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:

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:

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

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. 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

[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.

After some simple steps to prepare the data, the team began its EDA. [EDA plots and findings - need to truncate this and decide what's actually important to what we are trying to say]

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

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

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

Models

[The model used for the first research question, the prediction question, is linear regression]

[Describe how we fit the model and the process]

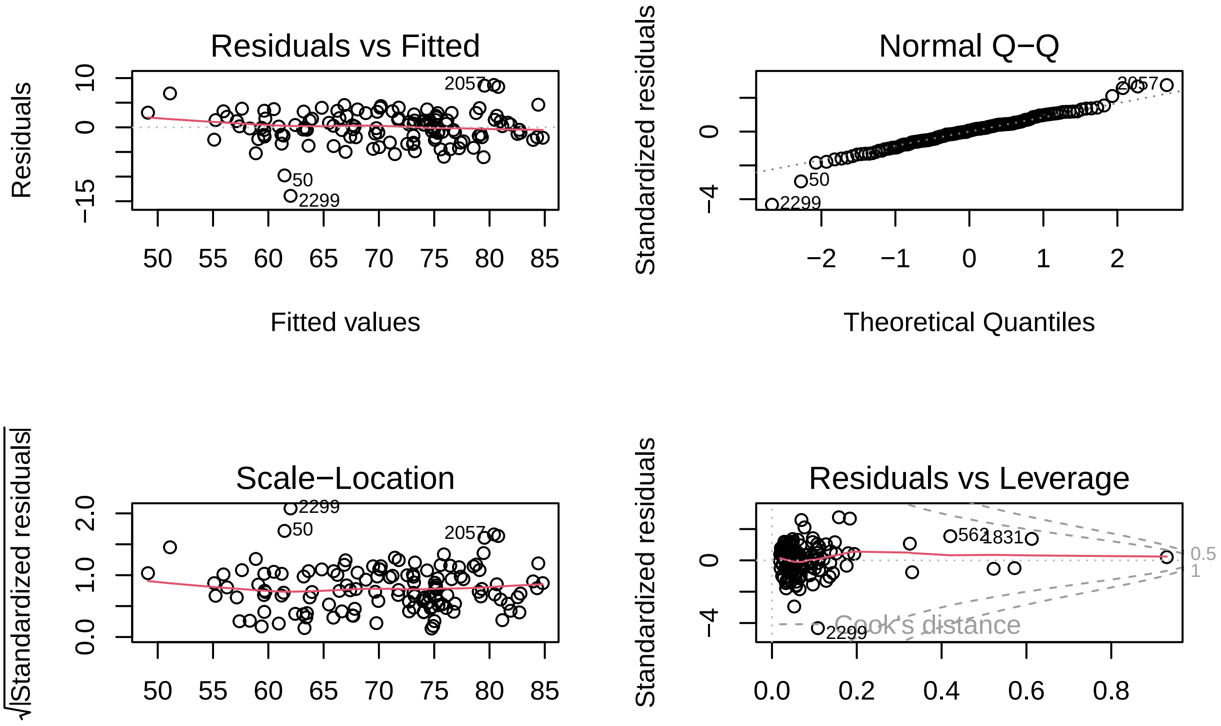
```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on the
## right-hand side and was dropped
```

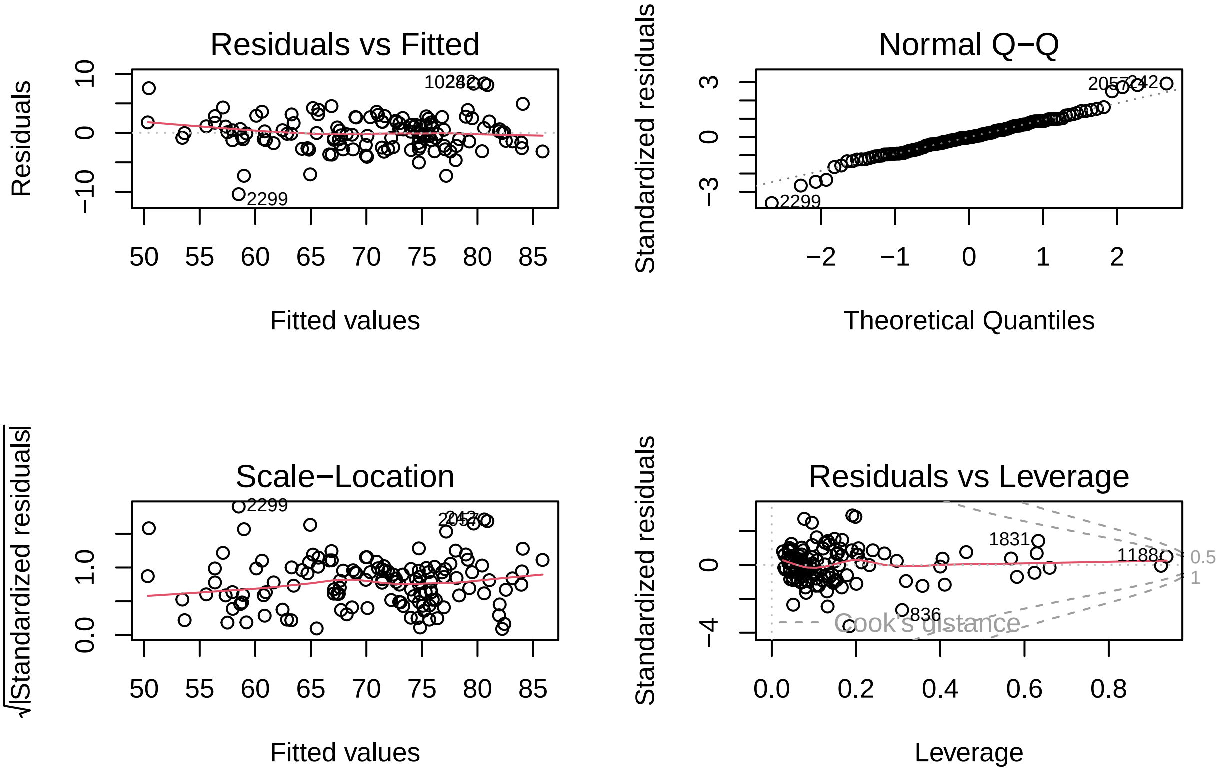
```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 3 in
## model.matrix: no columns are assigned
```

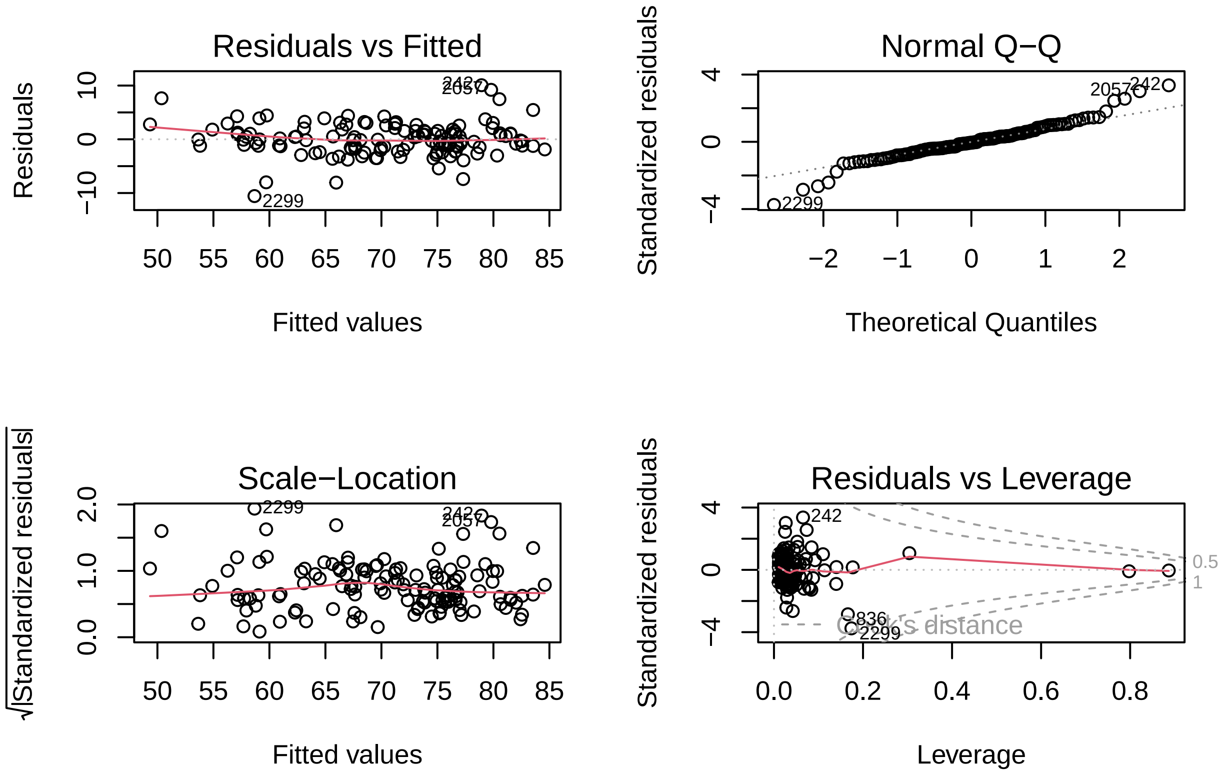
```
##
## Call:
## lm(formula = Life.expectancy ~ Status + Population + Life.expectancy +
##     percentage.expenditure + Measles + Polio + HIV.AIDS + GDP +
##     Schooling + Income.composition.of.resources + BMI + Total.expenditure,
##     data = df_life_expectancy_2014)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.910  -1.885   0.186   1.848   8.598
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.262e+01  2.751e+00  15.492  < 2e-16 ***
## StatusDeveloping                -9.977e-01  1.070e+00  -0.932   0.3531
## Population                       1.394e-09  3.545e-09   0.393   0.6949
## percentage.expenditure           5.012e-04  4.793e-04   1.046   0.2978
## Measles                         -2.977e-05  4.247e-05  -0.701   0.4846
## Polio                            1.008e-02  1.605e-02   0.628   0.5310
## HIV.AIDS                        -1.365e+00  2.267e-01  -6.020 1.99e-08 ***
## GDP                             -6.893e-05  6.930e-05  -0.995   0.3219
## Schooling                       -2.175e-01  2.811e-01  -0.774   0.4407
## Income.composition.of.resources  4.547e+01  5.906e+00   7.699 4.47e-12 ***
## BMI                             -5.827e-03  1.941e-02  -0.300   0.7645
## Total.expenditure                2.722e-01  1.329e-01   2.047   0.0428 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.404 on 119 degrees of freedom
## Multiple R-squared:  0.8567, Adjusted R-squared:  0.8435
## F-statistic: 64.69 on 11 and 119 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = subset(df_life_expectancy_cc_2014,
##     select = -c(Country, Year)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4098  -1.7264  -0.0392   1.7715   8.3880
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.122e+01  3.314e+00  15.457  < 2e-16 ***
## StatusDeveloping                -1.170e+00  1.035e+00  -1.130 0.261006
## Adult.Mortality                 -1.724e-02  4.148e-03  -4.157 6.36e-05 ***
## infant.deaths                    8.287e-02  5.619e-02   1.475 0.143057
## Alcohol                          5.674e-03  9.749e-02   0.058 0.953689
## percentage.expenditure           4.627e-04  4.639e-04   0.997 0.320716
## Hepatitis.B                      1.205e-02  2.808e-02   0.429 0.668582
## Measles                         -3.361e-05  4.823e-05  -0.697 0.487345
## BMI                             -7.576e-03  2.000e-02  -0.379 0.705531
## under.five.deaths               -6.014e-02  3.838e-02  -1.567 0.119989
## Polio                           -8.746e-03  2.117e-02  -0.413 0.680327
## Total.expenditure                2.878e-01  1.274e-01   2.259 0.025833 *
## Diphtheria                       7.644e-03  3.445e-02   0.222 0.824805
## HIV.AIDS                        -8.363e-01  2.470e-01  -3.385 0.000984 ***
## GDP                             -5.980e-05  6.656e-05  -0.898 0.370911
## Population                      -1.729e-09  6.804e-09  -0.254 0.799816
## thinness..1.19.years            -1.300e-01  2.267e-01  -0.574 0.567462
## thinness.5.9.years               5.458e-03  2.227e-01   0.025 0.980489
## Income.composition.of.resources  3.597e+01  6.228e+00   5.775 7.11e-08 ***
## Schooling                       -1.617e-01  2.740e-01  -0.590 0.556279
## Status_num                              NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.186 on 111 degrees of freedom
## Multiple R-squared:  0.8829, Adjusted R-squared:  0.8629
## F-statistic: 44.06 on 19 and 111 DF,  p-value: < 2.2e-16
```

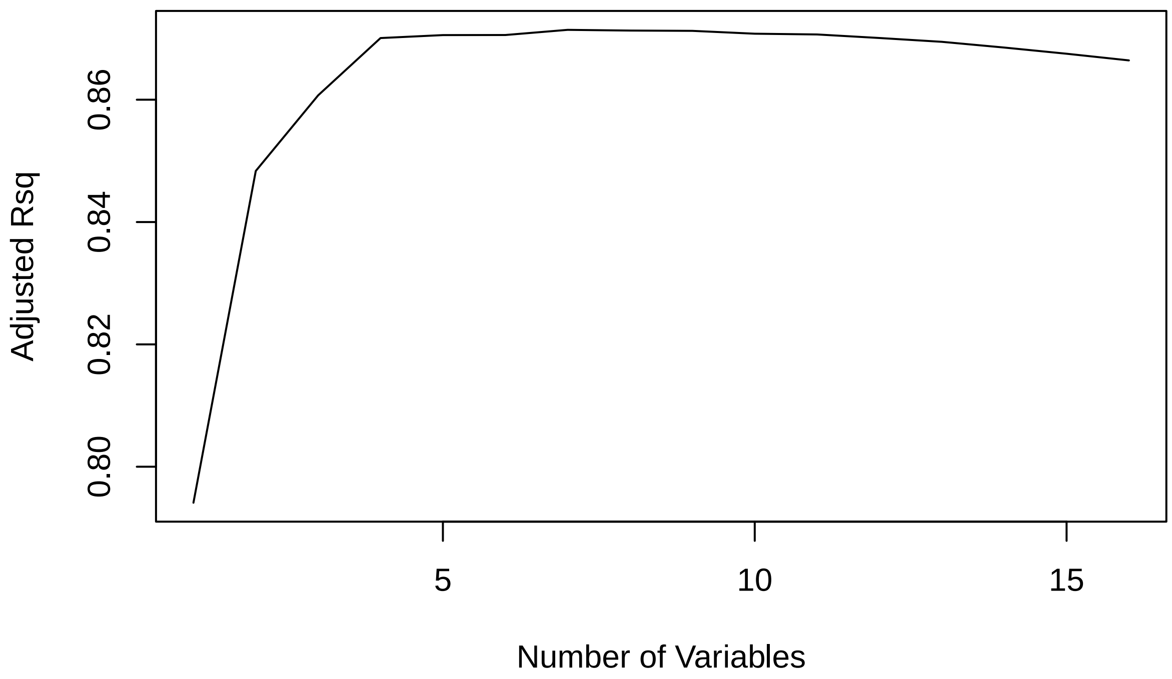
```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
##     under.five.deaths + Total.expenditure + HIV.AIDS + Income.composition.of.resources,
##     data = subset(df_life_expectancy_cc_2014, select = -c(Country,
##         Year)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.568  -1.569  -0.127   1.561  10.060
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     47.860575   2.066580  23.159  < 2e-16 ***
## Adult.Mortality                 -0.017405   0.003866  -4.502 1.53e-05 ***
## infant.deaths                    0.042287   0.029527   1.432 0.154625
## under.five.deaths               -0.033561   0.022614  -1.484 0.140331
## Total.expenditure                0.349095   0.111930   3.119 0.002258 **
## HIV.AIDS                        -0.809261   0.231621  -3.494 0.000661 ***
## Income.composition.of.resources 35.911609   2.502726  14.349  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.096 on 124 degrees of freedom
## Multiple R-squared:  0.8765, Adjusted R-squared:  0.8705
## F-statistic: 146.7 on 6 and 124 DF,  p-value: < 2.2e-16

## Warning in model.matrix.default(object, data = structure(list(Life.expectancy =
## c(59.9, : the response appeared on the right-hand side and was dropped

## Warning in model.matrix.default(object, data = structure(list(Life.expectancy =
## c(59.9, : problem with term 3 in model.matrix: no columns are assigned
```



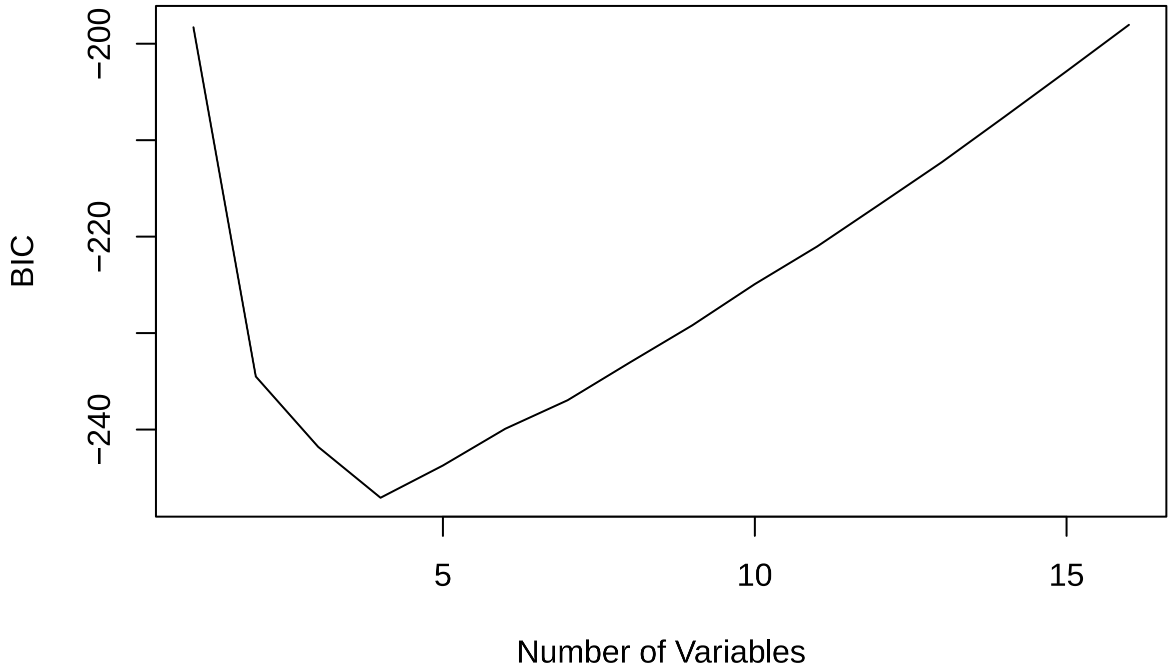

Residuals vs Fitted
Residuals
10
0
−10
1028242
2299
50 55 60 65 70 75 80 85
Fitted values
Normal Q−Q
Standardized residuals
3
0
−3
2057242
2299
−2 −1 0 1 2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
0.0
1.0
2299
2042
50 55 60 65 70 75 80 85
Fitted values
Residuals vs Leverage
Standardized residuals
0
−4
1831
1188
836
0.5
1
Cook's distance
0.0 0.2 0.4 0.6 0.8
Leverage



```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 1 linear dependencies found
```

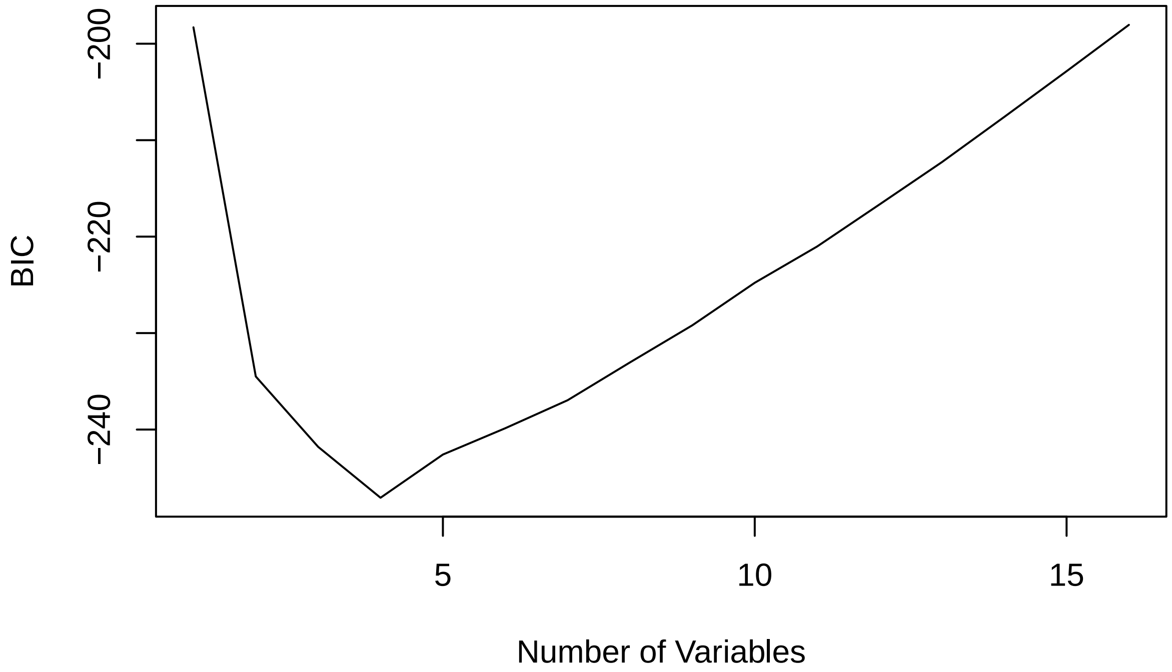


```
## [1] 7
```



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

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 1 linear dependencies found
```



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

[The second question, the inferential one, will be approached with a logistic regression model]

[Dingkun, what is all of this]

[Multicolinearity concerns]

[Following are the ones updated]

EDA & Pre-possing:

When conducting Exploratory Data Analysis (EDA) for WHO BALABALALALA dataset, there are 2938 observations with 22 variables in total ; [WE MAY COPY FROM OUR PART 1]

Due to missingness of some variables in the data, we subset our data from 183 to 131 observations (IF WE DO NOT MERGE POPULATION DATA FROM OTHER SOURCE)

FOR BETTER INTERPERTATION, *Std. Income Composition of Resources* (IT RANGES FROM 0 TO 1), After transformation, instead of saying one unit, we say 1 percent

TABLE NUMBER MAY CHANGE (DUE TO QUESTION 1 PART), TAKING THAT IN MIND WHEN REFERENCING

(num): meaning each respective model, and why we did that

(1) Since the population gaps may be large, , we use log transformation on the predictor variable, *Population.*

(2) Since VIF for Model 1 shows that *Health Expenditure/GDP per capita* and *GDP per capita* are the ones have high VIF being over 10, WE DECIDE TO DELETE *Health Expenditure/GDP per capita* by taking another variable *Health Government Expenditure Percentage* into consideration, since the later carry some portion of the same information as the former one.

(3) For better interpretation we use log transformation on *GDP per capita* (THIS IS OUR FINAL MODEL)

(4) Just out of curiosity (SURPRISE)

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:

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

As Table XXXX: Logistic Regression Models above shown, out of 8 predictor variables, only one predictor variable has p-values less than 0.05. The interpretation of this predictor's coefficient in terms of the odds of a country being identified or labeled as developed country is that while holding all other predictor variables constant, one percent increase in *Std. Income Composition of Resources* (prepossessed by timing 100), the odds of that country being identified as developed country increases about 1.55 times ($e^{0.44}$).

Table 2: 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	
Health Gov. Expenditure Percentage	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	0.0000 (0.0001)	−0.0000 (0.0000)		
	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 = 0.0000***
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 = 0.0000***
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 3: 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

Confusion Matrix for our final model

When calculate the accuracy of the model, we use confusion matrix (as Table XXXX shown below) to calculate True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) count by comparing predicted values and actual values. For prediction, we use the threshold of 0.5 to classify games as Win or Loss. As you can see, the accuracy of this model is 0.94.

Table 4: Confusion Matrix for Final Model

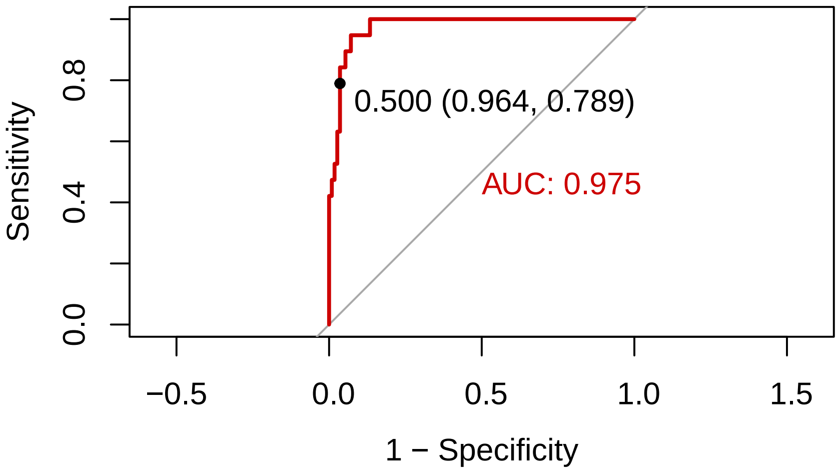
	True Developed	True Developing
Predicted Developed	15	4
Predicted Developing	4	108

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

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

ROC Curvey for our final model

Receiver Operator Characteristic (ROC) curve (Sensitivity vs 1 - Specificity) is shown below, and the Area Under the Curve (AUC) is 0.975, when we set the threshold at 0.5.



Using Model to predict out-of-sample probabilities

The confusion matrix of using Final Model to predict out-of-sample probabilities for Year 2013 dataset using 0.5 as the cutoff for inferring developed or developing countries is shown below as Table XXXX. The accuracy is 0.92, which shall be considered as a decent model.

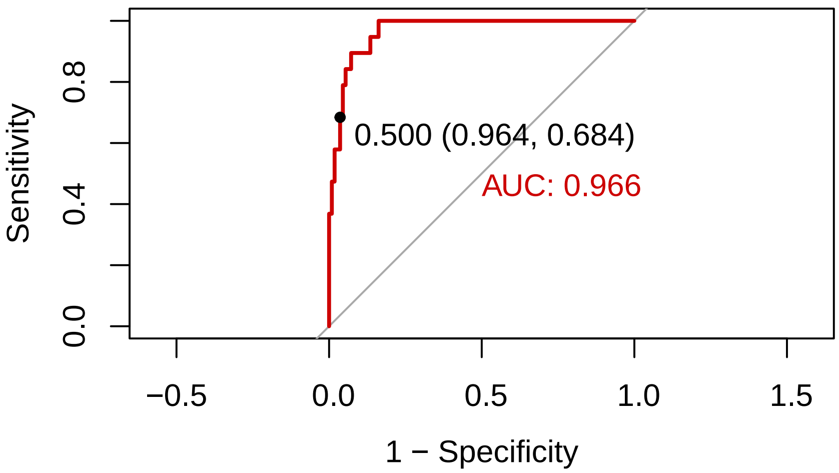
Table 5: Confusion Matrix for Infering Year 2013 Data

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

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

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

Receiver Operator Characteristic (ROC) curve (Sensitivity vs 1 - Specificity) is shown below, and the Area Under the Curve (AUC) is 0.966. The fact that AUC is so close to 1 confirms that this model does a good job identifying country development status for another year, 2013.



SURPRISE MOLDE

Out of curiosity, we compare the final model with the model with only one predictor variable, *Std. Income Composition of Resources*, the ANOVA result surprisingly shows that two models are not statistically different in terms of “inference” accuracy.

As we can see on Table XXXX: 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.

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?]

Table 6: 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

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]

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

Model Results [What does the model output tell us]

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 [What does the model output tell us]

Assessment [What kid of accuracy did we find from the model]

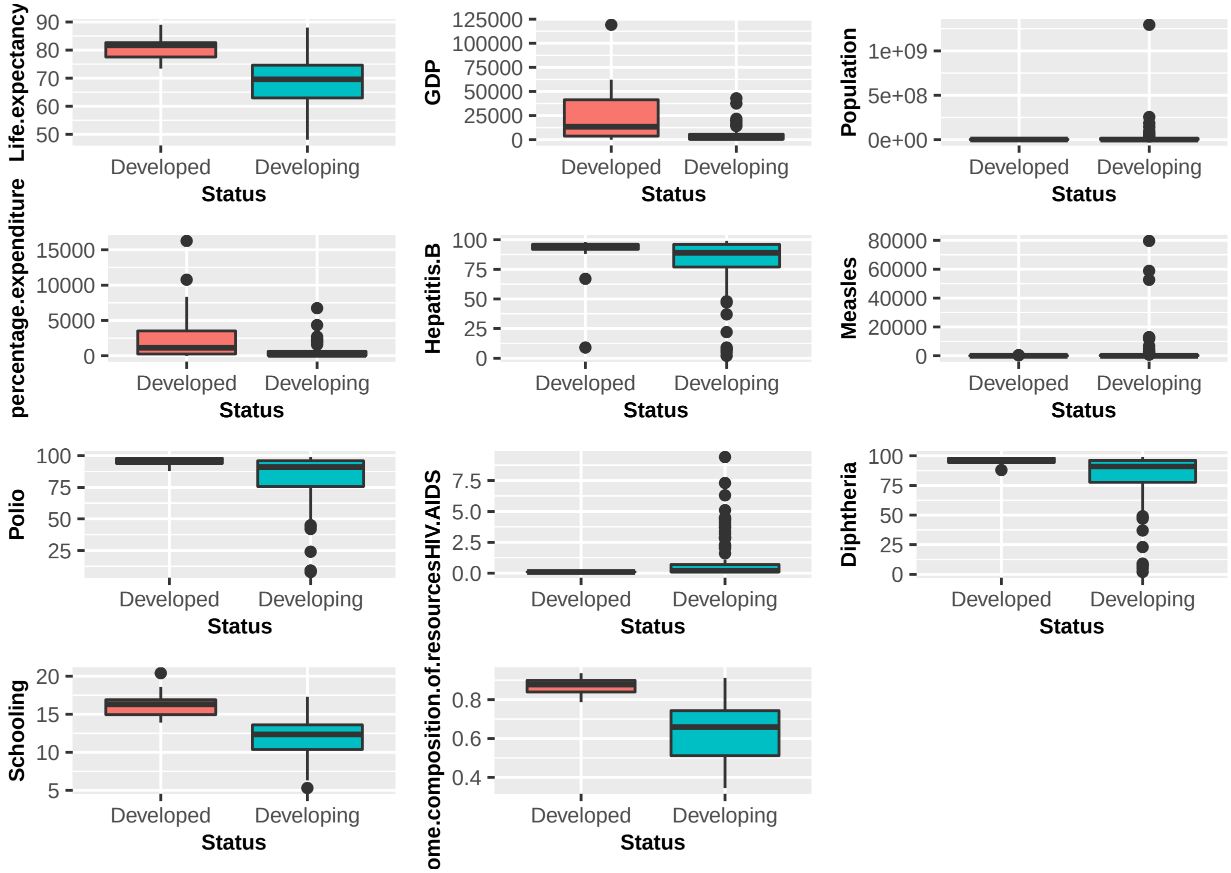
Conclusion

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.