MACHINE LEARNING APPLICATIONS IN PREDICTING LIFE EXPECTANCY AND A NATION'S DEVELOPMENT STATUS

Tristan Luu

Denison University – Data Analytics Department

Abstract

The goal of this project is to apply different Machine Learning methods to predict the life expectancy of countries around the world and to classify whether a nation is a developed or developing country. The Life Expectancy data set was downloaded from Kaggle but was originally collected from World Health Organization (WHO) and the United Nation (UN) websites. The type of Machine Learning algorithms implemented is supervised learning for all models. This project is divided into two main sections. In the first part, a fixed effect linear regression is used to predict the life expectancy of different nations. In the second section, a parametric approach (logistic regression) and a non-parametric approach (decision tree) is implemented to classify a nation's development status. Multiple indexes are used to comprehensively evaluate the prediction results, such as confusion matrix, MSE, adjusted R-squared, receiver operating characteristic (ROC) curve, and are under the curve (AUC) score. After comparing those values, it is concluded that decision tree performs slightly better than logistic regression when classifying a nation's development status.

1. Introduction

The world has been suffering from the Covid-19 pandemic for the past few years, which drastically affected each and every countries in many ways. During these unprecedented times, the

two major concerns are health and the economy, which serves as an inspiration for this project. This project aims to answer the following questions:

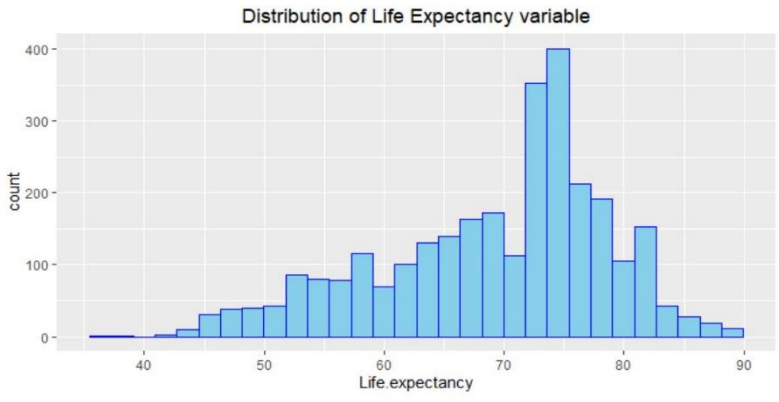
1) What variables most significantly affect the life expectancy of a country? Did the life expectancy increase or decrease from 2000 to 2015.
2) Does a parametric or non-parametric model classify a nation's development status better?

The data set was downloaded from Kaggle. However, the data related to life expectancy and health factors for 193 countries was originally collected from The Global Health Observatory (GHO) data repository under World Health Organization (WHO) website and its corresponding economic data was originally collected from United Nation website. It is made available to public for the purpose of health data analysis.

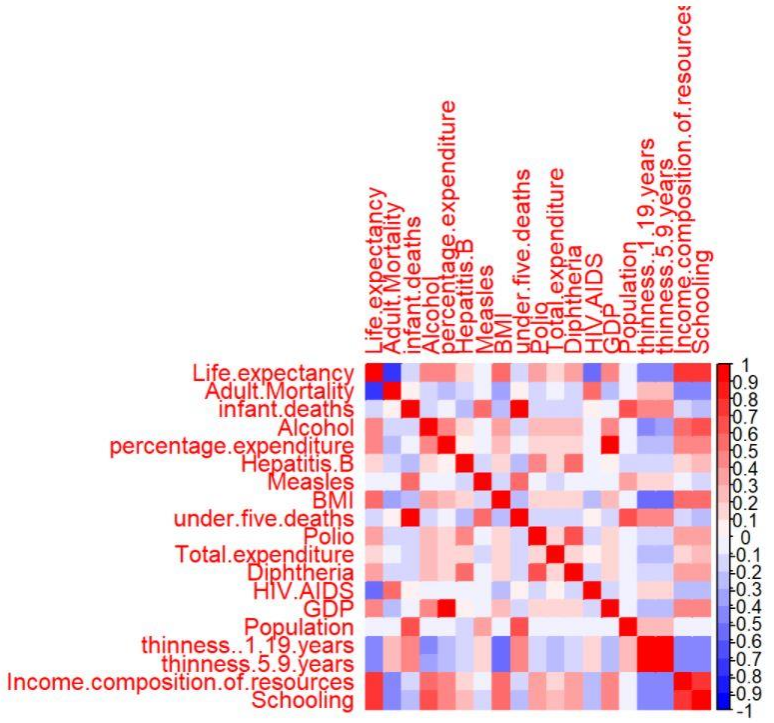
2. Methods

2.1 Exploratory Data Analysis

The dataset from this project consists of 22 columns and 2938 rows with information related to economic and health factors in 193 countries from 2000 to 2015. There are 20 predicting variables, which can be divided into several broad categories: Immunization related factors, Mortality factors, Economical factors and Social factors. The full list of definitions for all of the variables is given in the Appendix A.1.



This figure shows the distribution of the chosen dependent variable, life expectancy, for the fixed effect linear regression. The distribution seems to be left-skewed and is not quite normally distributed. There might be a few outliers on the left side of the graph, which means that there are a few countries with a very low life expectancy at only around 40 years-old.



This figure presents a heatmap correlation matrix for all of the variables in the dataset. The highly positively correlated pairs are in bright red, and the highly negatively correlated pairs are in dark blue. Some highly correlated pairs are life expectancy – adult mortality, life epectancy – schooling, infant deaths – BMI, percentage expenditure – GDP, under-five deaths – infant deaths, and thinness for age 10 to 19 – thinness for age 5 to 9.

In order to prepare for the modelling stage, the false types of some variables are changed to their correct types, and some missing data points are also removed. This data set is then splited into a training and a test set by a 80/20 ratio. Additionally, the full list of the basic statistics, such as the mean, median, minumum and maximum and types of all of the variables is given in Appendix A.2.

2.2 Fixed effect linear regression

The first model being used in this project is fixed effect linear regression to predict the life expectancy of different coutries. Linear regression assumes that there is a linear relationship between the independent and dependent variables, which means it follows a straight line. The form of the equation is y = a + b*x. Since this is a panel data set (a combination of time-series and cross-sectional data), the year variable is used as the fixed effect in the model so that it is possible to compare the life expectancy differences of a country in different years between 2000 and 2015.

The variable selection procedure starts with using all of the predicting variables in the data set as the independent variables and life expectancy as the dependent variable. Then the insignificant variables with p-values larger than 0.1 are removed one at a time. The final model is reached when all of the independent variables are significant with p-values smaller than 0.1. The results and validation of the final model will be discussed in part 3 of this paper.

2.3 Logistic regression

The parametric approach being chosen to classify whether a nation is a developed or developing nation is logistic regression. Logistic regression also assumes that there is a linear relationship between the response variable and the predictors. However, unlike linear regression, this method is applied when the dependent variable is not continuos. Logistic regression is used to model the probability of a binary outcome. The form of the equation is P = (e^{a+bx}) / $(1+e^{a+bx})$.

The variable selection procedure is similar to the fixed effect linear regression's variable selection procedure. The modelling process starts with using all of the predicting variables in the data set as the independent variables and status as the response variable. Then the insignificant variables with p-values larger than 0.1 are removed one at a time. The final model is reached when

all of the independent variables are significant with p-values smaller than 0.1. The results and validation of the final model will be discussed in part 3 of this paper.

2.4 Decision tree

The non-parametric approach being chosen to classify a nation's development status is decision tree. The structure of a decision tree is very similar to a flowchart. On every node there is a condition/test, and depending on the test result, the tree will continue to grow in one of few possible directions until it reaches the final answer for the introduced problem at the start. The decision tree model can be set-up using numerical, categorical or a combination of both variables. The results and validation of the final model will be discussed in part 3 of this paper.

3. Results & Discussion

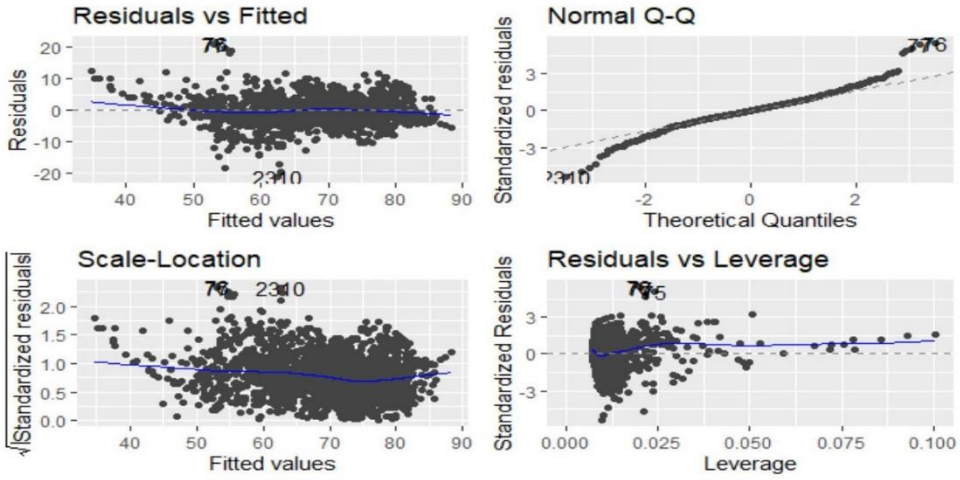
3.1 Fixed effect linear regression

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.229e+01  5.802e-01  90.134  < 2e-16 ***
Year2001                      -8.745e-01  4.944e-01  -1.769 0.077053 .
Year2002                      -1.414e+00  4.790e-01  -2.952 0.003191 **
Year2003                      -1.321e+00  4.786e-01  -2.761 0.005820 **
Year2004                      -8.811e-01  4.848e-01  -1.817 0.069292 .
Year2005                      -1.201e+00  4.841e-01  -2.481 0.013177 *
Year2006                      -1.382e+00  4.754e-01  -2.907 0.003689 **
Year2007                      -1.448e+00  4.851e-01  -2.985 0.002870 **
Year2008                      -1.280e+00  4.824e-01  -2.654 0.008011 **
Year2009                      -1.302e+00  4.808e-01  -2.707 0.006837 **
Year2010                      -1.323e+00  4.906e-01  -2.696 0.007074 **
Year2011                      -1.707e+00  4.921e-01  -3.469 0.000532 ***
Year2012                      -1.834e+00  4.871e-01  -3.766 0.000170 ***
Year2013                      -1.493e+00  4.865e-01  -3.068 0.002182 **
Year2014                      -1.377e+00  4.859e-01  -2.834 0.004637 **
Year2015                      -1.106e+00  4.938e-01  -2.240 0.025219 *
Adult.Mortality               -1.755e-02  9.096e-04 -19.294  < 2e-16 ***
percentage.expenditure         3.415e-04  4.668e-05   7.316 3.60e-13 ***
BMI                            3.408e-02  5.324e-03   6.402 1.88e-10 ***
under.five.deaths             -2.058e-03  5.485e-04  -3.752 0.000180 ***
Diphtheria                     4.378e-02  4.106e-03  10.663  < 2e-16 ***
HIV.AIDS                      -5.041e-01  1.972e-02 -25.557  < 2e-16 ***
Income.composition.of.resources  9.236e+00  7.198e-01  12.831  < 2e-16 ***
Schooling                      9.237e-01  4.743e-02  19.474  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.94 on 2151 degrees of freedom
  (175 observations deleted due to missingness)
Multiple R-squared:  0.8207,    Adjusted R-squared:  0.8188
F-statistic: 428.1 on 23 and 2151 DF,  p-value: < 2.2e-16
```

The figure above shows the final fixed effect model predicting life expectancy of different countries. The final fixed effect linear model is life expectancy = year + adult mortality +

percentage expenditure + BMI + under five deaths + diphtheria + HIV AIDS + Income composition of resources + Schooling. To answer the question at the start of the paper, the variables in the equation are the most significant ones that have a major impact on a country's life expectancy, compared to other variables in the data set. Additionally, the base year in this fixed effect model is 2000. Therefore, the life expectancy increased from 2000 to 2015 because the year's coefficient increases after each year.

All of the variables are significant at the convention levels, and the model itself also has a very low p-value. The null hypothesis here is that all of the coefficients in the model are zeros, and the alternative hypothesis indicates that at least one coefficient is not zero. The low p-value illustrates that the null hypothesis is rejected, which means there is a linear relationship between the response and the predicting variables. Therefore, this model is significant. Additionally, the adjusted R-squared of 0.8188 means that nearly 82% of the variation in the life expectancy variable can be explained by the independent variables in this model. Moreover, accoding to the correlation matrix, non of the predicting variables are highly correlated to each other, which illustrates that this model does not have a problem with multicollinearity.



This figure is used to validate the final model. The points spread out quite randomly in the residual plot, does not follow any , and tend to cluster towards the middle of the plot, which proves

that this is a good model. In the normal Q-Q plot, the data points form an upward trend but the line is not exactly 45 degrees. The two tails are a little off, which means there might be some outliers in the data set.

```
        Shapiro-Wilk normality test

data:  linear4$residuals
W = 0.97365, p-value < 2.2e-16
```

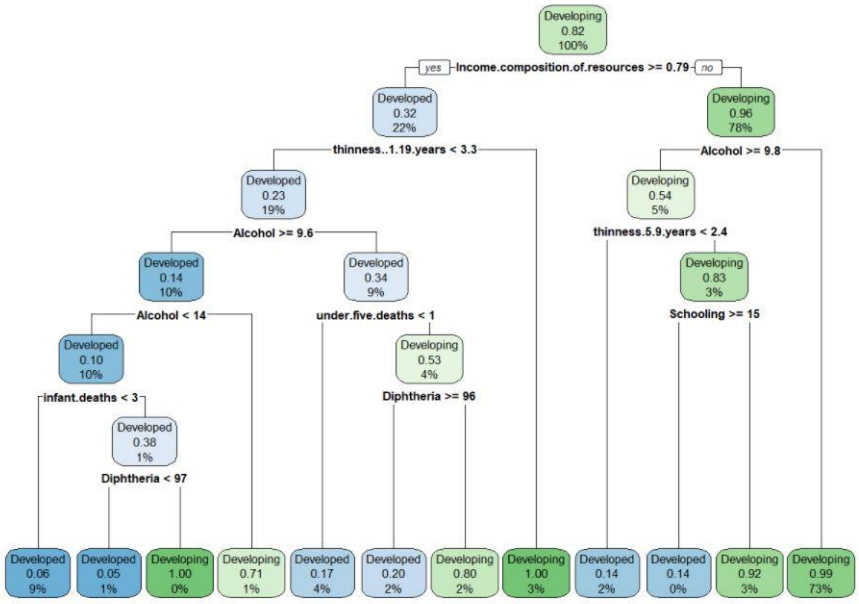
```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 238.5172, Df = 1, p = < 2.22e-16
```

A Shapiro-Wilk test is run to check the normality of the model. For this test, the null hypothesis is that the data follows a normal distribution, and the alternative hypothesis is the opposite. The P-value is very small, which means the null hypothesis is rejected. Therefore, the data is not proven to be consistent with normality. Moreover, a non-constant variance score test is used to check for homoscedasticity. The null hypothesis is that the data points are homoscedastic, and the alternative hypothesis is the opposite. P-value is very small indicating that the null hypothesis is rejected. Therefore, the data points are not very consistent with homoscedastcity. Lastly, this model is used on a test set, which results in a MSE of appoximately 13.138.

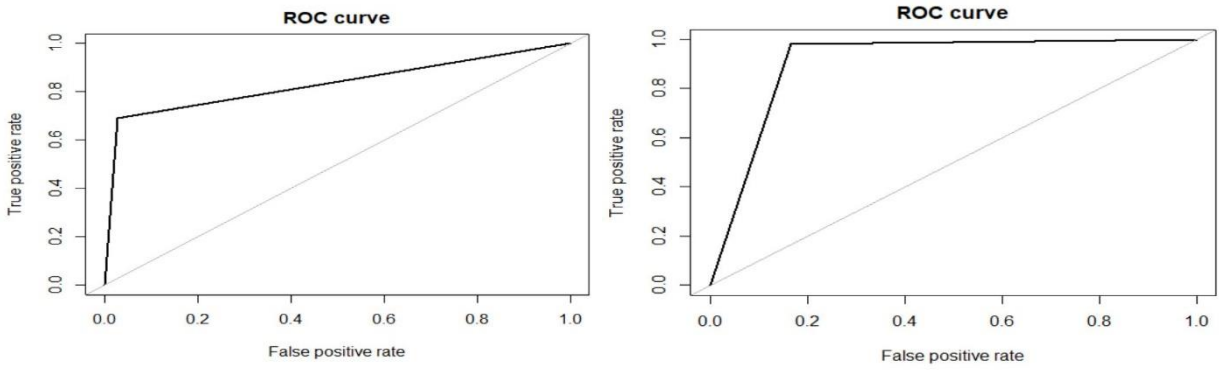
3.2 Logistic regression & Decision tree

```
glm(formula = Status ~ Alcohol + Hepatitis.B + under.five.deaths +
    thinness + Income.composition.of.resources, family = binomial,
    data = train2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.63820  -0.12993  -0.00189   0.00000   2.76453

Coefficients:
                                 Estimate Std. Error z value
(Intercept)                    -20.576247   2.190817  -9.392
Alcohol                          0.274699   0.034282   8.013
Hepatitis.B                      0.029341   0.005458   5.376
under.five.deaths               -0.278277   0.071210  -3.908
thinness                        -0.098608   0.024411  -4.040
Income.composition.of.resources  20.196507   2.378604   8.491
                                Pr(>|z|)
(Intercept)                      < 2e-16 ***
Alcohol                         1.12e-15 ***
Hepatitis.B                     7.63e-08 ***
under.five.deaths               9.31e-05 ***
thinness                        5.35e-05 ***
Income.composition.of.resources  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1368.41  on 1683  degrees of freedom
Residual deviance:  462.75  on 1678  degrees of freedom
  (666 observations deleted due to missingness)
AIC: 474.75

Number of Fisher Scoring iterations: 12
```

The above figure shows the final logistic regression to classify a nation's development status. On a special note, thinness is an interaction term where thinness = thinness age 5-9 * thinness age 10-19. Both of these variables are significant in the model but they are highly correlated according to the correlation matrix. Therefore, an interaction term is created so that it is possible to keep the thinness factor in the final model. Additionally, all of the variables are significant at the 1% level with p-values close to zero.



The above figure presents the final decision tree to classify whether a nation is a developed or developing country. The top five most significant factors from the decision tree that contribute to a nation's development status are income composition of resources, thinness, alcohol, under-five deaths, and schooling.



These are the two ROC curves for logistic regression (left) and decision tree (right). The area under the curve (AUC) for logistic regression is 83.1, while the AUC for decision tree is 90.7. Moreover, logistic regression results in an accuracy of 93.3% when being used on the test set, whereas decision tree achieves a slightly higher accuracy of 96.2%. From these statistics, it is concluded that decision tree performs better than logistic regression when classifying a nation's development status.

4. Conclusion & Recommendation

In conclusion, from the results of the fixed effect linear regression, people lived longer during the past years because life expectancy increased from 2000 to 2015. Additionally, in terms of classifying whether a nation is a developed or developing country, the decision tree method is statistically proven to perform better than the logistic regression approach.

As this data set only had data from year 2000 to 2015, which is pre-Covid time, the predictions might be different now. During the past few years, Covid-19 has caused many youth and adult deaths and heavily affected the income and a country's GDP. Taking into account those major changes caused by Covid-19, deaths caused by Covid-19 could be introduced as a new variable in the model. Moreover, the variables that are already in the models could also lose their significance because of this. Therefore, new data should be updated into this data set in order to correctly predict the life expectancy and classify a nation's development status now.

5. References

Rajarshi, Kumar. "Life Expectancy (WHO)." Kaggle, 10 Feb. 2018, https://www.kaggle.com/kumarajarshi/life-expectancy-who.

Appendix A

A.1 Variables' definitions

- Country: Country name
- Year: Year (from 2000 to 2015)
- Status: Developed or Developing status
- Life expectancy: Life Expectancy in age
- Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- infant deaths: Number of Infant Deaths per 1000 population
- Alcohol: Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
- percentage expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita (%)
- Hepatitis B: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- Measles: Measles - number of reported cases per 1000 population
- BMI: Average Body Mass Index of entire population
- under-five deaths: Number of under-five deaths per 1000 population
- Polio: Polio (Pol3) immunization coverage among 1-year-olds (%)
- Total expenditure: General government expenditure on health as a percentage of total government expenditure (%)
- Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
- HIV/AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years)
- GDP: Gross Domestic Product per capita (in USD)

- Population: Population of the country
- thinness 1-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
- thinness 5-9 years: Prevalence of thinness among children for Age 5 to 9 (%)
- Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- Schooling: Number of years of Schooling (years)

A.2 Summary of the variables

```
  Country               Year         Status          Life.expectancy
Length:2938        Min.   :2000   Length:2938        Min.   :36.30
Class :character   1st Qu.:2004   Class :character   1st Qu.:63.10
Mode  :character   Median :2008   Mode  :character   Median :72.10
                   Mean   :2008                      Mean   :69.22
                   3rd Qu.:2012                      3rd Qu.:75.70
                   Max.   :2015                      Max.   :89.00
                                                     NA's   :10
Adult.Mortality infant.deaths       Alcohol
Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100
1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775
Median :144.0   Median :   3.0   Median : 3.7550
Mean   :164.8   Mean   :  30.3   Mean   : 4.6029
3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025
Max.   :723.0   Max.   :1800.0   Max.   :17.8700
NA's   :10                       NA's   :194
percentage.expenditure  Hepatitis.B      Measles
Min.   :    0.000      Min.   : 1.00   Min.   :     0.0
1st Qu.:    4.685      1st Qu.:77.00   1st Qu.:     0.0
Median :   64.913      Median :92.00   Median :    17.0
Mean   :  738.251      Mean   :80.94   Mean   :  2419.6
3rd Qu.:  441.534      3rd Qu.:97.00   3rd Qu.:   360.2
Max.   :19479.912      Max.   :99.00   Max.   :212183.0
                       NA's   :553
     BMI        under.five.deaths     Polio       Total.expenditure
Min.   : 1.00   Min.   :   0.00   Min.   : 3.00   Min.   : 0.370
1st Qu.:19.30   1st Qu.:   0.00   1st Qu.:78.00   1st Qu.: 4.260
Median :43.50   Median :   4.00   Median :93.00   Median : 5.755
Mean   :38.32   Mean   :  42.04   Mean   :82.55   Mean   : 5.938
3rd Qu.:56.20   3rd Qu.:  28.00   3rd Qu.:97.00   3rd Qu.: 7.492
Max.   :87.30   Max.   :2500.00   Max.   :99.00   Max.   :17.600
NA's   :34                        NA's   :19      NA's   :226
  Diphtheria       HIV.AIDS           GDP
Min.   : 2.00   Min.   : 0.100   Min.   :     1.68
1st Qu.:78.00   1st Qu.: 0.100   1st Qu.:   463.94
Median :93.00   Median : 0.100   Median :  1766.95
Mean   :82.32   Mean   : 1.742   Mean   :  7483.16
3rd Qu.:97.00   3rd Qu.: 0.800   3rd Qu.:  5910.81
Max.   :99.00   Max.   :50.600   Max.   :119172.74
```

```
  Population        thinness..1.19.years thinness.5.9.years
 Min.   :3.400e+01   Min.   : 0.10        Min.   : 0.10
 1st Qu.:1.958e+05   1st Qu.: 1.60        1st Qu.: 1.50
 Median :1.387e+06   Median : 3.30        Median : 3.30
 Mean   :1.275e+07   Mean   : 4.84        Mean   : 4.87
 3rd Qu.:7.420e+06   3rd Qu.: 7.20        3rd Qu.: 7.20
 Max.   :1.294e+09   Max.   :27.70        Max.   :28.60
 NA's   :652         NA's   :34           NA's   :34
 Income.composition.of.resources   Schooling
 Min.   :0.0000                  Min.   : 0.00
 1st Qu.:0.4930                  1st Qu.:10.10
 Median :0.6770                  Median :12.30
 Mean   :0.6276                  Mean   :11.99
 3rd Qu.:0.7790                  3rd Qu.:14.30
 Max.   :0.9480                  Max.   :20.70
 NA's   :167                     NA's   :163
'data.frame':   2938 obs. of  22 variables:
 $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
 $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
 $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
 $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
 $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
 $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
 $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
 $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
 $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
 $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
 $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
 $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
 $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
 $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
 $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
 $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
 $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
 $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ..
 $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```