Modelling Effect of Pollution on Mortality

Noam Benkler & Serafina Chen

Introduction:

Over the past hundred years or so, pollution has emerged as one of the most serious concerns to modern humanity. The components of pollution, can be either natural contaminants or foreign substances. Once the environment is polluted, it can lead to serious health problems and may cause death. In this study we examine the relationship between relative pollution potentials and human mortality rates, controlling for socioeconomic and climatological effects on mortality. More specifically, we will look at relative pollution potential of hydrocarbons, oxides of nitrogen and Sulphur dioxide to determine the relationship between the mortality level and the pollution.

Data:

Our data set consists of 60 observations on 17 variables. However, after examining VIF values and finding evidence of multicollinearity we condensed the data set into four: “Climate,” which accounted for annual precipitation, mean January temperature, mean July Temperature, and percent relative humidity, “Pop,” which accounted for population per household, population density, percentage of the household that is sound with all facilities, and percentage of the population over 65 years old, “Wealth,” which accounted for percentage of the population that was not white, percentage of the population employed in white collar occupations, percentage of households with annual incomes under $3,000, and median number of school years completed for persons 25 or older, and “Pollution,” which accounted for relative pollution potential of hydrocarbons, relative pollution potential of oxides of nitrogen, and relative pollution potential of sulfur dioxide. While our Influential index plots showed some outliers, it did not appear that we could eliminate them from our data set without diving down a rabbit hole of continuously eliminating points.

Results:

Our model includes four variables: Pollution, Climate, Wealth, Pop.

$$\boldsymbol{\mu}[\text{Mortality} \mid \text{Pollution} + \text{Climate} + \text{Wealth} + \text{Pop}] = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1(\text{Pollution}) + \boldsymbol{\beta}_2(\text{Climate}) + \boldsymbol{\beta}_3(\text{Wealth}) + \boldsymbol{\beta}_4(\text{Pop})$$

We found that, one unit of relative pollution potential increase in “pollution” is associated with 0.020 unit decrease in mortality level when holding all other components. One unit of relative pollution potential increase in “climate” is associated with 0.193 unit increase in mortality level when holding all other components. One unit of relative pollution potential increase in “pop” is associated with 0.014 unit increase in mortality level when holding all other components. One unit of relative pollution potential increase in “wealth” is associated with 2.154 unit increase in mortality level when holding all other components. This model explains 24.6% of variability of response variables.

The estimates of model parameters are given in Table 1.

Discussion:

While our model coefficients reported a negative relationship between relative pollution potentials and mortality, the p-value obtained by our model showed that the relationship is not significantly different from zero, meaning we cannot conclude that relative pollution potentials have any significant effect on mortality level.

Our model presents several other problems. While we decided not to remove some of the outliers from the data set due to inability to conclude they belonged to a population other than the one under investigation, there were several outliers with non negligible leverage values in our data which could have strongly affected our model. Furthermore, the accuracy of our model is questionable as it seems counterintuitive that an increase in pollution potentials would have no significant effect on mortality level.

At the end of our analysis we calculated a model using the logged values of NOX, HC, and SO2 instead of the combined “pollution” variable that yielded results showing definitive effects of relative pollution rates on Mortality. However, we made a conscious decision not to use this model as the VIF values for the new coefficients seriously indicated multicollinearity(VIF values greater than 10), meaning the model was biased, and we could not conclude the new model was an accurate representation of the true relationship between relative pollution potentials and mortality level.

Table 1. Model Coefficients

	Dependent variable: Mortality
climate	0.193
	(0.525)
pop	0.014***
	(0.005)
wealth	2.154***
	(0.732)
pollution	-0.020
	(0.046)
Constant	669.318***
	(85.908)
Observations	60
R2	0.297
Adjusted R2	0.246
Residual Std. Error	54.008 (df = 55)
F Statistic	5.815*** (df = 4; 55)
Note:	*p<0.1; **p<0.05; ***p<0.01

Case Study 2: Code Supplement

Noam Benkler & Serafina Chen

May 6, 2018

A.

```
pollution <- Sleuth3::ex1217
tidy(pollution [, 2:17])
```

```
##       column  n        mean           sd   median    trimmed         mad
## 1  Mortality 60  940.356500   62.2018851  943.680  940.05604   65.656941
## 2     Precip 60   37.366667    9.9846775   38.000   38.20833    8.154300
## 3   Humidity 60   57.666667    5.3699309   57.000   57.31250    3.706500
## 4    JanTemp 60   33.983333   10.1688985   31.500   32.91667    8.154300
## 5   JulyTemp 60   74.583333    4.7631768   74.000   74.52083    4.447800
## 6     Over65 60    8.798333    1.4645520    9.000    8.80000    1.556730
## 7      House 60    3.263167    0.1352523    3.265    3.27125    0.118608
## 8       Educ 60   10.973333    0.8452994   11.050   11.00417    0.889560
## 9      Sound 60   80.913333    5.1413731   81.150   81.15417    4.077150
## 10   Density 60 3874.550000 1454.7267252 3567.000 3729.18750 1043.009100
## 11  NonWhite 60   11.870000    8.9211480   10.400   10.68333    8.006040
## 12  WhiteCol 60   46.081667    4.6130431   45.500   46.12292    4.670190
## 13      Poor 60   14.373333    4.1600956   13.200   13.68333    2.075640
## 14        HC 60   37.850000   91.9776732   14.500   18.58333   12.602100
## 15       NOX 60   22.650000   46.3332896    9.000   13.18750    8.895600
## 16       SO2 60   53.766667   63.3904678   30.000   41.10417   32.617200
##         min     max   range        skew    kurtosis          se
## 1    790.73 1113.06  322.33  0.09304575 -0.05048039   8.0302288
## 2     10.00   60.00   50.00 -0.76251780  0.94585817   1.2890163
## 3     38.00   73.00   35.00  0.22576474  3.61575954   0.6932551
## 4     12.00   67.00   55.00  0.91320964  0.77195316   1.3127992
## 5     63.00   85.00   22.00  0.12994643 -0.18455468   0.6149235
## 6      5.60   11.80    6.20 -0.03242915 -0.73046441   0.1890728
## 7      2.92    3.53    0.61 -0.46512225 -0.15533172   0.0174610
## 8      9.00   12.30    3.30 -0.21380710 -0.86161922   0.1091277
## 9     66.80   90.70   23.90 -0.39638391  0.14761806   0.6637484
## 10 1441.00 9699.00 8258.00  1.31211848  2.99073522 187.8044127
## 11     0.80   38.50   37.70  1.07520034  0.63708836   1.1517153
## 12    33.80   59.70   25.90  0.09358734  0.29855627   0.5955413
## 13     9.40   26.40   17.00  1.39116100  1.16599439   0.5370660
## 14     1.00  648.00  647.00  5.31685870 30.61396611  11.8742666
## 15     1.00  319.00  318.00  4.91014708 26.68669488   5.9816020
## 16     1.00  278.00  277.00  1.81727009  2.96033854   8.1836742
```

Creat merged variables

```
pollution$climate <- pollution$Precip + pollution$Humidity + pollution$JanTem
p + pollution$JulyTemp
```

```
pollution$pollution <- pollution$HC + pollution$NOX + pollution$SO2
pollution$pop <- pollution$House + pollution$Density + pollution$Over65 + pol
lution$Sound
pollution$wealth <- pollution$Poor + pollution$WhiteCol + pollution$NonWhite
+ pollution$Educ
```

creating linear models

```
pol.lm <- lm(Mortality ~ Precip + Humidity + JanTemp + JulyTemp + Over65 + Ho
use + Educ + Sound + Density + NonWhite + WhiteCol + Poor + HC + NOX + SO2, d
ata = pollution)
pol2.lm <- lm(Mortality ~ climate + pop + wealth + pollution, data = pollutio
n)
pollog.lm<- lm(Mortality ~ Precip + Humidity + JanTemp + JulyTemp + Over65 +
House + Educ + Sound + Density + NonWhite + WhiteCol + Poor + log(HC) + log(N
OX) + SO2, data = pollution)
stargazer(pol.lm, pol2.lm, pollog.lm, type = "text")
```

```
##
## =============================================================================
==============
##                                               Dependent variable:
##                       -----------------------------------------------------
--------------
##                                                    Mortality
##                                (1)                     (2)
(3)
## -----------------------------------------------------------------------------
--------------
## Precip                       1.905**                                       2.
657***
##                              (0.923)                                       (
0.844)
##
## Humidity                      0.106
0.193
##                              (1.169)                                       (
0.990)
##
## JanTemp                      -1.935*                                       -2
.467**
##                              (1.108)                                       (
0.947)
##
## JulyTemp                      -3.102                                       -
3.478*
##                              (1.901)                                       (
2.052)
##
## Over65                        -9.045                                       -1
```

```
4.255*
##                              (8.483)                                                    (
7.884)
##
## House                        -106.502                                                -13
6.766**
##                              (69.766)                                                 (6
6.641)
##
## Educ                         -17.068                                                   -
14.373
##                              (11.861)                                                 (1
0.540)
##
## Sound                         -0.659                                                   -
0.791
##                              (1.768)                                                    (
1.639)
##
## Density                       0.004
0.003
##                              (0.004)                                                    (
0.004)
##
## NonWhite                     4.460***                                                  3.
778***
##                              (1.326)                                                    (
1.280)
##
## WhiteCol                      -0.192                                                   -
0.188
##                              (1.661)                                                    (
1.488)
##
## Poor                          -0.165
0.744
##                              (3.225)                                                    (
2.975)
##
## HC                            -0.672
##                              (0.491)
##
## NOX                           1.340
##                              (1.005)
##
## log(HC)                                                                                -3
3.657**
##                                                                                         (1
5.569)
##
```

```
## log(NOX)                                                                 42
.239***
##                                                                          (1
5.093)
##
## SO2                          0.086
0.064
##                             (0.147)                                       (
0.116)
##
## climate                                          0.193
##                                                 (0.525)
##
## pop                                            0.014***
##                                                 (0.005)
##
## wealth                                         2.154***
##                                                 (0.732)
##
## pollution                                       -0.020
##                                                 (0.046)
##
## Constant                  1,762.487***         669.318***            1,89
9.218***
##                            (437.113)            (85.908)              (4
18.735)
##
## ----------------------------------------------------------------------------
--------------
## Observations                   60                   60
60
## R2                           0.765                0.297
0.793
## Adjusted R2                  0.685                0.246
0.722
## Residual Std. Error    34.913 (df = 44)      54.008 (df = 55)       32.783
(df = 44)
## F Statistic         9.552*** (df = 15; 44) 5.815*** (df = 4; 55) 11.227***
(df = 15; 44)
## ============================================================================
==============
## Note:                                                   *p<0.1; **p<0
.05; ***p<0.01
```

```
vif(pol.lm)
```

```
##     Precip  Humidity   JanTemp  JulyTemp    Over65     House
##   4.113808  1.906540  6.144998  3.967545  7.470930  4.309909
##       Educ     Sound   Density  NonWhite  WhiteCol      Poor
##   4.866145  3.998022  1.660457  6.773090  2.842303  8.714508
```

```
##          HC         NOX         SO2
##   98.637392 104.981032   4.229207
```

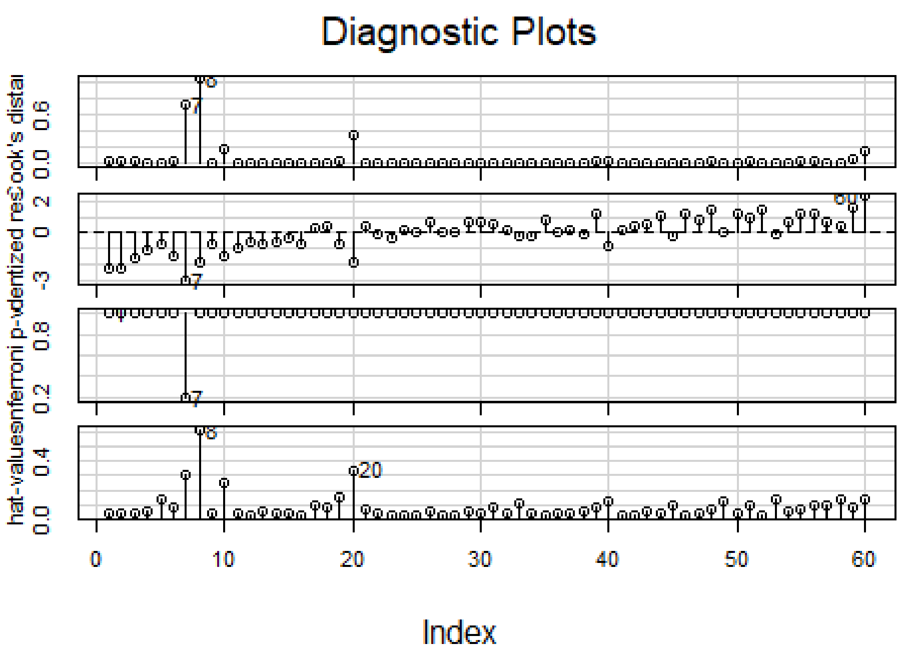
```r
vif(pol2.lm)
```

```
##    climate       pop    wealth pollution
##   1.844726  1.093076  1.770729  1.208244
```

```r
vif(pollog.lm)
```

```
##     Precip  Humidity   JanTemp  JulyTemp    Over65     House      Educ
##   3.903120  1.552752  5.093656  5.244655  7.318854  4.460011  4.357410
##      Sound   Density  NonWhite  WhiteCol      Poor   log(HC)  log(NOX)
##   3.899891  1.729636  7.161898  2.587694  8.408287 18.398052 17.543127
##        SO2
##   2.943944
```

```r
influenceIndexPlot(pol2.lm)
```



work with the final model

```r
final_model <- lm(Mortality ~ climate + pop + wealth + pollution, data = poll
ution)
stargazer(final_model, type = "text")
```

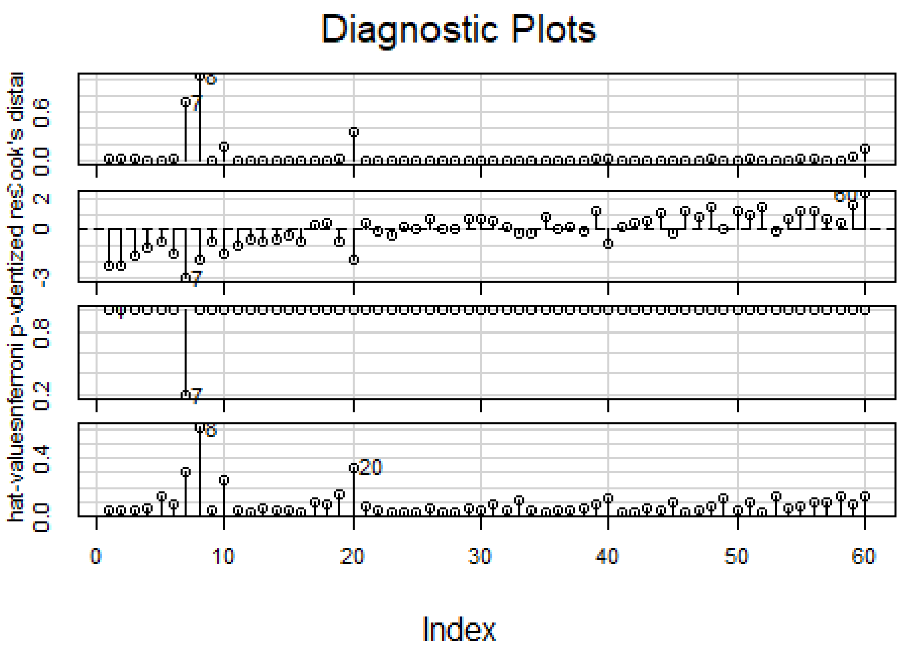
```
##
## ===============================================
```

```
##                              Dependent variable:
##                          ---------------------------
##                                   Mortality
## -----------------------------------------------------
## climate                             0.193
##                                    (0.525)
##
## pop                               0.014***
##                                    (0.005)
##
## wealth                            2.154***
##                                    (0.732)
##
## pollution                          -0.020
##                                    (0.046)
##
## Constant                         669.318***
##                                   (85.908)
##
## -----------------------------------------------------
## Observations                          60
## R2                                  0.297
## Adjusted R2                         0.246
## Residual Std. Error          54.008 (df = 55)
## F Statistic               5.815*** (df = 4; 55)
## =====================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

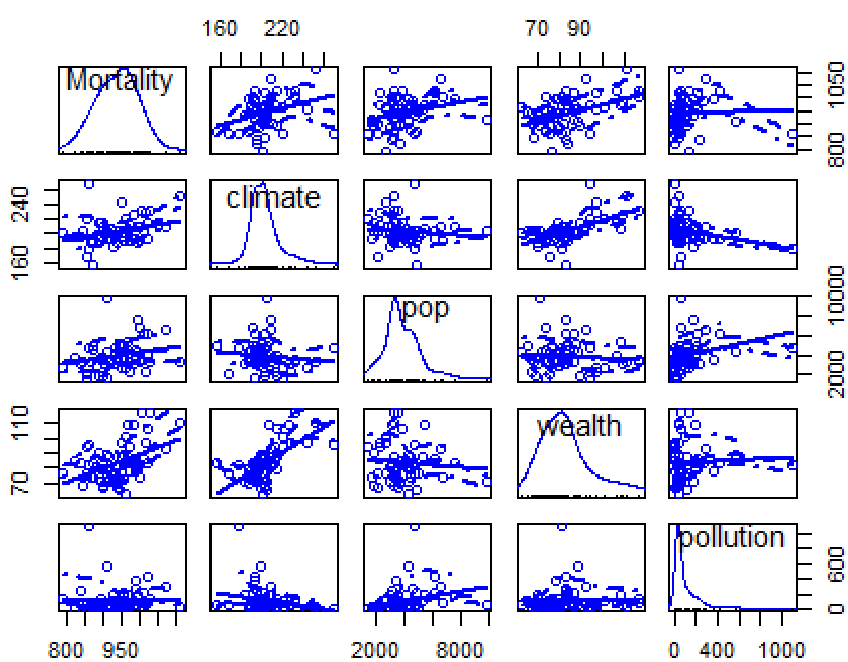
```
vif(final_model)
```

```
##   climate       pop    wealth pollution
##  1.844726  1.093076  1.770729  1.208244
```

```
influenceIndexPlot(final_model)
```

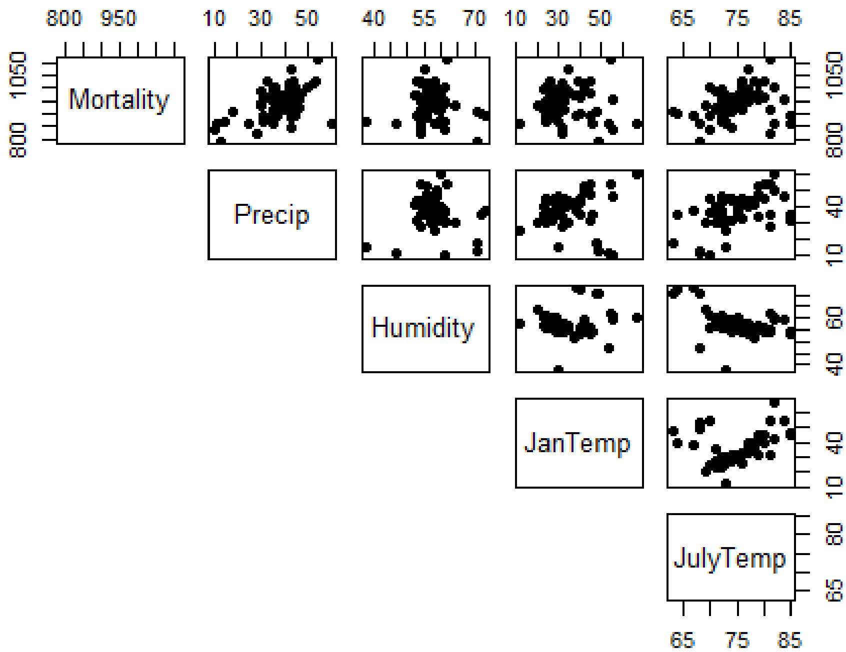



```r
scatterplotMatrix( ~ Mortality + climate + pop + wealth + pollution, data = p
ollution)
```

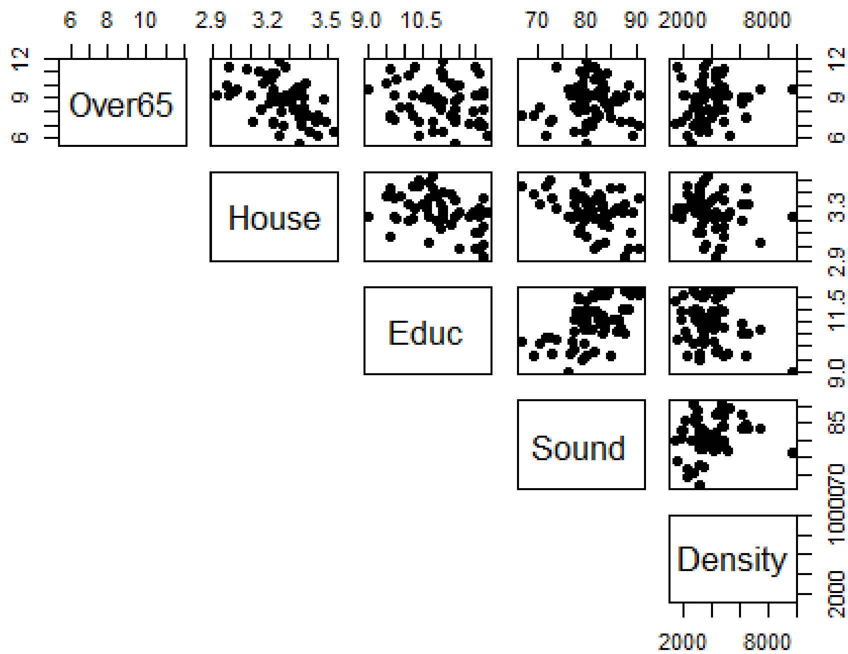


tests to look for multicollinearity

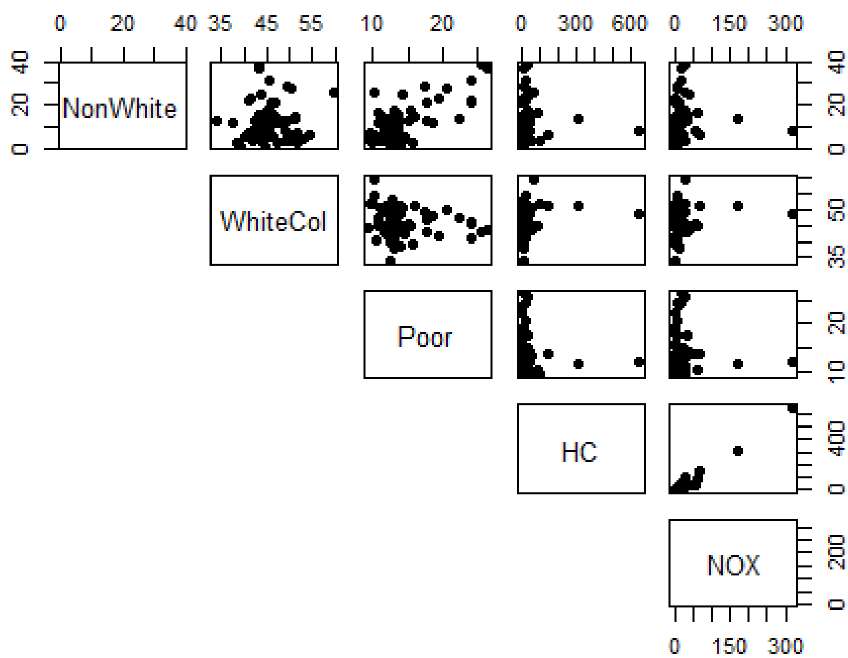
```
pairs(pollution[,2:6], pch = 19, lower.panel = NULL)
```



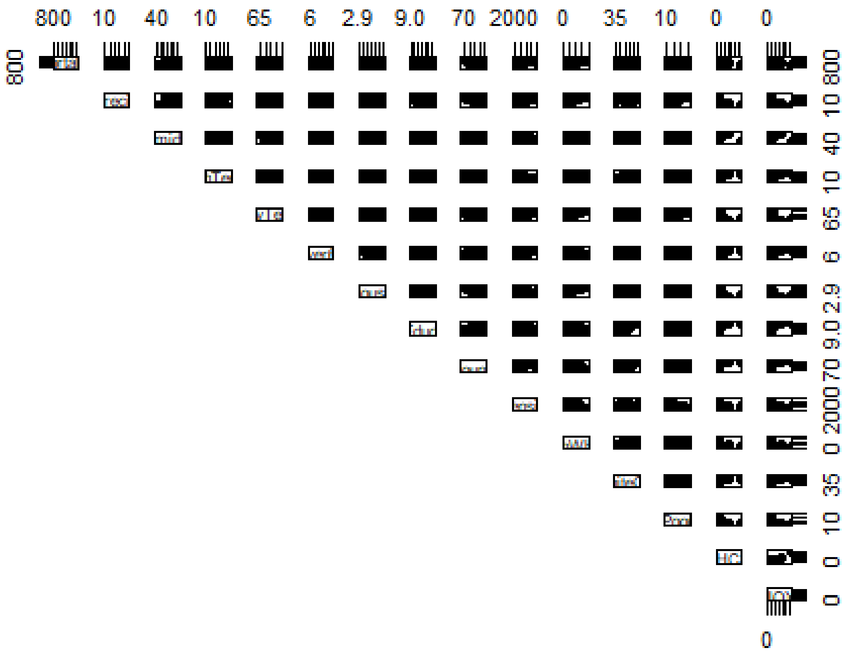

```
pairs(pollution[,7:11], pch = 19, lower.panel = NULL)
```



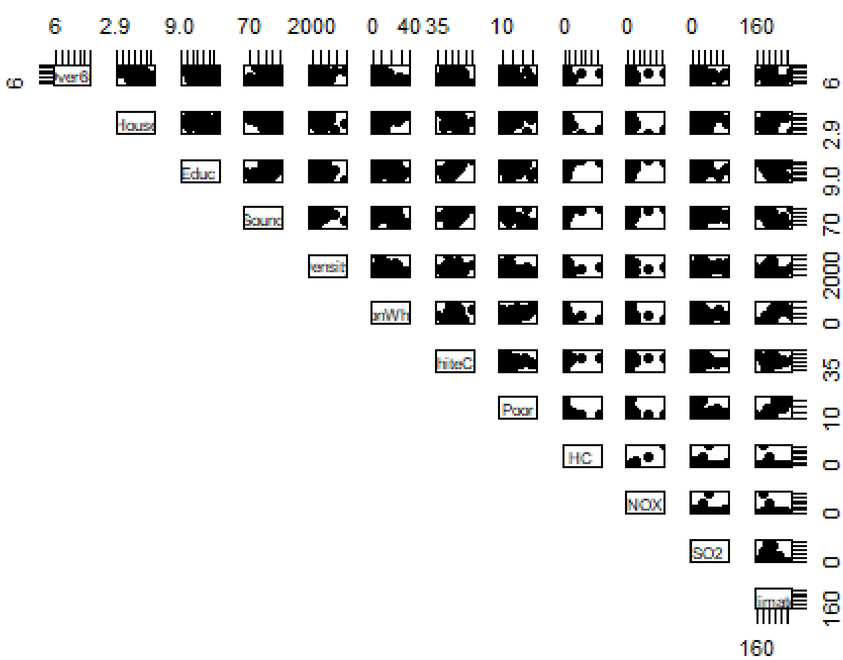

```
pairs(pollution[,12:16], pch = 19, lower.panel = NULL)
```



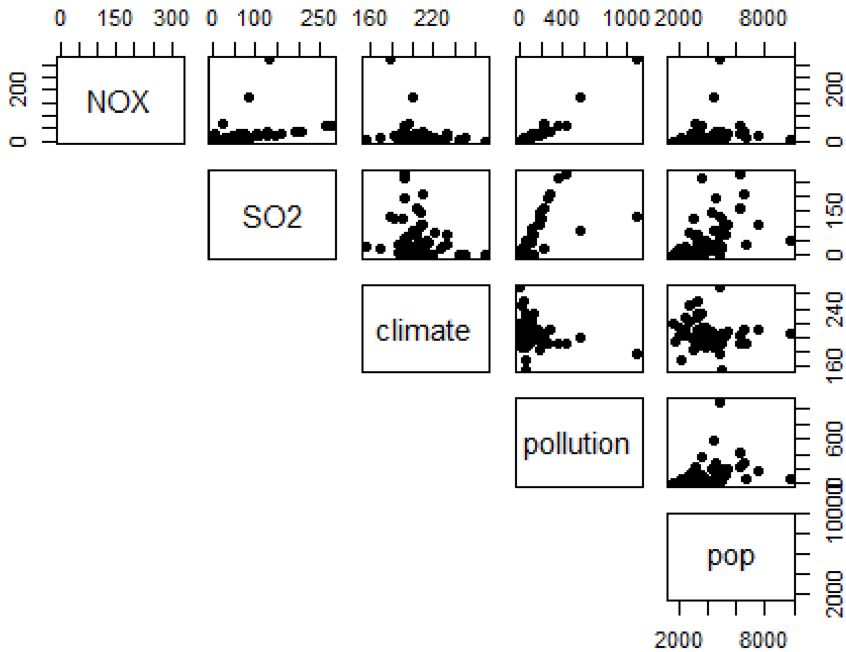

```
pairs(pollution[,2:16], pch = 19, lower.panel = NULL)
```



```r
pairs(pollution[,7:18], pch = 19, lower.panel = NULL)
```




```r
pairs(pollution[,16:20], pch = 19, lower.panel = NULL)
```



backwards elimination

```
full_mod <-lm(Mortality ~ climate + pop + wealth + pollution, data = pollutio
n)
belim <- stepAIC(full_mod,
                 scope = list(lower = ~ 1),
                 direction = "backward" , trace = 0)

tidy(belim)

##          term    estimate   std.error statistic      p.value
## 1 (Intercept) 696.72276862 51.210547213 13.605064 1.502623e-19
## 2         pop   0.01299881  0.004783368  2.717502 8.696644e-03
## 3      wealth   2.30569588  0.545066775  4.230116 8.564582e-05
```

Forward selection

```
null_mod <- lm(Mortality ~ 1, data = pollution)
fselect <-
  stepAIC(null_mod,
        scope = list(upper = ~ climate + pop + wealth + pollution, data = pol
lution),
        direction = "forward" , trace = 0)

tidy(fselect)
```

```
##          term     estimate    std.error statistic      p.value
## 1 (Intercept) 696.72276862 51.210547213 13.605064 1.502623e-19
## 2      wealth   2.30569588  0.545066775  4.230116 8.564582e-05
## 3         pop   0.01299881  0.004783368  2.717502 8.696644e-03
```

Stepwise selection

```
step_credit <-
  stepAIC(null_mod,
        scope = list(lower = ~ 1,
                     upper = ~ climate + pop + wealth + pollution, data = pol
lution),
        direction = "both"
        , trace=0)

tidy(step_credit)
```

```
##          term     estimate    std.error statistic      p.value
## 1 (Intercept) 696.72276862 51.210547213 13.605064 1.502623e-19
## 2      wealth   2.30569588  0.545066775  4.230116 8.564582e-05
## 3         pop   0.01299881  0.004783368  2.717502 8.696644e-03
```

Additional model to check if log may help

```
pollution$climatalogical <- pollution$Precip + pollution$Humidity + pollution
$JanTemp + pollution$JulyTemp
pollution$socioeconomic <- pollution$House + pollution$Density + pollution$Ov
er65 + pollution$Sound + pollution$Poor + pollution$WhiteCol + pollution$NonW
hite + pollution$Educ

final_model2 <- lm(Mortality ~ climate + pop + wealth + log(HC) + log(NOX) +
log(SO2), data = pollution)
stargazer(final_model2, type = "text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              Mortality
## -----------------------------------------------
## climate                        0.768
##                               (0.466)
##
## pop                            0.004
##                               (0.005)
##
## wealth                       1.834***
##                               (0.628)
##
## log(HC)                     -44.395***
```

```
##                                (16.318)
##
## log(NOX)                         34.353*
##                                (18.393)
##
## log(SO2)                        22.132***
##                                 (6.758)
##
## Constant                       587.391***
##                                (82.036)
##
## -------------------------------------------------
## Observations                       60
## R2                                0.541
## Adjusted R2                       0.489
## Residual Std. Error         44.465 (df = 53)
## F Statistic              10.409*** (df = 6; 53)
## =================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

```r
vif(final_model2)
```

```
##   climate       pop    wealth   log(HC)  log(NOX)  log(SO2)
##  2.138998  1.330982  1.922185 10.986295 14.161401  3.056769
```

```r
influenceIndexPlot(final_model2)
```