Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Chenyin Lin Date

Time-series analyses of the association between mortality and ambient $PM_{2.5}$ concentration

By

Chenyin Lin

MPH

Emory University

Rollins School of Public Health

Department of Biostatistics

_____ [Chair's Signature]

Howard H Chang

_____ [Member's Signature]

Lance A Waller

Time-series analyses of the association between mortality and ambient $PM_{2.5}$ concentration

By

Chenyin Lin

Emory University, 2013

MPH, Emory University

Rollins School of Public Health

2015

Advisor: Howard H Chang.

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2015

Time-series analyses of the association between mortality and ambient $PM_{2.5}$ concentration

By: Chenyin Lin

Abstract:

Estimates of adverse health effect due to outdoor air pollution from epidemiological studies can be used in setting the regulatory standards and help improve public health. The objective of this paper to is to use time-series analysis to examine the association between counts of deaths and ambient $PM_{2.5}$ concentration accounting for confounders including meteorology and long-term and seasonal trends in mortality. Multiple models including Poisson generalized linear models, Bayesian Poisson models, and Bayesian negative binomial models were used to examine the health effects associated with $PM_{2.5}$ concentrations.

We found positive associations between mortality and ambient $PM_{2.5}$ concentrations but none of the estimates from the three models are statistically significant. We also found that the negative binomial model fits the data better compared to a Poisson regression model, suggesting the importance of accounting for over-dispersion in mortality count data.

Keywords:

Air pollution, particulate matter 2.5 ($PM_{2.5}$), Generalized Linear Model, Poisson Regression, Negative Binomial Regression, Time-series Analyses

Time-series analyses of the association between mortality and ambient $PM_{2.5}$ concentration

By

Chenyin Lin

Emory University, 2013

MPH, Emory University

Rollins School of Public Health

2015

Advisor: Howard H Chang

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2015

Acknowledgements

First, I really appreciate my supervisor Howard Chang. Thanks for his patience and instruction in the process of writing this thesis. His advice and support helped me a lot in all steps of the project including paper review, data analyses and paper writing write this thesis. I also appreciate Lance Waller for taking his time to read my thesis.

At last, I want to thank the faculty and staff in Biostatistics Department. They provide a good study environment with diligence and hard work.

Chapter I

Introduction

1. Applications of time-series air pollution and mortality study

Ambient air particulate matter (PM) is formed by particles directly emitted or due to chemical reaction of gases. High levels of particulate matters were found to be associated with many diseases including respiratory, lung related disease, cardiovascular diseases, which may lead to increased mortality (EPA, 2003). $PM_{2.5}$, the particles with diameters below $2.5\mu m$, draw most attention due to their small size and maybe more relevant to adverse health effects than other particulate matter (Schlesinger, 2007) In recent years, studies provide evidence that $PM_{2.5}$ levels relate to potential negative health effects in both epidemiology and toxicology studies (Schlesinger et al, 2003). However, the specific connection between PM and mortality is still debated (Levy et al., 2000).

Early studies on the health effects of PM were conducted when PM levels were extremely high for several days, leading to high mortality and morbidity in a short period (Goldberg et al., 2001). However, modern air pollution researches have focused on investigating the associations when PM concentration was at a much reduced level (Bell et al, 2004). Time-series analysis methods are the most commonly used approach to detect the small short-term health risk associated with day-to-day $PM_{2.5}$ concentration variation, while controlling for confounders (Bell et al., 2004). These potential confounding effects include other pollutants, weather, and seasonal variations in the health outcome (Wyzga, 1978). Since seasonal variations can only be tested over a relatively long period, we need time-series data collected for a long period of time

(Chang et al, 2012). The data used in this paper are from multiple regions where large fractions of the population are at risk. Such Multi-city data have advantage in reducing bias and stabilizing estimates from previous studies (Stieb et al., 2009). The time-series approach has been facilitated by the increasing accessibility of public data sources in many countries. Results from time-series studies have played an important role in improving human health by setting appropriate air quality standards for particulate matter pollution such as the Clean Air Act and the National Ambient Air Quality Standards. (Greenbaum et al., 2001)

2. Poisson Regression

To explore these effects, we use Poisson regression (Kuhn 1994). We begin by defining a series of event counts at a particular time *i* with density given by:

$$g(y_i; \lambda_i) = \frac{e^{-\lambda_i}\lambda_i}{y_i!}$$

where y_i is the observed number of counts for i =1, 2, …, n; λ_i is the mean of the Poisson distribution. The number of events follows a Poisson distribution. The mean λ_i is related to a vector of explanatory variables *Xi* as follows:

$$\log(\lambda_i) = X_i \beta$$

where β is a vector of unknown coefficient to be estimated (Loomis, 2005).

One important and common analytic issue in Poisson model is over-dispersion. Over-dispersion appears when the variance of the fitted Poisson model is larger than the variance of the Poisson model. One reason for over-dispersion can be that one or more important explanatory factors are missing. Another potential reason is that the incorrect distribution is specified (Hastie, 1986). Because of the existence of over-dispersion, it is

important to evaluate the model by comparing the results of Poisson regression to estimates from other models, such as a generalized Poisson linear model, or a negative binomial linear model (Ismail et al., 2007).

3. **Negative binomial Regression**

Negative binomial (NB) regression models have been widely used for regression of discrete count response because they accommodate overdispersion in the count data. Assume the random variable Y follows a negative binomial distribution Y ~ NB (μ, k), the mean and variance of Y satisfy following requirements:

$$E(Y) = \mu,$$

$$\text{Var}(Y) = \mu + k^{-1}\mu^2,$$

where $\mu > 0$, $k > 0$. Here, the over-dispersion is determined by term $1 + \mu/k$, which depends on μ. k is the dispersion parameter. If k equals zero, the mean and variance of Y are equal, and then the distribution is reduced to a Poisson distribution. If $k > 0$, the variance is larger the mean and the distribution allows for over-dispersion (Ismail, 2007). NB log-linear regression models assume the mean count response (μ) relates to explanatory variables through $\log(\mu) = X\beta$, where X is a series of explanatory variables and β is a vector of unknown regression coefficients to be estimate (Mi, 2015).

In this study we investigate the association between $PM_{2.5}$ concentration and mortality using a time-series approach. The modeling process of time-series analyses is conducted in both Bayesian and non-Bayesian frameworks. To address concerns with overdispersion we examined both Poisson and negative binomial regression under a Bayesian framework.

Chapter II

Methods

1. Mortality and Air Pollution Data Descriptions

The dataset contains 1823 observations and 23 variables. Table 1 describes a summary of original and created variables in the study. Mortality data were obtained from the National Center for Health Statistics. The study region consists of five counties in New York City (Bronx, Kings, New York, Queens and Richmond). The dataset includes daily number of deaths from 2001-2005. Only deaths due to cardiovascular and respiratory diseases were included as defined by the International Statistical Classification of Diseases 10^{th} revision. We acquired mean daily temperature and dew point temperature from the National Oceanic Atmospheric Administration's National Climatic Data Center. Daily ambient $PM_{2.5}$ data were downloaded from Statistically Fused Air Quality database (http://www.epa.gov/esd/land-sci/lcb/lcbsfads.html).

From Table 1, we see the mean average ambient $PM_{2.5}$ concentration for previous three days is 15.12 $\mu g/m^3$ with standard deviation is 6.31 $\mu g/m^3$. The mean daily temperature is 55.58 °F with standard deviation 17.38 °F. The average temperature for the previous three days is 55.61 °F with standard deviation 16.97 °F. The mean daily dew-point temperature is 42.56 °F with standard deviation of 18.58 °F. The average dew-point temperature for previous three days is 42.60 °F with standard deviation of 18.57 °F. The mean daily total number of deaths has mean of 143.92 with standard deviation of 17.35.

2. Mortality Model and Risk Estimation

We consider three statistical models for the daily counts of death: Poisson generalized linear models, Bayesian Poisson models and Bayesian negative binomial models. The goal of considering a negative binomial model is to account for any over-dispersion in the outcome.

Poisson Generalized linear Model

Regression models are used in time-series analysis to estimate associations between observed changes in mortality associated with changes in ambient air pollution level on a short-term basis. The mortality model specification follows Chang et al. (2012). We considered the average $PM_{2.5}$ levels for the previous three days as the exposure of interest. Time-series studies are fairly robust against confounding by population characteristics which remain constant over the study period (Bell et al., 2004). However, health outcome and pollution can be affected by confounding from some other factors that vary on shorter time scales. In order to apply time-series analysis to these data, we need to control for seasonal and temperature factors. Potential confounding by these seasonal trends and weather effects could be controlled by assigning natural cubic splines to the following variables: present day temperature (d=6) and average temperature of previous three days (d=6); present day dew-point temperature (d=3) and average dew-point temperature for the previous three days (d=3); calendar date (d=40, 8 per year); and indicators for day of the week. We assume the daily number of deaths y_i follows an over-dispersed Poisson distribution with expected value E $(y_i) = \mu_i$. The regression model is given as follows:

$$y_i \sim \text{Poisson}(\mu_i)$$

$$\ln(\mu_i) = \beta_0 + \beta_1 * avpm25 + \beta_2 * ns(temp, 6)_1 + \cdots + \beta_7 * ns(temp, 6)_6 + \beta_8 * ns(avtemp, 6)_1 + \cdots + \beta_{13} * ns(avtemp, 6)_6 + \beta_{14} * ns(DpTemp, 3)_1 + \cdots + \beta_{16} * ns(DpTemp, 3)_3 + \beta_{17} * ns(avDpTemp, 3)_1 + \cdots + \beta_{19} * ns(avDpTemp, 3)_3 + \beta_{20} * ns(date, 40)_1 + \cdots + \beta_{59} * ns(date, 40)_{40} + \beta_{60} * (dow = Monday) + \cdots + \beta_{66} * (dow = Sunday)$$

where avpm25 denotes average ambient $PM_{2.5}$ concentration for previous three days; temp denotes current day temperature; avtemp denotes average temperature for previous three days; dptemp denotes current day dew-point temperature; avdptemp denotes average dew-point temperature for previous three days; dow is the indicator for day of the week (from Monday-Sunday).

Bayesian Poisson and Negative Binomial Models

For the Bayesian Negative Binomial Model analyses, we assigned the following prior distributions:

$$\beta_i \sim N(0, 0.00001)$$

$$\log(\alpha) \sim N(0, 0.0001)$$

For the Bayesian Poisson analyses, we used the following prior distributions:

$$\beta_i \sim N(0, 0.00001)$$

We used the R Package "R2winBUGS" to run WinBUGS from R to implement both models. We used the coefficients from the Poisson generalized linear model as initial values in the Bayesian analyses. The total number of iterations was 20,000 with a burn-in sample of 10,000. The WinBUGS code is given below:

Bayesian Poisson model

```
model{
for (i in 1:1823){
Y[i]~dpois(mu[i])
mu[i]<-exp(mut[i])
mut[i]<-inprod(beta[1:66], X[i,1:66])}
for(i in 1:66){
beta[i]~dnorm(0, 0.00001)}
}
### Bayesian negative binomial model #####
model{
for (i in 1:1823){
Y[i]~dpois(mu[i])
mu[i]<-exp(rho[i]*mut[i])
mut[i]<-inprod(beta[1:66], X[i,1:66])
rho[i]~dgamma(alpha,alpha)}
for(i in 1:66){
beta[i]~dnorm(0, 0.00001) }
alpha<-exp(logalpha)
logalpha~dnorm(0,0.0001)
}
```

Chapter III

Results

In this study, the comparison between the negative binomial model and the Poisson model is through the deviance information criterion (DIC). The Bayesian negative binomial model has a smaller DIC of 14420.4 compared to that of the Bayesian Poisson model, which is 14427.7. Thus, we find that the negative binomial Bayesian model fits our data better.

The estimate of odds ratio from the Bayesian Poisson model of 1.000766 is similar with the estimate of non-Bayesian generalized linear model of 1.000769. The 95% confidence interval of generalized linear model is (0.9998945, 1.001644). The Poisson Bayesian model has a wider 95% posterior interval, which is (0.999855, 1.001707). The Bayesian negative binomial model gives a smaller estimate for the association of mortality and $PM_{2.5}$ concentration, which is 1.000764. The 95% posterior interval of negative binomial model is between the range of Poisson general linear model and Poisson Bayesian model, which is (0.9998635, 1.001655). However, the 95% interval estimates of the log relative risk from all three models include 1. Thus, in this study, we do not find evidence of association between ambient $PM_{2.5}$ concentration and daily mortality.

Chapter IV

Discussion

We identified a positive association between daily mortality and ambient PM2.5 concentration in the study; however, estimates from all three models (Poisson GLM, Poisson Bayesian, negative binomial Bayesian) were not statistically significant. We found that the DIC was smaller for the negative binomial model, demonstrating that over-dispersion may be present in the mortality data and it is important to account for it.

Several additional analysis on this dataset can be considered. For example, we could consider different pollutant exposures and lag effects, or use the Generalized Poisson distribution which can better capture the tail distribution (Ismail et al., 2007). Future work could also focus on spatial variations in time-series associations to find local areas of highest risk.

References

1. Bell, M. L., Samet, J. M., & Dominici, F. (2004). Time-series studies of particulate matter. *Annu. Rev. Public Health*, *25*, 247-280.

2. Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical science*, 297-310.

3. EPA. 2002. Fourth External Review Draft of Air Quality Criteria for Particulate Matter. EPA/600/P-95/001aF-cF, EPA office of Research and Development. Research Triangle Park. NC

4. Loomis, D., Richardson, D. B., & Elliott, L. (2005). Poisson regression analysis of ungrouped data. *Occupational and environmental medicine*, *62*(5), 325-329.

5. Chang, H. H., Fuentes, M., & Frey, H. C. (2012). Time series analysis of personal exposure to ambient air pollution and mortality using an exposure simulator. *Journal of Exposure Science and Environmental Epidemiology*, *22*(5), 483-488.

6. Ismail, N., & Jemain, A. A. (2007). Handling overdispersion with negative binomial and generalized Poisson regression models. In *Casualty Actuarial Society Forum* (pp. 103-158).

7. Kuhn, L., Davidson, L. L., & Durkin, M. S. (1994). Use of Poisson regression and time series analysis for detecting changes over time in rates of child injury following a prevention program. *American Journal of Epidemiology*, *140*(10), 943-955.

8. Wyzga, R. E. (1978). The effect of air pollution upon mortality: a consideration of distributed lag models. *Journal of the American Statistical Association*, *73*(363), 463-472.

9. Greenbaum, D. S., Bachmann, J. D., Krewski, D., Samet, J. M., White, R., & Wyzga, R. E. (2001). Particulate air pollution standards and morbidity and mortality: case study. *American journal of epidemiology*, *154*(12), 78-90.

10. Schlesinger, R. B. (2007). The health impact of common inorganic components of fine particulate matter (PM2. 5) in ambient air: a critical review. Inhalation toxicology, 19(10), 811-832.

11. Schlesinger, R. B., & Cassee, F. (2003). Atmospheric secondary inorganic particulate matter: the toxicological perspective as a basis for health effects risk assessment. Inhalation toxicology, 15(3), 197-235.

12. Levy, J. I., Hammitt, J. K., & Spengler, J. D. (2000). Estimating the mortality impacts of particulate matter: What can be learned from between-study variability?. Environmental health perspectives, 108(2), 109.

13. Goldberg, M. S., Burnett, R. T., Bailar, J. C., Brook, J., Bonvalot, Y., Tamblyn, R., ... & Valois, M. F. (2001). The association between daily mortality and ambient air particle pollution in Montreal, Quebec: 1. nonaccidental mortality. Environmental Research, 86(1), 12-25.

14. Stieb, D. M., Szyszkowicz, M., Rowe, B. H., & Leech, J. A. (2009). Air pollution and emergency department visits for cardiac and respiratory conditions: a multi-city time-series analysis. Environ Health, 8(25), 10-1186.

Appendix

Table 1. Mean and SD of total mortality, daily $PM_{2.5}$ ambient concentrations ($\mu g/m^3$) and confounders.

Characteristics	Mean $\pm$ SD	Description
Alldeaths	143.92 ± 17.35	Total mortality
Pm25	15.12 ± 8.35	Ambient $PM_{2.5}$ concentration
Avpm25	15.12 ± 6.31	Average ambient $PM_{2.5}$ concentration for previous three days
Temp	55.58 ± 17.38	Present day temperature
Dptemp	42.56 ± 18.58	Present day dew-point Temperature
Avtemp	55.61 ± 16.97	Average temperature for previous three days
Avdptemp	42.60 ± 18.57	Average dew-point temperature previous three days
Dow		Indicator for day of the week (Monday-Sunday)

Table 2. Comparison between different models

Point Estimate	Odds Ratio	95% interval for OR	AIC/DIC
0.0007689698	1.000769	(0.9998945, 1.001644)	14427.0
0.0007658876	1.000766	(0.999855, 1.001707)	14427.7
0.0007632831	1.000764	(0.9998635, 1.001655)	14420.4