Data Exam Two – Data Description

36-402 Advanced Methods for Data Analysis

Due Friday, May 7, 2021 at 3pm EDT

The dataset `chicago.csv` contains measurements of air quality in Chicago, Illinois each day from January 1987 to the end of December 2000. It also includes the total number of non-accidental deaths recorded in Chicago each day and the mean temperature each day.

Relevant Variables

The data file `chicago.csv` contains 5,114 observations of 6 variables. The variables are:

time The date of observation, given as the number of days before or after December 31, 1993

death The number of non-accidental deaths on that date

pm10median The median density of PM_{10} pollution, i.e. particulate matter with diameter less than 10 micrometers (milligrams per cubic meter)

o3median The median concentration of ozone (parts per billion)

so2median The median concentration of sulfur dioxide (SO_2)

tmpd The mean temperature (Fahrenheit)

Note that you can convert `time` to a Date object using

```
as.Date(chicago$time, origin = as.Date("1993-12-31"))
```

which allows you to use it in plots and to determine when specific observations happened. Note also that some of the pollution variables have been shifted, so they contain negative values.

The pollution and temperature variables also have versions whose names start with `lag_`. These versions are 7-day averages of the original variables; for example, `lag_pm10median`

on day 7 is the average of `pm10median` on days 1 through 7. We might expect deaths on each day to be related to the amount of pollution over the past several days, rather than pollution on that day alone, so these variables will help you explore that possibility.

A few observations contain `NA` values. You can ignore these observations.

Your Goals

Doctors and epidemiologists believe that air pollution is an important contributor to mortality; people who live in areas with higher levels of air pollution tend to have higher death rates. There are several kinds of air pollution:

Ozone Ozone is O_3. The ozone layer high in the stratosphere famously absorbs most ultraviolet light from the Sun, preventing it from harming us, but ozone at ground level is dangerous because it is strongly chemically reactive and can damage the lungs.

Sulfur dioxide Produced as a byproduct of combustion, such as by burning coal, sulfur dioxide is both toxic and a contributor to acid rain.

Particulate matter Small particles of solid material suspended in air, such as soot from fires, sea salt from the ocean, or mineral dust from mining and construction. Small particles can get deep into the lungs and are associated with lung problems.

It is also known that extreme temperatures tend to lead to higher death rates, for instance among elderly residents who do not have heating or air conditioning.

Famous billionaire Preston Jorgensen, searching for ways to extend his life, is interested in knowing which pollutants are most strongly associated with increased deaths. Specifically, he would like to know:

1. Is there evidence that these pollutants are associated with increased mortality? Which pollutant seems to be most strongly associated with mortality?

2. Is the effect of pollution instantaneous (i.e., the amount of pollution on a certain day only affects death rate on that same day), or does it extend over time (i.e., it affects the death rate over the next few days)?

3. If the level of each pollutant were somehow lowered to match the lowest value ever recorded, what mean death rate would we estimate for Chicago on a 70-degree day? Should Jorgensen spend his enormous fortune to try to reduce pollution this much?

The template Rmd file **guides you through answering these questions** one step at a time, so you should follow the template as you work. It also gives more specific detail about what specific models and methods you should use.

Your Introduction and Conclusion should give your results in plain English, so Jorgensen can read your answers to the questions; the rest of your report should be written so it can be ready by someone moderately familiar with regression.

Note that our focus is inference, not prediction. You must address the questions in the R Markdown template—submissions that do not address these questions will not receive credit.