Investigating Controllable Factors of Life Expectancy

Dillan Sant, Chase Mathis

Introduction

Humans have experienced a tremendous rise in life expectancy in the past 200 years thanks to major advances in public health, but recently the momentum has stalled if not reversed.[1] Children are taught from a young age to have a healthy diet, exercise regularly, and keep up personal hygiene to stay healthy and thus live longer. However, we want to give advice to entire countries on how to stay healthy. Countries cannot collectively have good hygiene or healthy, diverse diets, so we rely on aggregate macro-factors to predict longevity.

This research project aims to investigate the relationship between these macro-factors and life expectancy for developing and developed countries alike. Experts in public policy and public health are our intended audience as we attempt to give them further evidence on what should be prioritized in the the struggle to keep their citizens healthy. Our research questions are the following:

(1) *Given a country is developing, what can they do to increase their life expectancy?* This question hopes to guide methods for public policy and health experts in developing countries. Generally, developing countries have a lower life expectancy, so what does the data say about which significant factors cause this, and how can life expectancy be increased in these countries?

(2) *For countries that already have a high life expectancy, is it economically beneficial to attempt to marginally increase life expectancy?* Developed countries have had the advantage of modern medicine for quite some time, so this question investigates if incremental increases in life expectancy are "worth" the increase in global health expenditure. Should countries focus on research in healthcare innovation and finding a "miracle" vaccine, or is there still work to be done for other factors like schooling or alcohol abuse?

Similar to how each person gets individual treatment from their primary care physician on their health, we think it is important to divide the research into the categories outlined above so the findings can be more specific and beneficial for nations who fall into those categories. We've also split up our predictors in a likewise fashion. We've categorized "Control" variables as features that public policy and health experts can somewhat control. We then categorized "Nuisance" variables as those which governments have little control over such as Population, BMI, and GDP.

Our first research question hopes to tackle the issue of inequality in life expectancy based on where one was born. Developing countries must increase their life expectancy so to match that of other developed countries. The economic benefit of a healthy, long-living country is clear and thus important to understand how to cultivate. The second question hopes to answer an important question of the utility of investing in public health measures. In other words, is there some law of nature that reasonably, or financially limits how old one can get? Can we find if there is some limit through data analysis? How beneficial is it to increase health expenditure, and instead should governments invest in cutting-edge research to find miracle cures?

We will first show our exploratory data analysis, which will help us understand how to fit the data later on in the modeling stage. After EDA, we will explore our first research question using various interpretable models. Next, we will explore our second research question through similar methods as the first. Finally, we will conclude and give advice to researchers in the field of public health and public policy in what direction they should prioritize.

[1] Why life expectancy in the US is falling

Data

The Global Health Observatory (GHO) under the World Health Organization (WHO) collected the data and has made it public in their data repository for global health analysis. The features of this data contain global health data for specific countries collected by GHO and WHO as well as economic data collected from the United Nations' website. The data has 21 features which are outlined in the data dictionary in the appendix. Each observation in our data represents a country, its macro-factor summary statistics, and the year. As the features are summary statistics, we are predicting averages from averages. We will *not* use black-box models such as random forests or bagging to get high predictive accuracy, as this question investigates aggregate relationships. We will therefore be using statistical modeling techniques such as linear regression, regularized regression, trees, and GAMS, as we aim to quantify the magnitude and type of relationships between life expectancy and each of our features.

Data Cleaning and Limitations in Our Data

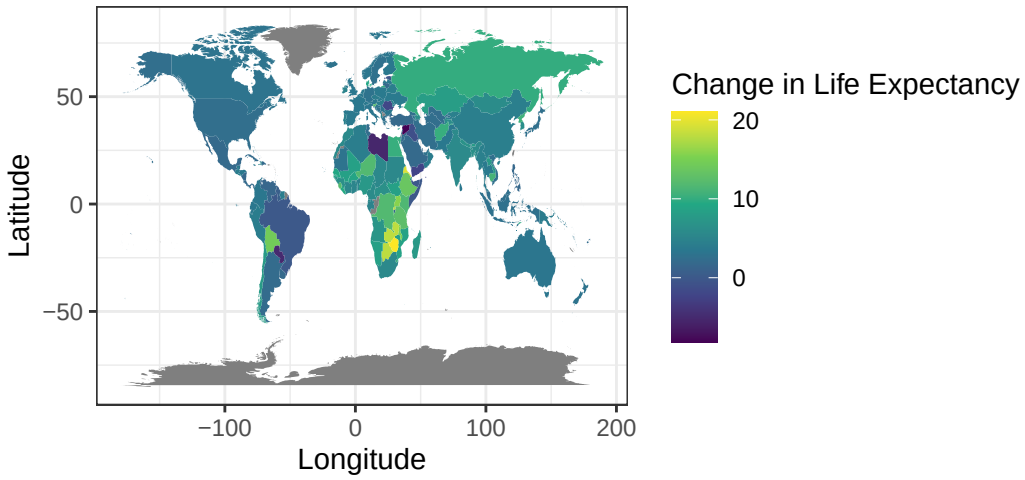
Table 1: Variables with Missing Values

Term	Percent Missing
Population	22.19
hep_b	18.82
GDP	15.25
tot_expend	7.69
Alcohol	6.60

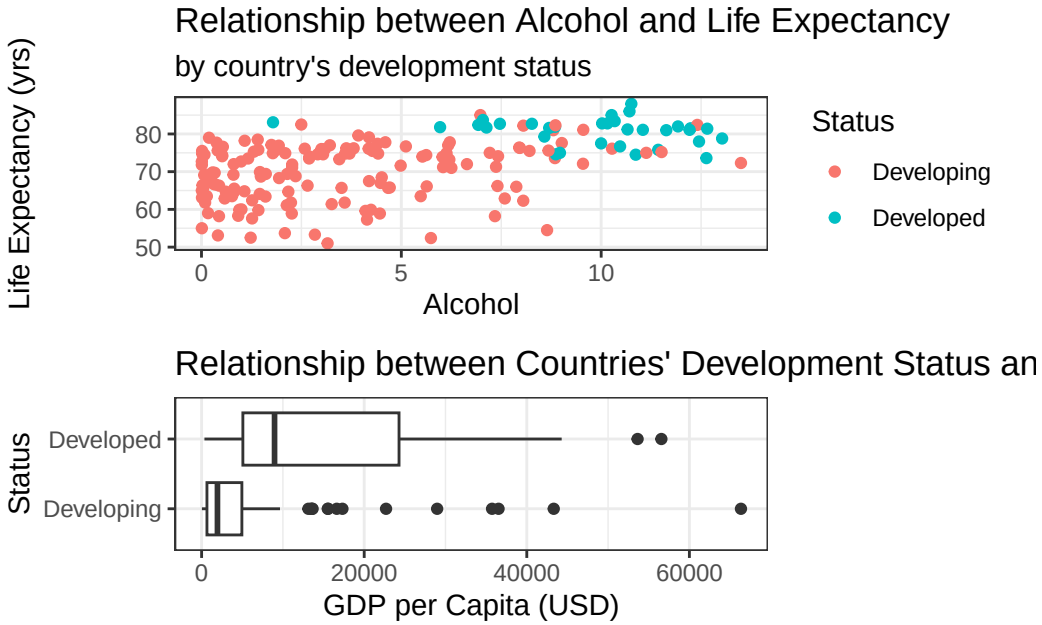
The data is mostly complete, but certain features have missing values. The missing data will have an impact on our modeling and interpretation which is something to consider for future work. For the time being, we will split up our missing values into two categories: (1) Missing values for entire countries and (2) Missing values for time ranges within a certain country. In regards to the first type of null value, we are left with little options. Many types of models will throw errors if there is missing data, so when fitting models that depend on predictors with missing data, we will throw away observations where missing data is present. In the other case, we propose using the mean of the other samples in that country to fill in the data. For instance, if Algeria has data on its alcohol consumption for all years except 2006, we assume that we should estimate 2006 alcohol consumption using the mean of Algeria's other years' alcohol consumption measures. We continue with this methodology.

After filling in null values with the mean, we are able to decrease the percent of na-values from 43.9% to 27.8%. This reduction will help us use more of the data so that we can find better conclusions for policy makers.

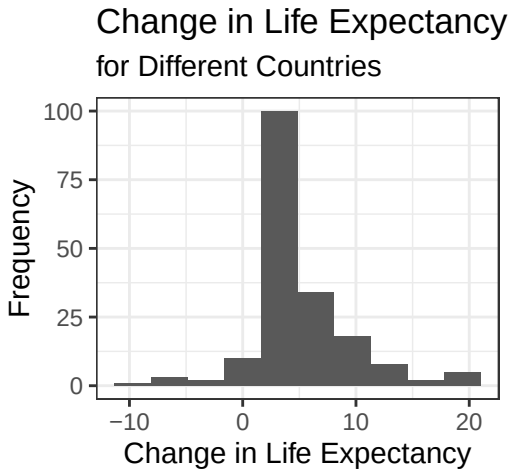
Exploratory Data Analysis

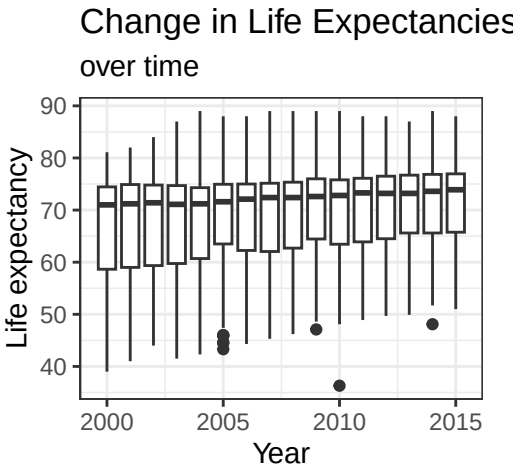


Generally, countries experienced an increase in life expectancy from 2000-2015, especially in Sub-Saharan Africa. Many of the developed countries experienced only a slight increase in life expectancy.



Looking at this plot, it is clear that developed countries consume more alcohol than developing countries, on average. Logically, this makes sense since developed countries generally have a higher GDP than developing nations, thus they are better off, and consumers have more economic freedom to purchase and consume alcohol. We will analyze this relationship concerning developed, higher life expectancy nations during our modeling, as perhaps alcohol could have detrimental effects on developed countries' life expectancy.





As stated earlier, generally, the globe has seen a universal increase in life expectancy from 2000-2015. The distribution of nations' change in life expectancy is centered above 0, and is skewed right, indicating that there are many more nations that saw an increase in life expectancy than a decrease over this time period. Even though life expectancy generally increased from 2000-2015, it is interesting that life expectancy started to slightly decline from 2010-2015.

Change in Life Expectancy	Country
-8.1	Syrian Arab Republic
-5.8	Saint Vincent and the Grenadines
-5.3	Libya
-5.0	Paraguay
-2.3	Yemen
-2.0	Romania
-1.1	Iraq
-0.4	Estonia
-0.4	Grenada

From 2000-2015, the few nations that experienced a decrease in life expectancy are Syria, St. Vincent and the Grenadines, Libya, Paraguay, Yemen, Romania, Iraq, Estonia, and Grenada. All of these nations except for Romania are developing. For countries that experienced drastic life expectancy changes, major events outside of the scope of health policy are to blame. For instance, Yemen experienced drought and famine. Many other countries have experienced political turmoil and revolutions. We believe it is then useless to model for these specific outliers and thus don't make them a focus in the research.

Research Question 1: Given a country is developing, what can they do to increase their Life Expectancy?

Introduction and Pre-Modeling

In this section, we will apply interpretable statistical models to explore the relationships between life expectancy and various *controllable* predictors given that the country is developing. Before modeling, we create

a recipe using the `tidymodels` framework. The recipe instructs the data to first filter only countries that are marked as `Developing` , then select the response variable and the variables we noted as *controllable*. We will also keep the `Country` variable as a way to ID certain observations.

Linear Regression

We first fit a simple linear regression model predicting life expectancy from our control variables we outline in our data dictionary. As one can see from the output, hepatitis B vaccination rate and total expenditure are not statistically significant predictors, while the rest are.

Term	Estimate	P-Value	Significant?
(Intercept)	39.8454847	0.0000000	Significant
Schooling	2.1005290	0.0000000	Significant
Alcohol	-0.5173249	0.0000000	Significant
tot_expend	-0.1554012	0.0544660	Not Significant
Diphtheria	0.0328494	0.0007438	Significant
Polio	0.0288214	0.0009836	Significant
hep_b	0.0065012	0.4664656	Not Significant
pct_expend	0.0017945	0.0000000	Significant

At first glance, we notice a few interesting insights. For one, increasing schooling by one year seems to have the largest real effect on life expectancy. Schooling, which very few public health experts discuss under the lens of life expectancy, seems to have the largest impact. Second, alcohol is naturally inversely related with life expectancy, while alcohol and a country's GDP may be related to one another, as stated earlier during EDA.

A shortcoming to linear regression in this setting is that it requires many assumptions, and some of these the data does not meet. For instance, linear regression assumes that the data is *independent*. However, because the data was sampled every year, each observation is dependent on the the one before it. This is a shortcoming in the model, and provides inspiration to future research in ways we can mitigate the dependency between observations, perhaps by a bootstrapping method.

Find a Sparse Model

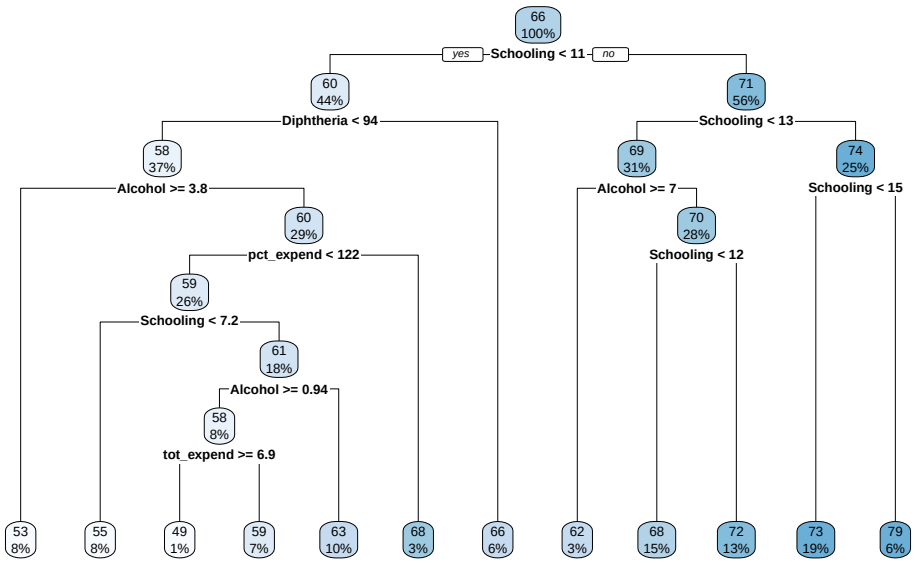
We believe a priori that life expectancy is a function of only a few of these variables given that there is such a wide variance in life expectancy that depend on factors not included in this dataset. Thus, with this belief, finding a sparse model is a natural step. Lasso regression will help us select important variables, by regularization.

Term	Estimate	Penalty
(Intercept)	39.9034294	0.01
Schooling	2.0937077	0.01
Alcohol	-0.5079514	0.01
tot_expend	-0.1454601	0.01
Diphtheria	0.0329144	0.01
Polio	0.0285251	0.01

In fitting the lasso model, we see that `Schooling`, `Diphteria`, `Polio`, `Alcohol`, and `Total Expenditure` are the variables selected. Schooling we saw had the greatest impact in our linear regression model above, which further hints at it being an important predictor. It is interesting that total expenditure on healthcare (as a percentage of total government expenditure) has an inverse relationship with life expectancy and is an important predictor.

Interaction Effects Through Trees?

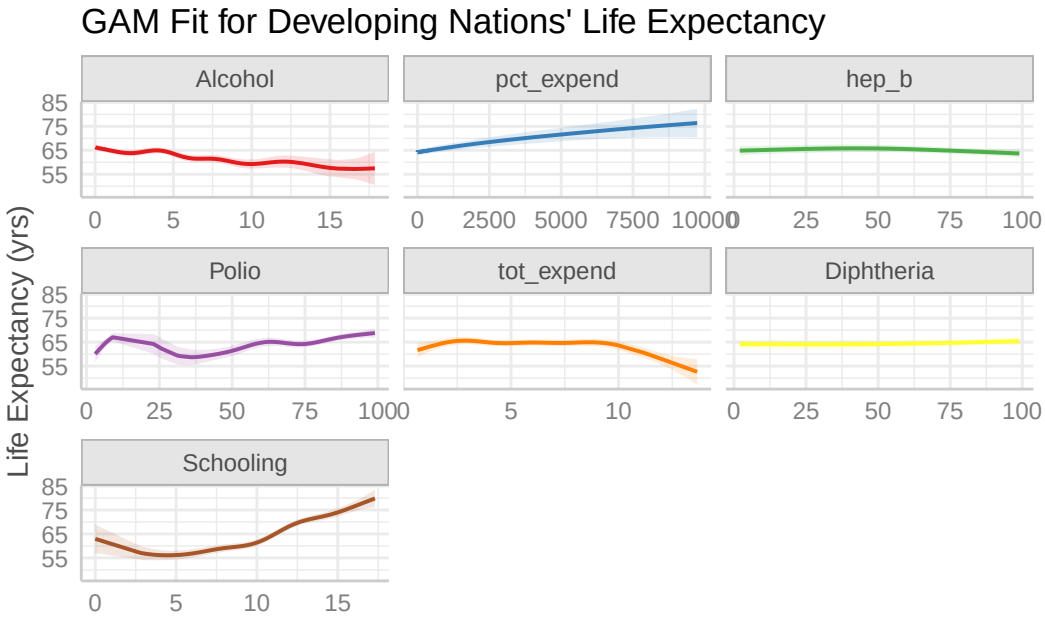
In fitting the two regression models above, we fail to see any interaction effects. Using a tree based model, we can fit a complex, nonlinear model to predict life expectancy, yet also maintain its interpretability.



We see again the importance of schooling in increasing life expectancy. Schooling is the first split implying the greatest importance of the variables. In addition, following the tree we see the same trend that we did with increases in years of educated related to increases in life expectancy. In addition, schooling is the most prevalent decision the tree makes. Following the left sub-tree, interactions between `Schooling` and other variables emerge. This is interesting given that the right sub-tree has much less interactions, and instead attempts to predict based off more schooling decisions.

GAMs

As promised, we fit a Generalized Additive Model to find nonlinear affects for each of our controllable variables. Schooling has been our most important variable this far, so we begin with analyzing the schooling variable. From the plot shown below, we see that Schooling has a quadratic shape in that increasing years of education from zero to five actually *decreases* life expectancy, but once schooling changes from five onward, life expectancy drastically shoots up. Alcohol has a clear negative relationship that seems generally linear, and percent expenditure also has a slight positive relationship with life expectancy.



These models illustrate the importance of schooling in increasing life expectancy, and we conclude that schooling is the most important aspect for developing countries to focus on in increasing life expectancy. We also suggest that given that the average number of years of education is less than 11 years, focusing on Diphtheria vaccination rates will also help increase life expectancy as noted in our tree model. Increasing government expenditure on healthcare does not simply increase life expectancy, so it is important for governments of these nations to spend money wisely on the factors we listed out, instead of spending freely on anything.

Research Question 2: For countries that already have a high life expectancy, is it economically beneficial to attempt to marginally increase life expectancy?

Introduction and Pre-Modeling

According to the Centers for Disease Control (CDC), the average life expectancy globally is roughly 75 years for women and 70 years for men, as of 2022. Since our data only contains life expectancy measures up to 2015, we will classify the top quartile of 2015 life expectancies as "high life expectancy". Subsequently, 34 countries make up our subset of nations we will consider as having high life expectancy in 2015. Interestingly, 19 of these 34 nations are classified as developing nations. This result is most likely due to the fact that a vast majority of the countries are classified as developing, so even when we take a subset of nations with the highest life expectancy, we still expect a lot of these nations to be developing. As with our first research question, we will fit models with the controllable variables as predictors on our 34 nations' data from 2000-2015 to determine just how cost efficient (or inefficient) it would be for a high life expectancy nation to further improve life expectancy. We will also do the same to the remaining nations to compare results.

Fit Lasso Models to Assess Magnitude of Effect of Each Controllable Variable

Like earlier, we fit lasso regression models to both the high-life expectancy nations data set and the non-high-life expectancy data set. These lasso models will provide interpretable results of not only which controllable variables are significant, but also how much a change in one of those variables would alter the expected life expectancy of a nation. The $\widehat{\beta}$'s for these variables allow us to assess the expected effect of a government policy affecting one of the controllable variables on life expectancy.

Term	Estimate	Penalty
(Intercept)	65.1772530	0.01
Schooling	0.7292362	0.01
tot_expend	0.0892498	0.01
Alcohol	0.0421388	0.01
Diphtheria	0.0161732	0.01

Fitting to nations with an already high-life expectancy, the significant variables selected are `Schooling`, `tot_expend`, `Alcohol`, and `Diphtheria`. The coefficient for `Schooling`, $\hat{\beta}_1$ is only 0.729. Comparing this to the `Schooling` coefficient from fitting the earlier lasso model on developing nations (2.094), it is clear that an increase in years of school results in diminishing marginal increases to life expectancy. An additional year of school for developing nations results in an expected increase in life expectancy of about 2 years, while an additional year of schooling for nations with a high life expectancy results in an expected increase in life expectancy of less than a year.

Furthermore, it would take an expected increase of government expenditure on healthcare as a percentage of total government expenditure by 11% to increase life expectancy by only 1 year, based on $\hat{\beta}_2$, the coefficient for `tot_expend`. This lasso model's output and results clearly illustrate that a marginal gain in life expectancy is not worth the required investment. The governments of these 34 nations would be better off maintaining their current expenditure on schooling and healthcare and focus their policy in other areas.

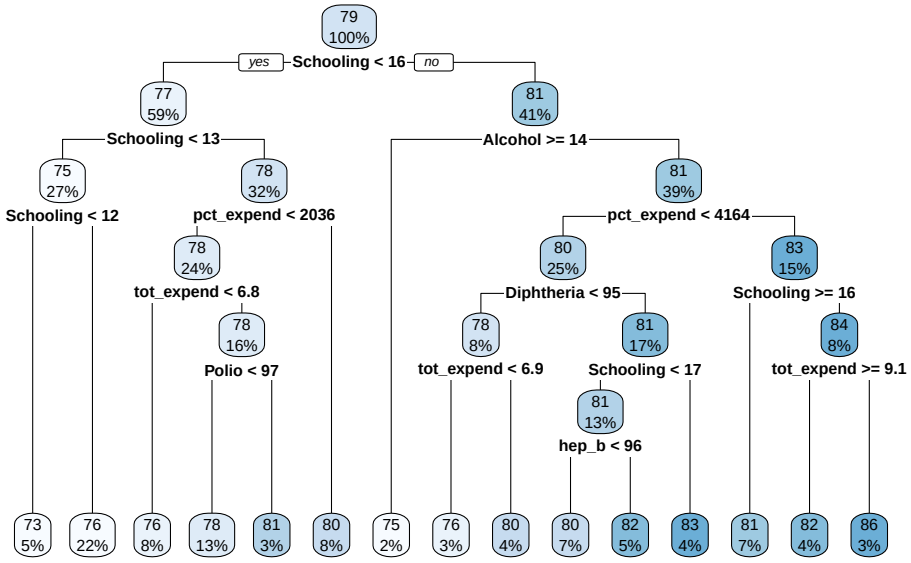
One problem to this model's fit is the significant, positive coefficient for the `Alcohol` variable, which contradicts previous results. This implies that an increase in nationwide alcohol consumption would lead to an expected increase in life expectancy, an illogical claim.

Term	Estimate	Penalty
(Intercept)	41.9469485	0.01
Schooling	1.8916517	0.01
Alcohol	-0.4691431	0.01
tot_expend	-0.3915740	0.01
Diphtheria	0.0360454	0.01
Polio	0.0348726	0.01

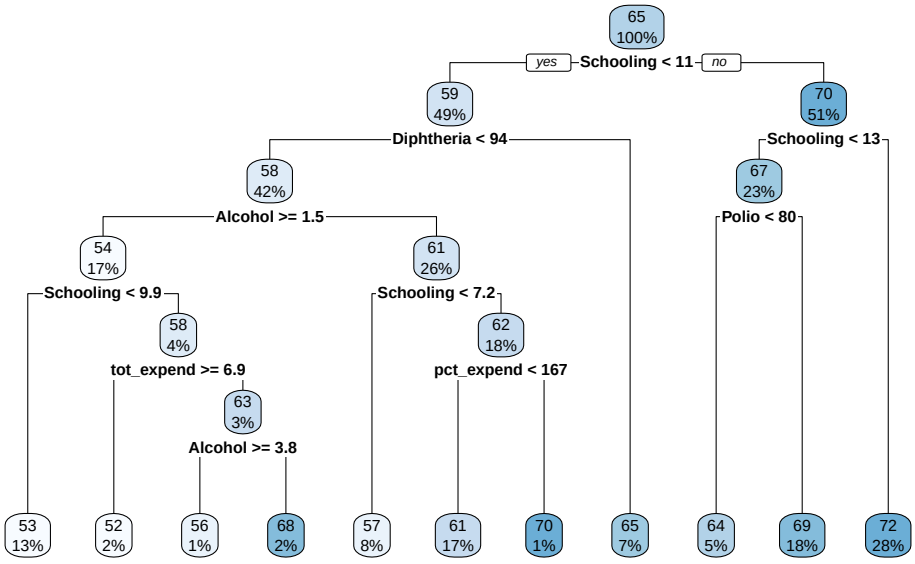
The results of the remaining nations' lasso regression model support our conclusions from the developing nations research conducted earlier. Here, we see a much more significant effect of an increase in schooling for the nations with a life expectancy outside the global top 25%. Interestingly, the model displays a negative coefficient for `tot_expend`, indicating that an increase in spending in healthcare for this subset of nations might decrease life expectancy. Regardless, this further proves that length of education is the most important influencer of life expectancy.

Fit Decision Tree to Visualize Possible Decisions

The interpretability of decision trees is very beneficial to government officials and global health officials. The visualizations of these fitted model explicitly outline the decisions a government could make to increase life expectancy. While these models are for a subset of nations rather than individual nations, a policy-maker in a nation could ultimately use these generalized results for nation similar to theirs and employ corresponding policy or spending packages to increase their nation's life expectancy.



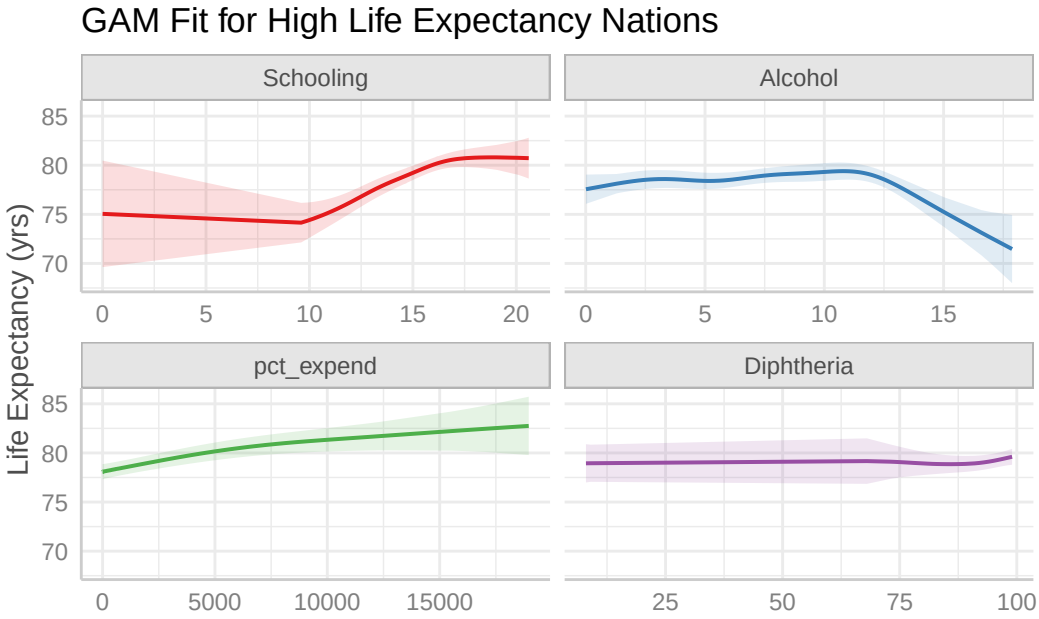
This decision tree's results mostly follows our previous conclusions. The first decision, thus the most important predictor is again `Schooling`. However, it is worth noting that this primary decision is made off years of education being greater or less than 16 years. The mean years of schooling for high life expectancy nations is only 14.74 years. The fact that this primary decision is being made at a value of `Schooling` that is well above the mean is concerning, indicating that it is expected to take a significant increase in education to increase life expectancy further.



Again, this decision tree for non-high life expectancy nations indicates that `Schooling` is the most important predictor of life expectancy. The rest of the tree is similar to the tree from the first research question, that after schooling, raising immunization rates is a method of increasing life expectancy in the long run.

Generalized Additive Model to Analyze Nonlinear Relationships

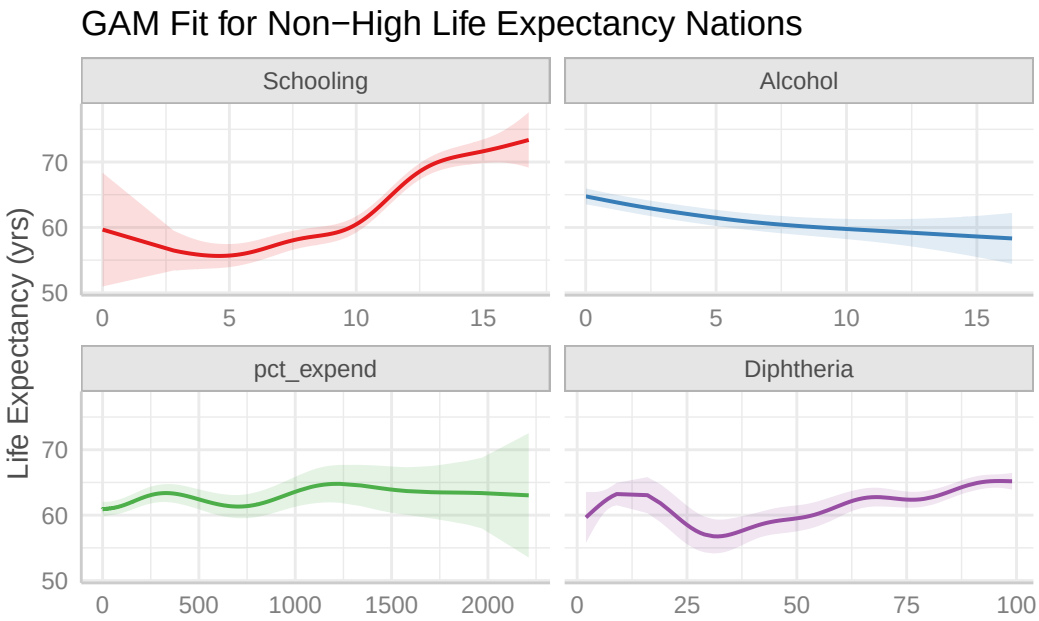
The results from lasso and decision trees indicate that increases in years of education lead to diminishing increases in life expectancy, and we considered that it may not be worth the further investment into education for nations with already high life expectancy. Also, both models' results indicate that there could possibly be nonlinear relationships between the predictor variables and the response of life expectancy. So, we decided to fit a generalized additive model (GAM) on the four selected significant predictors chosen by lasso earlier, as to not overcomplicate our mdoel. We trust the feature selection of the lasso model, and thus fit a GAM predicting life expectancy from `Schooling`, `Alcohol`, `pct_expend`, and `Diphtheria`. Using this model will allow us to understand more complex relationships between the predictors and the response, while maintaining interpretability for government officials and policy makers.



The plot of the generalized additive model fit outlines nonlinear relationships for all four chosen predictors. The results for `Schooling` follow that similar of the lasso plot, that increasing years of education increases life expectancy, but we now see that increased alcohol consumption decreases life expectancy. Also, vastly increasing expenditure on healthcare (and specifically Diphtheria immunizations) for nations with high life expectancy only marginally increases life expectancy. While these results are telling, we would like to compare these results to the results of a GAM fit on the non-high life expectancy nations.

ModelType	MSE
Lasso	5.91
Tree	5.15
Generalized Additive Model	4.84

The GAM for nations with high life expectancy provides a new lens of interpretability, but also better predictive performance than the lasso and tree model.



With the non-high life expectancy nations, the effect of schooling is much stronger. It is clear that from the fitted GAM that increasing years of education has a drastic effect on life expectancy for these nations, indicated by the sharp increase in life expectancy going from 5-12 years of schooling, as well as the low standard error around that portion of the fit. Based on the results of these fitted GAMs, increasing schooling has a lesser effect of increasing life expectancy for countries with high life expectancy. For these nations, it might not be in their best national interest to further invest in education to improve global health. Also, the detrimental effect of alcohol consumption on life expectancy is much stronger for high life expectancy nations. Perhaps, high life expectancy nations that are generally better off and more economically independent and consume more alcohol and have a more unhealthy relationship with alcohol. This raises the question of whether these better-off nations' governments should address their people's alcohol consumption from a global health perspective.

Fitting these generalized additive models tells us that it is tough for both groups of nations to alter life expectancy by solely increasing spending on healthcare and immunizations. Also, they further prove that increasing years of education leads to only marginal increases in life expectancy for healthy nations. Finally, the GAMs helped us discover this sharp, detrimental effect of alcohol consumption for high-life expectancy nations, a relationship the other collective nations do not necessarily have.

Conclusion

In this study, we investigated the relationship between life expectancy and numerous factors using interpretable statistical models. For countries that are still developing, schooling has the strongest positive relationship with life expectancy, and we suggest to countries that are developing to fund and prioritize education. The regression models and tree model both provide the necessary evidence to support this conclusion. Outside of funding education, we suggest governments of developing nations to tactically spend their healthcare expenditure allocation on immunizations, like Diphtheria, as we also found that factor to be a significant predictor of life expectancy. Our results show that it is not neccessary for governments of these nations to increase `pct_expend`, or the proportion of total government spending on healthcare, rather, it is crucial for them to priortize education and make sure an approriate portion of their public healthcare allocation is going towards immunization research and production.

For nations that already have a high life expectancy, schooling also has the strongest positive relationship to life expectancy, but this relationship is a lot weaker than that for nations with non-high life expectancy.

For these well off nations, an increase of one year of education is expected to have a lot lesser effect on life expectancy than other nations. As stated earlier, the average number of years of schooling of these nations is roughly 15 years of education, corresponding to a little more than a high school education. In Australia, a nation with high life expectancy, an average year of at an undergraduate university can be up to $45,000 (AUD), according to Study Australia. From a global health perspective, it would be ridiculously costly for the government of Australia to spend billions of dollars on public, collegiate education with the sole purpose of increasing life expectancy, especially since the expected corresponding increase in life expectancy is marginal.

We found that these nations that are typically well-off also have a detrimental relationship with alcohol. In fact, our modeling suggests that the negative effect of alcohol consumption outweighs any positive effect of other factors on life expectancy. Thus, we also suggest that the governments of these well-off nations shift funding towards government programs related to public alcohol and other drugs education and guidance, safety, and rehabilitation.

Finally, our research demonstrates that expenditure on healthcare shows a slight positive relationship with life expectancy. It would not be optimal for these high life expectancy nations to heavily shift their expenditure towards healthcare, rather these nations' global health innovation could be driven by slow and steady increase in investments in healthcare, immunizations, and global health. Aside from the numbers and expenditure, it is most important that all nations' governments, developing or not, high life expectancy or not, well off or not, recognize the relevance of global health studies and innovation in order to further progress this world's collective society.

Appendix

Data Dictionary

Controllable Variables

- `Alcohol`: Alcohol consumption per capita (liters of pure alcohol)
- `pct_expend`: Expenditure on health as a percentage of GDP per capita
- `hep_b`: Hepatitis B immunization rate among 1-year-olds (%)
- `Polio`: Polio immunization rate among 1-year-olds (%)
- `tot_expend`: Government expenditure on healthcare as a percentage of total government expenditure
- `Diphtheria`: Diphtheria tetnus toxoid and pertussis immunization rate among 1-year-olds (%)
- `Schooling`: Average number of years of schooling

Uncontrollable Variables (Nuisance Variables)

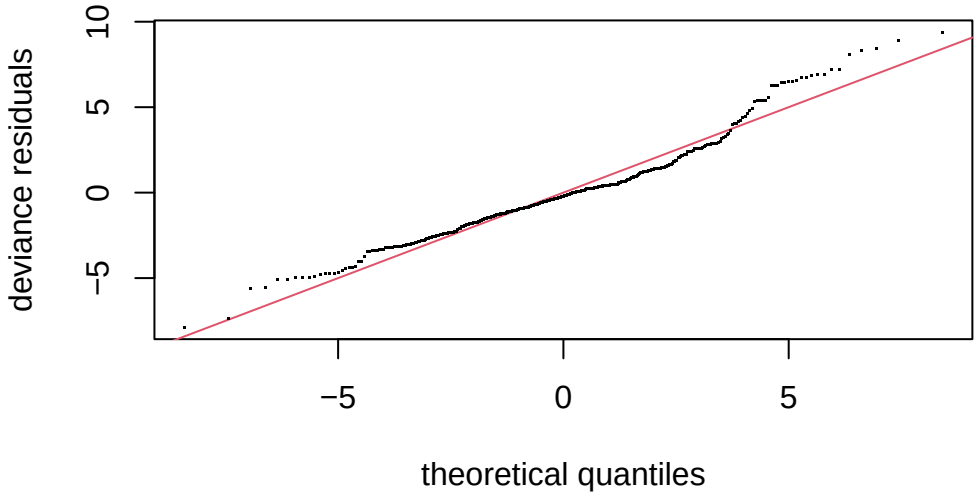
- `BMI`: Average BMI (Body Mass Index) of entire country's population
- `GDP`: GDP per capita
- `Population`: Total population of country
- `thinnes_adole`: Prevalence of "thinness" among adolescents aged 10-19 (%)
- `thinness_infant`: Prevalence of "thinness" among infants aged 5-9 (%)
- `income_comp`: Human Development Index in terms of income composition of resources (0 to 1)
- `Status`: Developmental status of country (Developed or Developing)

Indicator Variables

- `under_five_deaths`: Number of deaths of 5-year-olds or younger per 1000 people
- `life_exp`: Average expectancy in country (years)
- `adult_mortality`: Number of deaths of people aged 15-60 per 1000 people
- `infant_mortality`: Number of infants deaths per 1000 infants
- `hiv_aids`: Number of deaths of 0-4 year-olds from HIV/AIDS per 1000 live births

Model Diagnositcs

QQ-Plot of GAM Fit for High-Left Expectancy Nations



QQ-Plot of GAM Fit for Non-High Life Expectancy Nations