Analysis of Global Life Expectancy and Related Factors

Echo Chen, Andrew Kroening, Pooja Kabber, Dingkun Yang

November 23rd, 2022

Contents

Abstract

This analysis seeks to understand which factors influence life expectancy and the development status of countries worldwide. The dataset used for this research consisted of national disease, economic, and social factors for 2013 and 2014. The primary source is the World Health Organization (WHO), with data augmented from the World Bank where necessary. We consider two research questions in the analysis: one prediction and one inference. Unique models are fit for each problem approach, and the results are analyzed for a utility to policymakers. We find that investments in education pay great dividends to overall social well-being, with a significant upside in both questions. Public health investments also carry positive upsides, although the effect is consistent with expectations.

Introduction

Life expectancy averages are increasing worldwide and have been for some time. The specific drivers of this newfound human longevity are the topic of much debate in academic and policy circles. Despite differing motivations, both groups seek to understand and find insights to improve population well-being. While some obvious indicators and areas for improvement would pay high dividends, this analysis aims to find high-payoff factors that reach beyond the economic aspects that recent studies have popularized.

This analysis uses data from the WHO and World Bank. We seek to understand the variables that most influence societal well-being. From the two research questions, we aim to improve insights into factors that drive a country's developmental status and the population health indicators that lead to improved life expectancy. Below are the questions we aim to answer in this analysis:

Research Questions

Question #1 (Prediction)

"How did major disease, economic, and social factors impact life expectancy around the globe in 2014?"

Question #2 (Inference)

"How did disease and mortality rates, along with national economic factors, contribute to a country's development status in 2014?"

Data

The dataset for this analysis contains national-level observations of variables related to life expectancy around the globe for a period spanning the early portion of the 21st century. The complete dataset includes observations beginning in the year 2000 and ending in the year 2015. As a full dataset, there are 2,938 observations for 22 variables. Practically, each country has approximately one observation each year, averaging 183 for each of the 16 years encompassed by the data. The dataset effectively contains 20 variables for each country and year combination, covering significant disease, economic, and social factors.

During the initial round of exploratory data analysis (EDA), the team identified many missing values from the dataset. These missing values included: 41x Population, 10x Hepatitis B, and 10x Schooling observations, with a large number of missing population values being the most concerning. The team considered several approaches for mitigating the problems posed by this data, as it precludes several potentially influential countries from being included in this analysis. We considered multiple imputation and scholastic imputation for mitigating these issues but ultimately decided against both approaches.

An alternative approach for missing data was identified as theoretically feasible early in the analysis. Because of the national level of our dataset, we could find suitable replacement values from another source with high integrity in these areas, such as the World Bank, the International Monetary Fund, or the CIA World Factbook. While those sources had existing data for some of our missing values, we opted not to use them for replacement. Most replacement candidates we found did not match the surrounding data points (i.e., GDP figures for the country/year in question needed to be closer to consider a match). Thus we have low confidence that those values would be consistent with the WHO's data collection methods. While options are certainly available, the team assesses that the potential gain from including the additional countries does not offset the possible bias or skew introduced by imputation.

[@DINGKUN - ADD THE POPULATION FROM WORLD BANK PART HERE]

After the initial treatments to factor variables, we reduced our dataset to complete cases only and subset for our analysis's two years of interest. After these steps and decisions, we subset the data to make two sub-datasets: one for 2014 and one for 2013, which we will use in our second research question. These two datasets are nearly identical in size, with the 2014 dataset consisting of [@DINGKUN 131] observations and the 2013 dataset having [@DINGKUN 130].

Methods

To analyze this data and accomplish the research objectives, the team began with an *a priori* variable selection. We then checked certain variables for multicollinearity, after which we conducted EDA to examine the distributions of key variables. We conducted these steps to prepare our data and methods for fitting and assessing the models used to answer both questions.

A Priori variable selection

To answer both research questions, we began by pre-selecting variables that the team believed would be the most impactful and insightful to analyze. When investigating the factors that influence life expectancy, we chose to include the following:

- Country development status.
- Population.
- Measles incidence.
- Polio immunization coverage.
- HIV/AIDS deaths among children.
- Gross domestic product (GDP).
- Income composition of resources.
- Average body mass index (BMI).
- Government expenditure on healthcare (% total expenditure).
- Government expenditure on healthcare (% GDP).
- Average years of schooling.

The second question selected slightly different variables. We will attempt to answer the second question, whether a country is developed or developing, by this set of variables:

- Life expectancy.
- Average BMI.
- HIV/AIDS deaths among children.
- Average years of schooling.
- Population.
- Income composition of resources.
- GDP.
- Government expenditure on healthcare (% GDP).

We assess that these variables provide coverage of the broad range of economic, social, and health factors included in our data and will be insightful when fit to a model.

Correlation and Multicollinearity

While conducting EDA, we found that a few variables in our dataset were potentially correlated. To determine the actual effect, we investigated the most troubling relationships. We used a correlation matrix plot and a VIF function to analyze the correlations. We found that these variables were highly correlated:

- Infant deaths and deaths under age five
- Government expenditure on healthcare (% GDP) and GDP
- Hepatitis B immunizations and Diphtheria immunizations
- Thinness variables (measuring starvation in different age groups)
- Schooling and income composition of resources

Since percentage expenditure is highly correlated with GDP and schooling is highly correlated with income composition of resources, we dropped percentage expenditure and schooling from our initial list of *a priori* variables.

Exploratory Data Analysis

The dataset for our analysis has 22 variables, as previously described. For a complete description of all variables, see Appendix A. The categorical and outcome variable for our second question is country development status. We check the relationship between this variable and the outcome variable for our first question, life expectancy. We use boxplots to find that the life expectancy in developed countries is higher, which means the categorical variable might be a good predictor for the model in the first question. The full summary of EDA plots is available in Appendix B.

For the continuous variables, we selected the ones that we believed were most insightful and examined those relationships. First, we examined the schooling, income, thinness, GDP, and HIV variables and found linear relationships consistent with the initial prediction of the variables of interest. After we dropped correlated pairs showing significantly high VIFs, we were left with the resulting dataset for analysis. Descriptive information about the variables used to address question 1 can be found in Table 1 (below).

Table 1: Summary of Variables

Statistic	N	Mean	St. Dev.	Min	Max
Population	131	22,269,096.000	116,699,866.000	41.000	1,293,859,294.000
Life.expectancy	131	70.520	8.605	48.100	89.000
percentage.expenditure	131	850.874	2,071.444	0.443	16,255.160
Measles	131	2,042.863	9,842.341	0	79,563
Polio	131	83.496	20.966	8	99
HIV.AIDS	131	0.810	1.562	0.100	9.400
GDP	131	7,256.847	14,741.400	12.277	119,172.700
Schooling	131	12.676	2.750	5.300	20.400
Income.composition.of.resources	131	0.670	0.151	0.345	0.936
BMI	131	40.476	20.734	2.000	77.100
Total.expenditure	131	6.107	2.533	1.210	13.730

Almost all the data have significant differences between the minimum and maximum values, suggesting that the country measurements are spread across a wide range. The population, GDP, and Measles all have large standard deviations, and the mean is exceptionally close to the maximum, which may indicate skewed data. As we prepare for model fitting and assessment, these observations are essential to keep in mind. This is our data ready for model selection.

Results

Question 1 (Prediction)

Results

We then proceeded to model fitting. To answer the first research question, we fit a simple linear regression model with life expectancy as our response variable. We fit three models to study the relationship between the independent variables and life expectancy, including only our *a priori* variables, including all variables in the data and using backward stepwise selection. To find the best fit parsimonious model, we evaluated the models using adjusted R^2 and BIC. In our final model, we included the variables selected by backward stepwise selection and our *a priori* variables but excluded the variables with high correlation that we discovered during EDA.

Finally, we obtain diagnostic information about the performance of our models.

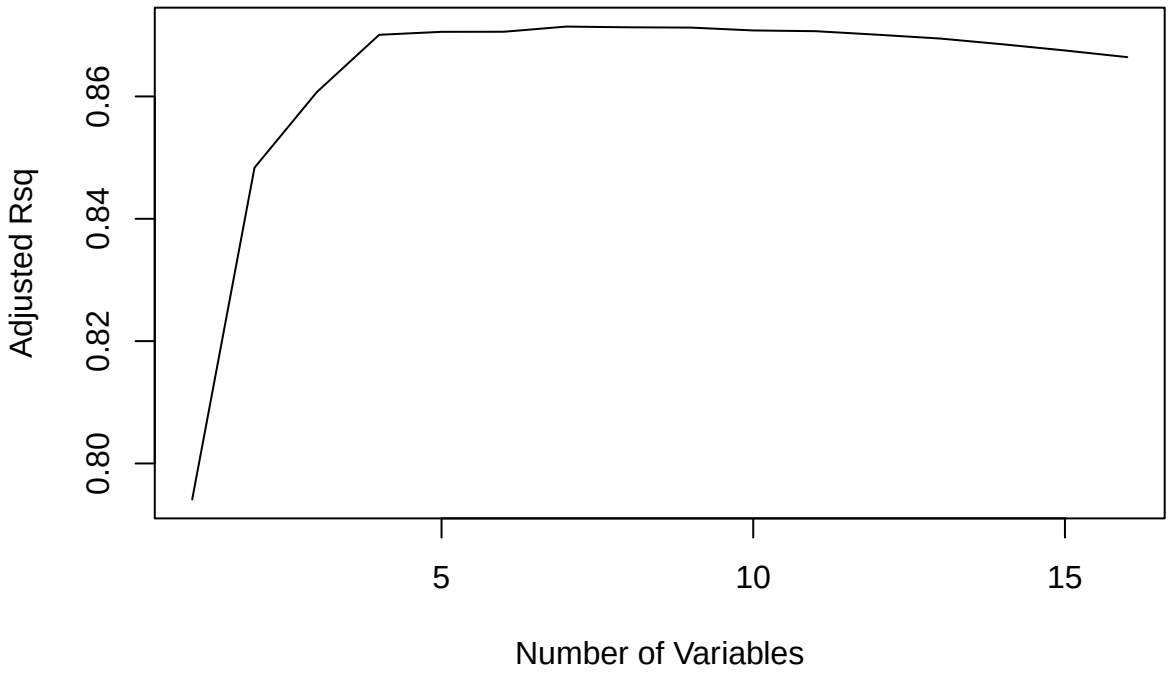
Model selection based on AICc:

```
                K   AICc Delta_AICc AICcWt Cum.Wt      LL
```

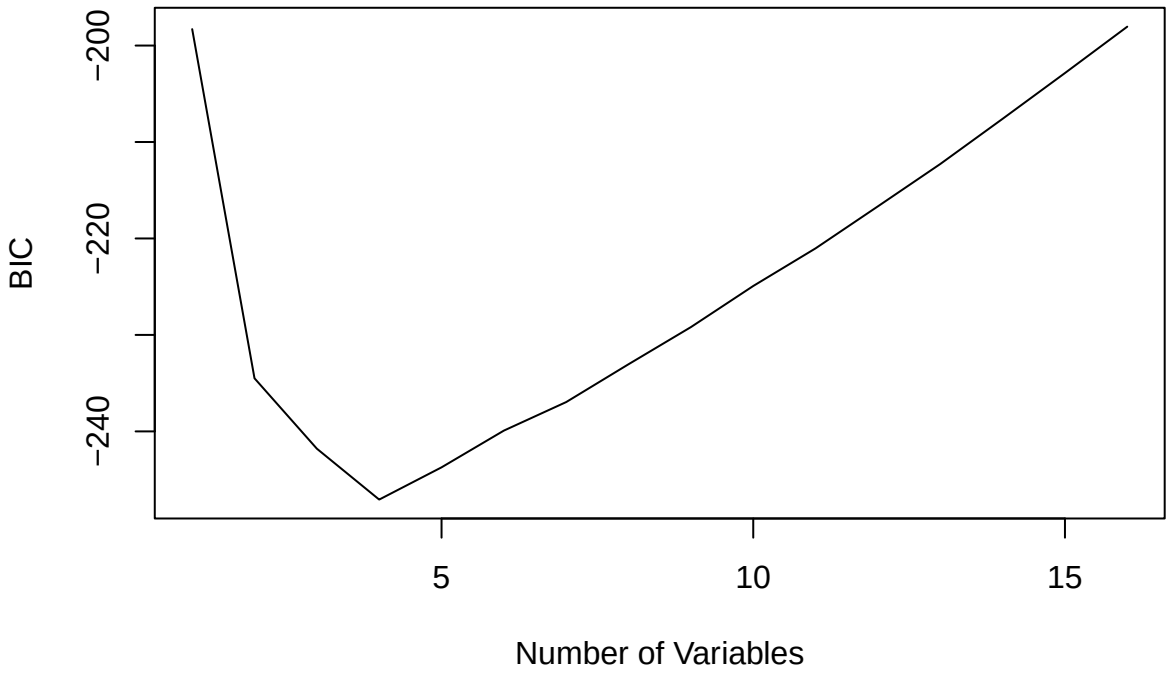
backwardstepwise 8 677.87 0.00 1 1 -330.34 all 21 704.17 26.30 0 1 -326.85 apriori 11 706.28 28.41 0 1 -341.03

% latex table generated in R 4.2.2 by xtable 1.8-4 package % Thu Dec 1 22:09:14 2022

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	49.5586	2.9829	16.61	0.0000
StatusDeveloping	-1.0507	0.9883	-1.06	0.2898
Population	0.0000	0.0000	0.33	0.7416
Measles	-0.0000	0.0000	-0.38	0.7043
Polio	0.0026	0.0148	0.18	0.8609
HIV.AIDS	-0.8647	0.2390	-3.62	0.0004
GDP	0.0000	0.0000	0.21	0.8356
Income.composition.of.resources	34.6152	3.5565	9.73	0.0000
BMI	0.0001	0.0180	0.00	0.9964
Total.expenditure	0.3286	0.1168	2.81	0.0057
Adult.Mortality	-0.0179	0.0040	-4.52	0.0000




```
## [1] 7
```




```
## [1] 4
```

Model Assessment

To check if the model that we selected follows the linear regression assumptions, we will plot the model summary plots and also interpret the F-statistic in the model results. The resulting plots are available in Appendix C.

The linearity assumptions are met throughout the model. We can note that all linear regression assumptions are roughly satisfied.

- Residuals and Fitted Plots: There is no pattern in the residual plot. We can assume a linear relationship between the predictors and the outcome variables. (Linearity assumption)
- Scale-Location Plot: The residuals are equally distributed over the range of predictors. (Homogeneity assumption)
- Normal Q-Q plot: All points are distributed along the reference line; again, it is good to assume that the data have the normality of the residuals. (Normality of residuals assumption)
- Relationship between residuals and leverage: Any point at 0.5 kitchen distance from the boundary is influential. However, there is no significant point here. (Linearity assumption)

Model Interpretation [This is not consistent with the brief]

We devised the final model based on the three linear models 'apriori', 'all,' and 'backward stepwise.' The variables that have a crucial impact on population longevity are [@POOJA @ECHO development status, population, Measles, Polio, and HIV.AIDS, GDP, Income.composition.of.resources, BMI, Total.expenditure, Adult.Mortality]

. According to the statistics, the first information is the residual summary statistics, and we can see that the median should be close to 0, i.e., the mean of the residuals. 3Q and 1Q should be quantitatively close to each other and symmetrically distributed, and the errors follow a Gaussian distribution. For the second observation, Estimate, all coefficients are very small, and we may make the coefficients easier to observe by adjusting the unit size. Total.expenditure, HIV.AIDS, Income.composition.of.resources, StatusDeveloping The larger absolute values of the coefficients for these four variables indicate that a one-unit change in these four variables would have a more severe effect on average life expectancy than the other variables, other things being equal. Judging the p-values of these variables according to our setting of a=0.05 revealed that

among them, Income.composition.of.resources, HIV.AIDS and Adult. Mortality is much less than 0.05 and has a very significant effect on life expectancy. Total. expenditure is close to 0.05 and is also a very important influence and somehow significant. Finally, we can observe that by looking at R-squared: 0.8758, Adjusted R-squared 0.8654 Our final model matches the data to a degree of 87%. In summary, our final model using ten variables has 87% accuracy in meeting all assumptions, where the most critical influences on mean life expectancy are these four variables: - HIV.AIDS (-8.647e-01) - Income.composition.of.resources (3.462e+01) - Total.expenditure(3.286e-01) - Adult.mortality(-1.794e-02) ranked in order.

Question 2: (Inference)

Results

Table XXXX: Logistic Regression Models (below) shows the output of 8 predictor variables regressed onto country development status in four models. From left to right those models are:

1. the initial full model
2. a model with Health Expenditure/GDP per capita dropped for multicolinearity concerns
3. the potential model fit after all logistic transformations with Income Composition of Resources included
4. the final model fit after dropping Income Composition of Resources

From the model output we observe that only one predictor variable has a p-value less than 0.05 denoting it as a significant predictor: *Years of Schooling*. The interpretation of this predictor's coefficient in terms of the odds of a country being identified or labeled as a developed country is: while holding all other predictor variables constant, a one year increase in *Years of Schooling* make it 2.39 ($e^{1.05}$) times more likely that country being identified as developed country.

We also examine multicolinearity concerns for all four model fits. From the table below it is clearly observable that the first model has significant concerns, with two variables scoring a VIF of over 10. After we remove one variable, all subsequent models are satisfactory, with no variables scoring high on this test.

Assessment

Before assessing the accuracy of the potential model fit (3rd model) for this research question, we realize the interpretation of *Income Composition of Resources* variable might be troublesome. It recorded as Human Development Index in terms of income composition of resources (index ranging from 0 to 1), however the meaning behind it is confusing. We decided to adjust the model to omit *Income Composition of Resources* variable, and try to predict the development status of each country in the dataset. Those predictions are used to build a confusion matrix, shown in the table below. From this table, we can determine the True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) rates by comparing predicted values and actual values. For prediction, we first use the threshold of 0.5 to classify countries as developed or developing. The accuracy of this model is approximately 91%, with 95% Confidence Interval of within 83% ~ 95%.

A useful tool for measuring our predictions is the Receiver Operator Characteristic (ROC) curve with the axis as Sensitivity (true positive rate) and 1 - Specificity (true negative rate). We find out the optimal prediction cut-off for year 2014 dataset is 0.232, and we can see that the area under the curve is 0.959, a very strong number. We leave the value as described because the small number of countries designated as developing in our dataset means that each missed prediction carries greater weight in this group. We observed 5 false positives and 8 false negatives in our matrix, and are satisfied with this result.

Table 2: Logistic Regression Models

	Dependent variable:			
	Development Status			
	(1)	(2)	(3)	(4)
Life Expectancy	−0.06 (0.12)	−0.07 (0.12)	−0.07 (0.12)	0.19 (0.10)
	p = 0.62	p = 0.58	p = 0.59	p = 0.06*
Health Expend. /GDP per capita	−0.0003 (0.0005)			
	p = 0.59			
BMI	−0.04 (0.03)	−0.04 (0.03)	−0.03 (0.03)	−0.03 (0.02)
	p = 0.20	p = 0.21	p = 0.23	p = 0.25
Gov. Expend. on Healthcare	0.22 (0.17)	0.19 (0.16)	0.24 (0.17)	0.16 (0.14)
	p = 0.21	p = 0.24	p = 0.17	p = 0.27
HIV/AIDS Deaths/1k live births	−151.19 (25,152.25)	−151.66 (25,302.50)	−154.85 (25,727.43)	−165.21 (16,977.18)
	p = 1.00	p = 1.00	p = 1.00	p = 1.00
GDP per capita	0.00002 (0.0001)	−0.00002 (0.00003)		
	p = 0.80	p = 0.55		
Log of GDP per capita			−0.39 (0.26)	−0.26 (0.26)
			p = 0.14	p = 0.33
Std. Income Composit. of Resources	39.24 (15.62)	38.65 (15.56)	38.73 (14.71)	
	p = 0.02**	p = 0.02**	p = 0.01***	
Log of Population	−0.27 (0.28)	−0.26 (0.28)	−0.34 (0.30)	−0.39 (0.26)
	p = 0.34	p = 0.35	p = 0.26	p = 0.13
Years of Schooling	0.07 (0.43)	0.12 (0.44)	0.27 (0.50)	1.05 (0.41)
	p = 0.87	p = 0.80	p = 0.59	p = 0.02**
Constant	−9.32 (2,515.24)	−8.84 (2,530.27)	−7.16 (2,572.75)	−6.20 (1,697.73)
	p = 1.00	p = 1.00	p = 1.00	p = 1.00
Observations	129	129	129	129
Log Likelihood	−19.66	−19.83	−18.86	−23.59
Akaike Inf. Crit.	59.33	57.66	55.72	63.19

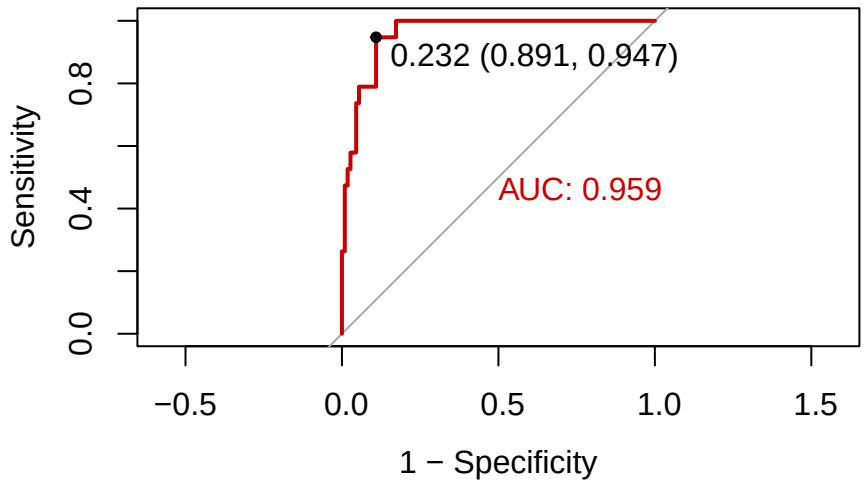
Note: *p<0.1; **p<0.05; ***p<0.01

Table 3: Variance Inflation Factors

	(1)	(2)	(3)	(4)
Life Expectancy	2.10	2.09	2.13	1.40
Health Expenditure/GDP per capita	10.77			
BMI	1.28	1.28	1.19	1.26
Health Gov. Expenditure Percentage	1.20	1.11	1.22	1.26
HIV/AIDS Deaths/1000 live births	1.00	1.00	1.00	1.00
GDP per capita	10.59	1.54		
Std. Income composition of resources	4.13	4.20	3.52	
Log of Population	1.29	1.31	1.40	1.52
Years of Schooling	2.28	2.31	2.50	2.17
Log of GDP per capita			1.54	1.72

Table 4: Confusion Matrix for Final Model

	True Developed	True Developing
Predicted Developed	11	5
Predicted Developing	8	105

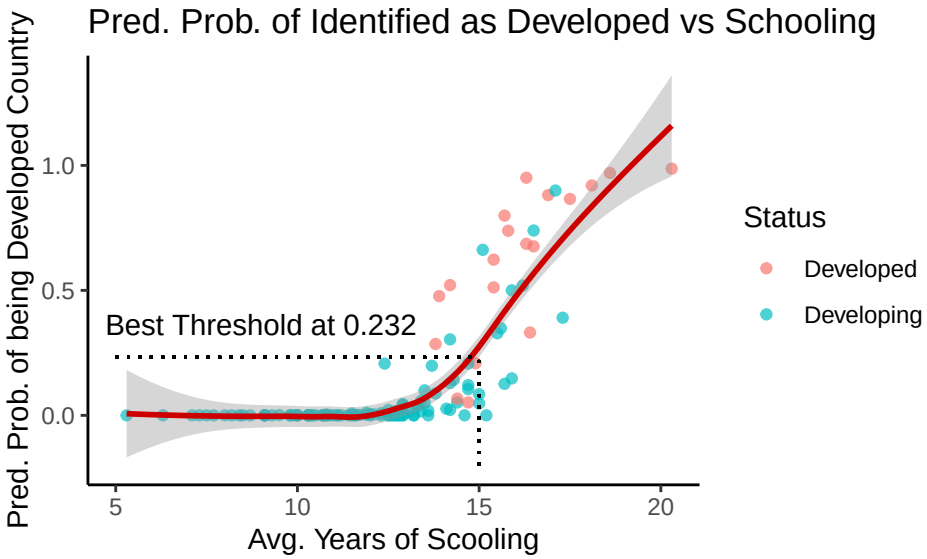


Out-of-Sample Predictions

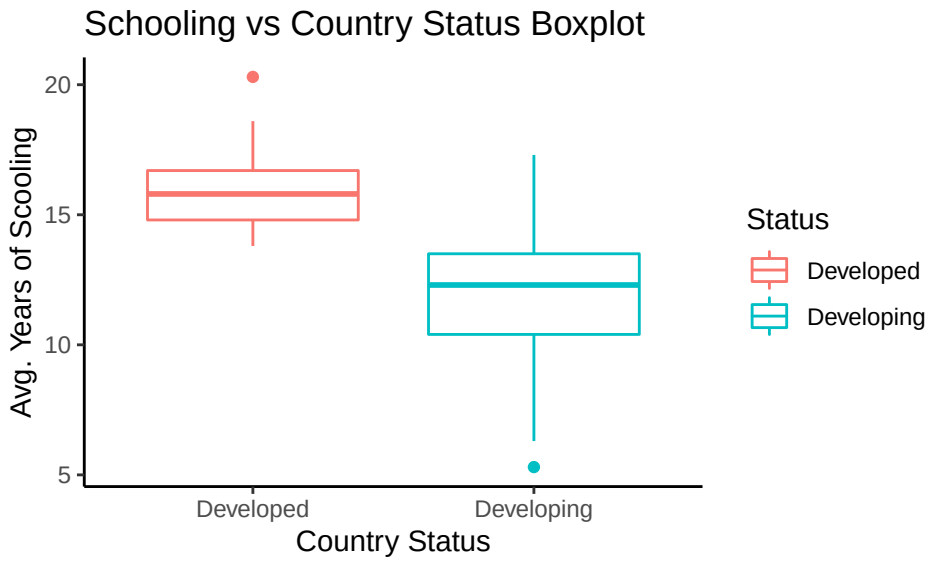
A final validation step for this model was to predict out-of-sample probabilities for the Year 2013 dataset using "optimal" threshold, 0.232, as mentioned above for inferring developed or developing status. The result of this experiment is shown in the table below. The accuracy of these predictions is still 0.91 with 95% Confidence Interval of being within the range of 84 ~ 95% , which is approximately the same accuracy as our training data set, and should be considered a pretty good fit.

Table 5: Confusion Matrix for Inferring Year 2013 Data

	True Developed	True Developing
Predicted Developed	16	9
Predicted Developing	3	100



As you can see from the plot above, with keeping all the other variable constant, given a country with people's average years of schooling being larger than around 15 years, we may infer that the country would be more likely identified as a developed country than developing ones. On the other hand, if a country with people's average years of schooling being lower than 15 years, according to our model, we shall infer the country as developing country.



Final Model Evaluation

Out of curiosity, we compare the final model with a model with only one predictor variable, *Years of Schooling*. An ANOVA result surprisingly shows that two models are not statistically different in terms of "inference" accuracy. As we can see in the table below (Analysis of Deviance: Final Model vs Model w/ One Predictor Variable), the p-value being 0.07 is greater than 0.05, meaning *Years of Schooling* variable by its own is a great indicator of whether the country should be considered as developed or developing country.

Table 6: Analysis of Deviance: Final Model vs Model w/ One Predictor Variable

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	121	47.19			
2	127	58.92	-6	-11.73	0.07

Conclusion

Describe the key takeaways from your analysis, limitations, and future work that can be done to advance knowledge in this area.

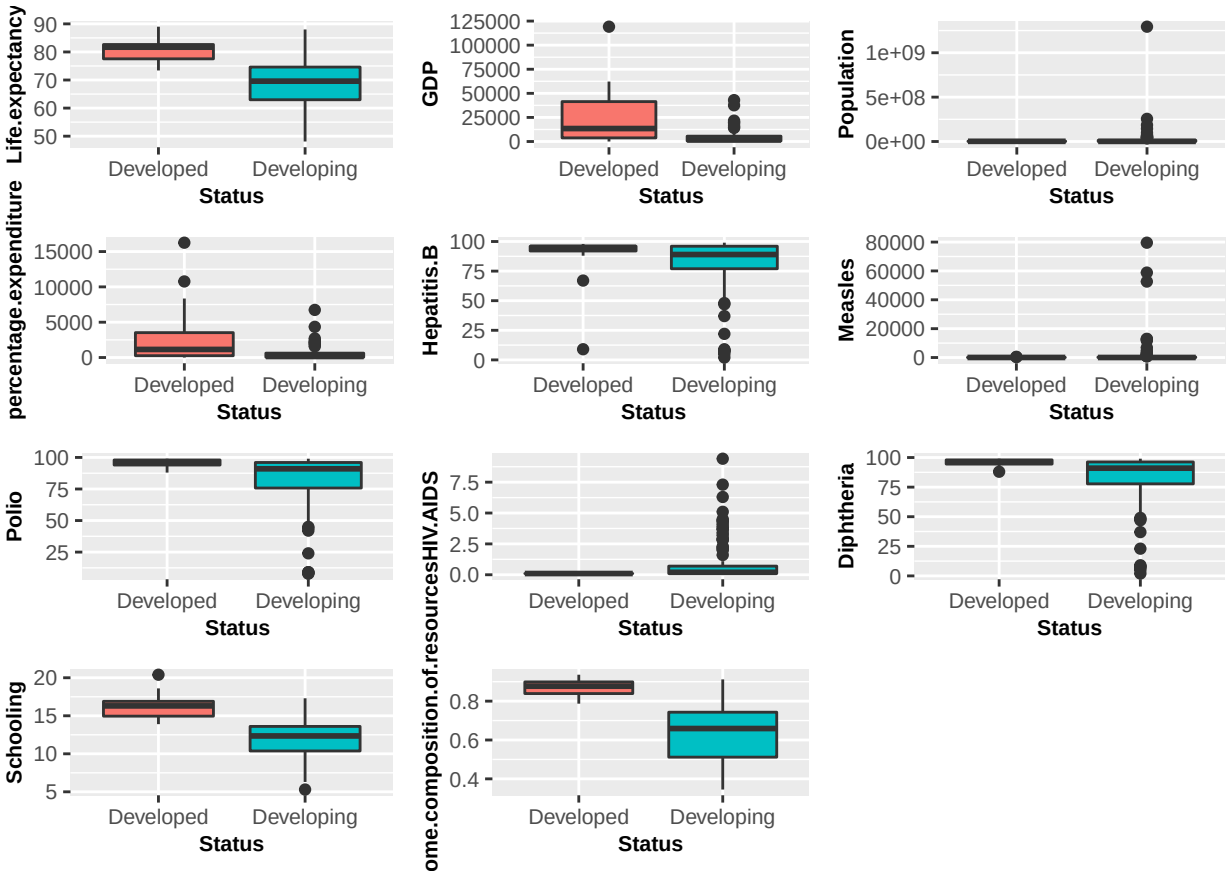
[What is the impact of this analysis, do we think it is insightful or not?]

Appendix

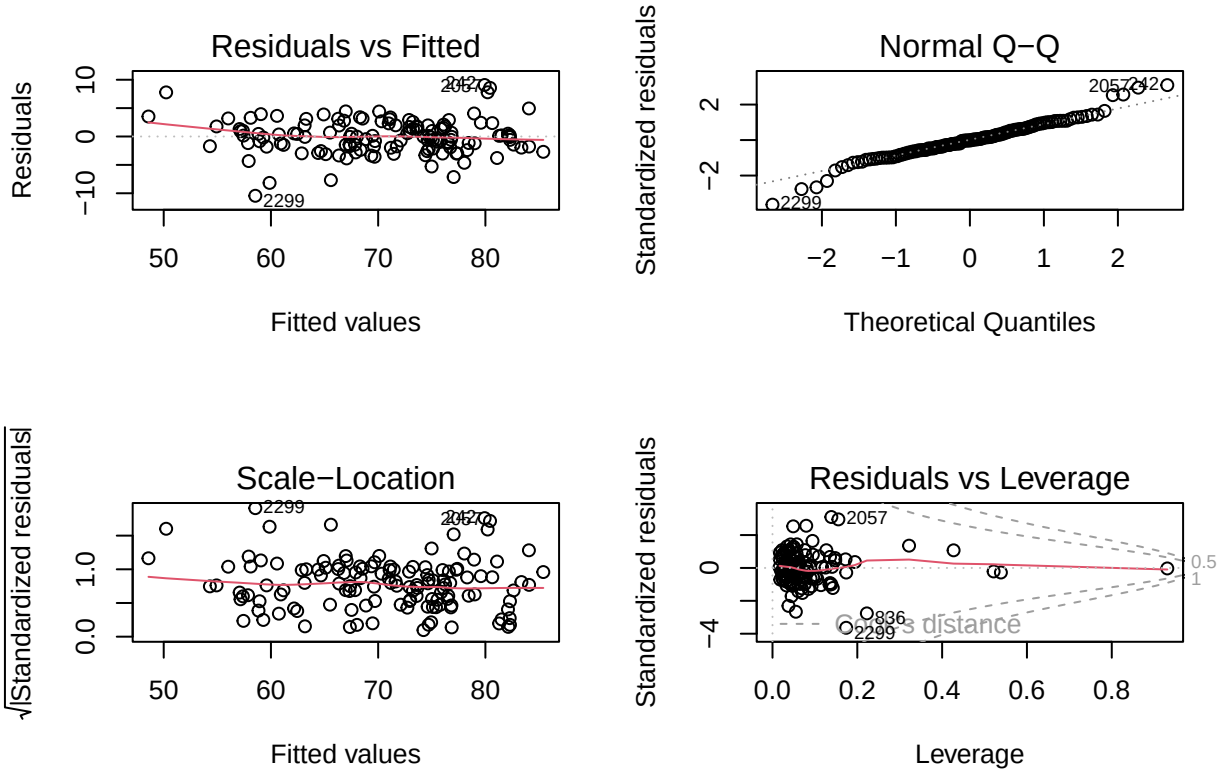
Appendix A - Variable Descriptions

Variable	Type	Description
Country	factor	Country name
Year	numeric	Year of the data
Status	factor	Country status of developed or developing
Life_Expectancy	numeric	Life expectancy in age
Adult_Mortality	numeric	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
infant.deaths	numeric	Number of Infant Deaths per 1000 population
Alcohol	numeric	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
percentage.expenditure	numeric	Expenditure on health as a percentage of Gross Domestic Product per capita(%)
Hepatitis.B	numeric	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Measles	numeric	number of reported cases per 1000 population
BMI	numeric	Average Body Mass Index of entire population
under.five.deaths	numeric	Number of under-five deaths per 1000 population
Polio	numeric	Polio (Pol3) immunization coverage among 1-year-olds (%)
Total.expenditure	numeric	General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria	numeric	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
HIV.AIDS	numeric	Deaths per 1 000 live births HIV/AIDS (0-4 years)
GDP	numeric	Gross Domestic Product per capita (in USD)
Population	numeric	Population of the country
thinness..1.19.years	numeric	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
thinness.5.9.years	numeric	Prevalence of thinness among children for Age 5 to 9(%)
Income.composition.of.resources	numeric	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	numeric	Number of years of Schooling(years)

Appendix B - EDA Plots



Appendix C - Residual Plots, Question 1



[Presently, a dumping ground for all our images and lots until we know what we want to keep]