Modelling a Country's Average Life Expectancy

Parker Johnson, Teagan Johnson

11/23/2021

Introduction

Life expectancy has been a measure of a country's well-being in countless studies. Because we were curious about predicting a country's well-being, we decided to use a dataset collected from WHO about countries' health-related data to analyze what predictors are most associated with life expectancy. Studies have been conducted in the past that focus on life expectancy at an individual level. For example, a study in 2016 concluded that life expectancy for women can be greatly influenced by whether or not they smoke (Canudas-Romo 2016). Our data, however, focuses more on country-level life expectancy by including terms like immunization rates and average schooling experience. Using the data from WHO, we created two multiple linear regression models setting life expectancy as the response variable and 21 other health-related predictors as the explanatory variables: one for 2004 and one for 2014. Using our multiple linear regression models, we were able to conduct significance tests and determine which of our explanatory variables were most significant on life expectancy and speculate on changes in terms from 2004 to 2014.

Data

The dataset we used was collected from the WHO website on which there were individual datasets that listed health-related data separately for each country. WHO collects data such as infant deaths per 1000, GDP, and population on a yearly basis. Corresponding economic data was collected from the United Nations website. Deeksha Russell and Duan Wang combined all the individual datasets into one and posted it to Kaggle.com, a website where users can post datasets. We downloaded the data from Kaggle onto our local machines.

Our variables include:

life expectancy - measured in years

country

year - spans 2000 to 2015

status - whether the country is developed or not

adult mortality rates of both sexes - probability of dying between 15 and 60 years per 1000 population

infant deaths - per 1000 population;

Alcohol - recorded per capita consumption in litres of pure alcohol;

Expenditure on health as a percentage of Gross Domestic Product per capita;

Hepatitis B immunization coverage - among 1-year-olds as a percentage;

Measles - number of reported cases per 1000 population;

Average Body Mass Index of entire population;

Number of under-five deaths per 1000 population;

Polio immunization coverage among 1-year-olds as a percentage;

General government expenditure on health - as a percentage of total government expenditure as a percentage

Diphtheria tetanus toxoid and pertussis immunization coverage among 1-year-olds as a percentage

HIV/AIDS Deaths per 1000 live births HIV/AIDS

Gross Domestic Product per capita in US dollars

Population of the country

Prevalence of thinness among children and adolescents for Age 1 to 19 as a percentage

Prevalence of thinness among children for Age 5 to 9 as a percentage

Human Development Index in terms of income composition of resources - index ranging from 0 to 1

Number of years of Schooling in years.

However, not every country was included in this data. Smaller countries (like Andorra, Dominica, Vatican City, Liechtenstein, Marshall Islands, Monaco, Nauru, Palau, Saint Kitts and Nevis, San Marino, and Tuvalu) do not have easily accessible data. Regions that were not universally recognized as a country were also excluded, such as Kosovo and Taiwan.

Results

This data set covered years from 2000 to 2015. Ideally, our "newer model" would have been from the most recent year possible in order to have up-to-date information. Using this newer model, we planned to compare it with an older model representing the oldest year in our data. However, the dataset for the most recent year, 2015, contained significant missing values, making this year unusable for interpretations. Similarly, the years 2000, 2001, 2002, and 2003 had significant missing hepatitis B values which meant these years would be less accurate for our interpretations. Because of these missing values, we decided to use the years 2014 and 2004 for our models, since they did not have as many missing values as the other years.

After deciding that 2014 and 2004 were good fits for our study, we had to remove the rows with significant missing terms. First, we removed the countries that contained columns with missing entries. This narrowed our 2014 model to 131 countries, and our 2004 model to 103 countries. However, we noticed that some terms contained unreasonable values that were equal to 0. These terms were infant deaths, measles, and under 5 deaths, and they all contained around 30 values that were equal to 0. Removing the countries that had these 0 values led to models that were less accurate and contained more variation, so we decided to remove these 3 terms.

We then investigated whether or not transformations were necessary for our models. Looking at the predictor vs. residual plots for the predictors, we found that we needed to transform percentage expenditure (logged),

Table 1: Final Model for 2014

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	59.487	2.734	21.762	0.000	54.076	64.898
StatusDeveloping	-2.164	0.963	-2.247	0.026	-4.070	-0.258
Adult.Mortality	-0.022	0.004	-5.938	0.000	-0.030	-0.015
log(percentage.expenditure + 0.5)	0.698	0.177	3.933	0.000	0.347	1.049
Total.expenditure	0.228	0.119	1.913	0.058	-0.008	0.464
log(HIV.AIDS)	-2.076	0.320	-6.479	0.000	-2.710	-1.441
log(thinness.5.9.years)	-0.614	0.331	-1.854	0.066	-1.271	0.042
Schooling	0.743	0.174	4.283	0.000	0.400	1.087

hepatitis B (squared), Polio (squared), Diphtheria (squared), HIV/AIDS (logged), GDP (logged), Population (logged), thinness from ages 1-19 (logged), and thinness from ages 5-9 (logged) for both the 2014 and 2004 models. These transformations allowed our models to satisfy the assumptions for a multiple linear regression model, as the issue of nonconstant variance was fixed and linearity was met with these transformations. We decided that our model had little potential for interactive terms due to the nature of our terms, so we didn't include any.

We investigated whether the data had any potential influential cases that were skewing our models and analysis. After looking at residuals and Cook's distances, we concluded that there were no significantly influential cases in our data for both the 2014 and 2004 models, as all Cook's distance values were fairly low.

Next, we looked at the summary statistics to determine which variables were insignificant and could be removed from our models. For the 2004 model, the terms that were deemed insignificant due to high collinearity or low significance included StatusDeveloping, Alcohol, BMI, $Hepatitis.B^2$, $Polio^2$, Total_expenditure, log(GDP), log(Population), and log(thin_1_19). For the 2014 model, the terms that were deemed insignificant due to high collinearity or low significance included Alcohol, BMI, $Hepatitis.B^2$, $Poli0^2$, $Diphtheria^2$, log(GDP), log(Population), Income.Composition.Of.Resources, and log(thin_1_19).

It's worth noting that for the 2004 model, Income.Composition.Of.Resources had a p-value of 0.0542, but since this was very close to 95% significance, we decided to keep it in the model. This borderline p-value was also the case for Total_expenditure and log(thin_5_19) for the 2014 model.

Our final models are listed below (see tables 1 and 2):

$$\begin{aligned}\hat{\mu}(LifeExpectancy2014 \mid X) &= 59.484 - 2.164StatusDeveloping - 0.022AdultMortality \\ &\quad + 0.698log(PercentageExpenditure + .5) + 0.228TotalExpenditure \\ &\quad - 2.076log(HIV/AIDS) - 0.614log(Thin - 5 - 9) + 0.743Schooling\end{aligned}$$

$$\begin{aligned}\hat{\mu}(LifeExpectancy2004 \mid X) &= 57.949 - 0.021AdultMortality + 0.437log(PercentageExpenditure + .5) \\ &\quad + 0003Diphtheria^2 - 1.905log(HIV/AIDS) - 0.852log(Thin - 5 - 9) \\ &\quad + 3.974IncomeCompOfResources + 0.529Schooling\end{aligned}$$

Table 2: Final Model for 2004

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	57.949	2.241	25.856	0.000	53.499	62.398
Adult.Mortality	-0.021	0.003	-6.127	0.000	-0.028	-0.014
log(percentage.expenditure + 0.5)	0.437	0.209	2.091	0.039	0.022	0.853
I(Diphtheria^2)	0.000	0.000	2.529	0.013	0.000	0.001
log(HIV.AIDS)	-1.905	0.288	-6.621	0.000	-2.477	-1.334
log(thinness.5.9.years)	-0.852	0.391	-2.178	0.032	-1.629	-0.075
Income.composition.of.resources	3.974	2.038	1.949	0.054	-0.073	8.020
Schooling	0.529	0.185	2.851	0.005	0.160	0.897

With our final models, we could now compare each term to determine how the effects of each term on the mean life expectancy within a country has changed from 2004 to 2014. With the 2004 model, schooling is associated with an estimated increase of 0.529 years on mean life expectancy, and we are 95% confident that this estimated effect is associated with an increase of 0.160 and 0.897 years in mean life expectancy (see Figure 1). Meanwhile, with the 2014 model, schooling is associated with an estimated increase of 0.743 years on mean life expectancy, and we are 95% confident that this estimated effect is associated with an increase of 0.400 and 1.087 years on mean life expectancy (see Figure 2).

Figure 1: Life Expectancy in 2004 versus Years of Schooling

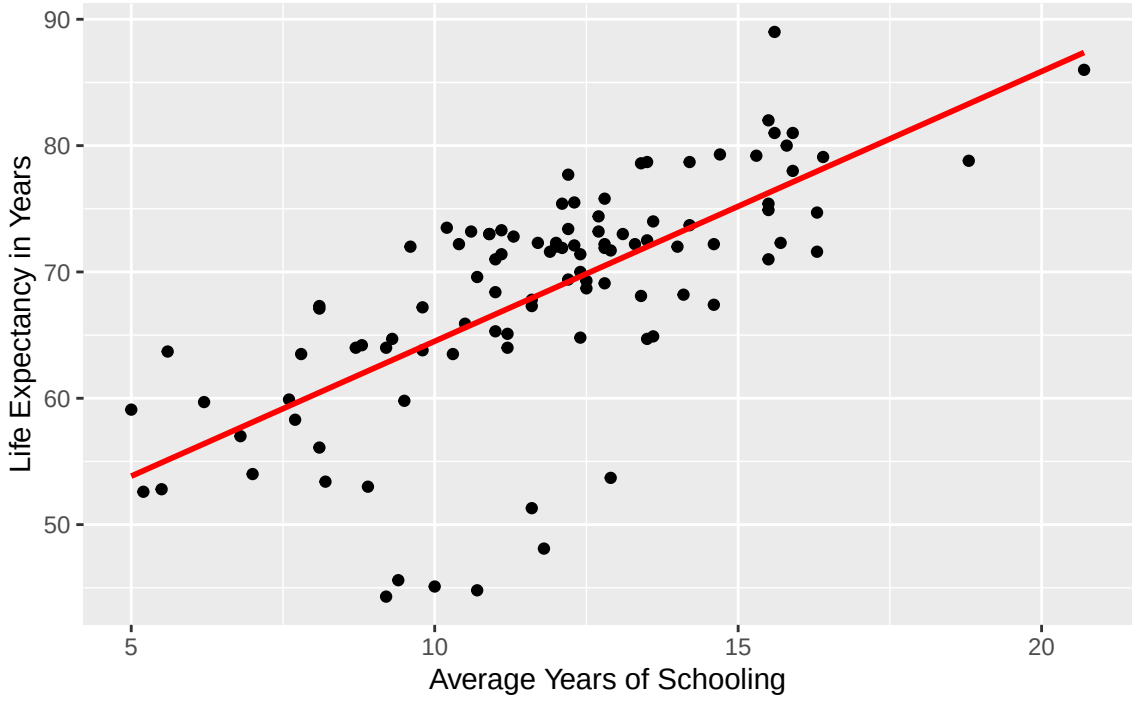
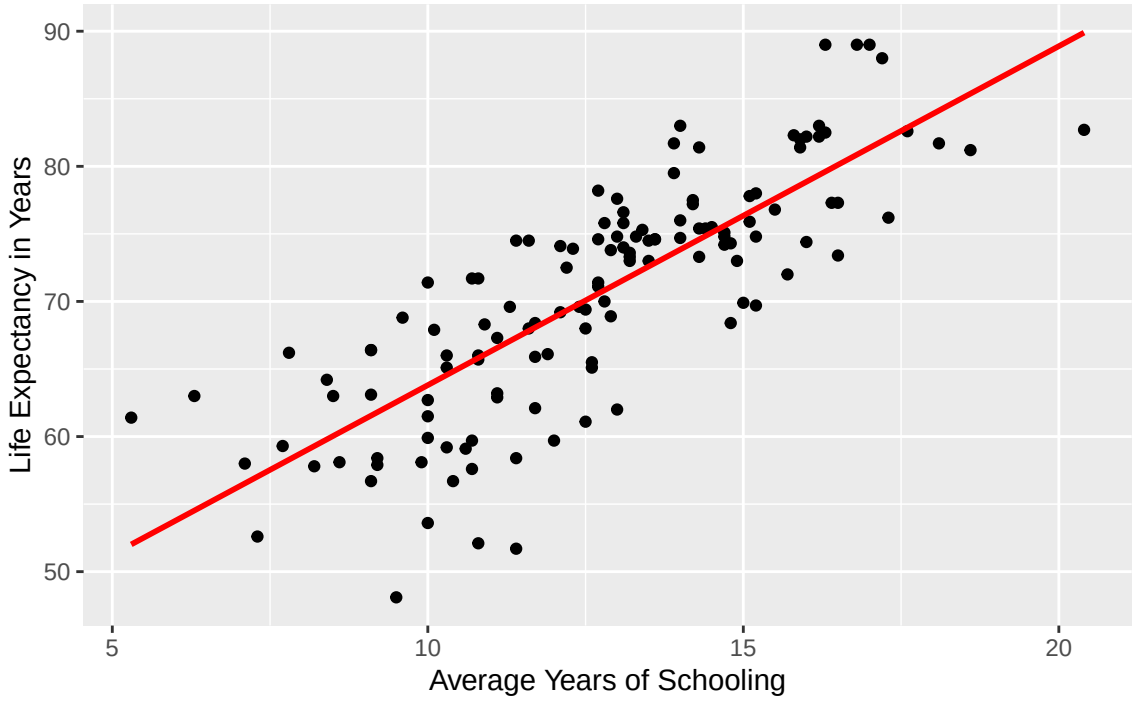


Figure 2: Life Expectancy in 2014 versus Years of Schooling



For HIV/AIDS, the 2004 model shows that the estimated log death rate for HIV/AIDS is -1.905 years, which means that the estimated death rate for HIV/AIDS in 2004 is associated with a decrease of 0.149 years on mean life expectancy. With the 2004 model, we are 95% confident that the estimated death rate for HIV/AIDS is associated with a decrease of 0.084 to 0.263 years on mean life expectancy (see Figure 3). For the 2014 model, the estimated death rate for HIV/AIDS is associated with a decrease of 0.125 years on mean life expectancy, and we are 95% confident that the estimated death rate for HIV/AIDS is associated with a decrease of 0.067 to 0.237 years on mean life expectancy (see Figure 4).

Figure 3: Life Expectancy in 2004 versus log of HIV/AIDS

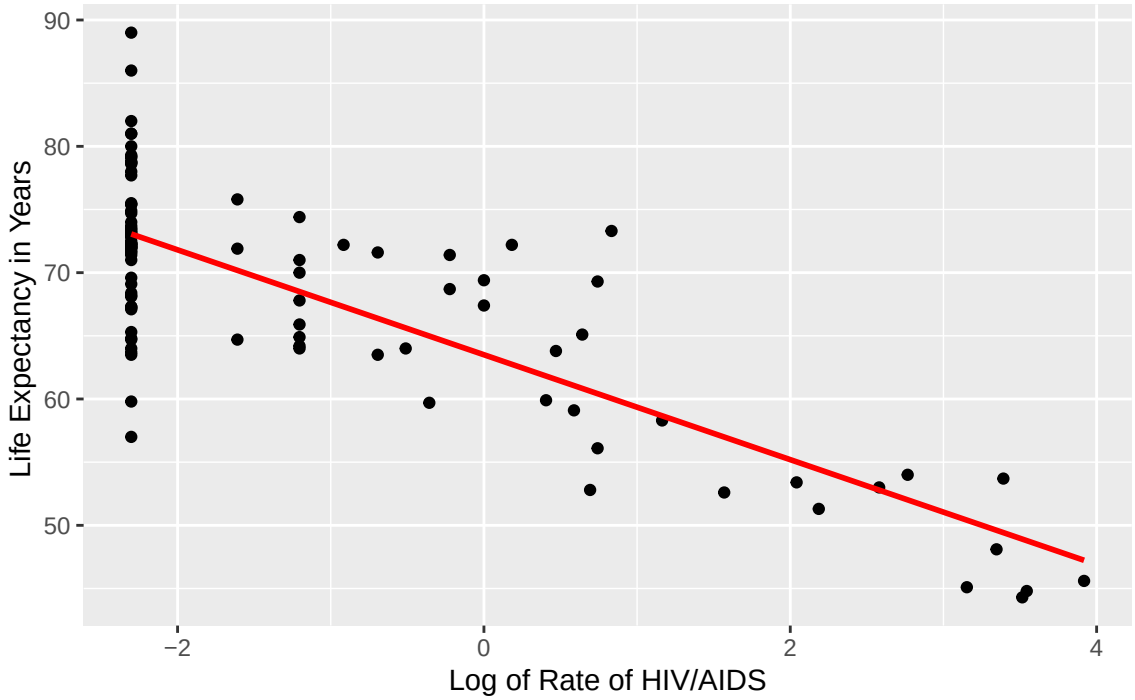
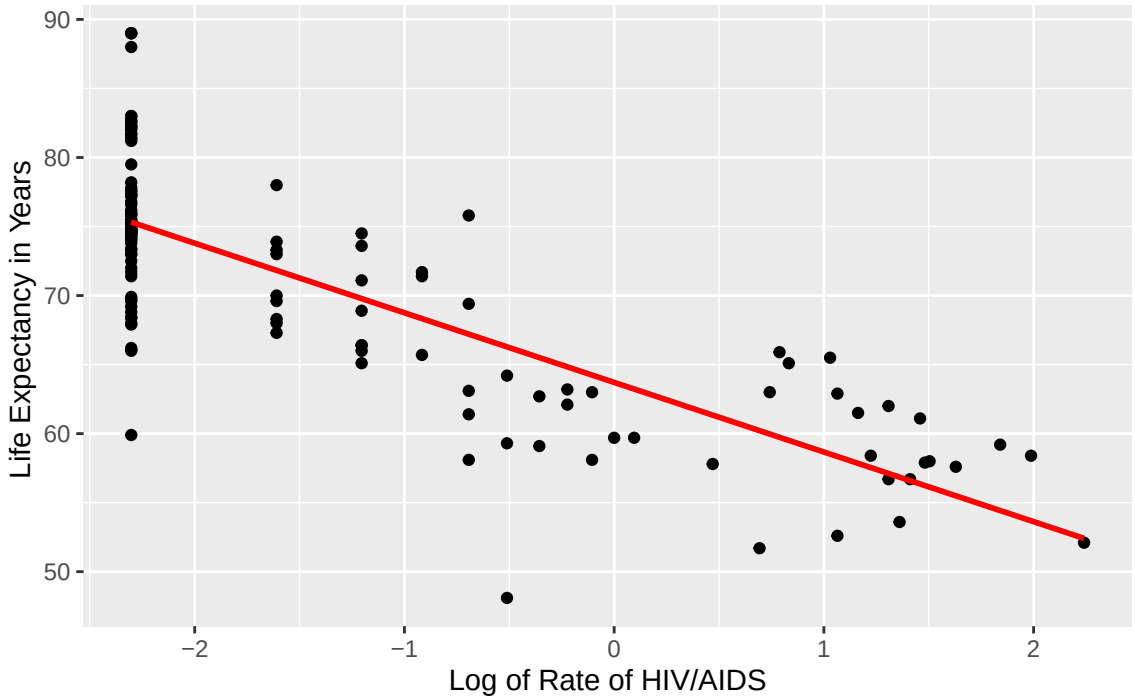


Figure 4: Life Expectancy in 2014 versus log of HIV/AIDS



For the status of a country, the 2014 model shows that the a country being labeled as developing versus developed is associated with an estimated decrease of 2.164 years on mean life expectancy. With the 2014 model, we are 95% confident that a country being labeled as developing versus developed is associated with an estimated decrease of 0.258 to 4.070 years on mean life expectancy (see Figure 5).

Figure 5: Life Expectancy in 2014 versus Status of a Country

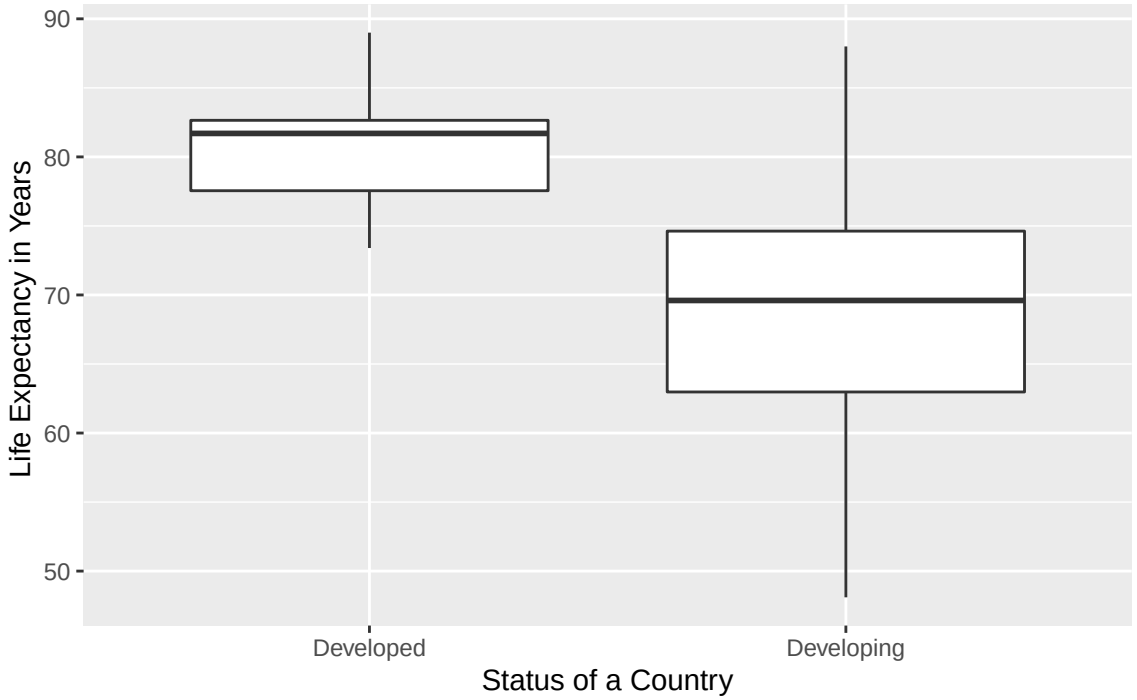
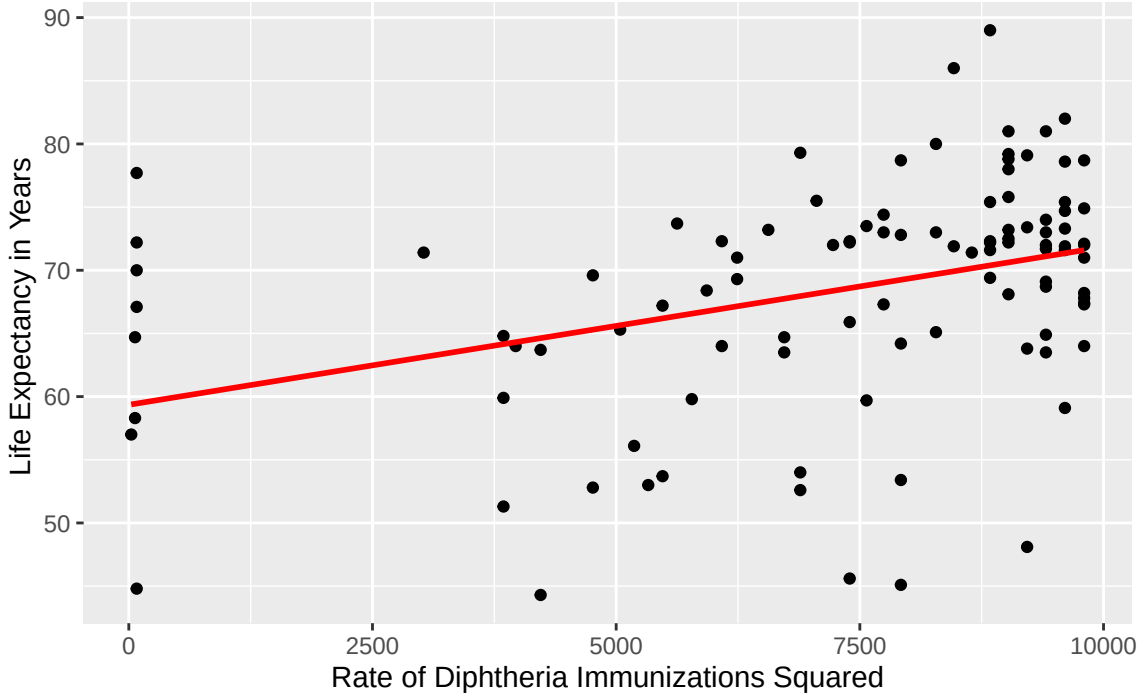


Figure 6: Life Expectancy in 2014 versus Diphtheria Immunizations ^2



Discussion

We concluded that the most significant factors for determining life expectancy in 2014 out of the 22 terms were country status (whether it's developing or developed), adult mortality, percentage expenditure, total expenditure, HIV/AIDS, thinness_5_9, and level of schooling. For 2004, we concluded that the most

significant factors for determining life expectancy out of the 22 terms were adult mortality, percentage expenditure, diphtheria, HIV/AIDS, thinness_5_9, income composition of resources, and level of schooling.

Schooling had one of the highest positive effects on life expectancy both in 2014 and in 2004 (see results for estimated effects). This supports a previous study from 2006 concluding that "in Russia and Estonia an improved educational structure prevented an even greater decline in life expectancy" (Shkolnikov). There appears to be a clear positive relationship between the average length of schooling and life expectancy. It's worth mentioning the estimated effect of schooling on life expectancy has increased from 2004 to 2014, but we cannot conclude with 95% confidence that this estimated difference is statistically significant. Because countries with more schooling tend to have higher life expectancies, it is reasonable to conclude that in order to increase life expectancy, countries should place more emphasis on providing opportunities for education for citizens.

On the other hand, the status of a country had one of the most negative effects on life expectancy in 2014, with status being associated with an estimated 2.164 year decrease on mean life expectancy. This means that if a country was considered developing, its life expectancy was estimated to be lower than if it were developed. Countries' status in 2004, however, was not statistically significant with respect to a country's life expectancy. This indicates that the disparity between developing and developed countries has become more significant since 2004. It's possible that this trend may continue in the future, and the disparity could become even more significant. To combat this increasing disparity, it may be beneficial for developed countries to provide more support for developing countries.

Diphtheria immunization coverage had a statistically significant effect on life expectancy in 2004 (see figure 6), but not in 2014. In a study conducted in 1995, it was noted that "several developing countries where coverage has been high for 5-10 years have reported diphtheria outbreaks" (Galazka). The study notes that these outbreaks (characterized by high fatality rates and widespread occurrence) were likely due to lack of herd immunity due to lower immunization rates in developing countries. Although this paper was written in 1995, its analysis on low immunization rates in developing countries could shed light on how diphtheria has become less statistically significant on life expectancy. It's possible that countries, specifically developing countries, have been able to increase their immunization rates since 2004.

The scope of our model is limited to the countries we included in the model. This excludes countries like the United States and the United Kingdom. Therefore, it is ill-advised to make generalizing conclusions about every country in the world. Instead, it is expected that our conclusions relate to our model's smaller group of countries.

One limitation that we faced during this study was that the original data set had some clear inaccuracies. For example, in the 2014 model, the GDP of Romania in 2014 was listed as 12.277, Sri Lanka had a population of 2771, and Australia's adult mortality rate was listed as 6 per 1,000 people. While the majority of the data set had values that seemed to be accurate, values similar to those listed above seem suspiciously incorrect. With this in mind and since the data collected from the WHO and United Nations was most likely fairly accurate, it's likely that there was an issue when merging the data sets into the data set we obtained from Kaggle that affected some of these values. While no individual country was found to be an outlier in our inference, inaccurate data limits the strength of the validity of our results if our models would have differed with more accurate data points.

Another limitation was that not every term was included in our initial study. We had to remove infant

mortality, measles cases, and under 5 mortality rates because there were too many values inaccurately listed as 0, and there are countless factors that could impact life expectancy that were not measured in the first place. By having a data set with fewer initial predictors, we may be missing out on inferences to be made about infant mortality or other predictors not encapsulated in this study. This also decreases the validity of our conclusions regarding whether these terms are the most significant with regards to life expectancy in 2004 and 2014.

Keeping these limitations in mind, we can speculate on how the significance of these terms might change in the future with regards to life expectancy. The significance of variables did change from 2004 to 2014 with our model, with diphtheria no longer being significant in 2014. This could indicate that diseases in the future may become less of a significant factor on life expectancy as countries become better well-equipped to treat diseases. Predicting these changes would be worth studying. A variable that could have potential significance on life expectancy is the prevalence of type II diabetes in a country, which is a condition more often associated with developed countries. Another area of interest in the future based on this study would be to examine how significant predictors vary from developing to developed countries.

Conclusion

We sought out to analyze the most significant predictors associated with life expectancy. Our results indicate that the most significant terms associated with life expectancy in 2014 were adult mortality, percentage expenditure, diphtheria, HIV/AIDS, thinness_5_9, income composition of resources, and level of schooling. There is a change in significant terms from 2004 to 2014 in our model that indicated changes in country-wide programs, like diphtheria immunization rates. We hope that our results spark interest and open doors to future research on life expectancy, including the study of predictors and also the general impact of life expectancy on countries.

Works Cited

Vladimir Canudas-Romo, Heather Booth, Marie-Pier Bergeron-Boucher. (2019) Minimum Death Rates and Maximum Life Expectancy: The Role of Concordant Ages. North American Actuarial Journal 23:3, pages 322-334.

Shkolnikov VM, Andreev EM, Jasilionis D, et alThe changing relation between education and life expectancy in central and eastern Europe in the 1990sJournal of Epidemiology & Community Health 2006;60:875-881.

Galazka AM, Robertson SE. Diphtheria: changing patterns in the developing world and the industrialized world. Eur J Epidemiol. 1995 Feb;11(1):107-17. doi: 10.1007/BF01719955. PMID: 7489768.