Analysis of Coronary Heart Disease Data

1 Introduction

In this project, we are going to explore the prevalence of Coronary Heart Disease (CHD) among 18+ adults in Pennsylvania in 2020. According to CDC, coronary heart disease is the most common type of heart disease in the United States. CHD is a serious disease, and for some people the first sign of CHD is a heart attack. Thus, it is very important to learn about its spatial distribution and explore possible risk factors that are associated with it.

In the following parts, we are going to map the observation and expected counts of the CHD data first, and further compute the SMRs to explore the spatial distribution of the CHD morbidity, and fit non-spatial and spatial Poisson-Lognormal smoothing models to smooth the relative risks.

Further, we will explore five county-level variables that are possibly related to CHD morbidity, which are high cholesterol rate, smoke rate, low education rate, family income, and PM2.5 level. We will explore the associations of these variables with CHD using non-spatial and spatial Poisson-Lognormal models, and discuss the difference of the results from these two models.

The results show that the western and northern part of Pennsylvania tends to have higher CHD morbidity, and the southeastern part of Pennsylvania has relatively lower CHD morbidity. The association analysis shows that high cholesterol rate, smoke rate are positively associated with CHD morbidity, and PM2.5 level has negative association with CHD.

2 Data Description

All of the data that we are going to analyze is obtained from the Interactive Atlas of Heart Disease and Stroke section on the website of Centers for Disease Control and Prevention. The URL of the website is: https://nccd.cdc.gov/DHDSPAtlas/.

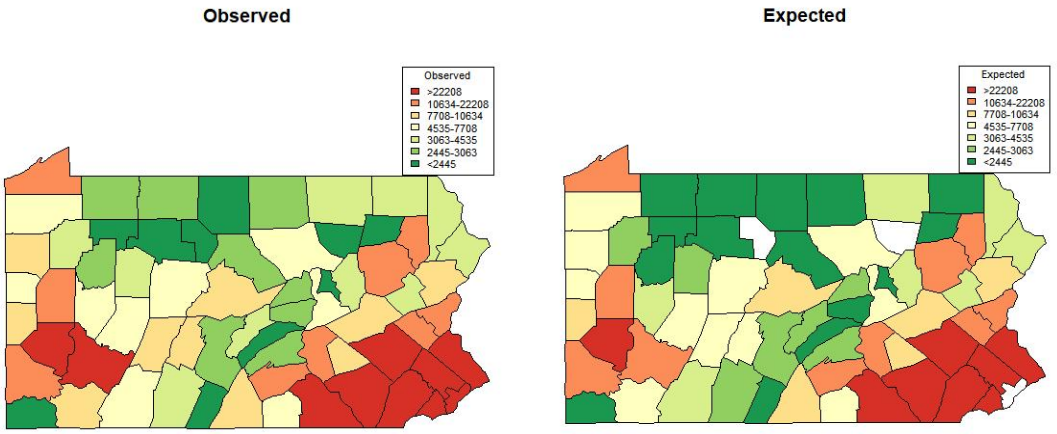
We are going to use 7 tables obtained from the website: (1) A table containing FIPS codes and county-level CHD observations among 18+ adults in Pennsylvania in 2020. (2) A table containing FIPS codes and county-level population of 18+ adults in Pennsylvania in 2020. (3) A table containing FIPS codes and county-level high cholesterol rate among 18+ adults in Pennsylvania in the past 5 years. (4) A table containing FIPS codes and county-level smoke rate among 18+ adults in Pennsylvania in 2020. (5) A table containing FIPS codes and county-level low education (lower than high school) rate in Pennsylvania in the past 5 years. (6) A table containing FIPS codes and county-level family income in Pennsylvania in 2020. (7) A table containing FIPS codes and county-level PM2.5 level in Pennsylvania.

Before importing the data into R, I merge the seven tables into one table according to the FIPS codes. Among the variables, the CHD rate, high cholesterol rate, smoke rate, and low education rate are presented by percentage. The neighbor relationships of counties in Pennsylvania is obtained here: https://faculty.washington.edu/jonno/SISMIDmaterial/.

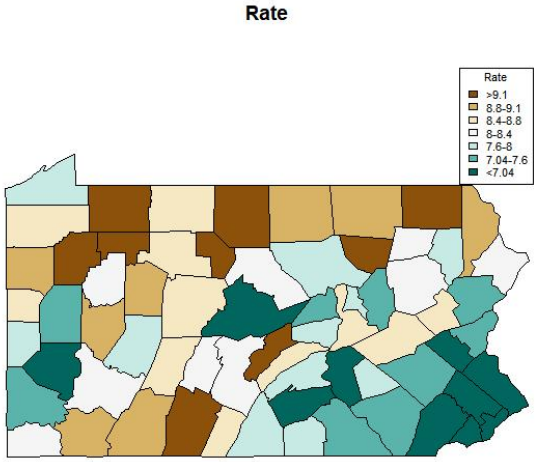
3 Analysis Methods and Results

3.1 Disease mapping

In this part, we are going to map the CHD data and analyze the distribution and variation of the SMR and the smoothed ratio. First, we map the observed count and expected count to have a look at the distribution of the CHD data in Pennsylvania. The maps are shown below.



As shown in the maps, we divide the data counts into 7 levels, and we can see that the southeast part seems to have more CHD cases, and the north part has relatively less cases. However, we are more interested in the rate (prevalence) of the disease, so we compute the ratio of the CHD and map it as below.



As we expected, the distribution of the rate is much different than the count distribution. We can see that the north part seem to have higher ration of CHD, and on the contrary, the southeast part now has lower CHD rate.

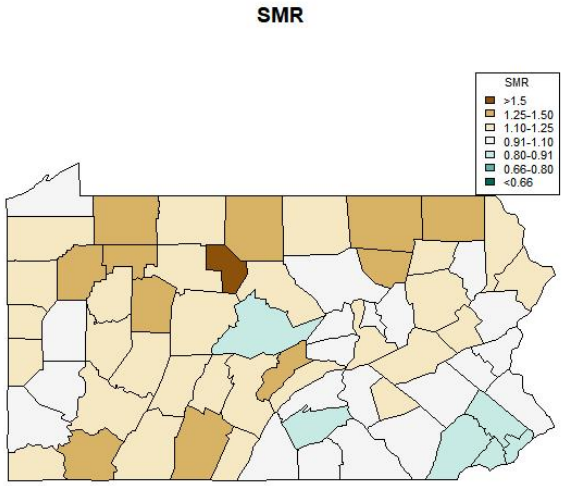
To obtain more meaningful results, we compute the standardized morbidity rates (SMRs) as below:

$$\mathrm{SMR_i} = \mathrm{Y_i}/\mathrm{E_i}$$

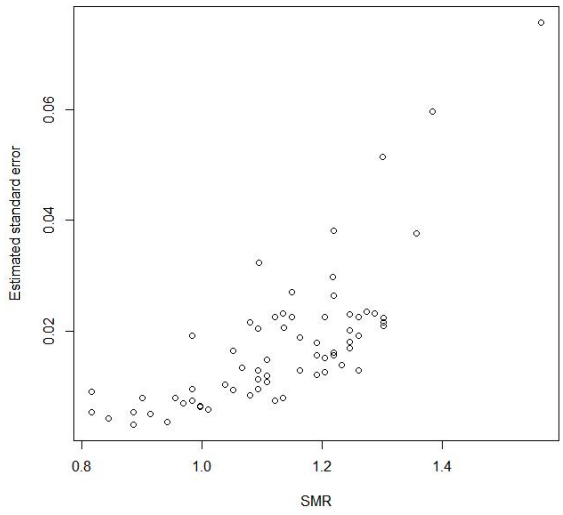
where $\mathrm{Y_i}$ is the observed CHD count in area i and $\mathrm{E_i}$ is the expected CHD count in area i, which can be computed as

$$\mathrm{E_i} = \frac{\sum_i \mathrm{Y_i}}{\sum_i \mathrm{P_i}} \times \mathrm{P_i}$$

where $\mathrm{P_i}$ is the population of area i. We can map the SMRs as below.



We can see that compared to the naive rate, the SMRs are less variable. Now we compute the estimated standard errors of SMRS and plot the SMRs versus the estimated standard errors as below.



We can observe that larger SMRs tends to have larger standard errors, although the estimated standard errors of all SMRs are quite small. Next, we fit a Poisson-Lognormal non-spatial model to smooth the SMRs:

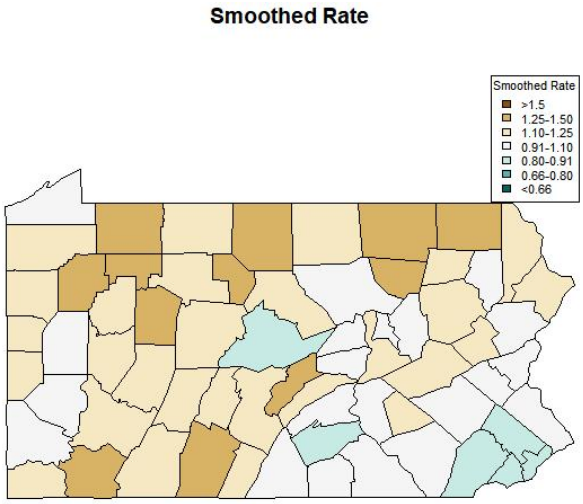
$$Y_i|\beta_0, e_i \sim_{ind} Poisson(E_i e^{\beta_0} e^{e_i})$$

$$e_i|\sigma_e^2 \sim_{iid} N(0, \sigma_e^2)$$

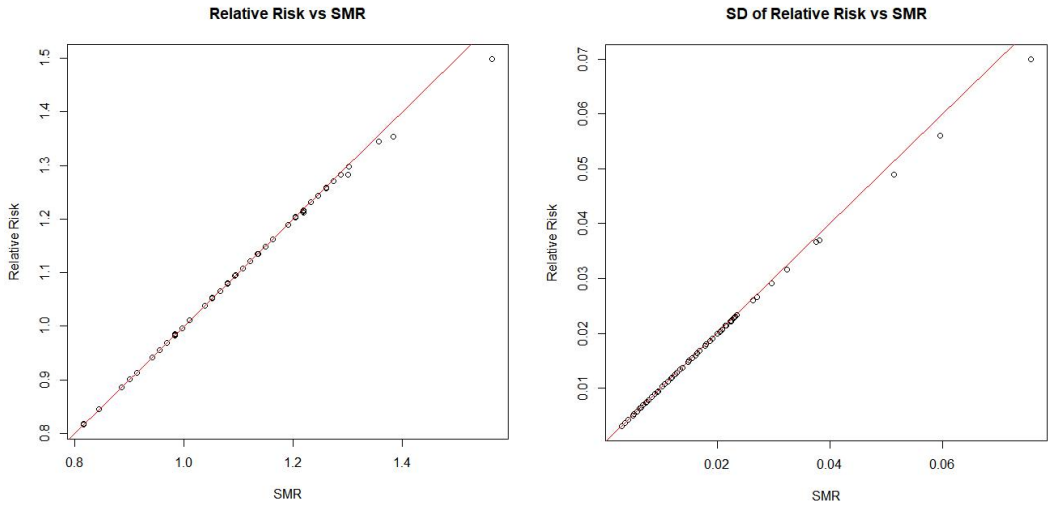
where e_i are area-specific random effects for residuals or unexplained relative risks of the CHD. We use the INLA in R to fit the model, and show the posterior medians of the estimates and the 95% confidence intervals as below.

	Median	**CI_lower**	**CI_upper**
$\boldsymbol{\beta_0}$	0.11	0.08	0.14
$\boldsymbol{\sigma_e}$	0.13	0.11	0.15

Specifically, we extract the posterior medians of the relative risks and map them as below.



We can see that, the relative risks are very similar to the SMRs, though we can still see evidence of smoothing as the only county in the SMRs map that is larger than 1.5 is now lower than 1.5. To check the exact relations, we plot the relative risks versus the SMRs and the standard errors of them as below.



We can see that when SMRs are small, the relative risks are very similar to the SMRs, and the extreme large SMRs show evidence of shrinkage as the relative risks are smaller and have smaller standard errors. The plots make sense since the standard errors of SMRs are quite small, and when SMRs are small, the standard errors are even smaller, thus the smoothing when SMRs are small is not very obvious.

The Poisson-Lognormal non-spatial model allows area-specific random effects, but can't deal with spatial correlations. Furthermore, we now fit a Poisson-Lognormal spatial model with a penalized complexity prior to take the spatial dependence into consideration.

$$Y_i|\beta_0, S_i, \epsilon_i \sim_{iid} Poisson(E_i e^{\beta_0} e^{S_i+\epsilon_i})$$

$$\epsilon_i|\sigma_\epsilon^2 \sim_{iid} N(0, \sigma_\epsilon^2)$$

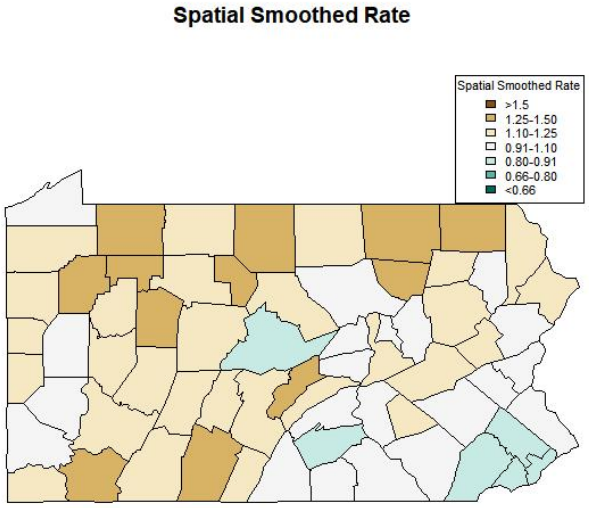
$$S_1, \ldots, S_n \sim ICAR(\sigma_s^2)$$

We fit the model with INLA and report the posterior medians and 95% intervals for β_0, the total variance of the random effects, and the proportion of the total variance attributed to the spatial random effect as below.

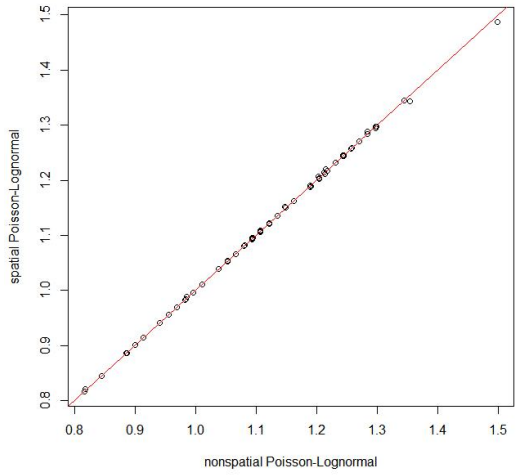
	Median	CI_lower	CI_upper
β_0	0.11	0.09	0.13
total_var	0.01	0.01	0.02
prop_spatial	0.60	0.28	0.88

We can see that the posterior median for β_0 is the same as the non-spatial smoothing model, but the 95% confidence interval is narrower now. We can also see that the total variance of the random effects is quite small, and the the proportion attributed to spatial random effects has a very wide range.

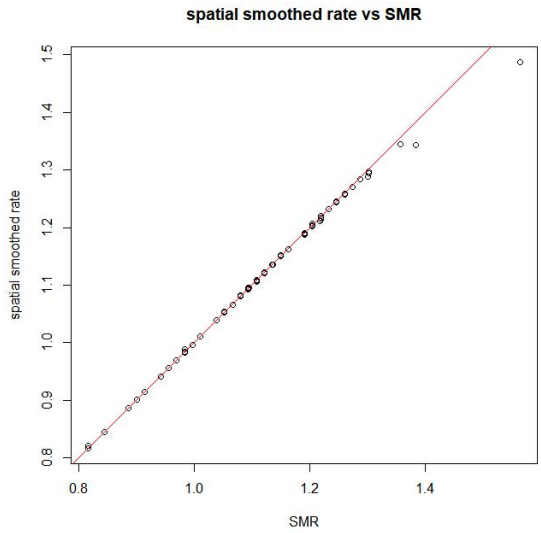
Now we map the spatial smoothed rate as below and compare it with the non-spatial smoothing method.



We can see that the map is quite similar to the non-spatial smoothing method. We plot the relative risks of the spatial versus the non-spatial Poisson-Lognormal models as below.



We can see that the estimates of the two models are indeed very similar. Next we plot the smoothed rate of the Poisson-Lognormal spatial smoothing model versus the SMRs as below.



We can see that similar to the non-spatial mode, the smoothed rates are very similar to the SMRs when SMRs are small, and when the SMRs are extreme, the spatial smoothing method shows shrinkage.

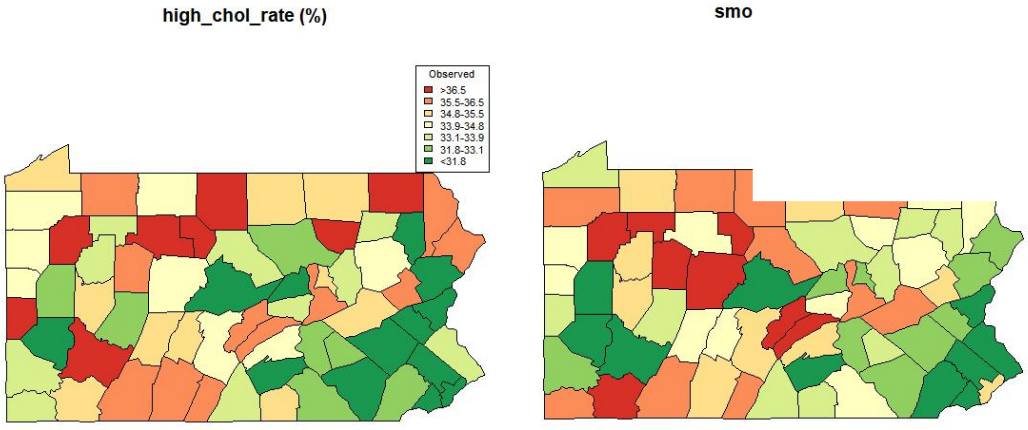
To sum up, we can see that combing the maps of SMRs, non-spatial smoothed rate, and spatial smoothed rate of CHD, we find that most of the counties of the western and northern part of Pennsylvania have morbidity rate higher than 1, which can be interpreted as the excess morbidity, and in the southeastern part of Pennsylvania, the counties have morbidity rates close to or lower than 1. The maps also show evidence of spatial dependence, as the morbidity rates are similar in many adjacent counties. We can also see that the overall variance of the morbidity rates of CHD in Pennsylvania is small.

3.2 Association of CHD and other variables

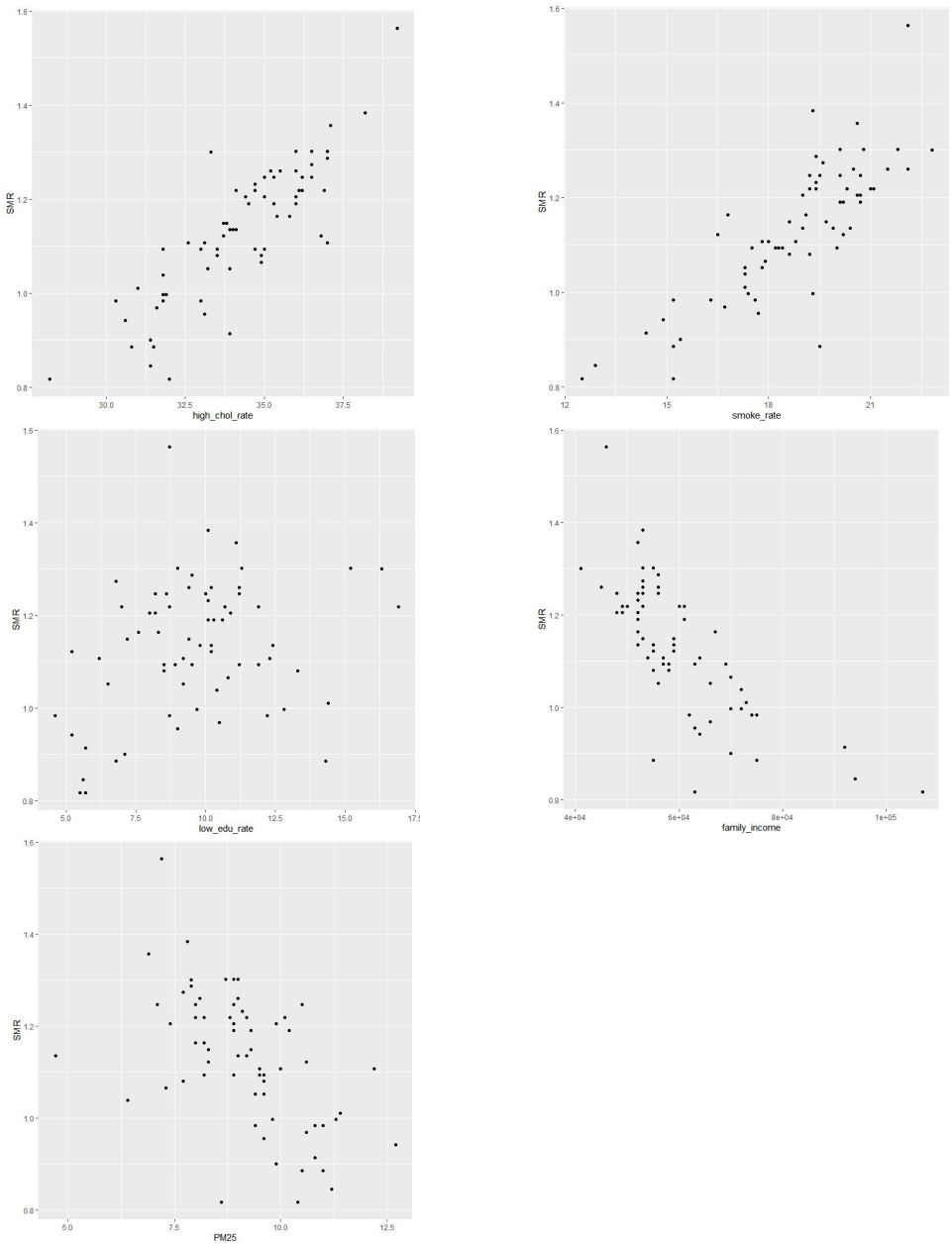
In this part, we are going to explore the associations of the CHD morbidity with a few risk factors and social and natural factors. We are going to consider 5 variables that are possibly correlated to the CHD morbidity:

high_chol_rate: Rate of high cholesterol level among 18+ adults (%)
smoke_rate: Rate of smokers among 18+ adults (%)
low_edu_rate: Rate of education level lower than high school (%)
family_income: The median of the family income in the county ($)
PM25: The PM2.5 level in the county

We map the first two of the variables to have a brief understanding of the distributions and variance of the variables.



We can see that these two maps are a little similar to the SMRs map, so maybe they are associated with CHD. Now, in order to understand the relationships of the variables with the CHD morbidity, we plot the variables versus the SMRs of CHD as below.



From the plots, we can see that high cholesterol rate and smoking rate have obvious positive relation with SMR of CHD, and family income has a obvious negative relation with SMR. PM2.5 seems to have a negative relation with SMR, though the relation is now very obvious. There seems to be no relation of low education rate with SMR of CHD.

Now, we fit the non-spatial Poisson-Lognormal models with a covariate to assess the association between the variables and CHD.

$$Y_i|\theta_i \sim_{ind} Poisson(E_i\theta_i)$$

$$\log \theta_i = \beta_0 + x_i\beta_1 + e_i$$

$$e_i|\sigma_e^2 \sim_{iid} N(0, \sigma_e^2)$$

where x_i is one of the covariates that we are interested in. We also assume independent priors of the hyperparameters β_0, β_1, and σ_e^2. We fit the model for every variables we discussed before and report the exponential of the medians and 95% confidence intervals of the estimates for β_0 and β_1 and the posterior medians of the 95% residual relative risk (RRR) intervals for σ_e.

x_i: high cholesterol rate

	Median	**CI_lower**	**CI_upper**
Exp(β_0)	0.1880298	0.1425533	0.2479692
Exp(β_1)	1.0534663	1.0449837	1.0620278
Median of RRR	-	0.8753691	1.1423753

x_i: smoke rate

	Median	**CI_lower**	**CI_upper**
Exp(β_0)	0.4323441	0.3724166	0.5015612
Exp(β_1)	1.0520163	1.0437356	1.0604120
Median of RRR	-	0.8753393	1.1424142

x_i: low education rate

	Median	**CI_lower**	**CI_upper**
Exp(β_0)	0.9585472	0.8550977	1.074789
Exp(β_1)	1.0160819	1.0044739	1.027815
Median of RRR	-	0.7871204	1.2704537

x_i: family income

	Median	**CI_lower**	**CI_upper**
Exp(β_0)	1.8699693	1.6775140	2.0864939
Exp(β_1)	0.9999914	0.9999896	0.9999932
Median of RRR	-	0.8508309	1.1753216

x_i: PM2.5

	Median	**CI_lower**	**CI_upper**
Exp(β_0)	1.7534808	1.4664870	2.0982763
Exp(β_1)	0.9519133	0.9336388	0.9704771
Median of RRR	-	0.8067266	1.2395773

From the regression results, we can see that as we expected, the medians and 95% intervals of exp(β_1) in the high cholesterol rate case and smoke rate case show that these two variables are positively associated with the CHD morbidity, and these associations are significant, which makes sense as these two variables are marked as risk factors for cardiovascular diseases according to CDC. These two variables also have the narrowest 95% RRR intervals for σ_e. Low education rate seems to have a little positive association with CHD morbidity, but the association is not obvious and not significant. The median and the 95% confidence interval of the exp(β_1) in the family income case are very close to 1, which indicates that family income has no ecological association with CHD morbidity. PM2.5 level has a significant negative association with CHD morbidity, although this case has the widest 95% confidence interval of RRR.

Then, we consider the spatial dependence of the data and fit a Poisson-Lognormal spatial model:

$$Y_i|\theta_i \sim_{iid} Poisson(E_i\theta_i)$$

$$\log \theta_i = \beta_0 + x_i\beta_1 + S_i + \epsilon_i$$
$$\epsilon_i|\sigma_\epsilon^2 \sim_{iid} N(0, \sigma_\epsilon^2)$$
$$S_1, \ldots, S_n \sim ICAR(\sigma_s^2)$$

where x_i is one of the covariates that we are interested in. We also assume independent priors of the hyperparameters. We fit the model for every variables and output the median and 95% confidence intervals for exp(β_0), exp(β_1), and other hyperparameters. The results are shown below.

x_i: high cholesterol rate

	Median	**CI_lower**	**CI_upper**
Exp(β_0)	0.258818	0.1951586	0.3387064
Exp(β_1)	1.043681	1.0355036	1.0523322
Precision for Region	211.5706	135.8028	319.11031
Phi for Region	0.6301267	0.2299249	0.9264561

x_i: smoke rate

	Median	**CI_lower**	**CI_upper**
Exp(β_0)	0.4802713	0.4134967	0.5570946
Exp(β_1)	1.0461168	1.0378638	1.0545189
Precision for Region	219.1662	139.1243	332.9024
Phi for Region	0.6646124	0.2574406	0.9426822

x_i: low education rate

	Median	**CI_lower**	**CI_upper**
Exp(β_0)	0.9158906	0.840493	0.9991204
Exp(β_1)	1.0209134	1.011844	1.0299783
Precision for Region	92.0913	60.4009	135.7923
Phi for Region	0.7000645	0.379292	0.9239445

x_i: family income

	Median	**CI_lower**	**CI_upper**
Exp(β_0)	1.7904013	1.5487662	2.0466339
Exp(β_1)	0.9999921	0.9999899	0.9999945
Precision for Region	137.4178	93.34950	200.2773
Phi for Region	0.1681264	0.01372456	0.6385386

x_i: PM2.5

	Median	**CI_lower**	**CI_upper**
Exp(β_0)	1.5106864	1.2089481	1.8697599
Exp(β_1)	0.9675637	0.9453907	0.9913537
Precision for Region	84.9924	56.9520	124.5446
Phi for Region	0.4196197	0.1213689	0.79455

From the results of the Poisson-Lognormal spatial model, we can see that the estimates are similar to the non-spatial Poisson-Lognormal model. High cholesterol rate and smoke rate have positive association with CHD morbidity, though the associations are not as significant as the non-spatial model. Low education rate in this model has a more significant positive association with CHD, though the association is still not very obvious. Family income has no association with CHD, which is same as before. PM2.5 has

significant negative association with CHD, though the association is not as significant as the non-spatial model.

We can see that the spatial model seems to smooth the associations between the risk variables and CHD morbidity, maybe because the excess association in the non-spatial model is caused by spatial dependence.

4 Discussion

In this project, we explored the Coronary Heart Disease (CHD) data among 18+ adults in Pennsylvania in 2020. After we map the disease data, we found that the western and northern part of Pennsylvania tends to have higher CHD morbidity, and the southeastern part of Pennsylvania has relatively lower CHD prevalence rate. By fitting non-spatial and spatial Poisson-Lognormal models, we smoothed the relative risk and obtained more accurate spatial distribution of the CHD data.

By analyzing 5 relative risk factors, we found that high cholesterol rate and smoke rate are positively associated with CHD morbidity, and PM2.5 level are negatively associated with CHD. The spatial Poisson-Lognormal model shows that low education rate has a positive association with CHD, but the association is not as obvious in the non-spatial model. Family income seems to have no association with CHD. By comparing the results of the two models, the spatial Poisson-Lognormal model seems can smooth the associations obtained from the non-spatial Poisson-Lognormal model, maybe because the excess association in the non-spatial model is caused by spatial dependence.

The project has limitations, such as that there might be associations between the risk factors, so we might need to fit a regression model containing all of the risk factors to explore their relationships with CHD morbidity as well as the interaction between these factors. Also, the prior values for the hyperparameters in the spatial Poisson-Lognormal model might not be optimal, so we might need to conduct more analysis to obtain better values. We can also spend more time on analyzing the spatial dependence of the data and the variances of the estimates in the Poisson-Lognormal models.

5 Reference

[1] Wakefield, J. C., Best, N. G., and Waller, L. A. (2000). Bayesian approaches to disease mapping. In P. Elliott, J. C. Wakefield, N. G. Best, and D. Briggs, editors, Spatial Epidemiology: Methods and Applications, pages 104 – 27. Oxford University Press, Oxford.

[2] Wakefield, J. (2008). Ecologic studies revisited. Annual Review of Public Health, 29, 75–90.

[3] Banerjee, S., Carlin, B., and Gelfand, A. (2015). Hierarchical Modeling and Analysis for Spatial Data, Second Edition. CRC Press.

[4] Centers for Disease Control and Prevention. Coronary Artery Disease (CAD). https://www.cdc.gov/heartdisease/coronary_ad.htm.

[5] Centers for Disease Control and Prevention. Interactive Atlas of Heart Disease and Stroke. https://nccd.cdc.gov/DHDSPAtlas/.

[6] Wakefield, J. Data of the neighbor relationships of the counties in Pennsylvania. https://faculty.washington.edu/jonno/SISMIDmaterial/.