Project proposal

Last!: Parker Dingman, Ethan Donecoff, Karam Oubari, Pei Yi Zhuo

2021-04-16

Section 1. Introduction

By telling us the average age of death in a population, life expectancy is a key metric for understanding a country's health. According to Max Roser, Esteban Ortiz-Ospina and Hannah Ritchie from Our World in Data, "Broader than the narrow metric of the infant and child mortality, which focus solely at mortality at a young age, life expectancy captures the mortality along the entire life course."[1] Over the course of history, life expectancy has risen dramatically. It is estimated that in pre-modern times, life expectancy worldwide was only about 30 years. Since the Industrial Revolution in the 18th and 19th centuries, many countries had huge gains in this number. And since the beginning of the 20th century, global average life expectancy has risen to about 70 years. However, there remain huge inequalities in this number. Currently (as of 2019), the Central African Republic has the lowest life expectancy of 53 years while Japan has the highest with 83.

In addition to the more obvious health-related connections to life expectancy, numerous pieces of academic literature have delved into the non-medical factors behind life expectancy. A major example is a longitudinal study conducted by Charles Lin, Eugene Rogot, Norman Johnson, Paul Sorlie, and Elizabeth Arias, which examined life expectancy by socioeconomic factors.[2] Academic literature such as this provides us with motivations to examine this topic on an international level, looking at various health-related and non-health-related factors that connect to life expectancy.

In this project we plan on using various regression analysis methods including, but not limited to, multiple linear regression, statistical inference, analysis of variance, and model selection in an attempt to understand country and region-level life expectancy, as well as the health, social, and economic relationships behind this number.

In terms of initial hypotheses of model selection, we expect that the strongest predictors of life expectancy will be `Adult Mortality`, `infant deaths`, and `GDP`. We also predict that countries that have `status` equal to "Developed" will have higher life expectancy than those that have `status` equal to "Developing". We also predict that, on average, life expectancy will have increased internationally in the 15 years that are documented in the data set.

Sources:
[1] https://ourworldindata.org/life-expectancy
[2] https://europepmc.org/article/med/12785422/reload=0#impact

Section 2. Data description

Our selected data set is comprised of information that others gleaned from the websites of the World Health Organization and the United Nations. Every entry describes the health, social, and economic conditions for one of 193 countries in a given year from 2000 to 2015. Additional information can be found on https://www.kaggle.com/kumarajarshi/life-expectancy-who.

Section 3. Analysis approach

The response variable we will be using in this project is `Life expectancy`. This is a measure of the average age of death in a year for the given country during that year.

Table 1: Summary Statistics of Response Variable, Life Expectancy

min	max	range	mean	median	first_quartile	third_quartile	iqr	standard_deviation
36.3	89	52.7	69.225	72.1	63.1	75.7	12.6	9.524

Table 2: 2000 World Average Life Expectancy

world_avg_2000
66.75

Table 3: 2000 World Median Life Expectancy

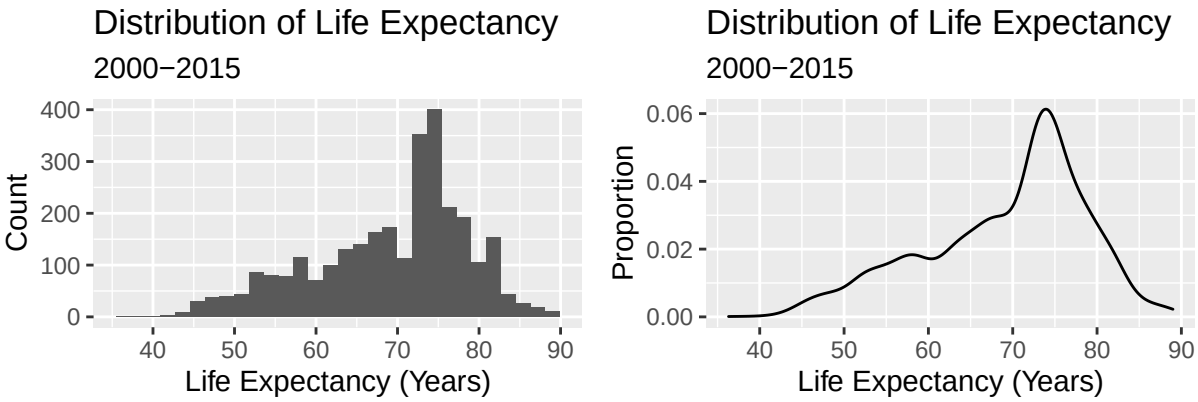
world_median_2000
71

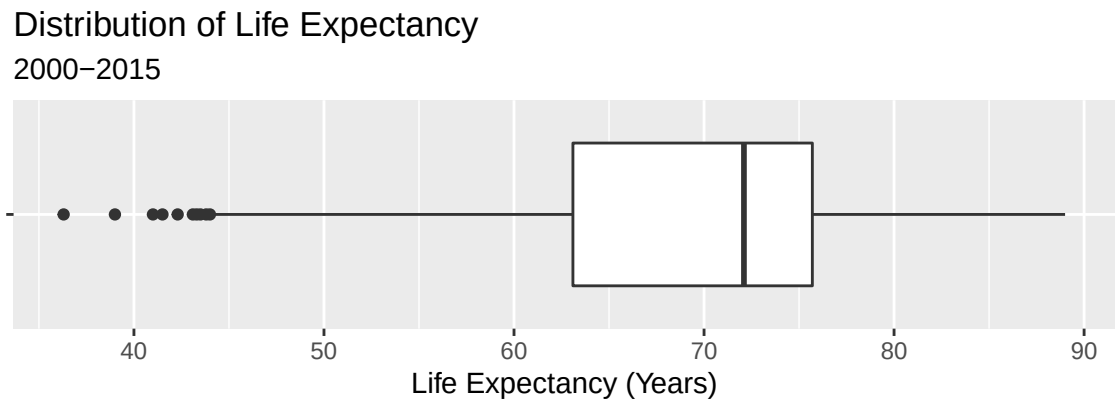
Table 4: 2015 World Average Life Expectancy

world_avg_2015
71.617

Table 5: 2015 World Median Life Expectancy

world_median_2015
73.9





In terms of predictor variables, the full model that will have all potential predictor variables will include all the variables in the data set apart from `Life expectancy`, which is the response variable, and `Country`, which is the identifier variable. In order to mitigate independence issues that arise from nearby countries being similar to one another, we will create and include in our model a new variable called `region` that identifies the area of the world in which the country is located. We will also investigate possible interactions between `region` and `Status`, our two categorical variables, and the quantitative predictor variables in our data set.

In terms of regression model technique, we will be using multiple linear regression because our response variable is quantitative.