Example Analysis

Roger D. Peng, Associate Professor of Biostatistics

May 18, 2016

Synopsis

In this report we aim to describe the changes in fine particle (PM2.5) outdoor air pollution in the United States between the years 1999 and 2012. Our overall hypothesis is that out door PM2.5 has decreased on average across the U.S. due to nationwide regulatory requirements arising from the Clean Air Act. To investigate this hypothesis, we obtained PM2.5 data from the U.S. Environmental Protection Agency which is collected from monitors sited across the U.S. We specifically obtained data for the years 1999 and 2012 (the most recent complete year available). From these data, we found that, on average across the U.S., levels of PM2.5 have decreased between 1999 and 2012. At one individual monitor, we found that levels have decreased and that the variability of PM2.5 has decreased. Most individual states also experienced decreases in PM2.5, although some states saw increases.

Loading and Processing the Raw Data

From the EPA Air Quality System we obtained data on fine particulate matter air pollution (PM2.5) that is monitored across the U.S. as part of the nationwide PM monitoring network. We obtained the files for the years 1999 and 2012.

Reading in the 1999 data

We first read in the 1999 data from the raw text file included in the zip archive. The data is a delimited file were fields are delimited with the | character adn missing values are coded as blank fields. We skip some commented lines in the beginning of the file and initially we do not read the header data.

```
pm0 <- read.table("pm25_data/RD_501_88101_1999-0.txt", comm
                  header = FALSE, sep = "|", na.strings = '
```

After reading in the 1999 we check the first few rows (there are 117,421) rows in this dataset.

Results

Entire U.S. analysis

In order to show aggregate changes in PM across the entire monitoring network, we can make boxplots of all monitor values in 1999 and 2012. Here, we take the log of the PM values to adjust for the skew in the data.

```
boxplot(log2(x0), log2(x1))
```

```
## Warning in boxplot.default(log2(x0), log2(x1)): NaNs pro

## Warning in bplt(at[i], wid = width[i], stats = z$stats[,
## $group == : Outlier (-Inf) in boxplot 1 is not drawn

## Warning in bplt(at[i], wid = width[i], stats = z$stats[,
## $group == : Outlier (-Inf) in boxplot 2 is not drawn
```