36-402 Data Exam 2

Raymond Li (rli3)

May 7, 2021

Introduction

Lately, the talk about employing products and habits that are "environmentally friendly" is getting more and more common. New markets such as vegan clothing, fake meat, and electric vehicles becoming bigger, using "better for the environment" as their lead selling point against their competitors. These kinds of products claim to emit less pollution than their counterparts either in use or in the production process. As a result, it is argued one way or another, these environmentally-friendly products will improve human health in the long term and perhaps even reduce mortality rate.

(1) Three of such common pollutants is Particle Matter, Ozone, and Sulfur Dioxide. The purpose of this study is use different data analysis and modeling techniques to determine whether or not the presence of these three pollutants influence mortality rate. In addition, this study aims to find whether or not if the effects of pollution are instantaneous, or if it causes death after a lag of time. Lastly, using a predictor, determine whether or not it is worth the efforts to minimize the count of these three pollutants specifically by estimating a death count when pollution is minimized.

(2) The dataset that will aid this study is in "chicago.csv", and it contains 5114 observations of 6 variables. The overall dataset aims to keep track of air quality of Chicago every day from January 1987 to December 2000.

The features in this dataset includes the count of each of the three pollutants mentioned before, in their default scientific measurements. The set also includes "tmpd", which is the mean temperature of that day, in Fahrenheit. "death" is the count of non-accidental deaths that occurred in Chicago on that day. And "time" is given by the number of days elapsed after the start of the study.

Later on, four new features were added- and they show the seven day moving averages of the three pollutants and the temperature. These fields aid to answer the second question regarding whether or not the effects of pollution happens instantly or over time.

(3) Overall, the study found that the presence of these three pollutants is associated with mortality rate in Chicago. However, the models used did not do a good job of finding that association. The effects of these pollutants did show to have a lagging effect, meaning the non-accidental deaths was typically result of the pollution count of a week before than of that day. However, despite the association between mortality rate and pollution levels, Jorgensen should not waste his time trying to lower the count of these pollutants as there is no statistical evidence that doing so would reduce the death count by a significant amount.

This study was also subjected to many limitations, which will be mentioned towards the end of the report.

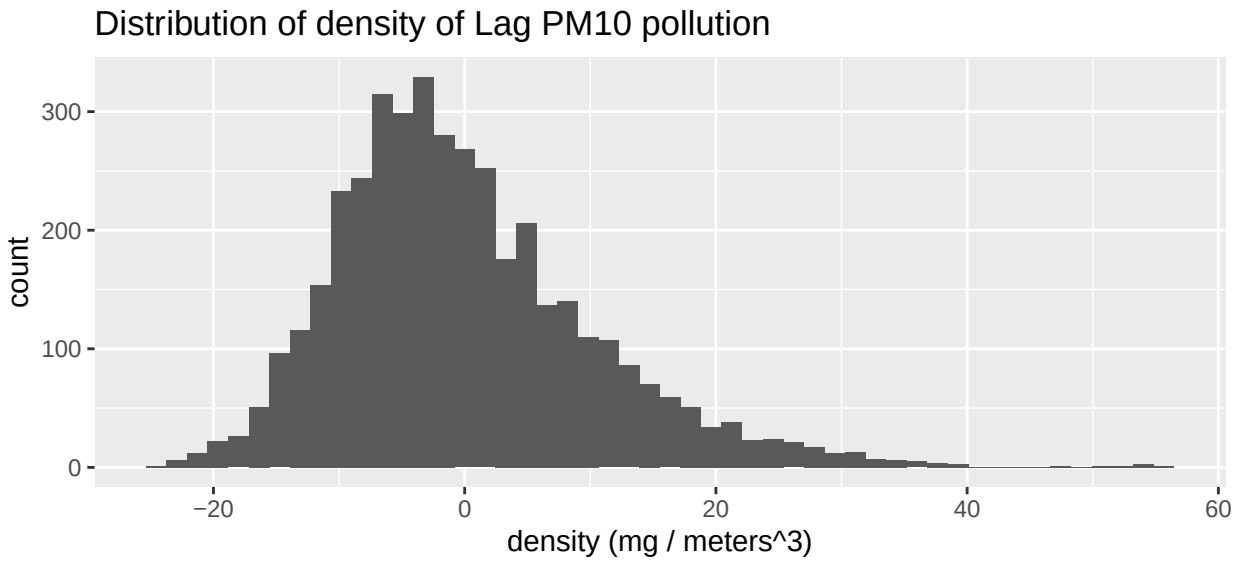
Exploratory Data Analysis

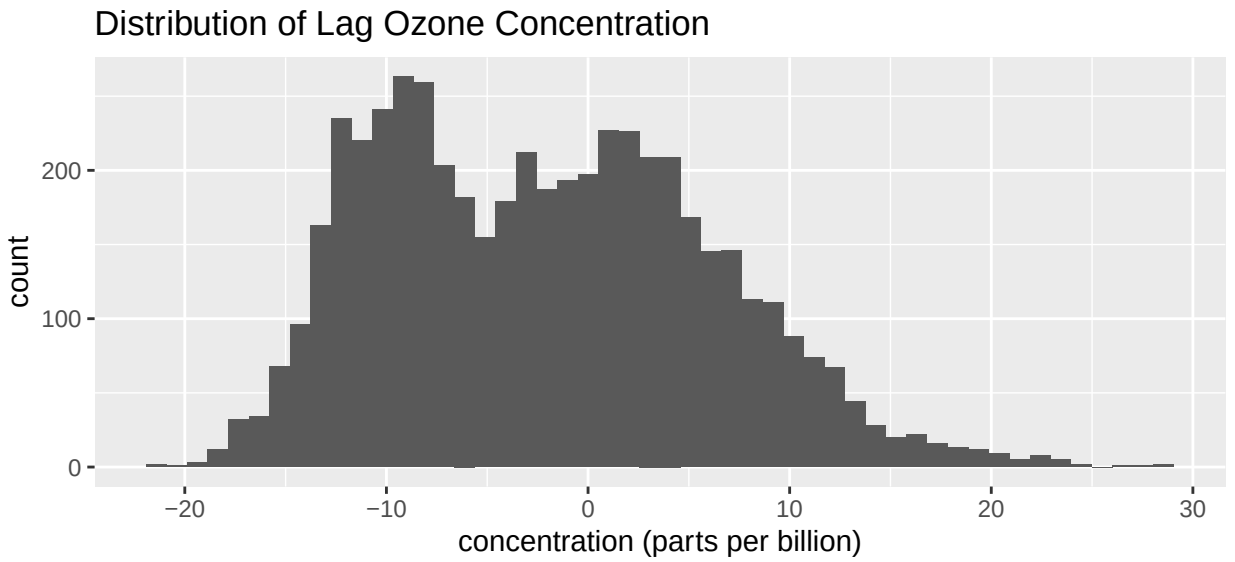
Some of the less important figures of this section will be hidden so that the report can be kept within 12 pages.

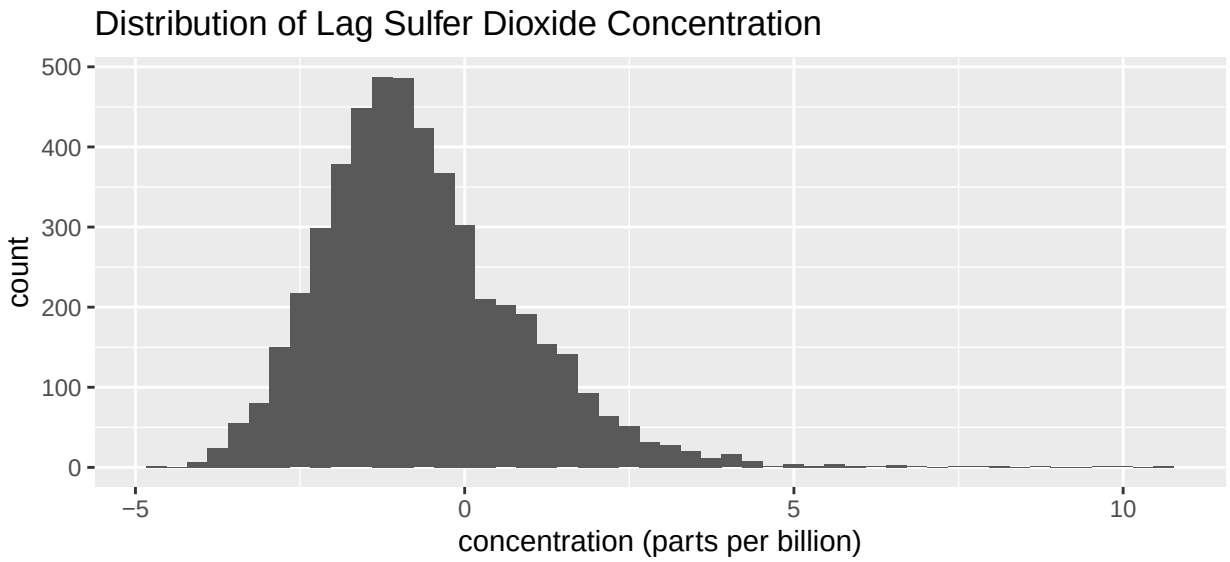
(1)

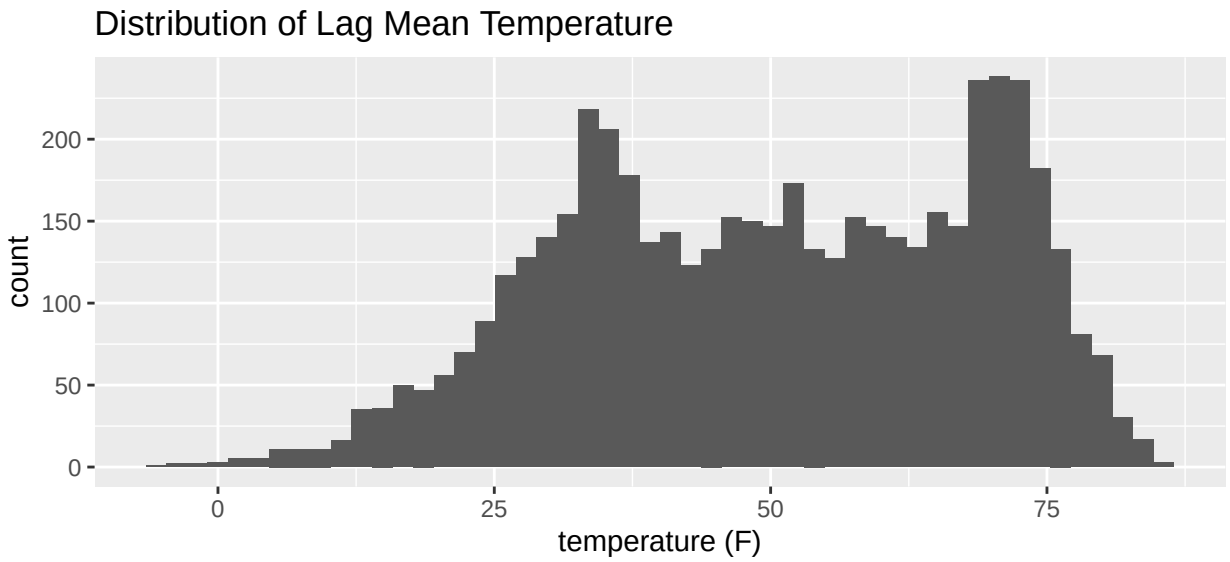
The distribution of PM10 and Sulfur Dioxide appear to have a right-skewed distribution with a few upper-level outliers. This distribution of Ozone also seems slightly right-skewed, but is closer to a normal distribution. The outliers for Ozone is also not as pronounced. This distribution for mean temperature is multimodal and appears to have a slight left skew, if any at all.

When examining the summaries, the left skew of PM10 and SO2 is reinforced by the difference in values between the median and min vs the median and max values. For O3 and temperature, those values are much closer. We also notice NA values for PM10 and SO2, which may indicate the possibility of removing rows later down the line. The summary and plots of the non-lag features will be hidden to save space for the report.


Distribution of density of Lag PM10 pollution
count
300
200
100
0
−20
0
20
40
60
density (mg / meters^3)


Distribution of Lag Ozone Concentration
count
200
100
0
−20
−10
0
10
20
30
concentration (parts per billion)





When examining the lag "version" of all the features discussed previously, the distributions still retain a similar skew, but the magnitude of that skew is much less. Overall, all four lag features appear to have a relatively more normal distributions than the non 7-day lag counterparts. This is expected, as the moving average of a value intuitively makes a better predictor. As it result, it is most likely the case the lag values will be used as the features for the models instead.

	lag_pm10median	lag_so2median	lag_o3median	lag_tmpd
Min.	-24.12	-4.643	-21.34	-5.714
1st Qu.	-7.035	-1.688	-9.031	35.14
Median	-1.683	-0.8747	-2.57	51

	lag_pm10median	lag_so2median	lag_o3median	lag_tmpd
Mean	-0.02429	-0.649	-2.167	50.22
3rd Qu.	5.465	0.1624	3.683	67.29
Max.	56.08	10.66	28.63	85.43
NA's	1054	139	6	6

The trend seen before is reinforced in the summary table- namely in the smaller "max" values. This indicates fewer and more manageable outliers. Since lag values cannot be calculated until after the first week, there are now NA values for every feature. It is most likely the case that all 1054 of those rows will be removed when performing operations later on.

(2)

The response variable is death. And its distribution is visibly close to normal, with a median value of approximately 115. The daily death count does have upper limit outliers, as seen in the summary, where the maximum value is 411- way above the median and the rest of the distribution. The non log-response plots and tables will be hidden to save space.

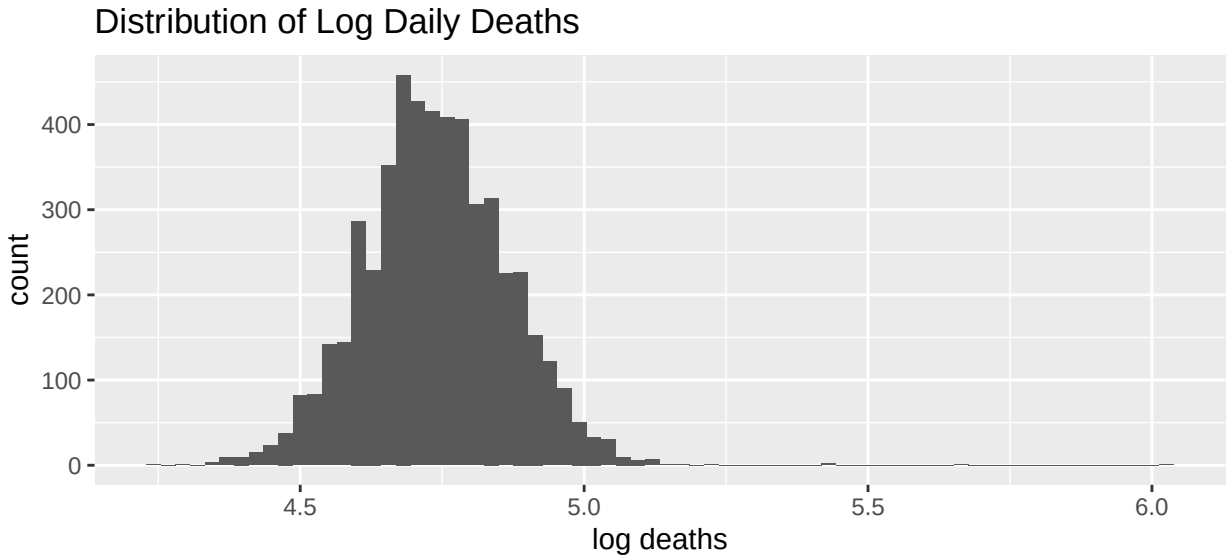


Figure 1: Figure 10: Distribution of Log Deaths by Day

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.234	4.654	4.736	4.74	4.82	6.019

When viewing the log distribution of deaths (for the use of Poisson GAMs), the distribution still looks normal with a median value close to 4.75. Since there is a log operation done, the outlier points do not appear to be far. Overall, the log transformation makes the distribution marginally closer to true normal. From now on, all operations involving deaths will be log transformed.

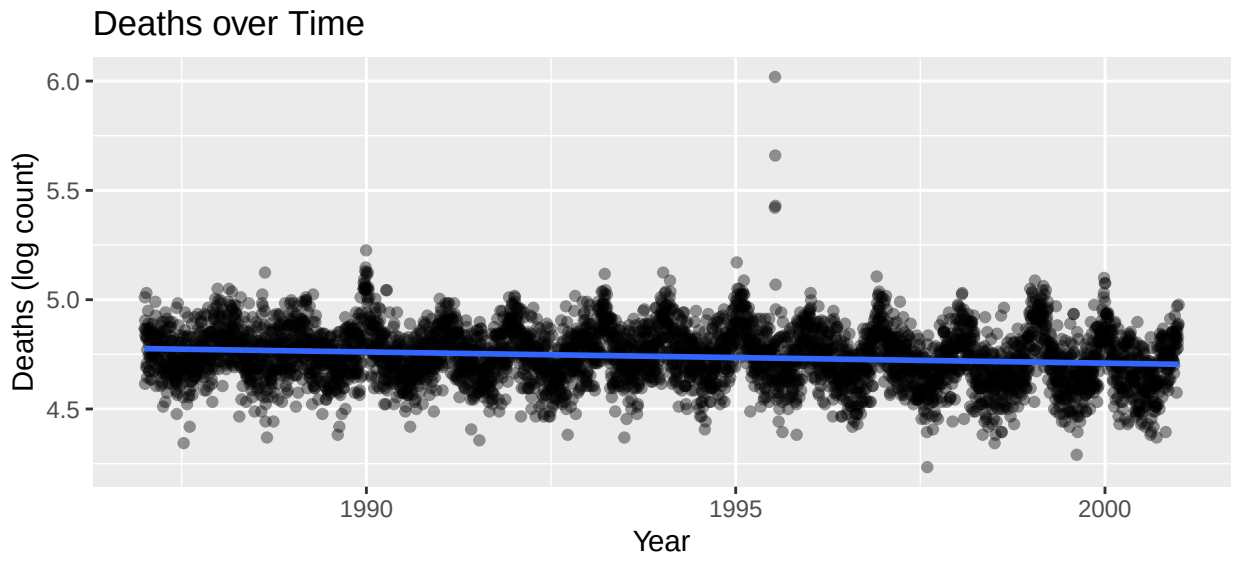
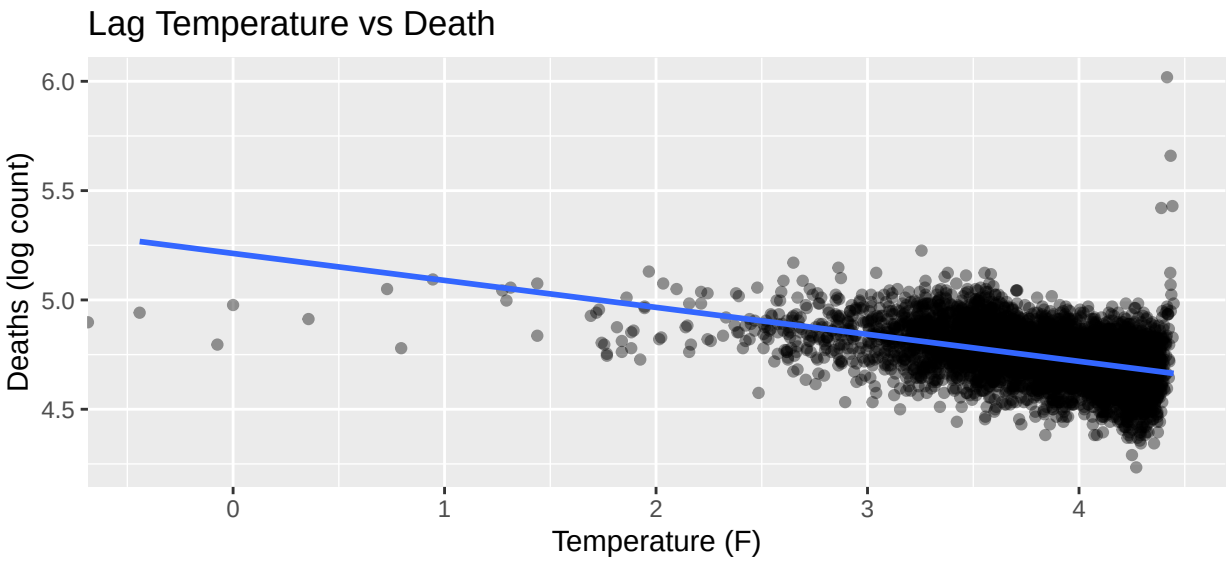


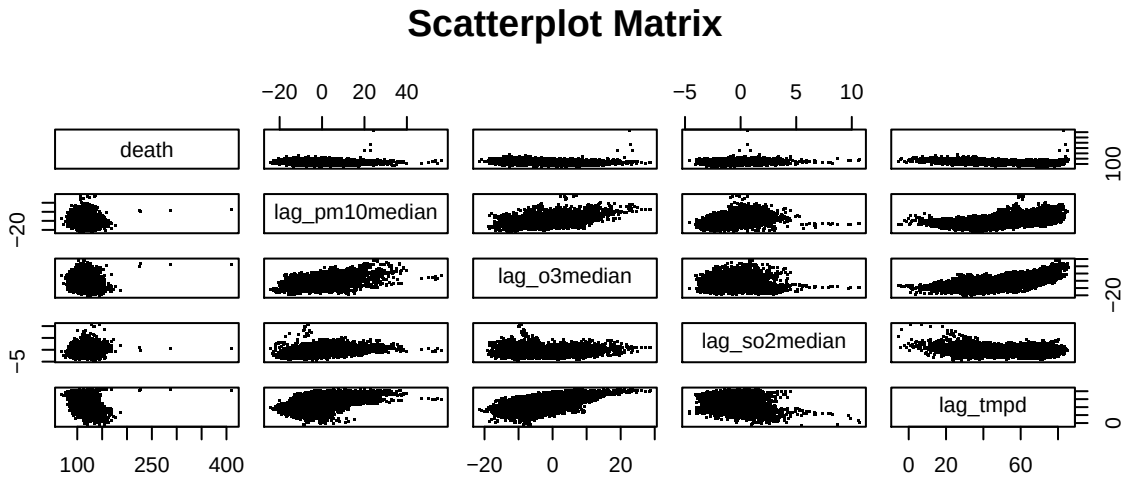
Figure 2: Figure 11: Distribution of time to log(death)

(3) The outliers can be explained by a heat wave that occurred in July of 1995. Since the deaths of the heat wave did not result from the rise of any of three pollutants, it makes more sense to remove the three rows from the dataframe to avoid biasing the regression.

The log death count was compared to the log of each of the four main features. The variable temperature seems to have highest correlation based on the scatterplots, as it has a clear negative relationship. Even then it is unclear. The scatterplots of the non-lag features with death are not displayed to save space.



When repeating the process with the lag variation of the features, the result is similar. However the relation between log death and each of the four features were slightly enhanced. Again, some of these scatterplots will be hidden to limit the length of the report.



(4) When looking at the pairs plot, it does not look like log death has a high correlation with any of the four standard features. A similar statement can be made when repeating the process with the 7-day time lag variation of the features.

```
##                      death lag_pm10median lag_o3median lag_so2median   lag_tmpd
## death           1.0000000     -0.1488917   -0.2574795     0.1331646 -0.4457561
## lag_pm10median -0.1488917      1.0000000    0.4806279     0.2968651  0.5024541
```

```
## lag_o3median   -0.2574795      0.4806279    1.0000000    -0.1124395  0.6693197
## lag_so2median   0.1331646      0.2968651   -0.1124395     1.0000000 -0.2286623
## lag_tmpd       -0.4457561      0.5024541    0.6693197    -0.2286623  1.0000000
```

(5) When looking at the correlation matrix, it does not appear that death has a high correlation with any of the aforementioned features, which supports the pair plots shown above. However, it is the case with both the lag and non-lag version of the features that temperature has a moderate correlation with both Ozone and PM10. This may lead to possible collinearity down the line, but it is unlikely. The pairs plots and correlation matrix with the non-lagged features will be hidden to save space.

However, as mentioned before, the 1045 rows of NA will be removed for future operations. The time lag version of the variables will also be used in the future, when applicable, because they have a better distribution and contain fewer outliers. Lastly, the four outlier rows that occurred as a result of a heat wave will be removed as well.

Modeling & Diagnostics

(1) The Poisson generalized additive model for the non lag features looks like this:

$\log(Death) \sim \beta_0 + s(pm10median) + s(o3median) + s(so2median) + s(tmpd)$

The GAM for the lag-version of the feature is similar and looks like this:

$\log(Death) \sim \beta_0 + s(lag_p m10median) + s(lag_o 3median) + s(lag_s o2median) + s(lag_t mpd)$

(2) After running the 5 fold CV for both models, we document the mean and standard deviation of the errors in the following table:

```
##           c.cv_mean.     c.cv_sd.
## Reg Model 0.01235252 0.0004039351
## Lag Model 0.01151782 0.0004077421
```

Based on the table shown above, the error value for the regular model is 0.01235252 The error value for the lag model is 0.01151782

(3) The cross validation error means appear to be very close, with the lag models having a slightly smaller error value (0.0008347). A similar thing can be said for the standard deviation (0.000003807), though in favor of the regular model. Despite being just marginally

better, if a model had to be selected for future operations, it would be model 2, the model containing the lag variables instead of the standard variables.

Based on the uncertainty of the estimates, the difference in error rate between the two models are NOT statistically significant. A t-test was done with the two sets of values to confirm this finding (p = 0.1839), though it is not required.

Results

The first procedure is to conduct a goodness-of-fit test of, model_lag.

(1) The value of the GoF test is 0.0876378, which is greater than the expected threshold. Therefore, it can be concluded that the model is NOT a good fit for the data. This implies the time lagged data points of the temperature and pollution levels are not meaningful influencers of death in Chicago, at least not by themselves.

(2) Next, to determine if just the three pollutants are related to mortality rates in Chicago, a reduced model including only the lag temperature is initialized and analyzed. That new model looks like this:

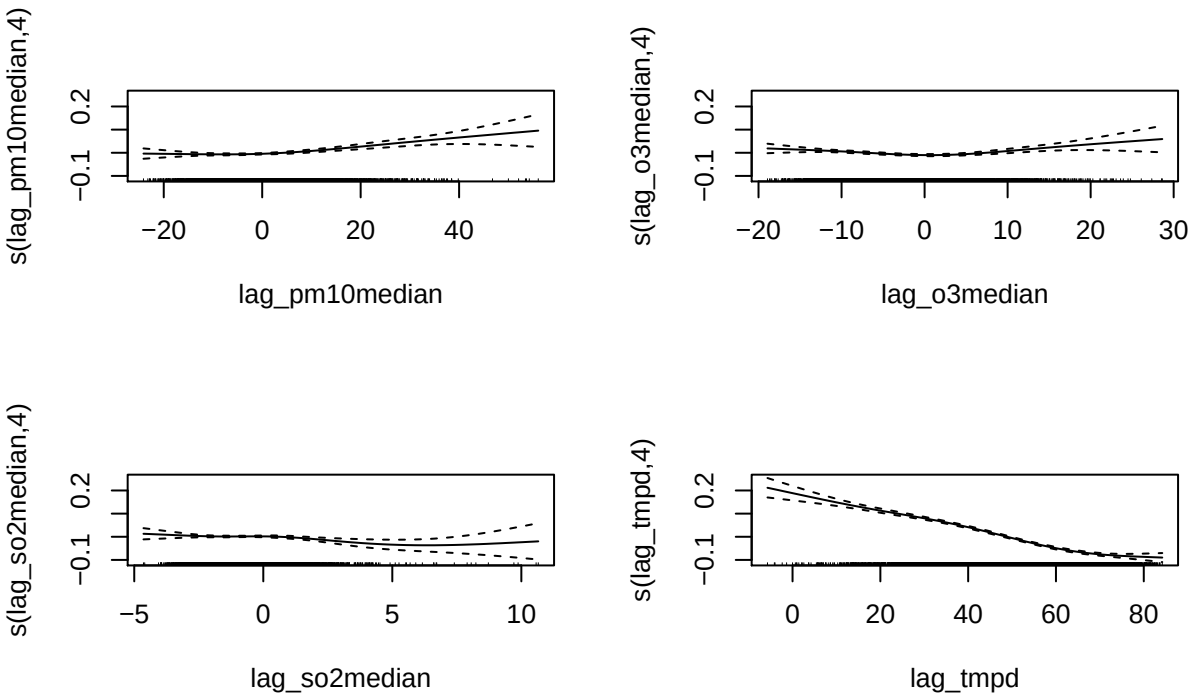
$\log(Death) \sim \beta_0 + s(lag_t mpd)$

An Anova test will be conducted between this new model and the chosen model to find a potential association between pollutants and mortality. The hypothesis test is testing H0 = s(lag_pm10median) + s(lag_O3median) + s(SO2median) = 0.

The p value of the anova test is 1.188 x 10^(-15), which is much smaller than 0.05. This indicates a rejection of the null hypothesis and it can be concluded that the three pollution factors have a statistically significant impact on mortality rate in Chicago.

(3) Based on the cross validation section, the effects of pollution appears to extend over time, as it has been concluded the model containing the time-lagged features is a marginally better representation of pollutants vs mortality.

(4) Plots of the time-lagged model can also be shown to demonstrate the average relationship between each pollutant and death.



Based on the figure shown above, it appears the pollutant most associated with mortality is Ozone. This indicates the possibility that the pollutant Ozone caused increases in mortality rates in Chicago more than both Particle Matter and Sulfur Dioxide.

Now, the smallest value observed for each pollution variable will be taken an using the lag model, a 95% confidence interval for number of deaths will be formulated with the lag temperature value fixed at 70. The minimum values for particle matter, ozone, and sulfur dioxide is -24.11919, -18.9239, and -4.643285, respectively. These values, along with the fixed value 70 for temperature will be the row of data used for the confidence interval based on the lag model.

Lower	Upper	Predict.fit
105.3	113.3	109.2

(5) The 95% confidence interval for the number of deaths using the predict.gam method is shown in the table above. The fit value predicted is 109.2 deaths, while the confidence interval is (105.3, 113.3). For reference, the standard error of the fit is 1.019.

Next the resampling bootstrapping method with 1000 replications will be used to find the 95% confidence interval of death count with the same input data as before. Then both the prediction fit value and confidence interval will be compared.

Lower	Upper	Predict
104.7	114.8	109.7

(6) The result of the 95% CI using the bootstrap is noted in the table above. The range is (104.7, 114.8). The bootstrapped range, which was found manually by obtaining the mean of the two values is 109.7.

(7) The predict value is very close, with the bootstrap method having a fit value of 0.5 higher. Overall, the bootstrap method had a higher range in the interval, meaning the lower end of the CI was lower than the predict and the upper end of the CI was higher than the predict. Otherwise, the upper and lower thresholds of the two measurements are very close, differing by a count of about 1. The wider (or less precise) range with the bootstrap method can possibly be explained by not having all of the model assumptions being met. It was showed earlier in part (1) of the section with the goodness of fit test, that the model was not necessarily a good fit. The model assumption that is suspected to be violated is homoscedasticity which happens when there is a pattern with the variance of the residuals in the residual plot. A residual plot was attempted to confirm this violation, but there were too many data points on the plot, making it unclear. Therefore, the plot will not be displayed.

Conclusions

(1) Overall, the association between atmospheric pollutants and mortality is statistically significant. This was shown from tests that revealed a difference in how death was predicted based on whether or not the temperature features was included. At the same time, however, the models used for this study did a poor job modeling that association as shown in the various plots and the goodness of fit test conducted. Lastly, mortality was impacted by lags in time, such that a change in pollutant level did not cause death right away. This is seen in how the model incorporating time lags had a smaller error in predicting the death count- though the difference is minuscule.

(2) Preston Jorgensen should conclude that reducing pollutants in an attempt to lower the mortality rate of Chicago is a big waste of time. Based on the prediction intervals, the death count was calculated to be about 109 to 110. According to the preliminary data exploration that was conducted, the median death count was 114. This means that reducing the pollution amount to the lowest amounts in the entire time span of the study

is predicted to reduce less than five deaths a day. This calculation was even done with the upper end outlier removed. Had those values not been removed, the resulting predictions would have even been less in Jorgensen's favor with a higher death prediction count.

(3) The first limitation in this study include the role of temperature. The temperature feature was utilized in the predictive model that is supposed to predict whether or not just pollution was causing extra deaths. The prediction interval was also done with the assumption that the temperature was 70 degrees, an arbitrary amount.

Next, about a fifth of the data set could not be used because it contained NA values from having using the time-lagged predictor variables. This limited the sample size, though it most likely did not have a big impact on the result.

Depending on the set seed value, another instance of the study conducted could have concluded that the non time-lagged model is more accurate because the error counts for both models were very close. Choosing the other model due to a lower error rate can possibly change the entire second half of the study. Though again, the two models turned out to be quite close in terms of predicting mortality, so this likely would not have had a big impact on the result either.

Lastly, this dataset used observational data specifically from Chicago. It is unclear whether or not collecting data from another location may undo/introduce the effects of a new confounding variable. Minute changes in data collection, such as collecting in different times of day may also influence a change in the dataset itself before any models are even run. Using observational data also prevents the ability to make causal predictions, limiting all results to just speculation and analyzing trends.