36-402 Data Exam 2

Yilin Wang (ID: yilinwan)

May 7, 2021

Introduction

It is widely acknowledged by doctors and epidemiologists that extreme temperature leads to a higher mortality rate, and it is believed by many (though not proved officially) that air pollution could impose threats on the heath. However, it is rarely studied how the effects of air pollution on mortality differ for different kinds of pollutants. In this report, we will focus on three kinds of air pollutants: ozone, sulfur dioxide (SO_2), and particulate matter with a diameter less than 10 micrometers (PM_{10}). We will try to answer the following research question: Are the levels of these air pollutants associated with mortality after controlling for time and temperature? If so, which air pollutant is most strongly associated with mortality? Is the effect of air pollution instantaneous, or does it propagates and extends over time? To address Mr.Preston Jorgensen's concerns, we will demonstrate our findings by estimating the mortality in Chicago when air pollution is at the minimal level in the historical record **(1)**.

To answer our research question, we will use the *Chicago* dataset. *Chicago* has 5,114 observations on air quality, temperature, and the number of non-accidental death in Chicago each day from January 1987 to the end of December 2000. More specifically, air quality is measured as three variables: the median density of PM_{10} pollution, the median concentration of ozone and SO_2 **(2)**.

We find that a decrease in the level of ozone and PM_{10} is generally associated with lower mortality in Chicago. However, it appears that a decrease in the level of SO_2 and extremely low levels of ozone are associated with an increase in mortality, which is worth further research. Among the three kinds of pollutants, ozone is most strongly associated with mortality. From our analysis, we have no evidence to conclude whether the effect of air pollution on mortality is instantaneous or does it propagates over time. However, the latter suggestion is more plausible as its corresponding model yields slightly better performance, and it is consistent with commonsense. If the average air pollution in the past 7 days

is at the minimal level (in the historical record) and the average median temperature is 70 degrees Fahrenheit, the expected mortality in Chicago will be about 107.25 people on 30 December 2000, according to our model. It is lower than the actual mortality on 30 December 2000, which is 145 people. However, we do not have sufficient evidence to conclude that reducing air pollution will lead to a decrease in mortality **(3)**.

Exploratory Data Analysis

There are 6 major variables in the *Chicago* data set. The first variable is *time*, which measures the number of days before or after December 31, 1993. Since it is a variable measuring time, it is clear that it follows the uniform distribution with the mean 0 **(1)**. Using *time*, we recover the *date* variable, which measures the date of the observation. We will mainly use *date* in EDA and use *time* in model-building. *death* measures the number of non-accidental death on particular dates, and it is our response variable **(2)**. From heuristics, we know that *death* will likely follow the Poisson distribution. Thus, we will apply log-transformation on *death*, and we will refer to the transformed variable as *death.log* **(2)**. As a measure of air pollution, $pm10$ measures the median density of PM_{10} pollution in milligrams per cubic meter. Similarly, *ozone* and $so2$ measures the median concentration of ozone/SO_2 (in parts per million) respectively, and *tmpd* measures median temperature in Fahrenheit **(1)**. Also, to facilitate the idea that the effect of pollution may propagates over time, we create the lagged variables, lag_tmpd, lag_pm10, lag_ozone, lag_so2. The lagged variables are 7-day averages of the original variables **(1)**. We noticed that there are missing values in our data. For simplicity, we will conduct a complete-case analysis, which means we will only retain observations that do not contain any missing values **(1)**. The variables measuring pollution have been shifted, which means they may contain negative values.

We first examine variables individually. Figure 1 shows the histogram of the response and explanatory variables. We notice that *death.log* generally follows the normal distribution, and its mean and median are both 4.736 **(2)**. This is a desirable feature. The distribution of $pm10$ appears to be right-skewed, with a mean of -0.2637 and a median of -3.5122 **(1)**. lag_pm10 follows similar distribution as $pm10$ but is less right-skewed, and its median is -1.71240 **(1)**. *ozone* seems to follow a right-skewed distribution, with a mean of -2.024 and a median of -3.164 parts per million **(1)**. The distribution of lag_ozone is quite different from that of *ozone*, as it appears to be bi-modal **(1)**. $so2$ and lag_so2 both follows right-skewed distribution. They have similar mean (around -0.64 to -0.66), but the median of $so2$ (-1.2696)

appears to be larger than that of *lag_so2* (-0.8725) **(1)**. *tmpd* and *lag_tmpd* follows similar bi-modal distribution, and they have the very similar mean (around 50.2) and median (both 51) **(1)**. In general, except for *tmpd*, for all explanatory variables, the distribution of the original variable and the lagged variable appears to be right-skewed. However, the lagged versions appear to have smaller variance, and their distributions are less skewed than the original variable **(1)**.

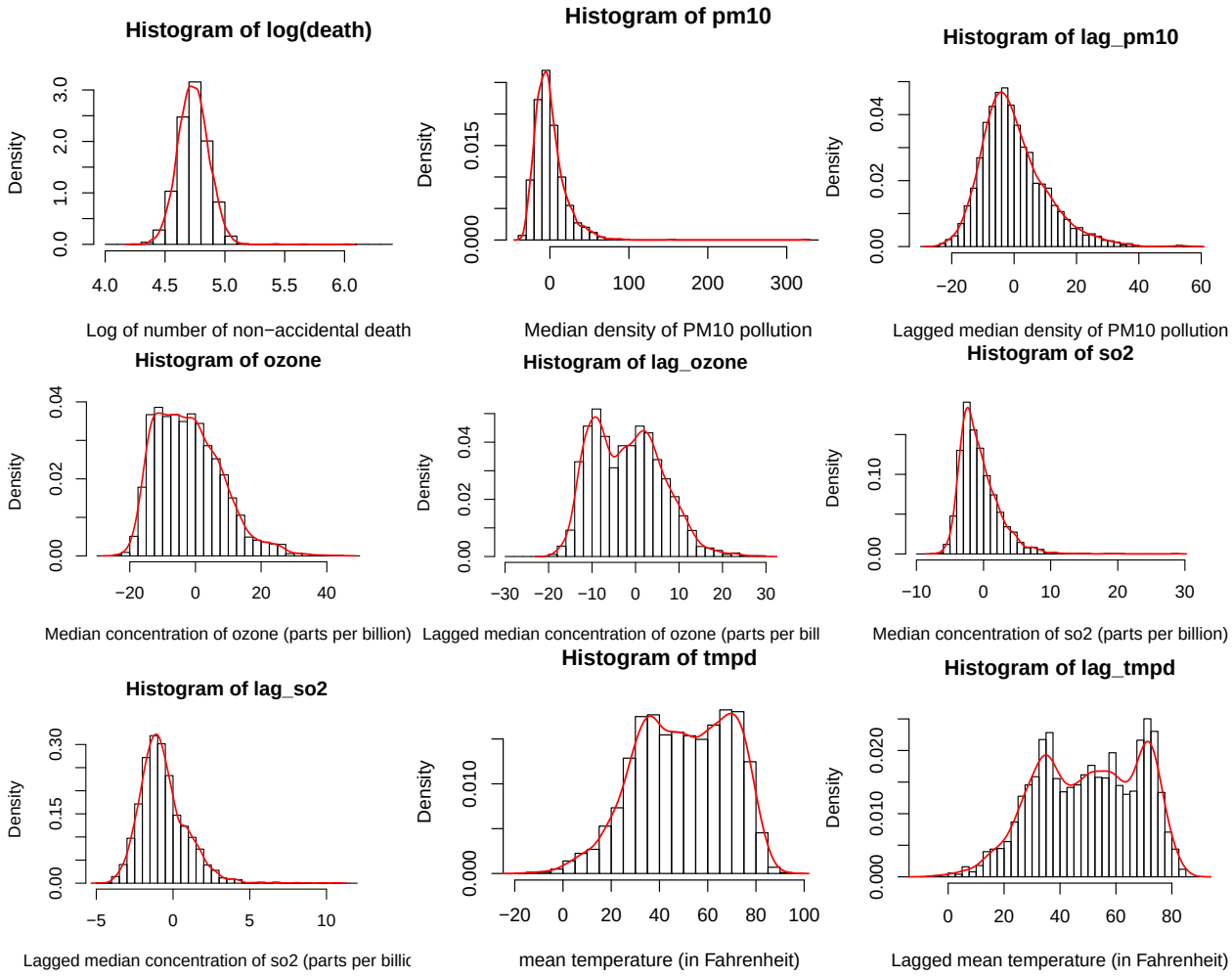

Figure 1: Histogram showing the distribution of response and explanatory variables

We then examine the relationship between variables. Figure 2 present a matrix of plots demonstrating the relationship between the response variable and the original explanatory variables **(3)**. The plots on the diagonal show the marginal distribution of the key variables, which is similar to the histograms in figure 1. While the plots on the upper part off the diagonal show the correlation coefficients between variables, the ones below the diagonal are scatterplots between variables. Notably, the plot on the (2, 1) cell shows how *death.log* is distributed along time. It is noticeable that the *death.log* is particularly high for some

dates in July 1995 **(3)**. Consequently, the scatter plots in figure 2 also suggest that the relationship between *death.log* and explanatory variables are highly affected by outliers (which is likely to be these observations in 1995). The exponential mortality in 1995 is likely a consequence of the 1995 Chicago heatwave, which caused 739 heat-related deaths in Chicago in 5 days **(3)**. As the *Chicago* data contains measures of temperature *tmpd*, we will preserve these outliers as they will likely be explained by our predictor variables **(3)**.

From figure 2, we see that there do not seem to be clear relationships between *death.log* and *pm*10 or *so*2. This is verified by the scatter plot (which is respectively on cell (3, 2) and (5, 2)) and the weak correlation coefficient (respectively -0.015 and 0.099) between them **(5)**. As suggested by the correlation coefficient (-0.192), *death.log* appears to be negatively associated with *ozone* **(5)**. However, the relationship between them seems weak as suggested by the magnitude of the correlation coefficient. In addition to measures of air pollution, we also see that the relationships between *death.log* and *tmpd* and *time* appears to be strong **(5)**. Also, it appears that *so*2 is strongly correlated with *tmpd* (correlation coefficient 0.542), which suggests that our models might suffer from multicollinearity **(5)**.

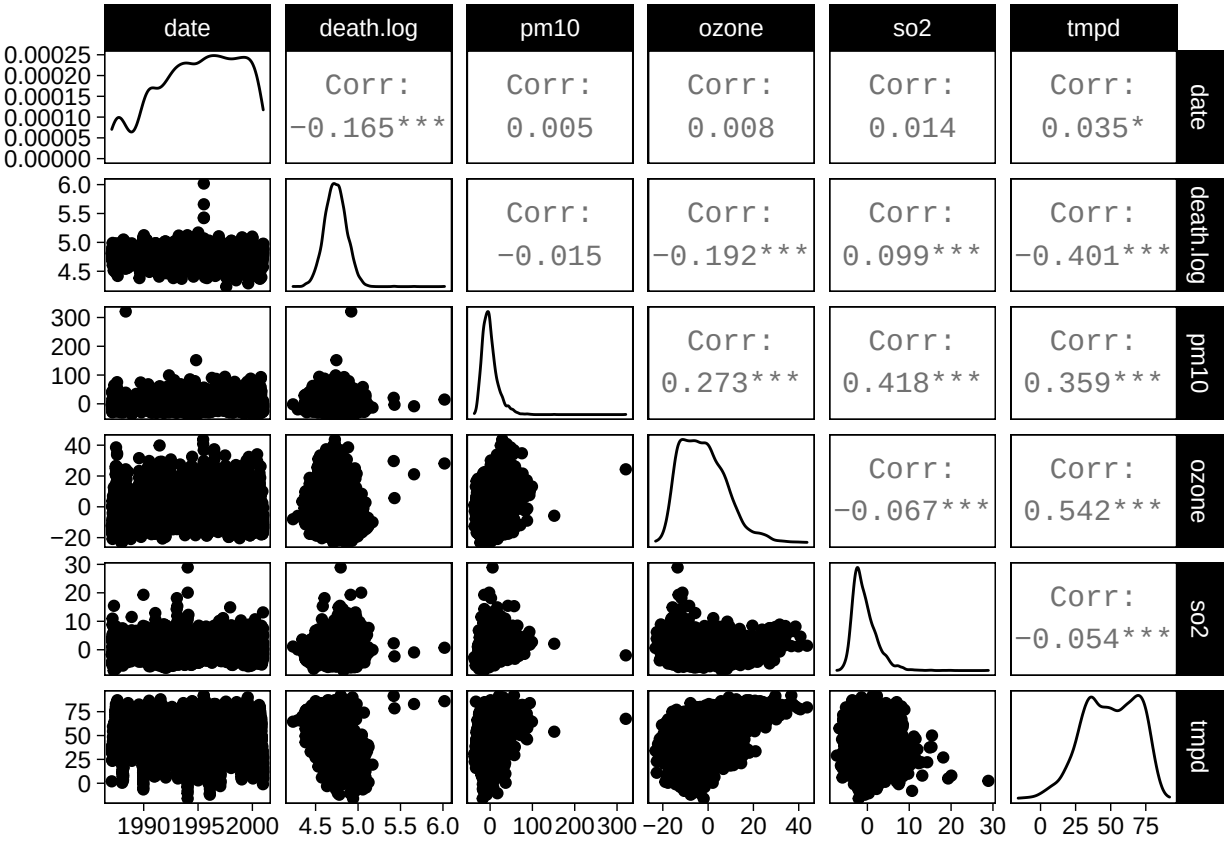

Figure 2: Relationship between response and origianl explanatory variables, presented in the form of scatter plots, boxplots, and correlation coefficients

Similar to figure 2, figure 3 present a matrix of plots demonstrating the relationship between the response variable and the lagged explanatory variables **(4)**. From figure 3, we see that the *death.log* appears to be negatively associated with *lag_pm*10, *lag_ozone*, *lag_tmpd*, and *date* **(5)**. This is verified by the correlation coefficients between *death.log* and these explanatory variables. Also, the magnitude of such correlation coefficients suggests that the relationship between *death.log* and the lagged variables are stronger than that for the original variables **(5)**. Thus, we expect that models using the lagged variables are likely to perform better than the model using the original variables **(5)**. In figure 3, we also see that the relationships between *death* and *lag_tmpd* and *time* to be strong **(5)**. It is noticeable that *death.log* are positively associated with *lag_so*2, which is counter-intuitive. This is likely a result of Simpson's Paradox **(5)**. From figure 2 and 3, we see that the relationships between *death.log* and explanatory variables (both lagged and non-lagged) does not appear linear. This means that we need to incorporate the non-linearity (e.g., use smoothing splines or include polynomial terms) in our model **(5)**.

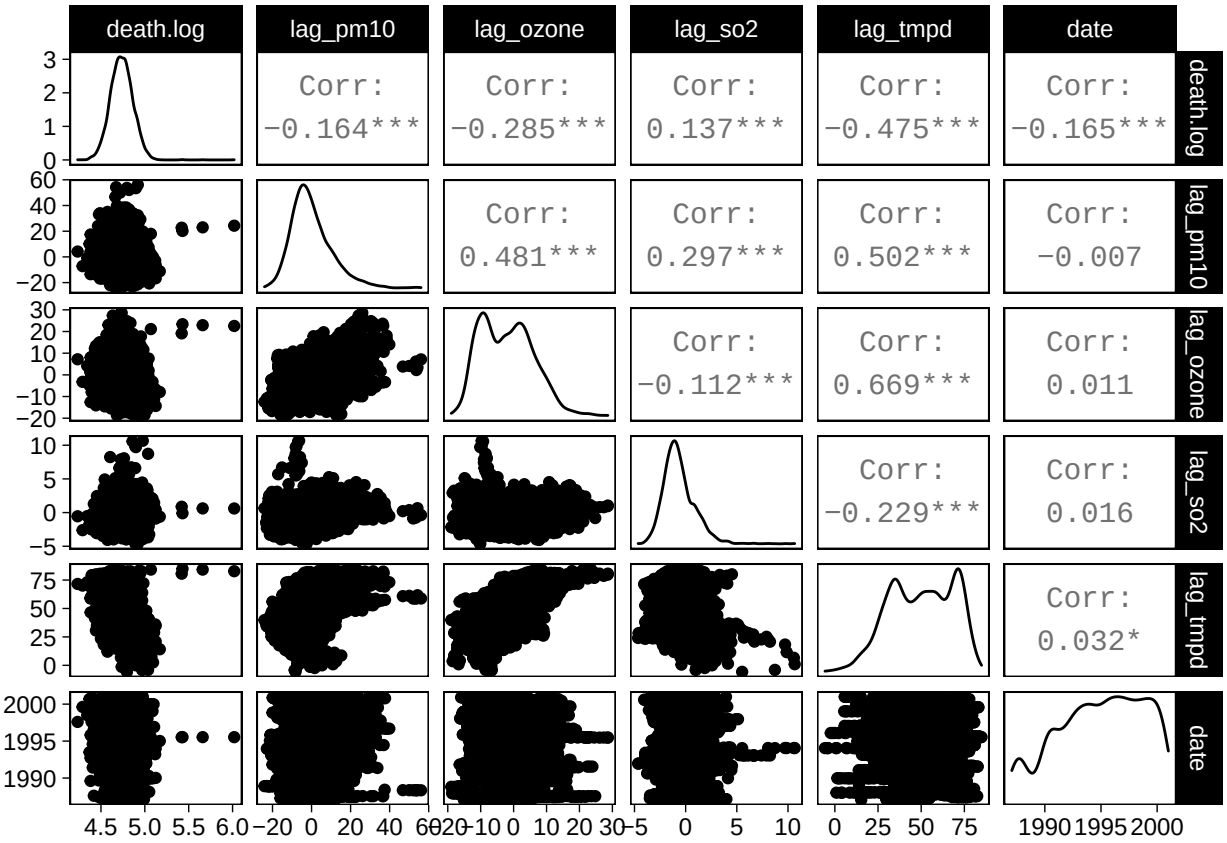

Figure 3: Relationship between response and lagged explanatory variables, presented in the form of scatter plots, boxplots, and correlation coefficients

Modeling & Diagnostics

To answer our research question, we will fit generalized additive models (GAM) to predict *death* from our explanatory variables using the log link from the Poisson family, as we assume *death* follows the Poisson distribution. First, we will build a GAM, called Model 1, to predict *death* from all non-lagged variables, which are *time*, *pm*10, *ozone*, *so*2, and *tmpd* **(1)**. Similar to Model 1, we also build a GAM, called Model 2, to predict *death.log* from all lagged variables, which are *time*, *lag_pm*10, *lag_ozone*, *lag_so*2, and *lag_tmpd* **(1)**. For both Model 1 and 2, we will use smooth spline terms for all individual variables with 4 degrees of freedom (df). In addition to the air pollutants, we also include measures of time and temperature in our model as we see in EDA that these two variables are correlated with the response variable and measures of air pollution **(1)**. We use smooth splines for individual variables as we see in EDA that the relationships between the response and explanatory variables do not appear linear **(1)**.

We will use cross-validation (CV) to select between Model 1 and 2. More specifically, we will use 10-fold cross-validation to estimate the MSE. Table 1 presents a summary of the 10-fold CV error of Model 1 and Model 2, with estimated standard error (SE) **(2)**. From table 1, we see that the CV error for Model 2 (178.4199) is less than the CV error of model 1 (194.1933) **(2)**. However, the difference in 10-fold CV error between Model 1 and Model 2 does not seem to be significant, as the difference in CV error (15.7734) is less than the SE of the estimated MSE (which is around 20-22) **(3)**. Despite Model 2 does not fit significantly better than Model 1, we still choose Model 2 as our final model **(3)**. As Model 1 and 2 have the same complexity ($df = 21$), we choose Model 2 as it performs slightly better.

Model	10-fold CV Error	SD(error)	SE(error)
Model 1	194.1933	69.98423	22.13096
Model 2	178.4199	64.28211	20.32779

Table 1: 10-fold cross validation error for Model 1 and Model 2, with standard deviation (SD) and standard error (SE). $SE = SD/\sqrt{n}$ and $n = 10$, which is the number of folds.

Figure 4 shows the residual diagnostic plots for Model 2. Here the residuals are the deviance residuals. We see from the normal-QQ plot that the residuals do not follow the normal distribution as they significantly deviate from the normal quantiles at the right end. We see from the residual plot that the residuals contain outliers, which means that the residuals are not identically distributed. It is noteworthy that the violations to the model assumptions are caused by very few outliers, which means the violations to model

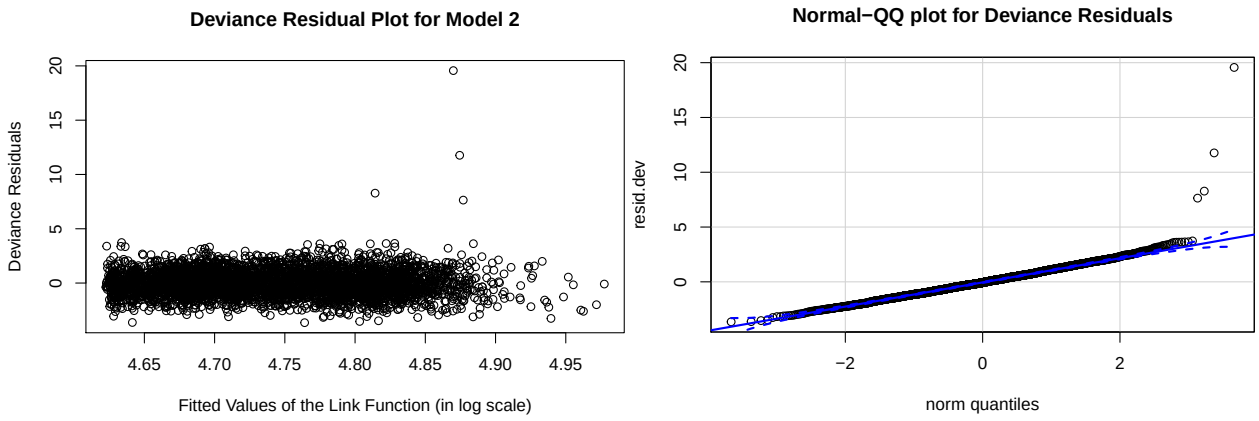

Figure 4: Model disgnostic plots (residual plot and normal-QQ plot for the deviance residuals) for Model 2

assumptions are minor.

Results

We want to assess how well our model fits the data. We conduct 2 global goodness-of-fit tests for our final model. We first test on H_0 : the null model (which only contains the intercept term) is correct versus H_1 : Model 2 is correct. The resulting p-value is less than 2.2×10^{-16}, which is close to 0 and suggests that Model 2 fits significantly better than the null model **(1)**. Similarly, we then test on H_0 : the Model 2 is correct versus H_1 : the saturated model is correct. The p-value is 0, which suggests that we reject H_0. Thus, although we could conclude that Model 2 fits better than the null model, there is significant evidence to reject Model 2 in favor of a saturated model **(1)**. While our model fits reasonably well, there is still room for improvement.

Having proved that our model fits better than the null model, we also want to assess whether pollution is an important factor in our final model. For consistency, we fit a GAM to predict *death* from *lag_tmpd* and *time* using the log-link from the Poisson family, and we call it Model 3 (see *Appendix* note 1). We then test on H_0 : Model 3 is correct versus H_1 : Model 2 is correct. The resulting p-value from the test is less than 2.2×10^{-16}, so we reject H_0 **(2)**. Therefore, we conclude that Model 2 fits better than Model 3, and the associations between mortality and air pollution is jointly significant **(2)**. Table 2 below summarizes the results of the 3 global goodness-of-fit tests.

In the previous section, we see that Model 2, which uses lagged variables, has a smaller estimated MSE than Model 1 that assumes the effects are instantaneous. However, we

H_0 Model	H_1 Model	**P-value**	**Result**
Null Model	Model 2	< 2.2e-16	Reject H_0
Model 2	Saturated Model	≈ 0	Reject H_0
Model 3	Model 2	< 2.2e-16	Reject H_0

Table 2: Summary of 3 global goodness-of-fit tests. H_0 Model and H_1 Model respectively refers to the models in H_0 and H_1.

have argued that the difference in CV error is not significant as it is less than the estimated SE of the test errors. Therefore, we do not have enough evidence to conclude whether the effect of pollution on mortality is instantaneous or propagates over time **(3)**. However, the proposition that the effect of air pollution extends over time is more plausible. Not only does Model 2 produces smaller MSE, but this proposition is more consistent with commonsense **(3)**. It is possible that we fail to see a significant difference because our test is not powerful enough **(3)**. In other words, given a larger sample size, we expect that Model 2 would fit significantly better than Model 1.

On the next page, Figure 5 shows the average relationship between each predictor and the outcome in our final model **(4)**. We see from figure 5 that the slope of the curves is the steepest for the plot of *lag_ozone*. Also, the standard error of the effect seems to be the smallest for the plot of *lag_ozone*. Thus, it appears that *ozone* (more specifically, *lag_ozone*) is most strongly associated with mortality **(4)**. Notably, we see that mortality seems to first decrease then increase as *lag_ozone* increase, which is not expected **(4)**. A possible explanation might be that extremely low levels of ozone at ground level are a signal of insufficient amount of ozone in the stratosphere, which could lead to excessive ultraviolet (which is detrimental to health) at ground level.

From figure 5, we see that decrease in air pollution (more specifically, *ozone* and *pm*10) is associated with decreased mortality in Chicago, in general **(4)**. Surprisingly, a decrease in the SO_2 level appears to be associated with an increase in mortality **(4)**. However, we see in figure 5 that the SE of SO_2's effect seems to be very large, which may indicate that SO_2 may not be a significant predictor of mortality. From the second plot in figure 5, we see that mortality first increases then decreases from 1987 to 2000. From the last plot in figure 5, we increase in temperature is very closely associated with a decrease in mortality **(4)**. This is consistent with the weather condition in Chicago, where extreme cold climates lead to massive death much more often compared to extremely hot weather.

To address Mr.Jorgensen's concern, we want to explore the level of mortality in Chicago

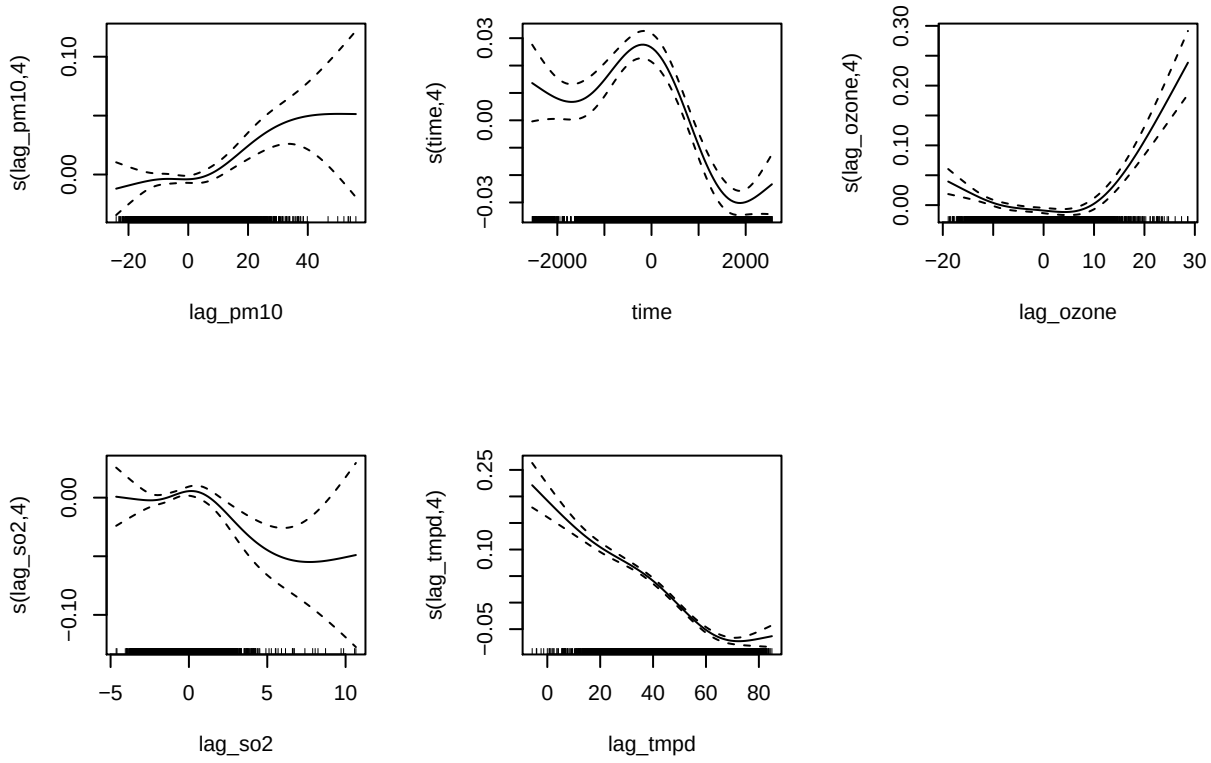

Figure 5: Average relationship between predictor variables and the outcome. Bounds representing standard error are represented as dashed lines

when air pollution is reduced to the lowest level in the historical record. For consistency, we will refer to the lagged air pollution level as the air pollution level. In the *Chicago* dataset, the lowest *lag_ozone* is -18.924, the minimum *lag_so*2 is -4.643, and the minimum *lag_pm*10 is -24.119 **(5)**. Our model controls for time. To promote the usefulness of the prediction, we will use the day 30 December 2000 (which is the latest to present) in our prediction.

Our model predicts that if air pollution is at such a minimum level, and the average temperature is 70 degrees in the past 7 days, the expected *death.log* in Chicago on 30 December 2000 is 4.675179 **(5)**. The SE of fit is 0.01934044. Consequently, a 95% confidence interval (CI) for our prediction for *death.log* is [4.637272, 4.713086]. To get the point prediction and CI for the *death*, we take the exponents of our prediction and CI for *death.log*. Thus, if air pollution is at such a minimum level, and the median temperature is 70 degrees in Chicago, our model predicts that the expected mortality on December 30 2000 is 107.2518 people. A 95% CI for our prediction is [103.2623, 111.3955] people **(5)**.

As we see from figure 4 in the last section, the residuals do not follow the normal distribution, which means we cannot use parametric bootstrap. Also, the residuals are not

identically distributed as it contains outliers. Although these violations result from very few outliers, which means the results may not be fundamentally biased if we use re-sample residuals, we will use non-parametric bootstrap (re-sample cases) for cautious concerns **(6)**. By bootstrapping, we conclude that a 95% pivotal CI for *death.log* in a setting described above is [4.632680, 4.721169] **(6)**. Consequently, A 95% pivotal CI for *death* in a setting described above is [102.7891, 112.2995] **(6)**.

Method	Point Prediction	CI for death.log	CI for death
SE of Fit	107.2518	[4.6373, 4.7131]	[103.262, 111.395]
Bootstrap	107.2518	[4.6327, 4.7212]	[102.789, 112.300]

Table 3: 95% confidence intervals constructed using SE of fit and non-parametric bootstrap. Subject of CI: estimated mortality in Chicago on 30 December 2000, if $lag\_ozone = -18.9239$, $lag\_so2 = -4.643285$, $lag\_pm10 = -24.11919$, and $lag\_tmpd = 70$. *death* refers to mortality and *death.log* refers to the log of mortality

Table 3 presents a summary of the CI calculated manually (under the assumptions of GAM), and the CI calculated using bootstrap **(7)**. We see that the two CI's are similar, and they cover approximately the same region **(7)**. However, it is noticeable that the CI calculated using bootstrap is slightly wider than the CI calculated using the SE of fit **(7)**. This is expected as non-parametric bootstrap relies on fewer assumptions and captures more variance in the data. The fact that the two CI's are highly similar suggests that the outliers we see in figure 4 do not constitute major violations to assumptions of our model **(7)**. Thus, we could conclude our model is correct, and the main assumptions (linearity (correctness), i.id, homoscedasticity, Gaussian error) of our model holds **(7)**.

Conclusions

In conclusion, an increase in air pollution is associated with an increase in mortality in Chicago, even after controlling for the effect of time and temperature **(1)**. More specifically, an increase in the average level of ozone and PM 10 is closely associated with an increase in mortality **(1)**. Our model suggests that an increase in the SO_2 level is associated with a decrease in mortality. However, since the standard error of the estimate of SO_2's effect is very large, we have no evidence to conclude that SO_2 can significantly predict mortality **(1)**. Also, we noticed that very low levels of ozone are associated with an increase in mortality, which is worth further study **(1)**. Among the three kinds of air pollutants, we discovered that the level of ozone is most strongly associated the mortality **(1)**. From our analysis,

we cannot conclude whether the effect of air pollution is instantaneous or it propagates over time **(1)**. However, the latter is more plausible as models using lagged variables yield slightly better performance, and this explanation is more consistent with commonsense.

As for Mr.Preston Jorgensen, we discovered that if the average air pollution in the past 7 days is at a minimal level (in the historical record), and the average median temperature is 70 degrees Fahrenheit, the estimated mortality in Chicago will be about 107.25 people on 30 December 2000, on average **(2)**. This number is lower than the actual mortality in Chicago on 30 December 2000, which is 145 people. Although we discovered that the decrease in air pollution is associated with lower mortality, we cannot conclude whether reducing pollution will lead to a decrease in mortality **(2)**. Unobserved confounding variables (like humidity) may still exist, which may invalidate causal inference drawn from our model. It is well known that high humidity increases air pollution, while high humidity itself is detrimental to health. Thus, the increase in mortality could be attributed to bad health caused by the high humidity instead of air pollution. Also, as we have argued, extremely low levels of pollutants (like ozone) might be associated with an increase in mortality. Thus, reducing the level of air pollution will not necessarily lead to a decrease in mortality.

Our model is not without limitations. A major limitation of our model is that it uses data from 1987 to 2000, which is not up-to-date **(3)**. The underlying societal and biological construct of interest may have changed over 20 years. Thus, our model could have low predictability when applied to a more modern data set. Also, our data only contains data in Chicago, which means that our model may lack external validity (that the result cannot be generalized to other regions) **(3)**. Our data only contains data on 3 kinds of pollutants, which means our model does not capture the effect of other pollutants on mortality **(3)**. As discussed above, our model does not control for all potential confounding variables, which means that casual analysis drawn using our model could be invalid **(3)**. Despite these limitations, our model provides a well-established framework for analyzing the relationship between air pollution and mortality. Further research should use more up-to-date data that contains records on more kinds of air pollutants. In particular, the effect of ozone level on mortality is both intriguing and strong, so ozone should be a major focus of future research.

Appendix and Note

1. The question says to only use temperature as feature. However, since we include *time* in our model, we only want to isolate the effect of air pollution. So we use *tmpd* and *time* in model 3. Also, since we reject model 3in favor of model 2, we would certainly have rejected model 3 if we do not include *time* in it.