

# 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.<sup>1</sup> 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.

---

<sup>1</sup><https://www.health.harvard.edu/blog/why-life-expectancy-in-the-us-is-falling-20210202835#:~:text=A%20dramatic%20fall%20in%20life,just>

## Data

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

\_\_\_\_\_ collected the data. The data has 21 features, which we will outline below. Each observation in our data is 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 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
```

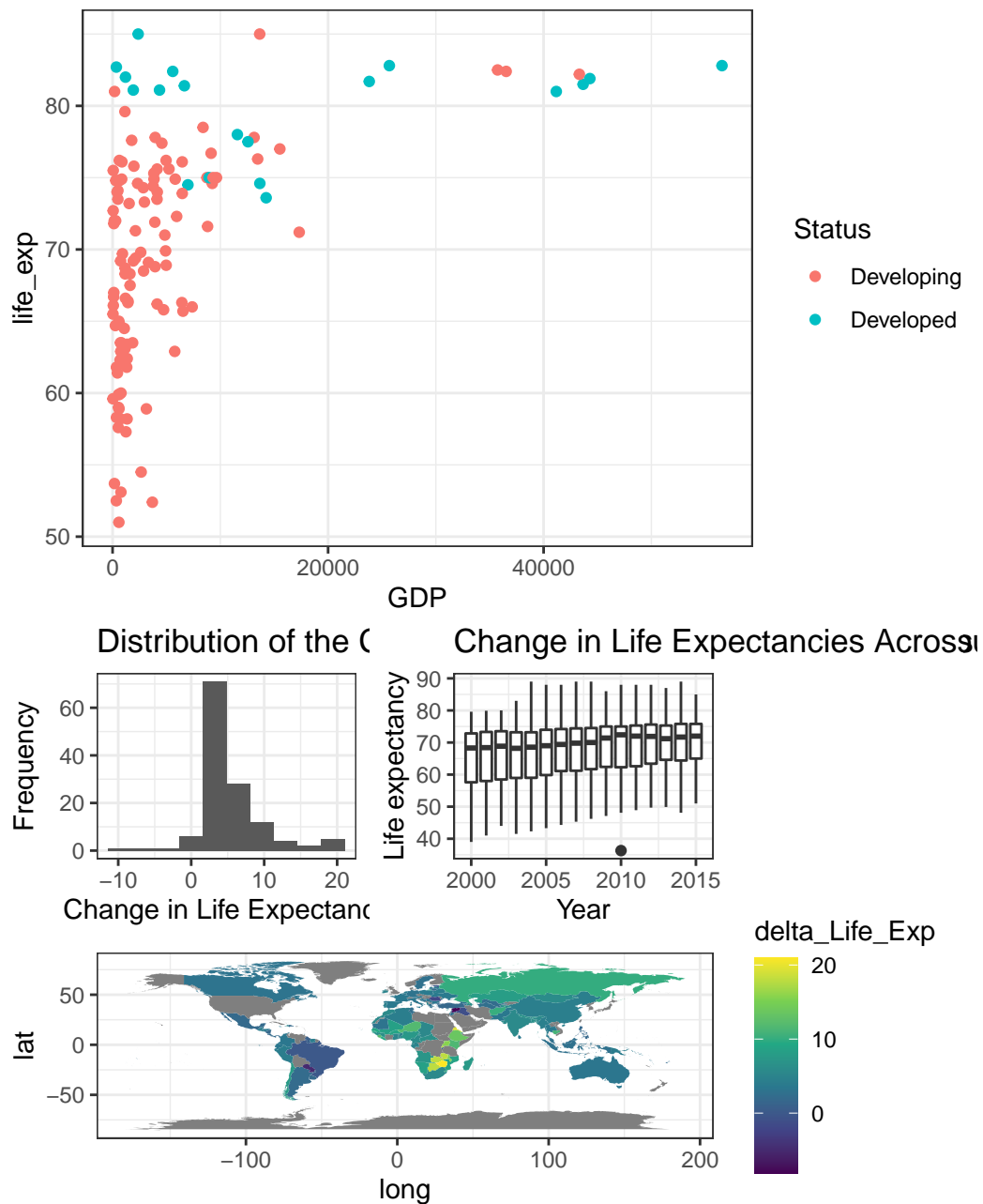
The data is mostly complete, but certain features are missing values. The missing data will have an impact on our modeling and interpretation which is something to consider for future work. For the time being, we will split up our missing values into two categories: (1) Missing values for entire countries 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
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 Grenadines, 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 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
pct_expend	0.0017945	0.0000000	Significant

### Find a Sparse Model

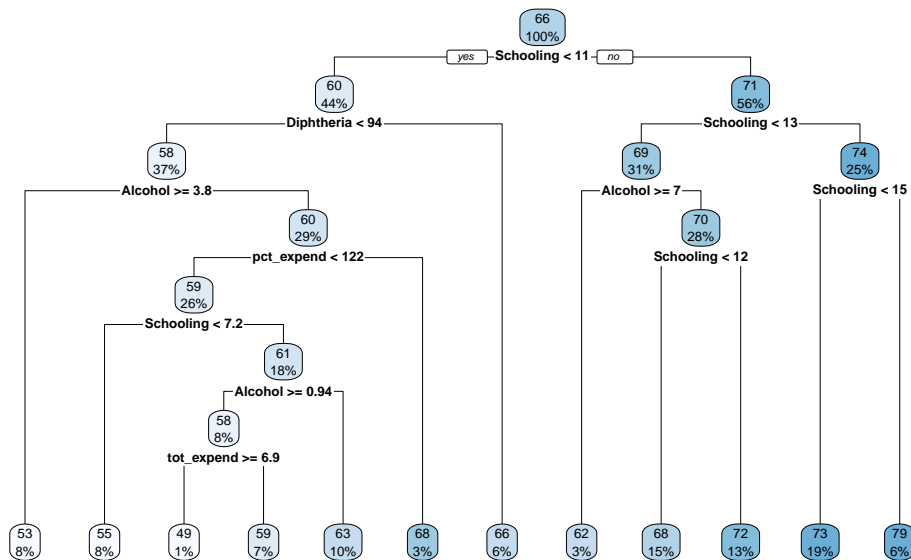
We believe apriori that average life expectancy is related to only a few of these variables given that there is such a wide variance in life expectancy that depend on factors not included in this dataset. Thus, with this belief finding a sparse model is a natural step. Lasso regression will help us select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

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

Term	Estimate	Penalty
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?

## 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 proportion 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 tetanus 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