Association Between PM2.5 and Lung Cancer Incidence Rates

(United States 1999-2013 Ecological Study with Generalized Linear Mixed Model)

By

Masahiro Nakanishi

M.Sc., Brown University, 2017

Thesis

Submitted in partial fulfillment of the requirements
for the Degree of Master of Science in Integrative Studies
in the Department of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND
MAY 2017

This thesis by Masahiro Nakanishi is accepted in its present form
by the Department of Applied Mathematics Integrative Studies Program
as satisfying the thesis requirements for the Degree of Master of Science

Date	Advisor: Tongzhang Zheng, Ph.D.
Date	Reader: Adam J. Sullivan, Ph.D.
Date	Director: Björn Sandstede, Ph.D.

Approved by the Graduate Council

Date	Dean of Graduate School: Andrew G. Campbell, Ph.D.

AUTHORIZATION TO LEND AND REPRODUCE THE THESIS

As the sole author of this thesis, I authorize Brown University to lend it to other institutions or individuals for the purpose of scholarly research.

Date _____ _____
Masahiro Nakanishi, Author

I further authorize Brown University to reproduce this thesis by photocopying or other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Date _____ _____
Masahiro Nakanishi, Author

Acknowledgement

I would like to take this opportunity to thank my thesis advisor Dr. Tongzhang Zheng for his guidance in cancer epidemiology. I enjoyed learning from the field-leading researcher like himself and his guest lecturers. I would also like to extend my immense gratitude to Dr. Adam J. Sullivan who has been a substantial part of my Brown University education. I met Dr. Sullivan in my Junior Year Fall Semester, and have taken three course taught by him. His passion in education inspires me, and he has been my great teacher, mentor, and a role model. I could not appreciate enough for all the education he has offered me throughout my last two years at Brown University. I would like to extend my tremendous appreciation to Dr. Björn Sandstede, the program director of Applied Mathematics and Integrative Studies. Without his generous support, my Concurrent M.Sc. & B.Sc. Degree could not have happened. I could not thank enough for his kindness. I would also like to thank Dr. Omar Galárraga for his advice and guidance on my research.

Abstract

Introduction:

Cancer is a leading cause of death worldwide, accounting for 8.8 million deaths in 2015, and among which Lung Cancer consists 1.69 million deaths (19.2%). [WHO] Particulate Matter was classified as IARC Group 1 carcinogens in 2013 by the International Agency for Research on Cancer (IARC), specialized cancer agency of the World Health Organization. [Press Release] The public interests in the potential detrimental health effect of air pollution, especially those caused by Particulate Matters (PM2.5, PM10, etc.), are soaring in the recent years. There have been rising number of studies conducted across the globe to report the association between PM2.5 and Lung Cancer Mortality Rates, but very few have successfully studied the association between PM2.5 and Lung Cancer Incidence Rates as suggested by [Vinikoor-Imler et al]. Hence, it is of immense interest to us to assess the association between PM2.5 and Lung Cancer Incidence Rates.

Methods:

Lung Cancer Incidence Rates, Tobacco, and Healthcare datasets obtained from CDC and PM2.5 dataset obtained from EPA were aggregated into one large longitudinal dataset (panel dataset or cross-sectional time-series dataset) with Lung Cancer Incidence

Rates, PM2.5, Tobacco, and Healthcare being the cross-sectional data and Year 1999-2013 (15 years) being the time-series data. Generalized Linear Mixed Model with Poisson Distribution Log Link was extensively employed in this study to estimate the association between PM2.5 and Lung Cancer Incidence Rates.

Results:

Through our analysis with Generalized Linear Mixed Model using Poisson Distribution Log Link, we obtained the Relative Risk of 1.183 (95% CI: 1.181, 1.186) as the association between PM2.5 and Lung Cancer Incidence Rates on State level. Our Modeled-Relative Risk's 95% CI overlapped with 10 out of 14 studies' 95% CIs quoted in the Systematic Review and Meta Analysis published by Harma et al. After obtaining the fixed and random effects of the covariates on Lung Cancer Incidence Rates for each of the 50 States and D.C., we projected the estimates onto the U.S. map to see if there is any geographical difference. Interestingly enough, even after alleviating the negative correlation between the intercepts and the slopes by centering the predictors, we observed a generally higher effect of PM2.5 on the East Coast and a generally higher effect of Tobacco and Healthcare on the West Coast.

TABLE OF CONTENTS

1. Introduction

1.1 Background

Cancer is a leading cause of death worldwide, accounting for 8.8 million deaths in 2015, and among which Lung Cancer consists 1.69 million deaths (19.2%). [WHO] Both tobacco smoking and second hand smoking are classified as exposures to numerous IARC group 1 carcinogens and these effects have been extensively studied. As [Zheng et al] have suggested, however, there are changing patterns of lung cancer incidence by histologic type recently with the decreasing trend in smoking prevalence. Squamous cell carcinoma (highly associated with tobacco smoking) incidence has been falling notably while the incidence of adenocarcinoma has been rising. With the decreasing proportion of lung cancer attributable to tobacco smoking, other factors are becoming relatively more important. Air pollution is one of them and in fact Particulate Matter was classified as IARC Group 1 carcinogens in 2013 by the International Agency for Research on Cancer (IARC), specialized cancer agency of the World Health Organization. [Press Release] The public interests in the potential repercussions of air pollution, especially those caused by Particulate Matters (PM2.5, PM10, etc.), are soaring in the recent years. There have been rising number of studies conducted across the globe to report the association between PM2.5 and Lung Cancer Mortality Rates, but very few have successfully studied the association between PM2.5 and Lung Cancer Incidence Rates as suggested by [Vinikoor-Imler et al].

1.2 Purpose

In this paper, we aim to employ Generalized Linear Mixed Model with Poisson Distribution Log Link to obtain the association between PM2.5 and Lung Cancer Incidence Rates on State level. We are aware of the potential repercussion of the ecological fallacy demonstrated by [Robinson] over half a century ago, and by no means do we intend to make a conclusion on the individual level. As [Curran et al] suggested, "Ecological fallacy only applies when an aggregate relation is misattributed to the level of the individual". That being said, however, there is a salient public health implication in studying the association between PM2.5 and Lung Cancer Incidence Rates on an ecological level. While there are only limited protective measures that individuals could take against PM2.5, since exposure levels to PM2.5 is highly dependent upon the air quality, the decisions that policy makers and business leaders render today will affect the ecology and thus the health of the people in the future.

It is unfortunate that we did not have access to individual level dataset regarding the Lung Cancer Incidence Rates and PM2.5 exposures, but these limitations do not preclude us from conducting this research, as we believe that it would provide valuable insights. Eventually, our result obtained on the State level will be compared with the meta-relative risk deduced on the individual level by [Harma et al.]. If the relative risks from the State level and the individual level reside within a consistent range, this research could potentially shed light on the high correlation between the State level and the individual level associations between PM2.5 and Lung Cancer Incidence Rates. Thus, this could potentially encourage many more researchers who have access to both individual and ecological level datasets to conduct Multilevel Analysis.

2. Methods

2.1 Data Collection

The Ecological Datasets for Lung Cancer Incidence Rates, Tobacco, and Healthcare were collected from CDC (Centers for Disease Control and Prevention), and these are all provided in yearly basis for 50 States and D.C. The Ecological Dataset for PM2.5 was collected from EPA (Environmental Protection Agency) and it was provided in daily basis for 50 States and D.C. Therefore, the simple means were taken for each of the states to aggregate the data into yearly basis. After these data manipulation, we eventually merged these respective datasets into one large longitudinal dataset (panel dataset or cross-sectional time-series dataset) for 50 States and D.C. from 1999 to 2013 in order to conduct the Generalized Linear Mixed Model analysis. All of these datasets met the high-quality data criteria stipulated by the respective institutions.

2.2 Variable Description

Table 1. Variable Names

	Variable	Description
Lung Cancer Incidence Rates	Dependent	Age Adjusted Rate (/100,000)
PM2.5	Independent	PM2.5 concentrations (mcg/m3)
Tobacco	Independent	Current Smoker Rate (%)
Healthcare	Independent	Healthcare Coverage (%)

Table 2. Descriptive Table

	Mean	SD	Min	Med	Max
Lung Cancer (All.Races)	67.48	12.1	25.6	68.5	103.6
Lung Cancer (White)	67.46	12.62	25	68.9	104.3
Lung Cancer (Black)	74.49	14.96	40.2	73.6	122.1
Lung Cancer (Hispanic)	41.52	12.87	17.9	38.8	136.9
PM2.5	10.49	2.69	3.2	10.13	19.73
Tobacco	20.84	3.68	9.1	20.8	32.6
Healthcare	85.2	4.43	69.4	85.4	95.7

2.3 Exploratory Figures

As we could note from the exploratory figures below, there are nontrivial variations across the States for each of the variables. This provides us the motivation to exploit the Generalized Linear Mixed Model as could model both within State variation and between States variation altogether.

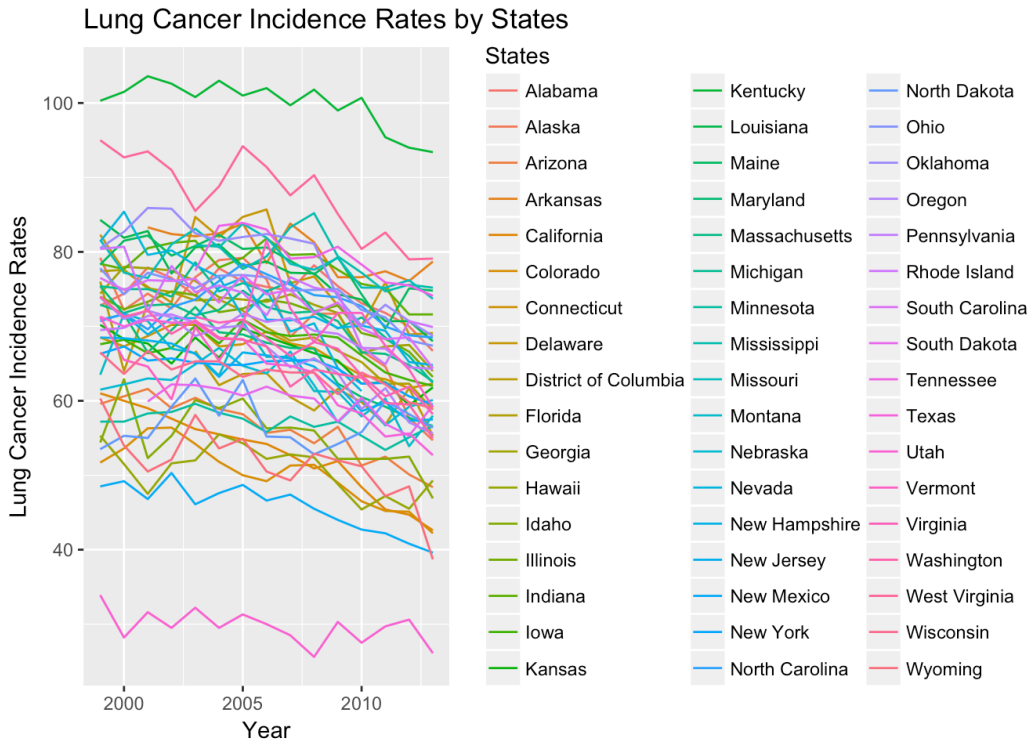

Figure 1. Lung Cancer Incidence Rates Line Graph

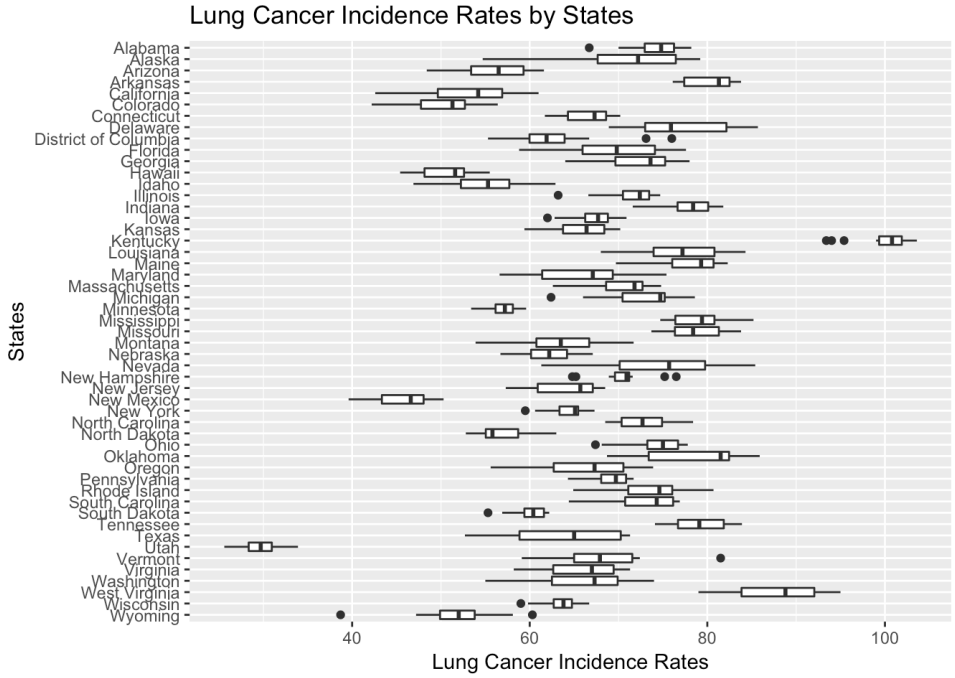

Figure 2. Lung Cancer Incidence Rates Box Plot

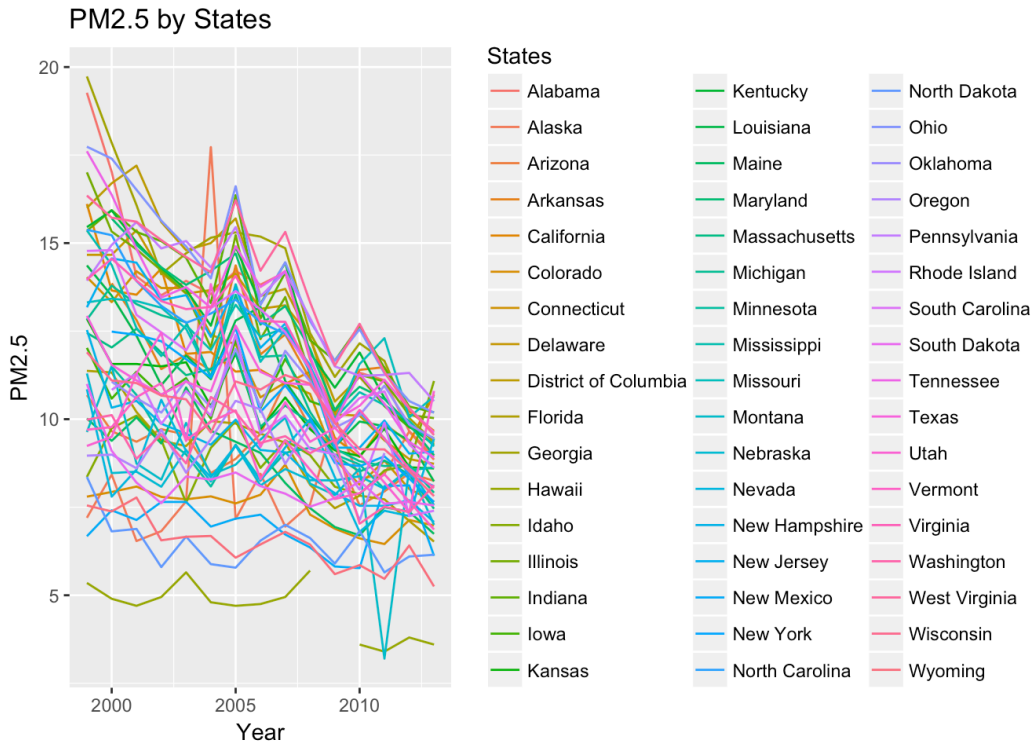

Figure 3. PM2.5 Line Graph

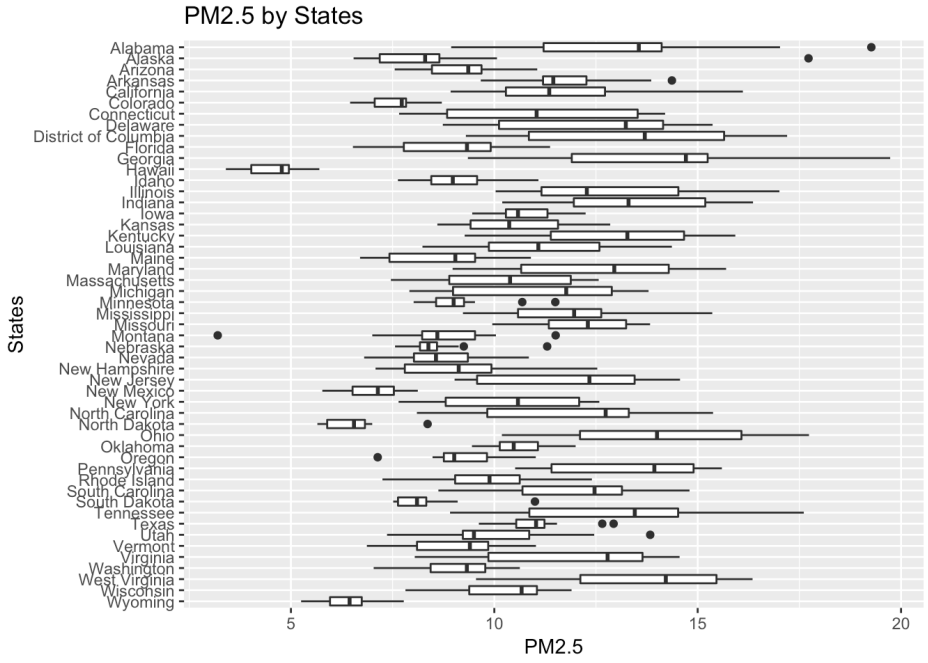

Figure 4. PM2.5 Box Plot

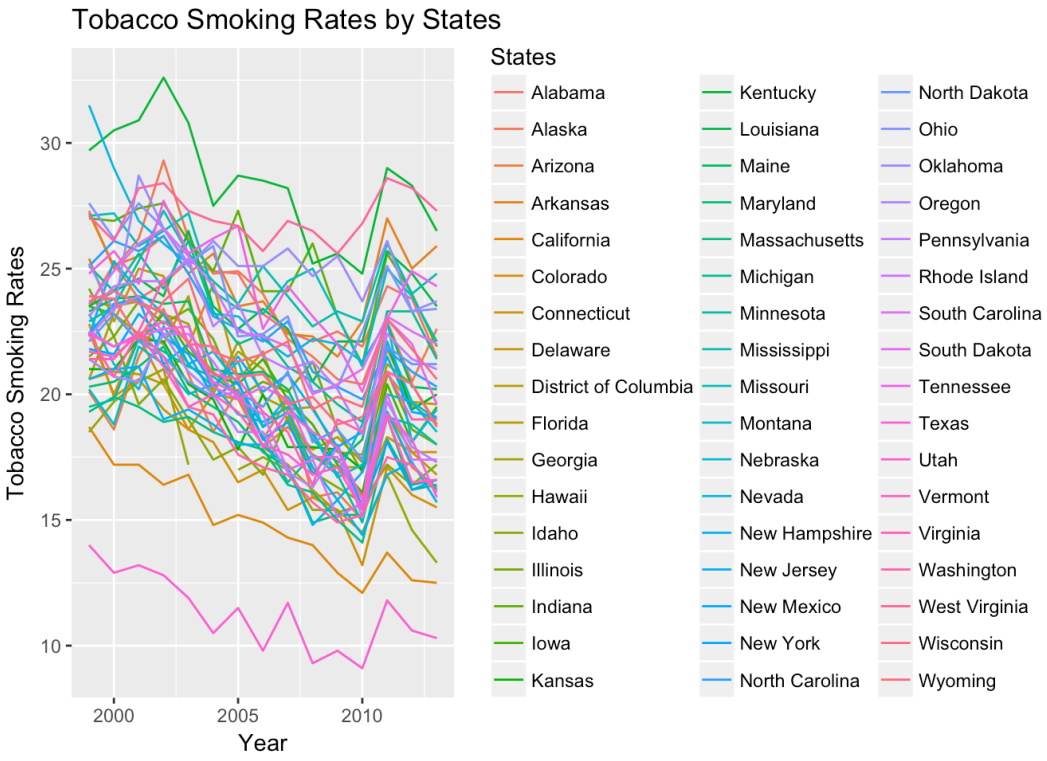

Figure 5. Tobacco Line Graph

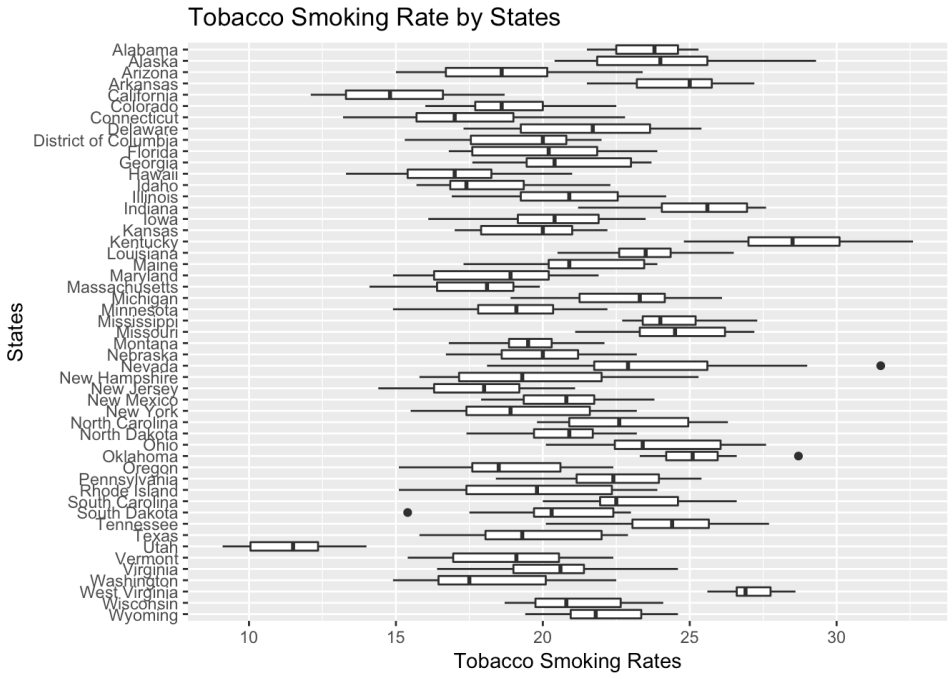

Figure 6. Tobacco Box Plot

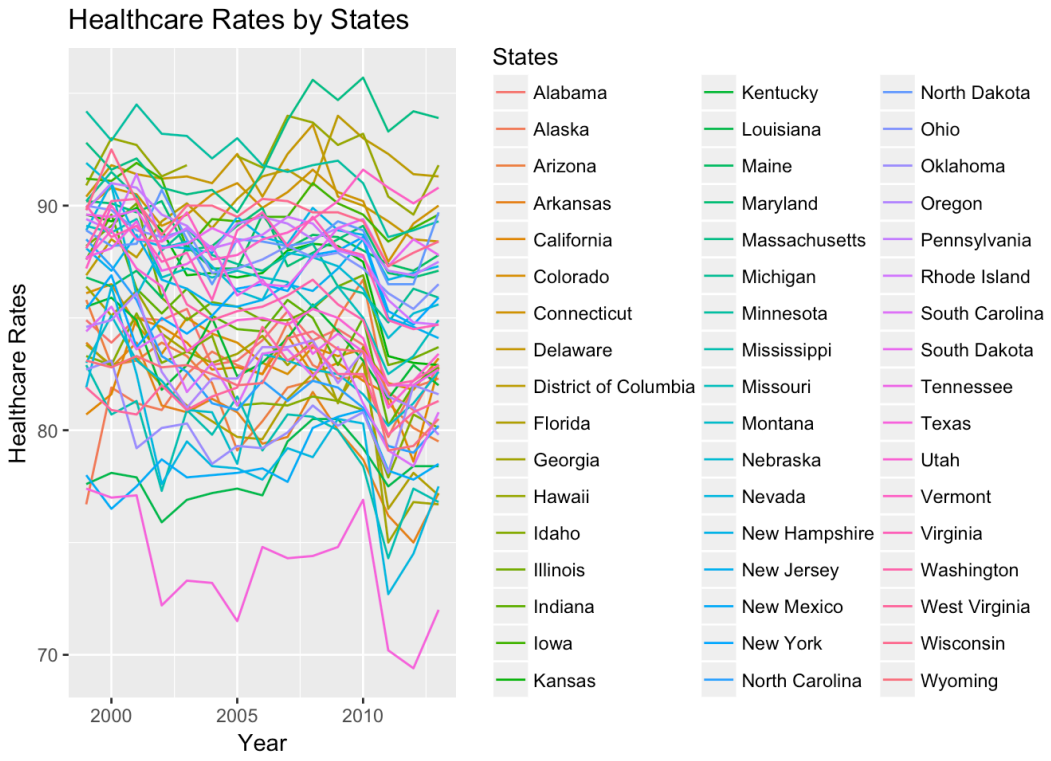

Figure 7. Healthcare Line Graph

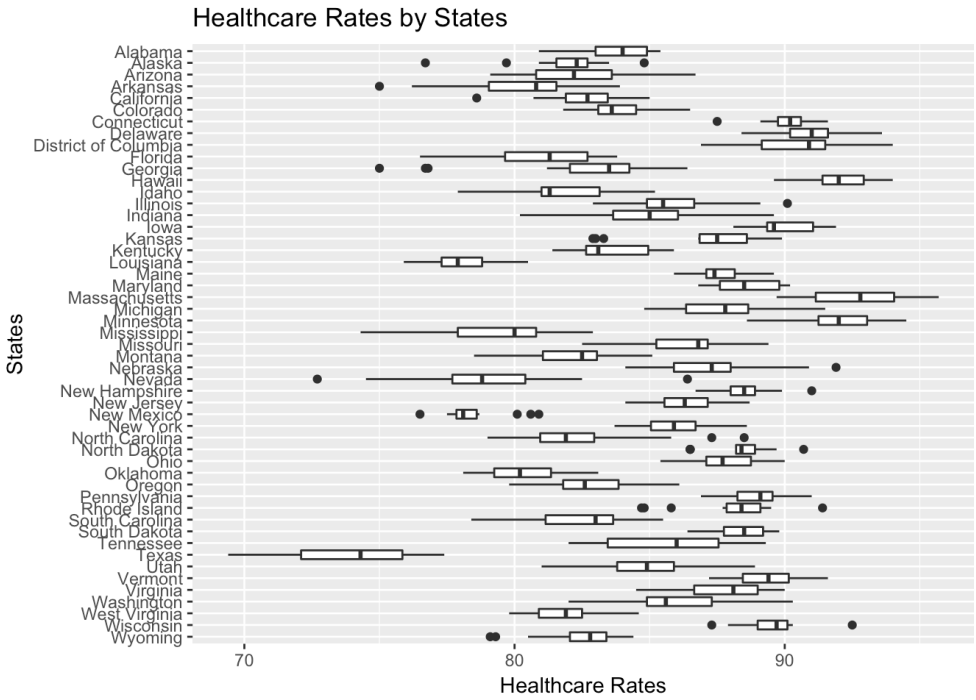

Figure 8. Healthcare Box Plot

2.4 Study Design

With the Multilevel Modeling mindset, let us define Level 1 as the individual level, Level 2 as the State level, and Level 3 as the Country level. Although we could not circumvent the ecological fallacy on the Level 1 / Level 2 interface, since we do not have individual level dataset, we could bypass the ecological fallacy on the Level 2 / Level 3 interface with the use of Generalized Linear Mixed Model to account for both within State variation and between States variation. Generalized Linear Mixed Model could be considered as the combination of the Generalized Linear Model (GLM) and the Linear Mixed-Effects Model (LMM). Linear Mixed-Effects Model is referenced as various names across different fields (Multilevel Models, Hierarchical Linear Models, Nested Data Models, etc.)

Poisson Distribution Log Link would be an appropriated Generalized Linear Model in this study since our dependent variable is a rate (Lung Cancer Incidence Rates) variable. Linear Mixed-Effects Model is also necessary since we would like to account for both the within State and between State variations. These necessities lead us to the employment of the Generalized Linear Mixed Model with Poisson Distribution Log Link.

Another critical advantage of using the Mixed Linear Effects Regression Model over the general Panel Data Regression Model is that the former does not require Balanced Dataset (All variables are observed for every single one of the cross-sectional variable and the time variable) while the latter does (or prefers to a much greater extent). This is because in the Mixed Linear Effects Regression Model, we consider the observations at each time point to be nested within the individual subject, in this case the 50 States + D.C., and there is no strong dependency on the Balanced Dataset condition.

2.5 Model Specifications

2.5.1 Notations

The following notations are used throughout this paper.

x	:	arbitrary variable
i	:	state
t	:	year
N	:	Number of states in US
T	:	Number of years
γ	:	fixed effect
u	:	random effect
PM	:	PM2.5
TO	:	Tobacco
HC	:	Healthcare
$White$	:	White
$Black$	:	Black
$Hispanic$	:	Hispanic

For an arbitrary variable x_{it}, $\bar{x}_{it}$ denotes its grand mean and x_{it}^{*} the variable demeaned by its grand mean.

$$\bar{x}_{it} = \frac{1}{N \cdot T} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it} \tag{1}$$

$$x_{it}^{*} = x_{it} - \bar{x}_{it} \tag{2}$$

We propose the following eight statistical models to be examined in this paper. As mentioned in the Study Design section, Generalized Linear Mixed Model with Poisson

Distribution Log Link will be extensively used in our analysis. These models will be evaluated in the Model Selection Section of this Methods Chapter.

The Rate_{it} denoted here is usually referred to as a count variable in other literatures, but we will continue to use this "Rate" notation since it was denoted as Incidence "Rate" per 100,000 person-years in the data source. Also note that these "Rate" are represented in decimal to the tenth since these were calculated by the number of new cases diagnosed over population of the State multiplied by$100{,}000 = 10^5$. As we know, Poisson Regression only takes in integers and thus decimal inputs could pose an issue. In order to circumvent this situation, we multiplied all the "Rate" by 10 and denominated by $1{,}000{,}000 = 10^6$ person-years. With this adjustment, offset ( ) command was also included in the glmer( ) function to run the Generalized Linear Mixed Effects Regression (GLMER) with Poisson Distribution Log Link.

2.5.2 Mean Centering

This mean centering procedure was conducted not only to make the interpretation of intercept to be more informative, but also to alleviate the correlation between the intercept estimates and the other covariate coefficient estimates. As we could observe from the before and after figures below, by centering the variables at their respective grand mean values, we have successfully removed most of the negative correlation between the intercept estimates and the other covariate coefficient estimates, especially for Tobacco and Healthcare. This concept will be revisited again in the Results Chapter Estimates Mapping Section.

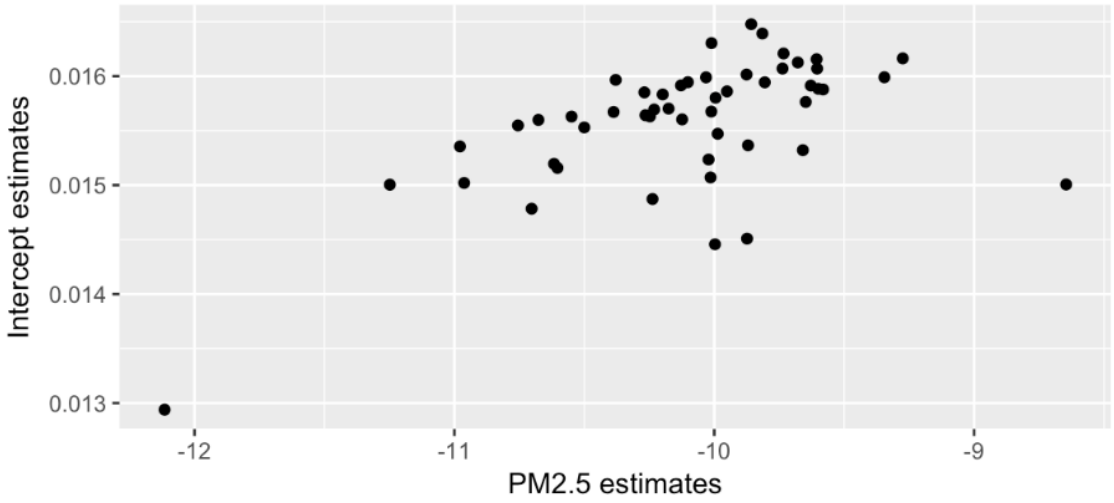

Figure 9. Before Mean Centering PM2.5 (The horizontal axis is $\ln(\text{Rate}/10^5)$.)

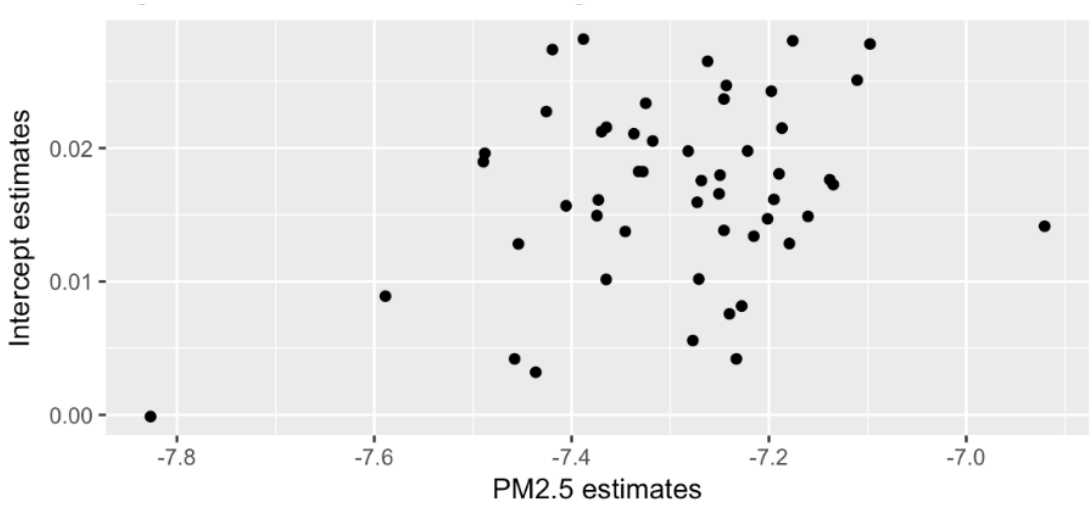

Figure 10. After Mean Centering PM2.5 (The horizontal axis is $\ln(\text{Rate}/10^5)$.)

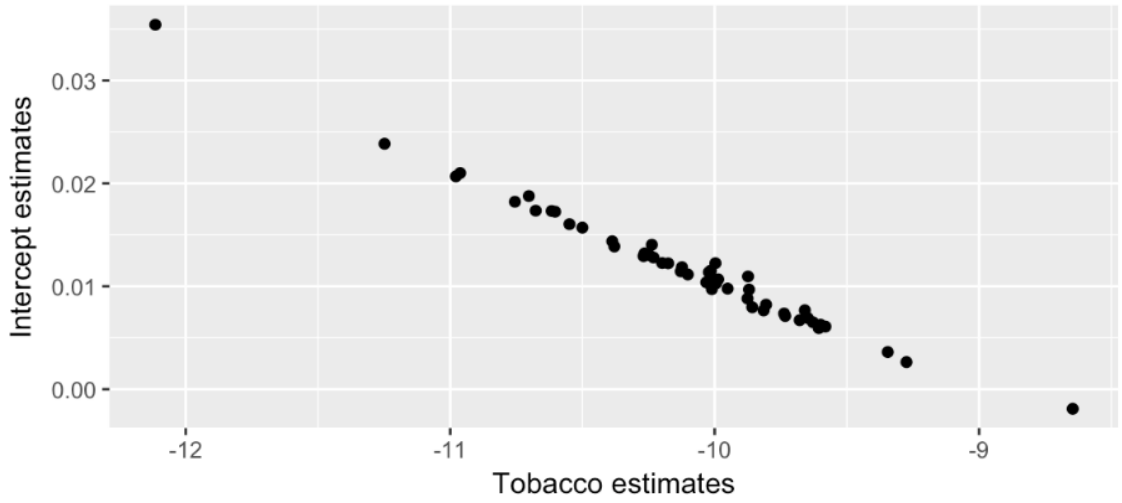

Figure 11. Before Mean Centering Tobacco (The horizontal axis is $\ln(\text{Rate}/10^5)$.)

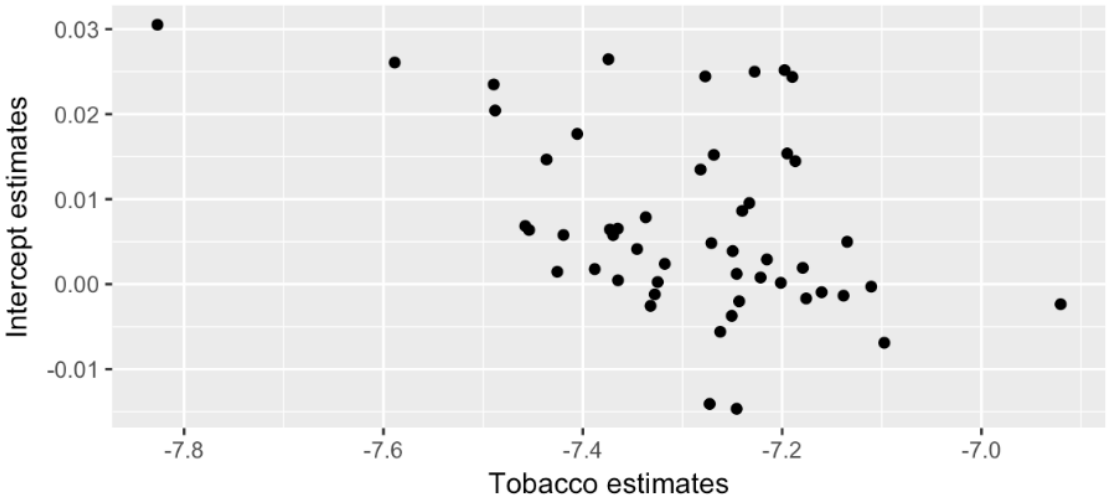

Figure 12. After Mean Centering Tobacco (The horizontal axis is $\ln(\text{Rate}/10^5)$.)

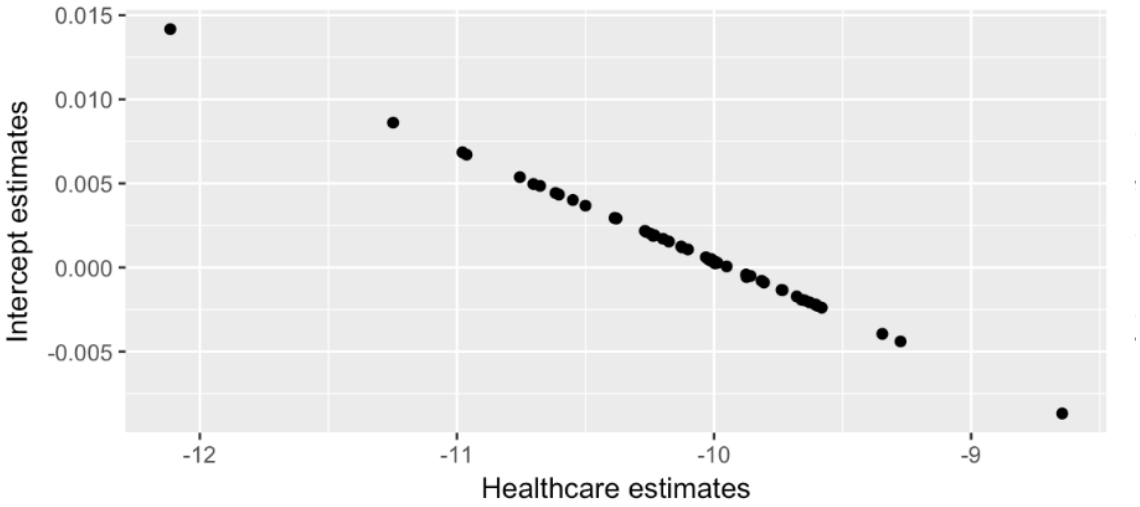

Figure 13. Before Mean Centering Healthcare (The horizontal axis is $\ln(\text{Rate}/10^5)$.)

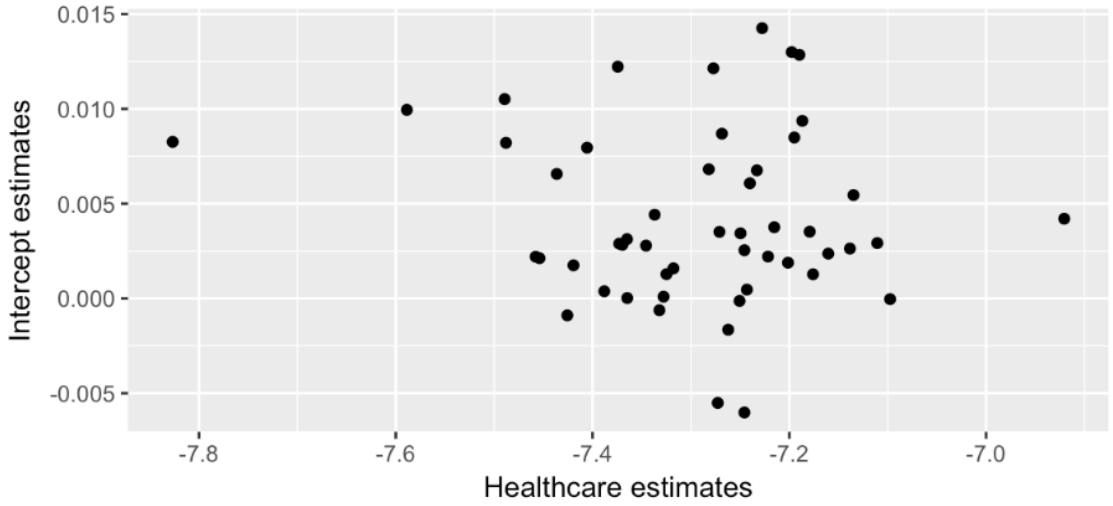

Figure 14. After Mean Centering Healthcare (The horizontal axis is $\ln(\text{Rate}/10^5)$.)

2.5.3 Model 1~4

In the model description below, the system of equations, the single line equations $Rate_{it} = \exp(X\gamma + Zu + \epsilon_{it})$, as well as the glmer( ) function from R lme4 package are presented for the respective Generalized Linear Mixed Model 1~4.

Model 1: Varying Intercept Effect

$$\begin{cases} \ln \dfrac{Rate_{it}}{10^5} = \beta_{0i} + \beta_1 PM^*_{it} + \varepsilon_{it} \\ \beta_{0i} = \gamma_{00} + u_{0i} \\ \beta_1 = \gamma_{10} \end{cases} \tag{3}$$

Consequently,

$$Rate_{it} = \exp\left\{\left(\gamma_{00} + \gamma_{10} PM^*_{it} + \ln\left(10^5\right)\right) + \left(u_{0i}\right) + \varepsilon_{it}\right\} \tag{4}$$

The corresponding glmer( ) function is

$$\text{glmer(Rate}\tilde{}\text{1+PM2.5+offset(logT6)+(1|State), family=poisson)} \tag{5}$$

Model 2: Varying Intercept and PM2.5 Effects

$$\begin{cases} \ln \dfrac{Rate_{it}}{10^5} = \beta_{0i} + \beta_{1i} PM^*_{it} + \varepsilon_{it} \\ \beta_{0i} = \gamma_{00} + u_{0i} \\ \beta_{1i} = \gamma_{10} + u_{1i} \end{cases} \tag{6}$$

$$Rate_{it} = \exp\left\{\left(\gamma_{00} + \gamma_{10} PM^*_{it} + \ln\left(10^5\right)\right) + \left(u_{0i} + u_{1i} PM^*_{it}\right) + \varepsilon_{it}\right\} \tag{7}$$

The corresponding glmer( ) function is

$$\begin{aligned}&\text{glmer(Rate}\sim\text{1+PM2.5+offset(logT6)}\\&\quad\text{+(1+PM2.5|State),family=poisson)}\end{aligned} \tag{8}$$

Model 3: Varying Intercept, PM2.5, and Tobacco Effects

$$\begin{cases}\ln\dfrac{Rate_{it}}{10^5}=\beta_{0i}+\beta_{1i}\,PM^*_{it}+\beta_{2i}\,TO^*_{it}+\varepsilon_{it}\\ \beta_{ki}=\gamma_{k0}+u_{ki}\ ;\ k=0,1,2\end{cases} \tag{9}$$

$$\begin{aligned}Rate_{it}=\exp\Big\{&\left(\gamma_{00}+\gamma_{10}\,PM^*_{it}+\gamma_{20}\,TO^*_{it}+\ln\left(10^5\right)\right)\\&+\left(u_{0i}+u_{1i}\,PM^*_{it}+u_{2i}\,TO^*_{it}\right)+\varepsilon_{it}\ \Big\}\end{aligned} \tag{10}$$

The corresponding glmer( ) function is

$$\begin{aligned}&\text{glmer(Rate}\sim\text{1+PM2.5+Tobacco+offset(logT6)}\\&\quad\text{+(1+PM2.5+Tobacco|State),family=poisson)}\end{aligned} \tag{11}$$

Model 4: Varying Intercept, PM2.5, Tobacco, and Healthcare Effects

$$\begin{cases}\ln\dfrac{Rate_{it}}{10^5}=\beta_{0i}+\beta_{1i}\,PM^*_{it}+\beta_{2i}\,TO^*_{it}+\beta_{3i}\,HC^*_{it}+\varepsilon_{it}\\ \beta_{ki}=\gamma_{k0}+u_{ki}\ ;\ k=0,1,2,3\end{cases} \tag{12}$$

$$\begin{aligned}Rate_{it}=\exp\Big\{&\left(\gamma_{00}+\gamma_{10}\,PM^*_{it}+\gamma_{20}\,TO^*_{it}+\gamma_{30}\,HC^*_{it}+\ln\left(10^5\right)\right)\\&+\left(u_{0i}+u_{1i}\,PM^*_{it}+u_{2i}\,TO^*_{it}+u_{3i}\,HC^*_{it}\right)+\varepsilon_{it}\ \Big\}\end{aligned} \tag{13}$$

The corresponding glmer( ) function is

$$\begin{aligned}&\text{glmer(Rate}\sim\text{1+PM2.5+Tobacco+Healthcare+offset(logT6)}\\&\quad\text{+(1+PM2.5+Tobacco+Healthcare|State),family=poisson)}\end{aligned} \tag{14}$$

2.5.4 Model 5~8

In the Generalized Linear Mixed Model 5~8, Race variable was added as a fixed effect to interact with other previously discussed main effects Intercept, PM2.5, Tobacco, and Healthcare. The Race data was available in the same dataset collected from CDC, and there were three types of races White, Black, and Hispanic. In our study, White was set to be the baseline race variable, and thus the coefficient estimates for Black and Hispanic were referenced with respect to White. There were considerably more missing values for Black and Hispanic compared to White, but this would not pose a problem thanks to the robustness of the Mixed-Effects Model against missing values. In the model description below, only the system of equations is shown and the single line equations are eliminated for the purpose of conciseness.

Model 5: Varying Intercept Effect with Race

$$\begin{cases} \ln \dfrac{Rate_{it}}{10^5} = \beta_{0i} + \beta_1 PM_{it}^{*} + \varepsilon_{it} \\ \beta_{0i} = \gamma_{00} + \gamma_{01} Black + \gamma_{02} Hispansic + u_{0i} \\ \beta_1 = \gamma_{10} + \gamma_{11} Black + \gamma_{12} Hispansic \end{cases} \tag{15}$$

The corresponding glmer( ) function is

$$\text{glmer(Rate}\tilde{}\text{Race*(1)+offset(logT6)+(1|State),family=poisson)} \tag{16}$$

Model 6: Varying Intercept and PM2.5 Effects with Race

$$\begin{cases} \ln \dfrac{Rate_{it}}{10^5} = \beta_{0i} + \beta_{1i}\, PM^{*}_{it} + \varepsilon_{it} \\ \beta_{0i} = \gamma_{00} + \gamma_{01}\, Black + \gamma_{02}\, Hispansic + u_{0i} \\ \beta_{1i} = \gamma_{10} + \gamma_{11}\, Black + \gamma_{12}\, Hispansic + u_{1i} \end{cases} \tag{17}$$

The corresponding glmer( ) function is

$$\text{glmer(Rate}\tilde{}\text{Race*(1+PM2.5)+offset(logT6)+(1+PM2.5|State),family=poisson)} \tag{18}$$

Model 7: Varying Intercept, PM2.5, and Tobacco Effects with Race

$$\begin{cases} \ln \dfrac{Rate_{it}}{10^5} = \beta_{0i} + \beta_{1i}\, PM^{*}_{it} + \beta_{2i}\, TO^{*}_{it} + \varepsilon_{it} \\ \beta_{ki} = \gamma_{k0} + \gamma_{k1}\, Black + \gamma_{k2}\, Hispansic + u_{ki} \;\; ; k = 0,1,2 \end{cases} \tag{19}$$

The corresponding glmer( ) function is

$$\begin{gathered} \text{glmer(Rate}\tilde{}\text{Race*(1+PM2.5+Tobacco)+offset(logT6)} \\ \text{+(1+PM2.5+Tobacco|State),family=poisson)} \end{gathered} \tag{20}$$

Model 8: Varying Intercept, PM2.5, Tobacco, and Healthcare Effects with Race

$$\begin{cases} \ln \dfrac{Rate_{it}}{10^5} = \beta_{0i} + \beta_{1i}\, PM^{*}_{it} + \beta_{2i}\, TO^{*}_{it} + \beta_{3i}\, HC^{*}_{it} + \varepsilon_{it} \\ \beta_{ki} = \gamma_{k0} + \gamma_{k1}\, Black + \gamma_{k2}\, Hispansic + u_{ki} \;\; ; k = 0,1,2,3 \end{cases} \tag{21}$$

The corresponding glmer( ) en is

$$\begin{gathered} \text{glmer(Rate}\tilde{}\text{Race*(1+PM2.5+Tobacco+Healthcare)+offset(logT6)} \\ \text{+(1+PM2.5+Tobacco+Healthcare|State),family=poisson)} \end{gathered} \tag{22}$$

2.6 Model Selections

2.6.1 Diagnostic Plots

As a process of model selections and diagnostics, we first present the diagnostic plots here to assess the validity of each model. For each of the models, the top right figure represents the Residual vs Fitted Plot, the top left figure represents the Scale-Location Plot, the bottom right figure represents the Residual vs Year Plot, and lastly the bottom right figure represents the Normal Q-Q Plot.

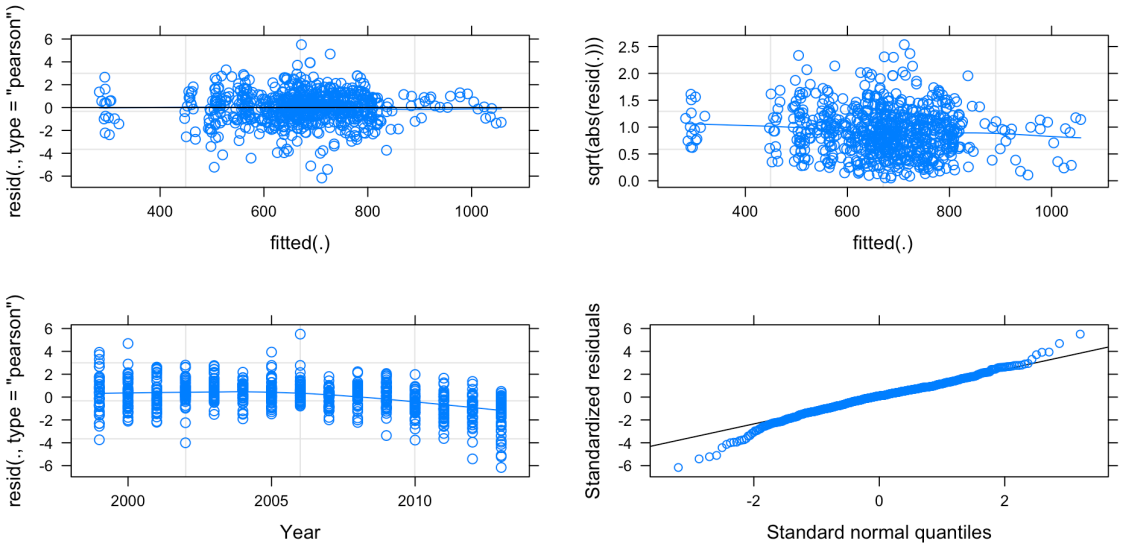

Figure 15. Model 1: Diagnostic Plots

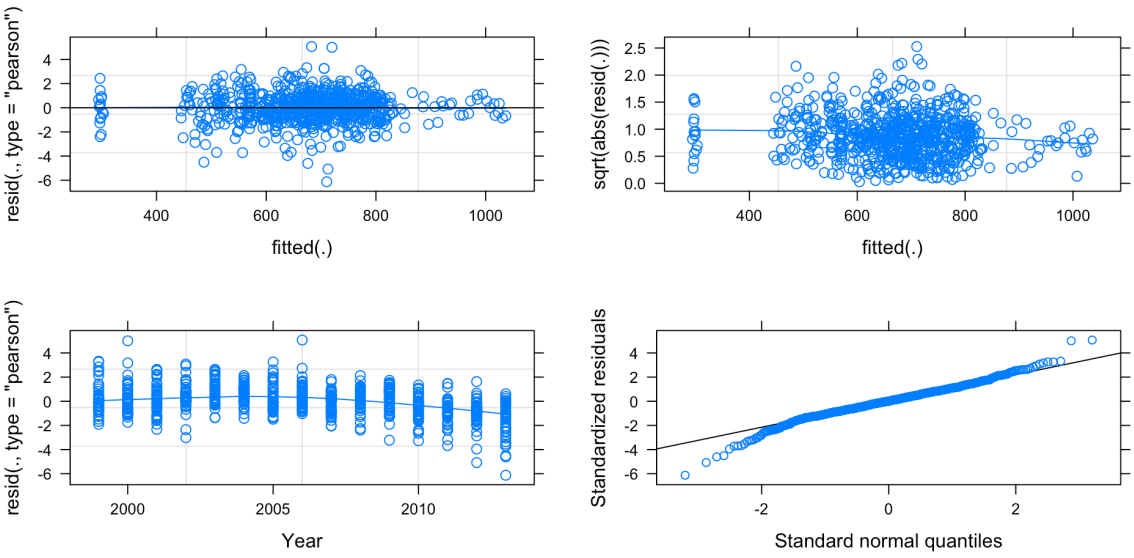

Figure 16. Model 2: Diagnostic Plot

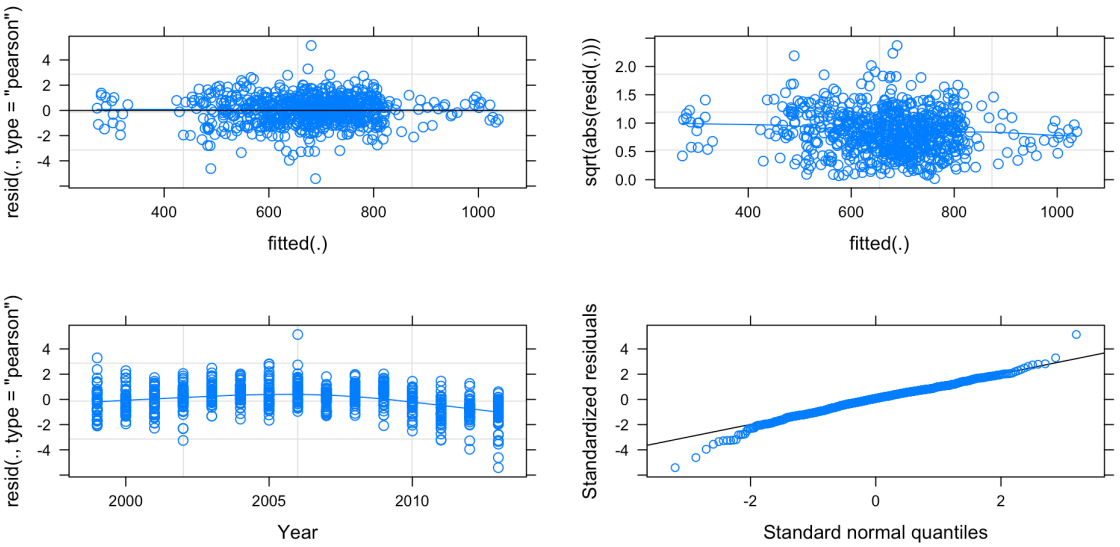

Figure 17. Model 3: Diagnostic Plots

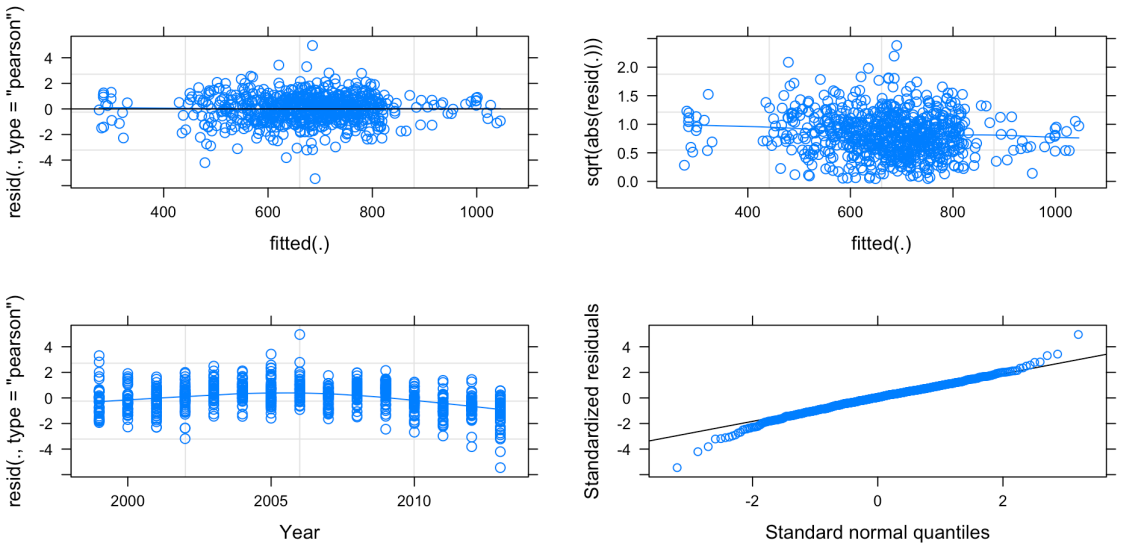

Figure 18. Model 4: Diagnostic Plots

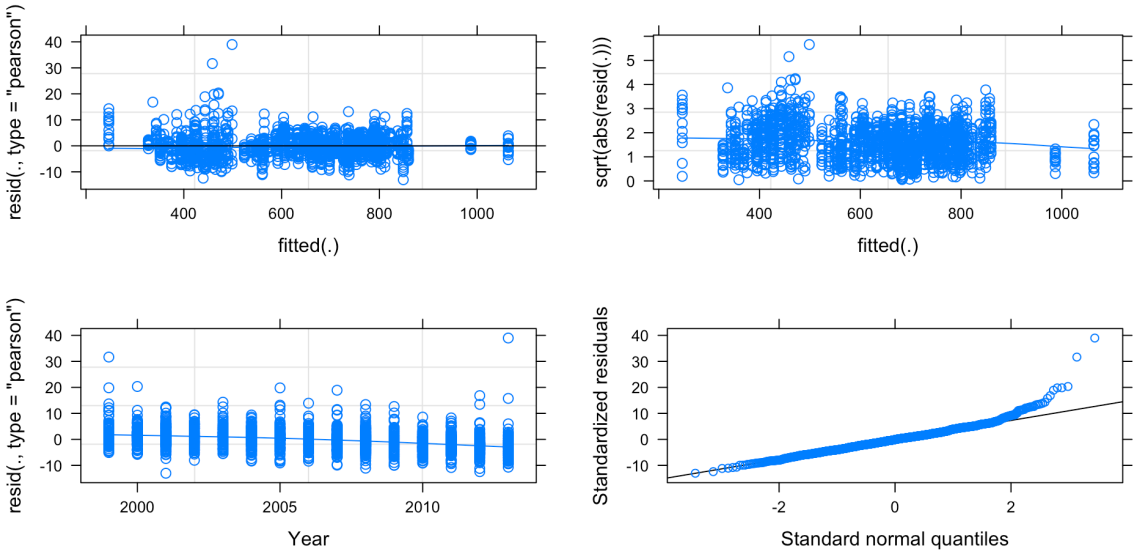

Figure 19. Model 5: Diagnostic Plots

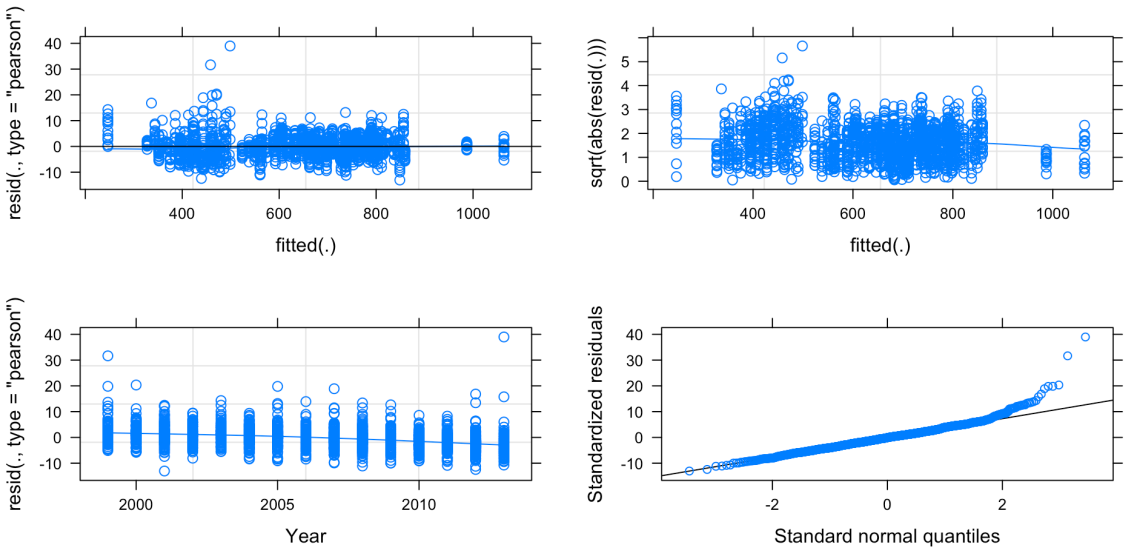

Figure 20. Model 6: Diagnostic Plots

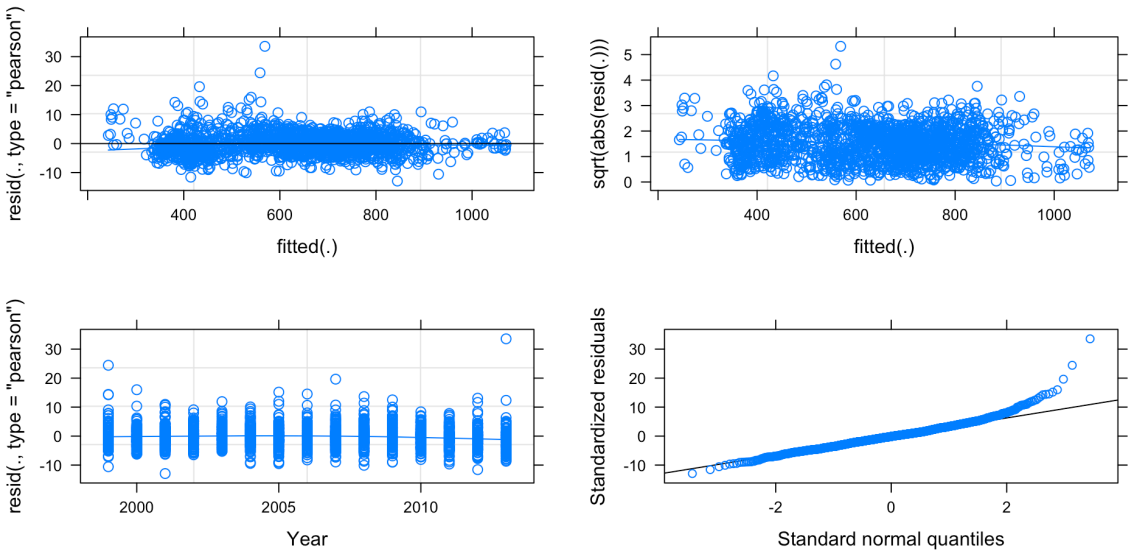

Figure 21. Model 7: Diagnostic Plots

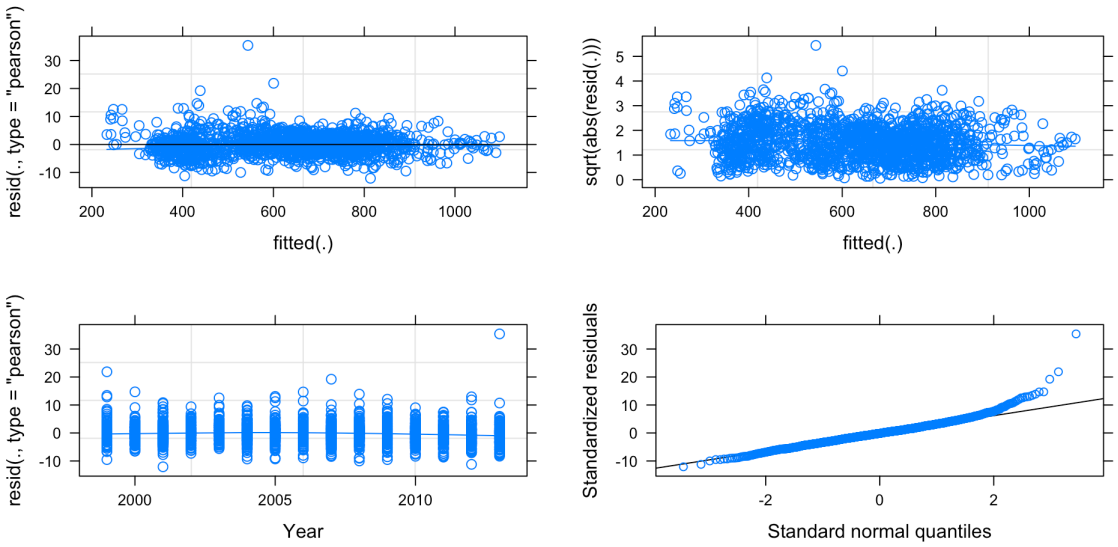

Figure 22. Model 8: Diagnostic Plots

From the diagnostic plots above, we observe that all the first four models Model 1~4 generally look good other than for the normality assumption. Although normality assumption is desired, there are also copious literatures stating that the normality assumption is not as important as the homoscedasticity assumption which is satisfied her for all the first four models Model 1~4. Normality test usually fails for large number of sample sizes, and indeed we do see some departures from normality here for extremely small and extremely large values. We would not regard this as a huge concern in this study, however, and we will proceed without removing any observations for the first four models Model 1~4.

On the other hand, however, the extent of deviation from normality for the last four models Model 5~8 is intolerable. As we would find out, there are also overdispersion issues for all Model 5~8. Hence, proper modification for Model 5~8 is urgent and required.

2.6.2 Multicollinearity Checks

VIF (Variance Inflation Factor) could be used to check the possible multicollinearity. If any of the values is greater than 5, then we should suspect potential multicollinearity. As we could see from the results below, none of the variables had VIF that is greater than 5 and thus there is no need to be concerned for multicollinearity here.

Table 3. Multicollinearity Checks

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Black	1	1	1.18	1.24	1.16	1.3	1.3	1.31
Hisp			1.18	1.47	1.16	1.17	1.17	1.2
PM2.5				1.38		1.01	1.17	1.25
Tobacco						1.25	1.17	1.24
Healthcare						1.12	1.62	1.03
Black:PM2.5							1.39	1.77
Hisp:PM2.5							1.51	1.44
Black:Tobacco							1.4	1.68
Hisp:Tobacco								1.46
Black:Healthcare								1.31
Hisp:Healthcare								1.22

2.6.3 Overdispersion Checks

The overdispersion check was conducted using the function overdisp_fun() quoted from the website developed by [Benjamin Bolker] (contributor of the R lme4 package which includes glmer() function). As suggested on the page, the rules of thumb for checking overdispersion for Poisson is to test whether the ratio $\lambda>5$. As we could see from the table below, Model 1~4 have the ratios $\lambda<5$ but Model 5~8 have the ratios $\lambda>5$. Therefore, we suspect that there are over dispersions for Model 5~8 and we need to make proper corrections for these issues. In an attempt to alleviate these overdispersions, we added observation-level random effects as suggested by [Harrison et al]. This procedure has indeed resulted in the reduction of the ratios but by exceedingly greater extent, and now we have the ratios well below zero, which could potentially be problematic.

Table 4. Over Dispersion Checks

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
chisq	1346.62	1130.58	886.42	826.14	31665.07	26905.36	24960.79	23779.93
ratio	1.83	1.54	1.22	1.14	18.16	15.47	14.4	13.78
rdf	734	732	728	723	1744	1739	1733	1726
p	0	0	0	0	0	0	0	0

2.6.4 Summary Table

Table 5. Model Selection Summary Table

	Dependent variable:								
	All Races (Not Stratified)					Races Stratified			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	
cPM2.5	0.0193*** (0.0177, 0.0210)	0.0226*** (0.0179, 0.0272)	0.0185*** (0.0149, 0.0222)	0.0168*** (0.0130, 0.0207)		0.0375*** (0.0289, 0.0460)	0.0279*** (0.0205, 0.0353)	0.0268*** (0.0202, 0.0334)	
cTobacco			0.0067*** (0.0029, 0.0105)	0.0069*** (0.0032, 0.0107)			0.0135*** (0.0082, 0.0191)	0.0120*** (0.0068, 0.0172)	
cHealthcare				0.0042** (0.0016, 0.0068)				0.0016 (-0.0023, 0.0055)	
RaceBlack:cPM2.5						-0.0078*** (-0.0096, -0.0061)	-0.0035*** (-0.0054, -0.0015)	-0.0090*** (-0.0111, -0.0069)	
RaceHispanic:cPM2.5						-0.0420*** (-0.0442, -0.0398)	-0.0374*** (-0.0398, -0.0349)	-0.0421*** (-0.0446, -0.0397)	
RaceBlack:cTobacco							-0.0064*** (-0.0078, -0.0051)	-0.0017*** (-0.0031, -0.0003)	
RaceHispanic:cTobacco							-0.0076*** (-0.0093, -0.0059)	-0.0032*** (-0.0050, -0.0014)	
RaceBlack:cHealthcare								0.0109*** (0.0099, 0.0119)	
RaceHispanic:cHealthcare								0.0148*** (0.0136, 0.0161)	
RaceBlack					0.0747*** (0.0704, 0.0791)	0.0781*** (0.0735, 0.0826)	0.0792*** (0.0746, 0.0838)	0.0815*** (0.0769, 0.0861)	
RaceHispanic					-0.4694*** (-0.4750, -0.4638)	-0.4615*** (-0.4672, -0.4559)	-0.4614*** (-0.4671, -0.4558)	-0.4532*** (-0.4590, -0.4475)	
Constant	-7.3159*** (-7.3623, -7.2696)	-7.3110*** (-7.3576, -7.2644)	-7.3031*** (-7.3406, -7.2655)	-7.2963*** (-7.3362, -7.2564)	-7.3110*** (-7.3513, -7.2706)	-7.3067*** (-7.3512, -7.2623)	-7.2986*** (-7.3396, -7.2576)	-7.2972*** (-7.3311, -7.2633)	
Observations	737	737	737	737	1,748	1,748	1,748	1,748	
Log Likelihood	-3,919.7500	-3,872.0460	-3,766.6780	-3,744.1510	-22,659.9000	-20,535.1400	-19,718.6500	-19,191.0700	
Akaike Inf. Crit.	7,845.4990	7,754.0920	7,551.3560	7,516.3010	45,327.8000	41,088.2700	39,467.2900	38,426.1400	
Bayesian Inf. Crit.	7,859.3070	7,777.1050	7,592.7800	7,580.7380	45,349.6600	41,137.4700	39,549.2900	38,546.4000	
Note:									*p<0.1; **p<0.05**; p<0.01

According to both AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), **Model 4** is preferred to all the models here because it has the lowest AIC and BIC. "The AIC penalizes the number of parameters less strongly than does the Bayesian information criterion (BIC). The BIC generally penalizes free parameters more strongly than the Akaike information criterion, though it depends on the size of n and relative magnitude of n and k." [Wiki] For the nested models, we could also use the Likelihood Ratio Test using anova( ) command to conduct the model selection.

Considering the normality assumption violation and the serious over dispersion of Model 5~8, it would be a fair judgement to conclude that Model 4 would be our best possible model among all Model 1~8. Therefore, **Model 4** will be used to conduct further analysis in the Results Chapter.

2.6.5 Summary Figures

As we could note from the Linear Regression figures, it would be cursory to just draw one line to represent the entire variation in the data. Therefore, instead, we exploited the features of Generalized Linear Mixed Model to account for both within and between States variations. Thus, we then realize that there is a remarkable improvement in the fit by switching from the general Linear Model Regression (lm( )) to the Generalized Linear Mixed-Effects Model Regression (glmer( )).

From all of the figures below, we confirmed that **Model 4** (Equations (12), (13), (14)) has one of the best fits out of Model 1~4. Therefore, we will conduct our analysis using **Model 4** in the Results Chapter.

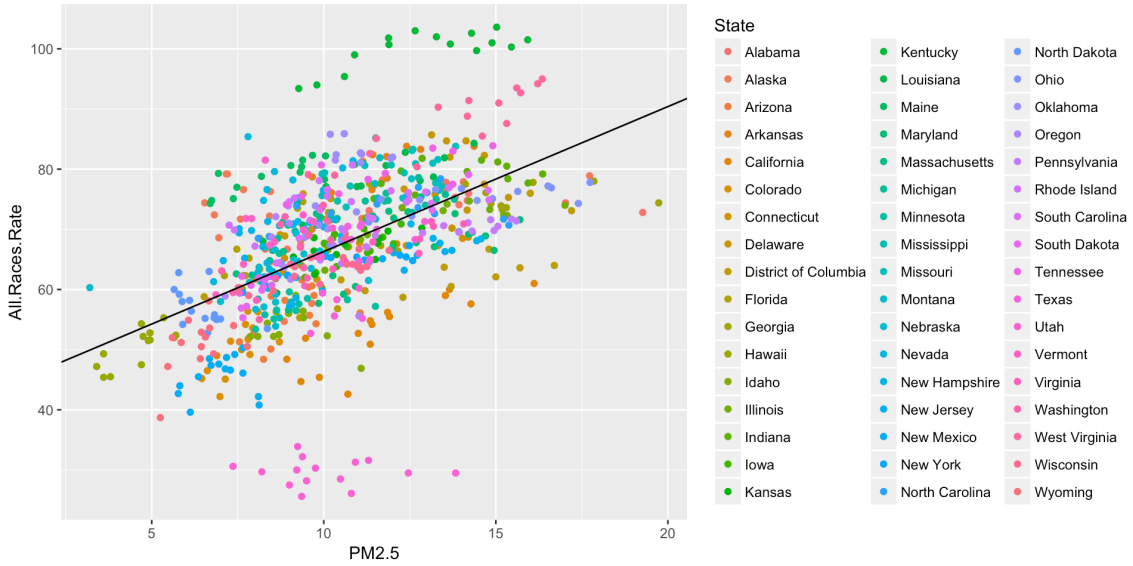

Figure 23. Linear Regression

Figure 24. Linear Regression Fit

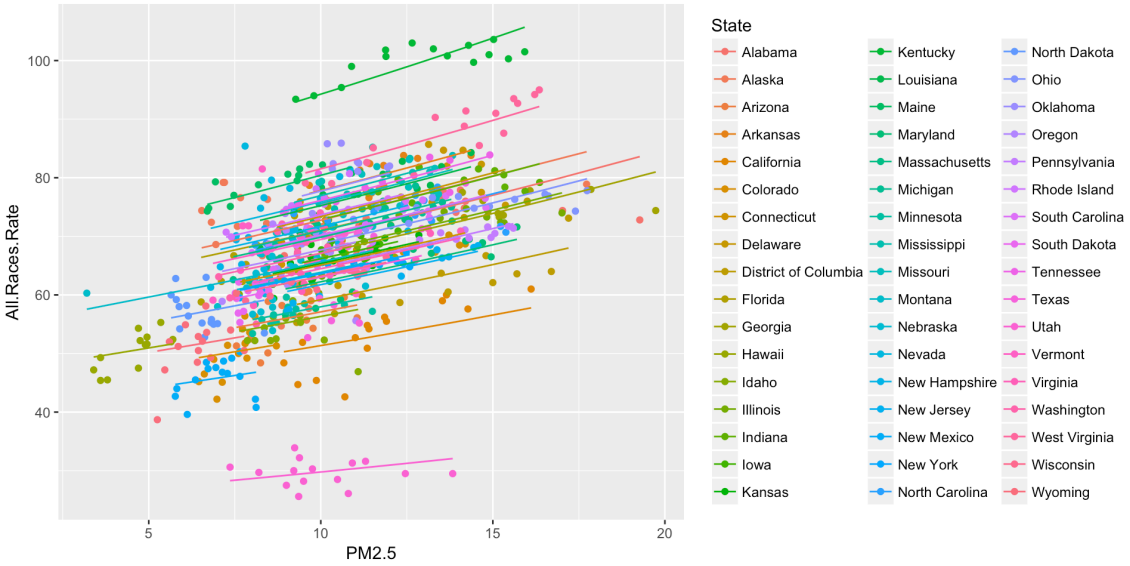

Figure 25. Model 1 Scatter Plot

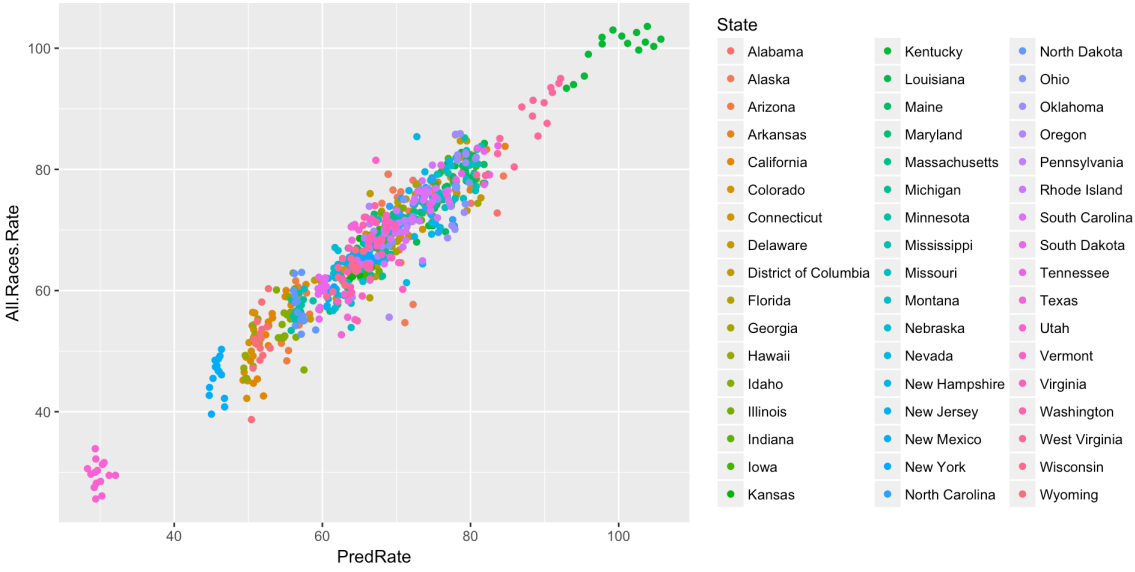

Figure 26. Model 1 Fit

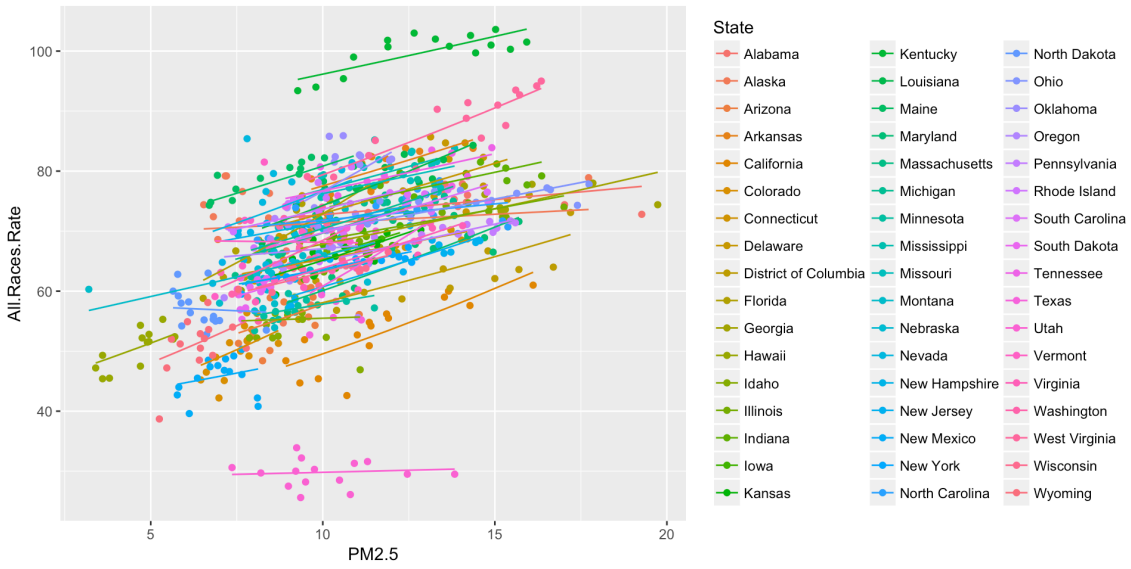

Figure 27. Model 2 Scatter Plot

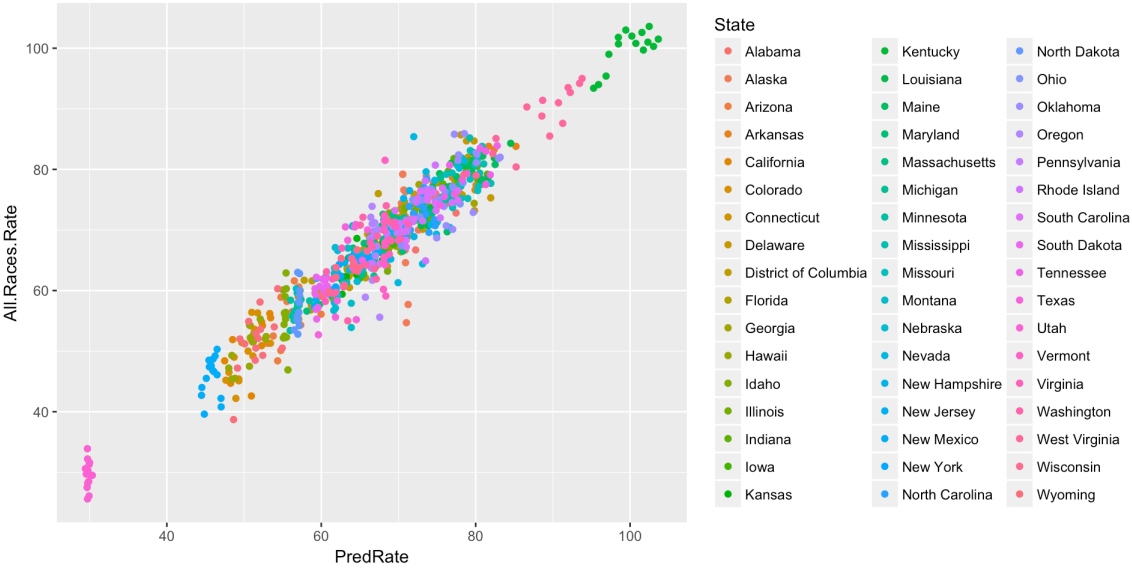

Figure 28. Model 2 Fit

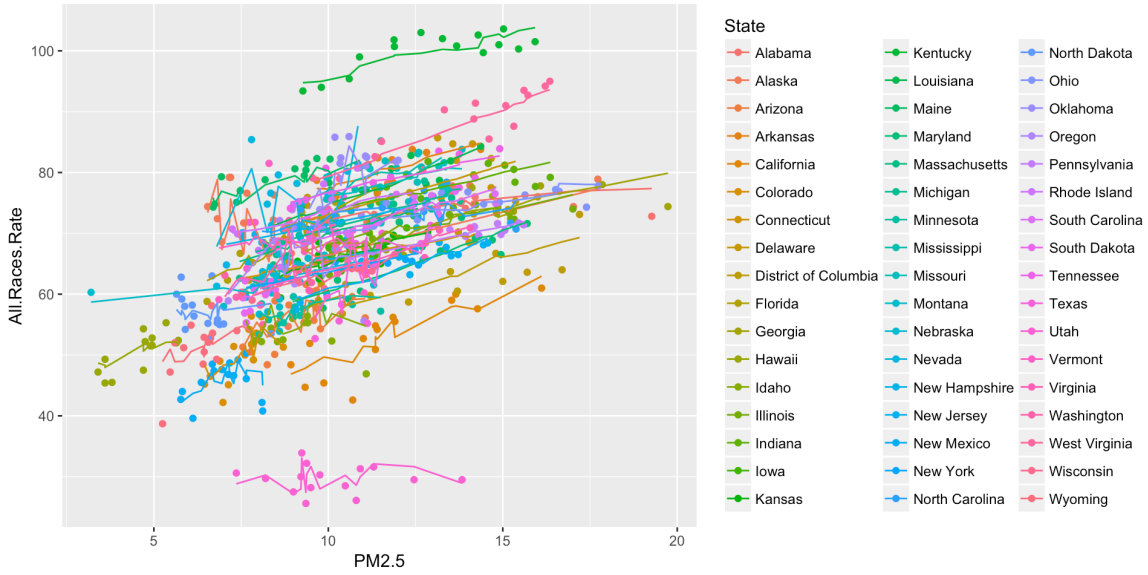

Figure 29. Model 3 Scatter Plot

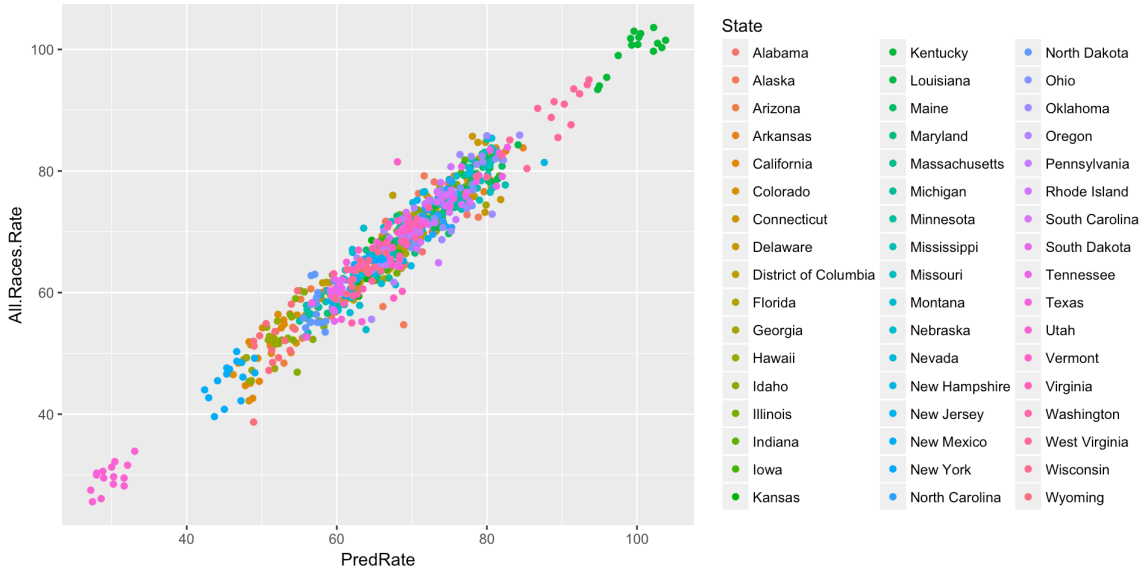

Figure 30. Model 3 Fit

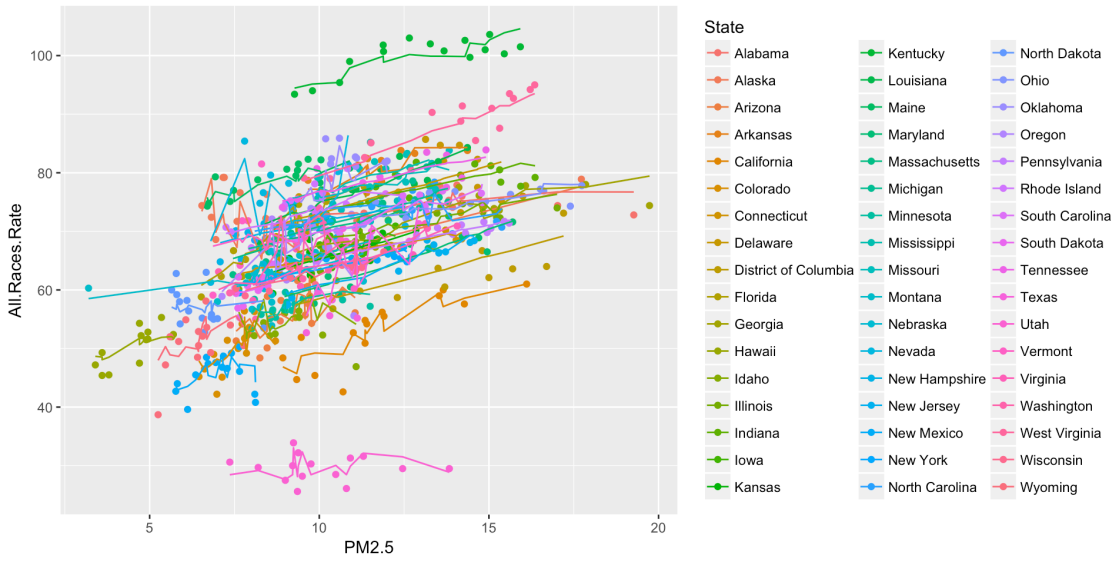

Figure 31. Model 4 Scatter Plot

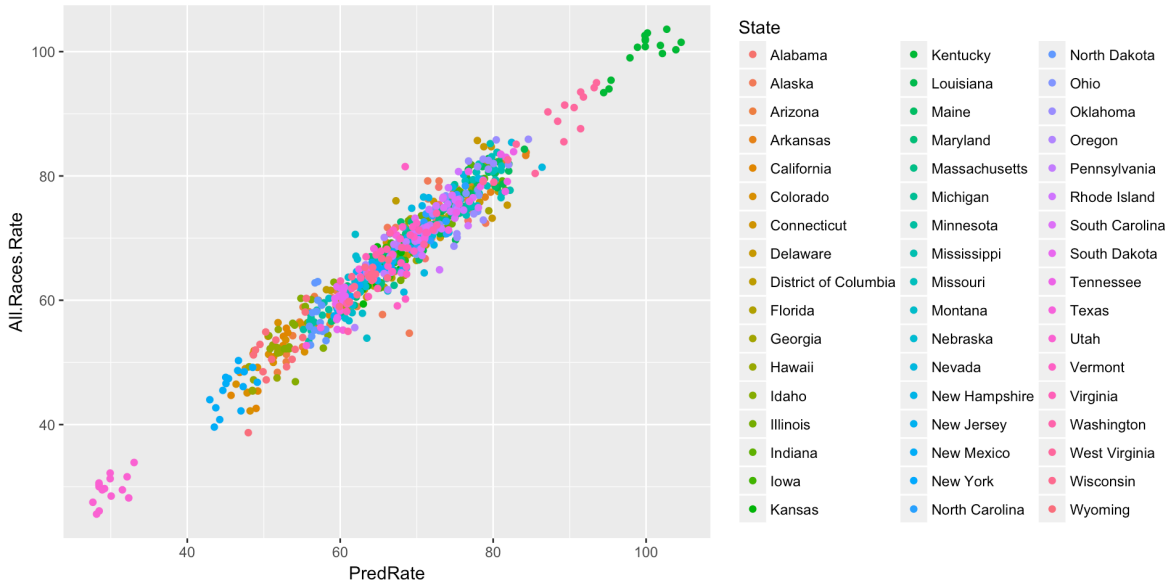

Figure 32. Model 4 Fit

3. Results

From here and onwards, the analysis will be proceeded using **Model 4** as it has the lowest AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) among all the Models. Our conclusions for this study will also be drawn from **Model 4**.

3.1 Statistical Tests

Table 6. Statistical Tests

	Dependent variable:			
	All Races (Not Stratified)			
	Model 1	Model 2	Model 3	Model 4
cPM2.5	0.01932*** (0.01768, 0.02096)	0.02256*** (0.01787, 0.02724)	0.01853*** (0.01488, 0.02218)	0.01683*** (0.01299, 0.02068)
cTobacco			0.00671*** (0.00292, 0.01051)	0.00695*** (0.00317, 0.01073)
cHealthcare				0.00417*** (0.00159, 0.00676)
Constant	-7.31593*** (-7.36225, -7.26960)	-7.31100*** (-7.35765, -7.26435)	-7.30305*** (-7.34063, -7.26547)	-7.29628*** (-7.33618, -7.25639)
Observations	737	737	737	737
Log Likelihood	-3,919.75000	-3,872.04600	-3,766.67800	-3,744.15100
Akaike Inf. Crit.	7,845.49900	7,754.09200	7,551.35600	7,516.30200
Bayesian Inf. Crit.	7,859.30700	7,777.10500	7,592.78000	7,580.73800
Note:				Model 1~4 are all presented for the purpose of references

The intercept and all the covariates PM2.5, Tobacco, and Healthcare were significant at α level of 0.05. The 95% Confidence Intervals for each of the variables are also presented in the table above. We also note that the coefficient estimates are roughly within the same range across the four models, and these are all significant at α level of 0.05.

3.2 Coefficient Estimates

3.2.1 Interpretation

Remember that our "Rate" was modeled with Poisson Distribution Log Link $log(Rate_{it}) = X\gamma + Zu + \epsilon_{it}$ or $Rate_{it} = \exp(X\gamma + Zu + \epsilon_{it})$. Hence, in order to evaluate the association between $10\mu g/m^3$ increases in PM2.5 and Lung Cancer Incidence Rates increase, we need to multiply the coefficient estimate by the factor of 10 and then exponentiate the value. With this calculation, for example, we obtain the estimated Relative Risk of PM2.5 on Lung Cancer Incidence Rates to be exp(10x0.017) = 1.183 (95% CI: 1.181, 1.186) on State level.

This is slightly higher than the Meta-Relative Risk of 1.09 (95% CI: 1.04, 1.14) deduced by [Harma et al] on an individual level. However, it is worth mentioning that our Modeled-Relative Risk of 1.183 (95% CI: 1.181, 1.186) overlapped with 10 out of 14 studies' 95% CIs quoted in the Meta-Analysis [McDonell et al. 2000, Hart et al. 2011, Lipsett et al. 2011, Lepeule et al. 2012, Hystad et al. 2013, Jerrett et al. 2013, Puett et al. 2014, Carey et al. 2013, Raaschou-Nielsen et al. 2013, Katanoda et al. 2011].

Likewise, the Modeled-Relative Risk of Tobacco on Lung Cancer Incidence Rates is 1.072 (95% CI: 1.07, 1.074) per 10% increase in the Tobacco prevalence on State level, and the Modeled-Relative Risk of Healthcare on Lung Cancer Incidence Rates is 1.043 (95% CI: 1.041, 1.044) per 10% increase in the Healthcare coverage on State level.

3.2.2 Confidence Intervals

BLUP (Best Linear Unbiased Prediction) for each of the coefficients from Model 4 are shown in the forest figure below.

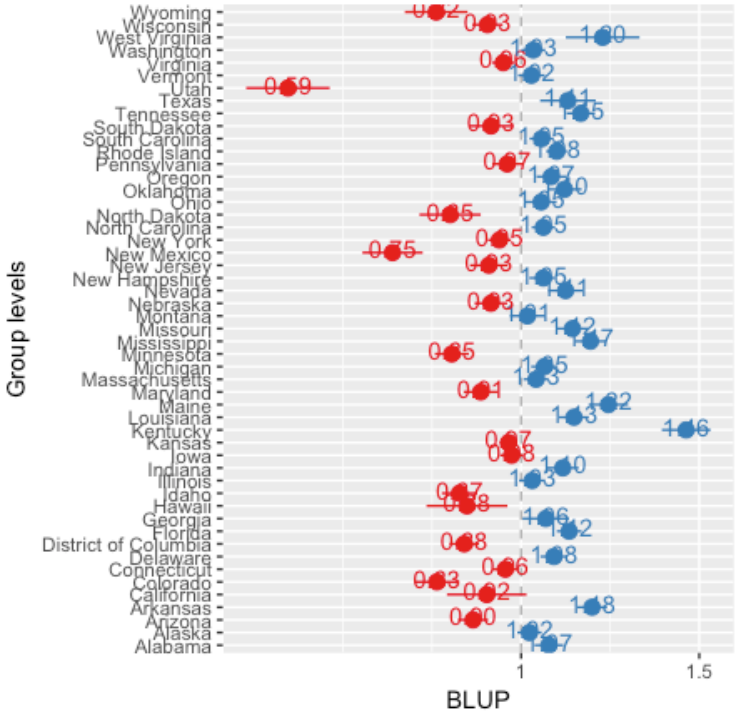

Figure 33. Random Effect of Intercept

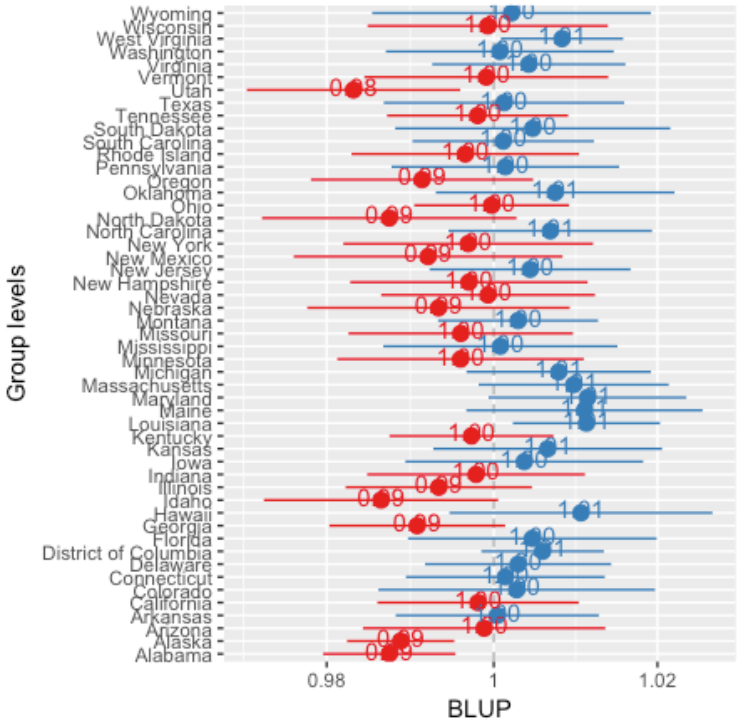

Figure 34. Random Effect of PM2.5

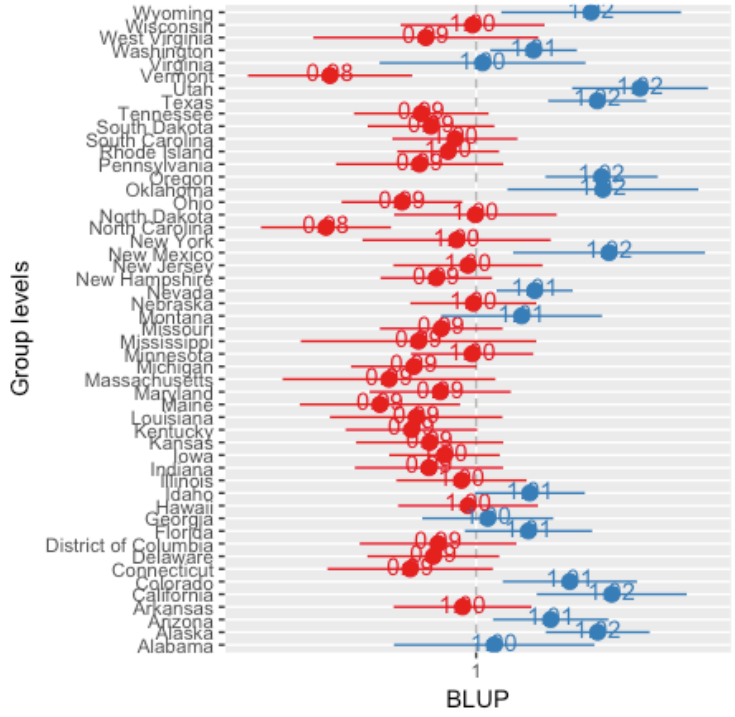

Figure 35. Random Effect of Tobacco

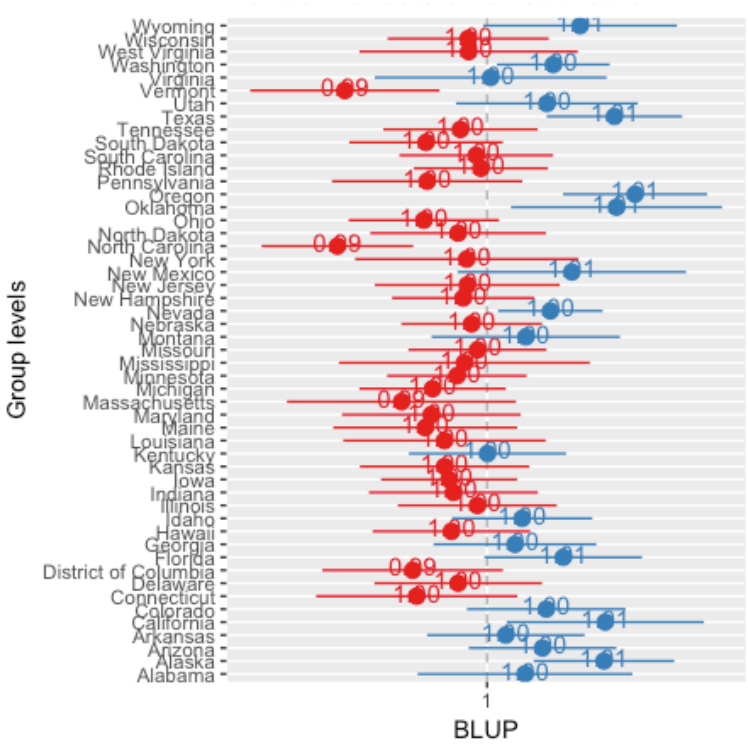

Figure 36. Random Effect of Healthcare

3.2.3 Estimates Mapping

The estimated Average Incidence Rates and the estimated respective covariates' Mixed Effects (Fixed Effects + Random Effects) are mapped onto the U.S. 50 States and D.C. Map in the following figures. Right below the individual figures, we have also included the actual Average Incidence Rates, PM2.5 Concentration, Tobacco Prevalence, and Healthcare Coverage by States for the purpose of comparison. Since there was not much change in these actual values' relative levels across States over time, Year 2003 was chosen as the representative to demonstrate the actual PM2.5 Concentration, Tobacco Prevalence, and Healthcare Coverage.

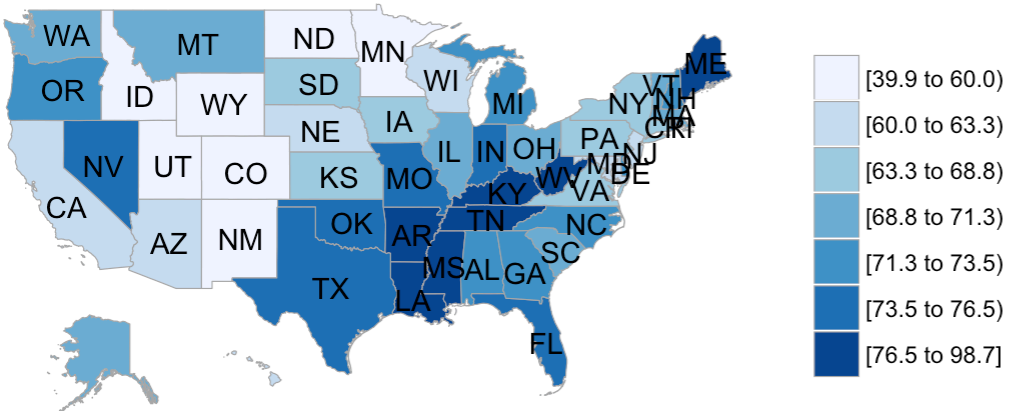

Figure 37. Intercept (Average Incidence Rate)

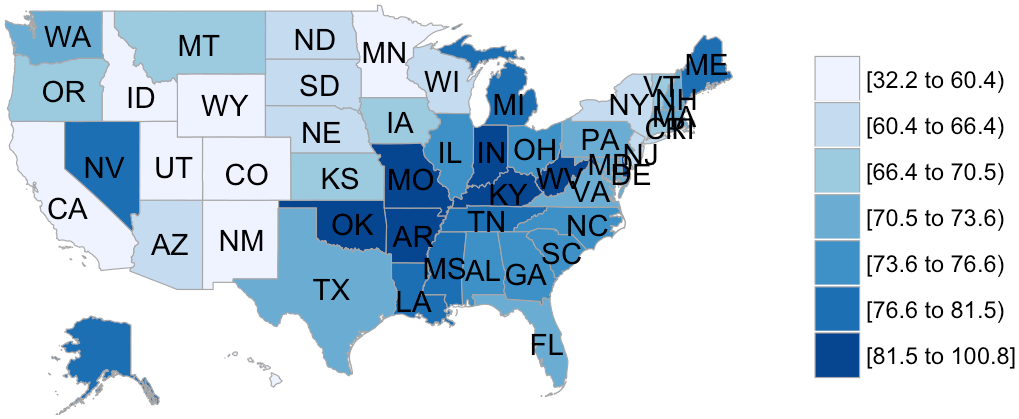

Figure 38. 2003 Incidence Rate

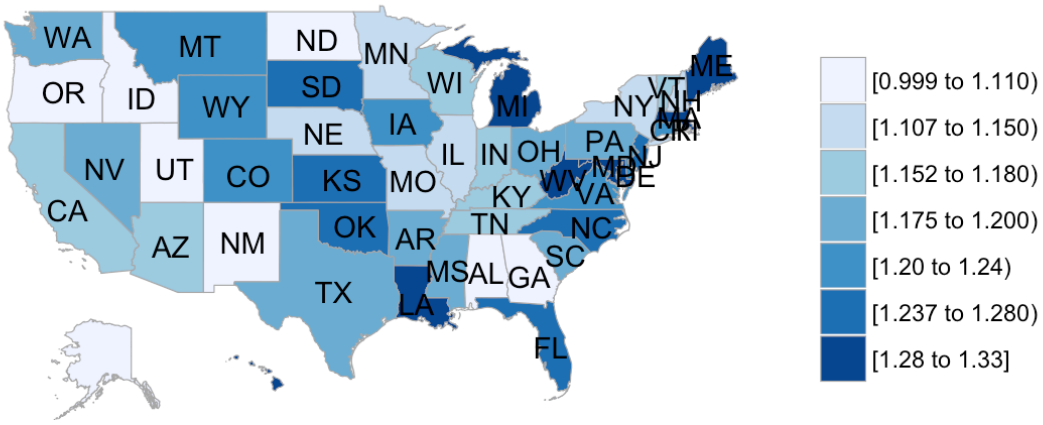

Figure 39. PM2.5 Effect

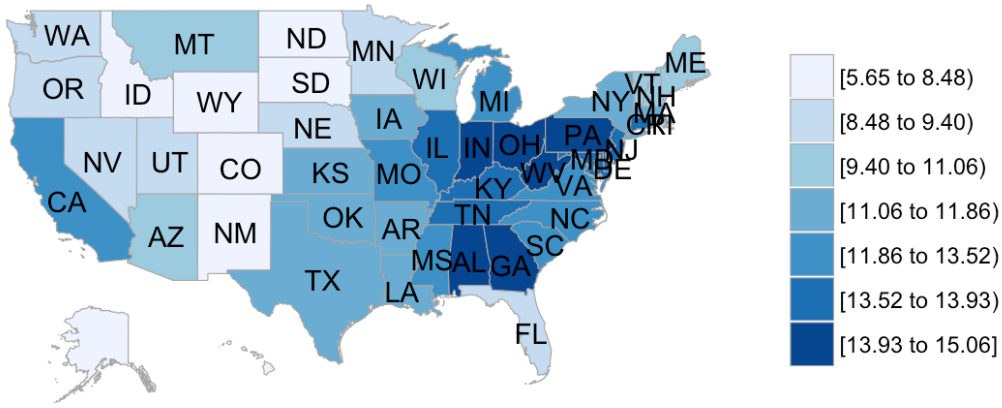

Figure 40. 2003 PM2.5 Concentration

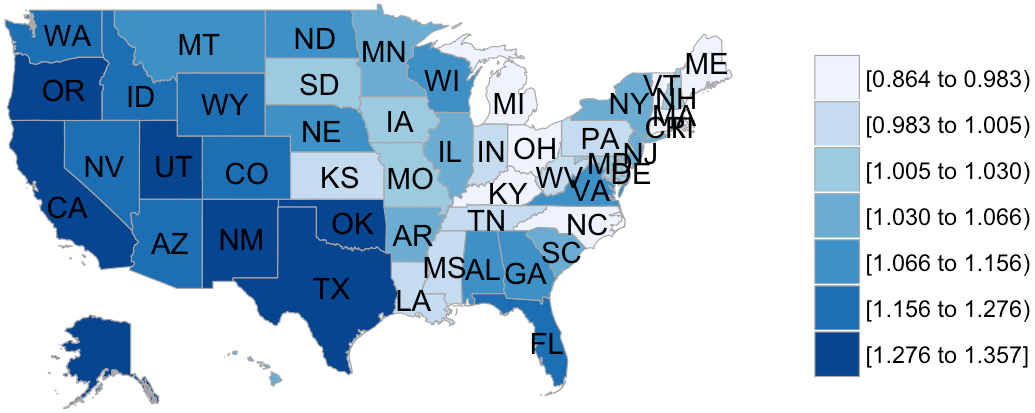

Figure 41. Tobacco Effect

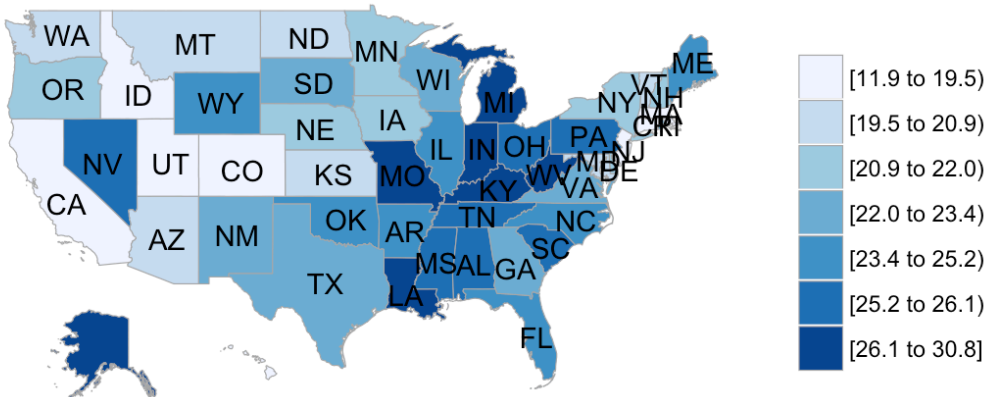

Figure 42. 2003 Tobacco Prevalence

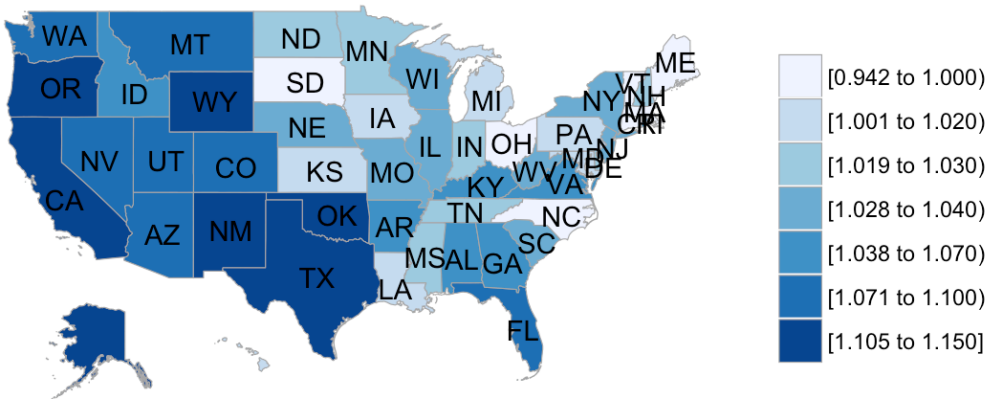

Figure 43. Healthcare Effect

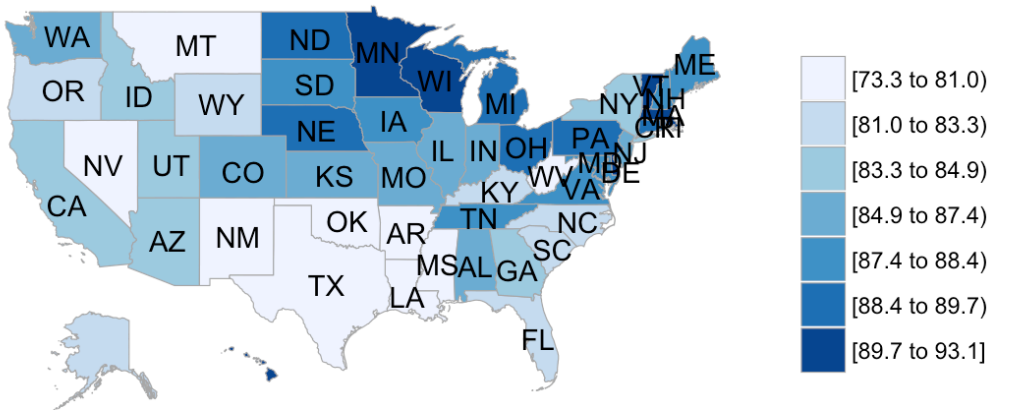

Figure 44. 2003 Healthcare Coverage

We note that intercept estimates and PM2.5 coefficient estimates are generally higher on the East Coast but Tobacco coefficient estimates and Healthcare coefficient estimates are generally higher on the West Coast. Please note that this trend was even severer before we demeaned the independent variables (center the independent variables around their respective means). This enigma will be further expanded in the following Discussion Chapter.

4. Discussions

4.1 Achievements and Challenges

First of all, it is worth noting that our study using the publicly available ecological datasets collected from CDC and EPA has successfully yielded an estimated Relative Risk of PM2.5 on Lung Cancer Incidence Rate that is consistent with many other literatures using privately owned individual level datasets. By no means does this insinuate that we could extrapolate the State level result to an individual level result, but it does demonstrate the power of Generalized Linear Mixed Model that accounts for both within and between subject effects. Though we were able to deduce the estimated Relative Risk of PM2.5 on Lung Cancer Incidence Rates on State level, this has also brought us a new puzzling question after mapping the predicted mixed effects for each of the States onto the U.S. 50 States and D.C. Map.

The estimated association between Lung Cancer Incidence Rates and PM2.5 was generally higher on the East Coast States, while the association between Lung Cancer Incidence Rates and Tobacco as well as Healthcare were higher on the West Coast States. In order to further delve into these phenomena, it would be helpful to understand the geographical and racial factors across the U.S. 50 States and D.C. In this section, we will attempt to provide some explanation for the higher association between Lung Cancer Incidence Rates and PM2.5 on the East Coast in comparison to the West Coast.

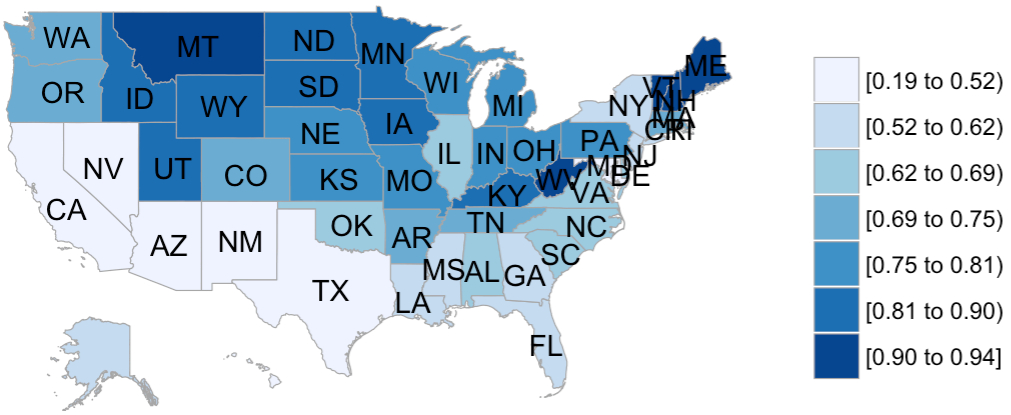

Figure 45. 2003 White Population

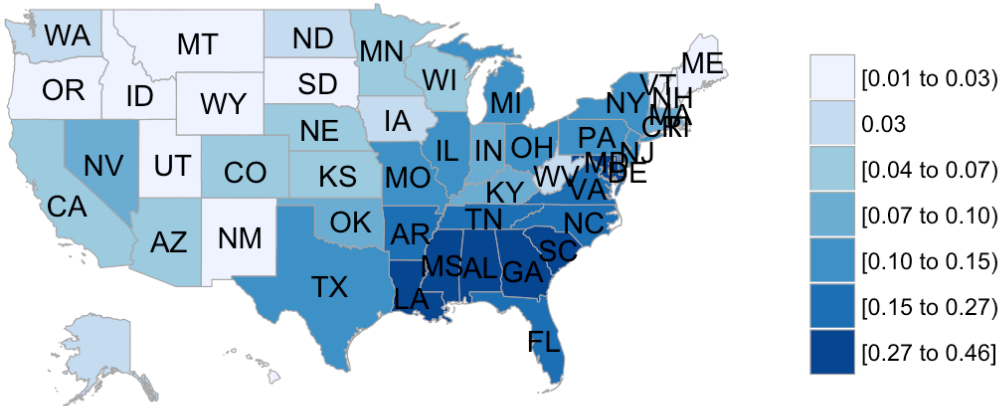

Figure 46. 2003 Black Population

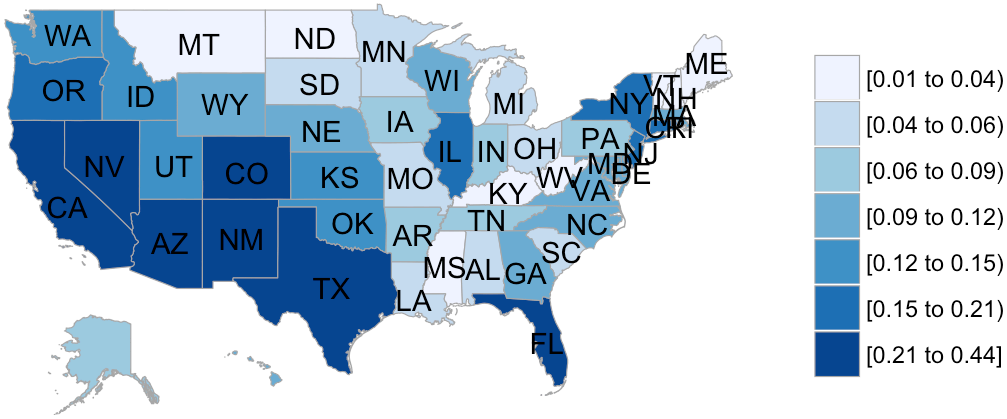

Figure 47. 2003 Hispanic Population

From the above figures, we observe that Northern regions generally have higher White population, Southern regions generally have higher Black population, and South Western regions generally have higher Hispanic population. By comparing these three maps with the maps from the Estimates Mapping Section, a natural first thought would be to conjecture that Whites are highly associated with the PM2.5 effect, Race Hispanic is highly associated with the Tobacco, and Race Black and Hispanic are highly associated with the Healthcare. This method of eyeballing, however, is not scientific and it needs to be complemented by some statistical analysis. Although there were overdispersion issues for Model 5~8, it would be useful to revisit these models time to time to observe the interaction terms between Race and each of the other covariate effects for suggestive purposes.

4.2 Possible Explanations

1. **Why does East Coast have higher effect of PM2.5 on Lung Cancer Incidence Rates than West Coast?**

- As we could note from Figure 40, East Coast has higher PM2.5 concentration than West Coast in general. We could possibly hypothesize that there is a certain cumulative threshold of PM2.5 intake for a person to start developing a cancer.
- It is also possible that there were sudden increase in the PM2.5 emission by factories and which could be seasonal. Hence, the data that we collected on yearly basis might have ignored the high variance of PM2.5 over season since could have been smoothen out by taking the yearly mean PM2.5 concentration. It is

possible that, for instance winter has much higher PM2.5 concentration and thus people on the East Coast are more likely to reach the threshold.

- This threshold argument is quite unlikely, however, since there is a general consensus among the researchers that there is no threshold exposure level below which no carcinogenic effect exists [Dockery et al and Pope et al].
- Indeed, WHO has also published that, "small particulate pollution have health impacts even at very low concentrations – indeed no threshold has been identified below which no damage to health is observed."
- It is also possible that Whites are more susceptible to PM2.5 than Blacks and Hispanics. In fact, from Model 8, we observed that Whites have higher Relative Risk of Lung Cancer Incidence Rates with respect to PM2.5 compared to Blacks and Hispanics.

2. Why does West Coast have higher effect of Tobacco on Lung Cancer Incidence Rates than East Coast?

- It is possible that Blacks and Hispanics are more susceptible to Tobacco induced Lung Cancer compared to Whites. Model 8, however, contradicts this hypothesis. Both interaction terms Black:Tobacco and Hispanic:Tobacco turned out to be lower than the White:Tobacco interaction term. [Haiman et al] studied the Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer, and found out the African American are more susceptible to Lung Cancer than Whites, but also showed that Latinos are the least susceptible.

3. **Why does West Coast have higher effect of Healthcare on Lung Cancer Incidence Rates than East Coast?**

- As we could note from Figure 44, North-Eastern States have higher Healthcare coverage compared to the other parts of U.S. We could conceivably hypothesize that improved healthcare coverage would encourage more people to be diagnosed more often and thus there is higher chance of Lung Cancer being detected and thus registered. This effect of improvement in Healthcare coverage on Lung Cancer Incidence Rate registration rate would be more conspicuous in the States that did not have good Healthcare coverage before. Hence, it makes sense that the Non North-Eastern States, which are South-Western States, have higher potential rise in Lung Cancer Incidence Rates when the Healthcare Coverage is enhanced by the same percentage increase.

5. Conclusion

Our goal was to study the association between Lung Cancer Incidence Rates and PM2.5 using the publicly available ecological data. In fact, our Modeled-Relative Risk of 1.183 (95% CI: 1.181, 1.186) overlapped with 10 out of 14 studies' 95% CIs quoted in the Meta Analysis conducted by [Harma et al]. This is a rather powerful result, and we note that if our ecological dataset is supplemented by some individual level dataset, we would also be able to make an inference on the individual level as suggested by NIH-PA Author Manuscript.

With the utilization of Linear Mixed Effects Model, we were able to explain both within and between State variations in the data, and eventually obtained the Relative Risk for each of the States since the random components of the model varies by States. Eventually, we mapped the Relative Risk of Lung Cancer Incidence Rates with respect to each of the covariates onto the U.S. 50 States and D.C. Map. We observed generally higher association between PM2.5 and Lung Cancer Incidence Rates on the East Coast and generally higher association between Tobacco as well as Healthcare and Lung Cancer Incidence Rates on the West Coast. These phenomena could be partially due to the marginal effect of geographical or racial factors as discussed in our study.

We believe that Multilevel Modeling (Hierarchical Modeling) using both individual and ecological level datasets could further advance the analysis and help us better understand these rather interesting yet puzzling phenomena.

Bibliography

[1] Albright, J. J., & Marinova, D. M. (2015). Estimating multilevel models using SPSS, Stata, SAS and R.

[2] Blakely, T.A., & Woodward, A.J. Ecological effects in multi-level studies. Journal of Epidemiology & Community Health 2000;54:367-374.

[3] Carey, Iain M., Richard W. Atkinson, Andrew J. Kent, Tjeerd Van Staa, Derek G. Cook, and H. Ross Anderson. "Mortality Associations with Long-Term Exposure to Outdoor Air Pollution in a National English Cohort." American Journal of Respiratory and Critical Care Medicine 187.11 (2013): 1226-233.

[4] Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. Annual review of psychology, 62, 583-619.

[5] Erickson, A. C., Ostry, A., Chan, H. M., & Arbour, L. (2016). Air pollution, neighbourhood and maternal-level factors modify the effect of smoking on birth weight: a multilevel analysis in British Columbia, Canada. *BMC Public Health*, *16*, 585. http://doi.org/10.1186/s12889-016-3273-9

[6] Fu, J., Jiang, D., Lin, G., Liu, K., & Wang, Q. (2015). An ecological analysis of PM2. 5 concentrations and lung cancer mortality rates in China. *BMJ open*, *5*(11), e009452.

[7] Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Vol. Analytical methods for social research, Cambridge University Press , New York .

[8] Hamra, G.B., Guha, N., Cohen, A., Laden, F., Raaschou-Nielsen, O., Samet, J.M., Vineis, P., Forastiere, F., Saldiva, P., Yorifuji, T., & Loomis, D., (2014). Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis. NLM-Export.

[9] Hystad, Perry, Paul A. Demers, Kenneth C. Johnson, Richard M. Carpiano, and Michael Brauer. "Long-term Residential Exposure to Air Pollution and Lung Cancer Risk." Epidemiology 24.5 (2013): 762-72

[10] Imai, K., & Kim, I. S. (2016). When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data? (Doctoral dissertation, Working paper, Princeton University, Princeton, NJ).

[11] Jerrett, M., R. T. Burnett, B. S. Beckerman, M. C. Turner, D. Krewski, G. Thurston, R. V. Martin, A. Van, E. Hughes, Y. Shi, S. M. Gapstur, M. J. Thun, and 3. R. Pope. "Spatial analysis of air pollution and mortality in California." American journal of respiratory and critical care medicine. U.S. National Library of Medicine, 01 Sept. 2013.

[12] Katanoda, Kota, Tomotaka Sobue, Hiroshi Satoh, Kazuo Tajima, Takaichiro Suzuki, Haruo Nakatsuka, Toshiro Takezaki, Tomio Nakayama, Hiroshi Nitta, Kiyoshi Tanabe, and Suketami Tominaga. "An Association Between Long-Term Exposure to Ambient Air Pollution and Mortality From Lung Cancer and Respiratory Diseases in Japan." Journal of Epidemiology 21.2 (2011): 132-43.

[13] Lepeule, Johanna, Francine Laden, Douglas Dockery, and Joel Schwartz. "Chronic Exposure to Fine Particles and Mortality: An Extended Follow-up of the Harvard Six Cities Study from 1974 to 2009." Environmental Health Perspectives 120.7 (2012): 965-70.

[14] Lipsett, Michael J., Bart D. Ostro, Peggy Reynolds, Debbie Goldberg, Andrew Hertz, Michael Jerrett, Daniel F. Smith, Cynthia Garcia, Ellen T. Chang, and Leslie Bernstein. "Long-Term Exposure to Air Pollution and Cardiorespiratory Disease in the California Teachers Study Cohort." American Journal of Respiratory and Critical Care Medicine 184.7 (2011): 828-35.

[15] Marí-Dell'Olmo, M., & Martínez-Beneito, M. Á. (2015). A multilevel regression model for geographical studies in sets of non-adjacent cities. PloS one, 10(8), e0133649.

[16] Mcdonnell, William F., N A O M I Nishino-Ishikawa, Floyd F. Petersen, Lie H O N G Chen, and David E. Abbey. "Relationships of mortality with the fine and coarse fractions of long-term ambient PM10 concentrations in nonsmokers." Journal of Exposure Analysis and Environmental Epidemiology 10.5 (2000): 427-36.

[17] Raaschou-Nielsen, Ole, Zorana Jovanovic Andersen, Steen Solvang Jensen, Matthias Ketzel, Mette Sørensen, Johnni Hansen, Steffen Loft, Anne Tjønneland, and Kim Overvad. "Traffic air pollution and mortality from cardiovascular disease and all causes: a Danish cohort study." Environmental Health 11.1 (2012): n. pag.

[18] Pollet, T. V., Stulp, G., Henzi, S. P., & Barrett, L. (2015). Taking the aggravation out of data aggregation: A conceptual guide to dealing with statistical issues related to the pooling of individual‐level observational data. *American journal of primatology*, *77*(7), 727-740.

[19] Puett, Robin C., Jaime E. Hart, Jeff D. Yanosky, Donna Spiegelman, Molin Wang, Jared A. Fisher, Biling Hong, and Francine Laden. "Particulate Matter Air Pollution Exposure, Distance to Road, and Incident Lung Cancer in the Nurses' Health Study Cohort." Environmental Health Perspectives (2014): n. pag.

[20] Richards, T. B., Johnson, C. J., Tatalovich, Z., Cockburn, M., Eide, M. J., Henry, K. A., & Ajani, U. A. (2011). Association between cutaneous melanoma incidence rates among white US residents and county-level estimates of solar ultraviolet exposure. Journal of the American Academy of Dermatology, 65(5), S50-e1.

[21] Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. International journal of epidemiology, 38(2), 337-341.

[22] Villeneuve, P.J., Goldberg, M.S., Burnett, R.T., et al. Associations between cigarette smoking, obesity, sociodemographic characteristics and remote-sensing-derived estimates of ambient PM2.5: results from a Canadian population-based survey. Occup Environ Med 2011; 68:920-927.

[23] Vinikoor-Imler, L. C., Davis, J. A., & Luben, T. J. (2011). An ecologic analysis of county-level PM2. 5 concentrations and lung cancer incidence and mortality. International journal of environmental research and public health, 8(6), 1865-1871.

[24] Wang, Y., Kloog, I., Coull, B. A., Kosheleva, A., Zanobetti, A., & Schwartz, J. D. (2016). Estimating Causal Effects of Long-Term PM2.5 Exposure on Mortality in New Jersey. Environmental health perspectives, 124(8), 1182.

[25] Zheng, T., Holford, T. R., Boyle, P., Chen, Y., Ward, B. A., Flannery, J., & Mayne, S. T. (1994). Time trend and the age-period-cohort effect on the incidence of histologic types of lung cancer in connecticut, 1960-1989. Cancer, 74: 1556–1567.

[26] "GLMM." DRAFT r-sig-mixed-models FAQ - GLMM. N.p., n.d. Web. 01 May 2017.

[27] "Media Centre - IARC NEWS." Media Centre - IARC News. N.p., n.d. Web. 01 May 2017.