Math 245

May 8, 2014

Lauren Yoo

**Case Study II: Environmental and Socioeconomic Variables for Mortality**

The goal of the analysis is to find out, after controlling other factors, how mortality is associated with SO2. Trying to explain which climate or socioeconomic factors influence mortality is a critical study to find out effective approaches for improving daily living environment or preparing for elderly support facilities and policies, because it allows us to predict mortality based on current conditions. Our data was collected from the 5 Standard Metropolitan Statistical Areas in the United States, and compose of 12 variables, which include mortality and various climate and socioeconomic factors.

Since we are ultimately interested in seeing whether mortality is associated with SO2, we first took a look at the two variables, mortality and SO2, and the association between the two. The SO2 is log-transformed since it was highly right-skewed. We looked at the scatter plot of logSO2 against mortality to see general association and check visible outliers. The visible outliers turn out to be observations from San Jose, CA, and New Orleans, LA, so we removed the observations.

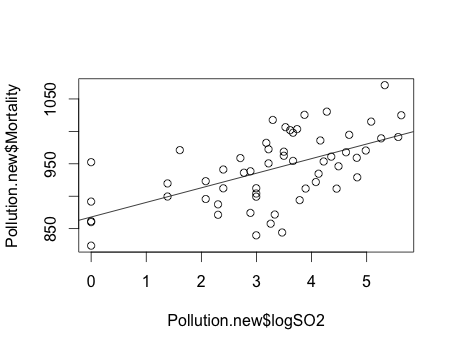
Min. 1st Qu. Median Mean 3rd Qu. Max.

Mortality 823.8 899.5 946.2 941.3 982.3 1071.0

SO2 1.00 15.00 32.00 54.25 68.00 278.00

Table1. Statistics of mortality and SO2 without outliers.

We find that the SO2 data is highly right-skewed, so I decided to log-transform the explanatory variable SO2.



Graph1. Scatterplot of mortality and logSO2 after taking out outliers.

From the graph, we can infer that there is a positive relationship between SO2 level and mortality. Our ultimate goal is to figure out the association between the SO2 and mortality after controlling other explanatory variables.

First, we need to come up with an optimal linear model to explain mortality without SO2 level. Since we have too many possibilities (2^12) for interaction terms, I first filtered out individual variables with at least marginal significance from the summary output of a full model that includes all the variables in the data. The summary output indicates significance for January temperature, education, population density, and non-white proportion.

Then, I proceeded to see if there is any interaction between the terms. The scatterplot matrix and trellis plots suggest quadratic and interaction terms that may be significant. In the summary test of linear model that includes all the quadratic terms, Educ^2 turns out significant. In the Trellis plots to describe interactions between two different variables, two terms turn out significant: Education:Density, Educaiton:JanTemp.

An anova test, between the earlier model without the interaction terms and the new model with the interaction terms, was conducted to figure out which model better explains mortality. The resulting p-value is 0.038, so we have sufficient evidence to prefer the fuller model that contains the interaction terms.

Diagnostics with leverage, cook’s distance, and Jackknife studentized residuals were performed to figure out if there are any influential outliers. The diagnostics indicates influential outlier observations from York, PA and Utica, NY. I removed the two outliers from the model. We now have four outliers removed, 2 outliers removed earlier from the Mortality-SO2 plot, and two from the diagnostics.

So, the resulting linear model is

E[Mortality | JanTemp, Educ, Density, NonWhite, Educ^2, Educ:Density, Educ:JanTemp]

= -4.386e+02 + 5.262e-01 (JanTemp) +2.788e+02(Educ) - 4.124e-02(Density)+4.021e+00(NonWhite) -1.432e+01I(Educ^2)+5.003e-03(Educ:Density) -1.733e-01(Educ:Jan Temp)

The summary output of the model suggests that only NonWhite is significant. Several anova tests to compare different pairs of smaller models were performed to come up with the best fitting model. (For the details of the pairs that were compared, and the results of each comparison, refer to the R code.)

As a result, I came up with the final linear model with all the containing variables significant.

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1130.88767 68.57828 16.490 < 2e-16 \*\*\*

JanTemp -1.74335 0.52367 -3.329 0.001624 \*\*

Educ -21.17918 5.96174 -3.553 0.000832 \*\*\*

Density 0.01314 0.00365 3.599 0.000721 \*\*\*

NonWhite 4.30343 0.60279 7.139 3.28e-09 \*\*\*

Table2. Summary output of the final model with all the factors other than SO2 controlled.

Again, I went through the diagnostics with leverage, cook’s distance, and jackknife student residuals, and removed two outlier observations from Miami, FL, and San Diego, CA. The resulting linear model has the following output.

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.094e+03 7.231e+01 15.133 < 2e-16 \*\*\*

JanTemp -1.742e+00 7.035e-01 -2.477 0.01662 \*

Educ -1.644e+01 6.088e+00 -2.700 0.00939 \*\*

Density 8.151e-03 3.323e-03 2.453 0.01763 \*

NonWhite 4.538e+00 6.727e-01 6.746 1.37e-08 \*\*\*

Table3. Summary output of the final model without outliers from Miami, FL, and San Diego, CA

There is no significant change being made due to outlier removal, since all variables are significant in both models(Table 2, Table3). However, it is worth noting that the estimate of the Education changes quite drastically, so we can infer that the observations from Miami and San Diego were influential outliers for mortality against education.

Now that all the other variables are controlled, log-transformed SO2 was put into the model. When having SO2 level in the model, the variable January temperature turns out as insignificant. The Avona two-model comparison test also suggests removing the January temperature, when SO2 is already in the model.

Thus, we have the final model

E[Mortality |Educ, Density, NonWhite, logSO2] = 1048 -19.25(Educ) +0.009092 (Density)+3.223(NonWhite)+9.876(log(SO2))

Variance Inflation factor test was performed to check for multi-collinearity, and the test result indicates that multi-collinearity is not a problem in our model. Thus, we keep our model.

The final model indicates that the doubling of SO2 level is associated with the mean mortality increasing by 6.846.