Applied GLM Final Project Proposal

Peter Shewmaker, Augie Ge, Ryan Buckland, Shawn Fries, Lauren Bergam

April 9, 2019

Data

For this project, our group plans to use data from the National Alzheimer's Coordinating Center's (NACC) Uniform Data Set. This data set is the "primary resource for researchers analyzing clinical and demographic from the NACC." Each row of the dataset represents a visit to the doctor from a patient, and each patient is associated with a unique id number. The data has also been filtered so that each row of the dataset represents an individual patient, by choosing each patient's last visit to represent them. Information about the different variables in the data set is provided at https://www.alz.washington.edu/WEB/rdd_uds.pdf.

Models

We came up with a few preliminary questions that we might be interested in. One simple question is if there is a binary variable that is a useful predictor of being diagnosed with Alzheimer's. One of the variables included in the dataset is "mmse," which stands for the Mini-Mental State Examination, a common questionnaire used for evaluating cognitive impairment. The scale of the test is 0-30, and "severe" cognitive impairment is a mmse score that is less than or equal to 9. We could then create a new boolean variable based on whether or not a patient has a mmse score less than or equal to 9, using this as a "screening test" for Alzheimer's diagnosis. Binary data analysis would then be performed on this new variable and the binary variable representing diagnosis with Alzheimer's. We could also implement a simple linear regression model predicting mmse score based on some of the demographic data in the set. Finally, for each patient we have multiple visits and the patient's age at each visit. We can manipulate this variable into person-years and run a poisson regression model to compare the rate of deterioration between Alzheimer's subjects and subjects without Alzheimers.

The main goal of the project would be to implement a model that predicts diagnosis with Alzheimer's based on the information we have about each patient. We would likely use a logistic regression model to do this, but we may also use other reasonable models. After creating the model, we will use K-fold cross validation to determine the model's error.

Analysis

The models will be created and analyzed using R. Time will be spent in the final paper considering the assumptions behind each model and whether or not they are met, as well as going over different diagnostic tests for the models and their results.