Project Proposal

Dingkun Yang, Echo Chen, Andrew Kroening, Pooja Kabber

November 4th, 2022

Overview

Dataset

The dataset used for this research is from the World Health Organization (WHO). The WHO compiles data on thousands of variables for as many countries as feasible and presents them for analysis. This particular dataset contains national-level observations of variables related to life expectancy around the globe for a period spanning the early portion of the 21st century. The complete dataset includes observations beginning in the year 2000 and ending in the year 2015. As a full dataset, there are 2,938 observations for 22 variables. Practically, each country has approximately one observation each year, averaging 183 for each of the 16 years encompassed by the data. The dataset effectively contains 20 variables for each country and year combination, covering significant disease, economic, and social factors.

Question #1 (Prediction)

"How did major disease, economic, and social factors impact life expectancy around the globe in 2014?"

Question #2 (Inference)

"How did disease and mortality rates, along with national economic factors, contribute to a country's development status in 2014?"

Models

Question #1 (Prediction)

Model

The dependent variable for the first question is Life Expectancy, which is continuous. Therefore, we will perform a linear regression analysis to model its relationship with the independent variables. We will assess different models and use measures like adjusted R^2 and AIC to determine the best one. First, we will clean and filter the data, removing variables of no interest and appropriately addressing missing data. We will also convert our categorical variables into factor variables and determine if mathematical transformations are required to handle the highly skewed variables. We will use the model plots to do this. Using the model plots, we will check for outliers and influential points and tackle them after deeper analysis. While model building, we will also conduct checks for correlation between independent variables and collinearity. During model building, we will also check if linearity assumptions are met.

Variable Selection

To select the model variables, we will use a combination of stepwise backward elimination to find the most significant variables. We will choose *a priori* variables whose effect on life expectancy we want to study, including at least one from the disease, immunization, economic, and social categories. We will also check for a correlation between independent variables and collinearity. If highly correlated variables are found, we will

remove one of them from the analysis based on variables of interest to ensure we build the most parsimonious model possible.

Question #2 (Inference)

Model

For this inference question, we will utilize logistic regression. We chose this because our outcome variable is a binary status (Developing or Developed), and there is no apparent ordering to these conditions. For this model, we will designate the status "Developing" as the baseline from which we will make inferences. As with our first question, we will clean and filter the data appropriately and examine our outputs to determine model fit and quality. We will use the 2014 subset of the data to train the model and will test its accuracy against country status' for the year 2013.

Variable Selection

A priori variable selection for this question will be complicated by the WHO's definition of a developing vs. developed country. From research, we can determine a rough framework for this designation, which likely includes an economic and social component in the context of this dataset. We will likely include Life Expectancy as a variable in this analysis. However, we may need to discard several variables, such as GDP or Income composition of resources, before fitting the final model. We may do this because several of these variables may be implicitly included in the categorical designation and therefore outweigh other dataset elements. To satisfactorily address these concerns, we will check for multicollinearity when fitting the final model using a VIF function. We will utilize a stepwise backward methodology from our list of *a priori* variables to perform the final fitting.

Challenge

The most concerning challenge for the second research question is missing values in the data. 41x Population, 10x Hepatitis B, and 10x Schooling observations are missing values. In total, these missing variables span 44 countries. When we remove Population as a variable, we still have 31 countries with missing data. In response, we will attempt to find verifiable and high-quality data from other sources to use in place of missing or sentinel values or other methods to deal with missing data. We may be forced to drop those observations if no suitable alternative data source is found.

Our next challenge will be to address the highly skewed variables. If, after conducting model diagnostics, we determine that the independent variables should be transformed, we will perform a square root or log transformation based on the distribution density of variables. An additional challenge here would be properly interpreting the transformed variables. We can also apply scaling to the continuous variables, which standardizes every feature over a normal distribution.