Life Expectancy

Dillan Sant, Chase Mathis

Introduction

Humans have experienced a meteoric 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] From a young age, we are taught to have a healthy diet, exercise regularly, and keep up personal hygiene to stay healthy and therefore live longer. However, macro-factors clearly play a role in how healthy ones life is. Health policy and its relationship with these macro-factors greatly impacts the health of individuals as this most recent pandemic has shown.

This research project aims to investigate the relationship between macro-factors and life expectancy for a 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 what is the most important?

(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 finding a "miracle" vaccine, or is there still work to be done for cheaper issues like vaccination and schooling.

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 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 off 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 one 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 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 the same method 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] https://www.health.harvard.edu/blog/why-life-expectancy-in-the-us-is-falling-202210202835#:~:text=A%20dramatic%20fall%20in%20life,just'

Data

The Global Health Observatory (GHO) 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 is a country 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, 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

```
# A tibble: 5 x 3
  name       value percent_missing
  <chr>      <int>           <dbl>
1 Population   652          0.222
2 hep_b        553          0.188
3 GDP          448          0.152
4 tot_expend   226          0.0769
5 Alcohol      194          0.0660
```

The data is mostly complete, but certain features are 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 na 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 using the mean. We continue with this methodology.

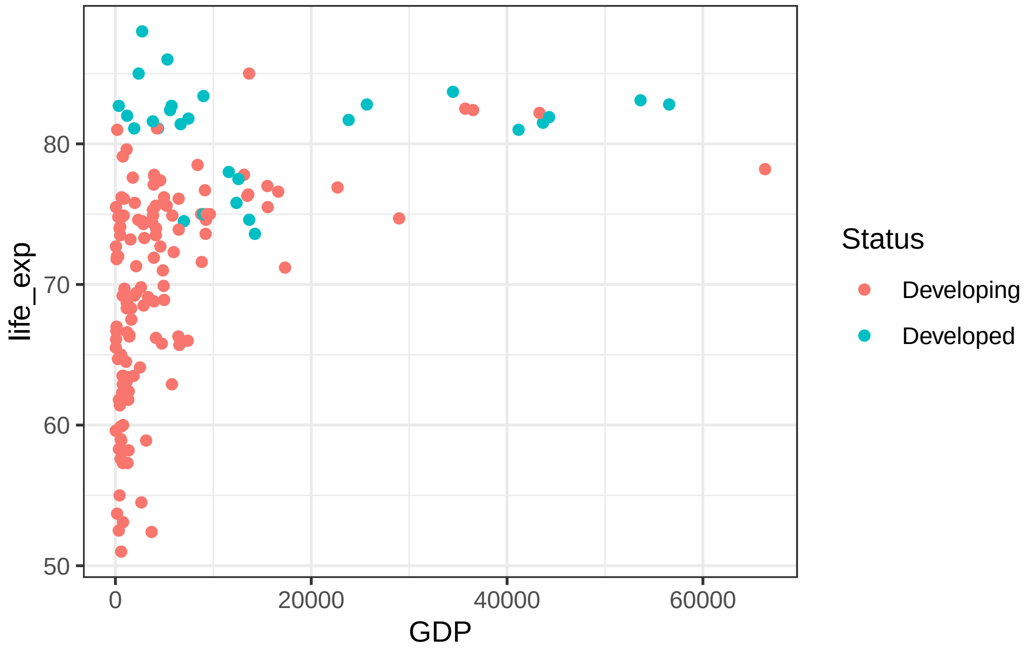
After filling in na values with the mean, we are able to decrease the percent of na-values from 43.9% to 27.8%. This reduction will help us use more of the data so that we can find better conclusions for policy makers. We did not check to see the distributions of before omitting the na values match the distributions after***.

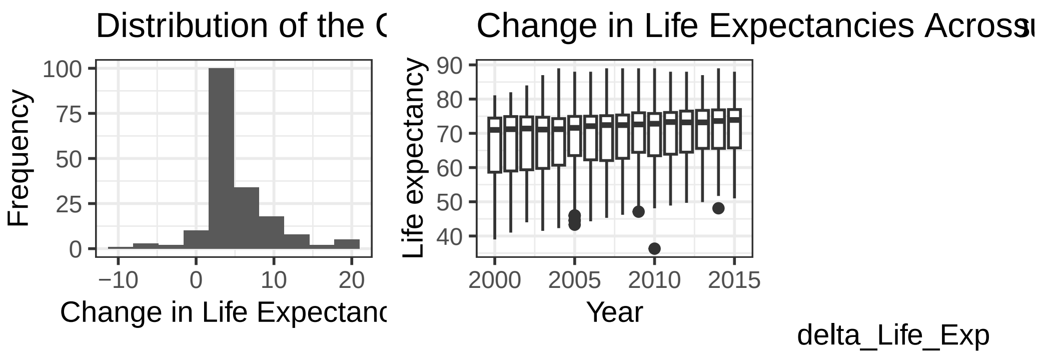
Exploratory Data Analysis

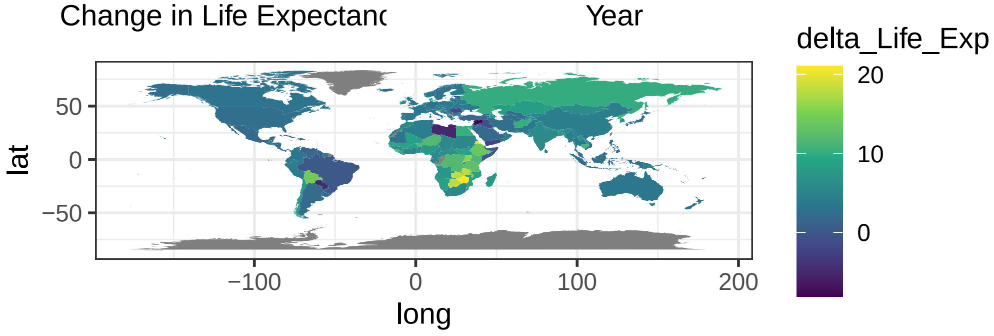
Table 1: Life Expectancy of the Top 10 Countries

Country	Life Expectancy
Slovenia	88.0
Denmark	86.0
Chile	85.0
Cyprus	85.0
Japan	83.7
Switzerland	83.4
Singapore	83.1
Australia	82.8

Country	Life Expectancy
Spain	82.8
Iceland	82.7







Generally countries increased life expectancy, especially in Sub-Saharan Africa. Which countries experienced a decrease in life expectancy.

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 nations that experienced a decrease in life expectancy are Syria, St. Vincent and the Grenandines, Libya, Paraguay, Yemen, Romania, Iraq, Estonia, and Grenada. All of these nations except for Romania are developing.

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 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.

At first glance, I 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 seems to have the largest impact! Second, Alcohol is naturally inversely related with life expectancy, while alcohol and a countries GDP may be related to one another.

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.

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

Term	Estimate	P-Value	Significant?
pct_expend	0.0017945	0.0000000	Significant

Find a Sparse Model

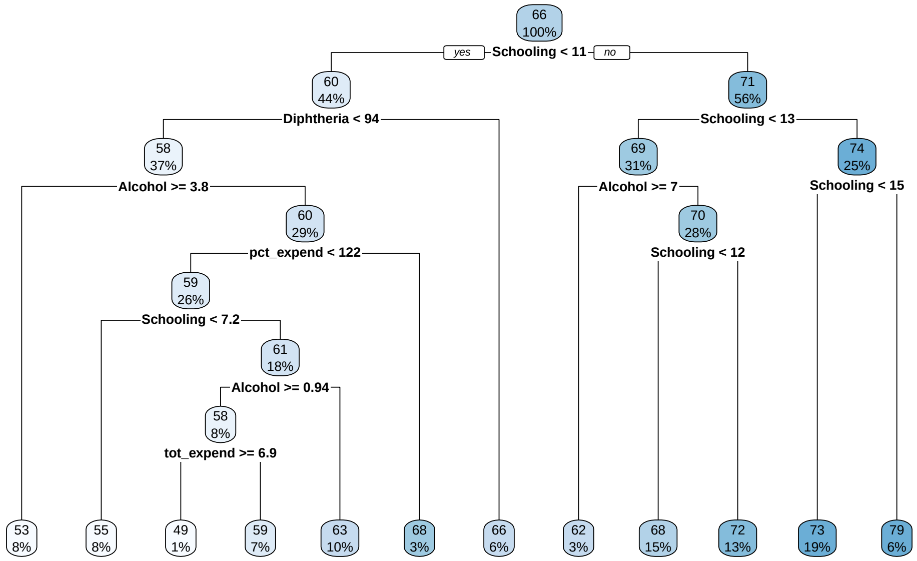
We believe a priori that average life expectancy is related to only a few of these variables given that there is such a wide variance in life expectancy that depend on factors not included in this dataset. Thus, with this belief finding a sparse model is a natural step. Lasso regression will help us select important variables, by regularization. In fitting the lasso model, we see that `Schooling`, `Diphteria`, and `Polio` 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. Rates of vacinnation for Diphteria and Polia are also deemed important predictors.

Chase, shouldn't you filter when abs(estimate) > 0.01? Cause alcohol is would be selected; feel like we shoudn't omit the negative betas

Term	Estimate	Penalty
(Intercept)	39.9034294	0.01
Schooling	2.0937077	0.01
Diphtheria	0.0329144	0.01
Polio	0.0285251	0.01

Interaction Effects Through Trees?

In fitting the two regression models above, we fail to use any interaction effects. Using a tree based model, we can fit a complex, nonlinear model to predict life expectancy, yet also maintain its interpretability. The tree model illustrates the importance of schooling in increasing a country's life expectancy. The first decision the tree makes is off of schooling. 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 School decisions.



Include in first research question: For the small subset of countries that saw a decrease in life expectancy from 2000-2015, what factors led to this decrease in life expectancy?

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, 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. Note that the United States is not included in this subset. 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 variable 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 ony 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`, $\widehat{\beta}_1$ 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. **Find online data on how much an extra year of high-level schooling costs the government in country x** Consequently, an increase of investment in education for these high life expectancy nations is not worthwhile…

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 $\widehat{\beta}_2$, 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.

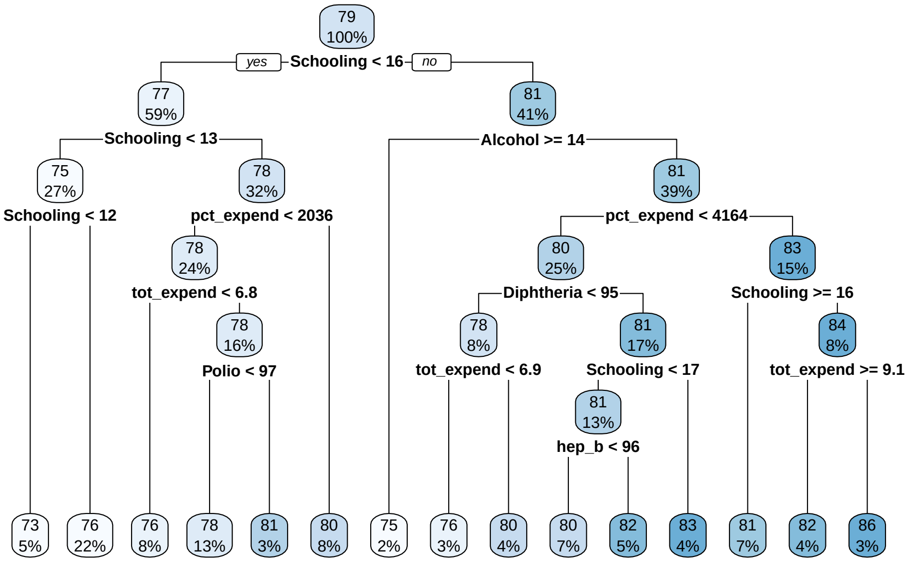
One problem to this model's fit is the significant, positive coefficient for the `Alcohol` variable, which contradicts previous results. This implies that an increase in nationwide alcohol consumption would lead to an expected increase in life expectancy, an illogical claim. It may be worth nothing that wealthier, high life expectancy nations could not experience the same magnitude negative effects of alcohol as other nations do.

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

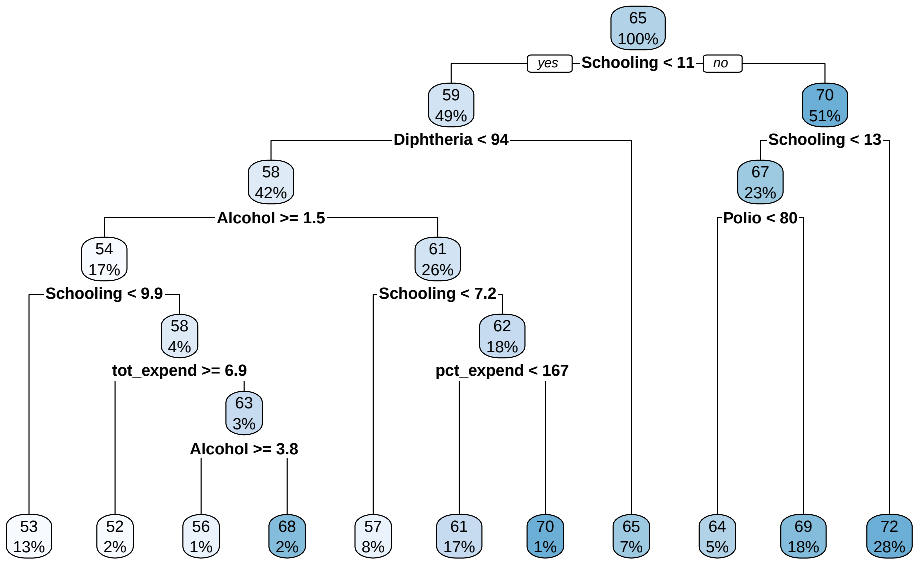
The results of the remaining nations' lasso regression model support our conclusions from the developing nations research conducted earlier. Here, we see a much more significant effect of an increase in schooling for the nations with a life expectancy outside the global top 25%. Interestingly, the model 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 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 significant increase in education to increase life expectancy further.

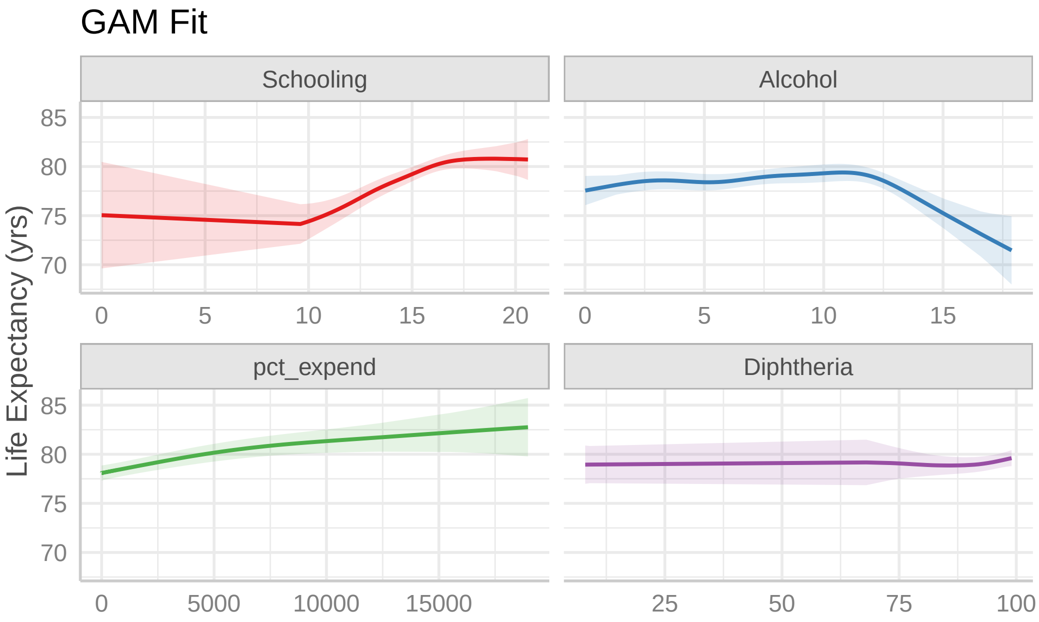


Again, this decision tree for non-high life expectancy nations indicates that `Schooling` is the most important predictor of life expectancy.

If space needed, move this tree to appendix

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`.



ModelType	MSE
Lasso	5.91
Generalized Additive Model	4.84

Notes

Introduction/EDA

- start with providing scientific context, refer to article
- shift towards problem, introduce research questions
- Context and previous research
- Why is this important? Goal
- Overview of the rest of the paper. Can do after
- give detailed description of data (see rubric), which predictors are uncontrollable, controllable, indicators
- start EDA, show some simple, interpretable plots regarding different predictors, find different relationships among controllable variables

Modeling

- Start with linear regression
- check out interaction effects
- ridge, lasso (for interpretable variable selection)
- trees
- stay away from uninterpretable methods like random forests, boosting

Within Research Questions

- focus on answering research question, using data/modeling merely as support for argument

- make sure models and its results would be interpretable for global health professionals and governments
- plot model diagnostics to assess models, make tables of results/predictions of models
- give suggestions based on results to policy makers (ex. "this nation should put a greater prorportion of their total expenditure into health care to increase life expectancy")

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 having 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.*** Question 2.

Conclusion

- suggest in which factors specific nations should invest their money in based on modeling during both research questions, or, suggest not to increase investment in health care for nations with already high life expectancy
- reference models, focus on interpretability and policy actions

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 (Nuiscance 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