STATS 415 Group Project

Members: Nicky Li Alyssa Yang Kiley Price Amanda Li

In our report, we plan on using the NHANES 2013-2014 dataset. Specifically, we will be using subsets of data from the questionnaire, laboratory, and examination surveys to answer the following questions.

Question 1 (Classification):

- Is the structure of variability the same between the two classes of those who self-identify as depressed and those who do not?
- We will be classifying the response of "feeling down/depressed/hopeless" (DPQ020) by building a model with the following predictors:
    - Quantitative:
        - Average # of drinks/day (ALQ130)
        - Vitamin D level (LBXVIDMS)
        - Amount of sleep per night in hours (SLD010H)
        - Monthly family income (IND235)
        - BMI (BMXBMI)
    - Qualitative:
        - Worried about food (FSD032A)
- We will first split the dataset into different folds. We will create and analyze both LDA and QDA models to determine which one provides a better fit using their respective cross-validated MSEs and the corresponding confidence intervals of the MSEs.
- Tools Used: LDA/QDA, Cross-Validation

Question 2 (Regression):

- What is the effect of each of the predictors on Body Mass Index? I.e. which ones are the most influential when predicting BMI?
- We will be regressing the response of Body Mass Index (BMXBMI) by building a model with the following predictors:
    - Quantitative:
        - Hours spent watching TV or videos over past 30 days (PAQ710)
        - Minutes spent outdoors 9am - 5pm not work day (DED125)
        - Direct LDL Cholesterol levels (mg/dL) (LBDHDD)
        - Direct HDL Cholesterol levels (mg/dL) (LBDLDL)
        - # of times in past year you had a sunburn (DEQ038Q)
        - # of meals not prepared at home (DBD895)
        - Systolic blood pressure (BPXSY1)
- We will split our dataset into training and testing sets. We will then be building two regression models using both lasso and ridge constraints (using cross-validation to select lambda) and determining the best model through comparison of the test MSE. Then, we will interpret the coefficients of our chosen model to answer the question of how each predictor influences the response variable.
- Tools used: lasso and ridge regression, cross-validation