Project Proposal

Gifty Osei

2024-10-28

Unveiling Temporal Health Patterns: A Bayesian MCMC and Clustering Approach to Health Indicators

Motivation

Health disparities between demographics and geographic areas remain a major public health concern. Understanding how different health indicators change over time by gender, race/ethnicity, and geography is critical for making informed policy decisions and effectively directing interventions. The goal of this project is to examine a large dataset comprising a variety of health indicators in order to discover underlying patterns, trends, and clusters. We use advanced statistical computing approaches such as Bayesian modeling with Markov Chain Monte Carlo (MCMC), longitudinal analysis, and clustering techniques to provide insights into the dynamics of health indicators and overcome difficulties such as missing data. The ultimate goal is to allow the data to disclose intrinsic structures and relationships, which will contribute to a better understanding of public health dynamics.

Questions:

In this project we will explore and answer the following questions;

- How does a Bayesian modeling technique using MCMC improve knowledge on the rate of occurrence for some health indicators considering demographic conditions?
- What are the temporal patterns or trends in health indicators throughout different years?
- How do health indicators differ based on gender and race/ethnicity?
- Can we discover clusters of areas with similar health indicators?
- How can we manage missing information in the 'Gender' category?

Dataset Description

The dataset consists of records of numerous health indicators that were gathered in many places over several years. A description of each variable is as follows;

- Indicator Category: The broad category that a health indicator belongs under, such as cancer, substance abuse, and HIV/AIDS.
- Indicator: The particular health metrics or results (e.g., Lung and heart cancer, Infant mortality rate, heart disease).

- Year: The year when the health indicator was measured or the data was gathered.
- Gender: The population group's gender (either male, female, or Both).
- Race and Ethnicity: The population's race or ethnicity (e.g., White, Black, Hispanic).
- Value: The dependent variable in our study, which is the health indicator's rate per 100,000 persons.
- Place (Location): The geographical area (in this case large cities in the US) where the data was gathered.
- BCHC Requested Methodology: The Big Cities Health Coalition's (BCHC) specific methods requested for gathering or analyzing the data.
- Source: Where the data came from (e.g., Census, hospital records, National Center for Health Statistics (NCHS) and CDC reports).
- Methods: The techniques employed to gather or calculate the values of the health indicators.
- Notes: Extra details or observations on the data points.

Statistical Methods

With "Value" as the dependent variable, we will use a number of sophisticated statistical computing techniques to answer the research questions and accomplish the project's goals. Methods like;

- Exploratory Data Analysis
- Descriptive Statistical Summary and Visualization (Histogram and Bar plots, Correlation Analysis)
- Bayesian Model with Markov Chain Monte Carlo (MCMC): Utilize MCMC techniques to sample from the posterior distributions, such as Metropolis-Hastings and Gibbs sampling..
- Longitudinal Analysis (Dynamics Across Year) : examines how the "Value" changes over time within the same entities (e.g., locations).
- Cluster Analysis: By using clustering analysis, data points are grouped so that they are more comparable to one another inside the same cluster than to those in other clusters.
- EM Algorithm for missing Data:

This project will allow to leverage a variety of statistical computing methods, offering a solid framework for analysis and valuable insights into some health indicator rate among large cities in the US.

Explore interesting structure :

This shows that analysis on the entire data will be challenging, so do in groups