HW3 Exercise 5

Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)

4/15/2020

```
library(SMPracticals)
```

```
## Loading required package: ellipse
```

```
##
## Attaching package: 'ellipse'
```

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

```
library(MASS)
```

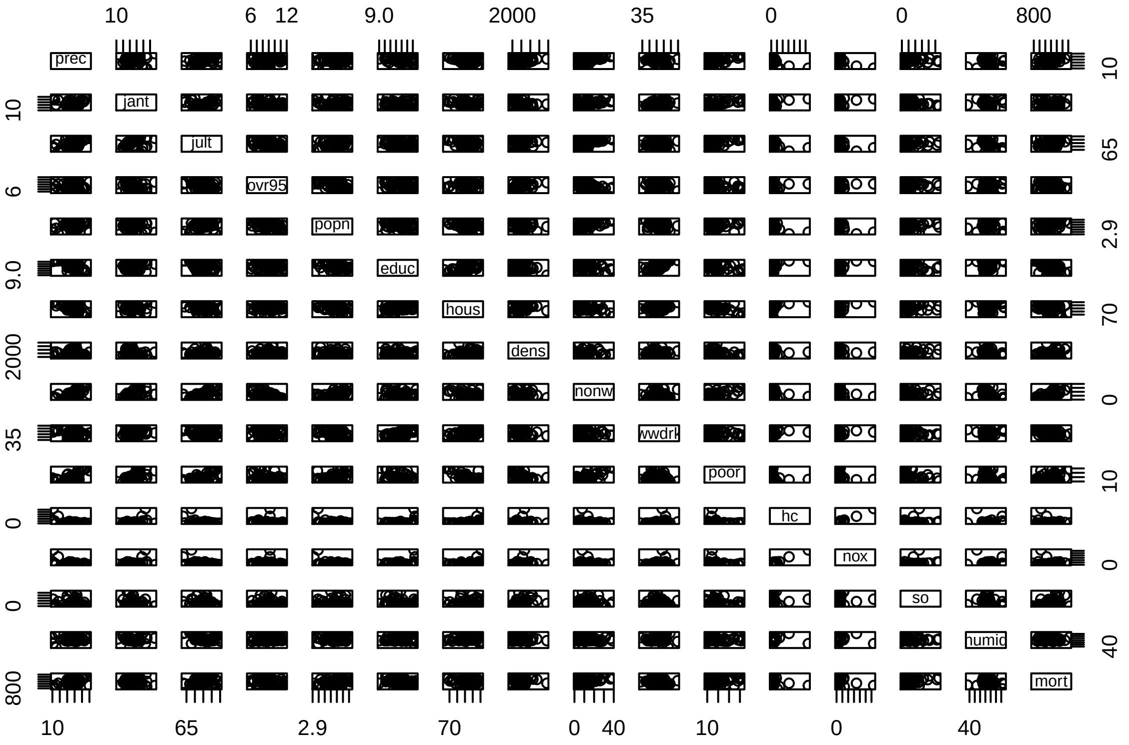
```
##
## Attaching package: 'MASS'
```

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

```
data(pollution)
```

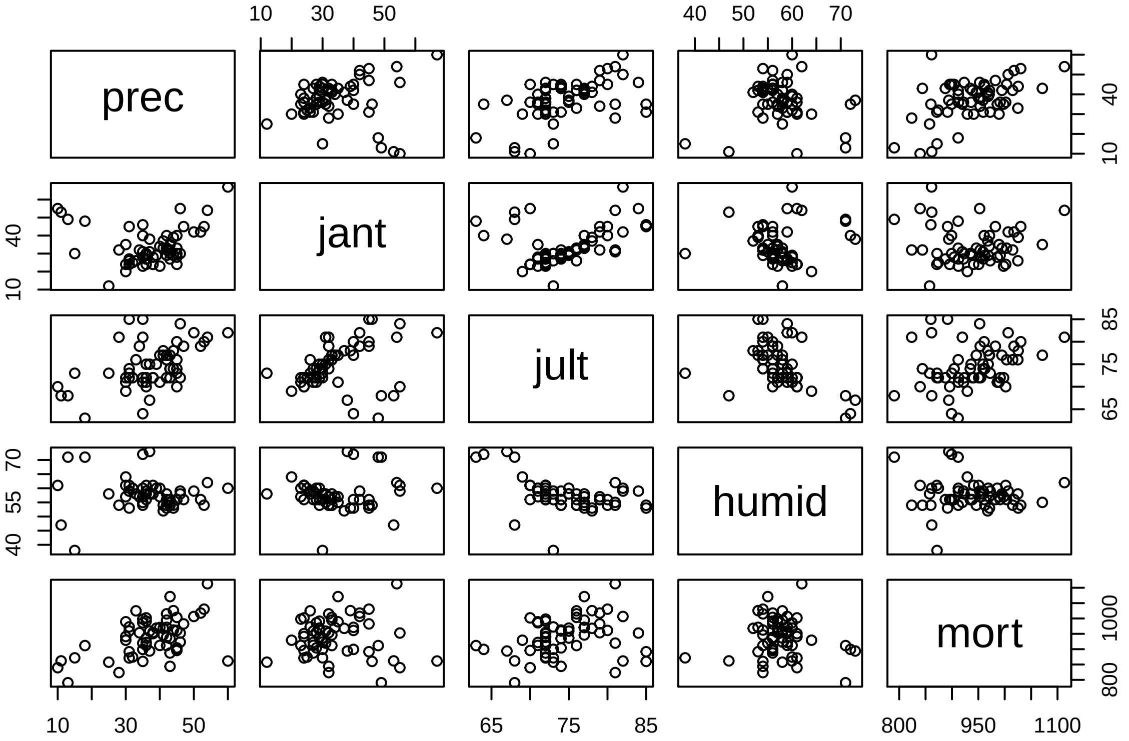
Problem 1

```
pairs(pollution)
```



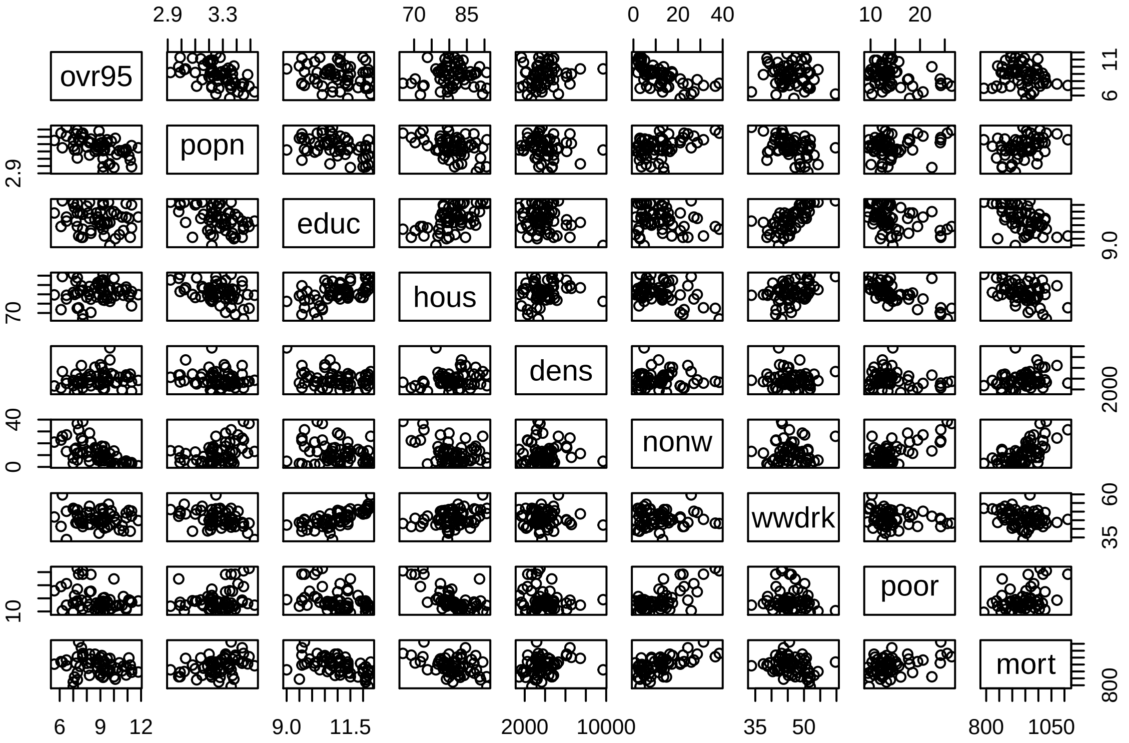
Since the plots are too dense, we can not distinguish any patterns from it.

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



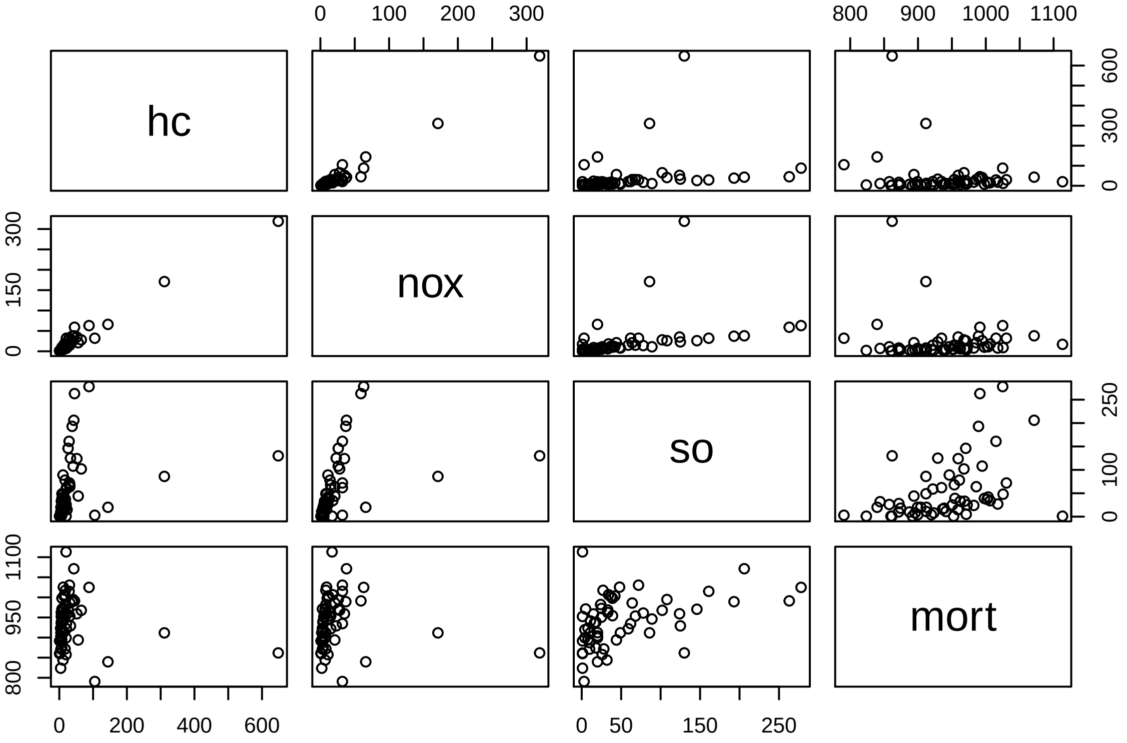
We do not see any patterns among weathers and mortality in these plots as the points are scattered across the plots; we could not recognize if there is outlier either.

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



We don't see any significant covariates from the plots and we could not recognize if there is outlier.

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



We still could not see any significant covariates from the plots, but we see that there might be possbile outliers in **hc** and **noc** since one or two points are apart from the rest of the points, which appear to form a cluster. We should consider transforming the dataset as most of the features don't seem to correlate with mortality. The issue of outliers and collinearity among the variables might arise in accounting for the effect of air pollution on mortality.

Problem 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
```

The model associated with the lowest AIC contains variables prec, jant, jult, popn, educ, dens, and nonw, so we want to run a reduced model using those 7 variables

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

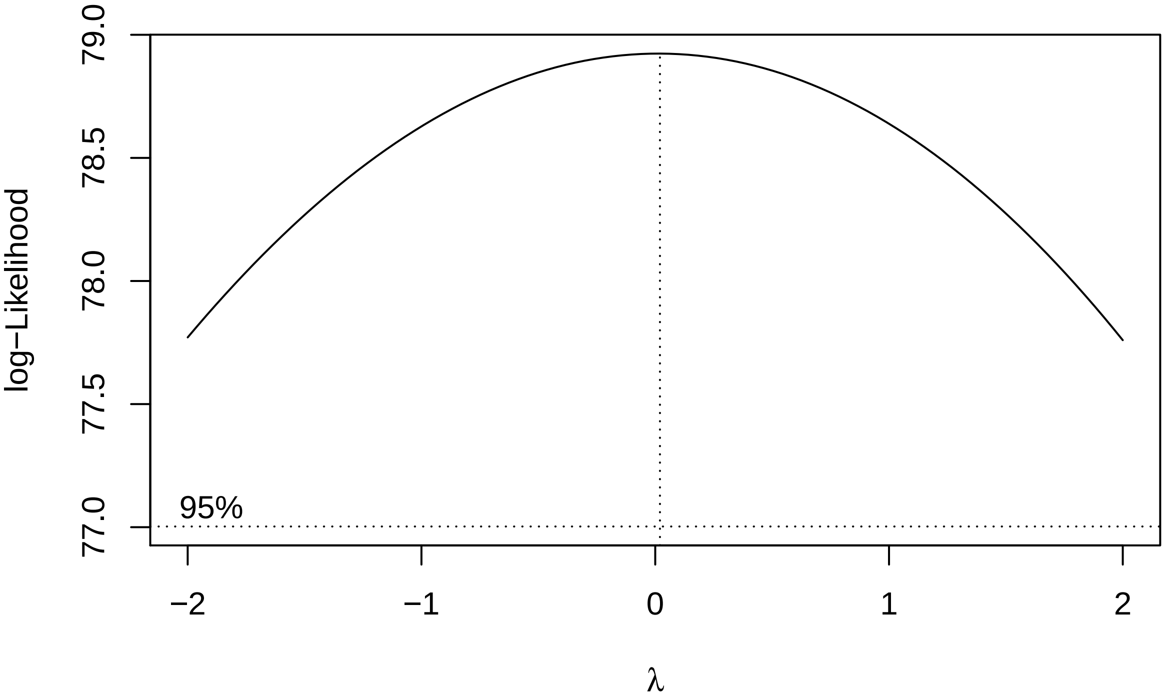
```
##
## Call:
## lm(formula = mort ~ prec + jant + jult + popn + educ + dens +
##     nonw, data = pollution)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.148 -20.837  -1.231  19.548  81.714
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  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
##
```

```
## Residual standard error: 35.09 on 52 degrees of freedom
## Multiple R-squared:  0.7196, Adjusted R-squared:  0.6819
## F-statistic: 19.06 on 7 and 52 DF,  p-value: 2.438e-12
```

The reduced model has an adjusted R-squared of 0.68, meaning 68% of the variance could be explained by those 7 variables. Keeping all else constant:

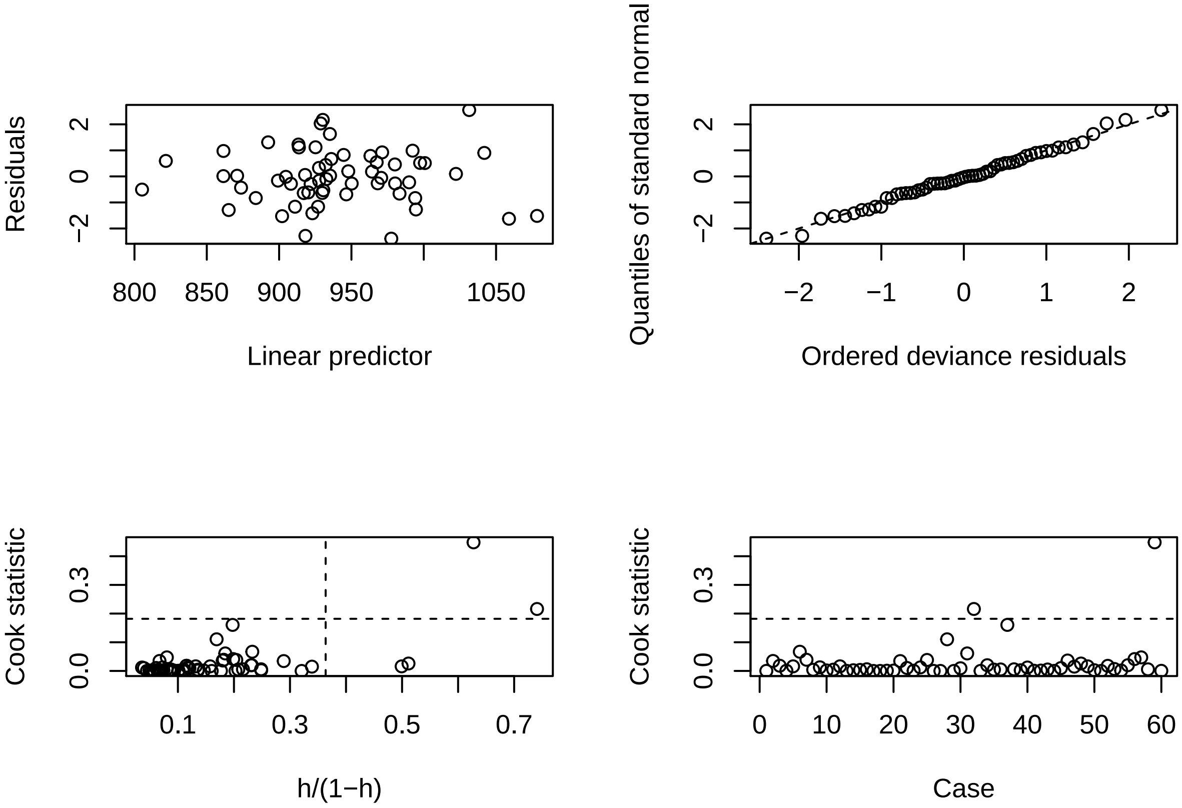
- an increase in average annual precipitation would increase the mortality and this change is statistically significant. It might sound plausible as high precipitation could cause flood.
- a decrease in average January temperature would increase the mortality and this change is statistically significant. It sounds plausible as the elderly are especially susceptible to cold weather.
- a decrease in average July temperature would increase the mortality and this change is statistically significant. It sounds plausible due to the same reason as the previous one.
- household size is not a significant variable, which is reasonable as household size has nothing to do with mortality.
- a decrease in median school years completed by those over 22 would increase the mortality and this change is statistically significant. It sounds plausible since the level of education might affect one's knowledge in nutrition and healthy diet.
- Population per square mile in urbanized areas in 1960 is not a significant variable.
- an increase in percentage non-white population in urbanized areas in 1960 would increase the mortality and this change is statistically significant, which is not reasonable.

```
boxcox(fit)
```



The box-cox plot suggests that a log transformation might be appropriate as $\lambda = 0$ falls inside the 95% confidence interval.

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



As we can see, homoscedasticity and normality condition are generally met as we don't see obvious pattern in the residual v.s. fitted value plot, nor do we see many points apart from the normal QQ-line. By checking the Cook's distance, we can see there are indeed some outliers.

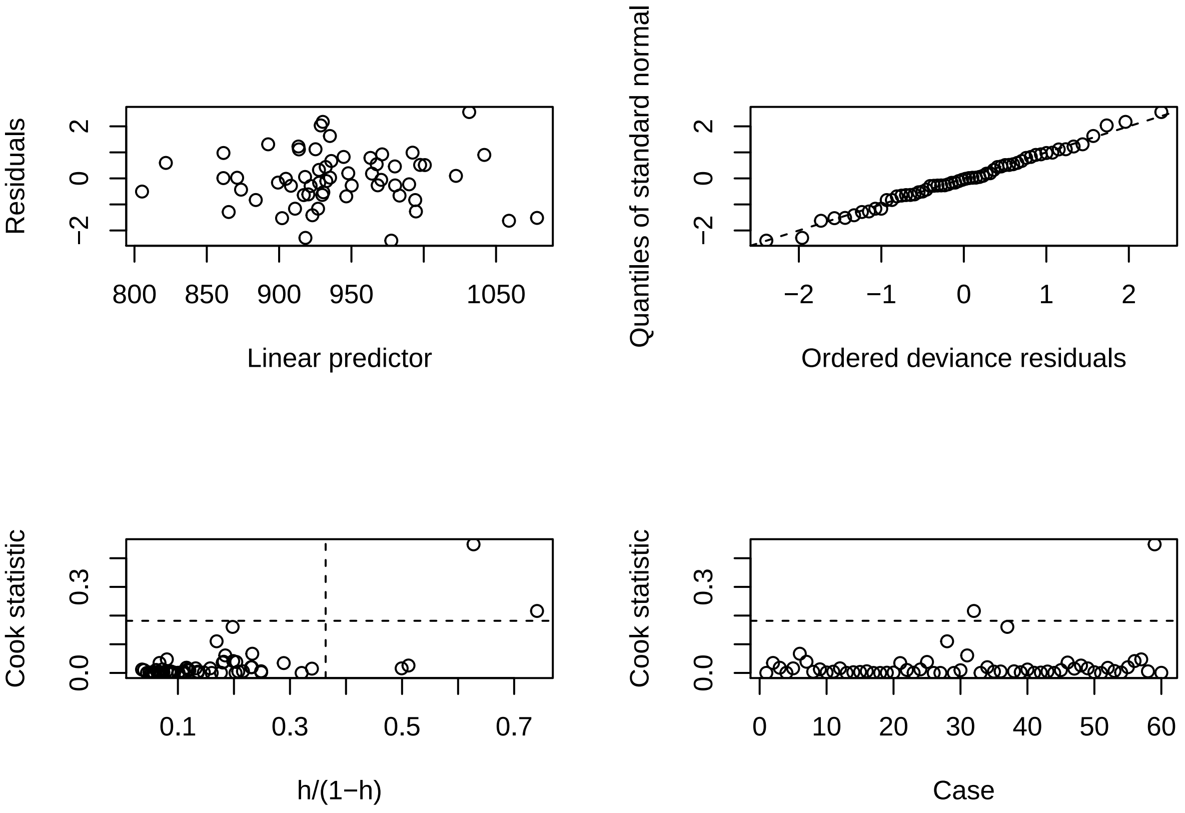
```
fit2 <- update(fit,log(mort)~.) # try log transform of response plot.glm.diag(fit) # model adequate?
summary(fit2)
```

```
##
## 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
```

We can see the conclusion is the same as the reduced model with no log-transformation applied.

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

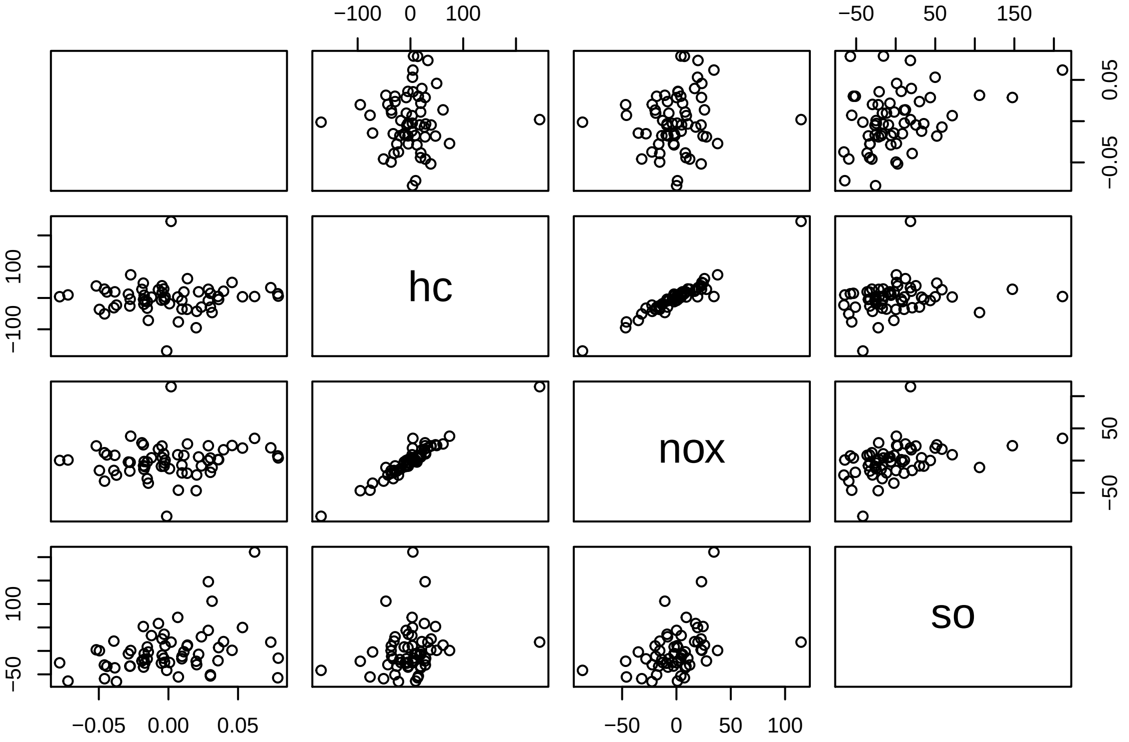


Again, homoscedasticity and normality conditions are satisfied, so the log transform version of the reduced model is adequate as well.

Hence, using the reduced model resulted from step function and apply log transformation as suggested by the box-cox plot would be the chosen model.

Problem 3

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



We see that **hc** and **nox** are not appropriate to be fitted using a linear model as there is obvious pattern in the residual plots, only **so** seems to have a significant linear relationship with the other variables and there seem to be outliers in all three pollution variables.

```
fit3 <- lm(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + hc + nox + so, data = pollution
summary(fit3)
```

```
##
## Call:
## lm(formula = log(mort) ~ prec + jant + jult + popn + educ + dens +
##     nonw + hc + nox + so, 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 ***
## hc          -6.469e-04  4.616e-04  -1.401  0.16743
## nox          1.242e-03  9.266e-04   1.340  0.18633
## so           9.900e-05  1.342e-04   0.737  0.46434
## ---
## 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
```

After adding the three pollution variables to the reduced model, the adjusted r-squared has improved by around 3%. From the previous conclusion, we want to log-transform **hc** and **nox** to eliminate the patterns in the residual plot.

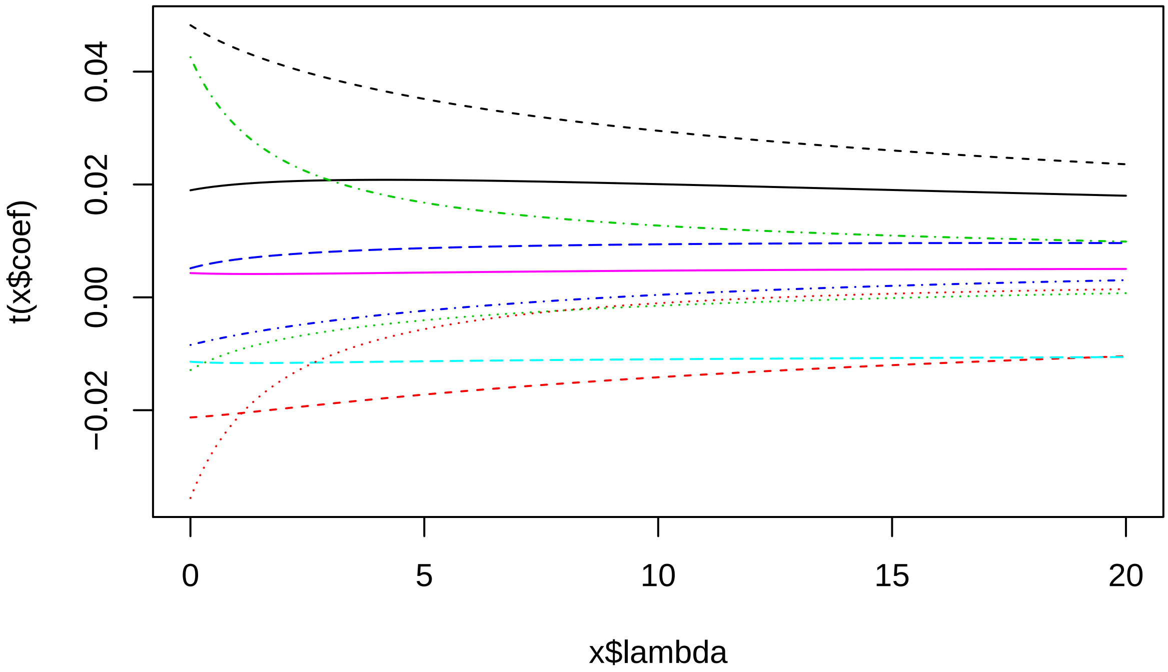
```
fit4 <- lm(log(mort) ~ prec + jant + jult + popn + educ + dens + nonw + log(hc) + log(nox) + so, data =
summary(fit4)
```

```
##
## Call:
## lm(formula = log(mort) ~ prec + jant + jult + popn + educ + dens +
##     nonw + log(hc) + log(nox) + so, 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 ***
## log(hc)     -3.051e-02  1.565e-02  -1.949  0.05700 .
## log(nox)     3.625e-02  1.484e-02   2.443  0.01822 *
## so           8.186e-05  1.169e-04   0.700  0.48716
## ---
## 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 the adjusted r-squared is further improved. Hence, the reduced model with added variables **so**, **log(hc)**, and **log(nox)** is a better model.

Problem 4

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



As we can see, as the penalty term λ increases, the coefficient of the varialbes tend to approache 0, and two of the variables are especially sensitive to the value of λ as their curvature are large.

```
select(rfit)
```

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

Those three values are estimates of the penalty term using different methods.

Problem 5

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

```
## Call:
## lqs.formula(formula = log(mort) ~ prec + jant + jult + popn +
##     educ + dens + nonw + log(hc) + log(nox) + so, data = pollution)
##
## Coefficients:
## (Intercept)         prec         jant         jult         popn
##   8.347e+00   -1.748e-04   -2.495e-03   -4.637e-03   -1.790e-01
##        educ         dens         nonw      log(hc)     log(nox)
##  -5.362e-02    9.315e-06    8.147e-03   -3.095e-02    3.263e-02
##          so
##  -2.237e-04
##
## Scale estimates 0.01837 0.01879
```

Using least trim square regression, we see that popn appear to be statistically more import than prec, which doesn't agree with what we have previously.

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

```
##
## Call: rlm(formula = log(mort) ~ prec + jant + jult + popn + educ +
##     dens + nonw + log(hc) + log(nox) + so, 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
## log(hc)     -0.0325  0.0155    -2.0976
## log(nox)     0.0378  0.0147     2.5739
## so           0.0001  0.0001     0.5128
##
## Residual standard error: 0.0259 on 49 degrees of freedom
```

From the t-values, we can see that using the robust M-estimation generally results in the same conclusion as what we have previously.