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:

Introduction - Mostly Good

Abstract

This analysis is conducted to understand which factors influence life expectancy and the development status of countries around the world. We seek to understand the drivers of both conditions in the framework of gaining better insights into what factors tend to lead to improved living conditions and general human welfare. Life expectancy and country status are two outcomes that we propose are good indicators for quality of like, and we investigate if there are "high-payoff" areas that have an outsized effect. Our intention is to understand these relationships better, and thus inform constructive policy discussions about improving societies at the national level. The dataset used for this research consists of national disease, economic, and social factors and was compiled by the World Health Organization (WHO). The version of the dataset that underpins this analysis was accessed at https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who. 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 human living conditions around the globe. We will approach this by first attempting to form a predictive model for life expectancy, and interpreting the components of that model in the context of what areas have positive effects on the outcome. We also attempt to find inferential value for determining the developmental status of a given country. The WHO has a formula for determining this, which we are unable to completely re-define, so we will attempt to explore this designation from the angle of finding outsized infleuencers that might explain how this designation is arrived upon. 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:

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?"

Methods needs abstract about model fitting added

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. To accomplish this objective, the team began with exploratory data analysis (EDA) to examine the distributions of key variables. From the EDA, missing data points were identified, and decisions made for courses of action to cope with those non-values. We then continued with EDA to examine the distributions of variables and identify any points that may be worthy of further examination. 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.

We then proceed to model fitting. [High-level how we do this.] Finally, we obtain diagnostic information about the performance of our models.

Data

During the course of this initial round of EDA, the team identified a number of missing values from the dataset. These missing values notably 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.Additionally, population values are estimates when expressed in between census windows. Due to the different possible approaches to those estimates, we choose to omit the missing population values and continue with our analysis.

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.

The final step in pre-processing the dataset is to create a new variable, called *Std. Income Composition of Resources* that is a transformation of the original *Income Composition of Resources* variable. *Income Composition of Resources* is a measurement that expresses a form of the country's Human Development Index from what the team is able to determine. We transform this variable from a 0-1 decimal scale by multiplying by 100. This allows the variable to be expressed between 0 and 100, and interpretations of changes in this variable during model assessment can thus be interpreted as 1% point changes.

Models

I removed Model 1 temporarily to make this easier

Question #1: ***"How did major disease, economic, and social factors impact life expectancy around the globe in 2014?"***

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

To address the second research question, an inferential one, we will choose a logistic regression model. We make this choice because of binary, categorical nature of the outcome variable maps appropriately to the assumed distribution that anchors a logistic regression model. We expect an output that will show the impact of predictors on the likelihood of a country's developmental status being "developed," as well as a model that can be used to infer country statuses to a high degree of confidence.

In fitting models, we begin with a model that has a number of variables selected *a priori* which we believe will provide the highest potential value to our inference question. From research, we are able to understand a loose framework for how the WHO is arrinving at its development status designation using a rough combination of socio-economic and public health factors. While, we do not believe all of those variables are captured in this dataset, we will still select variables ahead of model fitting that we believe avoid obvious inclusion in the baseline decision parameter so that we can ideally find other areas a country could use for improving its status in this area. We selected the following variables for our first model fitting:

- Life expectancy
- Health expenditure/GDP per capita
- Body Mass Index (BMI) - population average
- Percentage of Gov. Expenditure on Healthcare

- HIV/AIDS Deaths per 1000 live births
- GDP per capita / Logistic GDP per capita
- Standardized Income Composition of Resources
- Logisitic Population
- Average Years of Schooling

The first model fit to our data included all of the variables above. Prior to fitting, *Population* was transformed logistically to improve the distribution of the observations. We did this because there are some very large gaps in the data between some of the very large and very small countries with an obvious skew towards smaller countries.

After fitting the first model, we investigated a concern with multicolinearity in our variables. This concept is defined by a high amount of correlation between our predictors, which is present with *Health Expenditure/GDP per capita* and *GDP per capita*. We elect to remove *Health Expenditure/GDP per capita* because we are already capturing health expenditure data in our measurements of Percentage of Gov. Expenditure on Healthcare. We refit the model and examined the results once more.

After fitting the second model we elected to utilize a logistic transformation once more on the *GDP per capita* to improve the distribution of that variable, which has a heavy left skew. This was the last transformation that the team felt was necessary, and we considered this third model our final model for answering the second research question.

Model Assessment

[How did we assess the linear model and its assumptions? Plots, four key assumptions, etc.]

[How did we assess the validity of the logistic model?]

Results:

Here you should present results for all aspects of the analysis. The structure of this section should mirror the structure of the methods section. For example, you can start with a few key EDA results (e.g., a table of descriptive statistics), then present model results, then address assessment. This is the section where you will primarily refer to tables and figures. You should have at least 1 figure for each research question that illustrates a key result of the analysis.

[General insights. Were the models effective, set the stage for the discussion below]

Exploratory Data Analysis

[Does Table 1 have everything we want in it???]

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

[Anything else from EDA?? We need to look at the EDA report]

Question 1: *"How did major disease, economic, and social factors impact life expectancy around the globe in 2014?"*

Model Results

Table 4: 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

Assessment [Assess the validity of the outputs]

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 intial full model
2. a model with Health Expenditure/GDP per capita dropped for multicolinearity concerns
3. the final model fit after all logistic transformations
4. an additional model the team fit using only one predictor: Standardized Income Composition of Resources as a trial

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: Standardized Income Composition of Resources. 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 percentage point increase in *Std. Income Composition of Resources* will increase the odds of that country being identified as developed country by about 1.55 times ($e^{0.44}$).

The fitted logistic regression with *Life Expectancy*,*Health Expenditure/GDP per capita*, *BMI*, *Health Gov. Expenditure Percentage*, *HIV/AIDS Deaths/1000 live births*,*Log of GDP per capita*, *Log of Population*, *Std. Income Composition of Resources*, *Years of Schooling* as predictors for the dataset is:

$$\begin{aligned} ln(\frac{\widehat{Developed}}{Developing}) = &-8.88 - 0.14\ Life\ Expectancy - 0.03\ BMI + 0.27\ Health\ Gov.\ Expenditure\ Percentage \\ &- 145.58\ HIV/AIDS\ Deaths\ per\ 1000\ live\ births - 0.35\ ln(GDP\ per\ capita) \\ &+ 0.44\ Std.\ Income\ Composition\ of\ Resources - 0.21\ ln(Population) \\ &+ 0.20\ Years\ of\ Schooling \end{aligned}$$

We also examine multicolinearity concerns for the first three model fits. From the table below it is clearly observable that the first model has signficant conerns, 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 5: Logistic Regression Models

	Dependent variable:			
	Development Status			
	(1)	(2)	(3)	(4)
Life Expectancy	−0.12 (0.13)	−0.13 (0.13)	−0.14 (0.14)	
	p = 0.36	p = 0.33	p = 0.33	
Health Expenditure/GDP per capita	−0.0002 (0.001)			
	p = 0.71			
BMI	−0.03 (0.03)	−0.03 (0.03)	−0.03 (0.03)	
	p = 0.24	p = 0.25	p = 0.26	
Gov. Expenditure on Healthcare	0.24 (0.17)	0.22 (0.16)	0.27 (0.18)	
	p = 0.17	p = 0.18	p = 0.13	
HIV/AIDS Deaths/1000 live births	−142.51 (23,806.80)	−142.39 (23,786.91)	−145.58 (24,223.03)	
	p = 1.00	p = 1.00	p = 1.00	
GDP per capita	0.00001 (0.0001)	−0.00002 (0.00002)		
	p = 0.91	p = 0.50		
Log of GDP per capita			−0.35 (0.25)	
			p = 0.17	
Std. Income Composition of Resources	0.45 (0.16)	0.44 (0.16)	0.44 (0.15)	0.32 (0.07)
	p = 0.01***	p = 0.01***	p = 0.004***	p = 0.00001***
Log of Population	−0.19 (0.20)	−0.20 (0.19)	−0.21 (0.20)	
	p = 0.34	p = 0.31	p = 0.30	
Years of Schooling	0.04 (0.42)	0.07 (0.42)	0.20 (0.46)	
	p = 0.94	p = 0.87	p = 0.68	
Constant	−10.84 (2,380.69)	−10.37 (2,378.71)	−8.88 (2,422.32)	−26.77 (5.85)
	p = 1.00	p = 1.00	p = 1.00	p = 0.000005***
Observations	131	131	131	131
Log Likelihood	−19.70	−19.78	−18.99	−22.59
Akaike Inf. Crit.	59.41	57.56	55.99	49.17

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

Table 6: Variance Inflation Factors

	(1)	(2)	(3)
Life Expectancy	2.57	2.55	2.58
Health Expenditure/GDP per capita	14.75		
BMI	1.45	1.43	1.28
Health Gov. Expenditure Percentage	1.21	1.13	1.27
HIV/AIDS Deaths/1000 live births	1.00	1.00	1.00
GDP per capita	14.37	1.66	
Std. Income composition of resources	4.70	4.78	3.97
Log of Population	1.78	1.76	1.77
Years of Schooling	2.24	2.20	2.28
Log of GDP per capita			1.48

Assessment To assess the accuracy of the final model fit for this research question we use the model 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 use the threshold of 0.5 to classify countries as developed or developing. The accuracy of this model is approximately 94%.

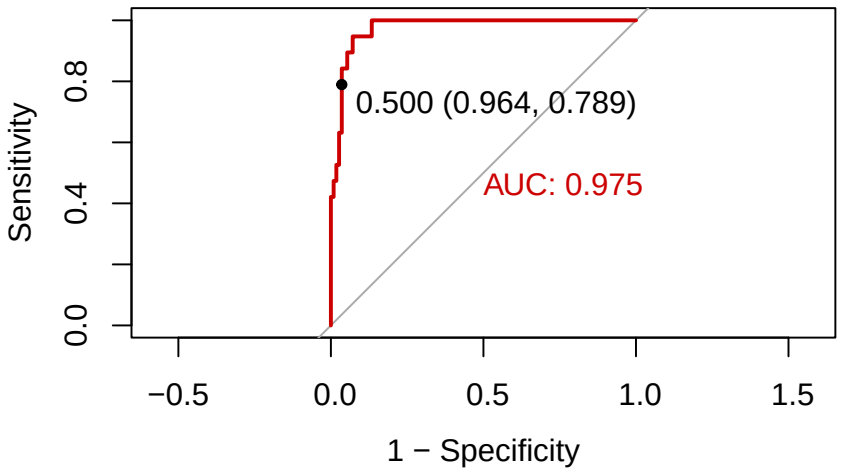
Table 7: Confusion Matrix for Final Model

	True Developed	True Developing
Predicted Developed	15	4
Predicted Developing	4	108

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + FN + FP + TN} \\ &= \frac{15 + 108}{15 + 4 + 4 + 108} \\ &= 0.94 \end{aligned}$$

$$95\%\ Confidence\ Interval\ of\ Accuracy : (0.88, 0.97)$$

A useful tool for measuring our predictions is the Receiver Operator Characteristic (ROC) curve with the axiis as Sensitivity (true positive rate) and 1 - Specificity (true negative rate). We we set the prediction cut-off at 0.5, we can see that the area under the curve is 0.975, a very strong number. This particular curve is very close to a right angle, and does prompt questions about whether or not a prediction value cut-off of 0.5 is indeed the best value. We leave the value as described because the small number of countries desginated as developing in our dataset means that each missed prediction carries greater weight in this group. We observed 4 false positives and 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 0.5 as the cutoff for inferring developed or developing status. The result of this experiment is shown in the table below. The accuracy of these predictions is 0.92, which is a slight drop-off in accuracy, but still considered a good fit.

Table 8: Confusion Matrix for Infering Year 2013 Data

	True Developed	True Developing
Predicted Developed	13	4
Predicted Developing	6	107

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + FN + FP + TN} \\ &= \frac{13 + 107}{13 + 4 + 6 + 107} \\ &= 0.92 \end{aligned}$$

$$95\%\ Confidence\ Interval\ of\ Accuracy : (0.8631, 0.9625)$$

Final Model Evaluation Out of curiosity, we compare the final model with a model with only one predictor variable, *Standardized Income Composition of Resources*. 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 0.41 is greater than 0.05, meaning *Std. Income Composition of Resources* variable by its own is a great indicator of whether the country should be considered as developed or developing country.

Table 9: Analysis of Deviance: Final Model vs Model w/ One Predictor Variable

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	122	37.99			
2	129	45.17	-7	-7.18	0.41

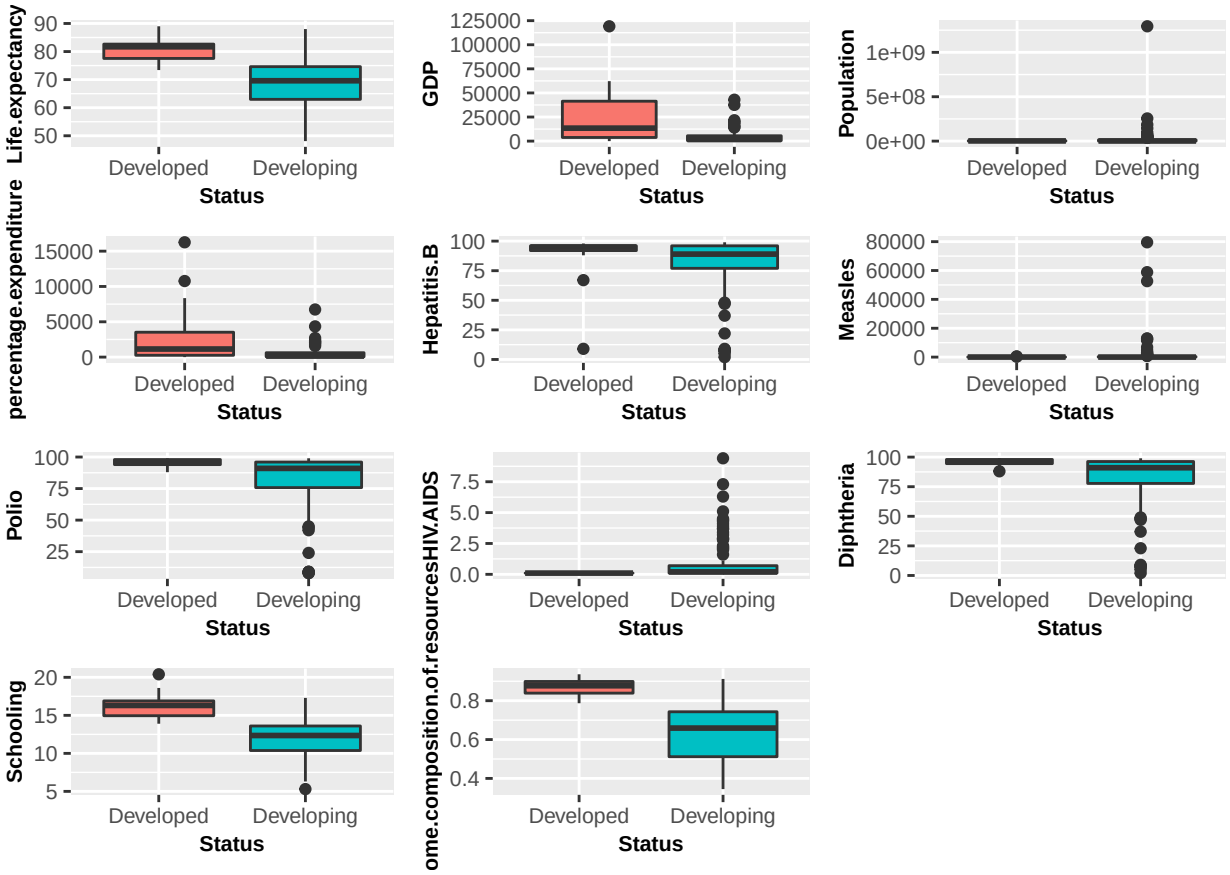
Conclusion

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]



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.