Your project title

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

2021-04-18

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]

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.

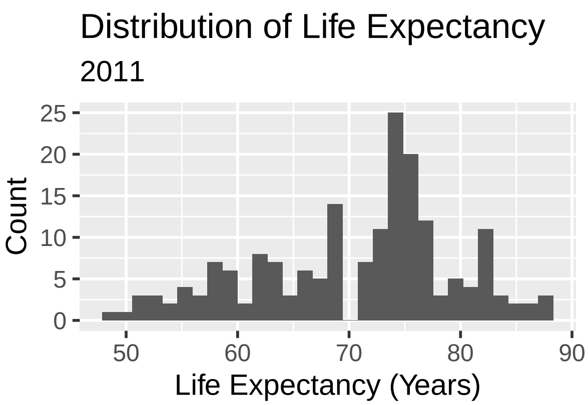
We will only analyze life expectancy for the year 2011, because it is the latest year among those years in which no variable has a missing rate greater than 22%.

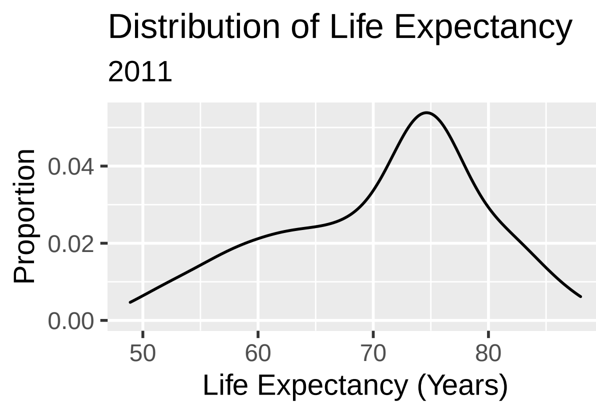
Table 1: Summary Statistics of Response Variable, Life Expectancy

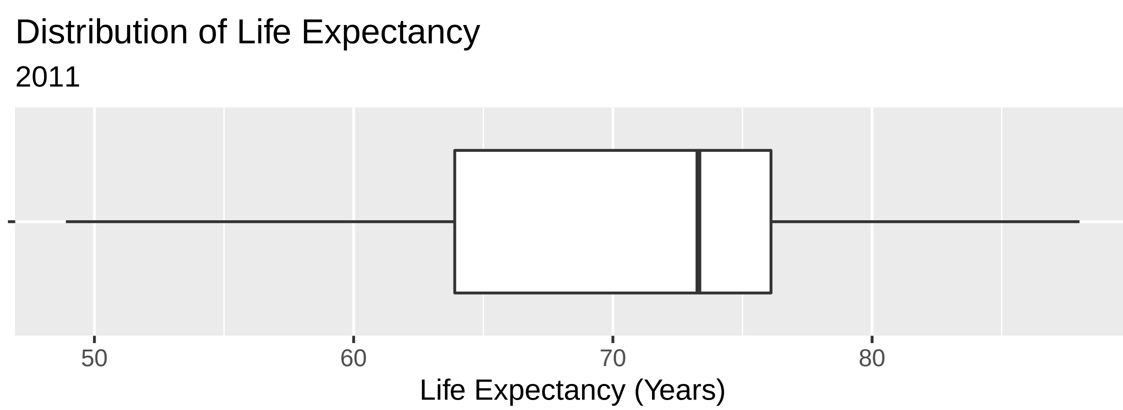
min	max	range	mean	median	Q1	Q3	iqr	sd
48.9	88	39.1	70.654	73.3	63.9	76.1	12.2	8.925

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.

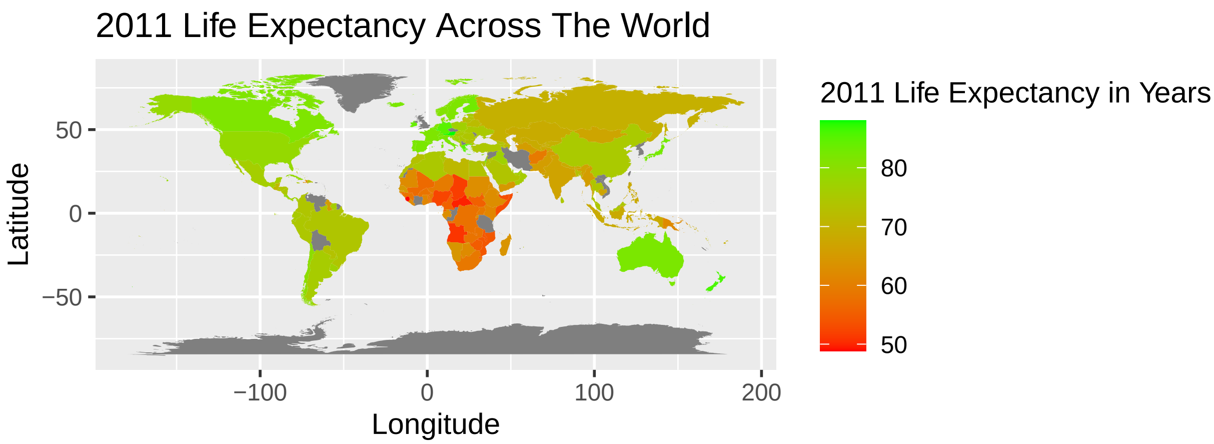
Now, we will look at some summary visualizations of life expectancy.







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.

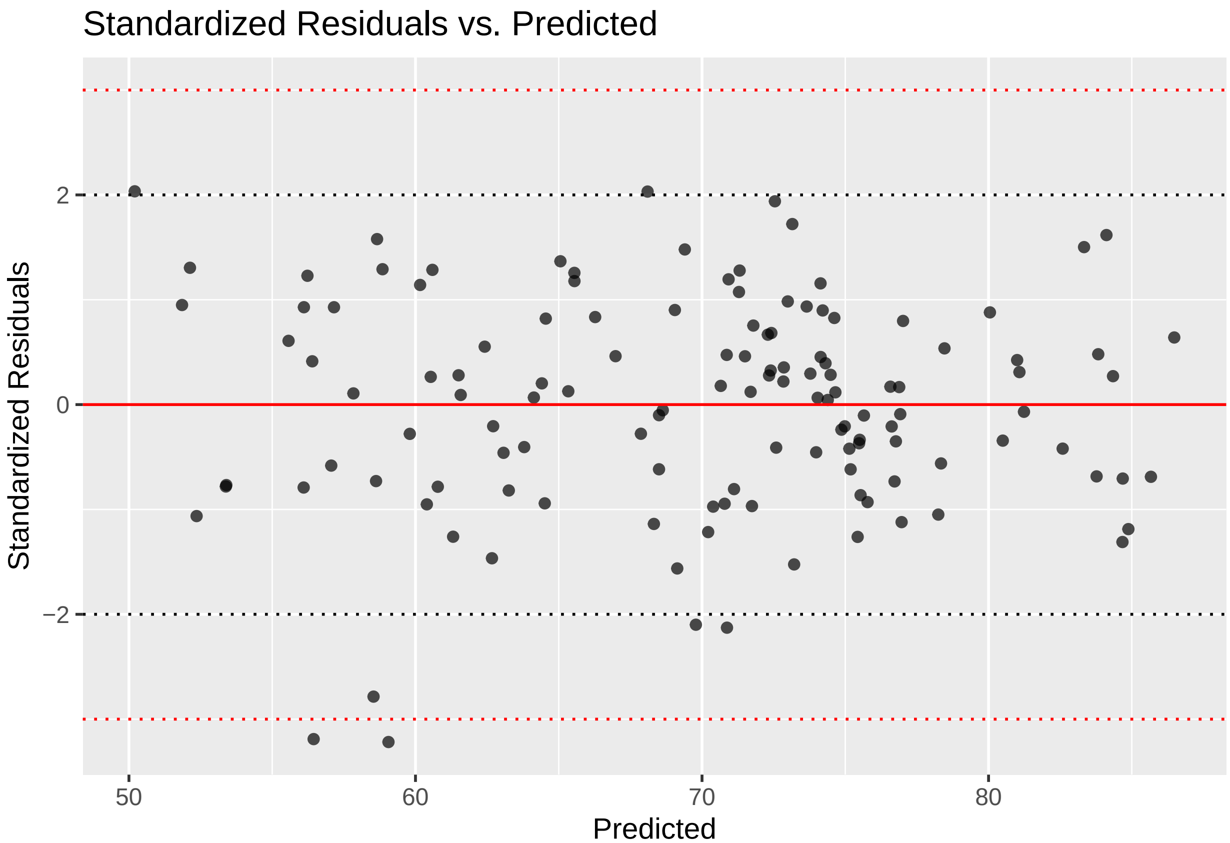
Table 2: Potential models

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.

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

Model Diagnostics



There are a few moderate outliers and two severe outliers. Will look into these later.

```
## [1] 0.3230769
```

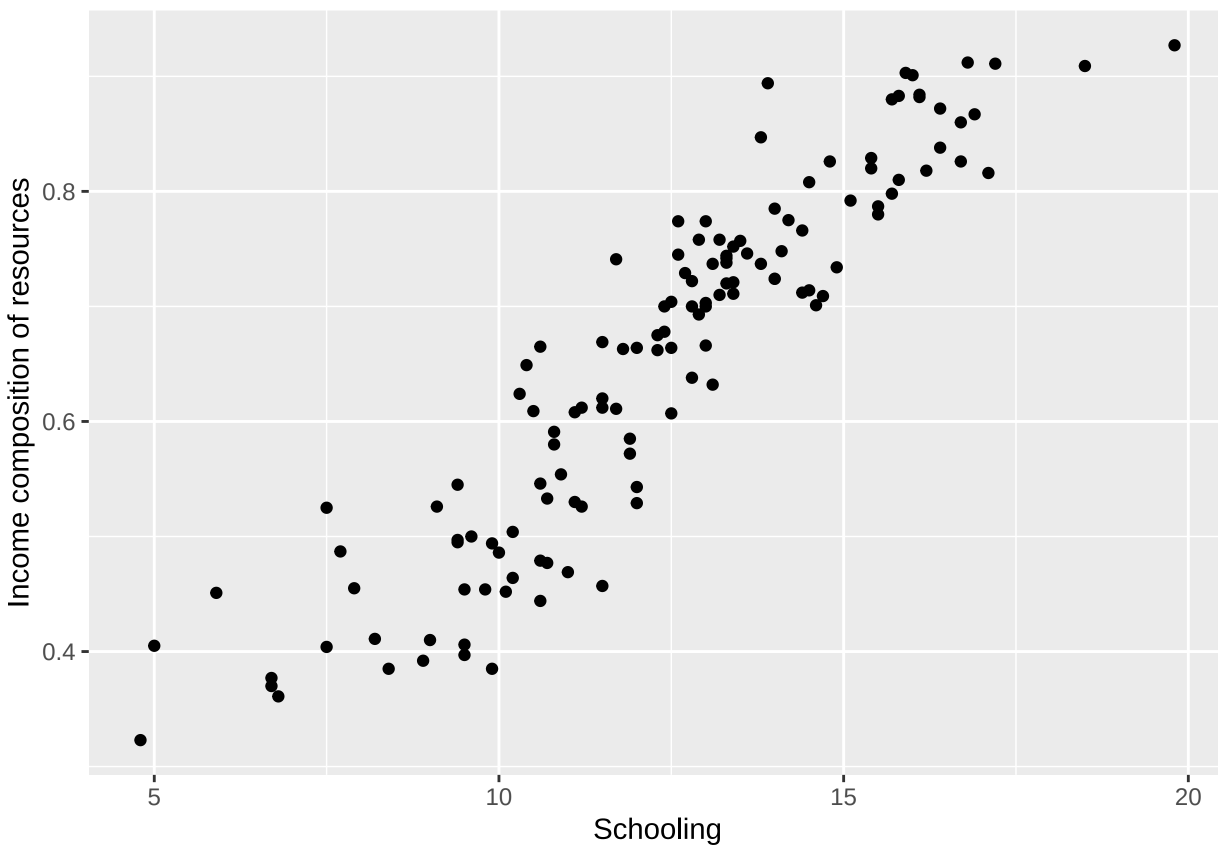
0.3230769 is the leverage threshold.

```
## # A tibble: 15 x 2
##    obs_num  .hat
##      <int> <dbl>
##  1       3 0.345
##  2      25 1
##  3      40 0.335
##  4      43 0.328
##  5      54 0.756
##  6      55 0.410
##  7      65 0.339
##  8      69 0.346
##  9      82 0.348
## 10      84 0.339
## 11      87 0.403
## 12      90 0.711
## 13      91 0.373
## 14     121 0.340
## 15     125 0.375
```

```
## # 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.

names	x
RegionBALTICS	1.408
RegionC.W. OF IND. STATES	1.899
RegionEASTERN EUROPE	1.920
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`.

We should try models without either `infant deaths` and `under-five deaths` and then use model comparison techniques to decide on which of these two variables should be removed.

```
## # A tibble: 1 x 3
##     AIC   BIC adj.r.squared
##   <dbl> <dbl>         <dbl>
## 1  638.  698.         0.916
```

```
## # A tibble: 1 x 3
##     AIC   BIC adj.r.squared
##   <dbl> <dbl>         <dbl>
## 1  637.  698.         0.916
```

```
## # A tibble: 1 x 3
##     AIC   BIC adj.r.squared
##   <dbl> <dbl>         <dbl>
## 1  696.  753.         0.868
```

```
## # A tibble: 1 x 3
##     AIC   BIC adj.r.squared
##   <dbl> <dbl>         <dbl>
## 1  642.  700.         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.

Conditions

Results

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