STATS 506, Fall 2019

Group 4 Proposal

Huayu Li, Kai Liu, Xiru Lyu

According to Centers for Disease Control and Prevention, in 2015-2016, more than 20% of adults at the age of 20 and above in the United States have cholesterol levels higher than 240mg/ml. High cholesterol levels can increase risks of cardiovascular diseases and stroke, and thus controlling the level of blood cholesterol is critical in preventing certain life-threatening diseases. There are two types of cholesterol, and low-density lipoprotein (LDL) is known as the “bad” one that builds up the level of blood cholesterol. The goal of this project is to explore factors that could potentially affect the level of the “bad” cholesterol in a human body. More specifically, how does the level of LDL in blood cholesterol vary for people in the US of different demographic backgrounds, body measures and diet habits?

We will use data from NHANES 2015-2016 to perform the analysis. The dependent variable is set as the level of LDL, chosen from Cholesterol - LDL & Triglycerides from the laboratory data. Since a high level of triglycerides is believed to be linked with a higher chance for cardiovascular diseases, we consider to use it as one of our independent variables. We also include blood pressure readings obtained from Blood Pressure in the examination data, amounts of intake of fat and cholesterol from Total Nutrient Take (both First Day and Second Day) of the dietary data, and weight and height measures from the examination data as our independent variables. To account for more individual differences, we will consider age, gender, and race from the demographic data as covariates as well.

We will then use multiple linear regression, GLM and mixed models to analyze relationships between the level of LDL and independent variables specified above. For mixed models, we plan to treat gender and race as random effects. Depending on the overall fit, we may perform transformations for the dependent variable and/or certain independent variables. As some independent variables may be ineffective to the level of LDL, we will use forward and backward selections to select proper independent variables. Finally, we will perform model selection for all attempted models using AIC and BIC. Inferences are drawn from the model that comparatively achieves a high goodness of fit score without being overly complex.

The analysis will be carried out in both R, STATA and Python. In R, we plan to use data.table for data cleaning, and lme4 for fitting generalized linear models. In Python, package sklearn linear\_model module seems to be good for regression.