Your project title

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

2021-04-26

Introduction and Data

Research question, background, and data

In this research project, we hope to most accurately predict life expectancy using economic, social, and health-related country level-data in conjunction with the regression analysis method, multiple linear regression. Therefore our research question is: how can we most accurately predict a country's average life expectancy?

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 strong predictors of life expectancy will be `Adult Mortality`, `infant deaths`, `under-five deaths`, and `Total expenditure`. 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 were found on Kaggle and credit is given to Deeksha Russell, Duan Wang, Fernando Lasso, and the above organizations for contributing to the creation of our datasets.

Definitions of Relevant Variables:

`Life expectancy`: Average life expectancy in years (Response variable)

`Country`: Name of country (Identifier variable)

Predictors 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

`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)

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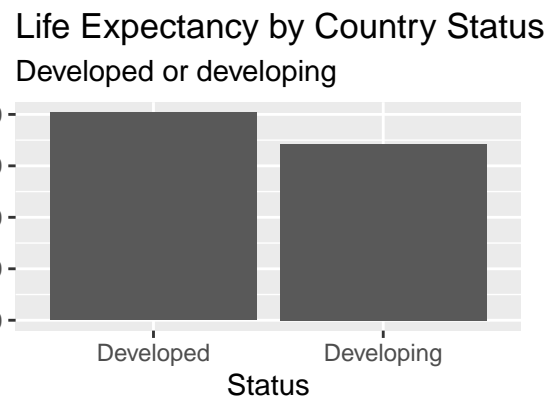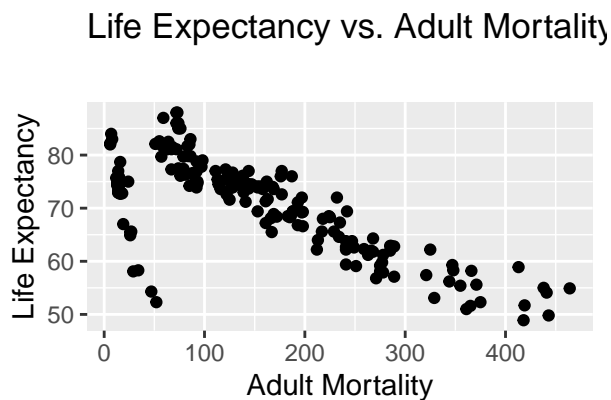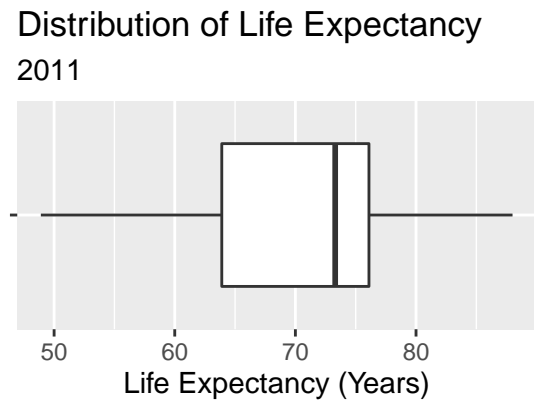
We will only analyze life expectancy for the year 2011, because it is the latest year among those years with the least missing data (no variable has a missing rate greater than 22%).

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, as well as other possibly relevant predictor variables.

Distribution of Life Expectancy
2011

Distribution of Life Expectancy
2011

Life Expectancy vs. Adult Mortality

Life Expectancy by Country Status
Developed or developing

In terms of the resposne variable, 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.

For relationships with predictor variables, we see that a strong negative relationship between life expectancy and adult mortality. We also a see a strange diversion in the points between countries that have lower and higher adult mortaltities. We predict that this is because some of those countries have higher infant mortalities, driving life expectancy down. We also see much higher life expectancies among countries with `status` "developed" than "developing" with a difference of about 12 years.

2011 Life Expectancy Across The World
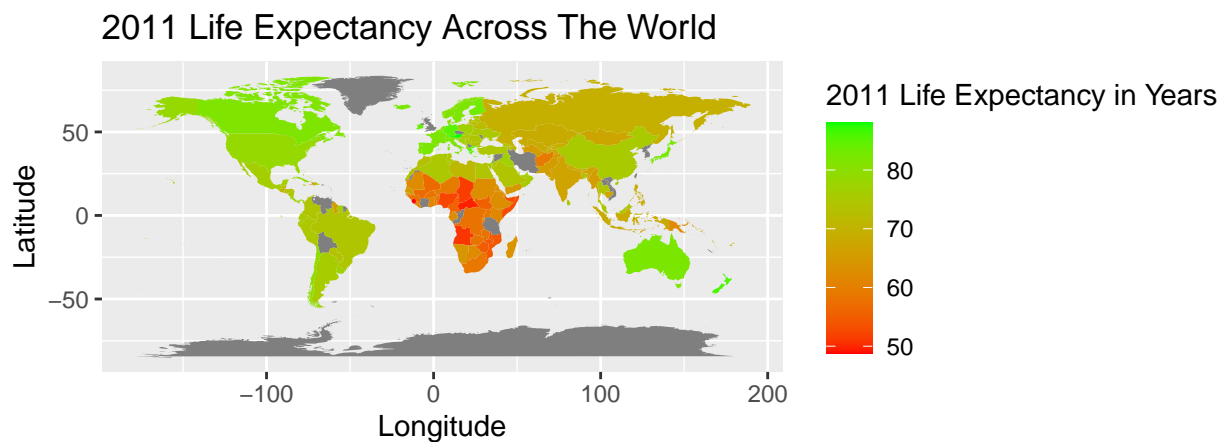
Here we can see the differences in life expectancy around the world. This map allows us to visualize how life expectancy varies between regions. It appears that that the Sub-Saharan African region has the lowest life-expectancy, while regions like North America and Western Europe have the highest life expectancy. This makes sense because certain regions around are more developed than others.

Correlated Predictors
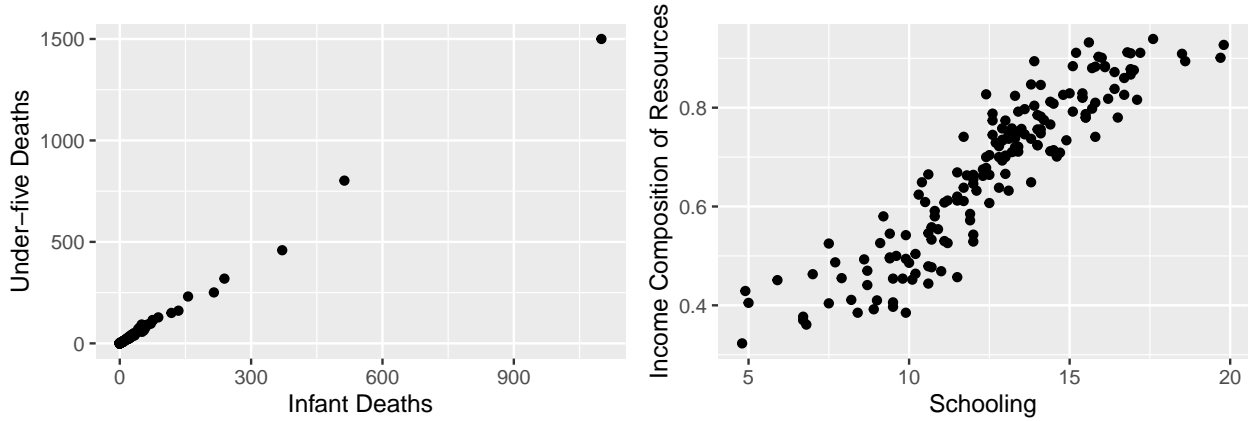
Methodology

In order to answer the research question, we will be using multivariable linear regression techniques due to the quantitative nature of our response variable, `Life expectancy`.

In order to create a model capable of predicting life expectancy, we use model selection to cull the number of potential predictor variables. We then assess whether certain interaction terms should be added to the model before addressing multicollinearity. Finally, with the set of predictor variables finalized, we ascertain whether any variable transformations are necessary through condition checking and identify outliers by examining model diagnostics.

Model Selection

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 criterion while backward selection means dropping variables one at a time, starting with our full model (Appendix Table 9), 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 out of all four models. Moreover, this model is the only model out of the four that is superior to the remaining three models in more than one metric (AIC and R^2).

Potential Interaction Terms

Before proceeding with the model, we will look at some potential interaction terms to see if they should be added to the selected model. We will begin by looking at the interaction between Region and HIV/AIDS. This is because we might suspect that the impact of HIV/AIDS on life expectancy changes by region due to disparate access to relatively new treatments. The interaction between Region and Income composition of resources is also shown. The table below shows the selected model including these interaction terms, but only displays the interaction terms.

Table 3: Selected Model Interaction Terms

term	estimate	std.error	statistic	p.value
RegionBALTICS:`HIV/AIDS`	NA	NA	NA	NA
RegionC.W. OF IND. STATES:`HIV/AIDS`	-8.530	14.324	-0.595	0.553
RegionEASTERN EUROPE:`HIV/AIDS`	NA	NA	NA	NA
RegionLATIN AMER. & CARIB:`HIV/AIDS`	-7.502	6.365	-1.179	0.241
RegionNEAR EAST:`HIV/AIDS`	NA	NA	NA	NA
RegionNORTHERN AFRICA:`HIV/AIDS`	NA	NA	NA	NA
RegionNORTHERN AMERICA:`HIV/AIDS`	NA	NA	NA	NA
RegionOCEANIA:`HIV/AIDS`	-6.167	6.775	-0.910	0.365
RegionSUB-SAHARAN AFRICA:`HIV/AIDS`	-4.655	5.696	-0.817	0.416
RegionWESTERN EUROPE:`HIV/AIDS`	NA	NA	NA	NA
RegionBALTICS:`Income composition of resources`	24.692	134.802	0.183	0.855
RegionC.W. OF IND. STATES:`Income composition of resources`	-7.027	17.522	-0.401	0.689
RegionEASTERN EUROPE:`Income composition of resources`	-32.847	27.206	-1.207	0.230
RegionLATIN AMER. & CARIB:`Income composition of resources`	-9.491	13.462	-0.705	0.483
RegionNEAR EAST:`Income composition of resources`	-15.406	16.364	-0.941	0.349
RegionNORTHERN AFRICA:`Income composition of resources`	-10.469	31.918	-0.328	0.744
RegionNORTHERN AMERICA:`Income composition of resources`	NA	NA	NA	NA
RegionOCEANIA:`Income composition of resources`	-4.197	11.762	-0.357	0.722
RegionSUB-SAHARAN AFRICA:`Income composition of resources`	-1.952	9.103	-0.214	0.831
RegionWESTERN EUROPE:`Income composition of resources`	-36.022	28.032	-1.285	0.202

As shown in the table above, some of the groups do not have enough data to generate model coefficients. For the model coefficients calculated, there are no significant interaction terms for the variables investigated. Thus, there is not enough evidence to say that there is a significant interaction term between Region and HIV/AIDS or Region and Income composition of resources.

Multicollinearity

Table 4: Predictor Variable VIFs

Variable	VIF
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`. 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 keep only one of the two.

Table 5: Infant Deaths vs. Under-five Deaths

| Included Variable	AIC	BIC	adj.r.squared
Infant Deaths	637.3430	697.5612	0.9164
Under-five Deaths	637.5806	697.7988	0.9162

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.

Table 6: 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

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

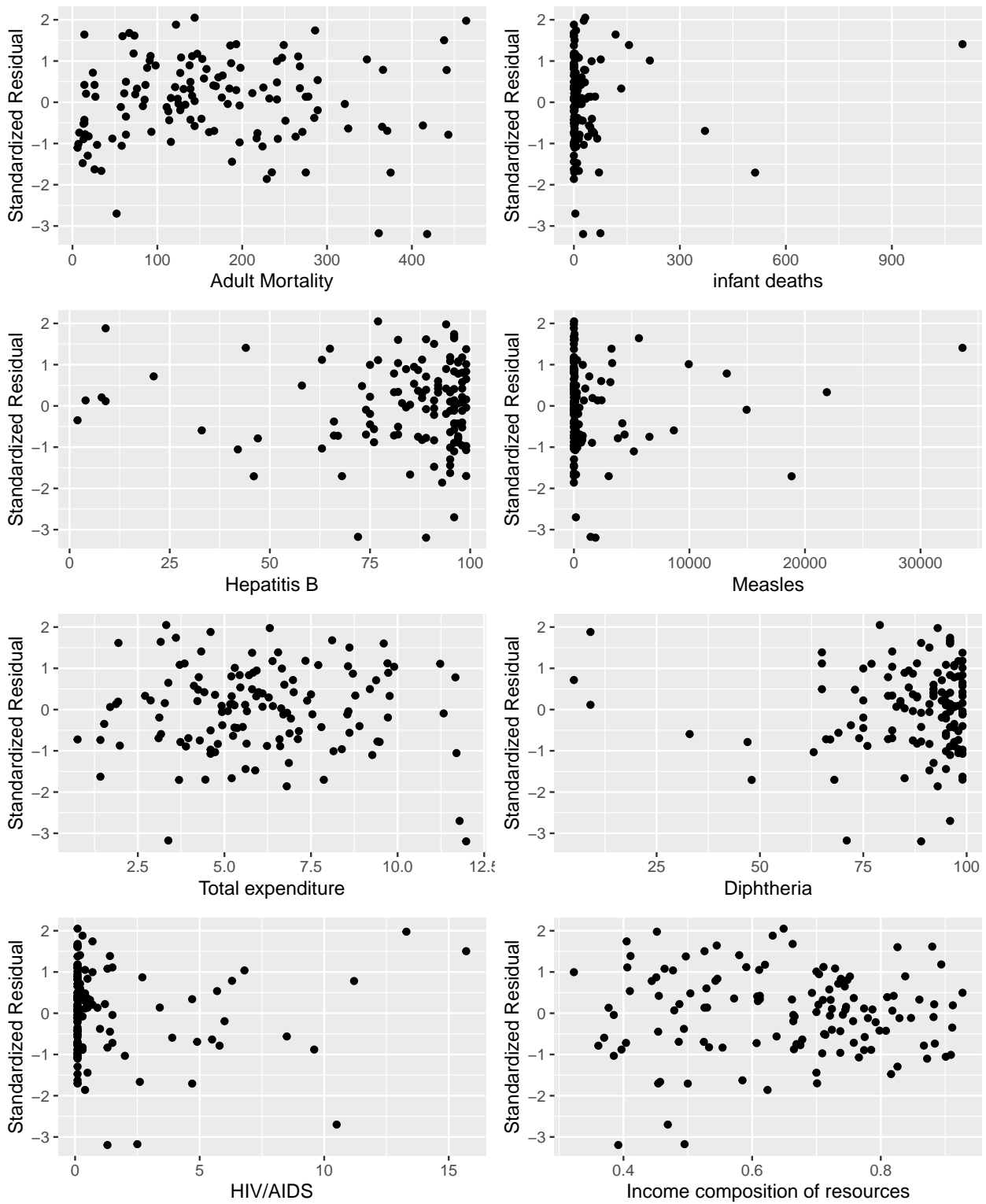Distribution of Standardized Residuals

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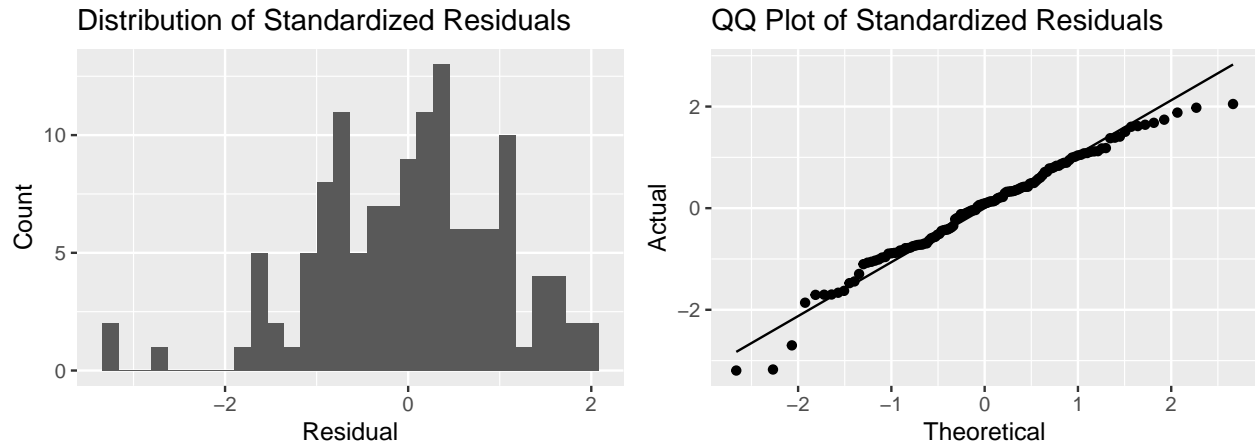
Judging from the histogram and QQ plot above, the standardized residuals appear to be normally distributed, thus fulfilling the normality condition.

Standardized Residuals vs. Predicted
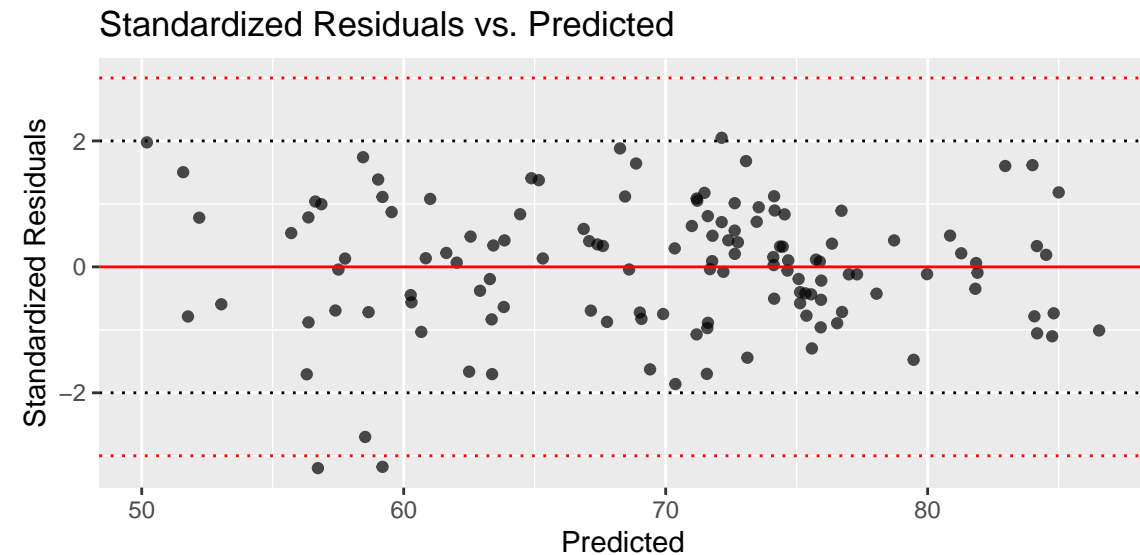
The plot above depicts standardized residuals on the y-axis and predicted values on the x-axis. We see no major departure from the constant variance assumption. However, two countries, Iraq and Lesotho are moderate outliers and two, Angola and Sierra Leone, are severe outliers in terms of their life expectancy. Specifically, Iraq has a much higher life expectancy than what we would expect given the values of its predictor variables while Angola, Lesotho, and Sierra Leone have far lower values for life expectancy than what was predicted.

Checking Linearity

The standardized residuals do not display a discernible pattern when plotted against any of the quantitative variables. This combined with the similar lack of any relationship in the plot of standardized residuals vs. predicted values above indicates that the linearity assumption holds.

Table 7: Predictor Variable Outliers

Country	Leverage
Algeria	0.345
Canada	1.000
Estonia	0.335
India	0.735
Indonesia	0.379
Latvia	0.339
Lithuania	0.345
Morocco	0.343
Myanmar	0.334
Netherlands	0.390
Tunisia	0.337
Ukraine	0.358

The 12 countries in the table above have high leverage (leverage > leverage_threshold = 0.3230769) while no countries are influential (Cook's Distance > 0.5).

Results

Table 8: 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

This is our final model. 92.47% of the variation in the `Life expectancy` is explained by the regression model above, which contains the predictor variables: `Region`, `Adult Mortality`, `infant deaths`, `Hepatitis B`, `Measles`, `Total expenditure`, `Diphtheria`, `HIV/AIDS`, and `Income composition of resources`.

Discussion + Conclusion

When looking at this selected model, a few things stand out. First, it was interesting to see the estimated coefficients of the different regions, which can be interpreted as the expected average life expectancy above

the baseline of Asia (excluding Near East), holding all other predictors constant. The region of Sub-Saharan Africa had the lowest coefficient of -2.576 while the region Western Europe had the highest coefficient of 4.905. According to our model, these discrepancies are solely because of the difference in geographic region, which logically doesn't make much sense. Therefore, it's reasonable to assume that some information on economic, social, or health-related factors that differentiate geographic regions were missing from our data. If this project, were to be expanded, it would be important to include more than 23 potential variables in the full model.

Additionally, it is interesting to see that according to our model, a unit increase in infant deaths and Hepatitis B is expected to also increase life expectancy, on average (due to the positive estimated coefficients of these variables). This clearly doesn't make much sense. However, this can be taken with a grain of salt since the p-values of these variables are relatively high. This means that assuming the null hypothesis that these variables' coefficients are equal to 0 is true, the probability of observing a coefficient at least as extreme as the one we've observed is not super unlikely. We have a similar situation for Measles, which according to our model, adds no additional information to model even though it made it past all the model selection methods we conducted.

Another interesting finding is the extremely high coefficient of `Income composition of resources`. With an estimated coefficient of 31.747, this means that for each additional unit of HDI (Human Development Index), we expect life expectancy to increase by 31.747 years, on average, holding all other predictor variables constant. However, this makes sense when it is understood that HDI is measured on a scale of 0 to 1, meaning that this outrageous increase will only be expected to happen if HDI goes all the way from 0 to 1. Therefore, it is more informative to interpret this variable in smaller units than an increase by one, such as an increase by 0.01. In that case, for each additional 0.01 increase in HDI, we expect life expectancy to increase by about 0.32 years or about 4 months, on average, holding all other predictor variables constant.

A further caveat for our model is that the data set used to fit the model contains outliers that have an outsized impact on the model itself. If we remove the four countries with large-magnitude standardized residuals from the data set and refit the model, we observe a dramatic effect on the p-value of `Total expenditure`. Whereas the coefficient had been positive and non-significant with a p-value above 0.2 in the model with the four countries, it becomes very significant with a p-value of around 0.008 once these observations are absent from the data set. Upon closer inspection, we found that individually removing Iraq, Lesotho, and Sierra Leone reduces the p-value of `Total expenditure`. However, removing Angola increases the p-value. Thus, it appears that it is the former three countries that are behind this effect. On the other hand, there are no differences between the model that excludes high leverage observations and the selected model that are as dramatic as the one between the model that excludes large-magnitude standardized residual observations and the selected model.

Appendix

Table 9: 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

References

[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