Data Analysis On Mortality Rate Of US Cities

Jake Denton

11/03/2020

Introduction

In this analysis we look at associations between the mortality rate of US cities (deaths per 100000) and other measurements. We find that pollution/precipitation levels explain the observed variability in mortality rate for this set of data. Income may have a link with mortality when considered on it's own, but is correlated with the other significant variables and so aren't associated to differences in mortality rate.

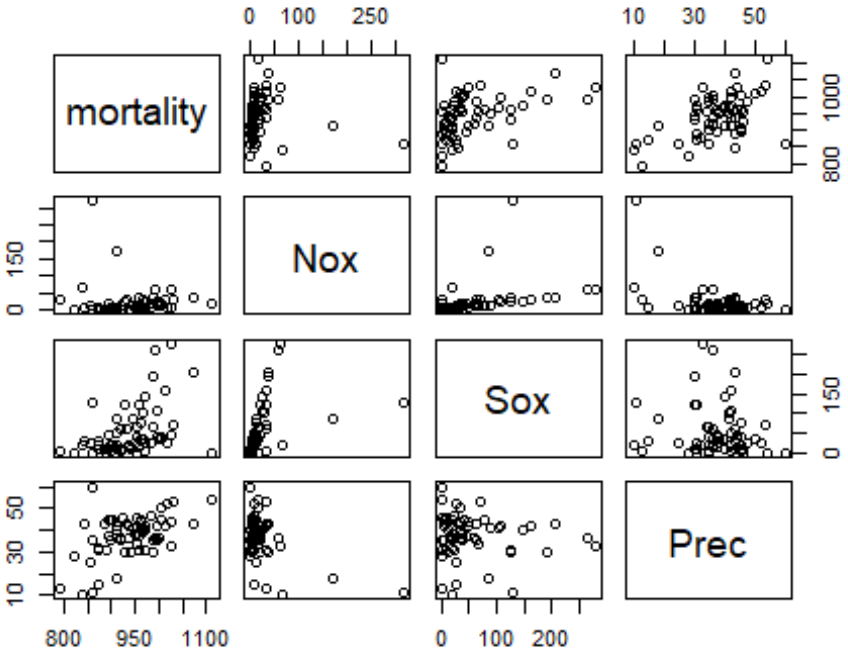
Exploratory Analysis

The data set given relates to 5 measurements of 60 US cities. For each city, the mortality rate (deaths per 100000 people) is measured along with: Nox (nitrogen oxide levels), Sox (sulfur dioxide levels), Prec (mean annual precipitation in inches) and Income ('High' or 'Low'). The first 4 variables mentioned above are continuous numeric measurements whilst Income is a factor with 2 levels, 'High' or 'Low'.

```
library(ggplot2) # Loads the ggplot2 package.
Pol<-read.table("PolData.txt",header=TRUE) # Reads the table given and
assigns a dataframe to it.
Pol$Income<-factor(Pol$Income)# Tells r that the variable Pol$Income is a
factor.
```

First of all, we consider pairwise scatter plots for the numerical measurements.

```
plot(Pol[,c(1,2,3,4)]) # This creates the matrix of scatter plots.
```

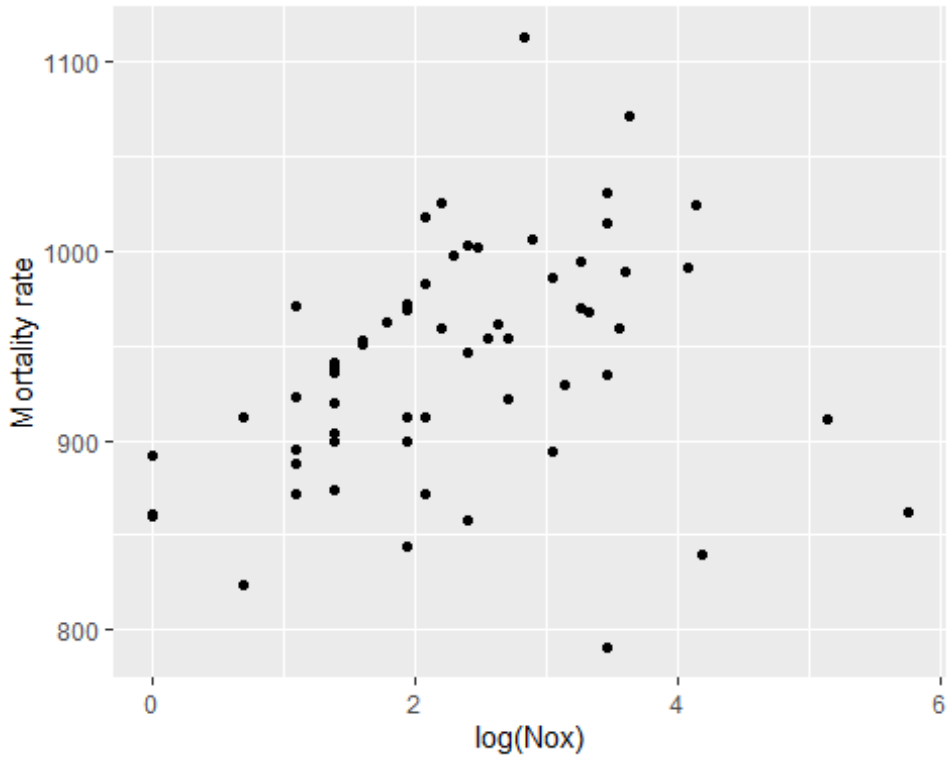



```
summary(Pol$Nox) #This provides us with the summary statistics for the
variable Nox.

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    4.00    9.00   22.65   23.75  319.00
```

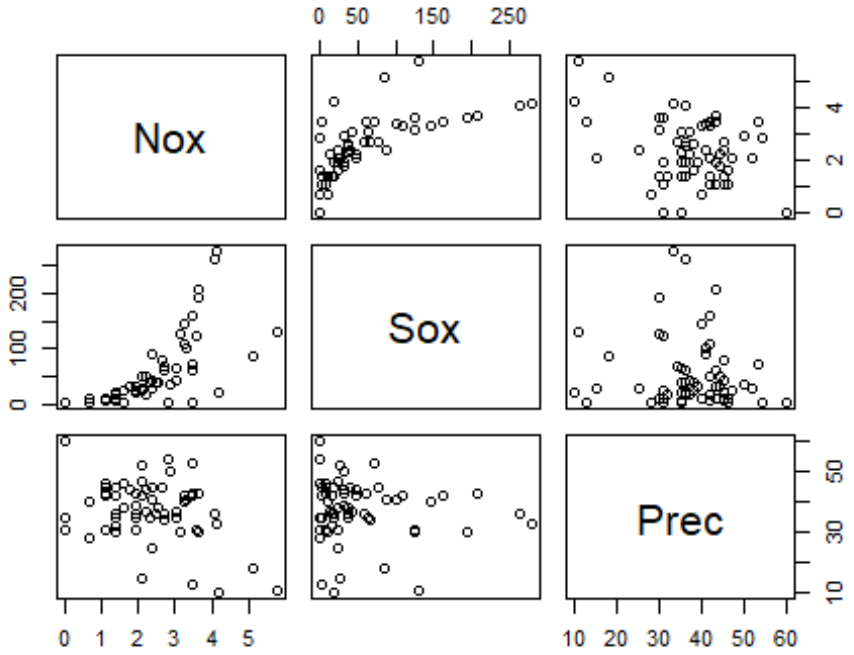
There are clear associations between the variables Prec/Sox and mortality. Higher sulfur dioxide levels and mean annual precipitation have higher mortality rates. As for the nitrogen oxide levels, the summary statistics tell us that the Nox data has a large range/standard deviation (considering distances from the mean), with minimum at 1 and maximum at 319. It also tells us that the majority of the data is in the region [4,23.75], so taking logarithms would bring the Nox data together and allow an association to be observed.

```
Pol$Nox<-log(Pol$Nox) # Applies log transformation to the Nox data.
qplot(Pol$Nox,Pol$mortality,data=Pol,xlab="log(Nox)",ylab="Mortality rate") #
Produces scatterplot.
```



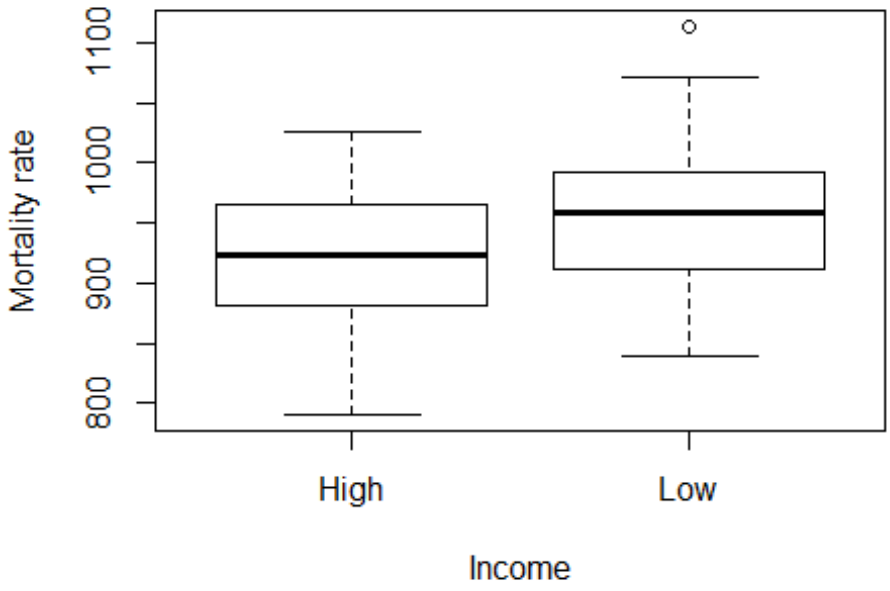
The new plot shows that mortality rates are higher when the log value of nitrogen oxide levels are higher, and so equivalently it shows when nitrogen oxide levels are higher mortality rate is higher. However, it should be noted from the pairwise scatter plot that there is an obvious correlation between sulfur dioxide levels and nitrogen oxide levels. Furthermore, there appears to be a slight negative correlation between Nox and the other explanatory variable Prec. For example, a city with high nitrogen oxide levels have high sulfur dioxide levels and lower levels of precipitation, as shown in the plots below. So, it could be possible that only one/two of these variables can explain the differences in mortality rate.

```
plot(Pol[,c(2,3,4)]) # Produces 3x3 matrix of scatterplots featuring the
variables log(Nox), Sox and Prec.
```



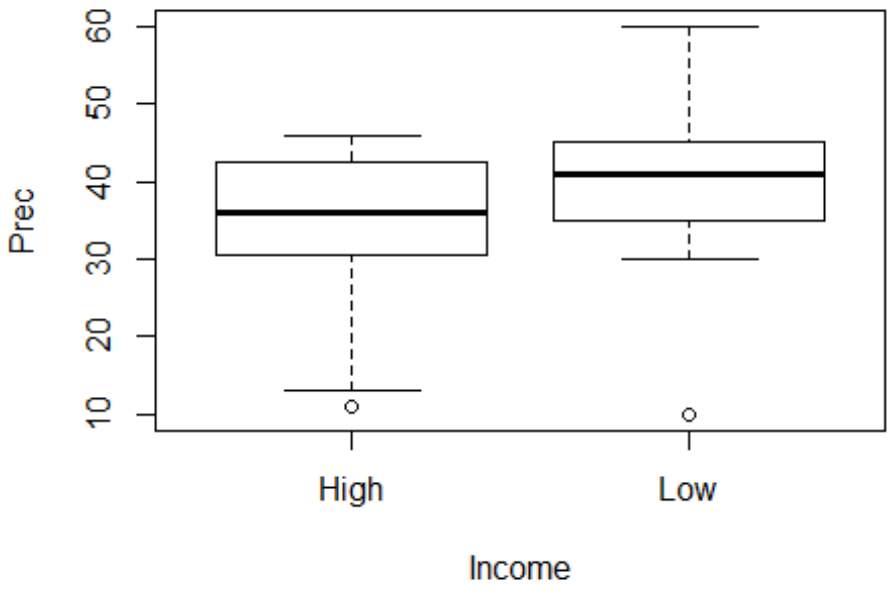
Next, we consider the Income variable.

```
boxplot(Pol$mortality~Pol$Income,xlab="Income",ylab="Mortality rate") #
Creates boxplot of mortality against income.
```



Thanks to this plot, we can conclude that cities with low income seem to have higher mortality rate. Still, we should investigate if a correlation exists between income and the other continuous variables. For instance, low income cities tend to have slightly larger amounts of precipitation, as shown below.

```
boxplot(Pol$Prec~Pol$Income,xlab="Income",ylab="Prec") # Creates boxplot of
precipitation against income.
```



The other plots have been omitted as the association between income and pollution levels is less clear and unlikely to be statistically significant. The relationship between income and precipitation can be considered later when we look at if the associations between the variables are significant.

Modelling

We'll begin by including nitrogen oxide levels and precipitation in the model.

```
m1<-lm(mortality~Nox+Prec,data=Pol) # Fits linear regression model to the
variables above.
summary(m1) # Produces a summary of the model, including p-values that
indicate significance.

##
## Call:
## lm(formula = mortality ~ Nox + Prec, data = Pol)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.862  -23.874    1.111   28.551   83.968
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 706.4473    29.9693  23.572  < 2e-16 ***
## Nox          29.1444     5.1529   5.656 5.23e-07 ***
## Prec          4.4476     0.6113   7.276 1.10e-09 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.58 on 57 degrees of freedom
## Multiple R-squared:  0.5257, Adjusted R-squared:  0.5091
## F-statistic: 31.59 on 2 and 57 DF,  p-value: 5.838e-10
```

This summary allows us to observe that both terms are worth keeping in the model (since the p-values are very small). In the exploratory section, we saw an association between sulfur dioxide levels and mortality. Due to this, we'll fit a model with just the Sox variable.

```
m2<-lm(mortality~Sox,data=Pol) # Fits linear regression model with Sox as the
independent variable
summary(m2)

##
## Call:
## lm(formula = mortality ~ Sox, data = Pol)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -128.408  -35.079   -8.669   34.338  194.851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 917.8874     9.6435  95.182  < 2e-16 ***
## Sox           0.4179     0.1166   3.585 0.000692 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.77 on 58 degrees of freedom
## Multiple R-squared:  0.1814, Adjusted R-squared:  0.1673
## F-statistic: 12.85 on 1 and 58 DF,  p-value: 0.0006922
```

Again the p-value here is small, so we should consider including this variable in our model and looking at whether or not it is significant after controlling for nitrogen oxide levels (cities with higher Nox had higher Sox from our exploratory plots).

```
m3<-lm(mortality~Nox+Prec+Sox,data=Pol) # Fits model including all 3
continuous explanatory variables
summary(m3)

##
## Call:
## lm(formula = mortality ~ Nox + Prec + Sox, data = Pol)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -115.712  -21.978    2.667   28.579  105.388
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 725.2998    30.9644  23.424  < 2e-16 ***
## Nox          19.6873     7.1013   2.772  0.00754 **
## Prec          4.1936     0.6128   6.843 6.28e-09 ***
## Sox           0.2346     0.1241   1.891  0.06386 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.63 on 56 degrees of freedom
## Multiple R-squared:  0.5542, Adjusted R-squared:  0.5303
## F-statistic: 23.21 on 3 and 56 DF,  p-value: 6.846e-10
```

```
summary(m1)$adj.r.squared
```

```
## [1] 0.5091062
```

```
summary(m3)$adj.r.squared
```

```
## [1] 0.5303177
```

The p-value for the sulfur dioxide levels is 0.06386, so considering a hypothesis test at the 5% level, we would conclude that there is insufficient evidence to reject the null (that Sox has no effect after controlling for log(Nox)/Prec), but the evidence is sufficient to reject the null at the 10% level. Furthermore, we should note that the value of adjusted R squared increases slightly upon addition of Sox to the model. So there are arguments for/against inclusion of this variable in the overall model, and because of this we will report the predictions of both models to the agency to see if the outcome changes.

We test the income variable similarly.

```
m4<-lm(mortality~Nox+Prec+Income,data=Pol)
anova(m1,m4) # Compares the two models and gives a p-value so we can decide
if income is worth including
```

```
## Analysis of Variance Table
##
## Model 1: mortality ~ Nox + Prec
## Model 2: mortality ~ Nox + Prec + Income
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     57 108276
## 2     56 104263  1    4012.7 2.1552 0.1477
```

The p-value is 0.1477, so we don't reject the null hypothesis. There is no evidence that income is associated with mortality after accounting for the effects of nitrogen oxide levels and mean annual precipitation.

So we have found that log(Nitrogen oxide levels) and precipitation are most heavily associated with mortality rate. However, adjusted R squared for this model is low, suggesting other variables not presented in the data could have an effect on mortality rate in US cities. Sulfur dioxide levels are also mildly significant after accounting for the other continuous explanatory variables.

Interpreting Parameter Estimates

```
summary(m1)$coef # Next two lines remind us of the parameter estimates for each model

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 706.447256 29.9693390 23.572334 4.459961e-31
## Nox          29.144351  5.1529161  5.655895 5.227778e-07
## Prec          4.447562  0.6112549  7.276118 1.102702e-09

summary(m3)$coef

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 725.2997797 30.9643638 23.423694 1.297210e-30
## Nox          19.6872905  7.1012621  2.772365 7.542477e-03
## Prec          4.1936160  0.6128057  6.843304 6.280121e-09
## Sox           0.2345559  0.1240673  1.890554 6.386231e-02
```

Therefore the models are (represent mortality rate by y, Nox by x[1], Prec by x[2], Sox by x[3]): (1) y[i]=706.4473+29.144log(x[i1])+4.4476x[i2] (2) y[i]=725.2998+19.6873log(x[i1])+4.1936x[i2]+0.2346x[i3]

Both models suggest that an increase in pollution levels and mean annual precipitation lead to more deaths per 100000 people. This makes intuitive sense since air pollution is linked to worsening respiratory and cardiac conditions, both of which can result in premature death. A tentative link could also be made between precipitation and mortality, as weather can factor into people's willingness to exercise e.g. catch a bus instead of walking in the rain. So if there is more rain people may exercise less and this could link to higher mortality rate (i.e. its likely that other variables could be responsible for the association between mortality and precipitation).

As suggested above, the continuous variables in these models may not fully explain the data. A few examples include: car use (which increases pollution), the number of times people exercise each week on average, the proportion of the cities population that regularly smoke/drink. All these are risk factors in mortality that haven't been considered in the model/data.

Prediction Regarding City 48

```
newdata<-data.frame(Nox=1.6021, Sox=40, Prec=18, Income="High") # Creates a data frame for city 48, with both pollution levels set to 40 (Nox is transformed by log in our model so is done so here)
predict(m1,newdata,interval="prediction",level=0.95) # Uses m1 to make a prediction for the above data

##        fit     lwr      upr
## 1 833.1955 741.047 925.3441

predict(m3,newdata,interval="prediction",level=0.95) # Uses m3 to make a prediction for the above data
```

```
##        fit      lwr      upr
## 1 841.7081 751.0875 932.3287
```

The predictions with the data for the city, setting the nitrogen oxide levels equal to 40 (and so log(Nox)=1.6021) and sulfur dioxide levels to 40, are 833.1955 and 841.7081 for respective models (1) and (2). This describes a decrease in mortality rate of 78.5 for model 1 (which does not include Sox) and 70 for model 2 (includes Sox), suggesting that lowering both emissions to 40 decreases mortality rate by a worthwhile amount.

However, the calculated prediction intervals for both models have an upper limit greater than the current value for mortality in the city. Hence, reducing both pollution levels to 40 may have no effect on the mortality rate at all. Upon these findings, I would advise the Agency to gather more data, which considers more variables that could be associated with mortality e.g those mentioned previously in this report. Inclusion of more explanatory variables should benefit the model, allowing more accurate conclusions to be drawn on the association between air pollution and mortality. Then, more insightful decisions can be made regarding whether it's worth investing resources to reduce pollution levels in city 48.

Summary

We found strong evidence that cities with higher nitrogen oxide levels and mean annual precipitation have higher mortality rates. There is also some evidence that cities with higher sulfur dioxide levels have higher mortality, after accounting for the effect of nitrogen oxide levels and precipitation.

Income was not found to be associated with mortality rate after controlling for NO levels and precipitation. Whilst we found that 'Low' income cities seemed to have higher mortality rate, we also saw that they tended to have more precipitation annually which could explain the association.

Lastly, we concluded that although NO levels/precipitation provided the most plausible explanation for differences in mortality rate, the variables investigated were not sufficient to allow us to predict accurately if there would be any significant effects of reducing pollution levels in city 48.