Joshua Ingram and Kyle Dennison

Statistical Learning: Data Questions and Models

**Workload Distribution:**

We will be distributing our workload between each other by questions so we can explore our individual areas of interest while maintaining a similar theme and using the same dataset. See our two sections below to see an outline of the questions, interests, methods, and workload of each. \*

\**Our questions should remain the same, but as we work through the data, we may alter our methods or variables of interest*

**Kyle Dennison:**

*Question(s):*

Can we predict the number of suicides per 100,000 people in the five lowest scoring HDI (Human Development Index) countries in our dataset? If so, how well?

*Variables of Interest and Data Alterations:*

Response variable: suicides per 100,000 (suicides/100k pop)

Predictor variables: year, sex, age, gdp\_for\_year

To focus on the lowest scoring HDI countries, Kyle will filter the data and select the 5 countries with the most recent HDI score in the dataset (year 2016).

*Models/Methods:*

Kyle is most interested in prediction rather than interpretation, so he will be using a regression tree model, but may consider other models depending on its performance. He will perform pruning of the tree to select the most optimal tree. To evaluate the performance of the model, he will be using cross-validation (likely 10-fold). However, He may consider using forests or random forests to lower the variance of the model.

Notes/Goal(s)

The final outcome of this part of the project will be an assessment of the model and its ability to predict the number of suicides (per 100,000 people) in a country with given information.

**Joshua Ingram:**

*Question(s):*

Are there any notable associations between the number of suicides (per 100,000) and the variables in our dataset? More specifically, are GDP, HDI, and region significant predictors? If so, could there be some form of interaction between GDP and HDI, or GDP and region?

*Variables of Interest and Data Alterations:*

Response variable: suicides per 100,000 (suicides/100k pop)

Predictor variables: year, sex, age (or generation), gdp\_for\_year, region (or continent)

To consider region, Joshua will be creating a new variable that categorizes countries into different continents or regions. He is also considering combining the suicide counts for each year and country into one observation, removing the age and sex variables to look at the overall effects of GDP, HDI, and/or region on aggregate suicide counts.

*Models/Methods:*

Joshua is focused on interpretation of the model rather than maximizing predictability, so he will be using a multiple regression model, possibly with interaction. He will be testing the significance of each variable in the model, as well as the overall model, to properly assess the effects of each variable. He will be testing for collinearity and multi-collinearity, as well as interaction. He will use k-fold cross-validation to test the accuracy of the model, but will not focus his efforts on this, as interpretation of the observed data is his interest.

Notes/Goal(s):

Since he is most interested in interpretation of the data and the model created, there is a lot of room to alter the variables of interest, dataset, and methods outlined in this paper. His interest lies in how GDP interacts with region or HDI to affect the number of suicides per 100,000 people in a given country. If there is a significant association found, this could have valuable insight into how the development, health, and location of an economy has a relationship with suicides and mental health of its members.

He will perform some background research on suicide rates the effects of economic performance of a country. Specifically, GDP and suicides, as well as HDI (or region) and suicides. Depending on the outcome of this background research, as well as what’s found in the model, finding an interaction between GDP and HDI (or region) could be valuable and interesting. Of course, one cannot assume a causal relationship between these variables considering the dataset, but some logical reasoning and speculation on why there could be this association may be a good idea.