Data Analysis for Opioids Drug Misuse in California 2019

Caffeinated Analysts: Alexander Shum, Jonas Lee, Minsung Kim, Zi Peng Liu

1. Objective/Preliminary

In this project we aimed to investigate and do inference on the prevalence of Opioids misuse within the state of California in 2019. Within the given dataset, a participant in the survey is considered to have misuse Opioids is when the participant have used Opioids not under the direction of any healthcare providers at any moment of the participant's lifetime.

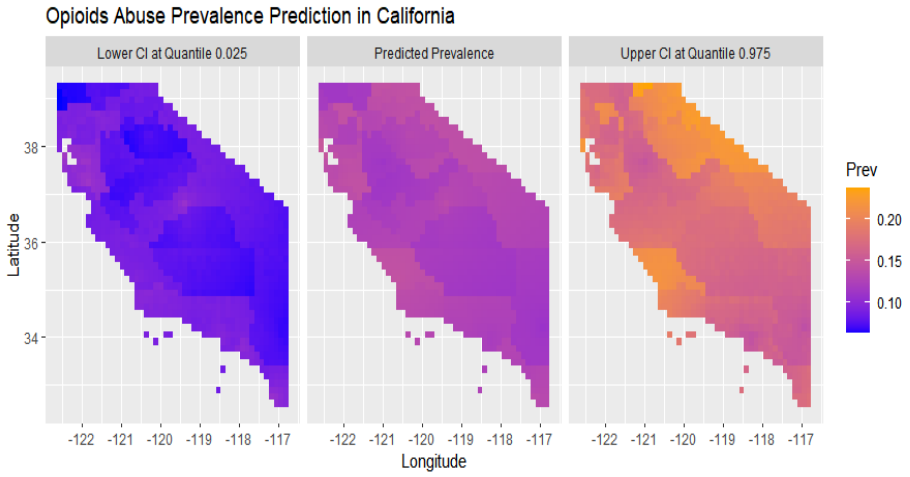
We attempted to model the overall Californian Opioids misuse cases under two approaches: a spatially dependent bayesian approach using the R-package INLA, and a spatially independent frequentist approach.

2. Modelling / Analysis

Spatially Dependent Logistic Regression

To fit the spatially dependent logistic model, the data set is first being restructured. We begin by finding the unique postal codes within California and computing the prevalence for each unqiue postal code by taking the sum of all opioids misuse in that postal area divided by the total amount of data in that postal area. We thereafter use Google geocoding API and the "GADM" geospatial dataset to find the longitude and latitude, and the county each postal code belongs to respectively. This is than use in conjunction with US county demographic dataset to obtain the demographic information to fit the following model.

$$\text{logit}(P(\mathbf{s}_i)) = d(\mathbf{s}_i)'\beta + Z(\mathbf{s}_i)$$



Spatially Independent Logistic Regression

Upon examining the data, we proceeded to select variables for our model. Several plots were constructed to measure patterns of non-medical use of Opioids and as result seventeen variables were selected. (examining the participants' demographic backgrounds and various health conditions.)

Initially a cursory approach was taken and a logistic regression "default" model was constructed. A summary of the model demonstrated that, there existed predictors that were not significant to the model.

So another model was constructed using forward and backward AIC method, and a likelihood ratio test was conducted to compare the two models. However, the likelihood ratio test revealed the alternative model to be not any more useful than the original "default" model and thus the "default" model was chosen for analysis in the end.

3. Conclusion and Limitations

The analysis of the logistic regression model has shown that the participants' overall involvement with drugs (legal or illegal) as well as pre-existing health conditions have the greatest impact on chances of Opioid misuse. Some of the notable findings include increases in Opioid misuse probabilities given alcohol/cigarette consumption as well as past history of mental illness.

It is important to keep in mind that interactions between variables and the possible presence of random effects were not considered due to time restraints.

$\mathbf{s}_i$ is an arbitrary point within California
$d(\mathbf{s}_i) = (1, d_1, ..., d_p)'$ is the vector of covariates with values basing on county the points belong to
β is the parameter of the fixed effect
$Z(\cdot)$ is the spatially structured random effect which follows a zero-mean Gaussian process with Matérn covariance function