36-402 Data Exam 2

Jacky Liu (jackyl1)

May 7, 2021

Introduction

Preston Jorgensen would like to to know which pollutants are most strongly associated with increased deaths. Given in-depth information and data recorded over a span of 13 years for over 5,114 observations (2), we will try to answer the following three questions. The first is whether air pollution associated with a higher death rate, and which pollutant has the strongest association with mortality. There are many variables that may also impact mortality, so we will attempt to control for time and temperature as well (studies have shown that lower temperature is associated with longer lifespans). Next, we plan to answer whether the effect of pollution happens instantaneously, as in affecting the death rate on the same day, or rather affecting the death rate over a period of time. For example, we want to know if a day with bad pollution will have more people dying on that same day compared to a day with light pollution. Lastly, using our model, we want to predict what the average death rate would be if we lower the level of each pollutant to match the lowest value recorded in Chicago given a fixed 70-degree temperature. (1)

Our analysis led us to the conclusion that pollution overall has a relationship with mortality, with two of the pollutants being significant in our model and one of them being slightly significant. With our bootstrap analysis and prediction, we concluded that lowering pollution would indeed likely result in a lower mortality rate, though there were limitations with our study and assumptions were made. Additionally, we found that the relationship between temperature and mortality is significant, and held temperature constant while predicting with our model. Lastly, we predicted that by lowering the mean pollution to the values of the lowest pollution day recorded in Chicago, the mean mortality rate is likely to be approximately between 101.5267 and 106.5222 deaths. (3)

Exploratory Data Analysis

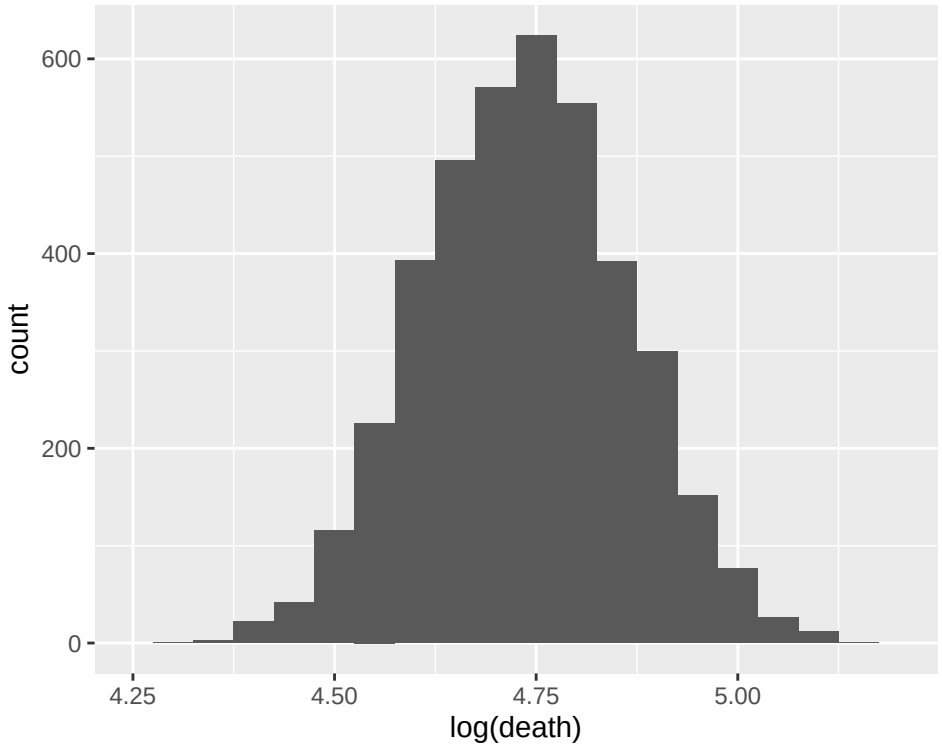

Figure 1: Marginal distribution of log mortality

The data being used is community data from Chicago, Illinois, which was recorded daily from January 1987 to December 2000. It also includes the total number of non-accidental deaths recorded in Chicago each day and the mean temperature each day, to study the affect of air quality on fatality and answer our three questions. We define several key variables. The first is our response variable, which we will call death. This is the number of non-accidental deafig.cap=ths recorded on a given date. Next, we have three variables which are types of pollutions: pm10median, which is the median density of particulate matter that are less than 10 mm; o3median, which is the median concentration of ozone; so2median, which is the median concentration of sulfur dioxide. Our predictor variables also include temperature, which is the mean temperature of that day in Fahrenheit units. (1)

The log of our response variable, death, is very normally distributed (figure _) with most data lying between 4.5 and 5 log deaths per day. (2) The predictor variables all seem to have somewhat but not perfect normal distributions, with pm10median and so2median being very noticibly right-skew, while o3median and temperature seem somewhat uni-

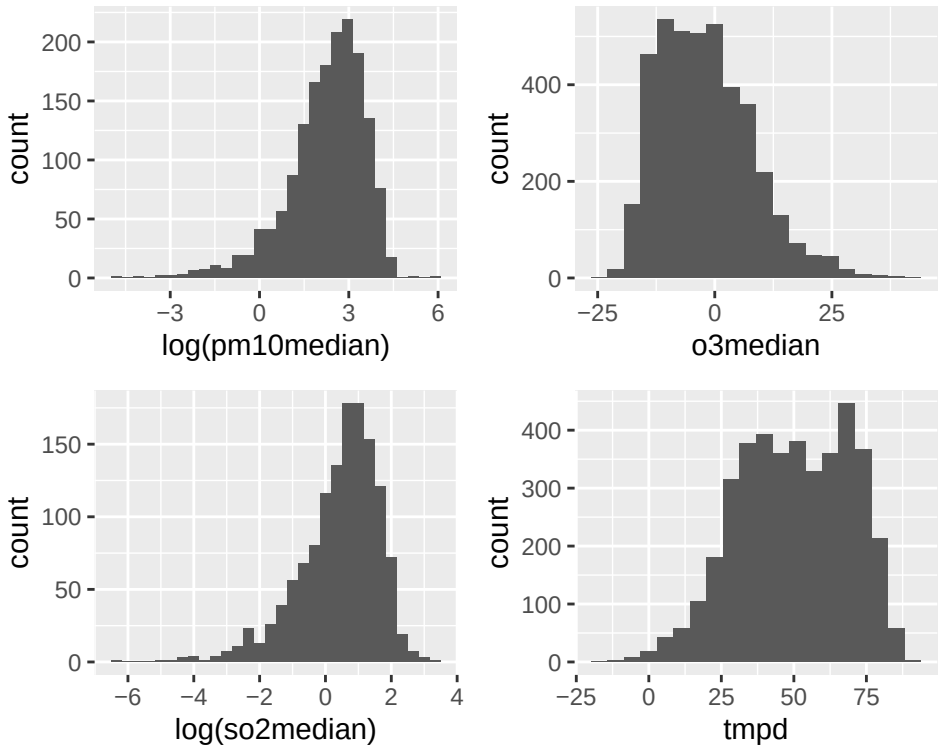

Figure 2: Marginal distribution of predictor variables

form with small tails. Temperature can also be seen as slightly bimodal instead of uniform or normal.

Looking at the ggpairs plot, we can see from here that the largest correlations are between the three pollution variables. This makes sense for the pollution variables to be associated with each other. An interesting thing to note is that there is notable correlation between temperature and death (5), which makes sense because as previously mentioned, studies have shown that temperature is associated with lifespan.

In fact, regarding temperatures and deaths, looking at the deaths over time plot, we can see that the number of deaths per day seem to remain constant through the 13 years while fluctuating up and down except for one outlier. Based on research, this spike that seems to occur around 600 days after December 31st 1993 is allegedly the 1995 Chicago Heat Wave, which led to 739 heat related deaths. (3) We have decided to keep this data in for modeling because this can technically be explained by one of our variables, temperature. Since studies have shown that higher temperature locations are associated with lower life expectancy, and the fact that heat waves are not uncommon in the US, we have decided to keep this data.

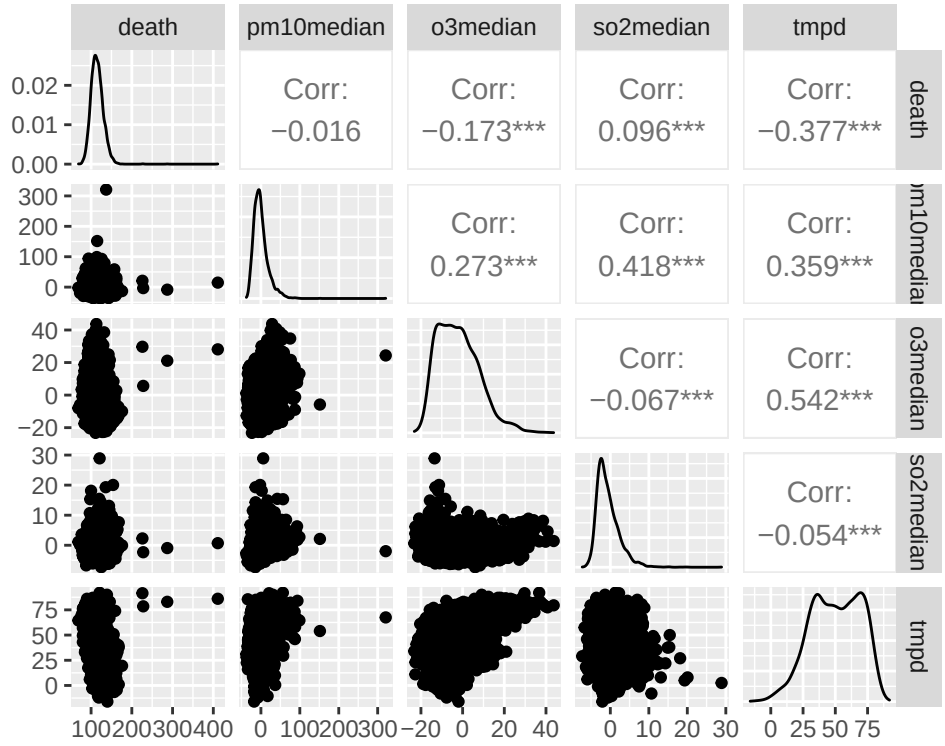

Figure 3: 2D Exploratory Data Analysis

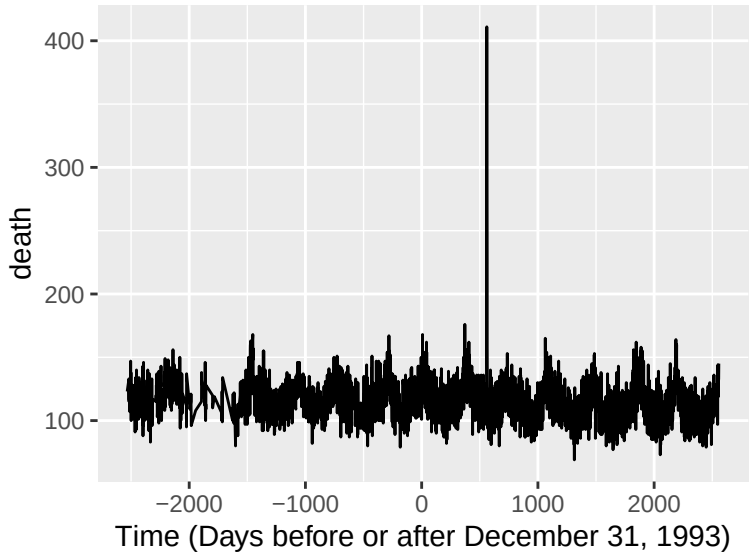

Figure 4: Deaths over time, year 1987 to 2000.

Modeling & Diagnostics

Due to the right-skewness of pm10median and so2median in our initial exploratory data analysis, we will be using the log of these variables from here on for modeling. o3median and temperature do not seem skewed enough to warrant a transformation so we will keep them as is. Time will not be included in our model as it does not seem to have an association with our response or any of our predictors.

We construct two models in order to answer the questions asked by Preston Jorgensen, both of which are General Additive Models. One of them uses non lagged covariates while one of them uses lagged variables. Our models are as follows: (1)

```
model1 = gam(log(death) ~ s(log(pm10median)) + s(o3median) + (log(so2median)) + s(tmpd))
model2 = gam(log(death) ~ s(log(lag_pm10median)) + s(lag_o3median) +
s(log(lag_so2median)) + s(lag_tmpd))
```

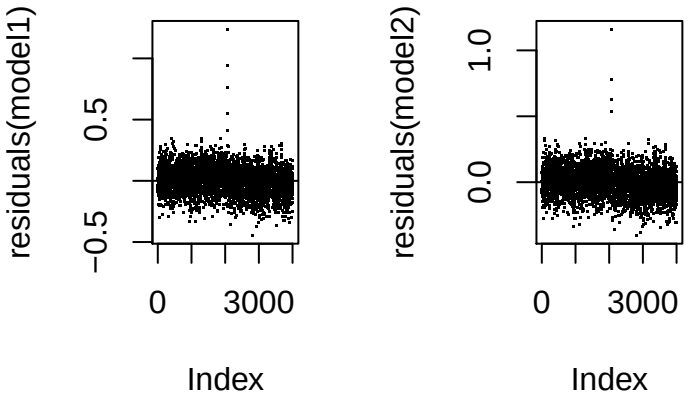

Figure 5: Diagnostic plots for our models

Looking at the residual plots for our two models, we get nearly identitcal results with the residuals evenly spread around the 0-line with constant variance, but the potential for a few outliers. Looking at the normality of the residuals (figures not shown), we again see very similar plots. Both plots show that a majority of the residuals are normally distributed with slightly non straight tails. The fact that the residuals are not completely normal slightly undermines the validity of each model's ability to perform inference but it is plausible enough where these can take place.

Results of CV	Non Lagged	Lagged
Estimated MSE (log)	0.01315587	0.01225358
Standard Error (log)	0.0006147497	0.0005032997

Using 5-fold cross validation, we have determined that model 2, the additive model with lagged covariates, fits best to the data with a mean square error of 0.01225358 (log) deaths and standard error of 0.0005032997 (log) deaths. (2)

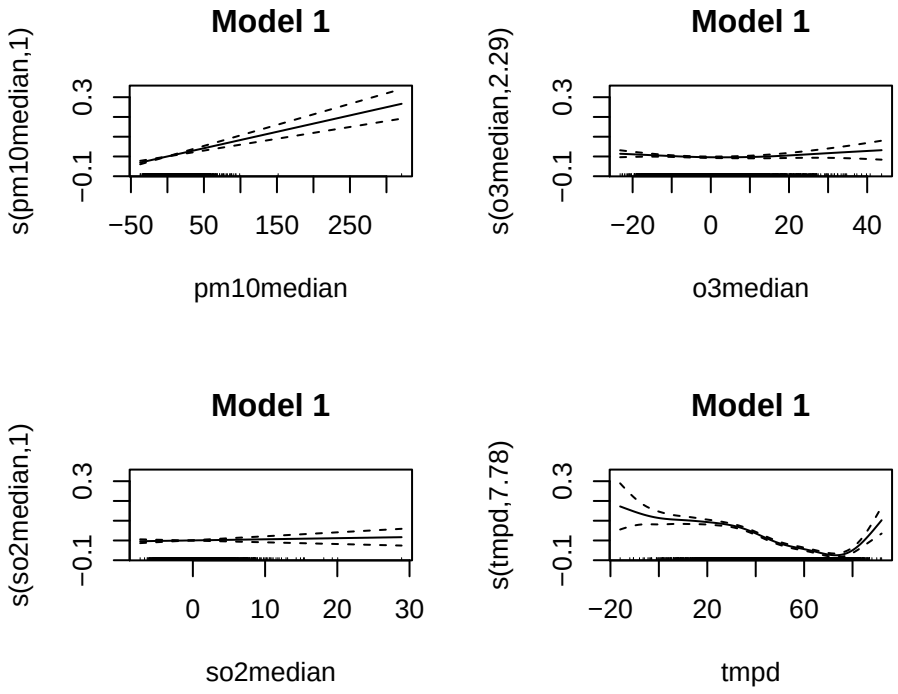

Figure 6: Plots for Model 1

The difference between models do not appear significant as the mean square errors of our 5-fold cross validation are fairly similar. (3) However, keep in mind that these results are based off the logged values, so the actual difference in days between our prediction errors is like from one to a few deaths. Despite that, we will still pick our lagged model (Model 2) as to use moving forward.

Results

Our anova F test suggests that model 2 is a better fit than model 1 (p-value = 2.2e-16). Therefore, we can conclude that model 2 fits our data fairly well for our purposes. (1) To determine whether the pollutants are associated with mortality, we fit a model that uses only temperature as a feature, not the pollutants:

```
model3 = gam(log(death) ~ s(tmpd))
```

From this model, we determine that there is indeed an association between pollutants and log mortality (p-value=2.2e-16). (2) Based on our cross-validation results and our model, including the time variable shows that it is insignificant. Therefore, we conclude that the effect of pollution is most like not instantaneous. (3)

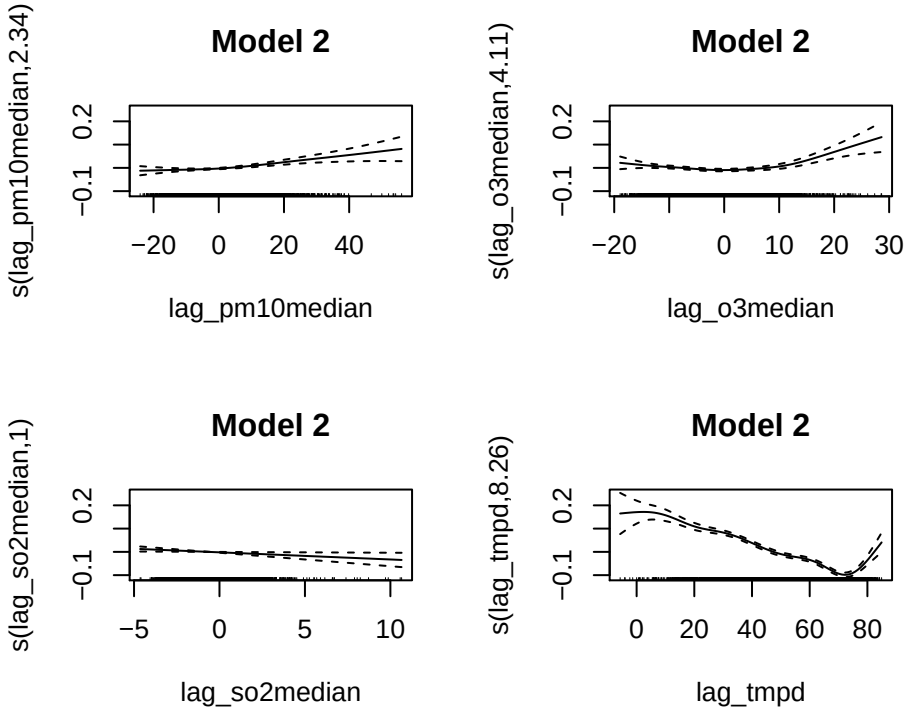

Figure 7: Plots for additive model

Looking at the plots for our additive model, we can see that o3median has the highest association with mortality, followed by pm10median. (4) The plot for so2median shows that there is possible but less associated compared to pm10median and o3median. This is confirmed by the summary of our model, which shows p=0.0241 for so2median, which

means that we can only conclude association under an a=0.05 level but not an a=0.01 level.

Using the minimum values for so2median, pm10median, and o3median, which are -7.30, -37.37, and -23.19 respectively, along with a temperature of 70 degrees, we found the 95% confidence interval for the mean number of log deaths is (4.620322, 4.668353). This means an untransformed interval of approximately (101.5267, 106.5222) deaths. (5) Using resample cases bootstrap to bootstrap for our sample mean, we get a 95% confidence interval of (114.4973, 115.4629). (6) This interval is higher and does not overlap with our 95% prediction interval using minimum values for pollution while holding temperature constant. This indicates about our model that our model predicts that minimum values of pollution leads to less deaths. This means our model is suggesting that there is an association of pollution and mortality, more specifically that less pollution is associated with lower mortality. (7)

Conclusions

In this study, we used an additive model to examine the relationship between mortality and pollution. We concluded that there was a significant relationship between pollution and mortality. (1) We accounted for confounding variables such as temperature, which is known through many studies to be correlated with mortality. In addition, there was a heat wave in 1995 of Chicago that caused an outlier amount of fatalities, which likely impacted our findings to some degree. Lastly, we found that by lowering the mean pollution to the values of the lowest pollution day recorded in Chicago would likely result in the mean mortality rate decreasing from 115 to between 101.5267 and 106.5222 deaths. Based on our model, Jeff Preston Jorgensen can indeed conclude that reducing pollutants would cause mortality to decrease. (2) However, there are limitations to our model, as there may be other confounding variables that our dataset does not cover; an example is that weather conditions, which can be associated with pollution, may also affect fatality. (3) In addition, there are also limitations because our data is only recorded from one city; it may be more helpful to analyze data from different cities to see if there is truly a casual relationship between pollution and mortality, or if Chicago is an outlier before Jorgensen spends an enormous fortune to reduce pollution this much. Finally, because we used observational data, even using a bootstrap analysis cannot guarantee that this relationship is causal; it may just be correlational.