Project Proposal

Jessica Moore

3/25/2020

Project background and premise

One indicator of a nation's health and development is infant mortality (Reidpath 2003). In 2017, the infant mortality rate in the United States was 5.80 deaths of a child under one year old per 1,000 live births. This level of infant mortality is comparable to countries with lower GDP and higher inequality, and concerningly higher than countries with similar demographics like Canada and the European Union, whose rates of infant mortality are 4.50 and 4.00 respectively (CIA World Factbook for 2017).

Within the United States, individual state's rates of infant mortality vary dramatically. Mississippi claims the highest likelihood of a child dying before their first birthday, at a rate of 8.6, while Massachusetts less than half of that frequency: a rate of 3.7 (CDC Stats of the States 2017). While each state is faced with unique risk factors for infant mortality and barriers to ameliorating the public health problem at hand, careful analysis of a single state may offer insight into drivers of infant mortality. Texas is a state with an infant mortality rate near the national mean (5.9 compared to the mean of 5.8, CDC) and, due to it's population size, a large number of observations upon which to conduct an analysis.

Data sources and compilation

To investigate factors which may relate to rates of infant mortality in Texas, this project aims to analyze data from three sources:

- Content from the Texas Department of State Health Services (DSHS), which publishes many measures of public health (including rate and cause of infant mortality) through its Center for Health Statistics
- The CDC's National Survey of Ambulatory Surgery
- Content from the Correlates of State Policy Project at IPSSR, which compiles state-by-state information about policies relevant to this analysis, including those related to inequality and health care

These three sources will compose a body of information including rates of infant mortality, measures of access to care in both urban and rural settings, other public health and hospital associated risk measures, and state-specific policy and demographic information.

Compiling the data into a uniform, usable format will require significant data wrangling due to the different origins, arrangements, and units of analysis comprising each component. The Texas DSHS data will be the most complex to obtain and will involve creating loops to scrape data directly from the website for some pieces of information, extracting pieces of data from other pre-available data files, and compiling data from multiple places into one data frame. The correlates of state policy project offers data in user friendly file format but will require extracting a subset of data from a large file for only Texas across multiple years. The CDC's NSAS data is simplest; it is available in a .csv file but will require modifying the unit of analyses, extracting the relevant information, and combining with the other pieces of data.

Project methods

For this project I will use fundamental data wrangling techniques to compile and arrange data from multiple sources. This code will rely on packages from the tidyverse, particularly reader, dplyr, and purr. Additionally, in order to visualize some descriptive representations of the data, I will use ggplot2 package code to (hopefully) visualize spatial data at the county level for the state, and basic relationships between some established indicator variables. Most critically, I will base my presentation and report on analysis using machine learning methods to develop a code for supervised learning to generate a quantitative model for factors which are associated with the rate of infant mortality at the state level in Texas.

Project success

This project can be considered "complete" if it fulfills the following:

- Finding a meaningful way to reconcile and represent the data from three different sources
- Successfully acquiring a machine learning result for the compiled data.

This model may be inconclusive or contradict the literature review and hypothesis; I would consider it a success (or a "finished" product) to achieve an output and effective representation, visually and statistically, of the compiled data.

The overall goal of this project is to demonstrate a successful reconciling of content from multiple sources in a way which offers meaningful data in the context of the literature and hypothesis, and to analyze this data in a way which adheres to best practices for statistical rigor and appropriate data visualization. The ultimate goal is that an illuminating model will emerge from the machine learning portion which offers new insight into factors related to the public health problem of high infant mortality in the United States.

Works Cited

Reidpath, D. D. (2003). Infant mortality rate as an indicator of population health. Journal of Epidemiology & Community Health, 57(5), 344–346. doi: 10.1136/jech.57.5.344

"Country Comparison: Infant Mortality Rate." CIA World Factbook. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/rankorder/2091rank.html

"Stats of the States: Infant Mortality Rates by State." National Center for Health Statistics, CDC. Retrieved from https://www.cdc.gov/nchs/pressroom/sosmap/infant_mortality_rates/infant_mortality.htm