Analysis of Factors that Influence Life Expectancy

Kevin Diaz Gochez

Introduction and Background

The current pandemic has forced the leaders of our nations to act, and the decisions we make now will be scrutinized by our descendants as they reverberate for years to come. During this time, many have been torn away from their own families and have had to endure the experience of losing loved ones. Others fear for their own lives as they struggle to remain afloat. The last couple of months have clarified what truly matters, living a long and healthy life.

Life expectancy is an average measure given to represent the expected number of years of life for a person depending on a variety of factors. As technology improves people have access to more information and governments can promote better treatment for the poorest of their communities. Some developing countries do not have the facilities to provide a basic level of healthcare and may be unable to adequately distribute the immunizations needed for their people. Within developing areas politics dictates who gets what resources and how the government moves forward to guarantee the health of their people. Lastly, rampant diseases have also contributed to lowering life expectancy and force officials to decide who receives treatments and vaccines.

Research Question

We intend to determine which variables will influence life expectancy age?

Motivation

Throughout the last year, health has become a larger concern around the world. Therefore, understanding which factors can enable a healthier life can help policy makers improve their allocation of resources. Although we are working with a limited dataset that cannot encapsulate all of the factors that truly affect our dependent variable, we aim to prove that there are a couple of prominent variables that can drastically improve life expectancy age and are worth looking into by government officials.

Dataset Description

Our team intends to analyze a dataset published by the Global Health Observatory (GHO) data repository under the World Health Organization. The GHO data repository is a streamlined source regarding global health and “Research and development activities”. Based on the Kaggle description (Rajarshi, 2018), the information within this database is publicized so that policy makers and researchers can ensure that public health needs are met and to secure the best allocation of their nation’s resources. The dataset was published onto Kaggle under the name “Statistical Analysis on factors influencing Life Expectancy.” According to the Kaggle webpage, the World Health Organization has observed that in the last fifteen years there has been a noticeable improvement in the predicted life expectancy within several countries over time. This dataset was generated with the intent of realizing which critical factors were the most influential in this development.

The dataset includes two types of variables, quantitative and qualitative regarding wellness, economic circumstances, and immunization records. Although the data set contains some missing values, the Kaggle page notes that most of the missing data are for the following predictors: population, Hepatitis B and GDP and from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc.

The dataset includes information for the World Health Organization's predicted life expectancy for 193 countries between the years of 2000-2015. Out of the 2939 observations, and 21 factors in our dataset, we decided to only use 18 predicting variables. The variable Life Expectancy is not included in our predicting variables since it will serve as the response. We decided to exclude the variable year from our model because a direct correlation between the current year and the life expectancy of a country does not exist. Our number of 18 also includes the variable 'Continent' which we added onto the dataset ourselves and serves as an alternative from using the Country variable moving forward. This was done using python where I created a mapping for each country and its respective continent based off classifications from 'Britannica.com.' Descriptions for all 18 of our variables can be found on Figure 16.

Data Summary Statistics

Figure 1 displays a table summarizing some important attributes for the dataset we used in our study. Notably, the dataset originally contained 2939 observations but after omitting any observations with missing values, we can see that the new dataset contains 1649 observations. The dependent variable we are highlighting of life expectancy demonstrates a large contrast between its minimum (44) and maximum (89) age values. Furthermore, we can also see that the standard deviation for life expectancy across countries is sizable. Our study aims to explain which factors are contributing towards the discrepancy in life expectancy expressed by figure 1.

In figure 2 and figure 3 we have two boxplots used in the exploratory analysis stage to gain an understanding of the two qualitative predictors not represented in Figure 1. Figure 2 displays values for life expectancy against status, whereas Figure 3 has Life Expectancy plotted against Continents. Figure 2 shows that generally we can expect a higher life expectancy from developed countries as opposed to developing ones. We came to this conclusion since the median value for life expectancy is much higher for developed compared to that of developing and there is a significant difference in the minimum expected life expectancy value between the two. On the other hand, from examining Figure 3 we can see that North America, South America, and Europe have the highest median value whereas Africa has the lowest median value. Therefore, we concluded that these continents generally benefit from a longer life expectancy.

Data Mining Method Description

Based off our results from the exploratory analysis stage, we are expecting the life expectancy age values to differ between the developing and developed countries as well as across continents. Therefore, we built two multi-linear regression models, the first model containing qualitative variable, status and the second model containing qualitative variable. Each model began with all predictors, then we utilized the Variance Inflation Factor to iteratively remove variables that displayed high risk of collinearity. Then we followed the backwards selection method to continue to remove predictors based on level of significance until we reached a model where all predictors were statistically significant at the 0.1 level of confidence. Afterwards, we decided to perform cross validation to get an estimate of test-set prediction error for both models and compare between the two which results in a more accurate prediction. We then used the model with the lower Mean Squared Error to build a regression tree, where we ran cross validation again to determine if we had to prune the tree.

Methods

We began with our two multi linear regression models, which both contained all the 18 predictors outlined in the dataset description portion. Starting off with the model containing status, we used the Variance Inflation Factor to find instances of multicollinearity. Two predictors stood out with high values of VIF, infant deaths (209.324074) and under five deaths (200.635530). Infant deaths were removed for having the higher VIF value then we ran it again and found that both GDP (13.563404) and percentage expenditure (12.849259) both displayed high VIF values. Following the removal of GDP, we now had a resulting model where all predictors were significant for at least at the 0.1 level of significance. The process was repeated for the model containing Continent but upon removing Infant deaths and GDP for their VIF values, we had a model where not all predictors were significant. To improve the model we used the backwards selection method where we removed total expenditure for having the highest P-value. The process was repeated for the removal of Hepatitis B and Population.

Afterwards, we began addressing the common problems with linear regression for both models, starting off with the "Non-Linearity of the response-predictor relationship assumption" for the model containing status. For this problem we plot the residuals against the predicted values of y in figure 4. Figure 4 shows some evidence of a discernible pattern but not enough to conclude a violation of the assumption. The second problem that we address is the Non-constant variance of error terms, which we can conclude does not show pattern of heteroscedasticity and does not violate assumption of equal variance from examining figure 5. Figure 6 shows that all outliers are within cook's distance and that the model is acceptable in terms of leverage points. Lastly, figure 7 shows that the residuals are relatively normally distributed by not deviating severely from the straight-line pattern.

Moreover, for the model containing Continents we can see from Figure 8 that there is not a violation of the linearity assumption when we plot the residuals against the predicted values of y. From Figure 9 we do not see any evidence of heteroscedasticity. Figure 10 shows that the residuals are once again relatively normally distributed and do not deviate severely from the straight line pattern. Lastly, Figure 11 does not show alarming evidence of residuals outside of cook's distance or high leverage points.

Figure 12 serves to summarize the output from both models, on the left hand side we have the model containing status whereas on the right we have the model containing continent. Notably, we can see that Status Developing variable has a coefficient of -0.991 indicating that developing countries will have a lower life expectancy than developed countries. On the right hand side, we can see that Europe, Asia, Oceania, North America, and South America have a higher life expectancy when each compared to Africa. Among all the continents, North America is expected to have the highest life expectancy since it holds the largest coefficient. For both models, Income composition of resources hold the highest coefficient amongst all predictors; therefore it is the predictor that has the largest positive effect on life expectancy. Predictors like under five deaths, alcohol, HIV/AIDS, Adult Mortality, Hepatitis B have a negative effect on life expectancy in both models according to figure 12. On the other hand, BMI, Schooling, Percentage Expenditure, Measles, Polio, Diphtheria have a positive effect on life expectancy across both models. Total expenditure, Hepatitis B, and Population are not significant in the continent model. Lastly, when comparing the R-Squared values for the two models, we can see that the regression model containing continents can explain the variability in life expectancy better than the model containing status.

We ran cross validation to decide which of the two models would give us the most accurate prediction based off the MSE values. We took the Validation Set approach which means that we split our dataset in half without replacement and set one half to represent the training data set. We ran linear regression for both models based on the training dataset, then we calculated the Mean Squared Error values for both based on the validation sample. Our hopes in cross validation were to be able to choose the better model out of the two by comparing their Mean Squared Error values. The model containing status had an MSE value of 13.54407 and the model containing Continent had an MSE value of 12.3996. Based off the MSE values, we chose the continent model since it had the lower MSE.

Aside from multilinear regression we decided to run an alternative model, regression tree to see if we get a better model and compare the results using the MSE values. We began with the validation set approach, splitting our observations in half. We then fit the model and took a look at the residual mean deviance which in the case of regression is simply the average sum of squared errors for the tree and it tells us how well the model fits the data. The value for residual mean deviance is 9.394 which is a relatively small value which means our model fits well. We then used cross validation to determine whether we had to prune our tree. Figure 13 displays a plot of the tree's size against its deviance. According to the plot, we can see that the value for deviance is minimized when the tree size equals 10. Therefore, our original value of 10 is the best tree size so we did not have to prune our tree to make prediction on the test data set. To generate the MSE for the regression tree, we calculated the predicted value of life expectancy on the test data using the trained model and the true value of life expectancy on the test data. Then we used the mean function to calculate the MSE and we got a result of 11.14319.

Model evaluation

We examined the MSE value of our tree to evaluate the model's accuracy and serve as a comparison to our previous MSE values. Our MSE value came out to 11.14319, which in comparison to our previous MSE values was significantly lower. Therefore, we determined that our regression tree served as the best out of the models we found. On Figure 14, we can see how the true life expectancy compares to our predicted life expectancy. The plot shows that our model's predictions follows a similar trend to the real values.

Conclusions

Figure 15 displays our finalized regression tree which was both more interpretable and had the best accuracy rate out of our regression models. If an observation has an Income composition of resources value greater than 0.8075 but an Adult Mortality lower than 110.5 then we get the highest predicted value of life expectancy at 81.25. On the other hand, if the value for Income composition of resources is less than 0.5765 but the value of HIV/AIDS is greater than 16.25 then we predict a value of 47.35 for life expectancy which is the lowest value among the terminal nodes.

The three most influential predictors to building the model were Income composition of resources, HIV/AIDS and Adult Mortality. According to the regression tree, the positive effect of income composition of resources on life expectancy was the most important factor contributing towards increasing life expectancy which is consistent with the results from our linear regression models.

Practical Implications

The importance of income composition of resources makes sense because it is an index for how countries use their resources effectively. However, since it is an index composed of multiple predictors then we cannot conclude there is a single predictor primarily responsible for life expectancy. Instead, this should communicate to decision makers that prioritizing health, social, and economic variables collectively can result in a high life expectancy.

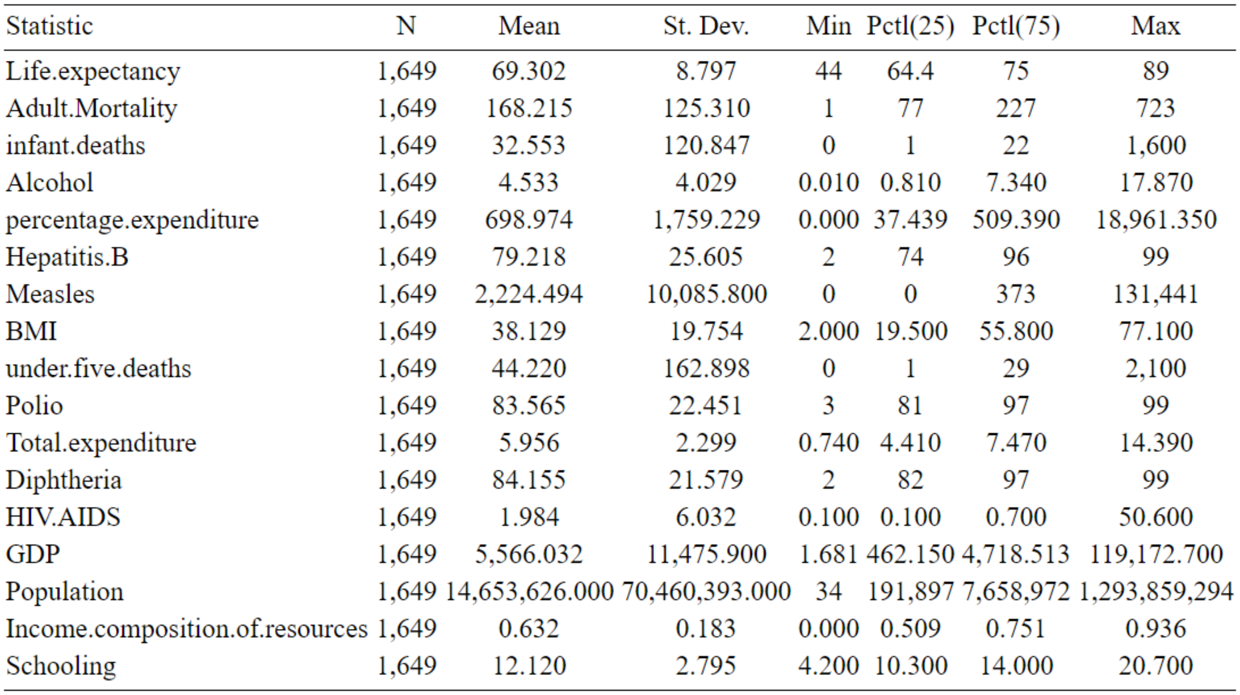
Furthermore, trying to predict an individual's precise life expectancy from our regression tree model is impractical, considering all the other variables in life that can't be accounted for. Rather this is a representation of what can be expected at a population level.

Our regression tree results are in line with the HIV/AIDS variable description which states that it measures death rates among 0-4 years old in 1000 live births in contrast to variables like Diptheria, Hepatitis B, and Polio which measure immunization coverage among 1 year-olds. The other virus that does not measure immunizations is measles but the difference is that there is a known measles vaccine whereas there is none for AIDS/HIV which could be a possible explanation regarding why it is less relevant towards predicting life expectancy. Therefore, our results are consistent in having both HIV/AIDS and Adult Mortality being two of the most influential predictors because they both measure forms of death and would directly detract from average life expectancy across a population.

We believe a similar study could be done on a much larger scale where the findings could be put into practice by government officials. Based off our analysis, we would advise representatives of all nations to prioritize raising their peoples' disposable income so that they can afford the health, social, and economic benefits that make up the income composition of resources values.

Figures:

Figure 1. Statistical Analysis on Factors Influencing Life Expectancy summary table



Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Life.expectancy	1,649	69.302	8.797	44	64.4	75	89
Adult.Mortality	1,649	168.215	125.310	1	77	227	723
infant.deaths	1,649	32.553	120.847	0	1	22	1,600
Alcohol	1,649	4.533	4.029	0.010	0.810	7.340	17.870
percentage.expenditure	1,649	698.974	1,759.229	0.000	37.439	509.390	18,961.350
Hepatitis.B	1,649	79.218	25.605	2	74	96	99
Measles	1,649	2,224.494	10,085.800	0	0	373	131,441
BMI	1,649	38.129	19.754	2.000	19.500	55.800	77.100
under.five.deaths	1,649	44.220	162.898	0	1	29	2,100
Polio	1,649	83.565	22.451	3	81	97	99
Total.expenditure	1,649	5.956	2.299	0.740	4.410	7.470	14.390
Diphtheria	1,649	84.155	21.579	2	82	97	99
HIV.AIDS	1,649	1.984	6.032	0.100	0.100	0.700	50.600
GDP	1,649	5,566.032	11,475.900	1.681	462.150	4,718.513	119,172.700
Population	1,649	14,653,626.000	70,460,393.000	34	191,897	7,658,972	1,293,859,294
Income.composition.of.resources	1,649	0.632	0.183	0.000	0.509	0.751	0.936
Schooling	1,649	12.120	2.795	4.200	10.300	14.000	20.700

Figure 2. Boxplot for Life Expectancy vs Status

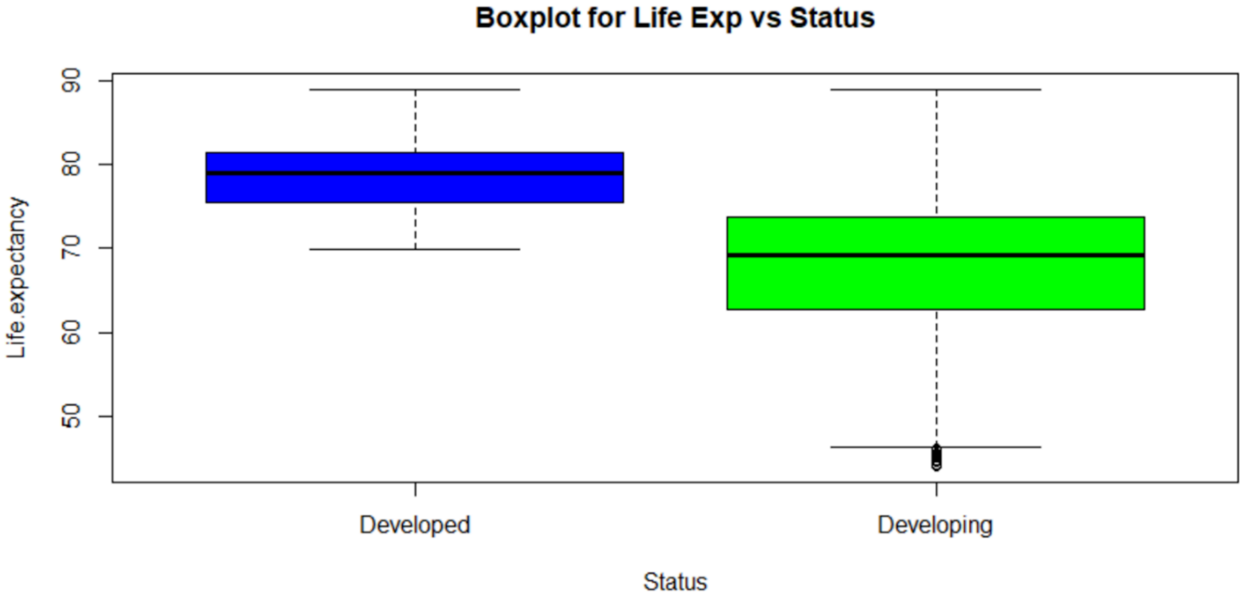


Figure 3. Boxplot for Life Expectancy vs Continent

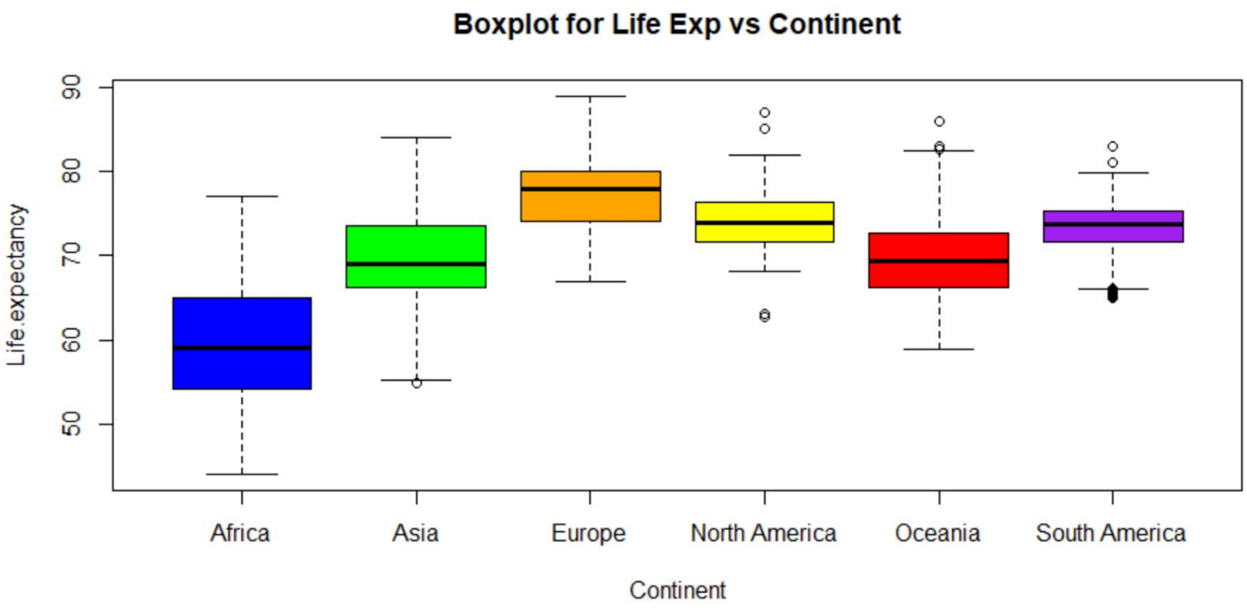


Figure 4. Residuals vs Fitted plot of status model

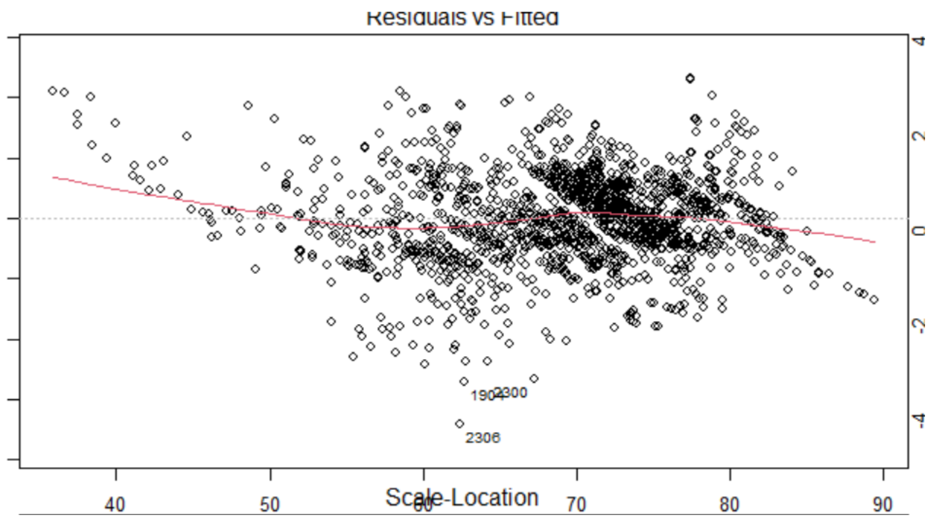


Figure 5. Scale-Location of status model

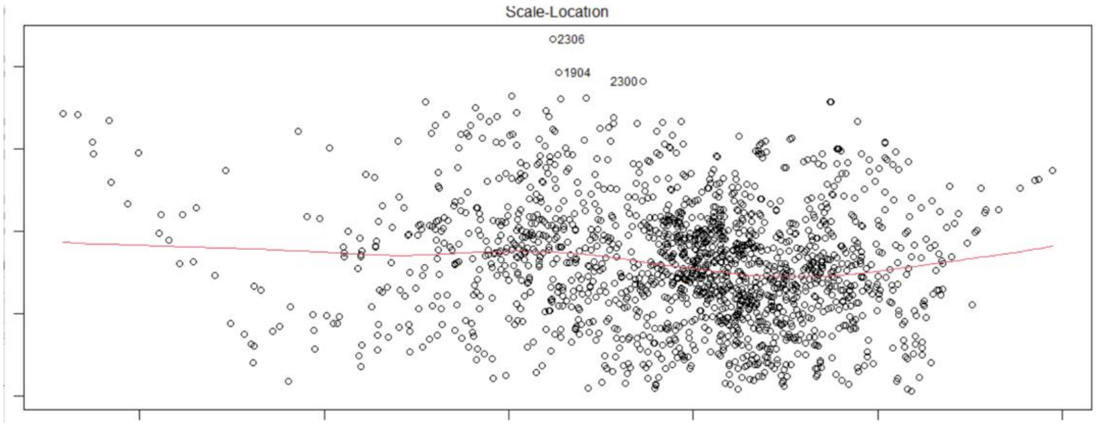


Figure 6. Residuals vs Leverage of status model

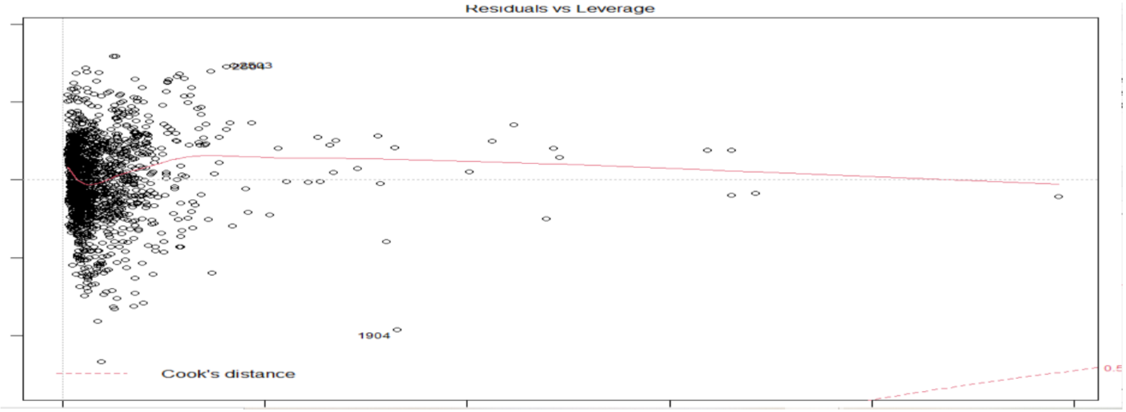


Figure 7. Normal Q-Q plot of status model

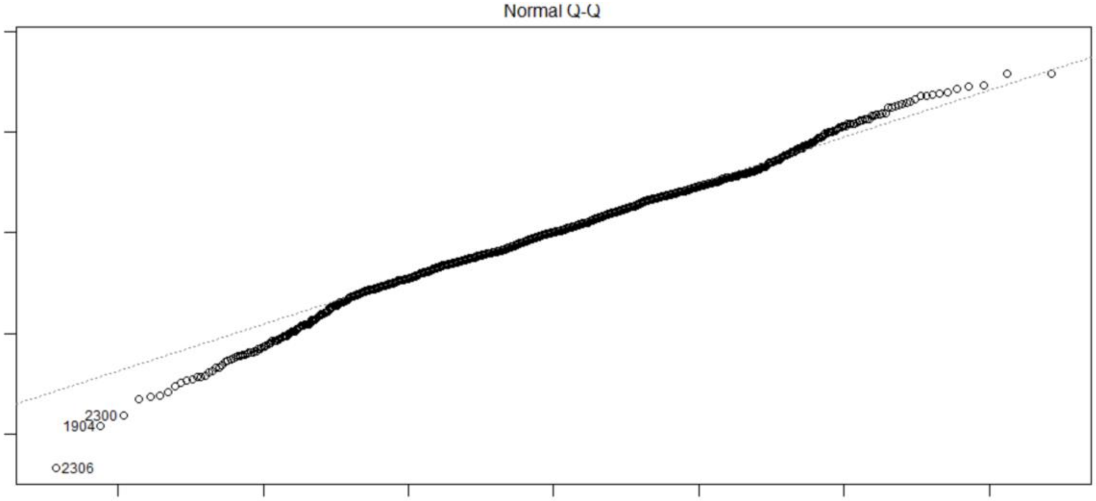


Figure 8. Residuals vs Fitted plot of Continent model

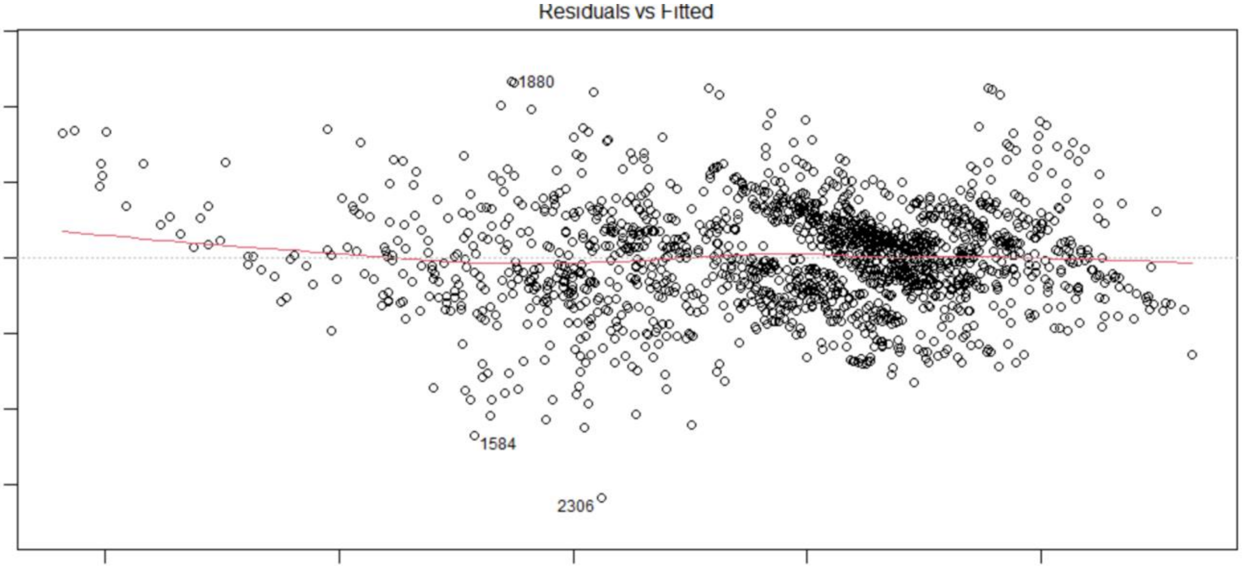


Figure 9. Scale-Location of Continent model

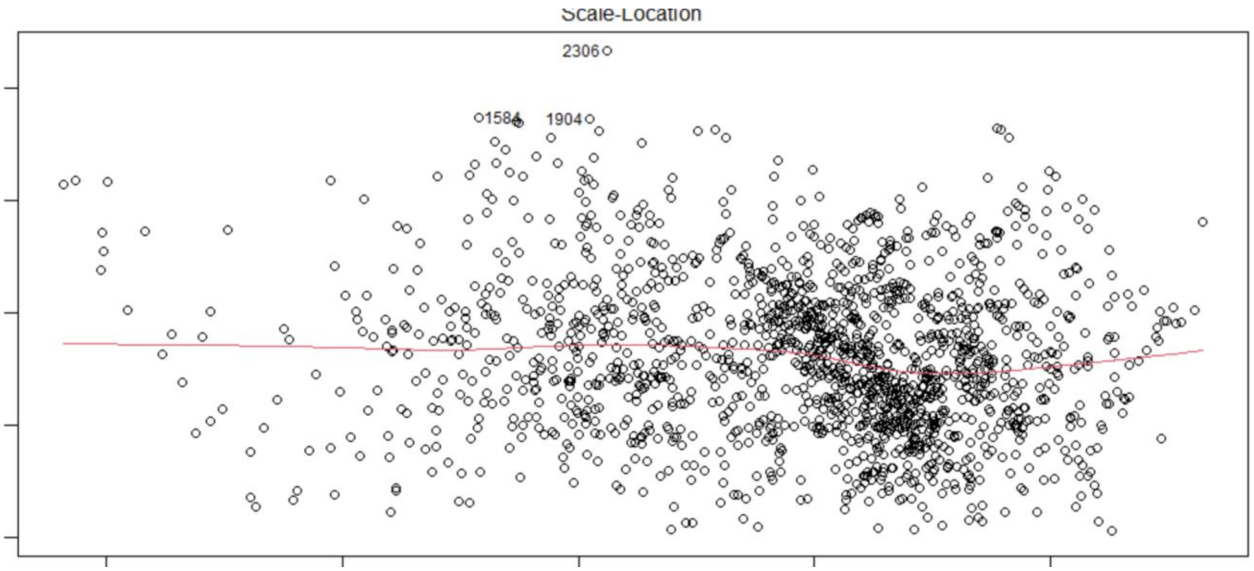


Figure 10. Normal Q-Q plot of Continent model

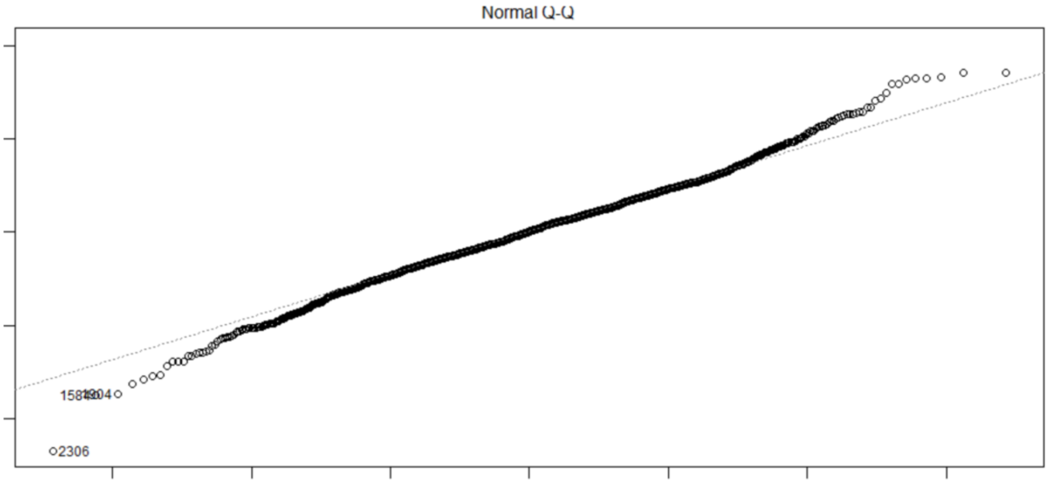


Figure 11. Residuals vs Leverage of status model

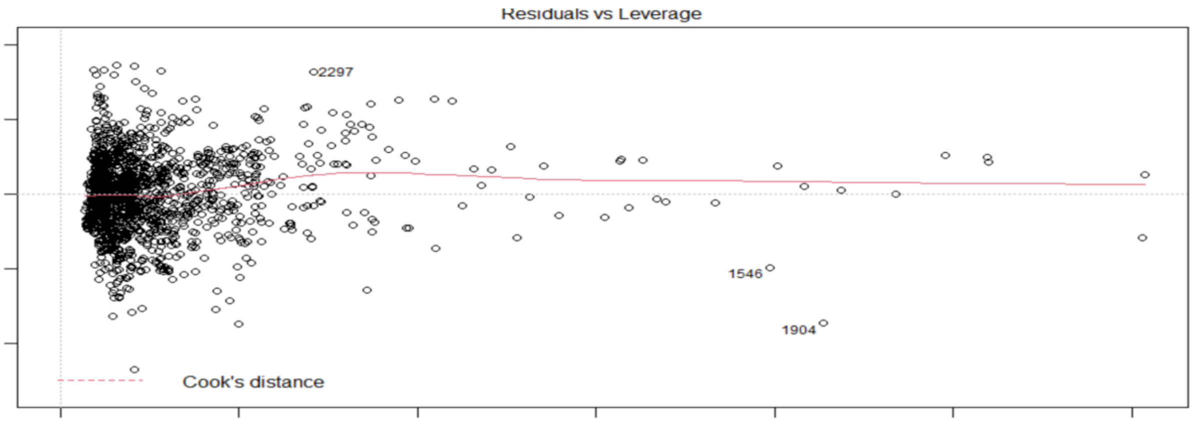
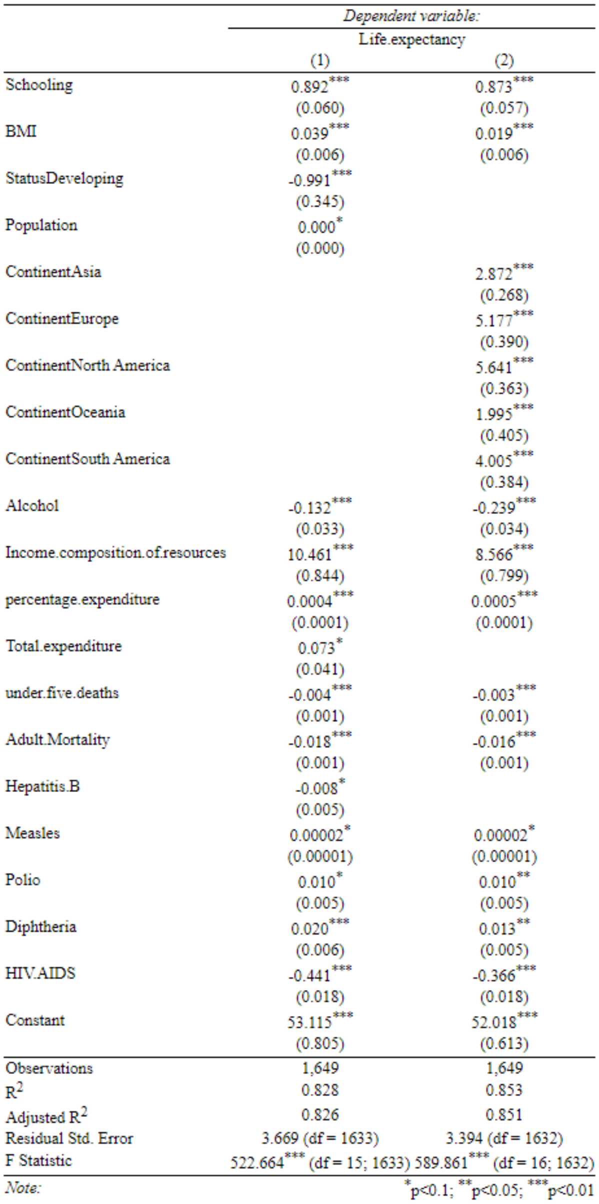


Figure 12 - Dependent Variables summary table



	Dependent variable:	
	Life.expectancy	
	(1)	(2)
Schooling	0.892***	0.873***
	(0.060)	(0.057)
BMI	0.039***	0.019***
	(0.006)	(0.006)
StatusDeveloping	-0.991***	
	(0.345)	
Population	0.000*	
	(0.000)	
ContinentAsia		2.872***
		(0.268)
ContinentEurope		5.177***
		(0.390)
ContinentNorth America		5.641***
		(0.363)
ContinentOceania		1.995***
		(0.405)
ContinentSouth America		4.005***
		(0.384)
Alcohol	-0.132***	-0.239***
	(0.033)	(0.034)
Income.composition.of.resources	10.461***	8.566***
	(0.844)	(0.799)
percentage.expenditure	0.0004***	0.0005***
	(0.0001)	(0.0001)
Total.expenditure	0.073*	
	(0.041)	
under.five.deaths	-0.004***	-0.003***
	(0.001)	(0.001)
Adult.Mortality	-0.018***	-0.016***
	(0.001)	(0.001)
Hepatitis.B	-0.008*	
	(0.005)	
Measles	0.00002*	0.00002*
	(0.00001)	(0.00001)
Polio	0.010*	0.010**
	(0.005)	(0.005)
Diphtheria	0.020***	0.013**
	(0.006)	(0.005)
HIV.AIDS	-0.441***	-0.366***
	(0.018)	(0.018)
Constant	53.115***	52.018***
	(0.805)	(0.613)
Observations	1,649	1,649
R^2	0.828	0.853
Adjusted R^2	0.826	0.851
Residual Std. Error	3.669 (df = 1633)	3.394 (df = 1632)
F Statistic	522.664*** (df = 15; 1633)	589.861*** (df = 16; 1632)
Note:		*p<0.1; **p<0.05; ***p<0.01

Figure 13 – size vs deviance plot

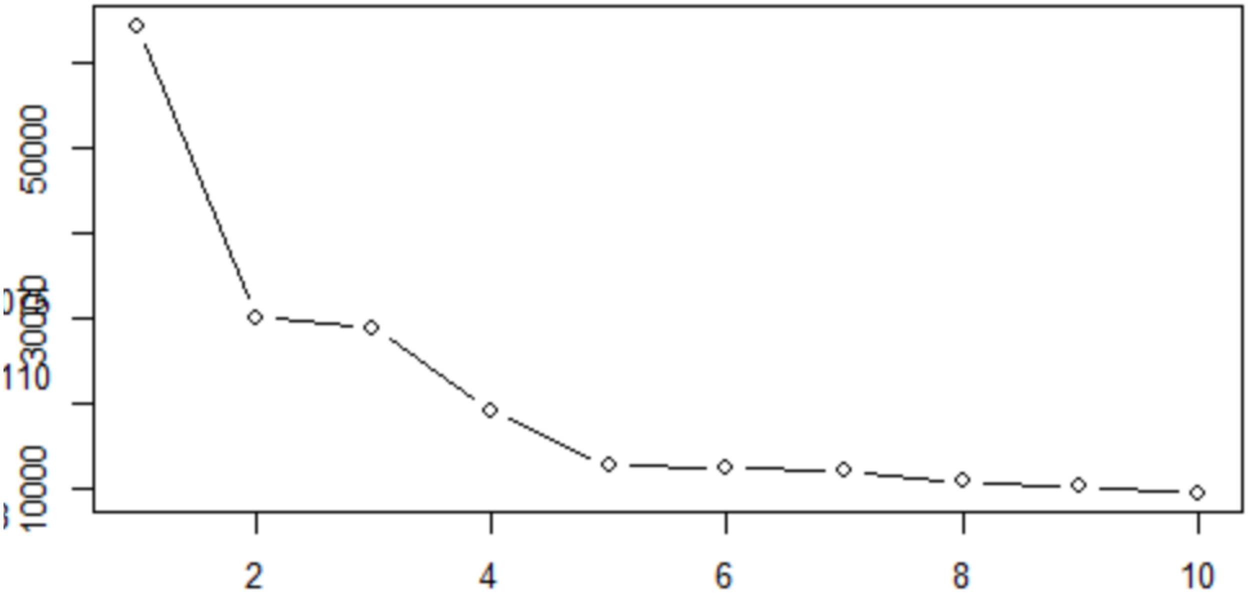


Figure 14 – **true y vs predicted y**

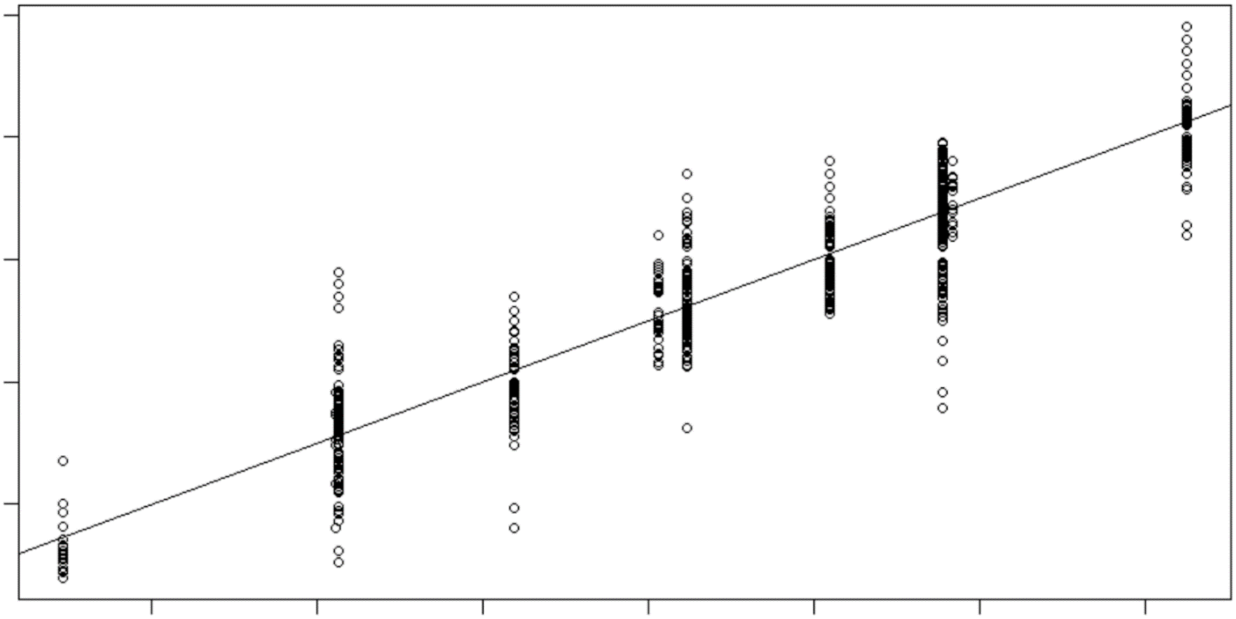


Figure 15 - Regression Tree for model containing Continent

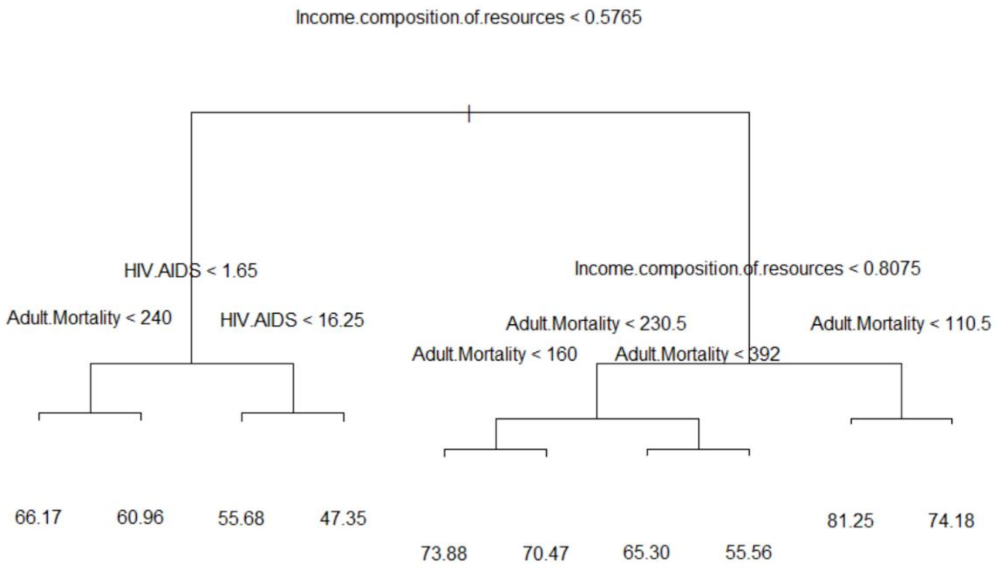
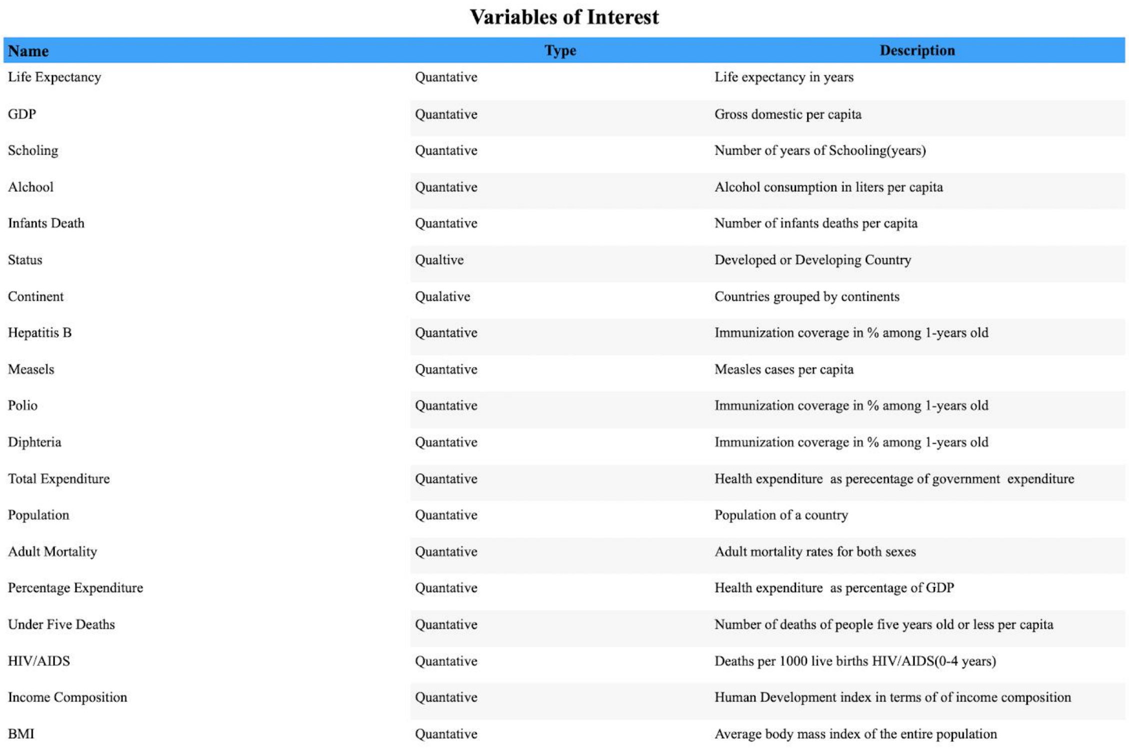


Figure 16 – Variables of interest



Variables of Interest

Name	Type	Description
Life Expectancy	Quantative	Life expectancy in years
GDP	Quantative	Gross domestic per capita
Scholing	Quantative	Number of years of Schooling(years)
Alchool	Quantative	Alcohol consumption in liters per capita
Infants Death	Quantative	Number of infants deaths per capita
Status	Qualtive	Developed or Developing Country
Continent	Qualative	Countries grouped by continents
Hepatitis B	Quantative	Immunization coverage in % among 1-years old
Measels	Quantative	Measles cases per capita
Polio	Quantative	Immunization coverage in % among 1-years old
Diphteria	Quantative	Immunization coverage in % among 1-years old
Total Expenditure	Quantative	Health expenditure as perecentage of government expenditure
Population	Quantative	Population of a country
Adult Mortality	Quantative	Adult mortality rates for both sexes
Percentage Expenditure	Quantative	Health expenditure as percentage of GDP
Under Five Deaths	Quantative	Number of deaths of people five years old or less per capita
HIV/AIDS	Quantative	Deaths per 1000 live births HIV/AIDS(0-4 years)
Income Composition	Quantative	Human Development index in terms of of income composition
BMI	Quantative	Average body mass index of the entire population

Works Cited

Rajarshi, K. (2018, February 10). Life Expectancy (WHO). Retrieved November 5, 2020, from https://www.kaggle.com/kumarajarshi/life-expectancy-who