Predicting Country-Level Life Expectancy

Team 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 of 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 large increases in life expectancy. 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 years.

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, where we can look 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 dataset 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:

Response Variable:

- `Life expectancy`: Average life expectancy in years

Identifier Variable:

- `Country`: Name of country

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
- `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.

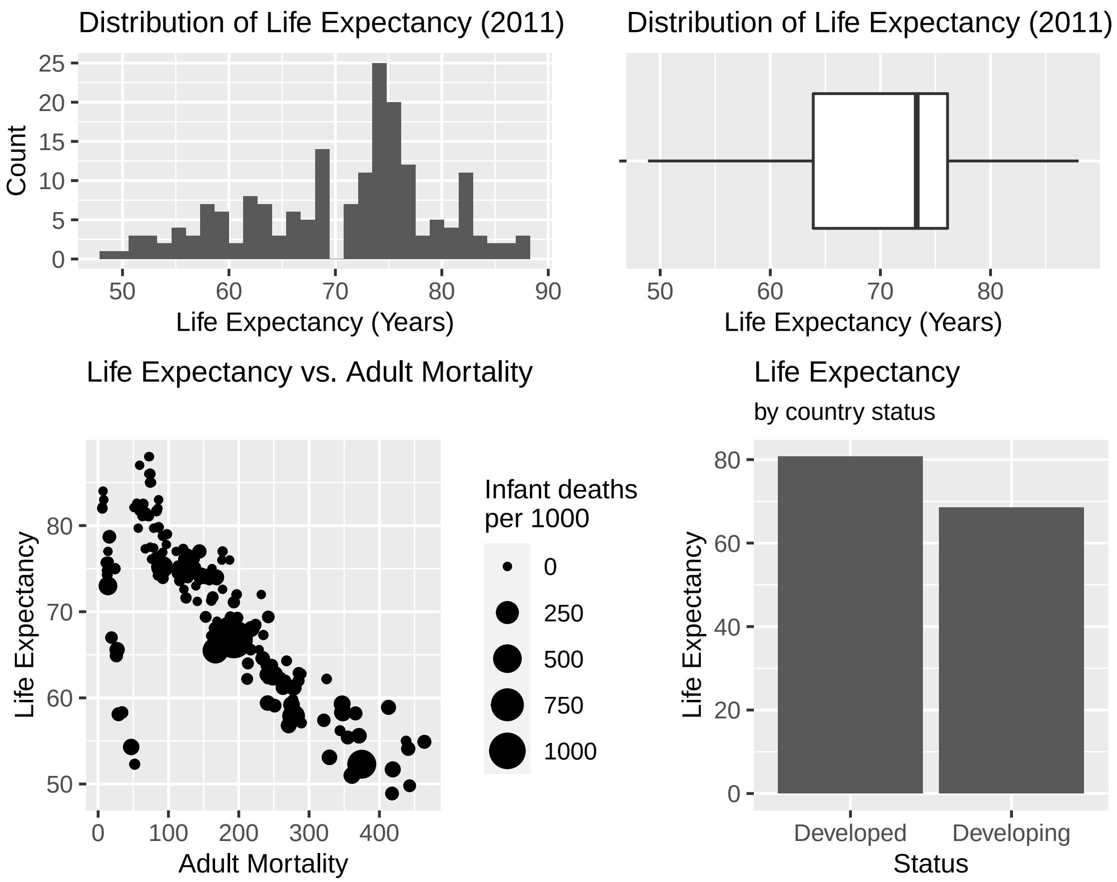
We will only analyze life expectancy for the year 2011 because it is the latest year among the years with the least missing data (no variable has a missing rate greater than 22%).

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

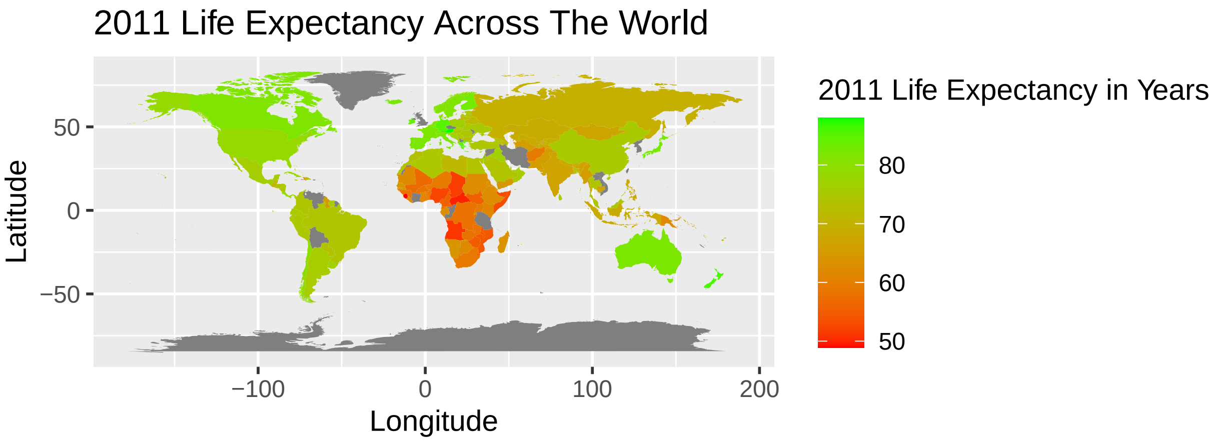
These summary statistics give a rough idea of the distribution of the response variable. The median life expectancy (~73.3 yrs) is almost 3 years greater than the mean (~70.7 yrs). Additionally, the median is closer to the third quartile than the first quartile. This suggests that life expectancy may 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 relavant predictor variables.

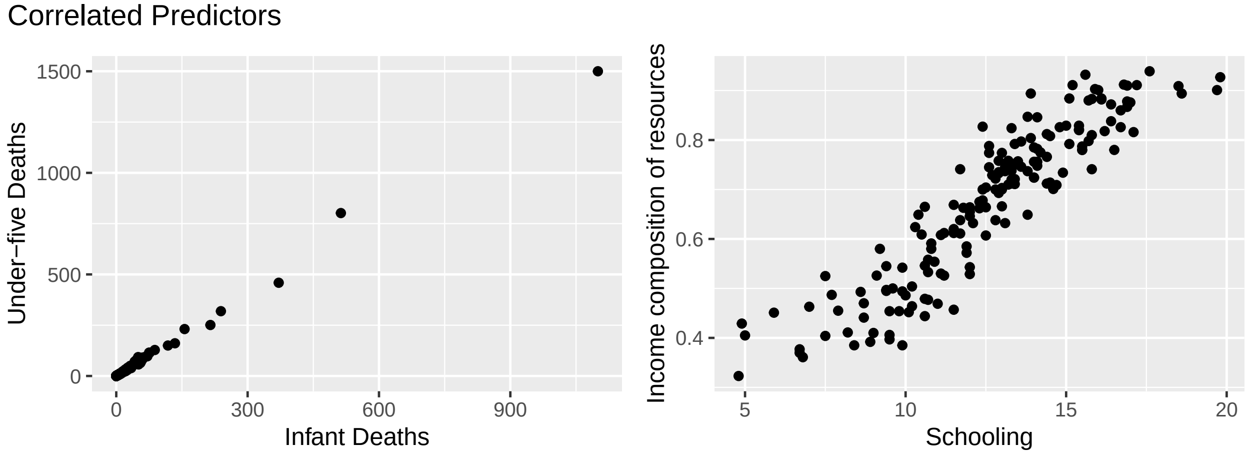


In terms of the response variable, we see that `Life expectancy` is skewed to the left with a center just over 70 years. `Life expectancy` ranges from about 50 years to 90 years.

For relationships with predictor variables, we see that a strong negative relationship exists between `Life expectancy` and `Adult Mortality`. We also a see a strange diversion in the points between countries that have lower and higher values for `Adult mortality`. There appears to be two groups of countries in the scatter plot, a smaller group that exhibits a greater `Life expectancy` penalty for a given increase in `Adult Mortality` than the larger group. We thought that this is because some of those countries have higher infant mortalities, driving life expectancy down. However, when points are sized according to `infant deaths`, this is shown not to be the case. We also see much higher `Life expectancy` among countries with the `status` “developed” than “developing”, with a difference of about 12 years.


2011 Life Expectancy Across The World
Latitude
50
0
−50
−100
0
100
200
Longitude
2011 Life Expectancy in Years
80
70
60
50

Above 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.



These two scatter plots demonstrate that the predictor variable `infant deaths` is correlated with the predictor variable `under-five deaths`. Likewise, two other predictor variables, `Income composition of resources` and `Schooling`, are clearly associated with one another. This finding will be addressed as we construct our model.

Methodology

In order to answer the research question, we will be using multiple 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 will use model selection to cull the number of potential predictor variables. We will 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 identifying outliers by examining model diagnostics.

Model Selection

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 adjusted 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. Although we cannot check all possible interactions, we see no significant interaction between the variables investigated above and in the EDA. As a result, we will not include any interaction terms in the model.

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 will 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 will 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

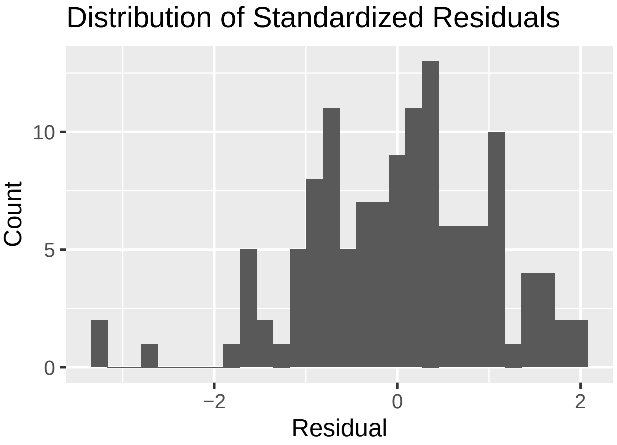
When comparing two models that are identical except that one includes `under-five deaths` and the other includes `infant deaths`, the one that includes `infant deaths` has lower values of AIC and BIC as well as a higher value of adjusted R^2.

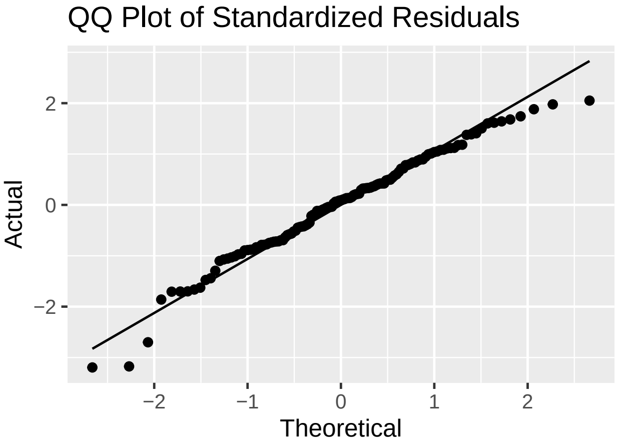
Table 6: Schooling vs. Income Composition of Resources

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





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

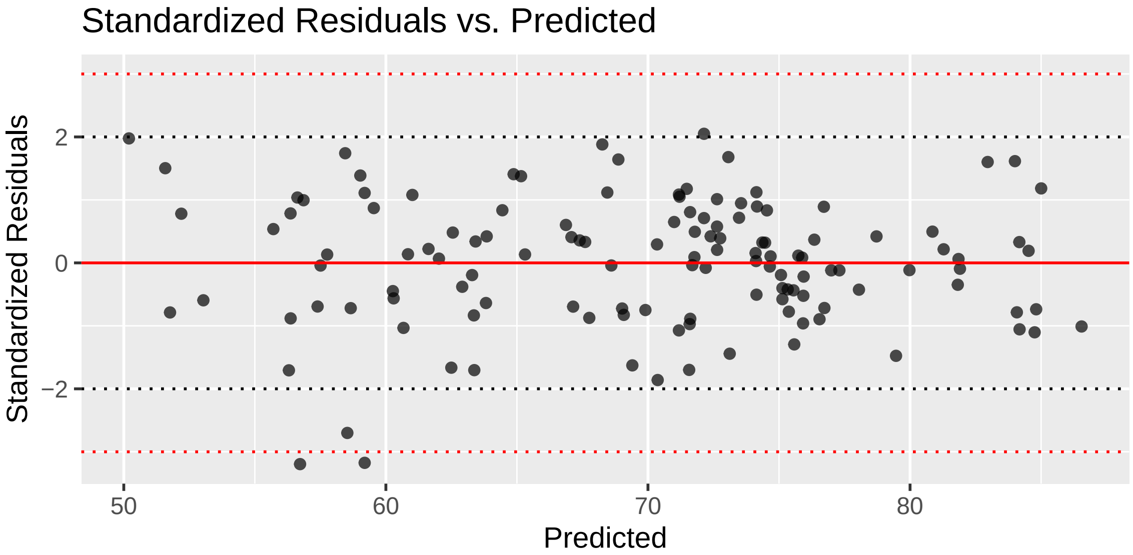


Table 7: Life Expectancy Outliers

Country	Standardized Residual
Angola	-3.175
Iraq	2.050
Lesotho	-2.700
Sierra Leone	-3.195

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, the two countries Iraq and Lesotho are moderate outliers. Two other countries, 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.

Moreover, the standardized residuals do not display a discernible pattern when plotted against any of the quantitative variables. The visualization illustrating this observation can be found below, in the appendix section. 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.

Lastly, we do not believe the independence condition for multiple linear regression can be considered fulfilled in our case. This is because countries are likely associated with one another in ways that we cannot capture with our model even though we did include `Region` as a predictor variable to try to address spatial correlation. For instance, some categories of `Region` are especially large (`SUB-SAHARAN AFRICA` encompasses 49 nations), yet others are quite small (`BALTICS` includes only 3 nations). Those countries within `SUB-SAHARAN AFRICA` may have yet more relationships with one another for which we simply cannot account with our current data set.

Table 8: 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

In terms of leverage and influential points, 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

Final Model

Table 9: 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

term	estimate	std.error	statistic	p.value
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. Around 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`. The AIC is about 642.34, and the BIC is approximately 699.69.

Key Findings and Conclusions

One interesting key finding when looking at the variables included in our selected model is that the predictors for life expectancy were extremely multifaceted. Variables that related to health-related, economic, and social measures of a country all play an important part in predicting life expectancy.

Something else that stands out is the extremely high estimated coefficient (31.747) of `Income composition of resources`. However, this makes sense when it is understood that the variable's unit (HDI) is measured on a scale of 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.

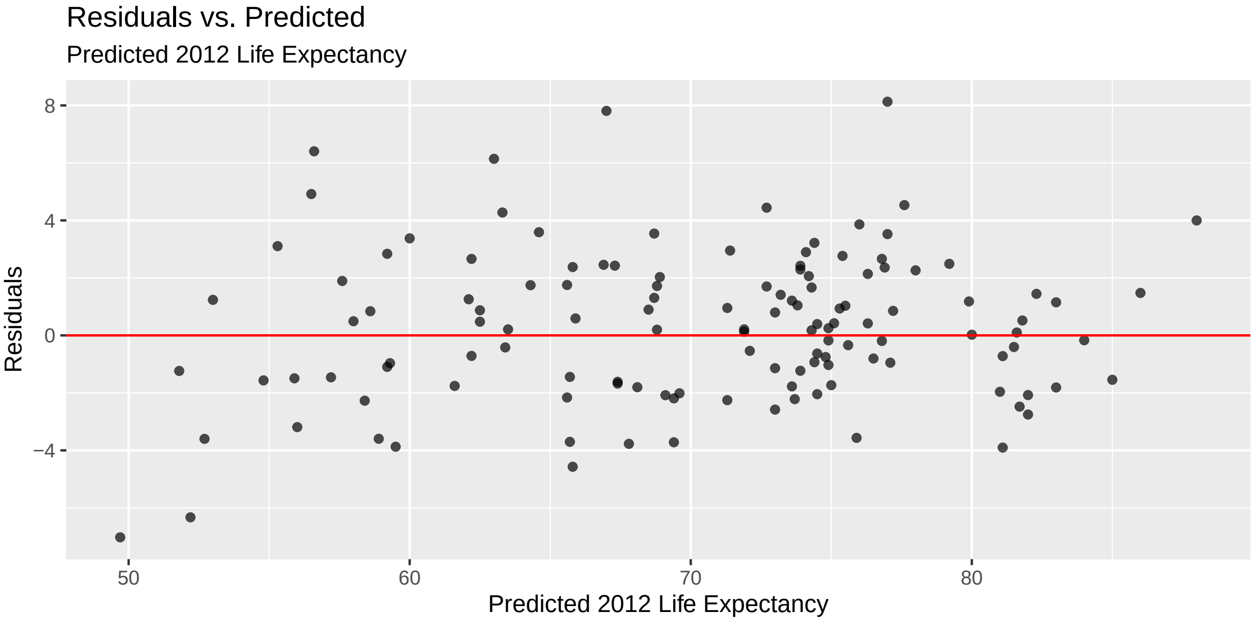
Additionally, when looking at the selected model, a few other key findings stand out. First, it was interesting to see the estimated coefficients of the different regions, which can be interpreted as the expected average life expectancy above the baseline of Asia (excluding Near East), holding all other 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 tell us the full story. Therefore, it's reasonable to assume that some information that differentiates geographic regions was 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.

Also, it is strange to see that according to our model, an increase in infant deaths is expected to increase life expectancy, on average (due to the positive estimated coefficient). This clearly doesn't make much sense. However, this can be taken with a grain of salt since the p-value is relatively high at 0.409. This means that assuming the null hypothesis that this variable's coefficient is equal to 0 is true, the probability of observing a coefficient at least as extreme as the one we've observed is not unlikely.

In terms of our initial hypotheses, the coefficient of only one variable that we thought would have a major effect on `Life expectancy` is significant: `Adult Mortality`. Per 100 adults (ages 15 to 60) out of 1000 that die, life expectancy is expected to decrease by around 1.34 years, holding all else constant. The other variables that we hypothesized at the beginning of the project, `infant deaths`, `under-five deaths`, and `Total expenditure` either are not included in the final model or are not significantly different from 0 at the 0.05 significance level. Other significant variables are `Income composition of resources` and `HIV/AIDS` with p-values near 0.

Assessing Predictive Power

To evaluate the predictive power of our model, we predicted the life expectancy for each country during 2012 and compared to the observed values from the original `life` dataset. We chose 2012 because we thought that one important application of a predictive model for life expectancy would be to predict life expectancy for the following year. The residual plot is shown below.



As seen above, the residuals for 2012 show no discernable pattern and display a relatively constant vertical spread. The residuals are positive on average, with a sum of 37.8482033. This indicates that the model slightly underpredicts life expectancy for the next year (2012), but appears fairly accurate overall. We also see that $R^2 = 0.9052248$. This means that over 90% of the variability in 2012 Life expectancy can be explained by our model, further demonstrating that it can predict future life expectancy relatively accurately.

Discussion + Conclusion

As mentioned in the Model Conditions and Diagnostics section, our model struggles when it comes to the independence condition. Indeed, the finding in the EDA that there are two distinct relationships between `Life expectancy` and `Adult mortality` among the countries seems to suggest that there is some unseen way in which countries are systematically related to each other.

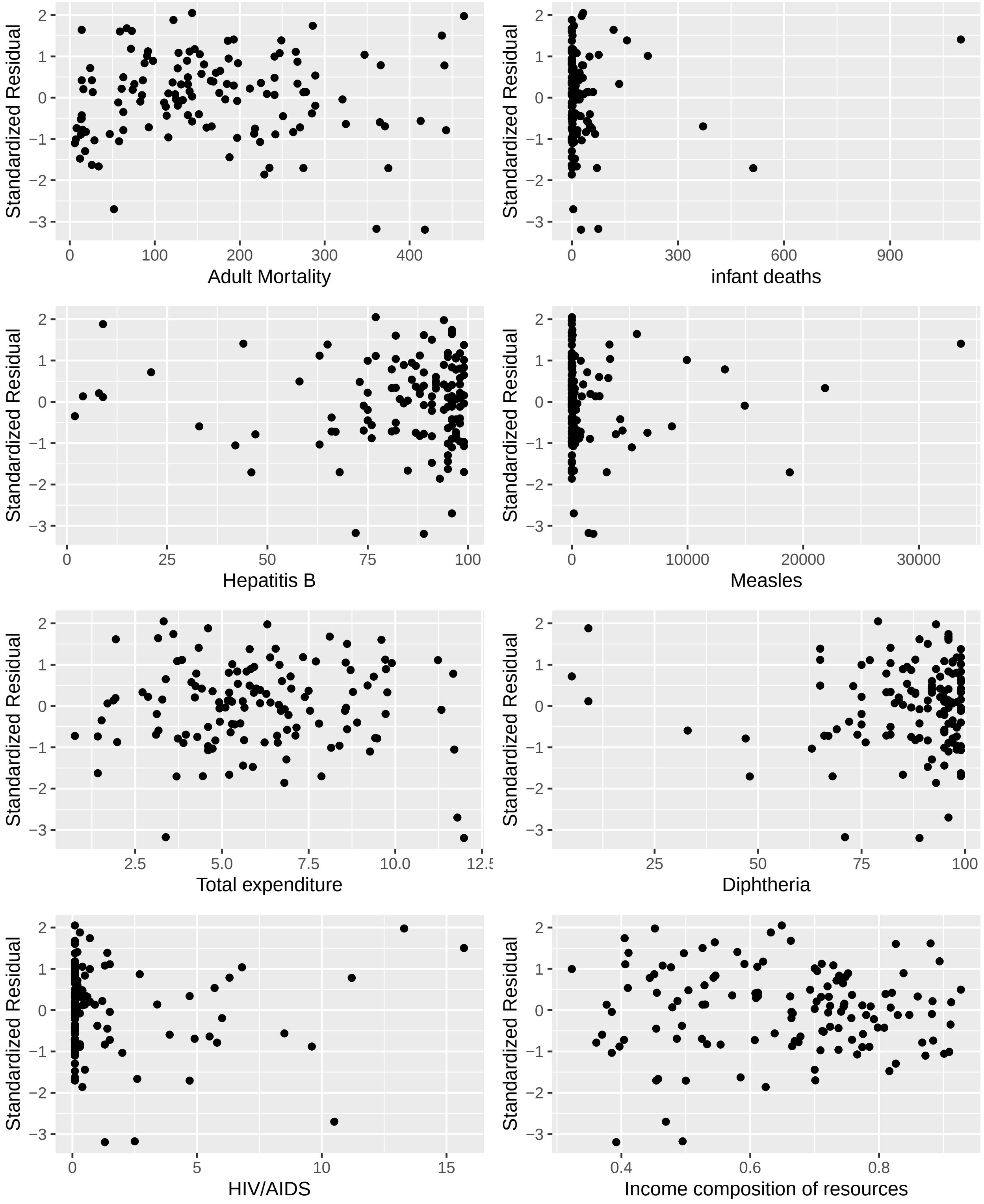
Another caveat for our model is that the data set used to fit the model contains outliers that have an outsized impact on the model itself. If we remove the four countries with large-magnitude standardized residuals from the data set and refit the model, we observe a dramatic effect on the p-value of `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 10: 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

Checking Linearity



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