6372: Project 1

Megan Ball, Matt Farrow, Neddy Nyatome

# Introduction

Using the World Health Organization (WHO) data compiled by Kumar Rajarshi, Deeksha Russell, and Duan Wang, we developed three different models:

* The first model was designed to be easily interpreted using linear regression.
* The second model was designed to be used as a predictive tool using linear regression.
* The third model was developed using non-parametric methods for prediction.

# Data Description

The description and context of the Life Expectancy (WHO) data set can be found [here](https://www.kaggle.com/kumarajarshi/life-expectancy-who). Data has been compiled from several different data sets into a final data set that represents health factors for 193 countries between the years of 2000-2015.

Looking at the data, there are 2,938 observations and 22 variables that cover various social, economic, and health-related factors. Add more detail about the variables within the data.

# Exploratory Data Analysis

We’ll start by taking a closer look at life expectancy.

# 

Figure 1 and 2. Histogram and Q-Q Plot of Life Expectancy data.

As we might expect, life expectancy tends to skew towards the older side. The Q-Q plot shows some slight deviations from normality towards the edges, but after trying various transformations, the deviations from normality that are evident in the distribution did not seem severe enough to warrant a transformation.

Next, we began looking at correlation to narrow down our variable list before examining specific relationships. Based on a cut-off of > 0.9 for correlation, we removed the variable in each correlation pair that had the higher number of NA values (see figure x). We then proceeded to look at what happens when we also remove population, since it has minimal correlation to life\_expectancy. In the end, we ultimately removed x due to no correlation to the response variable, and x, y, z due to collinearity.

Our next task was to address the missing values in the data set (Figure @ref(fig:missing-1)). We dropped all rows where life expectancy is NA, and limited the scope of our analysis to not include those countries (Figure @ref(fig:missing-2)). We have now excluded these countries from our scope: Cook Islands, Dominica, Marshall Islands, Monaco, Nauru, Niue, Palau, Saint Kitts and Nevis, San Marino, and Tuvalu.

Hepatitis B is now our variable with the most NA’s at 553. Let’s see if there is either a year or a country that has the majority of the NA’s.

In looking at the relationship between Hepatitis B and Life Expectancy (Figure @ref(fig:hep-life)), our options are to either drop all of the NA’s, impute the values, or fill with 0’s. For our interpretable model, we made the decision to drop the hepatitis\_b variable along with the remainder of the NA’s and we will revisit them for our predictive model (Figure @ref(fig:missing-3)).

# Objective 1:

## Restatement of Problem

## Additional Data Filtering

Now that we’ve subsetted our variables and dealt with NA’s, we noticed that our feature engineering dropped almost all the records from 2015. After several looks at the data, we decided to only use the most recent “good” sample size (2011-2014).

Model Selection

Type of Selection

Checking Assumptions

Compare Competing Models

## Interpretation of Regression Coefficients

We are 95% confident that the model’s intercept is between (45.41, 48.689) and the true regression coefficient’s for the predicted variables are: adult mortality (-0.016, -0.01), total expenditure (0.183, 0.359), HIV/AIDS (-1.062, -0.772), and income composition of resources (34.999, 38.894).

# Objective 2:

## Strategy

## Metrics

## Comparison to Objective 1

# Conclusion & Final Recommendations

# Appendix