STATS 506, Fall 2020

Final Project Proposal

Tianshi Wang

E-mail: wangts@umich.edu

In daily life, diet energy intake has become a crucial factor when judging whether someone maintains a healthy lifestyle. Meanwhile, the amount of household income is one of the most important criteria of household economic status. It might be a pretty interesting topic to explore the relationship between **income** and **diet energy intake**. In this project, we are going to explore the correlation between income and energy intake using NHANE 2017-2018 diet data. Additionally, we will also take some other demographic variables like **age**, **rurality**, and **gender** into consideration. By exploring the influence they play on our core relationship and the marginal effect might bring us some interesting findings.

- Question:

  > "Do people in the US from households with higher income consume more calories? "

- Datasets (NHANE)
  - NHANES 2017-2018 Dietary Data
  - NHANES 2017-2018 Demographics Data
- Variables:
  - id (SEQN)
  - energy intake(DR1KTCAL)
  - poverty income ratio (INDFMPIR)
  - gender(RIAGENDER)
  - age(RIDAGEYR), age_square
  - pergnancy (RIDEXPRG)
  - number of people in household (DMDHHSIZ)

    - number of people in family (DMDFMSIZ)

- Procedure:
  The whole project can be divided into 4 parts:
    - Data cleaning: Merge the dataset, select variables we want, and remove the NA values from our data.
    - Pre-analysis: Residual diagnosis, abnormal observation diagnosis, variable selection, collinearity detection.
    - Main analysis:
      There might be two possible approaches
        - Using **regression splines** to fit the non-linear model. For instance, some demographic variables like age might not have a uniform pattern on energy consumption (infants, teenagers, adults, and elders have a significant difference). We can use regression splines to divide the age variable into several intervals by using knots and combine different splines into a whole model. At the same time, we may try different variable combinations by adding or deleting variables. These might show the marginal effect on our problem.
        - Using **non-Gaussian GLM** consider using the **Gamma** link function. From the analysis we can see that the residuals seem to be non-uniform in variance. Meanwhile, the distribution of residual seem not come from a normal distribution. Taking consider of these factors, we may try to use a non-Gaussian Generalized Linear Model with link function (e.g. Gamma).
    - Summary

- Software/Tools:
    - **R** – tidyverse, data.table, ggplot2