Analysis of Global Life Expectancy and Related Factors

Echo Chen, Andrew Kroening, Pooja Kabber, Dingkun Yang

November 23rd, 2022

Report: Your report will be an 8-10 page self-contained document describing your analysis. It should be written as a professional document that can be understood by someone with limited statistics background (e.g., a client). You are also required to submit an RMD file that includes your code for the EDA and analysis. The report should be organized as follows:

intro - 2 pages eda - 1 page question 1 - 3 pages question 2 - 3 pages conclusion - 1 page

Abstract

This analysis is conducted to understand which factors influence life expectancy and the development status of countries around the world. The dataset used for this research consists of national disease, economic, and social factors and was compiled by the World Health Organization (WHO). [Need to link to the dataset from Kaggle] To conduct the analysis, two research objectives are formulated: one prediction question and one inference question. Unique models are fit to each approach and the results are analyzed for utility. [Add a concluding punchline]

Introduction

Provide more background on the data and research questions. Be sure to cite the data and background information appropriately (APA style is fine)

This analysis uses data from the WHO to better understand the drivers of life expectancy around the globe. We also attempt to find inferential value from the data for determining the developmental status of a given country. 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.

The particular 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. Below are the questions we aim to answer in this analysis:

(Cite data here)

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

[Data introduction: address missing, factors, cleanliness, subsetting, and eda]

During the course of this initial rouund of EDA, the team identified a number of missing values from the dataset. These missing values included: 41x Population, 10x Hepatitis B, and 10x Schooling observations, with the large number of missing population values the most concerning. The team considered several approaches for mitigating the problems posed by this data, as it precludes a number of potentially influential countries from being included in this analysis. We considered multiple imputation as well as scholastic imputation for ways to mitigate these issues.

An alternative approach for missing data was identified as theoretically feasible early in the analysis process. Because of the national-level of our dataset, it is possible that 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 opt to not 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 were not close enough to consider a match), and thus we have low-confidence that those values would be consistent with the WHO's data collection methods.

After the initial treatments to factor variables, our dataset is reduced to complete cases only and subset for the two years of interest in our analysis. We ultimately decide to preserve the original integrity of the data and bypass any available imputation methods. While there are certainly options available, the team assesses that the potential gain from the inclusion of the additional countries does not offset the possible bias or skew introduced from imputation. At the conclusion of these steps and decisions, we subset the data to make two sub-datasets: one for the year 2014 and one for the year 2013 which we will use in our second research question. These two datasets are nearly ideintical in size, with the 2014 dataset consisting of 131 observations, and the 2013 dataset having 130.

Methods

Describe the process you used to conduct analysis. This includes EDA and any relevant data cleaning information (e.g., did you exclude missing values? If so, how many? Did you collapse categories for any variables?) Then describe the models you fit, and any changes you made to improve model fit (e.g., did you exclude any influential points? Did you do have to address multicollinearity issues? Did you transform any variables?). Also describe model diagnostics. The organization of this section may depend on your particular dataset/analysis, but you may want to break it into subsections such as "Data," "Models," and "Model assessment." Note that you do not present any results in this section.

(The general methodology of our analysis centered around the ability to draw insights from the dataset without significant transformations or imputations. - do we need this?) To analyse this data / accomplish this objective, the team began with a priori variable selection, after which we conducted exploratory data analysis (EDA) to examine the distributions of key variables.

(A Priori variable selection) These are the variables we selected as part of our initial a priori variable selection which included selecting the variables whose relationship with Life expectancy we were interested in: Status + Population + Life.expectancy + Measles + Polio + HIV.AIDS + GDP + Income.composition.of.resources + BMI + Total.expenditure + percentage.expenditure + Schooling

(Dealing with missing data) From the EDA, missing data points were identified. Since imputing the missing values using mean or regression methods is not valid in this context where each value is for a country and using these methods may create an estimate significantly distant from the true value, we considered finding the missing data from other sources. We tried sources like (). Here we discovered that the data points for factors we had differed in these alternate datasets. Since the data collection time and methods may have been different for these sources, we did not use this method to impute the missing data. In the end we decided to drop the 52 rows with null values. At conclusion of data preparation, we subset for the year 2014, our focal point for this analysis, and the year 2013 as a testing validation dataset used in our second research question.

(Correlation and multicollinearity) While conducting EDA, we found that a few of the variables in our data were highly correlated. To analyse this correlation, we used a correlation plot and VIF. We found that these variables were highly correlated:

1. Infant deaths and Under five deaths
2. Percentage expenditure and GDP
3. Hepatitis B and Diphtheria
4. Thinness variables ()
5. 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 apriori variables.

(EDA interpretation) As mentioned before, this dataset has 22 variables. The categorical variable in our table is status. To understand their relation, we need to check the boxplots of life expectancy over the two categories developed and developing. We can check the 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. For continuous variables, by name only, we expected the first phase and selected the ones we were interested in. Here we can find statistical evidence by checking correlation and collinearity. Based on the correlation matrix and Variance Inflation Factor (VIF).First,schooling, income, thinness, gdp and hiv all find possible linear relationships, which is also consistent with the initial prediction of the variables of interest. Then,we have to drop correlated pairs to show significantly high VIFs, and we have data shown in table 1. Our Ready data has 131 rows of observations. Including 12 variables, one categorical and 11 continuous variables, their detailed information, like mean standard diviation min and max values, can be found in table 1.Almost all of the data have significant differences between the minimum and maximum values, suggesting that the country has a significant prior wealth gap, population gap, etc. The population, GDP, and Measles all have large std, and the mean is exceptionally close to the maximum, which may lead to skewed data. We may need to check the assumptions to exclude the effect of this situation on the model.This is our data ready for model selection.

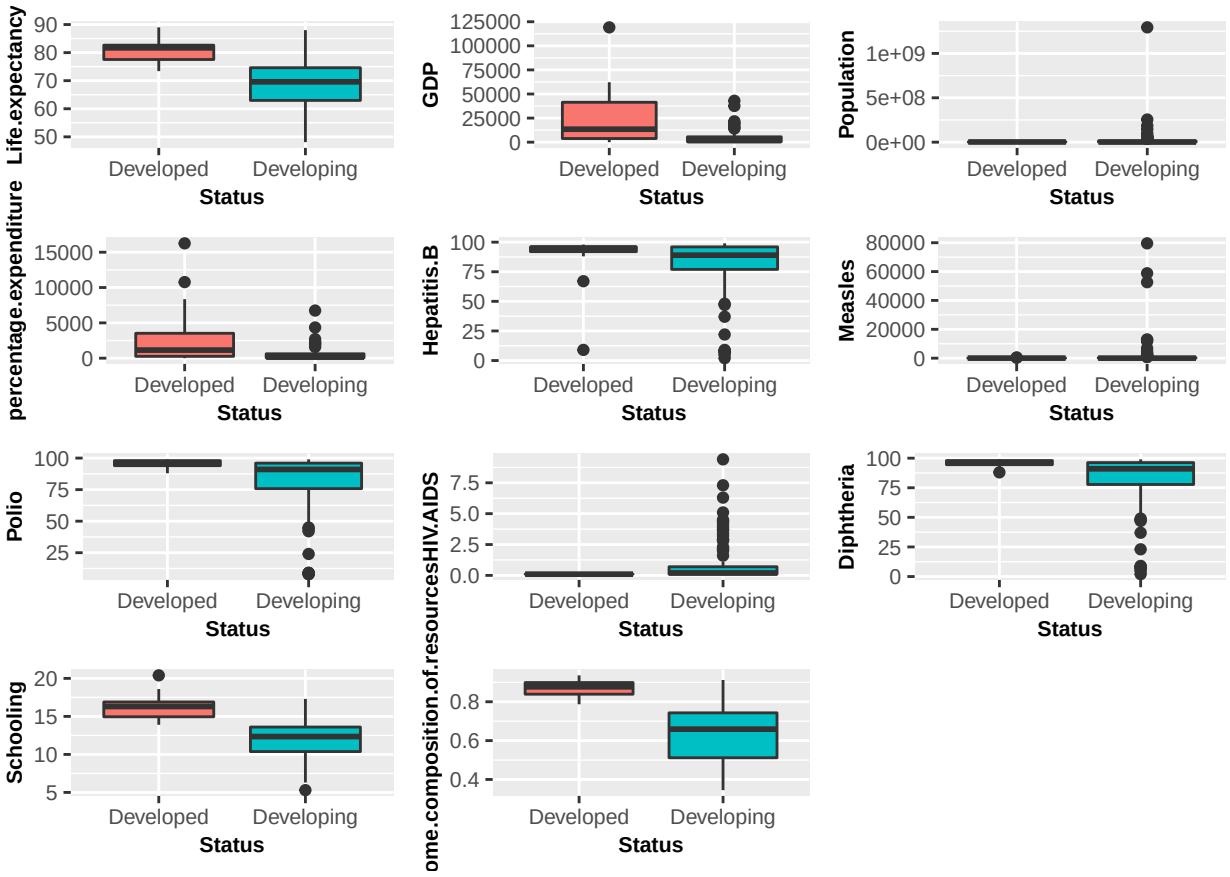


Table 1: Summary of Variables

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

[Describe Table 1 - feels like we should add more variables]

Models

Question 1

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 collinearity that we discovered during EDA.

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

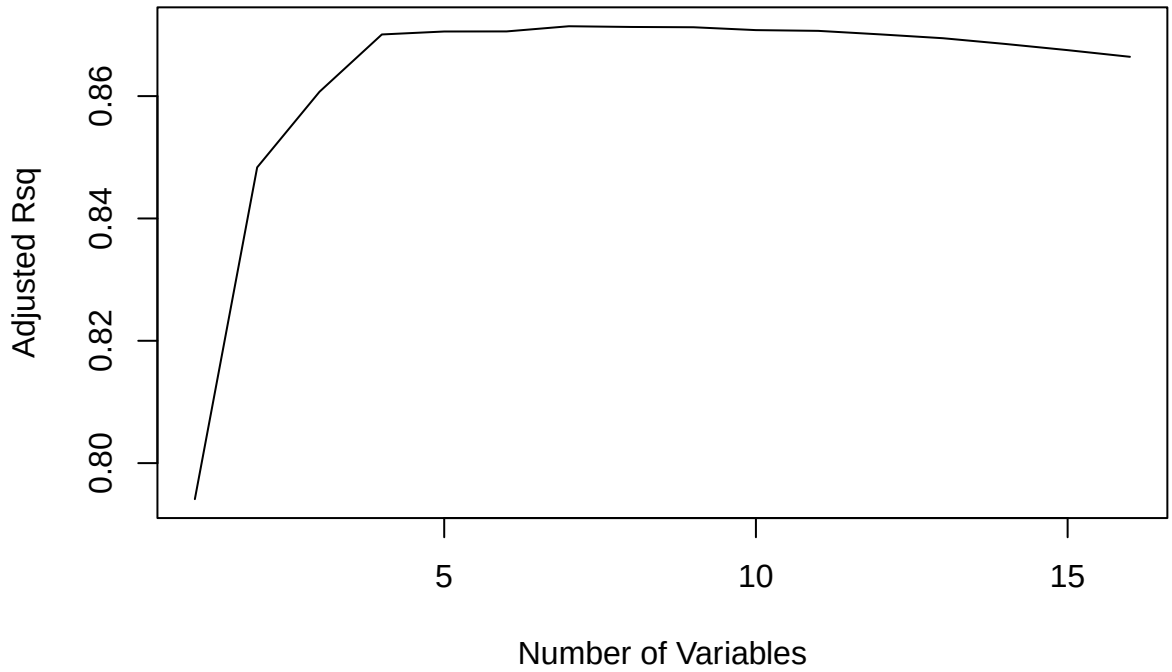
Model selection based on AICc:

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

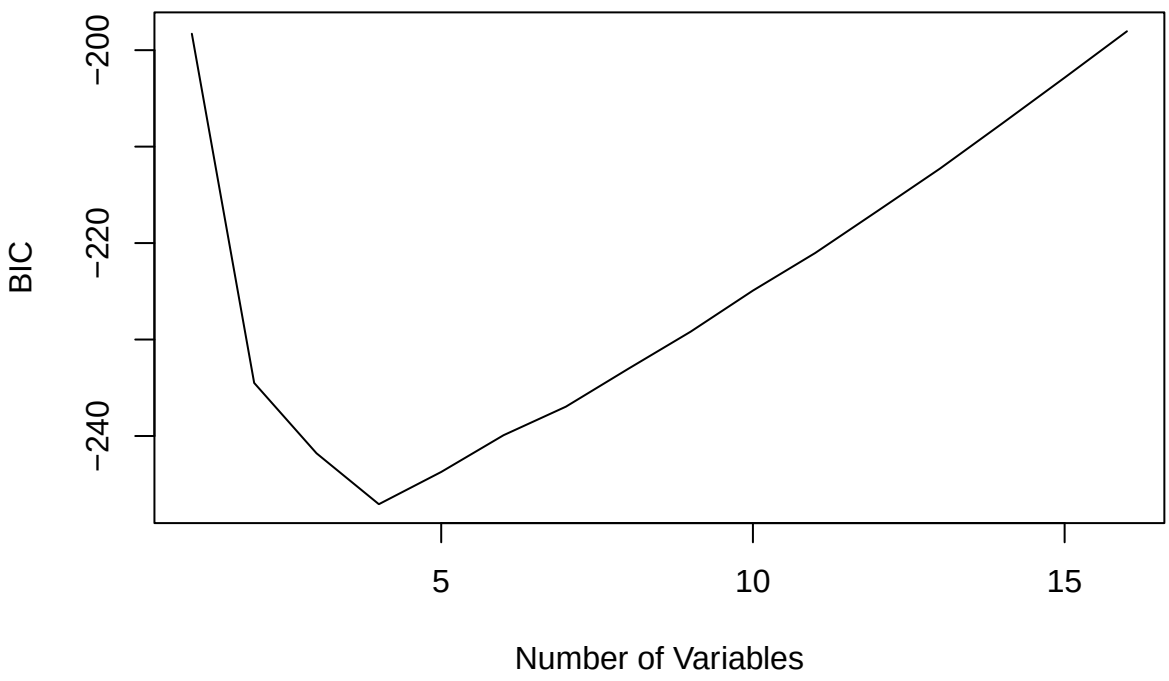
backwardstepwise 8 677.87 0.00 1 1 -330.34 all 21 704.17 26.30 0 1 -326.85 apriori 11 706.28 28.41 0 1 -341.03

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Wed Nov 30 23:04:50 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

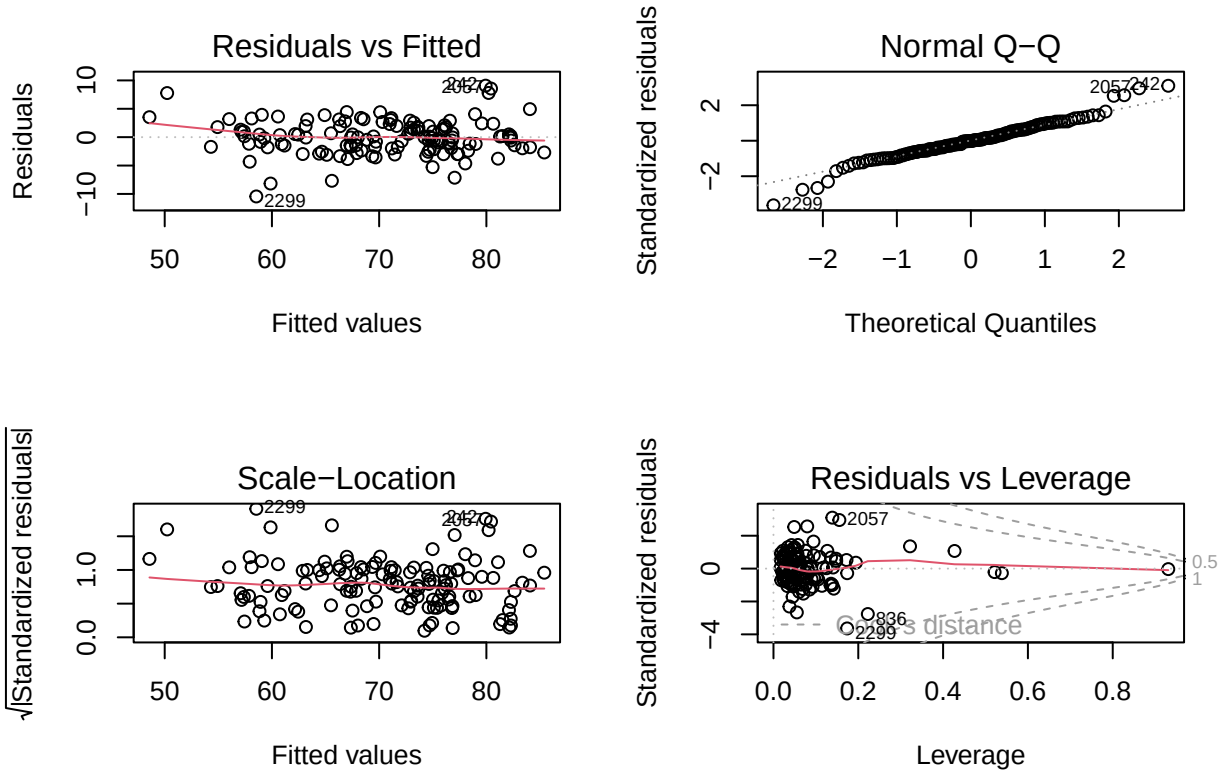


```
## [1] 7
```




```
## [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.



(interprtation for check assumptions) 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) 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 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.

Questions 2: ***"How did disease and mortality rates, along with national economic factors, contribute to a country's development status in 2014?"***

Model 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.

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

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.

Table 3: Variance Inflation Factors

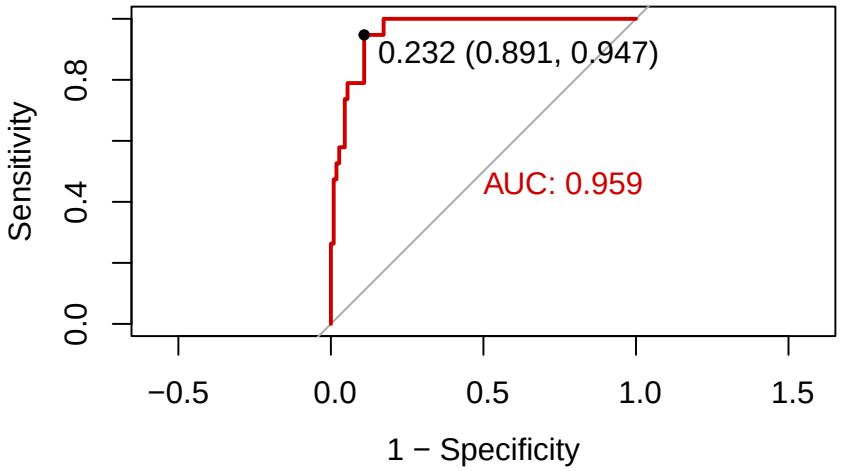
	(1)	(2)	(3)	(4)
Life Expectancy	2.10	2.09	2.13	1.40
Health Expenditure/GDP per capita	10.77			
BMI	1.28	1.28	1.19	1.26
Health Gov. Expenditure Percentage	1.20	1.11	1.22	1.26
HIV/AIDS Deaths/1000 live births	1.00	1.00	1.00	1.00
GDP per capita	10.59	1.54		
Std. Income composition of resources	4.13	4.20	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

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%.

Table 4: Confusion Matrix for Final Model

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

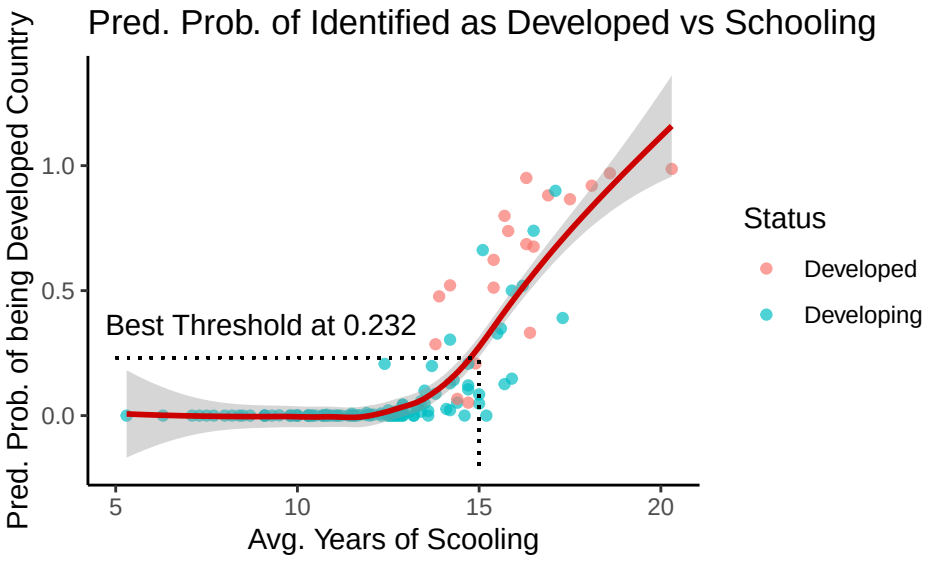
A useful tool for measuring our predictions is the Receiver Operator Characteristic (ROC) curve with the axis as Sensitivity (true positive rate) and 1 - Specificity (true negative rate). We find out the optimal prediction cut-off for year 2014 dataset is 0.232, and we can see that the area under the 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.



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

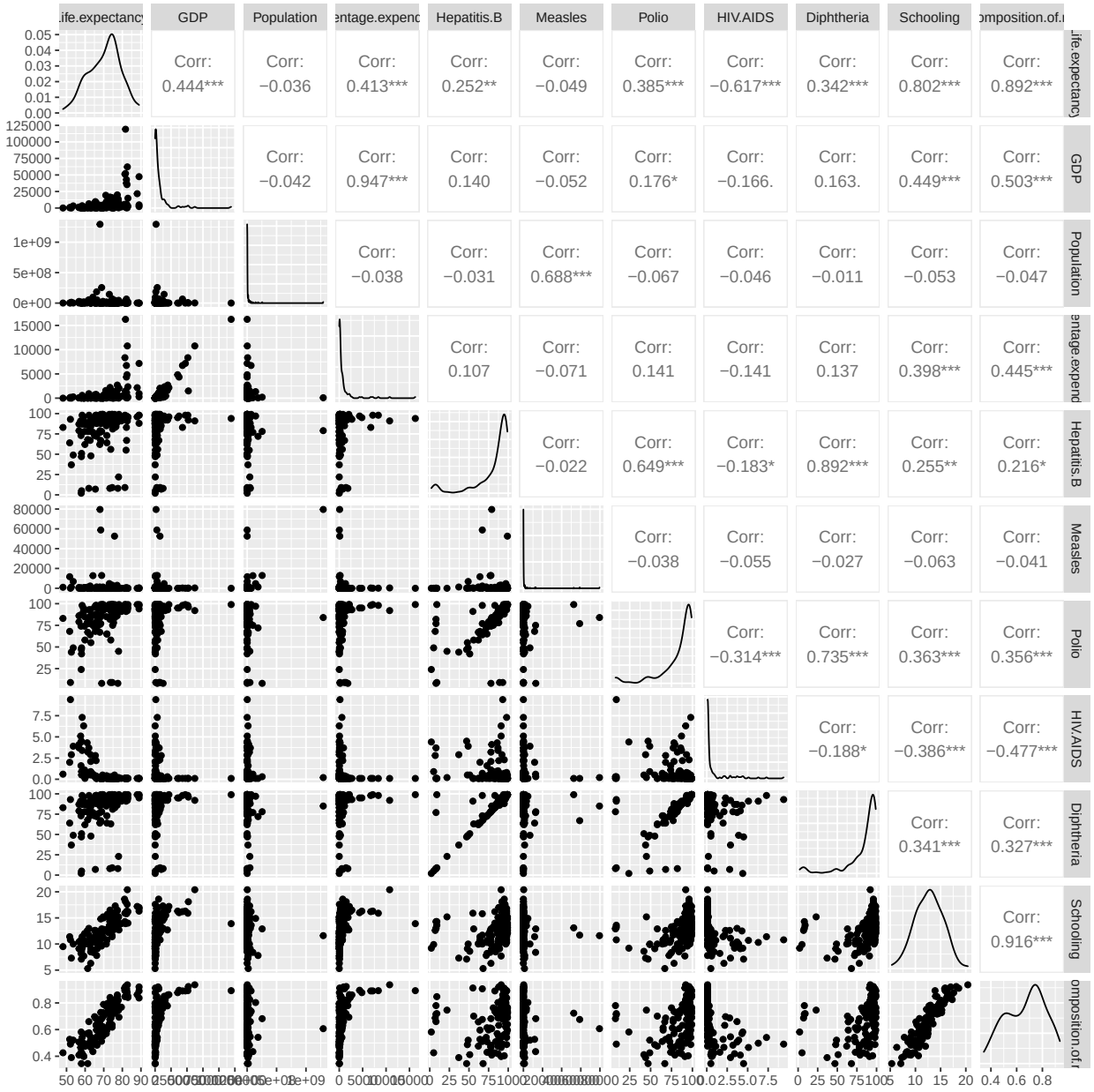
Conclusion

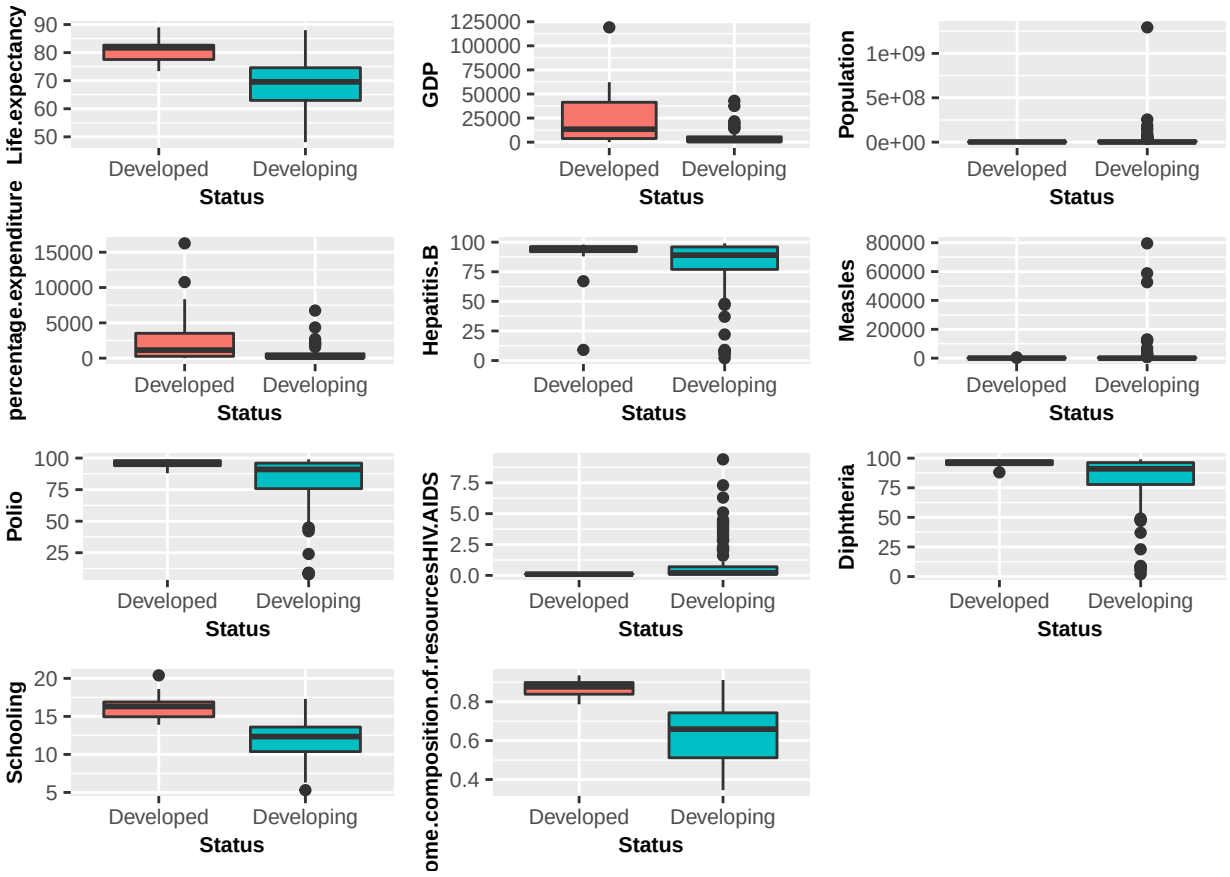
Describe the key takeaways from your analysis, limitations, and future work that can be done to advance knowledge in this area.

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

Appendix

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




Life.expectancy
90
80
70
60
50
Developed
Developing
Status
GDP
125000
100000
75000
50000
25000
0
Developed
Developing
Status
Population
1e+09
5e+08
0e+00
Developed
Developing
Status
percentage.expenditure
15000
10000
5000
0
Developed
Developing
Status
Hepatitis.B
100
75
50
25
0
Developed
Developing
Status
Measles
80000
60000
40000
20000
0
Developed
Developing
Status
Polio
100
75
50
25
Developed
Developing
Status
HIV.AIDS
7.5
5.0
2.5
0.0
Developed
Developing
Status
Diphtheria
100
75
50
25
0
Developed
Developing
Status
Schooling
20
15
10
5
Developed
Developing
Status
ome.composition.of.resources
0.8
0.6
0.4
Developed
Developing
Status

A few things to keep in mind: • You should never refer to actual variable names in the text, tables, or figures. For example, if a variable for height is called "ht___cm," you should always say "height," and the first time you mention it you should state that it is measured in cm. In plots and tables, it should say "height (cm)" • The report should be produced in R Markdown and knit to PDF. This may mean you need to create tables "manually" with knitr. I recommend this anyway because you can customize the labels and formatting. • Someone should be able to read the abstract and look at the tables and figures and have a pretty good idea of 1) the goals of your analysis, and 2) the key results. • I recommend using colorblind-friendly color palettes in your figures. It can be even better to differentiate with line types or symbols instead of relying on color.

Keep you audience in mind! A non-statistician should be able to read your report and have a good idea of what you did. • You can have an appendix if tables or figures are too large to fit into the main text. For example, if you have several predictors, you may want to put a table of model results in the appendix.