STAT431 Final Project Proposal

Hantang Qin, Wangdi Chen, Nuoxing Shang, Yuqiao Jiang

Introduction

Obesity is a global health issue linked to numerous chronic diseases for human beings. Understanding the factors contributing to obesity, like eating habits and physical activities, is important for public health interventions and for the future.

We chose Option 1 for our project. This project is meant to explore obesity levels using a Bayesian hierarchical model, which focuses on individual eating habits and physical conditions as predictors for obesity level across different age groups, while considering uncertainties in individual behaviors.

Data Description

The dataset we selected includes data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. 77% of the data was generated synthetically using the Weka tool and the SMOTE filter, 23% of the data was collected directly from users through a web platform.

Some key variables in the dataset that we will mainly analyze with are:

- Age (Continuous variable indicating the person's age)
- NObeyesdad (Categorical variable describing obesity level)
- FAVC (Binary variable indicating whether the person eat high caloric food frequently)
- FAF (Continuous variable indicating how often the person has physical activity)
- CALC (Categorical variable indicating how often the person drinks alcohol)

To prepare the data for Bayesian hierarchical modeling, non-numerical variables will be transformed into numerical formats as follows:

- NObeyesdad: Encode the obesity levels (e.g., Insufficient Weight, Normal Weight, etc.) into ordinal numeric values (e.g., 1 to 7) based on severity.
- FAVC: Convert the binary variable into numerical values (e.g., 0 = No, 1 = Yes).
- CALC: Map the frequency categories (e.g., 'Never', 'Sometimes', 'Frequently', 'Always') to numerical values (e.g., 0, 1, 2, 3) while maintaining ordinal relationships.

Methodology

I. overview

The study will analyze patient-level data, focusing on individual attributes and lifestyle choices that may impact obesity levels. We will implement a Bayesian hierarchical model incorporating both individual-level factors and group-level factors. At individual level, there are lifestyle factors that vary per individual, such as high caloric food consumption, physical activity, and alcohol consumption; the group-level factor we will be using is age groups, allowing for estimates to vary across age, capturing the interaction between age and lifestyle factors.

II. Gibbs Sampling Implementation

Priors will be chosen based on domain knowledge and previous studies on obesity, with sensitivity analyses conducted to test the robustness of the results to prior specification.

We will implement Gibbs sampling to estimate the posterior distributions of the parameters. Gibbs sampling is a Markov Chain Monte Carlo (MCMC) technique that iteratively samples from the conditional distribution of each parameter, given the current values of all other parameters. Choose an initial value for each parameter, potentially using prior means or random sampling. For each iteration, update each parameter by sampling from its full conditional distribution, for example, sample from $p(FAVC|FAF,\ CALC,\ Age,\ NObeyesdad,\ data)$, then repeat this for each parameter, updating all others iteratively. Once we have a sufficient number of samples, we will use them to estimate the posterior distributions for each parameter, which will inform the relationship between lifestyle factors, age, and obesity.

III. Model Validation

We will use convergence diagnostics such as trace plots and the Gelman-Rubin statistic to ensure that the MCMC chains have reached a stationary distribution. Additionally, we will conduct sensitivity analyses to evaluate the robustness of our model to prior specifications. This step will help verify that our findings are stable across a range of plausible prior distributions.

IV. Posterior Inference

To conduct posterior inference, we will analyze the posterior distributions of the parameters related to lifestyle factors and age groups. The Bayesian framework will allow us to obtain credible intervals for each parameter, providing insights into the strength and direction of the impact of lifestyle factors on obesity across age groups. We will interpret these results by focusing on how lifestyle factors like caloric food intake, physical activity, and alcohol consumption influence obesity levels within different age brackets, considering uncertainty in individual behaviors.

Challenges and Limitations

1. Data Quality and Synthetic Nature: Since a portion of the dataset was generated synthetically using SMOTE, there may be limitations in capturing true patterns in real-world behaviors and obesity levels. This could affect the generalizability of our findings.
2. Potential for Unmeasured Confounding: While we are focusing on specific lifestyle factors, other unmeasured variables (such as genetic factors or socioeconomic status) may also impact obesity levels but are not included in this analysis, which could introduce bias.
3. Computational Complexity: Bayesian hierarchical models with MCMC can be computationally intensive, especially when dealing with large datasets or complex hierarchical structures. Convergence of MCMC may require substantial time and computing resources.

Reference

- Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5H31Z.