STAT431 Project Proposal

Gan Yao and Tommy Tang

Nov. 3 2022

1 Data

We plan to base our analysis on a data set created and published by Schnell and Papadogeorgou [1]. This data set consists of CVD(cardiovascular disease) mortality, super market availability and demographic data on 3093 counties. And the problem that we intend to investigate is: how is CVD mortality related to county-level supermarket availability?

To be specific, the response Y is,

Name	Discription
Deaths	Deaths caused by CVD

The treatment Z, i.e. the variable of our particular interest,

Name	Description
PCT_HHNV1MI	Percent of households with no vehicle and more than 1 mile from a supermarket or large grocery store

Other covariates X include,

Name	Discription
PctUrban	Percentage of population in urban areas
PctWhite	Percentage of white population
PctBlack	Percentage of black population
PctHisp	Percentage of hispanic population
PctHighSchool	Percentage of population that attended high school
MedianHHIncy	Median household income (×1000 USD)
PctPoor	Percentage of impoverished population
PctFemale	Percentage of female population
PctMovedIn5	Percentage of population having lived in area for less than 5 years
PctOccupied	Percentage of housing units that are occupied
MedianHValuey	Median value of owner occupied housing (×1000 USD)
PopPerSQMy	Population per square mile
TotPopy	Total county population (×1000)
smokerate	Percentage of population that smokes

2 Main Model

We model the response with a Poisson regression model (with a log link function). That is, for a fixed county i, let Y_i represent the number of deaths in that county, P_i the population size of the county, Z_i represent the percentage of residents without convenient supermarket access, as detailed in Section 1, X_{ij} for $j = 1, .., J$ represent each of the other covariates associated to the county, we model the number of deaths as:

$$Y_i|X_i, Z_i \sim \text{Poisson}(P_i \exp\{\beta_z Z_i + \sum_{j=1}^{J} X_{ij}\beta_j\})$$

We will need to consider various priors for the β_j and most notably β_z. We may consider:

$$\beta_j, \beta_z \sim N(0, \sigma^2)$$

We might also reasonably expect that many of the β_j are 0 (especially conditional on other $X_{j'}$) and will also explore the use of a double exponential prior for the β_j but not for β_z. We may apply non-informative priors on the parameters for these priors.

Of particular interest (to, for example, public health policymakers) is the quantity β_z associated to the effect of (a lack of) supermarket access. As this represents the log proportional change in expected deaths for a one-unit change supermarket access.

Our goal is thus to obtain a model that will make accurate predictions for a "hidden" portion of the data and estimate β_z, the "effect" of a lack of supermarket access.

3 Plans for Analysis

3.1 Posterior inference

We will use MCMC to perform inference on β_z. We will separate roughly 20 percent of our data to be used as a test set, and validate our GLM by reporting test accuracy.

From our validated GLM we will perform inferences on the posterior mean of β_z and obtain credible intervals as well.

3.2 Sensitivity Analysis

As previously discussed, we have two potential classes of priors for the predictor values associated to our covariates: normal prior and double exponential prior.

We cannot expect "conjugate priors" or even closed form expressions for posterior distributions, as we are using a GLM. Thus we rely on MCMC and our choice of priors are based on encoding our domain knowledge - for example, a double exponential prior encodes our belief that many of the covariates will not influence the response, while a few will have a nonzero relationship.

Both of these choices of priors have parameters that allow for hierarchical modeling. We will perform modeling on both priors and compare the posterior results, and note the sensitivity of the results to our choice of prior.

Our current proposal is only a tentative set of prior distributions. We may test other priors as well (e.g., Jeffreys' prior).

References

[1] Patrick M Schnell and Georgia Papadogeorgou. "Mitigating unobserved spatial confounding when estimating the effect of supermarket access on cardiovascular disease deaths". In: *The Annals of Applied Statistics* 14.4 (2020), pp. 2069–2095.