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 four broad factors: immunization-related, mortality, economic, and social. various social, economic, and health-related factors. Each record in the data contains measurements for a single year within the country being measured.

# Exploratory Data Analysis

We began by plotting life expectancy into a histogram as well as a Q-Q plot.

# 

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

As we would hope, 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 and we proceeded using the original data.

Next, we began looking at correlation to narrow down our variable list before examining specific relationships (Figure 1). 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 (Figure 2). We then proceeded to look at what happens when we also remove population, since it has minimal correlation to life\_expectancy (Figure 3). In the end, we made the decision to remove under\_five\_deaths, gdp, thinness\_1\_19\_years, and population due to lack of correlation to the response variable or collinearity.

Our next task was to address the missing values in the data set (Figure 4). We then limited the scope of our analysis to not include those countries where life expectancy was missing (Figure 5). In doing that, we excluded the following countries from our scope: Cook Islands, Dominica, Marshall Islands, Monaco, Nauru, Niue, Palau, Saint Kitts and Nevis, San Marino, and Tuvalu.

Hepatitis B was now our variable with the most missing values. In looking at the relationship between Hepatitis B and Life Expectancy (Figure 6), our options with regards to the missing values were to drop them, impute them, or fill them in with 0’s. We chose different approaches based on each model.

### Interpretable Model

For our interpretable model, we made the decision to drop the hepatitis\_b variable along with the remainder of the NA’s (Figure 7). As a result of our feature engineering, we were left with only 2 records for 2015. After a number of looks at the data, we decided to only use the observations from the most recent four years (2011-2014).

### Linear Prediction Model

Knowing that we cannot have missing values for Ridge Regression or LASSO models, we examined the relationship of each variable that had more than 100 missing values to see which appeared to be significant.

* Hepatitis B (Figure 8)
* Total Expenditure (Figure 9)
* Alcohol (Figure 10)
* Income Composition of Resources (Figure 11)
* Schooling (Figure 12)

After reviewing the plots, we made the decision to remove Hepatitis B, total expenditure, and alcohol since the trend for those three variables was relatively flat. We then removed the remainder of the missing values from the data set before proceeding to modeling.

# 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

Chart, timeline, treemap chart

Description automatically generated

Figure 1: Correlation Matrix, original data

Chart, timeline, treemap chart

Description automatically generated

Figure 2: Correlation Matrix: excluding under 5 deaths, gdp, and thinness 1-19 years

Chart, treemap chart

Description automatically generated

Figure 3: Correlation Matrix: exluding all from Figure 2 along with population

Table

Description automatically generated

Figure 4: Missing Values, original data

Table

Description automatically generated

Figure 5: Missing values, removed life expectancy

Chart

Description automatically generated

Figure 6: Relationship between Hepatitis B and Life Expectancy

Table

Description automatically generated

Figure 7: Missing Data, remove Hep B & remaining NA's

Chart

Description automatically generated

Figure 8: Relationship between Life Expectancy & Hepatitis B

Chart, scatter chart

Description automatically generated

Figure 9: Relationship between Life Expectancy & Total Expenditure

Chart, scatter chart

Description automatically generated

Figure 10: Relationship between Life Expectancy & Alcohol

Chart, scatter chart

Description automatically generated

Figure 11: Relationship between Life Expectancy & Income Composition of Resources

Chart, scatter chart

Description automatically generated

Figure 12: Relationship between Life Expectancy & Schooling