36-402 Data Exam 2

Edwin Baik (ebaik)

May 7, 2021

Introduction

Yet again, we have been called on by Preston Jorgensen to help him in his journey to become immortal. This time, we have been asked by him to study specific pollutants like ozone, sulfur dioxide, and particulate matter can potentially decrease one's lifespan. We wish to first study these variables, and recognize if there is some significant relationship between any of these pollutants and the mortality rate, and if yes, which is the strongest. We then wish to see if the effect of these pollutants can have an instantaneous effect of death rate, or if it affects the death rate over some period of time. Ultimately, once we solve these questions, we wish to estimate the mean death of Chicago on a regular 70 degree day given the LEAST amount of pollution! This estimate can help us understand if it is worth it for Mr. Jorgensen to invest his fortune in bettering the pollution rates on Earth **(1)**. The data we were given is found from the Chicago area, and contains up to 5,114 observations over 6 variables. However, some of these observations contain NA values, and thus we will ignore these observations, so we may not use all 5,114 observations. Note also, with these 6 variables, we are given a "lagged" version of the variable, which will take the 7-day averages of each variable. Note the "unlagged" variables are pm10median, o3median, so2median, time, tmpd, and death. Here, we are trying to do inference on the death rate, thus death will mainly be our response variable **(2)**. Our final findings show that pollutants are significant in the death rates within the Chicago area, and also temperature. We found that pollutants do not have as much predictive power on the death rate than temperature does. Furthermore, it may not be worth it to try to lower pollutant values as even with the minimum pollutant values ever recorded with Chicago, on a regular 70 degree day, we still cannot reach the lowest death rate of 69 deaths. We found that the actual death rate would approximately be between 109.88 and 110.81 deaths **(3)**.

Exploratory Data Analysis

Before we do any explanatory data analysis, we wish to first remove all the NA values from the dataset. This will in turn help us later when we are doing any plotting/EDA as we do not have to deal with cases in which certain data is omitted and thus we cannot use. Note once we remove all the NA values from the dataset, we are left with 4,012 rows worth of data.

Using the data provided by the chicago database, and the observations within the data, we will use a portion of the variables given to answer the next few questions. We first look at the variables and what each represent. In this report, the death variable will be the response variable, and simply represents the non-accidental death count on that certain date. We then have our explanatory variables, which first starts with the time. Note there is no real distribution for time, as it is sequential, but it will be necessary to solve one of our essential questions: does pollutants affect the death rate instantaneously/over time. We also have the pm10median, which is the median density of the PM_10 pollution, the o3median, which is the median concentration of ozone at sea level, so2median, which is the median concentration of sulfur dioxide, and temperature, which is simply the mean temperature (measured in Fahrenheit) on that day. We also have the lagged version of those variables, but as described before, is simply the 7-day average of the medians. Temperature is also assumed to be another explanatory variable that can help us predict death rates. We first look at the response variable, death rate, and its histogram. At first glance, the death rate seems to be normally distributed, but the distribution more so represents a left-skew, as there are many more data points leaning towards the larger values of death rate as shown in Figure 1. This indicates that the median is greater than the mean of the response. Note since we will be using Poisson GLM distributions, we must check how the log of the death rate visualizes. When we simply log the death rate, there is a MUCH more normal distribution, and the left-skew from before almost disappears. One concern is that there quite the number of outliers towards the larger edges of death rate, and whilst it may be one or two points, may potentially be of wariness later on **(2)**.

When we look at the explanatory variables, we have 3 to mainly look at for potential inference/prediction: pm10median, o3median, so2median. Immediately, the pm10median and the so2median distributions are both right-skewed, with distributions looking similar to the death rate distribution. Also, some of these points within the two variables are outliers, so a log-transformation may be optimal to normalize the data. Note that this can

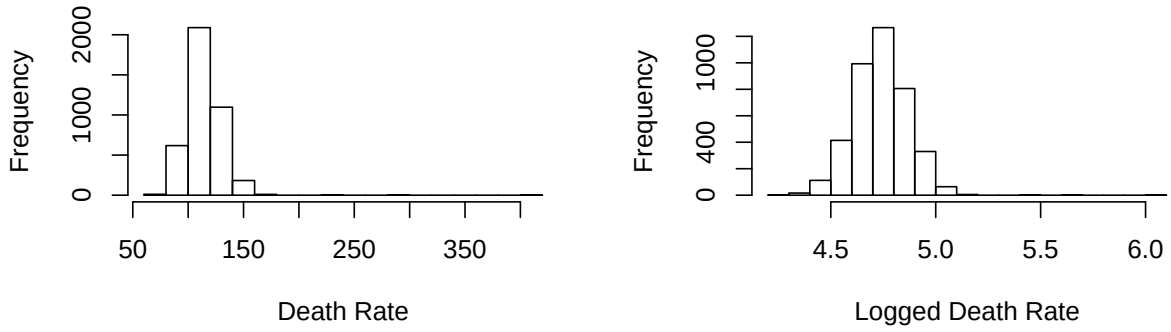


Figure 1: Histogram Regarding Death Rate & Logged Death Rate

all be found within the second figure, where the pm10 median and sulfur dioxide median distributions are defined. The o3median distribution is extremely unique. Note the distribution has no real single peak, but is consistent in frequency through multiple ozone values, as shown within Figure 3, with the histogram of the ozone median. Overall, it can be said that no general transformation has to be made regarding the o3median, but may be subject to change. Note that we do not add the temperature averages as a histogram because there is obviously going to be a clear pattern between the four seasons over time **(1)**.

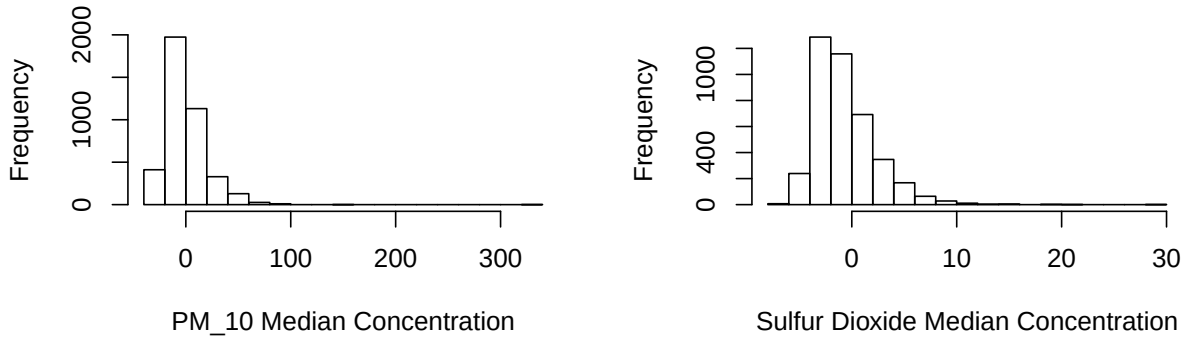


Figure 2: Histogram PM10 and Sulfur Dioxide Distribution

Next, we wished to see how the variables that we chose were related to each other so that when we do modeling, we can create a better idea of how to fit models to the data. When we compare our predictors to our response variable, we note a few ideas. Over-

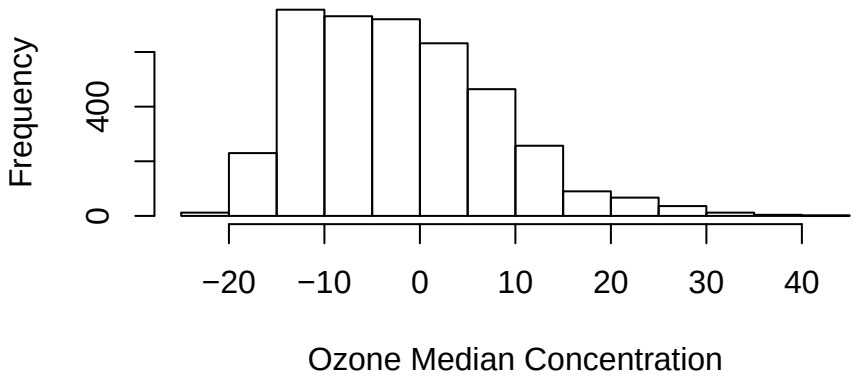


Figure 3: Histogram Regarding Ozone Median Distribution

all, there doesn't seem to be some linear relationship between death and any of the four explanatory variables. There are multiple concerns however about outliers. Note that the death vs pm10median has multiple outliers both above and to the right to the data mass. It is beneficial that the explanatory variables do not have any relationships with each other, as we can assume independence between the explanatory variables. However, as found within our EDA and the given scatterplot matrix, we would like to do multiple transformations on the data. After the transformations are done, as shown in Figure 5, the relationships between the PM10, SO2 and death (both log transformed) still does not have some linear relationship. Still similar to the plots pre-transformation, there are no noteworthy remarks to mention regarding the two scatterplots over the response variable and the two transformed explanatory variables **(4)**. We also add a plot showing the death rates increasing over time, and certain patterns are shown as shown in Figure 6. Overall, the death rate is consistent with some fluctuations through each year. An interesting point however is that around the year 1995, there is an uptick in death rates, and this may be because of the heat wave Chicago faced in the summer of 1995. I believe we should still keep the data from this time period however because we have temperature to explain this spike, and potentially other pollutants that may explain this tragic summer in Chicago **(3)**.

In terms of modeling decisions given the EDA and the scatterplots created, it is with no question that we must log transform the death rate. It normalizes the death rate and linearizes more relationships between the three main explanatory variables. We should

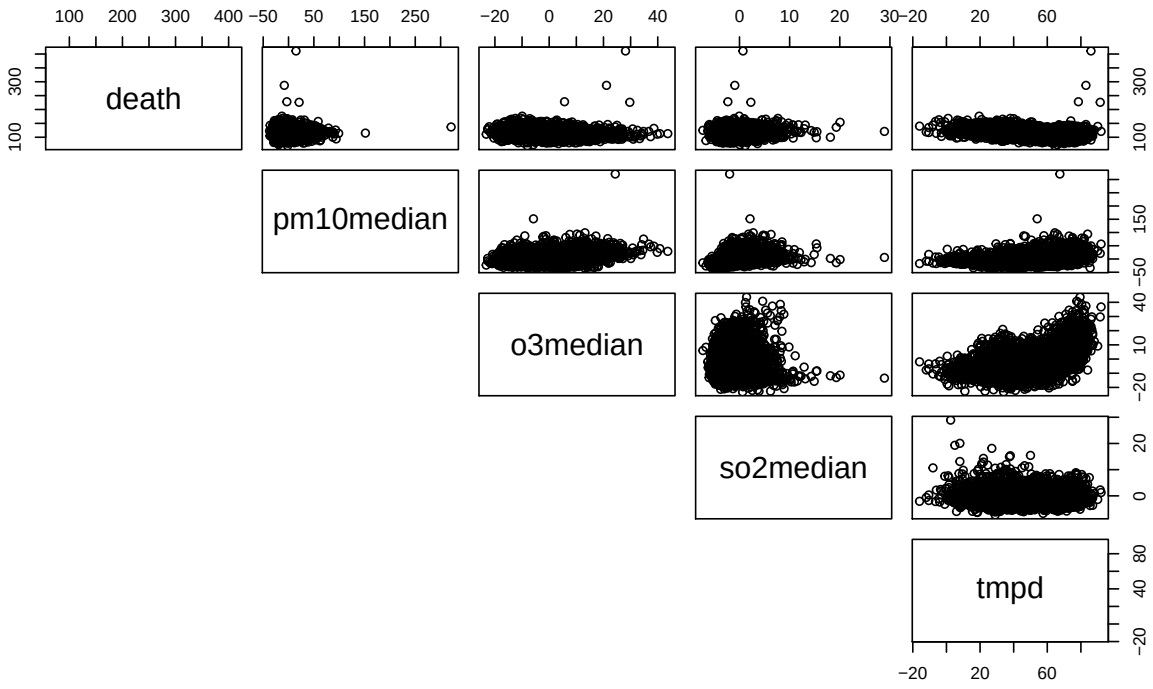


Figure 4: Scatterplot Matrix of Response/Explanatory Variables

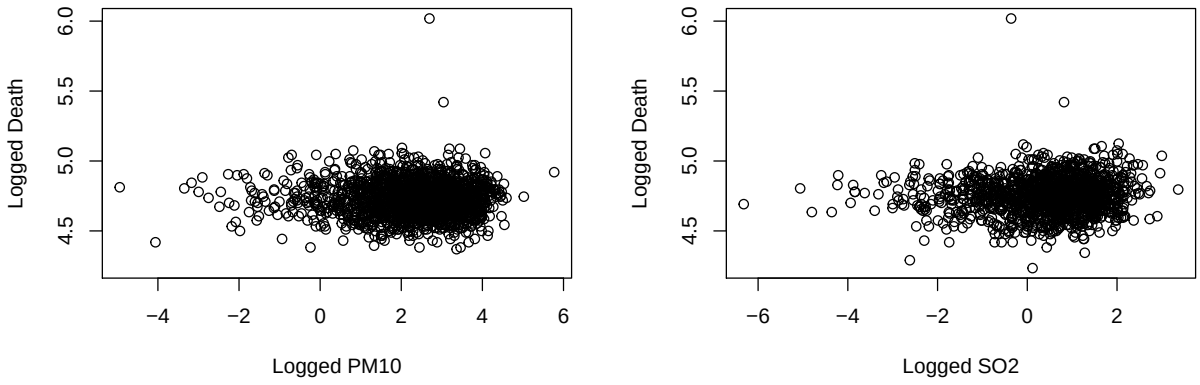


Figure 5: Scatterplot Between PM10, SO2 and Death Rates

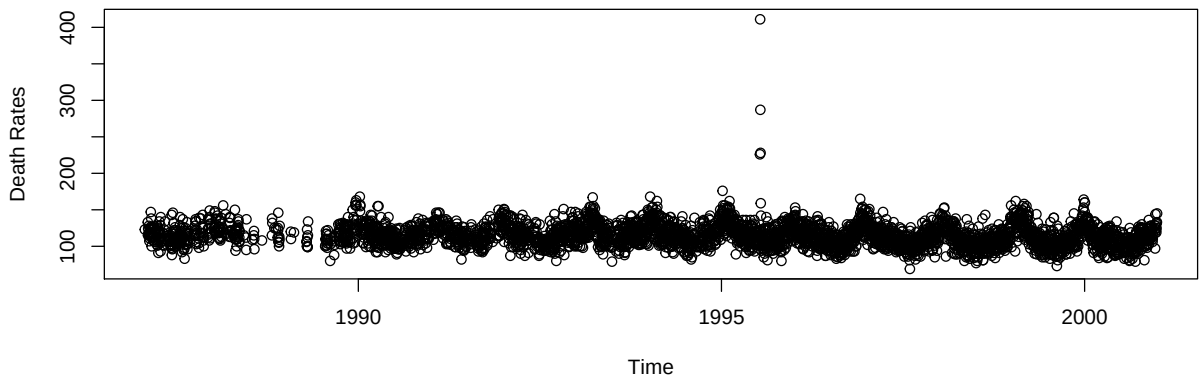


Figure 6: Scatterplot Between Time and Death Rates

also include all the explanatory variables: so2median, o3median, pm10, and temperature average. Although there may be no eye-opening relationship, there may be some correlation overall that might help us predict the death rate. For now, I do believe it NOT necessary to log-transform the so2 and pm10 medians as their EDA show if we log transform, no normalizations occur for the new logged histograms. Furthermore, logging the explanatory variables do not improve the relationship between the death rates, logged or non-logged. We do NOT log-transform the o3median or the temperature average **(5)**.

Modeling & Diagnostics

First, we create two Poisson generalized additive models, with both predicting the death rates. We include four different explanatory variables, the pm10median, o3median, so2median, and the tmpd. Note they are all smooth splined with the degrees of freedom equivalent to 4. The second model does NOT use the same explanatory variables, rather it uses the lagged versions of these variables **(1)**.

We now use cross-validation to choose which model will be best in predicting the death rates. Here, we will use 5-fold cross validation, as it is a proper balance between the amount of folds being created and the amount of validation data that we will use per iteration. Cross-validation is required because it helps to choose a model that is best fit for the data as the diagnostics and goodness of fit from a model is not enough to make that choice. After applying the 5-fold CV, we find that model 2 seems to be better in terms of error, both in MSE and SE, albeit very closely **(2)**. The errors are shown below:

a) (GAM with Non-Lagged) MSE: 200.1598, SE: 34.03415
b) (GAM with Lagged) MSE: 186.8195, SE: 31.74511

Based on the errors found from the cross-validation, it does not seem that the difference in errors seem significant. The difference in errors is miniscule at best, with a difference of 0.10 in the MSE and 0.15 with the SE. In terms of uncertainty, when the CV is run consistently, model 2 seems to be ahead of model 1 constantly, potentially indicating even the slightest betterness in models. However, based on the prediction error, we use model 2 as our model to move forwards on **(3)**.

Results

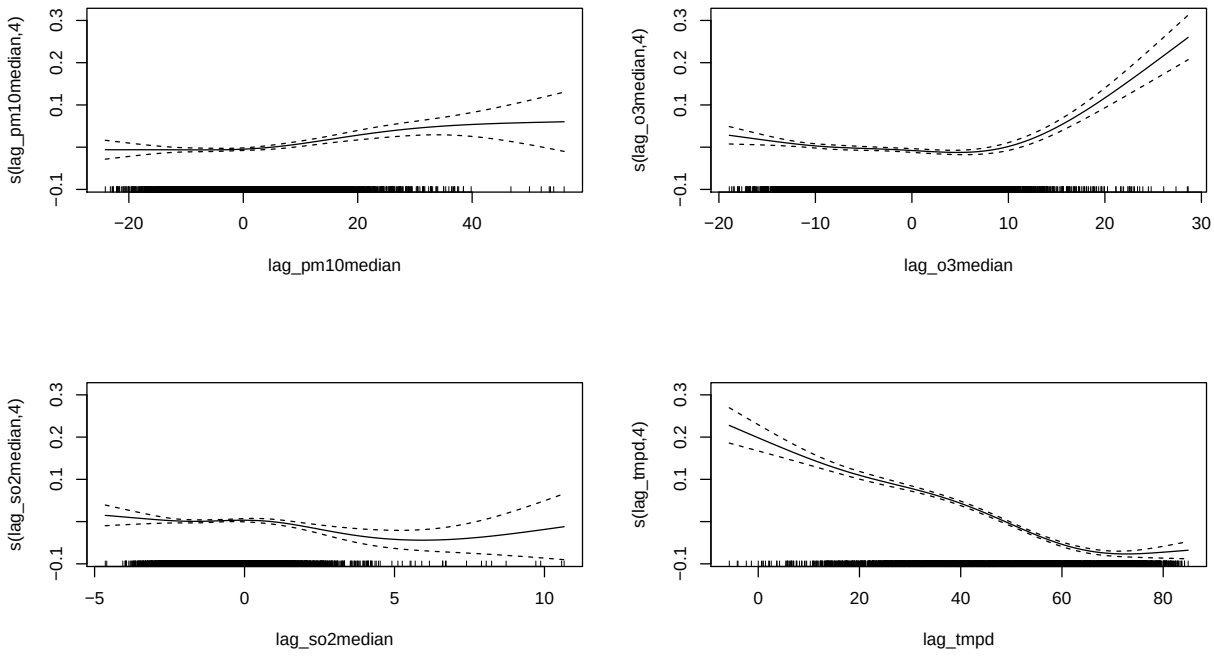
We first run an ANOVA model on the chosen model earlier (model 2), and see an interesting amount of details regarding it. For one, all the predictors, which are the lagged versions of the original explanatory variables, are all significant in predicting the death rate. Most variables have an extremely high chi-squared value, indicating their significance in helping predict the death rate. Overall, the test shows that the model fits well into the response variable **(1)**.

We then create a model using only the temperature average as the predictor for the death rate. Note for consistency, we have this model also be a GAM and the tmpd is also smoothened. When we compare the AIC values between our chosen model and this new model with only temperature as its predictor, we notice that the AIC value of model 2 is lower by 30 then the AIC value of the new model. We also run an ANOVA model between the two models, and notice that the model with the pollutants has a significant p-value. This then indicates model 2 is a better model. This in turn means that there is some significant importance between the pollutants and mortality in terms of prediction **(2)**.

Given the cross-validation results from before, we noted that model 2 seems to be a better model for predicting the death rate than model 1 by a smaller margin. This then indicates that the effect of pollution seems to be extended over time, as model 2 uses the lagged versions of the pollutants, which is a 7-day averaged value. As these values are being grabbed over time, and thus a better predictor for death rates, it seems the pollution affects mortality over time **(3)**.

There are multiple ideas that can be found from the plots created from our chosen model 2. For one, the pm10median and the so2median (both lagged), do

not show any real relationship. It's seen that whilst the line is close to the interval, the pm10median and so2median do not change drastically given that it increases over time. However, the o3 median (lagged) shows promise as the death rate potentially increases as o3median increases. Temperature seems to be the biggest change in death rate, as the line changes drastically as temperature increases. We also plot the residuals and note that apart from the residuals, the residuals are normal from the qqplot and have constant variance and zero mean **(4)**.



We wish to create a 95% confidence interval to predict the death rate GIVEN the fact that we have the lowest values of pollutants on a 70 degree temperature day. Note the minimum values for the lagged versions of the pm10median, o3median, and the so2median respectively all are -24.119, -18.923, and -4.6432. We do NOT use the minimum values of the non-lagged pollutant variables because model 2 consists of the lagged version for the predictors. When we produce the confidence interval, we recognize the death rate is approximately between 108.3517 and 112.3464 **(5)**.

We then attempt to do bootstrapping over 1000 iterations to create a pivotal confidence interval that tries to predict the death rate given the minimum values of the lagged pollutants and the fixed 70 degree day. We specifically resample the residuals within the parametric bootstrap for a few reasons. First, we can assume that our model fits the proper assumptions necessary for a parametric bootstrap as normality seems to be met within the residuals. Furthermore, we do not assume anything within the residuals, but we

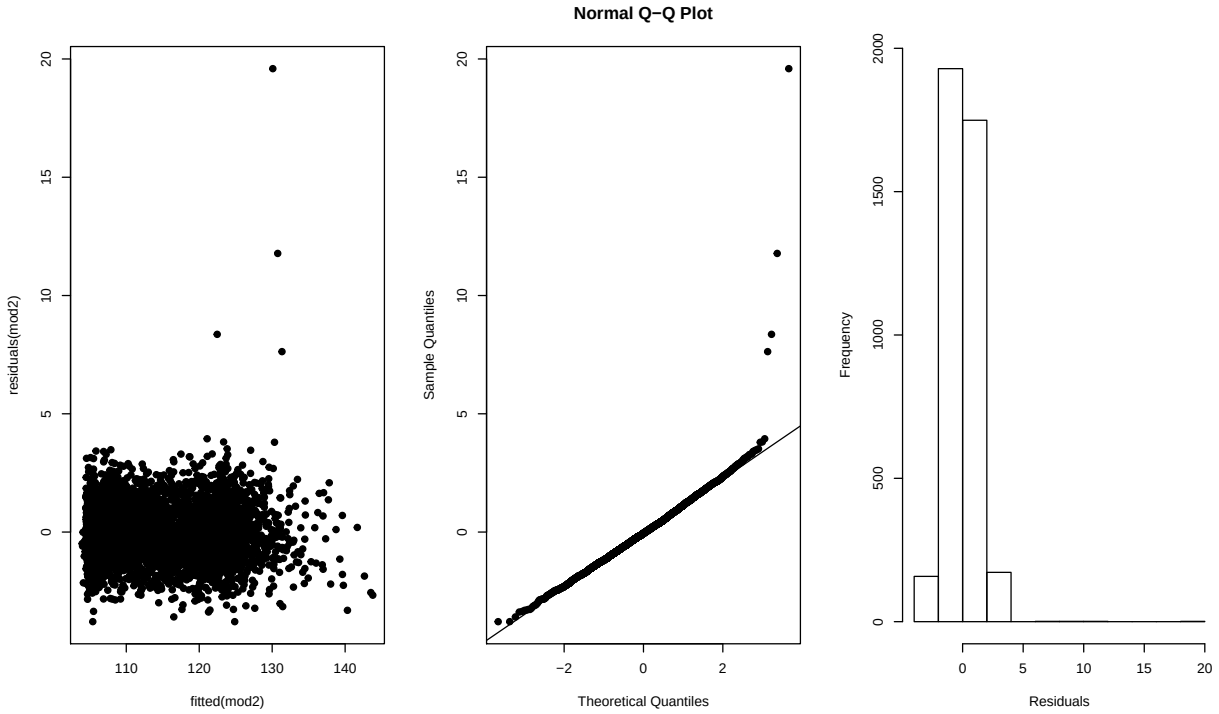


Figure 7: Residual Plots of Model 2

can say that our model 2 seems to be properly following the correct regression function. When we do the resampling residuals bootstrap, we recognize over the 1000 iterations that a pivotal confidence interval consisting of the death rate being between 109.8857 and 110.8156. This means that with 95% confidence, the mean number of deaths given that it is the minimum values of pollutants and a 70 degree day will be between 109.8857 deaths and 110.8156 deaths **(6)**.

I do see some differences between the two confidence intervals. This may be because the bootstrapped CI works under the assumptions that the model still follows the shape of regression function. This may not be the case within the first confidence interval, allowing us to make the overall assumption that the model is a good model in terms of predicting the death rate given its pollutant and temperature explanatory variables **(7)**.

Conclusions

Given our findings regarding the pollutants and the temperature against the mean death rates, there are multiple ideas that we can glean. First, we recognize that there does seem to be some association with atmospheric pollutants and mortality. When we created

multiple models, each containing some number of explanatory variables, we recognized the model that contained the atmospheric pollutants against the model that only had temperature averages was MUCH better in predicting the death rate. However, based on the partial function plots created from our chosen model, we do notice that temperature had the biggest "power" in predicting the death rates. The pollutants, while significant, did not have as much predictive power. Specifically, the higher the pm10 or the o3, the higher the death rate may potentially be. Inversely, the higher the so2, the lower the death rate seems to go. Temperature seems to also change inversely to the death rate **(1)**. Given this, we can assume to Mr. Jorgensen that reducing pollutants CAN cause mortality to decrease. However, it will NOT decrease to the lowest death rate Chicago ever recorded of 69 deaths, as ours calculated to be approximately 110 deaths **(2)**. There are limitations that must be considered for this report however. For one, our dataset is partially missing due to the massive frequency of missing data in both the non-lagged and lagged versions of the pollutants and the temperatures. This is of concern as we wish to have as many samples as possible for a much more accurate representation of the answer for the question provided. Furthermore, there are an extreme amount of variables that can lead to a lower or higher death rate, so they may act as confounding variables to our research. Transformations were also a large limitation. We did not employ the log transformation specifically to any of our explanatory variables as it did not necessarily help normalize our data. If we are able to find a specific transformation that can help to normalize our data, this may lead to more accurate results as we can assume much more about our variables and thus our model **(3)**.