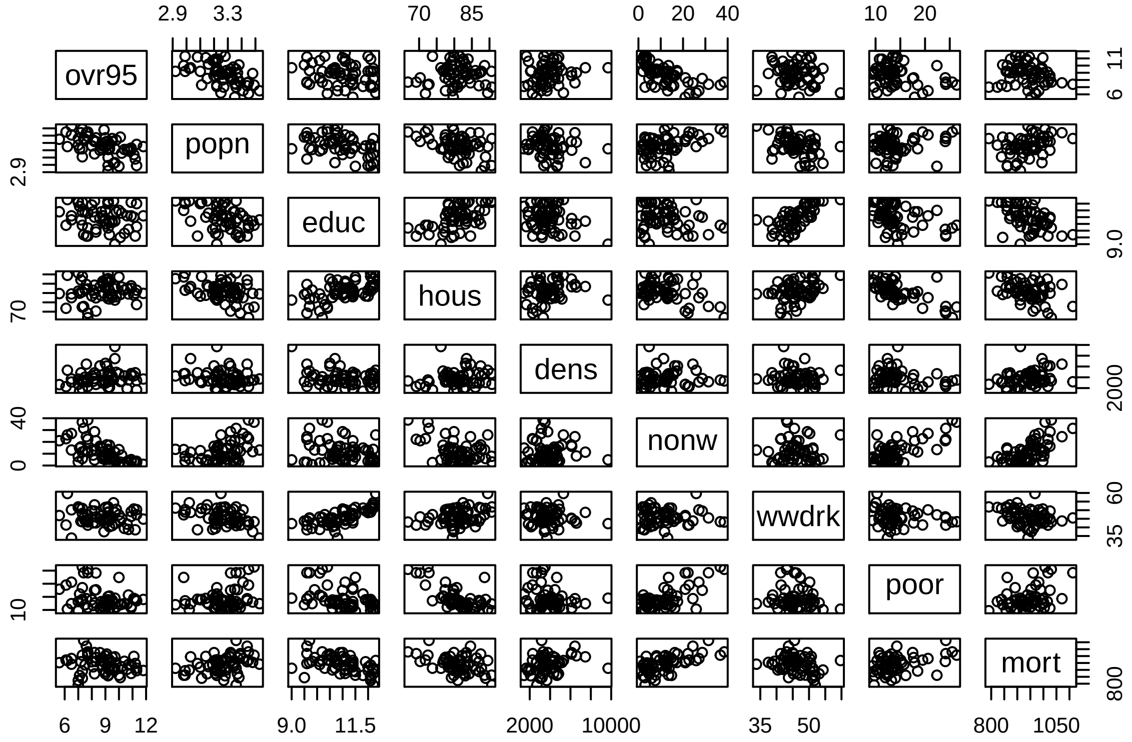
```r
pairs(pollution)
```



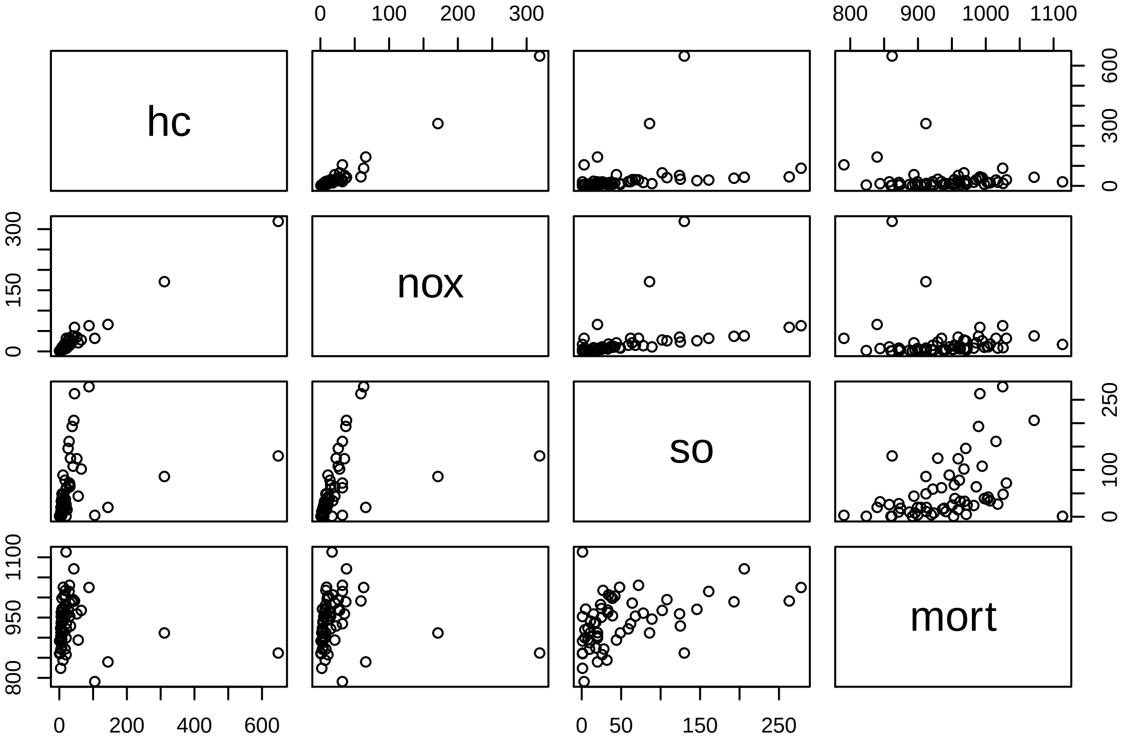
```r
pairs(pollution[,c(1:3,15:16)]) # weather
```



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



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



There are outliers in variables “hc” and “nox”, which can be observed from the pair plot of all variables. Some variables have quite different scales and distributions compared with others, so transformation is needed. By looking at the final plot, we can see that first the outlier may affect the regression results, second some pollutions seem to be strongly correlated, for example nox and so. Therefore, we may have multicolinearity problem.

Problem 2. Model selection

Should all the variables be included? Try various models, choose one or perhaps a few that you think are similarly adequate, give careful interpretations of the covariate effects, and discuss their plausibility. Check the adequacy of your model.

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

```
## 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
```

We choose the final model given by Step command with the lowest AIC value. In this model, only 7 covariates are selected. We can run a seperate linear model for this subset to see the summary of regression.

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

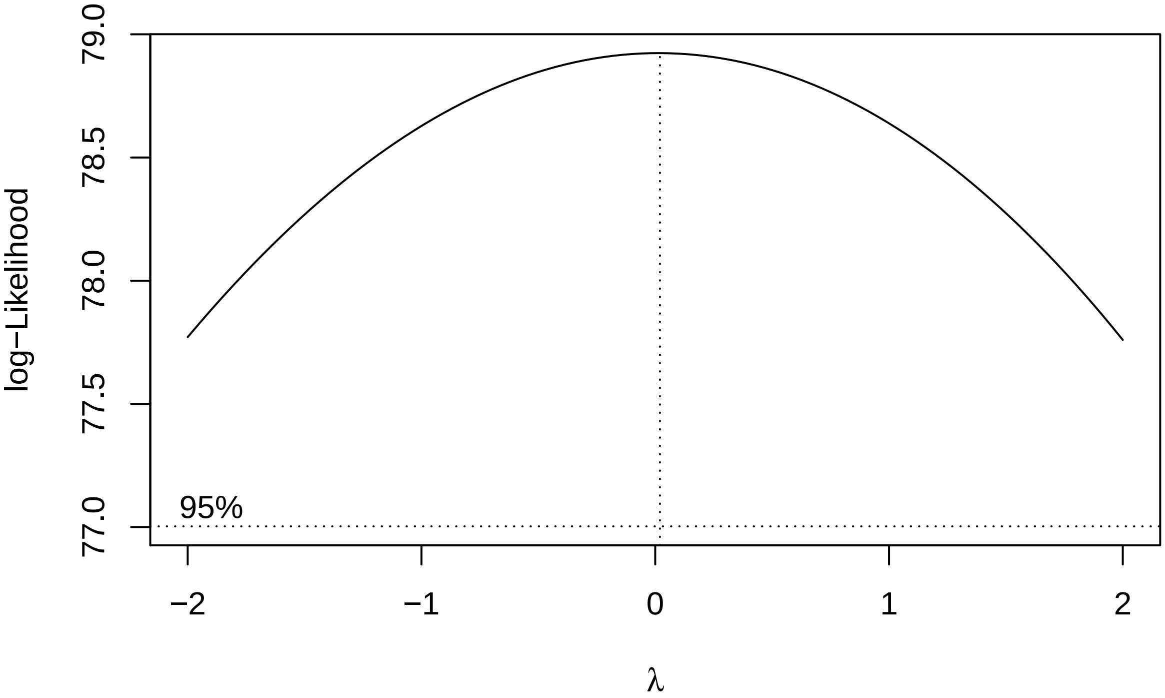
```
## 
## Call:
## lm(formula = mort ~ prec + jant + jult + popn + educ + dens + 
##     nonw, data = data, subset = -hc - nox - so)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -46.399 -24.065  -1.021  12.800  78.910 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.624e+03  2.783e+02   5.834 1.76e-06 ***
## prec         9.963e-01  8.101e-01   1.230  0.22772    
## jant        -1.720e+00  7.368e-01  -2.335  0.02600 *  
## jult        -2.973e+00  1.538e+00  -1.933  0.06208 .  
## popn        -6.280e+01  5.931e+01  -1.059  0.29762    
## educ        -2.806e+01  8.986e+00  -3.123  0.00378 ** 
## dens         3.898e-04  4.286e-03   0.091  0.92810    
## nonw         5.490e+00  9.402e-01   5.840 1.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 34.9 on 32 degrees of freedom
## Multiple R-squared:  0.7292, Adjusted R-squared:   0.67
## F-statistic: 12.31 on 7 and 32 DF,  p-value: 1.599e-07
```

As we can see here, although we have filtered out some covariates, only three remaining covariates are exactly significant, namely jant (temperature in Jan), educ (education) and nonw (non-white). For jant, it has a negative slope, which indicates that a low temperature in January will increase the mortality, which makes sense since a colder winter may kill more people. Similar pattern applies for educ variable, indicating that people with lower education level may have a higher risk of dying, which still makes sense. For nonw, the mortality increases with the percent of non white people in this district, which indicates that perhaps non-white people may have higher risk of dying. All these effects sound reasonable.

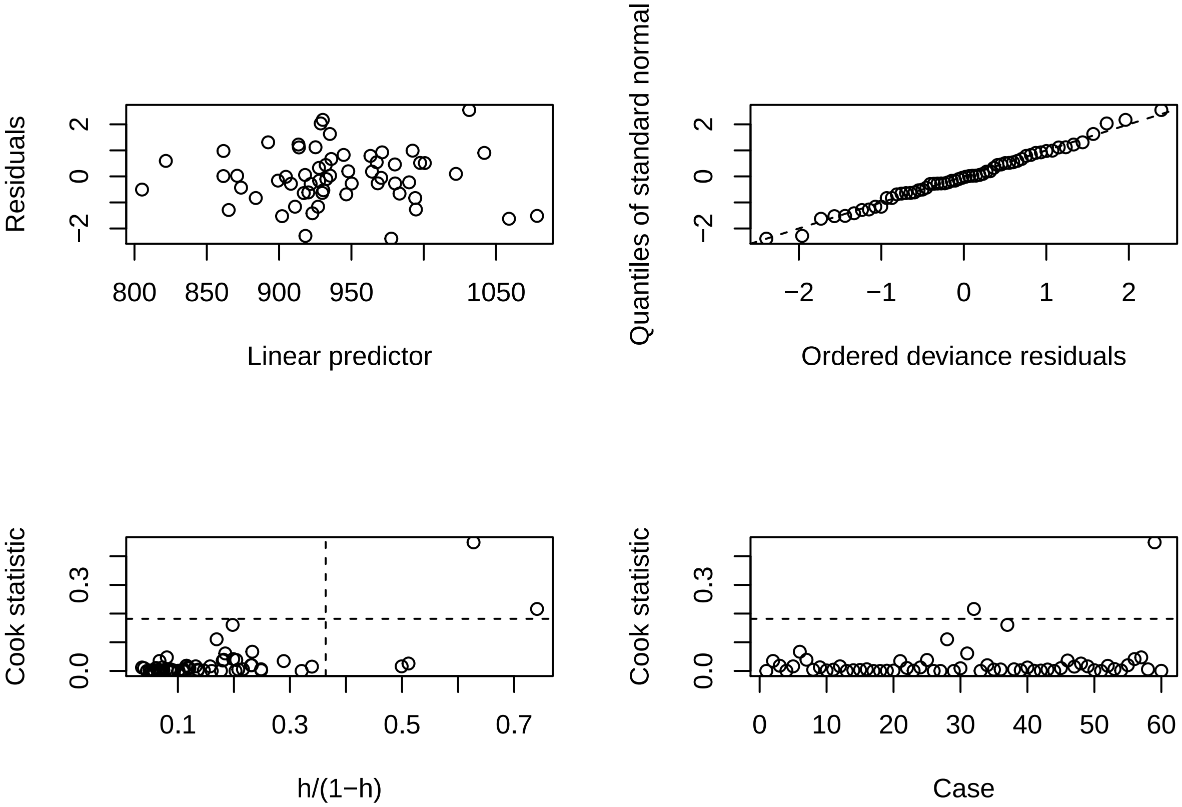
```r
library(MASS)
```

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

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

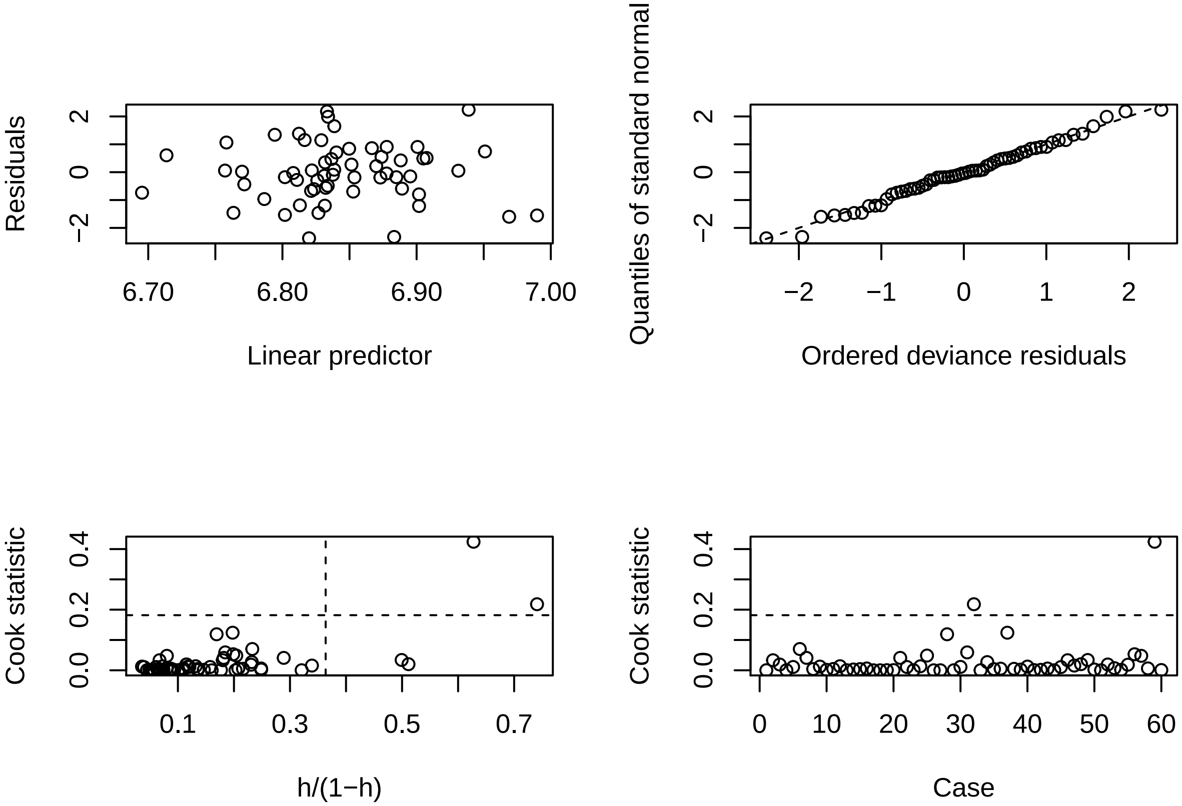
```r
boxcox(fit)
```



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



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

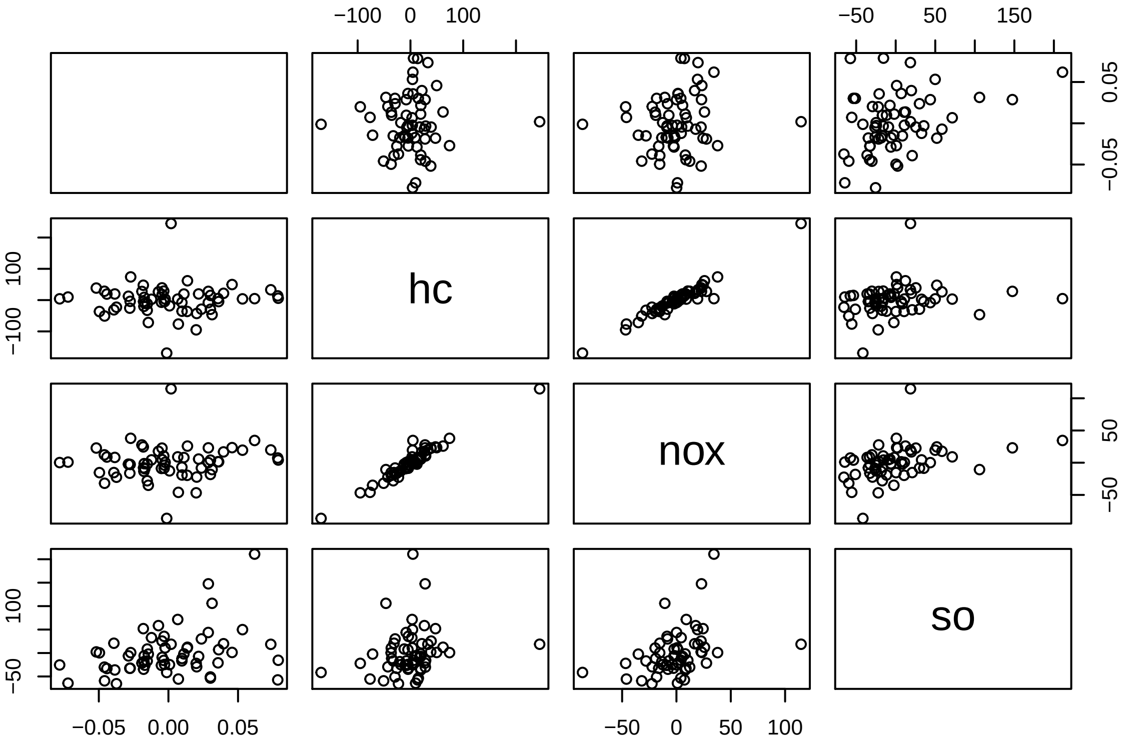


The likelihood plot of boxcox shows that $\lambda = 0$ is the best choice, which indicates a log-transform. However, whether applying this transform or not doesn't have a very significant difference on the final results. The residuals seem to have a normal distribution as our assumption. The linear model has a R^2 value of 0.7292. So using these variables, the linear model fits data well.

Problem 3. Added variable plots

What difficulties do you foresee for regression on all three pollution variables? Are outliers present? Try adding in these variables, or suitable transformations of them, to your chosen best model (or models) from above, and discuss the interpretation and fit of the various models.

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



The added variable plot can help us identify the correlation between some covariates and the response variable, when fix all the other covariates. It can be seen from the first line that only so seems to have a significant linear relation with the response. Also hc and nox contain outliers as we observed before. Now we add all three covariates to our model.

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

```
##
## Call:
## lm(formula = mort ~ prec + jant + jult + popn + educ + dens +
##     nonw + hc + nox + so, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.272 -17.715   0.389  16.729  87.737
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.402e+03  2.448e+02   5.727 6.16e-07 ***
## prec         1.304e+00  6.992e-01   1.865   0.0681 .  
## jant        -1.597e+00  6.592e-01  -2.423   0.0191 *  
## jult        -2.572e+00  1.332e+00  -1.932   0.0592 .  
## popn        -5.429e+01  4.857e+01  -1.118   0.2692    
## educ        -1.552e+01  7.106e+00  -2.184   0.0338 *  
## dens         3.512e-03  3.615e-03   0.971   0.3361    
## nonw         5.168e+00  8.530e-01   6.058 1.90e-07 ***
## hc          -5.857e-01  4.359e-01  -1.344   0.1852    
## nox          1.114e+00  8.748e-01   1.274   0.2087    
## so           9.988e-02  1.267e-01   0.788   0.4345    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 33.65 on 49 degrees of freedom
## Multiple R-squared:  0.757,  Adjusted R-squared:  0.7074 
## F-statistic: 15.26 on 10 and 49 DF,  p-value: 6.765e-12
```

Including all three pollutions doesn't seem to be a good choice, since none of them appear to be significant. So we will only include "so" next.

```
summary(lm(mort ~ prec + jant + jult + popn + educ + dens + nonw + so,data=data))
```

```
## 
## Call:
## lm(formula = mort ~ prec + jant + jult + popn + educ + dens + 
##     nonw + so, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -75.048 -18.365  -2.731  15.903  93.660 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.360e+03  2.321e+02   5.861 3.39e-07 ***
## prec         1.676e+00  6.063e-01   2.765  0.00790 ** 
## jant        -1.820e+00  5.975e-01  -3.046  0.00366 ** 
## jult        -2.309e+00  1.238e+00  -1.865  0.06789 .  
## popn        -4.981e+01  4.729e+01  -1.053  0.29723    
## educ        -1.543e+01  7.090e+00  -2.176  0.03417 *  
## dens         2.957e-03  3.585e-03   0.825  0.41341    
## nonw         5.147e+00  8.338e-01   6.173 1.10e-07 ***
## so           2.041e-01  8.575e-02   2.380  0.02108 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 33.61 on 51 degrees of freedom
## Multiple R-squared:  0.7476, Adjusted R-squared:  0.708 
## F-statistic: 18.89 on 8 and 51 DF,  p-value: 8.489e-13
```

By including "so", it seems significant in the new linear model and indicates that higher "so" pollution may increase mortality. Also it makes the prec (percipitation) significant, and it seems that the rain will increase mortality as well.

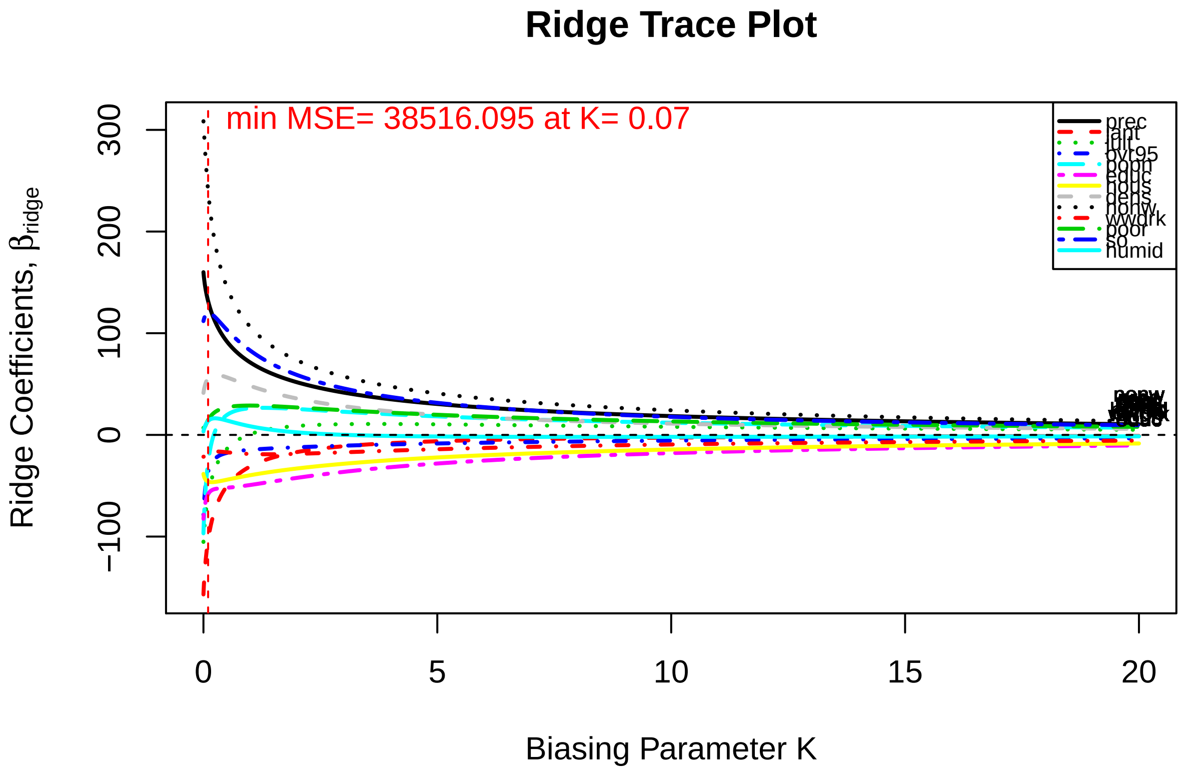
Problem 4. Fit ridge regression model

Discuss the interpretability of the resulting parameter estimates.

```r
library(MASS)
rfit <- lm.ridge(mort~.-hc-nox,data=data,lambda=seq(0,20,0.01))
select(rfit)
```

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

```r
mod <- lmridge::lmridge(mort~.-hc-nox,data, K=seq(0,20,0.01))
lmridge::plot.lmridge(mod)
```



From the trace plot we can see that, two covariates are super sensitive to the value of K, namely nonw and jant. Therefore, the the significance of these two covariates shown in previous results is questionable, since real significant covariates will maintain a similar level of coefficients when K changes.

Problem 5. Other models

Try using the functions lqs in library(lqs) for least trimmed squares regression, and rlm in library(MASS) for robust M-estimation, and see if your conclusions change.

```r
lqs(mort~.-hc-nox, data)
```

```
## Call:
## lqs.formula(formula = mort ~ . - hc - nox, data = data)
##
## Coefficients:
## (Intercept)         prec         jant         jult        ovr95
##   1.629e+03    4.114e+00    7.723e-01   -1.540e+00   -1.380e+01
##        popn         educ         hous         dens         nonw
##  -8.325e+01   -2.189e+01   -9.661e-01    7.982e-03    1.025e+00
##       wwdrk         poor           so        humid
##  -3.204e+00   -1.557e+00    3.832e-01    9.205e-01
##
## Scale estimates 21.13 23.02
```

```
rlm(mort~.-hc-nox, data)
```

```
## Call:
## rlm(formula = mort ~ . - hc - nox, data = data)
## Converged in 13 iterations
##
## Coefficients:
##   (Intercept)          prec          jant          jult         ovr95
## 1672.32908736    2.06162817   -1.62914419   -2.61950579   -7.26033069
##          popn          educ          hous          dens          nonw
##  -78.81287595  -10.72722304   -1.55319706    0.00551085    4.01945299
##         wwdrk          poor            so         humid
##   -1.07746593   -0.92756518    0.24593566   -0.25177190
##
## Degrees of freedom: 60 total; 46 residual
## Scale estimate: 26.5
```

They both suggest that there exists a positive correlation between “so” and mortality.