Effect of Air Pollutants on Temperature

Jason Lei

6/10/2020

Abstract

I wanted to analyze the effect air pollutants in Beijing had on the temperature. I used data from the Shunyi observation station, which included meteorological variables, to determine which had more significant effects on the Temperature. I found that there were not any significant effects the air pollutants had on the rising temperature, as some air pollutants, when other variables were held constance, even had a negative linear relationship to temperature. On the other hand, the meteorological variables had just as much relevance as the air pollutants, while the unit of times did not matter, other than the hours, which might simply correlate to the time of day.

Problem and Motivation

Is there really a correlation between air pollutants and temperature? Obviously, climate change is a real problem that will soon doom the entire world, but are the pollutants in the air really affecting the temperature in Beijing? Weather stations all around Beijing recorded the concentration of air pollutants such as O3, NO2, PM10, PM2.5, CO, and SO2, as well as some meteorological data including temperature, wind speed, pressure, dew point temperature, etc. These nationally-controlled air pollution monitoring sites recorded these data from over a period of almost four years. While there are many air-quality monitoring sites in Beijing, for my research, we will be using the data set from the Shunyi monitoring site. Almost everyone knows that China, the manufacturing powerhouse that it is, has a problem with pollution, air pollution specifically. When looking at relevant topics on climate change, various news articles can be found, reporting on how so-and-so countries are trying to pressure China (and other countries that produce heavy amounts of pollutants) into various climate-friendly policies. Essentially, I wanted to know how intense of an effect air pollutants have on rising temperatures, and if temperatures in Beijing are really rising, or if Beijing has always been a city of insanely hot summers.

Background

The data comes from the Shunyi weather station in Beijing. The temperature variable will be the response. I will include the all of the air pollutants (NO2, O3, CO, SO2, PM10, and PM2.5), as well as the times (year, month, day, hour) and the other meteorological variables (PRES, DEWP, WSPM). I will leave out RAIN, because the variable seems to be always 0, row number and station, for obvious reasons, and WD, because direction of wind should not have an effect on temperature.

Questions of Interest

I want to know the association of between time and the concentration of the air pollutants and whether or not there is a rise in temperature in Beijing. A long with that, I want to know if there is even a positive correlation between air pollutants and temperature.

Regression Methods

Because I want to know which variables in the data set would be significant, I first did fit the linear models, made an anova table and made scatterplots and av plots of the variables. I also used the step function

do decide the best model to use for multiple linear regression. As reflected in the AV plots, the variables SO2, year, month, and day have little to no effect on the model. Also, based off of the qq plot and the residual vs. fits plot, there is some violation of the LINE assumptions, when considering contant variance and normality. There likely need to be transformations

Important Details

First, I computed the coefficient estimates and their p-values and their test statistics to determine if the variables were sufficient to the model. Based off of the summary table and the step function, it was determined that the "day" variable had no significant effect on the model. I then used another ANOVA table to conduct a partial F-test. H0: b9 = 0 vs H1: b0 != 0 The F-value was 1.3451, and the p-value = 0.2461 > 0.05, which means we fail to reject H0, so b9, or we have sufficient evidence that the days are not significant to the model. However I also noticed the AV plots showed that S02, month, and year also suggested that these variables were not significant to the model. So, I used the subset() function to find the optimal model, which indeed correlated to the AV plots and excluded SO2, NO2, year, month, and day from the model. #Diagnostic Checks After the initial exploratory analysis, I also made scatterplots of optimal model and a residual vs fits plot, as well as a QQ-plot. Based on the scatterplots and residual vs fits plot, linearity and constant variance might be violated, but normailty is definitely violated. I used the yeo-johnson function to determine which variables should be transformed (or in otherwise, the variables which did not have a lambda close to 1). O3, CO, PM10, PM2.5, had lambda values close to 0, which meant that they should have log transformations. PRES had a lambda of -4, so I used the inverse transformation. Overall, based on the new scatterplots and the new residual vs fits plot, linearity and constance variance were much better. There was not a clear pattern in the residual vs fits model, and there were less outlying data points. The qq-plot, while still heavy tailed, is much closer to the line, which is an improved normality. #Interpretation Overall, for this data set in particular, there was not a significant association between air pollutants and temperature. SO2 and NO2, the regression model determined, were not significant to the response Temperature, and some air pollutants, such as CO2, even had a negative linear relationship with temperature, meaning that the more CO2 in the air, the colder. Additionally, the more air pollutants put into the air over time had little to no effect as only hour was the only unit of time that had some effect on the model, but it probably simply correlates to the time of day. Thus, from the Shunyi station, there is no significant correlation between air pollutants and rising temperature.

Conclusion

In terms of the data given by the Shunyi station in Beijing, the regression analysis gave some interesting perspectives. First, the air pollutants that were put in the air by, presumably, industries that gave off greenhouse gases such as the manufacturing industry, had little effect on the model. In fact, NO2 and SO2 were determined by the model to be irrelevant to the data set. Over the course of four years, the gradually increasing amount of pollutants in the air were irrelevant to the data set. The year, month, and day variables were not significant to the MLR model, and only hour was significant. However, this is probably because of the position of the sun during the day, and not the increase in air pollutants. The meteorological variables in the data fared far better, as all of them were considered relevant to the data. Once the data was fixed to fit the LINE assumptions of the regression model, only air pressure had to be adjusted. Wind speed and dew point temperature did not need to be transformed to fit the assumptions. Overall, meteorological variables seemed to have much greater effects on the temperature, while the air pollutants did not have great effects on the response, temperature.

Appendix

```
library(car)
```

```
## Loading required package: carData
```

```r
library(leaps)
library(bestNormalize)
beijing <- read.csv("PRSA_Data_Shunyi_20130301-20170228.csv", header = TRUE)
airquality <- na.omit(beijing)
attach(airquality)
airquality.lm <- lm(TEMP ~ O3 + CO + NO2 + SO2 + PM10 + PM2.5 + year + month + day + hour + PRES + DEWP
summary(airquality.lm)
```

```
##
## Call:
## lm(formula = TEMP ~ O3 + CO + NO2 + SO2 + PM10 + PM2.5 + year +
##     month + day + hour + PRES + DEWP + WSPM)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.8771  -2.6326  -0.1573   2.4096  18.7662
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.518e+02  4.201e+01  10.753  < 2e-16 ***
## O3           4.739e-02  6.103e-04  77.644  < 2e-16 ***
## CO          -1.240e-03  3.586e-05 -34.571  < 2e-16 ***
## NO2          9.390e-03  1.335e-03   7.035 2.04e-12 ***
## SO2          9.061e-03  1.522e-03   5.953 2.67e-09 ***
## PM10         1.066e-02  6.218e-04  17.140  < 2e-16 ***
## PM2.5       -1.924e-02  7.664e-04 -25.106  < 2e-16 ***
## year        -7.052e-02  2.100e-02  -3.358 0.000786 ***
## month        2.781e-02  7.395e-03   3.760 0.000170 ***
## day         -3.026e-03  2.609e-03  -1.160 0.246141
## hour         8.815e-02  3.658e-03  24.100  < 2e-16 ***
## PRES        -2.976e-01  3.911e-03 -76.088  < 2e-16 ***
## DEWP         4.937e-01  3.109e-03 158.786  < 2e-16 ***
## WSPM         8.666e-01  2.170e-02  39.944  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.952 on 30180 degrees of freedom
## Multiple R-squared:  0.8809, Adjusted R-squared:  0.8809
## F-statistic: 1.718e+04 on 13 and 30180 DF,  p-value: < 2.2e-16
```

```r
anova(airquality.lm)
```

```
## Analysis of Variance Table
##
## Response: TEMP
##               Df  Sum Sq Mean Sq    F value    Pr(>F)
## O3             1 1403710 1403710 89863.4428 < 2.2e-16 ***
## CO             1   58333   58333  3734.3717 < 2.2e-16 ***
## NO2            1   17705   17705  1133.4441 < 2.2e-16 ***
## SO2            1  162341  162341 10392.8538 < 2.2e-16 ***
## PM10           1   62690   62690  4013.3277 < 2.2e-16 ***
## PM2.5          1    1801    1801   115.2803 < 2.2e-16 ***
## year           1   88847   88847  5687.8381 < 2.2e-16 ***
## month          1   58824   58824  3765.8422 < 2.2e-16 ***
```