BST 6200 Spatial Statistics and Disease Mapping

Disease Models

Steven E. Rigdon

Spring 2020

Motivating Example

Sudden Infant Death Syndrome in North Carolina, 1979

County	BIR79	SID79
Alamance	5767	11
Alexander	1683	2
Alleghany	542	3
⋮	⋮	⋮

Method 1: Raw Rates

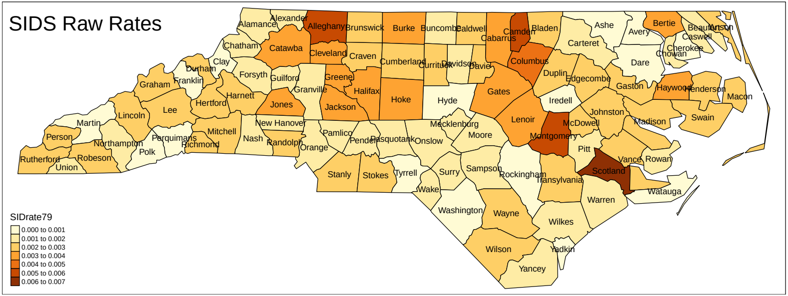
Y_i = number of cases in county i with population n_4.

$Y_i \sim \text{POI}(n_i \eta_i)$ where η_i is the risk in county i.

Assume $Y_1, Y_2, \ldots, Y_k$ are independent.

Under these assumptions, $\hat{\eta}_i = \frac{Y_i}{n_i}$. These are called the raw rates.

Choropleth Map of Raw Rates



Method 2: Hierarchical Model

The risks $\eta_1, \eta_2, \ldots, \eta_k$ are thought of as being selected from some distribution. (Parameters of this distribution are unknown.)

The technical term is that $\eta_1, \eta_2, \ldots, \eta_k$ are *exchangeable*.

This implies that the risks cannot be too different and results in smoothing of the risk estimates.

More Specific Model

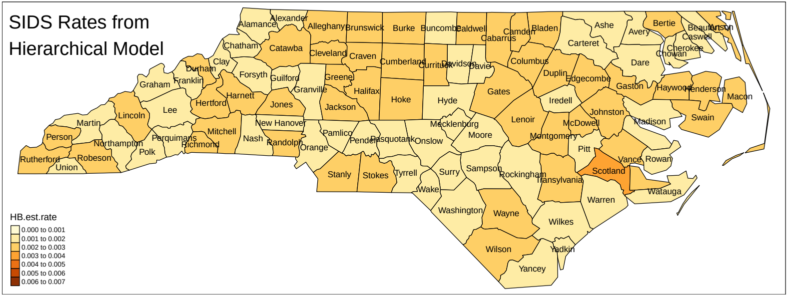
Let $\theta_i = \log \eta_i$.

Assume $\theta_1, \theta_2, \ldots, \theta_k \sim \mathrm{N}(\mu, \sigma^2)$ where μ and σ^2 are unknown parameters, called hyperparameters.

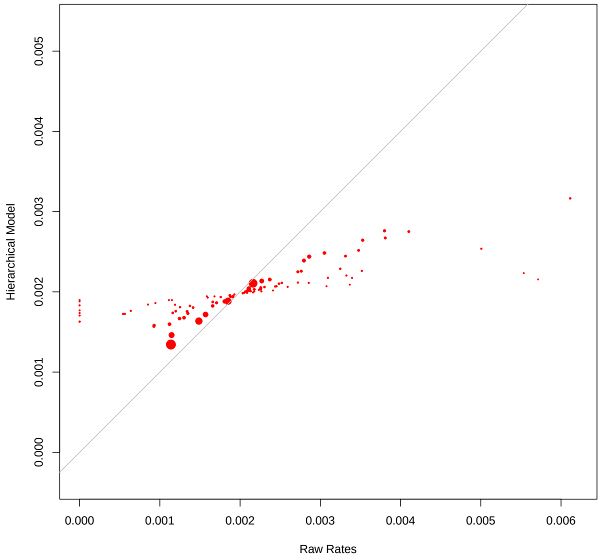
We will estimate μ and σ^2 using the Bayesian approach. (More on this to come.)

This approach ignores spatial information about the counties.

Hierarchical Bayes Model



Comparison of Raw Rates vs. Hierarchical Model



Method 3: Conditional Autoregressive (CAR) Model

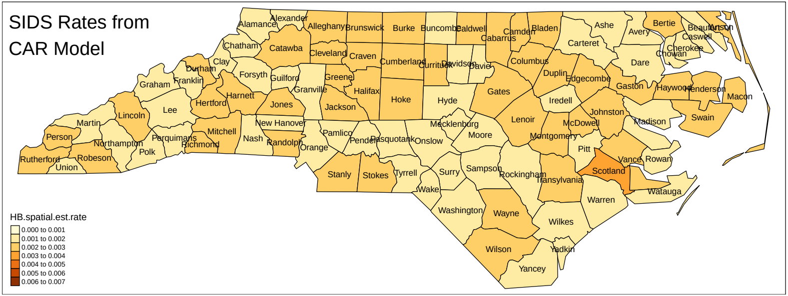
Similar to the hierarchical Bayes model, except the neighborhood structure of the counties is taken into account.

Assume

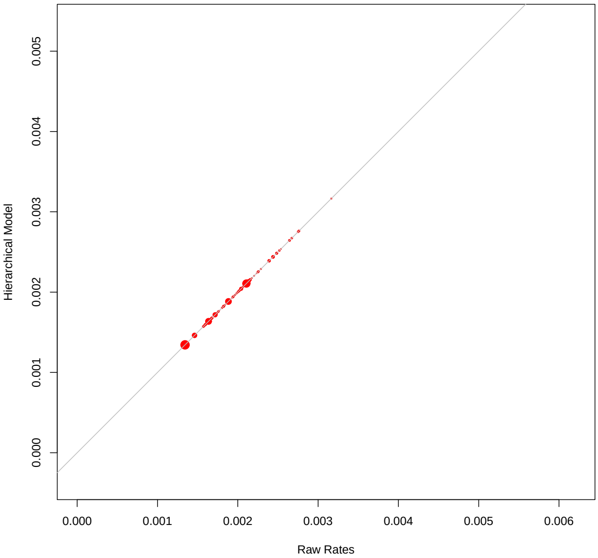
$$\theta_i = \underbrace{\mu}_{\text{overall mean}} + \underbrace{v_i}_{\text{uncorrelated heterogeneity}} + \underbrace{u_i}_{\text{correlated heterogeneity}}$$

Neighbor Structure

Choropleth Map for CAR Model



Hierarchical vs. CAR



Regression within CAR Model

$$\theta_i = \underbrace{\beta_0}_{\text{overall mean}} + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}_{\text{regression terms}} + \underbrace{v_i}_{\text{uncorrelated heterogeneity}} + \underbrace{u_i}_{\text{correlated heterogeneity}}$$

Example: Pennsylvania Lung Cancer Data

Possible model:

$$\theta_i = \beta_0 + \beta_1(\text{Smoking Rate})_i + \beta_2(\text{\% Nonwhite})_i + \beta_3(\text{Average Income})_i + v_i + u_i$$

Identifying a Project Topic

1. Identify a state and an outcome variable (usually a disease count)
2. Obtain the shape file for the state (via the `tigris` package)
3. Identify possible predictor variables.
4. Apply the CAR model to estimate parameters including disease rates

You may need two, three, or more sources for data. Shape file comes from the `tigris` package. Outcome variable comes from another source. Predictor variables come from one or more additional sources.

Plan for the remainder of Spring 2020

1. Bayesian inference
2. Simulation to perform Bayesian inference
3. R package `nimble` to do simulations
4. Using `nimble` to run the CAR model