Edward Shao

Ben Smith

Wenrui Zhao

# Final Group Project Proposal

The research question of interest is to examine a few different factors related to COVID-19 Deaths. These factors include geographical location (state), time, pre-existing conditions, and age. The dataset that will be used to examine this research question comes from the Center for Disease Control (CDC) and is linked below:

https://data.cdc.gov/NCHS/Conditions-Contributing-to-COVID-19-Deaths-by-Stat/hk9y-quqm

We will first carefully examine the dataset. We will summarize the dataset and note the ratio of complete / incomplete cases in the dataset. We will generate a table from these summarizations. Additionally, we will create graphics from the analysis of the dataset through the ggplot2 package.

Next, we will consider the use of lasso or ridge regression methods to help choose which individuals in the dataset we want to train the model on.

We will create several models to predict the COVID-19 infection rates. We will train various types of linear regression. For each model, we will employ stepwise regression to choose the predictors. To help train the model, we will employ cross-validation techniques, to make sure the trained models are robust.

We will collaborate through the github repository.