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. 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.

______ collected the data. The data has 21 features, which we will outline below. Each observation in our data is a a country, the year. As the features are summary statistics, we are predicting averages from averages. We will *not* use black-box models such as random forrests or baggging 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

** Dillan can you do as you found the data**

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

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

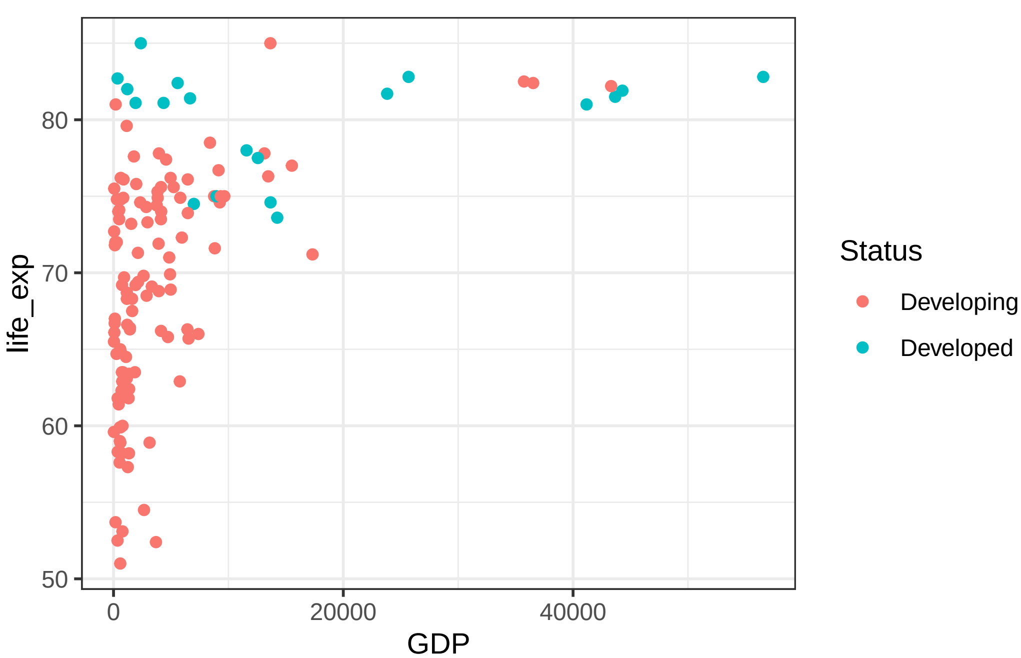
will split up our missing values into two categories: (1) Missing values for entire countries and (2) Missing values for time ranges within a certain country. In regards to the first type of na value, we are left with little options. Many types of models will throw errors if there is missing data, so 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.

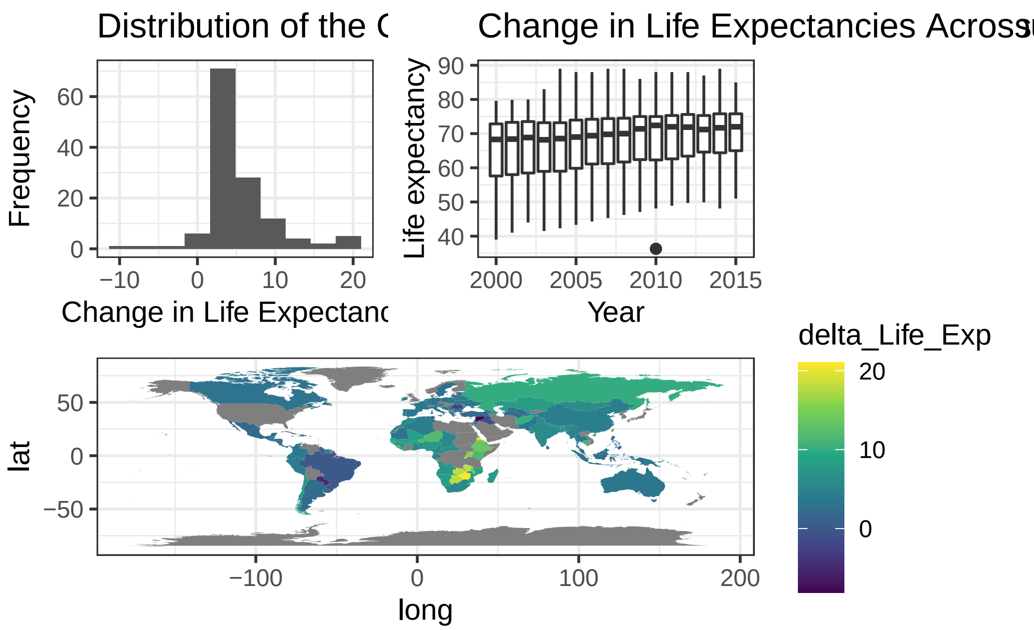
After filling in na values with the mean, we are able to decrease the percent of na-values from 43.9% to 27.8%. This reduction will help us use more of the data so that we can find better conclusions for policy makers. We did not check to see the distributions of before omitting the na values match the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

Exploratory Data Analysis

Table 1: Life Expectancy of the Top 10 Countries

Country	Life Expectancy
Chile	85.0
Cyprus	85.0
Australia	82.8
Spain	82.8
Italy	82.7
Israel	82.5
France	82.4
Sweden	82.4
Canada	82.2
Luxembourg	82.0





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
-5.0	Paraguay
-2.0	Romania
-1.1	Iraq
-0.4	Estonia

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?

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, this regression model shows that HIV/AIDS vaccination rates are inversely related to life expectancy. One conclusion of this finding is that in countries where HIV/AIDS is not a problem, citizens are not vaccinated for it. In countries where it is a problem, more people are vaccinated for it, and more people die from it. **This is a short come of our limited data and should be noted. Alcohol** is naturally inversely related with life expectancy, while alcohol and a countries GDP may be related to one another.

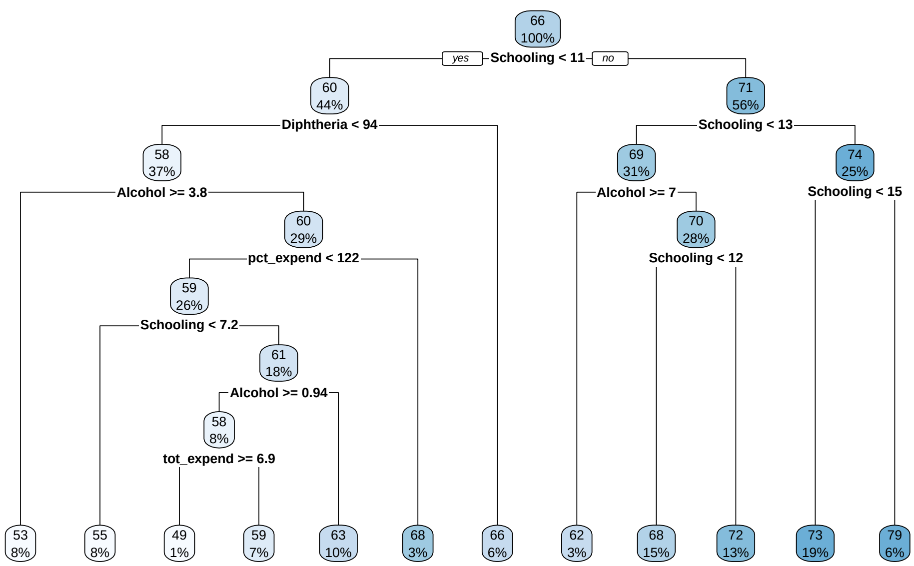
Term	Estimate	P-Value	Significant?
(Intercept)	39.8454847	0.0000000	Significant
Schooling	2.1005290	0.0000000	Significant
Alcohol	-0.5173249	0.0000000	Significant
tot_expend	-0.1554012	0.0544660	Not Significant
Diphtheria	0.0328494	0.0007438	Significant
Polio	0.0288214	0.0009836	Significant
hep_b	0.0065012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

Find a Sparse Model

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?

Very much depends on schooling. Also indicated in our other models. Schooling is THE most important thing to do in terms of increasing life expectancy.t



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?

Notes

Introduction/EDA

- start with providing scientific context, refer to article

- shift towards problem, introduce research questions
- 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

- 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