Your project title

Last!: Parker Dingman, Ethan Donecoff, Karam Oubari, Pei Yi Zhuo

2021-04-22

Your written report goes here! Before you submit, make sure your code chunks are turned off with `echo = FALSE` and there are no warnings or messages with `warning = FALSE` and `message = FALSE`

Introduction and Data

By telling us the average age of death in a population, life expectancy is a key metric for understanding a country's health. According to Max Roser, Esteban Ortiz-Ospina and Hannah Ritchie from Our World in Data, "Broader than the narrow metric of the infant and child mortality, which focus solely at mortality at a young age, life expectancy captures the mortality along the entire life course."[1] Over the course of history, life expectancy has risen dramatically. It is estimated that in pre-modern times, life expectancy worldwide was only about 30 years. Since the Industrial Revolution in the 18th and 19th centuries, many countries had huge gains in this number. And since the beginning of the 20th century, global average life expectancy has risen to about 70 years. However, there remain huge inequalities in this number. Currently (as of 2019), the Central African Republic has the lowest life expectancy of 53 years while Japan has the highest with 83.

In addition to the more obvious health-related connections to life expectancy, numerous pieces of academic literature have delved into the non-medical factors behind life expectancy. A major example is a longitudinal study conducted by Charles Lin, Eugene Rogot, Norman Johnson, Paul Sorlie, and Elizabeth Arias, which examined life expectancy by socioeconomic factors.[2] Academic literature such as this provides us with motivations to examine this topic on an international level, looking at various health-related and non-health-related factors that connect to life expectancy.

In terms of initial hypotheses of model selection, we expect that the strongest predictors of life expectancy will be `Adult Mortality`, `infant deaths`, and `GDP`. We also predict that countries that have `status` equal to "Developed" will have higher life expectancy than those that have `status` equal to "Developing".

Our primary data set is comprised of information that had been gleaned from the websites of the World Health Organization and the United Nations.[3] Each entry describes the health, social, and economic conditions for one of 193 countries in a given year from 2000 to 2015.[3] Because this data set lacks a variable that specifies the region of each country, we joined it with another data set that pairs country and region.[4] This secondary data set compiles information that can be found on the CIA World Factbook.[4] Both data sets came from Kaggle.

Definitions of Key Predictor Variables:

Region: Area, subcontinent, or continent where a country is located

Adult Mortality: Probability of dying between 15 and 60 years per 1000 population

Infant Deaths: Number of infant deaths per 1000 population

Hepatitis B: Percentage of hepatitis B immunization coverage among 1-year-olds

Measles: Number of reported measles cases per 1000 population

Under-Five Deaths: Number of under-five deaths per 1000 population

Total Expenditure: General government expenditure on health as a percentage of total government expenditure

Diphtheria: Percentage of DTP3 immunization coverage among 1-year-olds

HIV/AIDS: Deaths per 1000 live births HIV/AIDS (0-4 years)

Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

Schooling: Average years of schooling/education

Exploratory Data Analysis

First, we will look at some summary statistics of our response variable, `Life expectancy`. Though it is possible to analyze the entire data set over all years, this creates difficulties with creating models. As one might expect, the average global life expectancy increased from 2000 to 2015. This relationship might make it unclear whether a rise in life expectancy is explained by our predictor variables or if it is simply due to human development over time. As a result, we will only use data from one year to perform our analysis.

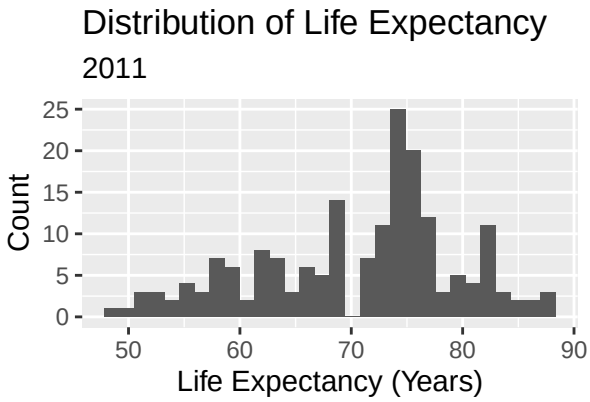
We will only analyze life expectancy for the year 2011, because it is the latest year among those years in which no variable has a missing rate greater than 22%.

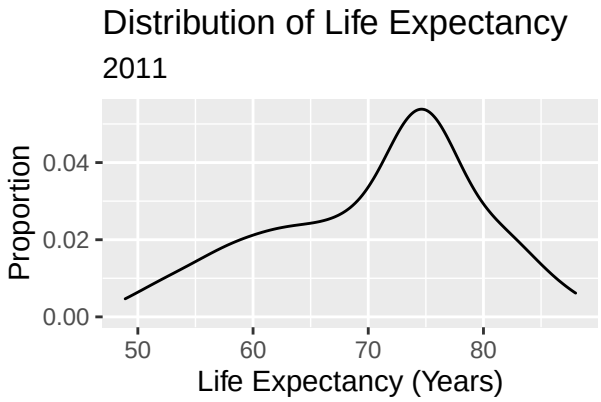
Table 1: Summary Statistics of Response Variable, Life Expectancy

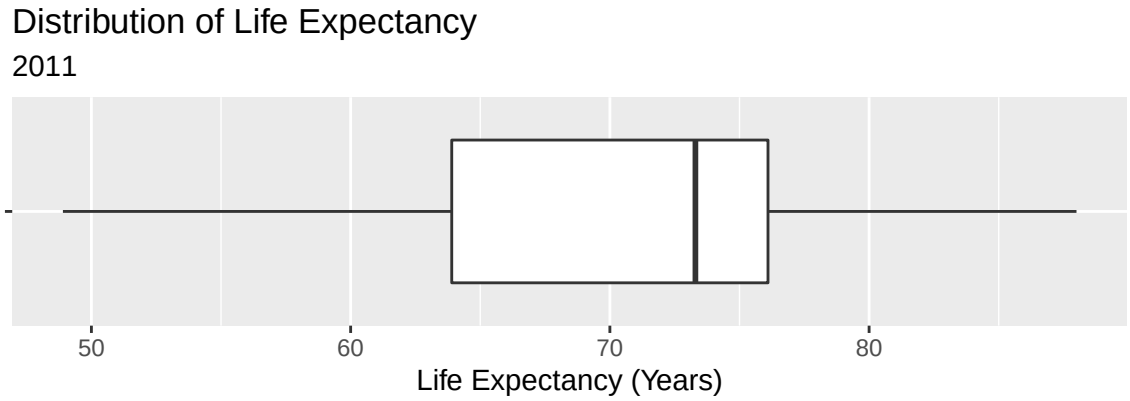
min	max	range	mean	median	Q1	Q3	iqr	sd
48.9	88	39.1	70.654	73.3	63.9	76.1	12.2	8.925

These summary statistics give a rough idea of the distribution of the response variable. The median life expectancy (~73.3 yrs) is almost 3 years larger than the mean (~70.7 yrs). Additionally, the median is closer to the third quartile than the first quartile. This suggests that life expectancy may be left-skewed, which we will evaluate further with visualizations.

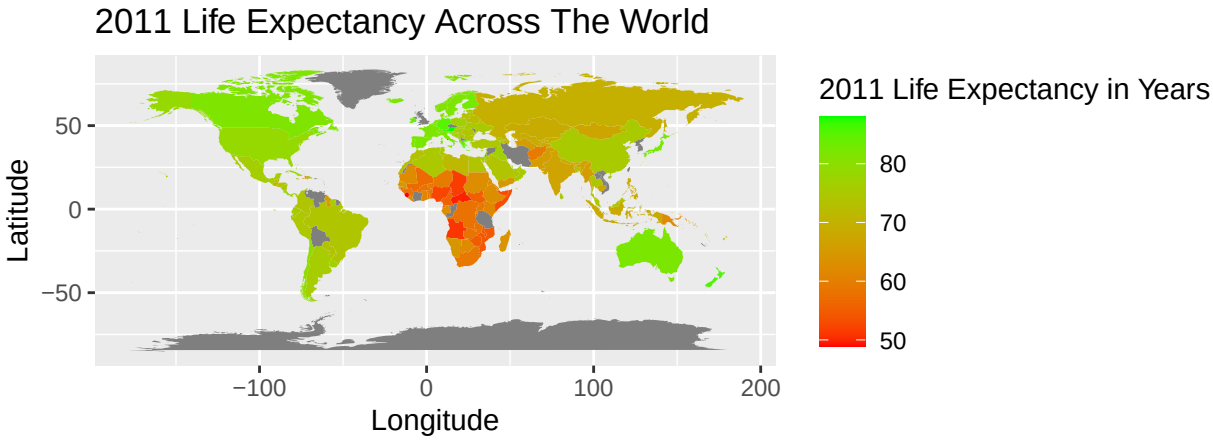
Now, we will look at some summary visualizations of life expectancy.



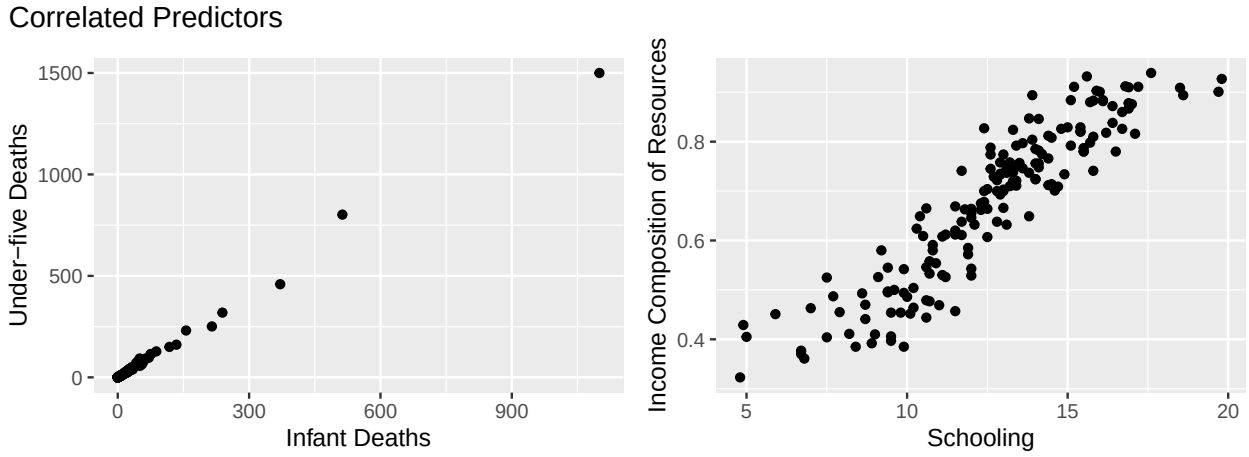




Here we see that life expectancy is left skewed with a center just over 70 years. Life expectancy ranges from about 50 years to 90 years.



Here we can see the differences in life expectancy across the world. For example, it is clear that life expectancy is much higher in North and South American countries when compared to African countries.



Methodology

Model Selection

In this project we plan on using various regression analysis methods including, but not limited to, multiple linear regression, statistical inference, analysis of variance, and model selection in an attempt to understand country and region-level life expectancy, as well as the health, social, and economic relationships behind this number.

The response variable we will be using in this project is `Life expectancy`. This is a measure of the average age of death in a year for the given country during that year. In terms of regression model technique, we will be using multiple linear regression because our response variable is quantitative.

The goal of our analysis is to calculate the most precise prediction of the response variable (life expectancy). To do this we will perform forward and backward selection using AIC and BIC. Forward selection entails adding variables recursively using AIC or BIC as the criteria while backward selection means dropping variables one at a time that are deemed irrelevant based on AIC or BIC.

Table 2: Potential Models

Selection Method	AIC	BIC	Adjusted R-squared
Backward (AIC)	636.282	699.368	0.918
Backward (BIC)	654.341	674.413	0.895
Forward (AIC)	636.413	688.029	0.915
Forward (BIC)	654.341	674.413	0.895

Based on AIC and adjusted R^2, we prefer the model we found through backward selection using AIC. This model results in the highest adjusted R^2 which means that the model's predictor variables explain the highest proportion of the variation in the response (`Life expectancy`) out of all four models. Moreover, this model is the only model of the four that is superior to the other models in more than one metric.

Multicollinearity

Table 3: Predictor Variable VIFs

Variable	VIF
RegionBALTICS	1.408
RegionC.W. OF IND. STATES	1.899
RegionEASTERN EUROPE	1.920

Variable	VIF
RegionLATIN AMER. & CARIB	2.742
RegionNEAR EAST	1.721
RegionNORTHERN AFRICA	1.260
RegionNORTHERN AMERICA	1.441
RegionOCEANIA	1.630
RegionSUB-SAHARAN AFRICA	4.657
RegionWESTERN EUROPE	3.002
`Adult Mortality`	2.254
`infant deaths`	244.366
`Hepatitis B`	2.972
Measles	3.156
`under-five deaths`	244.039
`Total expenditure`	1.328
Diphtheria	2.777
`HIV/AIDS`	2.016
`Income composition of resources`	13.166
Schooling	8.937

Variables with a VIF > 10 will have issues with multicollinearity. `infant deaths` and `under-five deaths` are clearly highly correlated (this makes a lot of sense in the context of the data). `Income composition of resources` appears to be correlated with `Schooling`. These two relationships were visualized in the EDA.

We should try models without `infant deaths` or without `under-five deaths` and then use model comparison techniques to decide on which of these two variables should be removed. Likewise, we ought to compare models with only `Income composition of resources` or `Schooling` and decide which variable to keep.

Table 4: Infant Deaths vs. Under-five Deaths

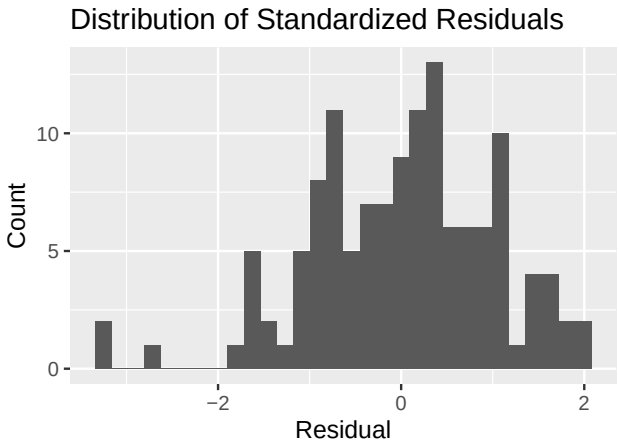
Included Variable	AIC	BIC	adj.r.squared
Infant Deaths	637.581	697.799	0.916
Under-five Deaths	637.343	697.561	0.916

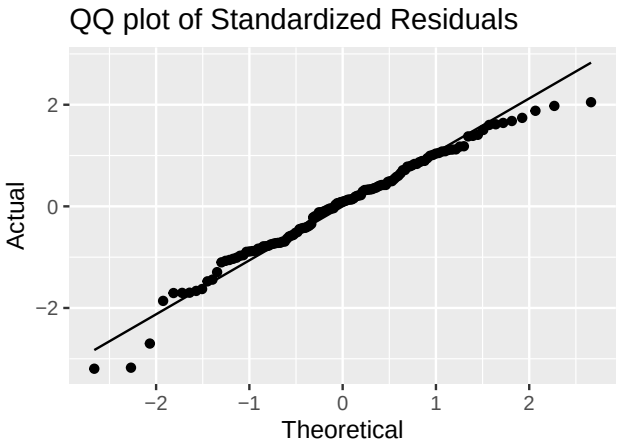
Table 5: Schooling vs. Income Composition of Resources

Included Variable	AIC	BIC	adj.r.squared
Schooling	695.555	752.905	0.868
Income Composition of Resources	642.335	699.686	0.913

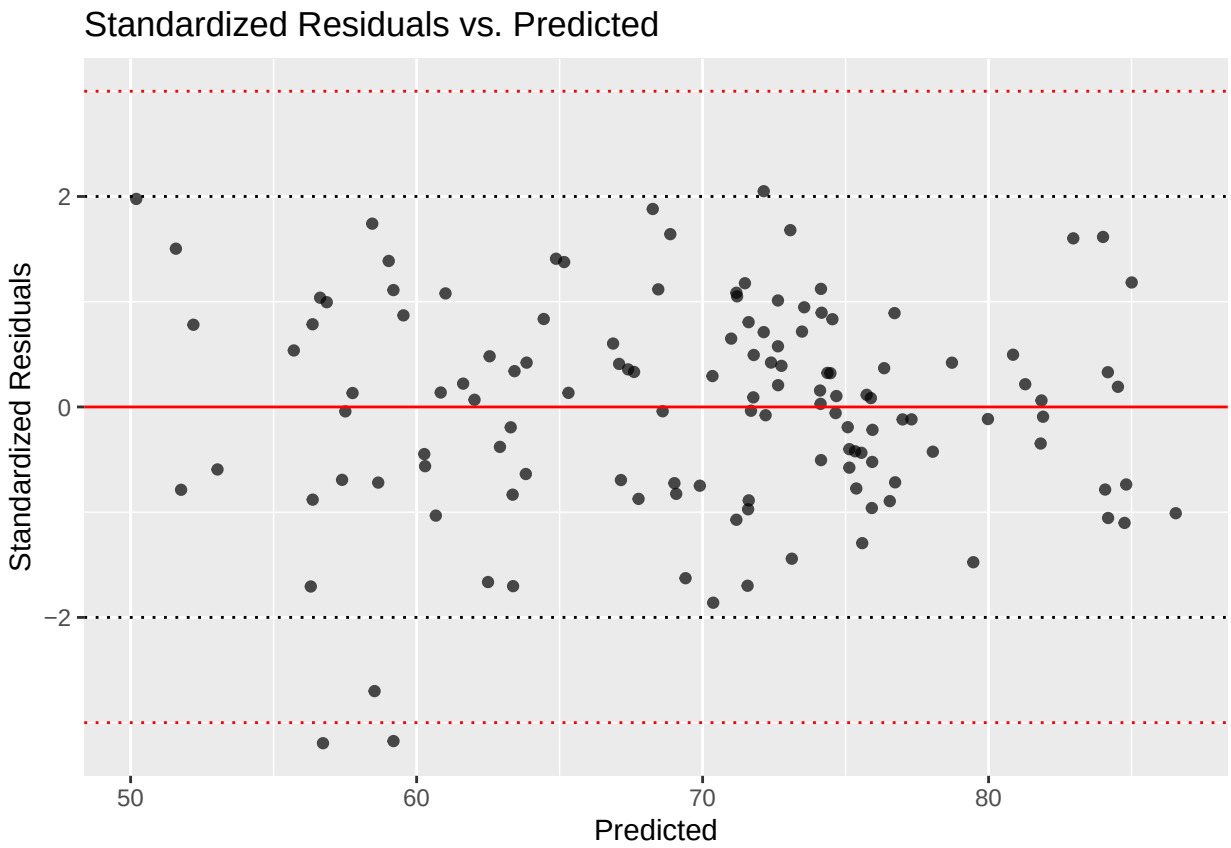
When comparing two models that are identical except that one includes `under-five deaths` and the other includes `infant deaths`, the one that includes `infant deaths` has lower values of AIC and BIC as well as a higher value of adjusted R^2. Likewise, given two models that have the same predictors except that one includes `Schooling` while the other includes `Income composition of resources`, the one that includes `Income composition of resources` is superior in terms of AIC, BIC, and adjusted R^2.

Model Conditions and Diagnostics





Judging from the histogram and QQ plot above, the standardized residuals appear to be normally distributed, thus fulfilling the normality condition.



The plot above depicts standardized residuals on the y-axis and predicted values on the x-axis. We see no major departure from the constant variance assumption. However, there are a few moderate outliers and two severe outliers. Will look into these later.

```
## # A tibble: 12 x 2
##    obs_num  .hat
##      <int> <dbl>
##  1       3 0.345
##  2      25 1
```

```
##  3      40 0.335
##  4      54 0.735
##  5      55 0.379
##  6      65 0.339
##  7      69 0.345
##  8      82 0.343
##  9      84 0.334
## 10      87 0.390
## 11     121 0.337
## 12     125 0.358
```

```
## # A tibble: 0 x 2
## # ... with 2 variables: obs_num <int>, .cooksd <dbl>
```

The points in the first table are points with high leverage (.hat > leverage_threshold = 0.3230769) while the points in the second table are influential points (.cooksd > 0.5). 15 observations have high leverage and 0 observations are influential.

Results

Table 6: Selected Model

term	estimate	std.error	statistic	p.value
(Intercept)	52.013	2.401	21.668	0.000
RegionBALTICS	-3.263	1.807	-1.805	0.074
RegionC.W. OF IND. STATES	-2.618	1.171	-2.237	0.027
RegionEASTERN EUROPE	-0.322	1.298	-0.249	0.804
RegionLATIN AMER. & CARIB	1.054	1.006	1.048	0.297
RegionNEAR EAST	1.110	1.392	0.798	0.427
RegionNORTHERN AFRICA	0.636	1.723	0.369	0.713
RegionNORTHERN AMERICA	4.466	3.187	1.401	0.164
RegionOCEANIA	-0.759	1.199	-0.633	0.528
RegionSUB-SAHARAN AFRICA	-2.576	1.030	-2.500	0.014
RegionWESTERN EUROPE	4.905	1.296	3.784	0.000
`Adult Mortality`	-0.013	0.003	-4.335	0.000
`infant deaths`	0.003	0.004	0.828	0.409
`Hepatitis B`	0.034	0.018	1.881	0.063
Measles	0.000	0.000	-1.679	0.096
`Total expenditure`	0.125	0.107	1.176	0.242
Diphtheria	-0.035	0.023	-1.517	0.132
`HIV/AIDS`	-0.620	0.118	-5.271	0.000
`Income composition of resources`	31.747	2.930	10.834	0.000

93.03% of the variation in the `Life expectancy` is explained by the regression model above, which contains `Region`, `Adult Mortality`, `infant deaths`, `Hepatitis B`, `Measles`, `under-five deaths`, `Total expenditure`, `Diphtheria`, `HIV/AIDS`, `Income composition of resources` and `Schooling`.

Appendix

Table 7: Full Model With All Possible Predictor Variables

term	estimate	std.error	statistic	p.value
(Intercept)	49.126	3.094	15.876	0.000
RegionBALTICS	-1.206	2.052	-0.588	0.558
RegionC.W. OF IND. STATES	-1.763	1.473	-1.197	0.234
RegionEASTERN EUROPE	0.980	1.700	0.577	0.565
RegionLATIN AMER. & CARIB	1.931	1.336	1.445	0.151
RegionNEAR EAST	1.333	1.678	0.795	0.429
RegionNORTHERN AFRICA	1.399	1.855	0.754	0.452
RegionNORTHERN AMERICA	5.640	3.376	1.671	0.098
RegionOCEANIA	0.458	1.641	0.279	0.781
RegionSUB-SAHARAN AFRICA	-1.361	1.225	-1.111	0.269
RegionWESTERN EUROPE	6.388	1.773	3.603	0.000
StatusDeveloping	1.105	1.243	0.888	0.376
`Adult Mortality`	-0.014	0.003	-4.443	0.000
`infant deaths`	0.091	0.041	2.249	0.027
Alcohol	0.008	0.106	0.079	0.937
`percentage expenditure`	0.001	0.001	1.204	0.232
`Hepatitis B`	0.050	0.019	2.623	0.010
Measles	0.000	0.000	-1.125	0.263
BMI	0.007	0.021	0.324	0.747
`under-five deaths`	-0.064	0.029	-2.184	0.031
Polio	-0.020	0.014	-1.440	0.153
`Total expenditure`	0.141	0.117	1.206	0.231
Diphtheria	-0.036	0.024	-1.520	0.132
`HIV/AIDS`	-0.529	0.121	-4.355	0.000
GDP	0.000	0.000	-1.036	0.303
Population	0.000	0.000	-1.292	0.199
`thinness  1-19 years`	0.199	0.272	0.731	0.466
`thinness 5-9 years`	-0.099	0.262	-0.378	0.706
`Income composition of resources`	43.132	6.029	7.154	0.000
Schooling	-0.548	0.250	-2.192	0.031

Sources

[1] https://ourworldindata.org/life-expectancy

[2] https://europepmc.org/article/med/12785422/reload=0#impact

[3] https://www.kaggle.com/kumarajarshi/life-expectancy-who.

[4] https://www.kaggle.com/fernandol/countries-of-the-world