Stats 503 Project Proposal

Ruixuan Deng, Xinyu Zhang, Jingyan Zhang

March 10, 2024

1 NHANES Datasets Used

The datasets to be analyzed come from NHANES 2017-March 2020 Pre-pandemic data, which include:

- Questionnaire Data: Diabetes
- Demographic Data: Demographic variables and sample weights
- Dietary Data: Dietary Interview
- Examination Data: Body Measures

2 Questions to Explore

1. **Trend Analysis: How have certain health conditions (e.g., diabetes) trended over the years within different demographic groups?**

   This analysis can reveal disparities in health outcomes, identify high-risk populations, and guide targeted interventions. With diabetes rates rising globally, understanding these trends can inform strategies to mitigate risk factors and manage the disease more effectively across diverse populations.

2. **BMI Prediction: How do dietary habits affect BMIs?**

   For example, diets high in fruits, and low in processed foods and sugars, are often associated with better health outcomes, including lower BMIs. The insights derived from exploring this question can inform healthcare providers, and nutritional experts in developing guidelines and recommendations that promote healthier eating habits within populations.

3 Response Variables

1. **DIQ010** - Diabetes diagnosis from professonal perspectives
2. **BMXBMI** - Body Mass Index (kg/m**2)

4 Approaches

1. Random Forest: We will apply random forest to the classification of diabetes diagnosis data to get inferences from the dataset.
2. Cross Validation: We will fit a cross-validated linear regression model with lasso penalty on the BMI data set and predict this index level knowing dietary habits of a person.
3. LDA/QDA: We will partition the BMI scores into multiple levels, and fit classification models like LDA and QDA to make predictions.