STATS 451 Final Project Proposal

Yuzhou Peng, Siyang Wu, Amy Sun

April 15, 2024

1 Problem of Interest & Goal

Birth rates critically influence population growth and structure, which is essential for demographic planning in areas like education and healthcare. This demographic indicator also affects economic development, as a youthful population requires investments in education and employment opportunities, while an aging population may necessitate a shift in economic strategies and healthcare services. Additionally, understanding birth rates aids in forecasting healthcare needs, particularly in maternal and child health, ensuring that appropriate resources are allocated to meet the demands of different population segments.

However, there are several states that due to some sort of limitation have less sufficient data related to birth rate, which lead to problems (eg. less informative) when making decisions out of the existing data. The birth rate in a state is not completely independent from each other due to the fact that they come from same country. We are going to propose a way to fully utilize the information in the dataset, and meanwhile conduct comparison between different regions across US (eg, west vs. east) during the process for additional insights.

In that case, we are going to divide the main goal into 2 section:

- We want to give evidence on whether there is significant differences of birth rate distribution among separated regions such as the West, the East and the South. Specifically, we want to compare two models: (1) A model that include all states in the US as a whole. (2) Separated models that apply to different divisions of the US.
- Given birth rate data during a period of time, say 2011-2016, we want to obtain tighter estimated confidence intervals for birth rates using inductive Bayesian models than the results from fitting a model once in one year.

2 Dataset

2.1 Dataset Basics

In this study we are going to work on two dataset:

1. Birth dataset from Center for Disease Control and prevention (CDC)
2. County Population Data Dictionary From National Cancer Institute.

2.2 Birth Data

The Birth data in the CDC WONDER database is collected through the National Vital Statistics System (NVSS), which compiles birth data from all birth certificates filed in the United States. This includes detailed information on birth and maternal demographics, collected through a standardized birth certificate form used across the United States. Hospitals, midwives, and other healthcare providers submit this information to state health departments, which in turn report the data to the federal government. The data is not a simple random sample. It is a comprehensive administrative

data collection that aims to capture all births in the United States. As such, it is a census of births rather than a sample. This means the data covers the entire population of interest (all births occurring in the U.S. within a given time frame) and is not subject to the same types of sampling errors as survey data might be.

2.3 County Population Data Dictionary

The data for the U.S. Census Bureau's Population and Housing Unit Estimates are collected using a variety of methods including administrative records and data from earlier censuses, specifically the most recent decennial census. This method is not a simple random sample but a comprehensive effort to include all individuals and housing units using established administrative sources to update and adjust the census counts annually.

2.4 Potential Defects

Both dataset are subject to the following points of defects:

- Reporting Errors: Misreporting or inconsistent reporting of data by the individuals filling out birth certificates can introduce inaccuracies.
- Data Entry Errors: Mistakes in the transcription of data from paper forms to digital databases.
- Missing Data: Not all fields on a birth certificate may be completed, leading to missing data issues.

2.5 Data structure

The two dataset are joined together on FIPs code in the form XXYYY, where XX = 2 digit state code and YYY = 3 digit county code. The finalized the dataset has the following structure shown in Table 1. The data contain 5619 data points containing all the available information about the county each year across the United State from 2011 to 2020.

Table 1: Structure of Processed Dataset

Column 1	Column 2	Column 3	Column 4	Column 5
County(name)	FIPs	Birth count	Year	Population

3 Method

3.1 Method Introduction

Given the dataset we have described in previous section, we have observed birth rates of each counties across all states of the US. We are going to model the underlying distribution of birth rates in each states according to the goal we want to achieve.

3.2 Notation

We define our notations that are to be used as follow:

- y_{ij} is the observed birth rates of county i in state j.
- ω_j, κ_j are the parameters of underlying distribution of birth rate in state j, which are our quantities of interest.
- ω, κ are the hyper parameters that follow a pre-designed prior distribution

3.3 Model Construction

To preserve the potential dependency structure within the dataset and to mitigate the effect of small sample size within some states, we apply Bayesian hierarchical model to estimate the distribution of birth rates in each states. The basic model structure is as follow:

$$y_{ij} \sim Beta(\omega_j(\kappa_j - 2) + 1, (1 - \omega_j)(\kappa_j - 2) + 1);$$
$$\omega_j \sim Beta(\omega(\kappa - 2) + 1, (1 - \omega)(\kappa - 2) + 1) \text{ and } \kappa_j \sim Gamma(S, R)$$
$$\omega \sim Beta(A, B) \text{ and } \kappa \sim Gamma(S, R)$$

Where A,B,S and R are pre-designed parameters.
Specifically, to study the two proposed problems, we design two slightly different set of models respectively.

3.3.1 Inductive Bayesian hierarchical model

To model the data from 2011 to 2016, instead of build the model separately, we want to use the inductive property of Bayesian model. In the 1st year, we assume the prior distribution of hyper parameters to be non-informative and generic vague. After that, The prior distribution of hyper parameters in i-th year is designed to be the estimated marginal posterior distribution of hyper parameters obtained in (i-1)-th year.

3.3.2 Comparison between full model and partially independent model

In this section, we geographically classified states in USA into several categories, say, West, East Coastal, South, etc. We want to study if there is any difference when we separately apply Bayesian hierarchical model to different partition of the states in the US comparing with fitting all states in one model at one time.

4 Potential Difficulties

1. The estimation of marginal posterior of hyper parameters is difficult since we cannot have an analytic form of this distribution.
2. The dataset might contain Nah value, we need to take action to address the missing value and keep the integrity of the data set the same time.

5 Responsibility and Task Division

5.1 Group Leader/Convener:

Yuzhou Peng

5.2 Meeting plan:

We are going to meet twice a week, Tuesday and Friday.

5.3 Division

5.3.1 Yuzhou Peng:

- **Role**: Model design, model coding and debugging, post modeling analysis and inference from the results.
- **Contribution so far**: Based on the problems we are interested in, I designed an initial model structure and features, including quantities of interest and distribution for estimation.
- **Future plan**: Fit the initial models to our data, perform diagnosis on the model. Make adjustments based on diagnosis results. Evaluate fitting results for final inference.

5.3.2 Siyang Wu:

- **Role**: Data collection and analysis, model coding and debugging, post modeling analysis.
- **Contribution so far**: Collect and compile the dataset into a csv file, proposal construction.
- **Future plan**: Perform EDA on existing data to gain better insight, provide support with model construction. Provide support on MCMC simulation.

5.3.3 Amy Sun:

- **Role**: Supported the team to conduct research topic background research by reading research papers on the relevant topic and summarizing the interesting findings.
- **Contribution so far**: Built a strong foundation for our research and identified the motivation and importance of our study. Helping the team formalize and specifically organize the proposal.
- **Future plan**: plan to continue to revise and discuss the contents that other teammates generated during weekly meetings. Our next step is to prepare the upcoming presentation and the report.

6 Way Forward Schedule

Table 2: Project Time Line

Date	Task Description
4/15/2024	Finalize and clean dataset and start model building
4/19/2024	Finish Modeling coding and start performing MCMC simulation
4/20/2024	Prepare for the presentation
4/21/2024	Collect simulation result and perform analysis
4/25/2024	Confirm result start compose report
4/29/2024	Finish report and retrospect
4/30/2024	Report submission and enjoy summer break