What Best Determines a Country's Life Expectancy?

GG Gods: Aaditya Warrier, Debosir Ghosh, Alexandra Lawrence, Julia Rosner

2021-04-27

Introduction

According to an article by NPR, "the average U.S. life expectancy dropped by a year in the first half of 2020" (Wamsley 2021). While COVID-19 and drug overdose are credited with being the leading causes of this trend, we were motivated to understand the factors determining life expectancy in general. By understanding the biggest contributors to life expectancy, we will hopefully get an idea of how to maximize a country's life expectancy and reverse the current trends in the US. Numerous studies have previously examined which factors are the most influential in determining life expectancy. In a study conducted by the International Institute for Applied Systems Analysis, education was deemed the most important predictor, even more than income (Lutz 2018). More specifically, it study suggested that better education results in both better health and higher incomes that in turn affect life expectancy; we will further explore the validity of this claim in our report.

The Life Expectancy dataset contains 19 variables and analyzes several factors affecting life expectancy. Unlike several past studies, it also takes into account components of the HDI such as mean years of schooling as well as predictors like immunization rates and health care expenditure. Each observation is a single country, and its corresponding information for each variable. This data was collected by the Global Health Observatory - a initiative of the WHO - from a period of 2000 to 2015 for 193 countries around the world and includes a variety of economic, social, and health related factors. The WHO uses a five step package to make usable health related data. This includes surveying the population for health risks, counting the number of births and deaths, and optimizing the health services data. Health facilities around the world are required to submit regular reports on observed conditions, and the quality of data is assured by reviews within the WHO. The economic factors in this dataset were found from data recorded by the United Nations, which is based primarily on national sources.

We plan to use life expectancy as our main response variable in order to determine which factors have the biggest impact on a country's life expectancy, also analyzing if any of these factors impact each other. Broadly speaking, we will focus on education, income and health care/public health, the factors we expect to have the biggest effect on a country's life expectancy.

GHO Data Repository: https://apps.who.int/gho/data/node.main

General research question: What best predicts the life expectancy of a particular country? Do education, income, and healthcare/public health account for the majority of the variability in a country's life expectancy, or are other factors responsible?

Hypothesis: We can best predict a country's average life expectancy by using its mean number of years of schooling in years, average immunization rates (a country's numerous immunization types - Hep B, Diphtheria, etc. - all averaged together), and the country's GDP per capita. Out of these three factors representing education, healthcare/public health, and income, we think that income will account for the most variability in a country's life expectancy.

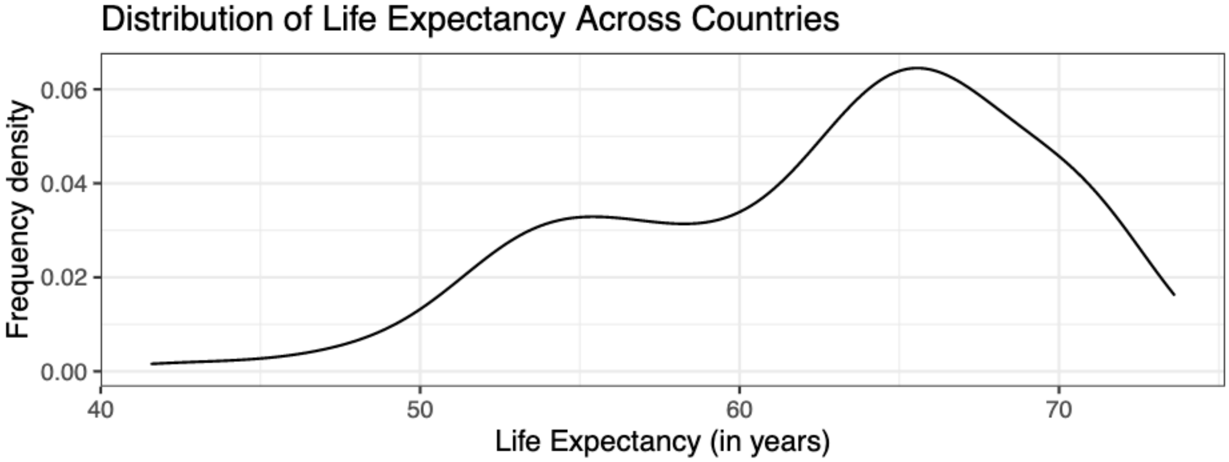
Methodology

Our goal is to produce a multiple linear regression model that can be used to accurately predict the life expectancy for a country. To begin, we wanted to explore our response and our hypothesized predictors to understand what we can expect to see in our final model. We also looked at other predictors related to the three areas we had previously identified, along with examining the Democracy Index - a metric provided by the Economist Intelligence Unit which gives an overall classification of how democratic a nation is - something we thought could provide interesting insights in today's fraught geo-political climate. For our analysis, we will start with a model containing all variables from the dataset, and then we will conduct backward selection to generate a model without unnecessary predictors. Finally, we will use a nested F test on any predictors in the final model that we are still skeptical of, to ensure everything is statistically significant.

EDA

A quick note on the data: Unfortunately, after delving deep into our initial data set we obtained from Kaggle that was purportedly from the GHO, we discovered numerous systematic errors scattered through the data; a large number of implausible values for certain predictors, misplaced decimal points, and more. Therefore, we had to reconstruct our data based on individual datasets from both the GHO and the World Bank. This led to some formatting issues and some data loss, both of have been dealt with in the markdown document. Additionally, missing values have been removed in order for analytic techniques to be applied.

Response variable Below, a density plot has been used to analyze distribution:

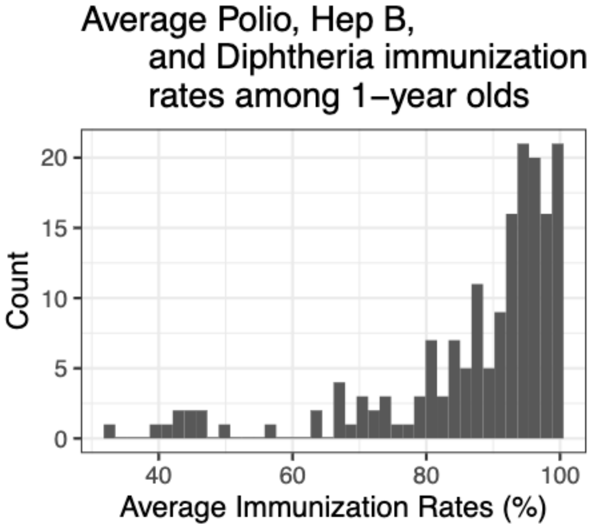


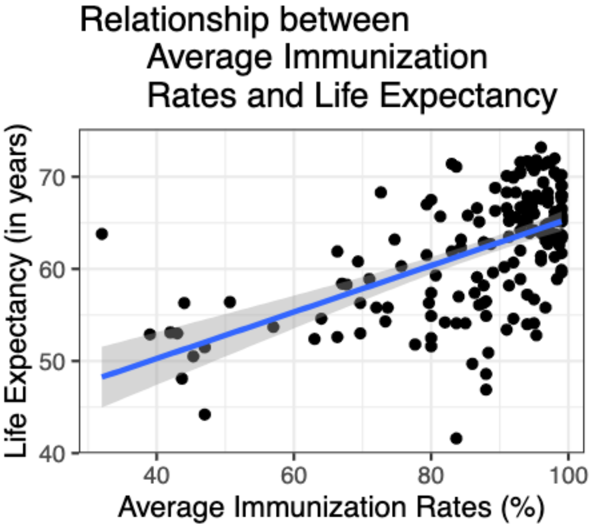
From the visualization above, we can see that the largest proportion of countries have an average life expectancy of between roughly 60-68 years old, with a median of 63.9 (appendix, ITEM 0). These are likely newly industrializing and recently industrialized countries, which make up the majority of the world's population. Another large group of countries can be found at the 52-57, with a peak of about 53 years old; these are likely less developed countries that are yet to fully industrialize. Finally, there is a small bump at about 70 years old and a subsequent trailing off, representing the most developed countries which make up an relatively small proportion of all countries. No country has a life expectancy of less than 41.6 years and greater than 73.6 years.

Predictors and their relationship to the response Due to the sheer number of predictors, we chose a selection of them to look at. To begin with, we will look at predictors related to **public health and healthcare**. First, we look at the prevalence of measles, using a newly mutated variable "Measles cases as proportion of population."

The median of measles cases as proportion of population is 1.9841959×10^{-6}, and the IQR of measles cases as proportion of population is 1.5077593×10^{-5}. Thus, most countries have an exceedingly small proportion of their population who had measles in the past year, given that the median of measles cases as proportion of population is very close to zero. Also, there is little variation among countries, as seen in the IQR and in the visual of the distribution of measles cases as proportion of population (the visual can found in the appendix, ITEM 1). Hence, this will likely have little impact on the response.

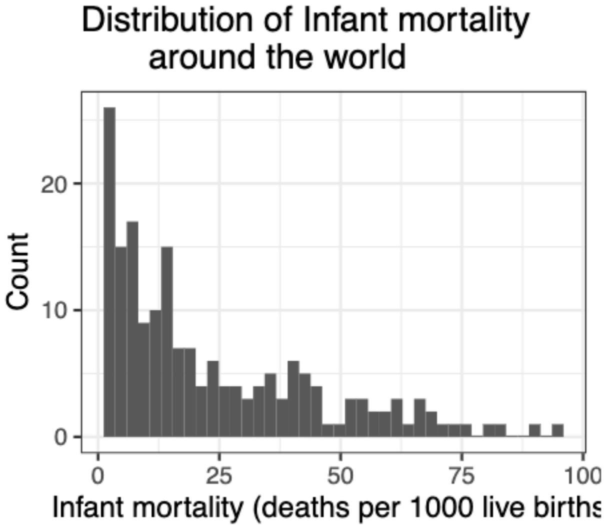
Next, we look at average immunization rates among infants across the countries. These will be averaged out using individual immunization rates from the polio, diphtheria, and hepatitis columns.

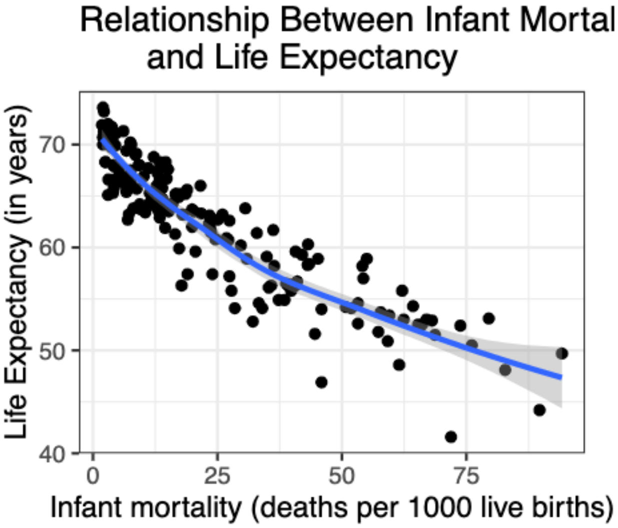




The graph on the left shows an incredibly left-skewed distribution. Far more countries have higher average immunization rates than lower ones, with many being clustered in the 80-100% region. We see a smaller, but significant, number of countries that have abysmally low rates of 35% - 45% and a similar group of countries with rates just above 60%. Within the 80% - 100% band, there are two prominent modes at around ~92-93 % and at near 100%. The variability in values indicates a possible relationship with the response. On the right, we do see a positive association, albeit only a weak-to-moderate one. However, it is notable that there *are* a few points that are anomalously large outliers (see the point at 32, 64 or 82, 42), which may be influential points significantly affecting the regression line. We will explore this predictor more in the model.

Our next predictor is Infant Mortality, the number of deaths per live births before an infant reaches the age of 1:



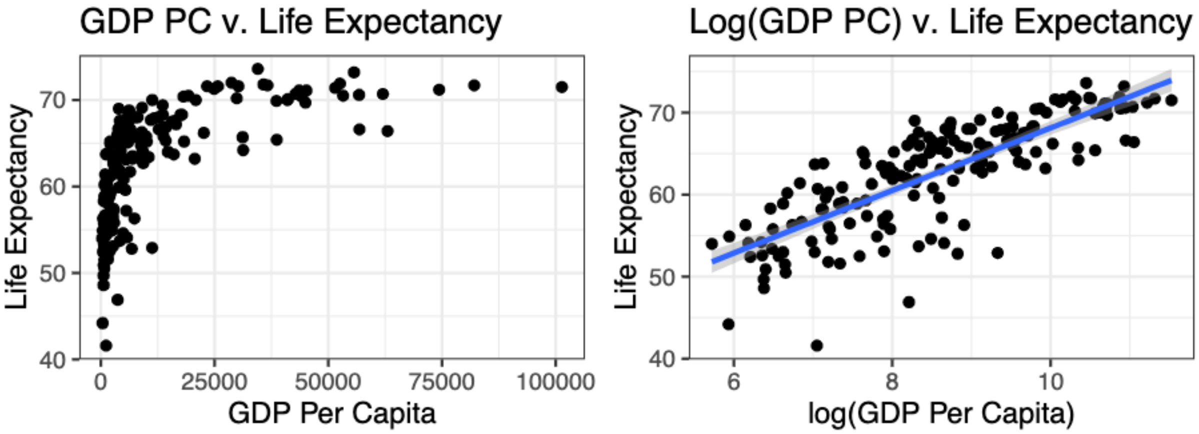


The visual on the left shows how the distribution of infant mortality predictor is extremely right-skewed, with infant mortality being somewhere between 0 and 25 deaths for most countries, and a significant trailing off of data past ~50 deaths. The highest infant mortality rate is at about 95 deaths per 1000 births. From the right visual, we observe a strong, negative, but not entirely linear association with the response variable, so we use a non-linear geom_smooth. We observe that as infant mortality increases, life expectancy decreases; the rate at which this happens slows as infant mortality increases.

Our next health-related predictor will be health expenditure as a % of GDP. This is not a direct measure of people's health, but better describes how a country's government values healthcare. We will add an interaction with the Democracy Index here; health expenditure in less democratic countries could be ineffective due to corruption. Due to the health expenditure as a % of GDP predictor's limited range, we log-transformed it to explore its relationship with the response.

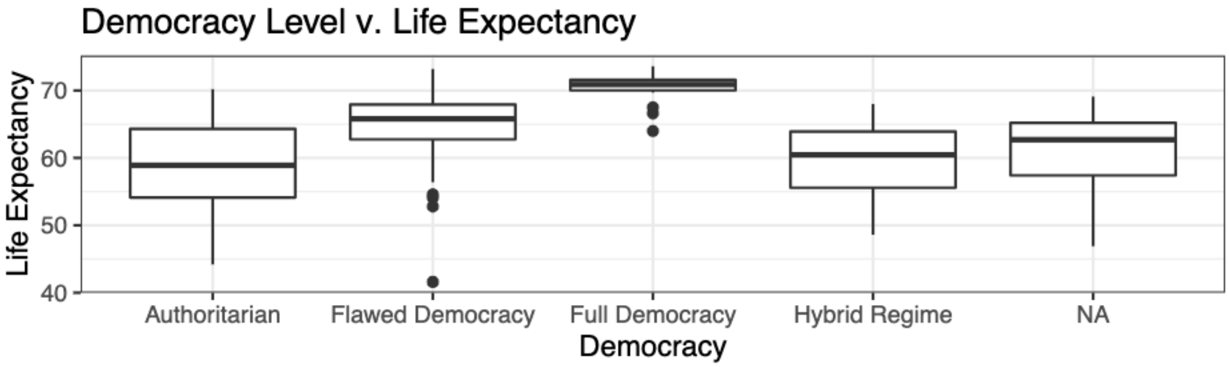
The distribution of CHE as % of GDP in different countries (see appendix ITEM 6) presents a nearly normal distribution centered at about 6% (median = 1.8229351) with an IQR of 0.5564918 (appendix ITEM 2). We notice two outliers at ~16% and ~21%. From the bottom, left visual (appendix ITEM 7), we observe that log of CHE has a weak, positive association with life expectancy; we will explore this further in the model. No significant interaction effects are seen, as all of the confidence intervals for the best fit lines overlap. We will include this interaction in the model for selection, but we expect it to be selected out.

Having looked at some healthcare predictors, we look at GDP per capita, an income predictor, next:



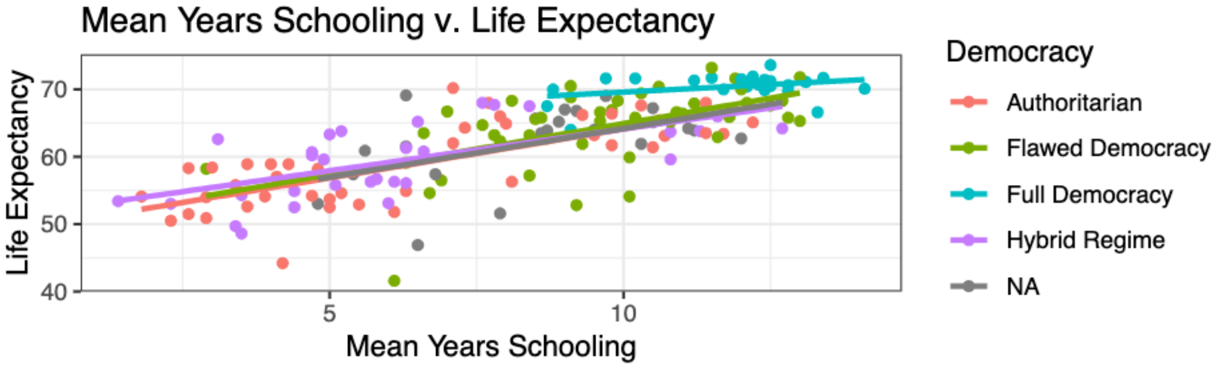
The leftmost plot, referred to as the Preston Curve in economics, demonstrates the expected trend that higher GDP per capita typically corresponds with higher life expectancy. This positive relationship between GDP per capita and life expectancy is more clear after log transforming the GDP predictor, as seen in the rightmost visual; this log transformation is required due to the highly non-linear relationship between the predictor and response. We will use the transformed version in the model.

Next, we look at the Democracy Index:



We observe that countries categorized as full democracies have higher life expectancies than those that are considered less democratic. Full democracies have a median life expectancy of 70.9 years, a value that is 12 years larger than their authoritarian counterparts which only have a median life expectancy of 58.9 years.

Finally, we look at the predictor for our final category education: mean years of schooling. We added the democracy index variable to our visual (allowing us to observe any interaction), as the relationship between education and life expectancy may change within different governmental structures.



It is clear that that for all types of democracies, an increase in mean years of schooling is associated with an increase in life expectancy. We see that full democracies show a different relationship than other democracies, however, with the effect of schooling being reduced. This is likely because a very small handful of countries are full democracies and are likely to be those that are already have high metrics for other predictors of life expectancy (such as income).

Model Selection - add interactions and diagnostics

After assessing the predictor variables thoroughly, we created a multiple linear regression model, discarding any indicator variables that were present and any predictors that had large amounts of missing data, as this could create problems in the selection process (a limitation we tackle later). We also log-transformed some predictors to reduce the effect of some extreme outliers (detailed in appendix).

Table 1: Predict a Country's Life Expectancy (initial model)

Term	Estimate	Std.Error	Statistic	P-Value
(Intercept)	61.846	5.091	12.148	0.000
avg_immune	0.002	0.019	0.097	0.923
InfantMortality	-0.189	0.021	-8.880	0.000
log_CHE	-0.151	0.639	-0.237	0.813
DemocracyFlawed Democracy	-0.761	0.634	-1.200	0.233
DemocracyFull Democracy	1.593	1.205	1.322	0.190
DemocracyHybrid Regime	1.032	0.570	1.812	0.074
log_gdp	1.483	0.376	3.945	0.000
`Mean Years Schooling`	-0.028	0.163	-0.171	0.865
Alcohol	-0.050	0.096	-0.523	0.602
`Mean BMI`	-0.550	0.196	-2.805	0.006
log(hiv_prop)	-1.052	0.177	-5.939	0.000
continentAmericas	2.956	0.851	3.473	0.001
continentAsia	-0.254	0.792	-0.321	0.749
continentEurope	0.990	1.098	0.902	0.370
continentOceania	-0.921	1.313	-0.702	0.485

To get the most precise model, we use backwards selection to achieve a model with the lowest Akaike's

Information Criterion (AIC) attainable. The reduced model is produced below.

In order to choose the most accurate model, we will consider the adjusted R^2 value, AIC, and BIC.

Table 2: Model Selection Criteria for Full Model

Adjusted R Squared	AIC	BIC
0.907	441.501	485.789

Table 3: Model Selection Criteria for Reduced Model

Adjusted R Squared	AIC	BIC
0.911	434.109	467.977

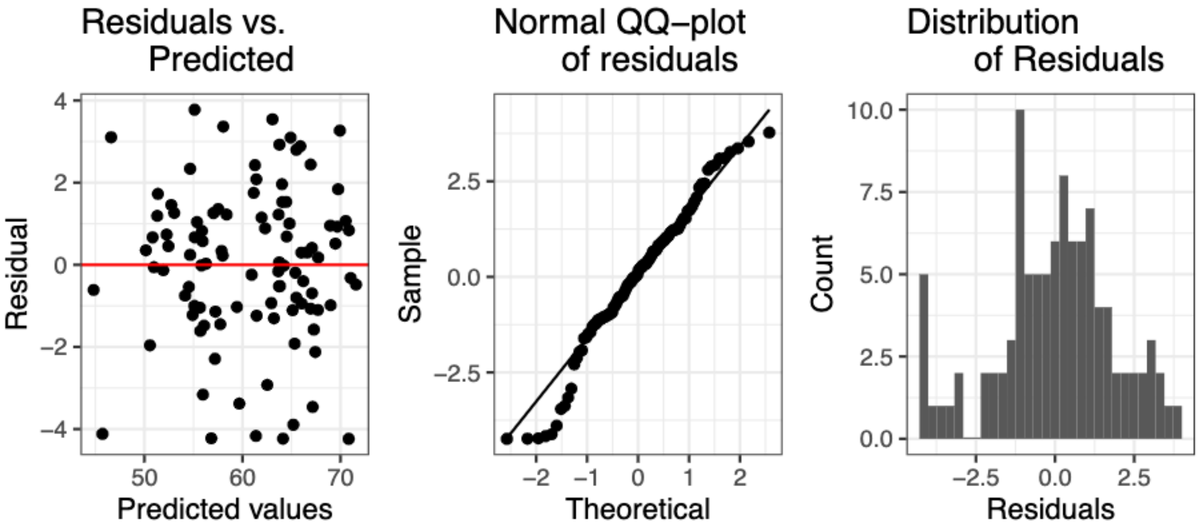
Because of its lower AIC and BIC values, we will move forward with the reduced model, which only includes the infant mortality rate, Democracy categories, log of GDP, mean BMI of a country, log proportion of HIV, and the continents. Additionally, this model has a slightly higher adjusted R^2 value, so it accounts for more variability.

Table 4: Prediction of a Country's Life Expectancy (reduced_cent)

Term	Estimate	Std. Error	Statistic	P-Value
(Intercept)	60.122	0.512	117.378	0.000
InfantMortality_cent	-0.188	0.016	-11.666	0.000
DemocracyFlawed Democracy	-0.785	0.616	-1.274	0.206
DemocracyFull Democracy	1.522	1.060	1.436	0.154
DemocracyHybrid Regime	0.970	0.545	1.779	0.079
log_gdp_cent	1.422	0.311	4.567	0.000
mean_bmi_cent	-0.568	0.167	-3.400	0.001
log_hiv_prop_cent	-1.090	0.163	-6.677	0.000
continentAmericas	2.881	0.767	3.757	0.000
continentAsia	-0.297	0.709	-0.419	0.676
continentEurope	0.618	0.889	0.695	0.489
continentOceania	-0.966	1.230	-0.785	0.434

Thus, we will move forward with the reduced model using mean centered numerical variables:

$$\hat{LifeExpectancy} = 60.122 - 0.188 \times infant\_mortality\_cent - 0.785 \times Democracy\ FlawedDemocracy + 1.522 \times Democracy\ FullDemocracy + 0.970 \times Democracy\ HybridRegime + 1.422 \times log\_gdp\_cent - 0.568 \times mean\_bmi\_cent - 1.090 \times log\_hiv\_prop\_cent + 2.881 \times continentAmericas - 0.297 \times continentAsia + 0.618 \times continentEurope - 0.966 \times continentOceania$$



The points do not appear to follow a particular pattern, so the condition of linearity is met by the data. Additionally, there doesn't appear to be a clear vertical pattern indicating that constant variance is not met. Normality does not appear to be met by the data. The points slightly diverge from the line on the normal QQ-plot and the histogram indicates that the data is left-skewed. However, normality is robust for linear regression, so we can continue with our analysis. Finally, independence is satisfied because it can be reasonably assumed that each variable used in the model does not have a direct impact on the others. While, spatial correlations are possible due to regional similarities between countries, we included continent as a predictor variable, thus resolving this issue. Therefore, the independence condition is satisfied.

When looking at the reduced model, it is apparent that the Democracy categories have larger p-values than the rest of the predictors. Therefore, we want to determine whether or not the categories for Democracy have a statistically significant impact on the life expectancy.

$H_0 : \beta_{FlawedDemocracy} = \beta_{FullDemocracy} = \beta_{HybridRegime} = \beta_{Authoritarian} = 0$

H_a : At least one $\beta_i \neq 0$, where i = Democracy categories (Flawed Democracy, Full Democracy, Hybrid Regime, Authoritarian)

Table 5: Compare Model With/Without Democracy

Res.DF	RSS	DF	Sum.sq	P-Value
91	407.729	NA	NA	NA
88	346.675	3	61.054	0.001

The p-value is less than an alpha level of 0.05, thus we have sufficient evidence to reject the null hypothesis. We can conclude that at least one level of Democracy has a statistically significant impact on life expectancy, so we will leave it in the model.

Additionally, because the exploratory data analysis revealed possible differences in the relationship between the log of Current Health Expenditure (CHE) and life expectancy for different types of democracy, we will explore whether there exists an interaction effect between the two predictors, by creating a new model and testing an interaction term between the log CHE and Democracy in our model. (Reference appendix ITEM 5 for the model we created with this new interaction term)

$H_0 : \beta_{FlawedDemocracy:log\_CHE} = \beta_{FullDemocracy:log\_CHE} = \beta_{HybridRegime:log\_CHE} = 0$

H_a : At least one$\beta_{Democracy:log_CHE} \neq 0$

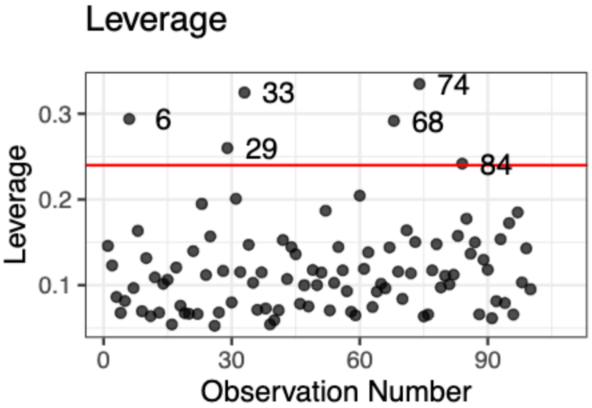
Table 6: Test log(CHE) x Democracy Interaction Term

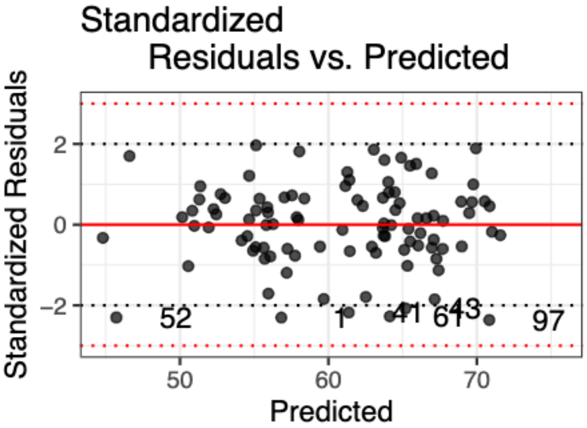
Res.DF	RSS	DF	SumSq	P-Value
88	346.675	NA	NA	NA
84	339.378	4	7.297	0.771

Our p-value is greater than 0.05, thus we fail to reject the null hypothesis. There is not sufficient evidence to assume that any of the coefficients for the interaction terms between Democracy and the log CHE are different than 0. We will continue with the original reduced model.

From the VIF chart (found in the appendix, ITEM 3), we observe no values greater than 10. Thus, we can assume that multicollinearity is not an issue for any of the variables used in the reduced model.

Model diagnostics





There are exactly *six* observations with high leverage: Australia, Fiji, Egypt, New Zealand, Papa New Guinea, Hungary, and Singapore (found with by observation number). We notice that most of these countries belong to the Oceania region; hence, their high leverage could either be because of some unique set of conditions for that region, or because the region itself has a very small number of countries. For Standardized Residuals we find numerous moderate outliers; on exploration, these are spread across numerous regions and country types, so no particular pattern can be found. There are no serious outliers. However, when looking at Cook's Distance (found in appendix, ITEM 4), none of these high leverage or Standardized Residual points translate into influential points for the model. Thus, we will keep all of them and do not have to make a decision of their inclusion in their model.

Results

Our final model, along with relevant fit statistics, is as follows:

Table 7: Predicting a Country's Life Expectancy (reduced_cent)

Term	Estimate	Std. Error	Statistic	P-Value
(Intercept)	60.122	0.512	117.378	0.000
InfantMortality_cent	-0.188	0.016	-11.666	0.000
DemocracyFlawed Democracy	-0.785	0.616	-1.274	0.206
DemocracyFull Democracy	1.522	1.060	1.436	0.154
DemocracyHybrid Regime	0.970	0.545	1.779	0.079

Term	Estimate	Std. Error	Statistic	P-Value
log_gdp_cent	1.422	0.311	4.567	0.000
mean_bmi_cent	-0.568	0.167	-3.400	0.001
log_hiv_prop_cent	-1.090	0.163	-6.677	0.000
continentAmericas	2.881	0.767	3.757	0.000
continentAsia	-0.297	0.709	-0.419	0.676
continentEurope	0.618	0.889	0.695	0.489
continentOceania	-0.966	1.230	-0.785	0.434

Table 8: Model Fit Stats

Adj. R Squared	AIC	BIC
0.911	434.109	467.977

As we previously mentioned, our model has a lower AIC than the model we started out with. Importantly, however, we also have a high Adjusted R Squared, with 91.1 % of the variability in life expectancy being accounted for after discarding useless predictors. This is essential as our initial goal was the prediction.

According to our analysis, the most significant predictors of life expectancy are infant mortality rate, the mean BMI of the country, the continent,level of democracy, national wealth (defined by GDP per capita), and the proportion of HIV cases within a country's population.

Based on our model, the key changes that a country should address to improve life expectancy are as follows:

- Reduce their infant mortality, as every 10-death reduction can lead to about a 1.9 year increase of life expectancy on average
- Increase their GDP per capita, as doubling their income can lead to a 0.986 increase in life expectancy.
- Work on eradicating HIV. Halving HIV infection rates can lead to a 0.756 increase in life expectancy. This is particularly important in Africa, where high HIV infection rates persist

Other changes to look at include:

- Focusing on keeping BMI at a sustainable and healthy level. Whereas incredibly low BMIs are probably harmful for life expectancy, we see that after adjusting for other factors, an increase in BMI by one unit can cause a marginal reduction in Life expectancy by about 0.57 years.
- Moving to a more democratic system. This may not be entirely feasible in many countries, but is a factor to keep in mind

Additionally, our findings contradicted the results of IIASA's study. Based on our final model, we concluded that the amount of education did not have a significant effect on life expectancy, but income (measure in GDP per capita) did.

Discussion + Conclusions

Our multiple linear regression model has shown us that after accounting for other factors, health and income are the two most important predictors of Life Expectancy. Every predictor in the model was some predictor related to these two areas, aside from region (which was included more to remove spatial correlations than as a predictor) and the Democracy Index, many levels of which had low p-values. These findings somewhat contradicted our hypothesis which also presumed education was important; however, it is possible that had we had more than a single predictor for education (i.e mean years of schooling) - like we did for health and income - we would have found more significant results. Further, we saw that the Americas had the highest life expectancy, likely because of the inclusion of USA and Canada, highly developed countries.

The main limitations of this model arose from having to reconstruct the dataset on our own, as this resulted in a large amount of missing data. This was most prevalent in the HIV predictor, so countries that had poor reporting of HIV or had moral reasons against its collection (e.g for Muslim countries) were systematically removed. Hence, our model likely shows some over-representation for countries that do not have this problem, or for any other predictors that were selected out. We briefly tried data imputation methods like predicting mean-matching imputation, but were unsure of how to use and interpret these results.

Another contradiction of our model was in its Democracy Index findings. As one would predict, countries that were considered Full Democracies had higher life expectancies than those classified as Authoritarian. However, flawed democracies had lower life expectancies than their authoritarian counterparts by 0.785 years. This was likely because Authoritarian countries had more missing values in various predictors, and hence the final number of observations that were Authoritarian was quite small. Some of these countries that remained in the data for use in our model such as China and Cuba were outliers with high life expectancies of 68 years, which consequently pulled up the mean expectancy for the authoritarian category.

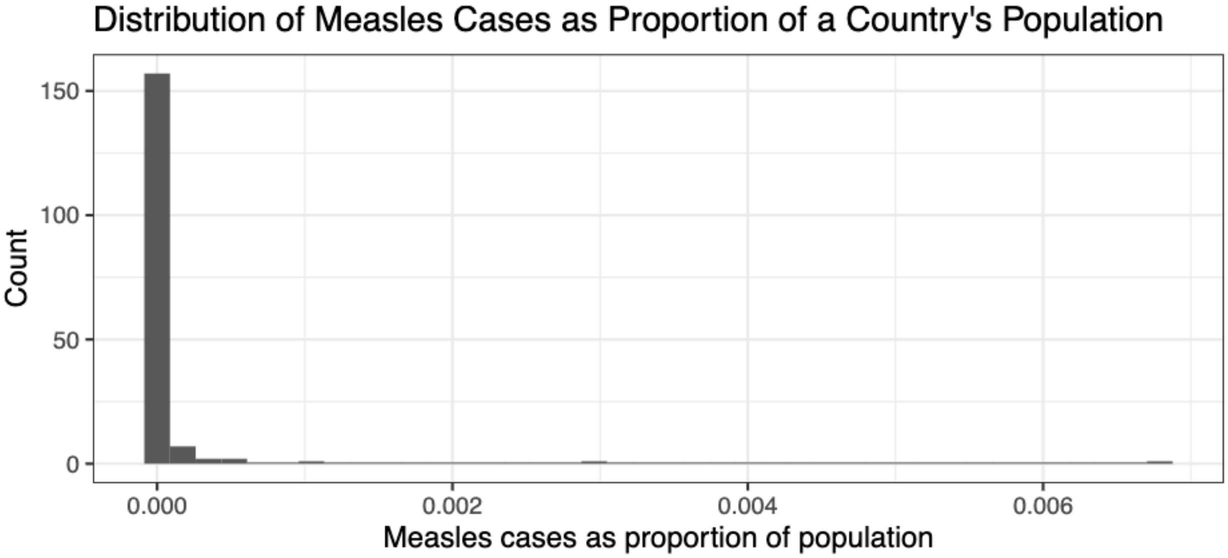
Appendix

ITEM 0:

Table 9: Life Expectancy Statistics

Min	Max	Median
41.6	73.6	63.9

ITEM 1:



ITEM 2:

Table 10: Log of CHE Statistics

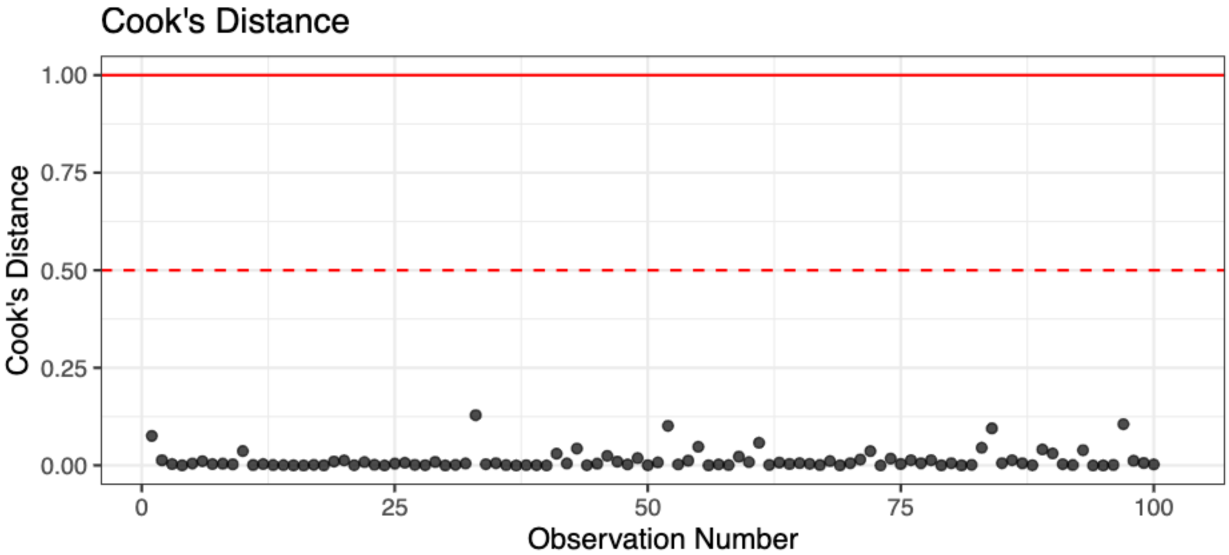
Median	IQR
1.823	0.556

ITEM 3:

Table 11: VIF Table

Names	x
InfantMortality_cent	3.446
DemocracyFlawed Democracy	2.096
DemocracyFull Democracy	2.567
DemocracyHybrid Regime	1.487
log_gdp_cent	4.178
mean_bmi_cent	2.775
log_hiv_prop_cent	1.747
continentAmericas	2.298
continentAsia	2.189
continentEurope	2.417
continentOceania	1.476

ITEM 4:

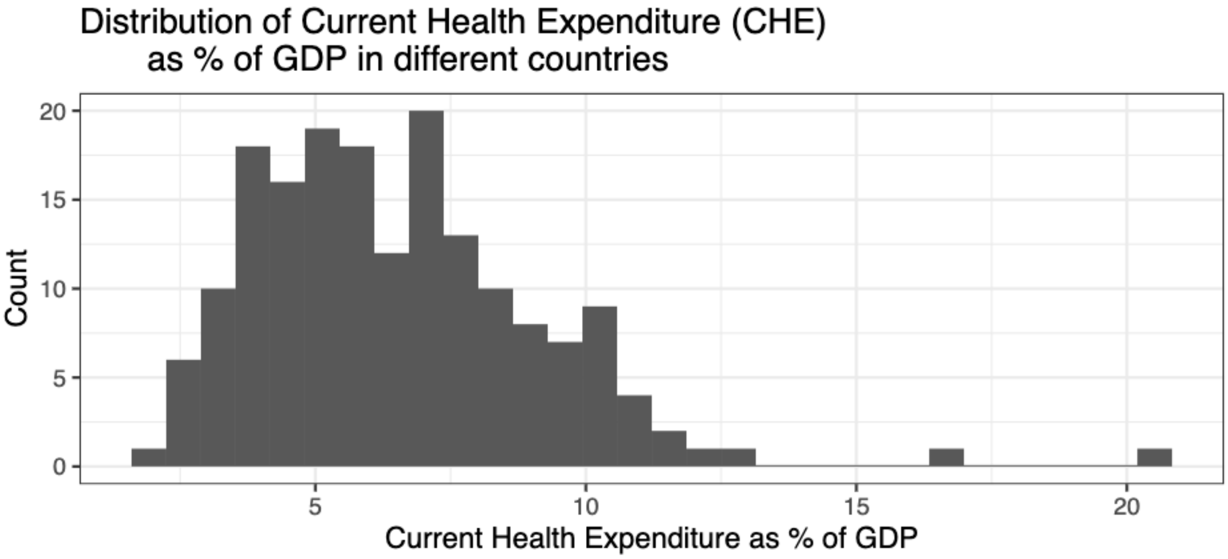


ITEM 5:

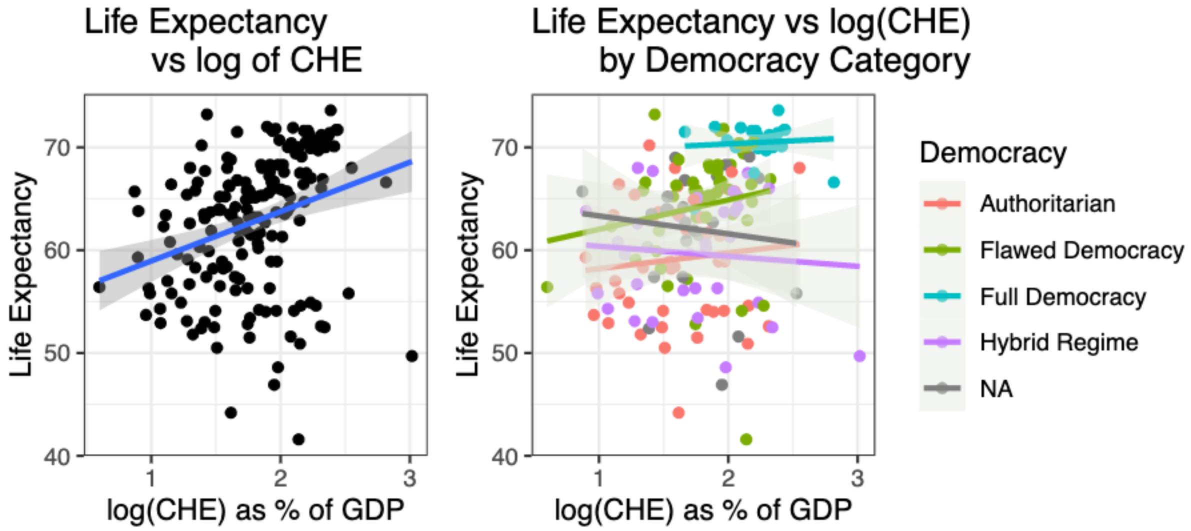
Table 12: Prediciton of a Country's Life Expectancy (interaction)

Term	Estimate	Std.Error	Statistic	P-Value
(Intercept)	62.151	5.113	12.155	0.000
InfantMortality	-0.186	0.017	-11.196	0.000
log_gdp	1.461	0.349	4.184	0.000
DemocracyFlawed Democracy	-0.413	2.744	-0.151	0.881
DemocracyFull Democracy	7.183	5.613	1.280	0.204
DemocracyHybrid Regime	-0.187	2.385	-0.078	0.938
log_hiv_prop	-1.096	0.168	-6.517	0.000
`Mean BMI`	-0.570	0.177	-3.227	0.002
continentAmericas	3.071	0.821	3.740	0.000
continentAsia	-0.287	0.734	-0.392	0.696
continentEurope	0.753	0.918	0.821	0.414
continentOceania	-0.965	1.346	-0.717	0.475
log_CHE	-0.255	1.051	-0.243	0.809
DemocracyFlawed Democracy:log_CHE	-0.229	1.594	-0.144	0.886
DemocracyFull Democracy:log_CHE	-2.533	2.649	-0.956	0.342
DemocracyHybrid Regime:log_CHE	0.670	1.347	0.497	0.620

ITEM 6:



ITEM 7:



Lutz, Wolfgang. 2018. “Education, Not Income, the Best Predictor of a Long Life.” *ScienceDaily*. ScienceDaily. https://www.sciencedaily.com/releases/2018/04/180416103428.htm#:~:text=Rising%20income%20and%20the%20subsequent,better%20predictor%20of%20life%20expectancy.

Wamsley, Laurel. 2021. “American Life Expectancy Dropped by a Full Year in 1st Half of 2020.” *NPR*. NPR. http://www.npr.org/2021/02/18/968791431/american-life-expectancy-dropped-by-a-full-year-in-the-first-half-of-2020.