Analyzing Significant Predictors of Life Expectancy Across Countries

Introduction

Worldwide, the average life expectancy in 2015 was 72.3 years ("Life Expectancy"). But in places like Sierra Leone and Angola, it was under 53 (Russel and Wang). Understanding why this staggering difference exists may reveal why some countries have a greater life expectancy than others. Life expectancy – the measure of the average number of years an individual is predicted to live – can be dependent upon many factors, including wealth, number of children, and level of education (Torjani). Thus, it is understandable that there exists such a large discrepancy in life expectancy across countries, particularly between developed and underdeveloped countries.

Breaking things down by region, the continent of Africa had the lowest life expectancy for both men and women in 2019. In contrast, North America had the highest for men, and Europe had the highest life expectancy for women. Many countries in the Sub-Saharan region had life expectancies of less than 60 years, while an individual born in Japan was expected to live over 80 years (Statista). Despite this disparity, there has been a global upward trend in life expectancy in recent years due to improved healthcare systems and the proliferation of vaccines (Torjani). Life expectancy rates have been steadily increasing since 2007, which has been a landmark sign of progress (Roser et al.).

What might be the root of the discrepancies in life expectancy between developed and developing countries? Much of the difference has been attributed to socioeconomic factors, including inequality, issues in the early stages of child development, and inconsistencies in access to adequate healthcare. This is supported by the work of Richard G. Rogers and Sharon Wofford. The pair discovered that mortality in developing countries can be attributed primarily to socioeconomic factors, such as urbanization, industrialization, and education, and secondarily to public health factors like access to safe water, health care physicians, and nutrition (Rogers et al.).

This paper considers a number of variables affecting life expectancy, broken down into four categories of variables: immunization, mortality, economic, and social factors. It aims to discover which factors and which categories have the most significant effect on life expectancy. Since life expectancy at birth is a widely accepted measure of a population's health status, the identification of factors that substantially contribute to life expectancy would allow countries to take a more targeted approach in improving the health of their citizens.

Methods

The data used in this paper was collected by Deeksha Russel and Duan Wang from the Global Health Observatory (GHO) data repository, as well as from the United Nations. The GHO data repository – which contains 1000 indicators on health topics, such as mortality, non-communicable diseases and risk factors, health systems, environmental health, and equity, among others – is under the oversight of the World Health Organization. It provides health-related statistics for all 194 Member States.

The data set compiled by Russel and Wang pulled data on life expectancy and health factors from the GHO data set, while the corresponding economic data was obtained from the United Nations website. Critical health factors – those that are more representative of life expectancy – were chosen for the data set. The health factors and economic factors data sets were then merged into one data set. There were several missing data points from lesser-known countries, such as Vanuatu, Tonga, Togo, Cabo Verde etc. Due to a lack of availability, Russel and Wang excluded these countries from the data set. The final data set aggregated the aforementioned factors, as well as life expectancy, for 193 countries for each year from 2000 to 2015, and it was made public on Kaggle (Russel and Wang).

As previously stated, the goal of this analysis was to determine which types of factors were most influential for life expectancy. Thus, the response variable was life expectancy, measured in years. The other variables in the data were divided into four categories of predictors: immunization-related factors, mortality factors, economic factors, and social factors.

The immunization-related factors included several metrics describing immunization percentages amongst one-year-olds, including vaccinations for polio, hepatitis-B, and diphtheria.

The mortality-related explanatory variables included adult mortality, represented as the probability of dying between 15 and 60 years per 1000 population. Infant deaths was another mortality variable collected; it was measured as the number of infant deaths per 1000 people in the population.

The economic explanatory variables included GDP per capita, total government expenditure on health, percentage government expenditure on health taken as a percentage of Gross Domestic Product per capita, and the country's development status. Development status was recorded as a binary variable named "Developed", where a value of 1 corresponded to a developed country and a value of 0 corresponded to a developing country.

The social explanatory variables included schooling, measured as the average number of years of education for the country's population. Additionally, alcohol was included as a social variable, and was measured as consumption (in liters of alcohol) recorded per capita for ages 15 and older.

While it may have been reasonable to believe that Russel and Wang's data collection process yielded reliable data, graphical inspection of several of the variables revealed some disturbing trends. In many cases, there seemed to be an unreasonable number of low outliers, seemingly correlated with one another. One example of this is shown in Figure 1, with Adult Mortality and Life Expectancy from 2015 on the axes (*Figure 1*).

Further analysis of the data suggested that these were not actually outliers, but values where the trailing zero(s) had been truncated in the data set. This conclusion was reached by comparing the questionable values to the values in recent years for the same country, and by observing that many of the outliers fit more accurately in the graph when multiplied by a factor of 10 (*Figure 2*). In light of this discovery, each variable that was to be incorporated in the original model was graphically inspected and historically cross-referenced in the manner described above. Data points that were determined to have been altered because of this truncating error were corrected to their presumably correct value.

During this data cleaning process, it was also observed that several data points within the Infant Deaths variable appeared to be significant high outliers, and a formal test for outliers was conducted. The results indicated that India, Nigeria, and Pakistan were outliers, and these three countries were subsequently removed from the data set. The rationale for their removal was that their values for Infant Deaths did not make sense contextually within the units of the variable; these data points were determined to have likely been erroneous values.

Analysis

The response variable for the data set was Life Expectancy, measured in years, and was a quantitative variable. As a result, an initial linear regression was executed, predicting Life Expectancy from the following explanatory variables: Adult Mortality, Infant Deaths, Hepatitis B, Polio, Diphtheria, GDP, Development Status, Total Expenditure, and Schooling.

Only data from the year 2015 was included in the model. The exception was that the 2014 data for Total Expenditure was used in place of the 2015 data, as there were only two 2015 data points present. The explanatory variables Alcohol and Percentage Expenditure were excluded altogether as both fields had many missing data points. Additionally, there were 53 observations with missing data points that were automatically removed during the construction of the model.

Once the initial model was created, Akaike information criterion (AIC) values were calculated and compared as a method for variable selection. Those variables whose removal would decrease AIC the most were removed one at a time until a model with the lowest AIC possible was left. This reduced model contained the optimal predictors Adult Mortality, Schooling, and Total Expenditure. The equation of this model was found to be as follows:

$$\textit{Life expectancy} = 70.46042 - (0.06422)(\textit{Adult Mortality}) + (0.84015)(\textit{Schooling}) + (0.18191)(\textit{Total Expenditure})$$

It was assumed that the average value of the residuals was equal to zero, and that the errors between any two different observations were independent, as only data points from a single year were included in the model. A residual plot graphing the fitted or predicted values against the residuals was then created (*Figure 3*). The analysis of the residual plot revealed that the residuals were normally distributed around

zero with an approximate constant variance. Next, a histogram of the residuals was created (*Figure 4*). It was approximately bell shaped, and so the residuals were assumed to be reasonably normal. Altogether, this analysis implied that the conditions for inference for linear regression were met and that no transformations of the response variable were needed.

In the final model, it was observed that Adult Mortality had a slight, negative correlation with Life Expectancy. Total Government Expenditure was found to have a positive, but weak, linear relationship with Life Expectancy. Finally, Schooling was shown to have a strong, positive and linear relationship with Life Expectancy (*Figure 5*). Out of the three variables in the final model, Adult Mortality was found to be the most significant variable in predicting Life Expectancy, as its removal would increase AIC the most. Total Expenditure was found to be the least significant, as its removal would increase AIC the least.

Discussion

Analysis of the previously mentioned immunization, mortality, social, and economic factors revealed that a country's mean life expectancy could best be predicted from a model that included Adult Mortality, Schooling, and Total Expenditure. Of these factors, Schooling and Total Expenditure were positively correlated with life expectancy, while Adult Mortality, measured as the probability of dying between 15 and 60 years per 1000 population, was negatively correlated. Adult Mortality proved to be the most significant variable in the prediction model, while Total Expenditure was the least significant.

While no test was explicitly done to determine the most predictive subset of factors, Adult Mortality was the most significant variable in the final model. Thus, mortality-related factors have a strong connection to life expectancy.

Generally, correlation tables displayed that higher immunization rates led to higher life expectancy. As mentioned previously, Total Expenditure, which is the general government expenditure on health as a percentage of total government expenditure, was positively correlated with life expectancy, with a correlation coefficient of 0.8089. Social factors, like the average number of years of schooling for the entire population, had a correlation value of 0.9002. Finally, mortality factors, such as Adult Mortality and Infant Deaths, were negatively correlated with life expectancy and displayed correlation values of -0.9417 and -0.3638 respectively.

It should be noted that the data set contained many missing or inaccurate data points, which made conducting an analysis on the un-edited, original data set difficult. The data set was reduced to only rows that had no missing values, and it also had to be cleaned in order to ensure that all of the data points were accurate (many of the values were missing 0's at the end, meaning that they were off by factors of 10). There was no data for the Total Expenditure variable in 2015, so the data from 2014 was used for the analysis. If a similar analysis is conducted in the future, it would be helpful to have a data set that is complete and error-free. Additionally, it would be interesting to execute an analysis that looked at the variables over time in order to see which are the most predictive of life expectancy across years.

Conclusion

Higher values for average years of education and increased percentage expenditure on health care, along with lower mortality rates, constitute the primary differences in countries with higher life expectancies. Thus, increasing life expectancy presents a significant challenge for underdeveloped countries. In these countries, citizens often cannot afford to send their children to schools, the government has fewer financial resources, and adult mortality rates are relatively high due to lack of health care, among other factors. Therefore, if countries desire to increase the average life expectancy of their population, they must make other adjustments that allow these three factors to change. They may benefit from aid so that they can provide better health care and education for their citizens.

Appendix

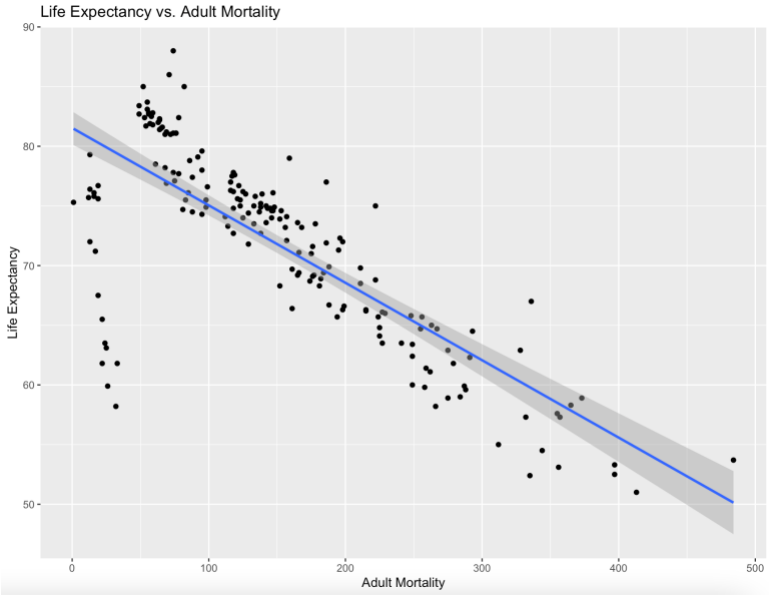

Figure 1. Life Expectancy vs. Adult Mortality (Prior to Data Correction)

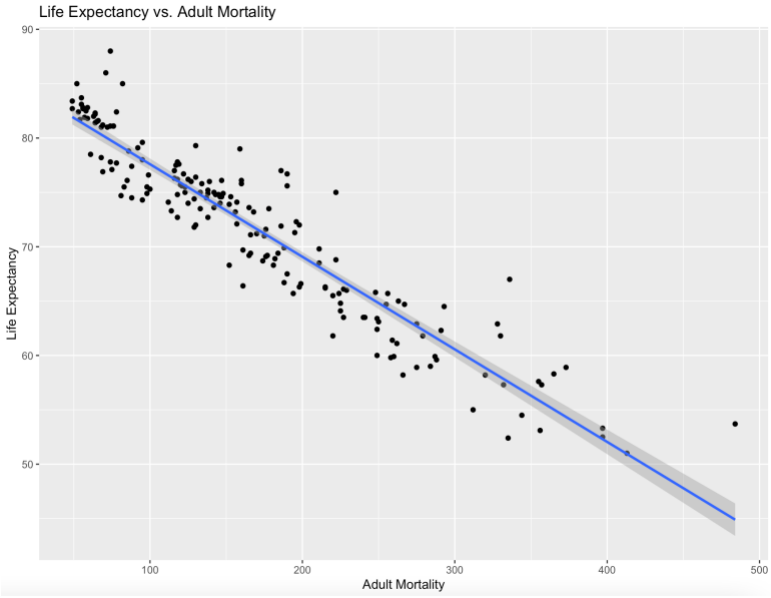

Figure 2. Life Expectancy vs. Adult Mortality (After Data Correction)

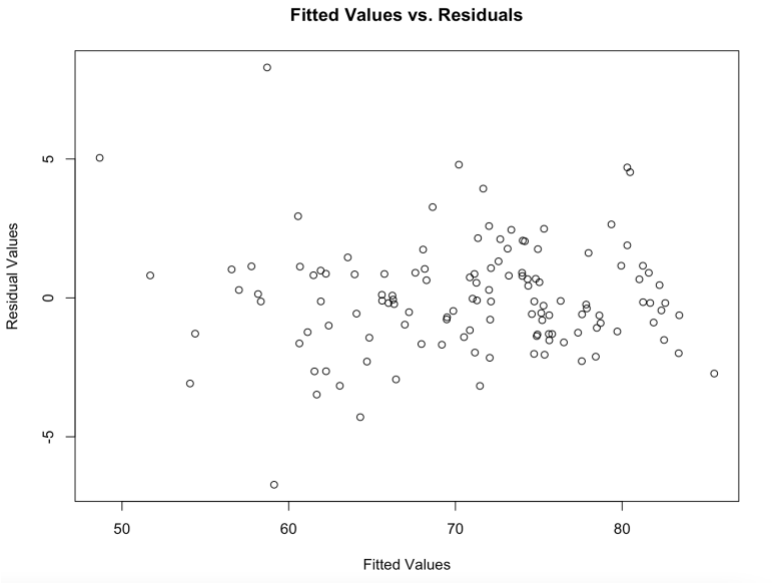

Figure 3. Fitted Values vs. Residuals

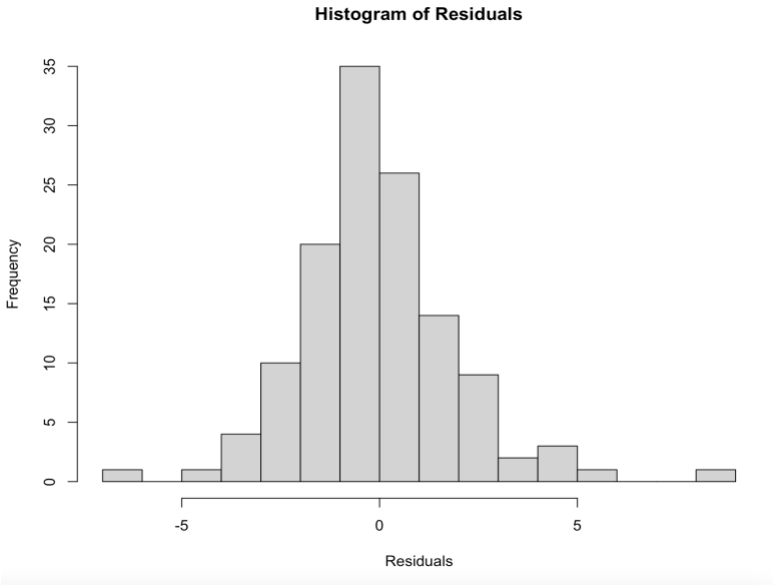

Figure 4. Histogram of Residual Values

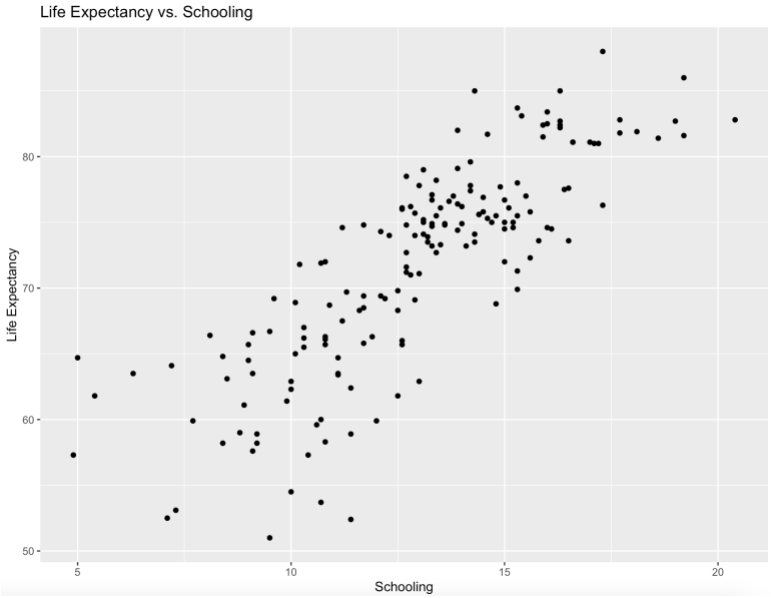

Figure 5. Life Expectancy vs. Schooling

Works Cited

"Life Expectancy of the World Population." *Worldometer*, https://www.worldometers.info/demographics/life-expectancy/.

Miladinov, Goran. "Socioeconomic Development and Life Expectancy Relationship: Evidence from the EU Accession Candidate Countries - Genus." *Journal of Population Sciences*, Springer International Publishing, 10 Jan. 2020, https://genus.springeropen.com/articles/10.1186/s41118-019-0071-0.

Published by M. Szmigiera. "Life Expectancy in Developed and Developing Countries." *Statista*, 8 Feb. 2022, https://www.statista.com/statistics/274507/life-expectancy-in-industrial-and-developing-countries/#:~:text=In%202021%2C%20the%20average%20life,and%2075%20years%20for%20females.

Rogers, R G, and S Wofford. "Life Expectancy in Less Developed Countries: Socioeconomic Development or Public Health?" *Journal of Biosocial Science*, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/2722920/.

Roser, Max, et al. "Life Expectancy." *Our World in Data*, 23 May 2013, https://ourworldindata.org/life-expectancy#:~:text=Life%20expectancy%20is%20the%20key,of%20death%20in%20a%20population.

Russel, Deeksha, and Duan Wang. "Life Expectancy (WHO)." *Kaggle*, https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?select=Life+Expectancy+Data.csv.

Torjani, Ava. "Life Expectancy: Discrepancies, Outcomes, and Future Directions | Princeton Public Health Review." *Princeton University*, The Trustees of Princeton University, 5 Nov. 2017, https://pphr.princeton.edu/2017/11/05/life-expectancy-discrepancies-outcomes-and-future-directions/.