# A statistical life expectancy analysis

Pol Abadia Conejos

June 2024

## 1 Abstract

The life expectancy is an important index that can be used to measure the lifestyle and healthcare of the population of a country or region. Higher life expectancy usually translates to better economies and health systems. However many other factors can influence the life expectancy of a population such as drug consumption, eating habits or social factors like schooling and education. It is primordial to have a deep understanding about how all these factors work together in order to assess governments and authorities which measures to take to boost the overall lifespan of their people.

## 2 Introduction

The goal of this project is to make a statistical analysis of how mortality, social and economical factors affect the lifespan of a population. That is to research whether there is a relation or not between life expectancy and a given variable, and if that is so, which type of relation there is (linear, exponential, logarithmic...)

## 3 Description of the dataset

The data for this project was extracted from a Kaggle dataset that collects data from the World Health Organization (WHO). The dataset contains data about 21 health and social features among 179 countries recorded from 2000 to 2015.

We will now take a first look on the dataset and explore its features and their meaning.

- Life expectancy: Average lifespan in years, our target variable.
- Status: Binary feature. It indicates the socioeconomic status of a country. A country can either be developing or developed based on whether it has a weaker or stronger economy, political institutions and democracy.

- GDP: Gross Domestic Product per capita (in USD), which approximately measures the concept of the living standards of the population in a given country. Mathematically it's defined as the country's total GDP divided by its population size.
- Population: Size of the population of the country.
- Percentage expenditure: Percentage of the total GDP destined to health.
- Schooling: Average years of formal education spent by people older than the age of 25.
- Adult mortality: Measures the number of deaths of adults per 1000 population.
- Infant deaths: Measures the number of deaths of infants per 1000 population.
- Under five deaths: number of deaths per 1000 population under the age of 5.
- Alcohol: Represents the consumption of alcohol recorded in liters of pure alcohol per capita of individuals older than the age of 15 years.
- Hepatitis B: Percentage of the coverage of Hepatitis B immunization among 1 year olds.
- Measles: number of measles cases per 1000 population.
- Polio: Percentage of the coverage of Polio immunization among 1 year olds.
- Diphtheria: Percentage of the coverage of Diphtheria tetanus toxoid and pertussis (DTP3) immunization among 1 year olds.
- Incidents HIV/AIDS: Number of HIV/AIDS incidents per 1000 population from the age of 15 to 49.
- BMI: Body mass index. Measures the nutritional status of an adult. Defined as the mass (Kg) of the person divided by the square of its height (in meters).
- Thinness 5-9 years: Prevalence (Percentage) of thinness among children from the age of 5 to 9. An individual of a given age is considered to suffer from thinness if its BMI is two standard deviations below the median.
- Thinness 10-19 years: Prevalence (Percentage) of thinness among preadolescents and adolescents from the age of 10 to 19.

Let's take a look on the distribution of some features and their relation with other variables. We can see some of the features present in the dataset such as life expectancy, schooling or total expenditure exhibit a normal distribution, even though some of it shows some skewness.

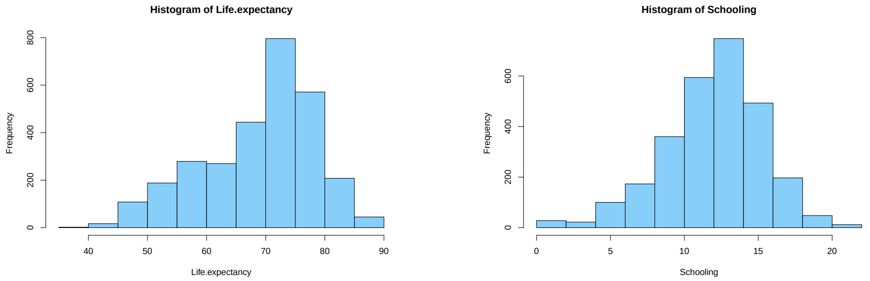


Figure 1: Frequency plots of life expectancy and schooling

The mode of life expectancy in the dataset is 70-75 years. The mode of schooling and education is 13-14 years.

We can see a special case of distribution in the BMI where there is a multimodal distribution, two peaks are shown on the histogram as if there was a mixture of two normal distributions. This suggests there are two values where BMI samples tend to fall around, which may be related to the wealth of the country.

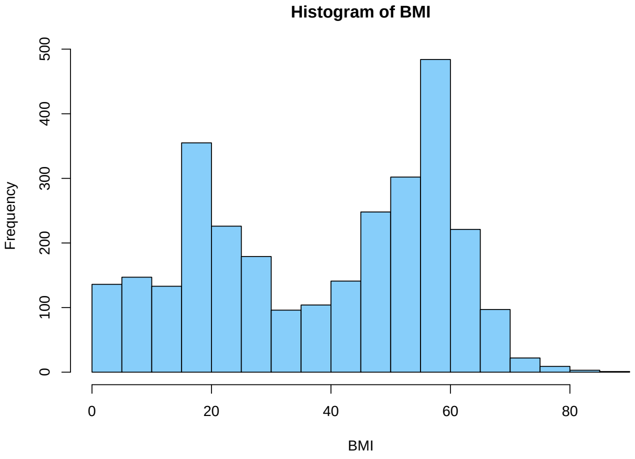


Figure 2: Frequency plot of the BMI

Many other variables exhibit left or right skewed distributions depending on their nature. For example, the distributions of infant deaths and measles incidents are completely left skewed, which makes sense since it shows the important global efforts there has been in the last decades to diminish these tragedies.

On the other hand the GDP histogram reflects the current global inequality

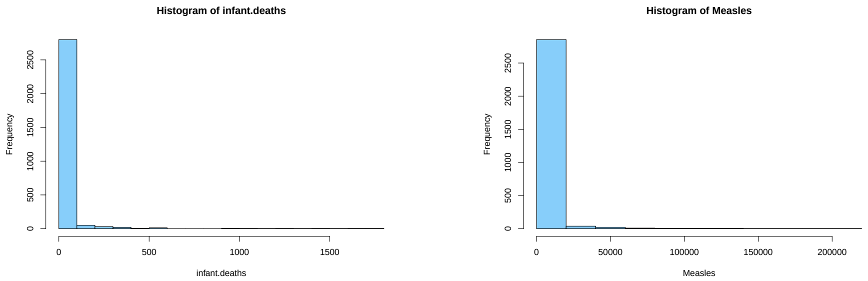


Figure 3: Frequency plots of infant deaths and measles incidents per 1000 population

where the vast majority of countries tend to live in poverty and only a few countries own most of the wealth. In fact, 82 per cent of the samples in this dataset have a GDP per capita below 10.000 USD. Notice how we are using the term "samples" and not "countries" since this dataset contains data of the same countries ranging from different years.

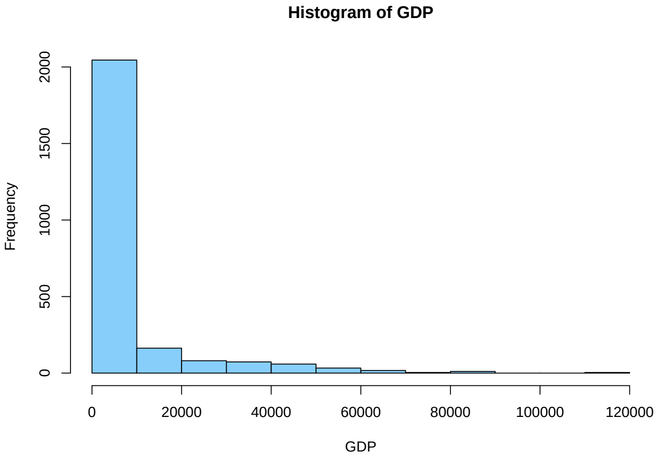
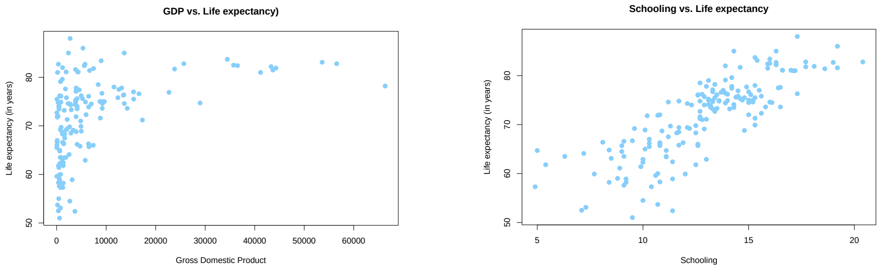


Figure 4: Frequency plot of the GDP per capita

We move on to explore the relation between the most important variables. The features with the most striking relation with life expectancy are the GDP and the schooling years. At a first glance just by looking at the scatter plot, one can observe a curve that resembles a logarithmic relation between GDP and Life expectancy. The schooling and life expectancy data also seem to have a clear linear relation. We will study these relations later on in the model evaluation step to get a mathematical value ($R^2$) of how much of the variance of the life


GDP vs. Life expectancy)

Schooling vs. Life expectancy

Figure 5: Scatter plots of GDP and schooling years vs life expectancy in 2015

expectancy data is explained by those variables.

Other variables were also considered by making more scatter plots against the target variable, where there's a faint relation as for example BMI and thinness suffered from age 10 to 19. There is a much clearer but unusual relationship with adult mortality. A seen in figure 6 there are two different trends represented by two linear relations, this situation could be modeled maybe with a broken line regression.

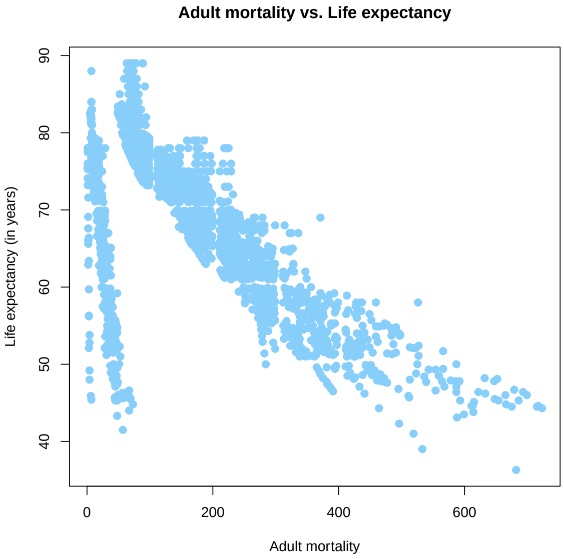

Adult mortality vs. Life expectancy

Figure 6: Scatter plot of adult mortality vs life expectancy

# 4 Model study

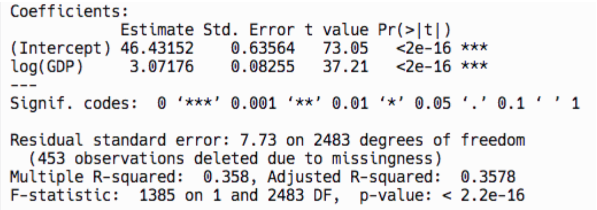
In this part of the project we'll perform a much more rigorous statistical analysis of the data. Different models have been employed to make predictions about the life expectancy of a country. Before that we'll make some considerations. To fit the models, the "Country" feature is removed from the dataset since, jokes aside, the name of a country shouldn't play a role in its socio-economical and health factors. Also, even though it might sound non sensical to train the models with the data of a same country for different years, our goal here is to predict the life expectancy of a hypothetical country given a set of their values (an instance of its predictors). Having the year of a country as an additional feature can help us predict how a country develops over time given an initial set of values.

Let's begin by finding out how well can some individual variables explain the life expectancy as we saw earlier in the scatter plots for GDP per capita and schooling. The following models were fitted:

$$\hat{Expectancy} = \beta_0 + \beta_1 log(\text{GDP}) + \varepsilon \tag{1}$$

$$\hat{Expectancy} = \beta_0 + \beta_1 \text{Schooling} + \varepsilon \tag{2}$$

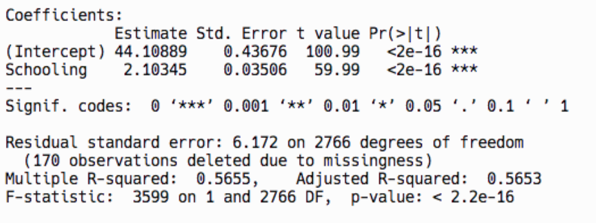
These were the results:



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.43152    0.63564   73.05   <2e-16 ***
log(GDP)     3.07176    0.08255   37.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.73 on 2483 degrees of freedom
  (453 observations deleted due to missingness)
Multiple R-squared:  0.358, Adjusted R-squared:  0.3578
F-statistic:  1385 on 1 and 2483 DF,  p-value: < 2.2e-16
```

Figure 7: Statistical summary of model (1)



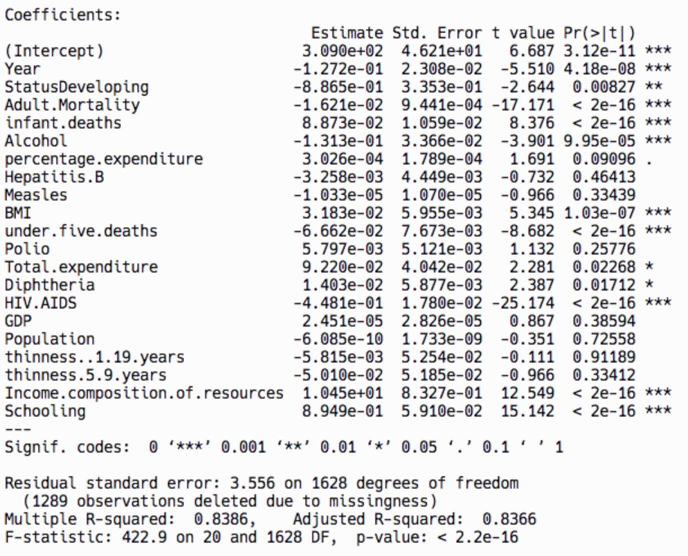
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.10889    0.43676  100.99   <2e-16 ***
Schooling    2.10345    0.03506   59.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.172 on 2766 degrees of freedom
  (170 observations deleted due to missingness)
Multiple R-squared:  0.5655,    Adjusted R-squared:  0.5653
F-statistic:  3599 on 1 and 2766 DF,  p-value: < 2.2e-16
```

Figure 8: Statistical summary of model (2)

As it can be seen in both summaries from figures 7 and 8, the $R^2$ coefficient indicates a clear relation. Even if these models by their own are not so good, it is notable how with just one variable so much of the life expectancy data can be explained, just like we suspected in the scatter plots before.

Fitting a full linear model with all the variables we get a pretty good performance ($R^2 = 0.8366$). Nevertheless, the p-values of some variables seem to indicate their contribution to the model is statistically insignificant, even the GDP variable has a high p-value although the previous log-model with itself alone as a predictor had a reasonable $R^2$ and a much lower p-value.



```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      3.090e+02  4.621e+01   6.687 3.12e-11 ***
Year                            -1.272e-01  2.308e-02  -5.510 4.18e-08 ***
StatusDeveloping                -8.865e-01  3.353e-01  -2.644  0.00827 ** 
Adult.Mortality                 -1.621e-02  9.441e-04 -17.171  < 2e-16 ***
infant.deaths                    8.873e-02  1.059e-02   8.376  < 2e-16 ***
Alcohol                         -1.313e-01  3.366e-02  -3.901 9.95e-05 ***
percentage.expenditure           3.026e-04  1.789e-04   1.691  0.09096 .  
Hepatitis.B                     -3.258e-03  4.449e-03  -0.732  0.46413    
Measles                         -1.033e-05  1.070e-05  -0.966  0.33439    
BMI                              3.183e-02  5.955e-03   5.345 1.03e-07 ***
under.five.deaths               -6.662e-02  7.673e-03  -8.682  < 2e-16 ***
Polio                            5.797e-03  5.121e-03   1.132  0.25776    
Total.expenditure                9.220e-02  4.042e-02   2.281  0.02268 *  
Diphtheria                       1.403e-02  5.877e-03   2.387  0.01712 *  
HIV.AIDS                        -4.481e-01  1.780e-02 -25.174  < 2e-16 ***
GDP                              2.451e-05  2.826e-05   0.867  0.38594    
Population                      -6.085e-10  1.733e-09  -0.351  0.72558    
thinness..1.19.years            -5.815e-03  5.254e-02  -0.111  0.91189    
thinness.5.9.years              -5.010e-02  5.185e-02  -0.966  0.33412    
Income.composition.of.resources  1.045e+01  8.327e-01  12.549  < 2e-16 ***
Schooling                        8.949e-01  5.910e-02  15.142  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.556 on 1628 degrees of freedom
  (1289 observations deleted due to missingness)
Multiple R-squared:  0.8386,	Adjusted R-squared:  0.8366 
F-statistic: 422.9 on 20 and 1628 DF,  p-value: < 2.2e-16
```

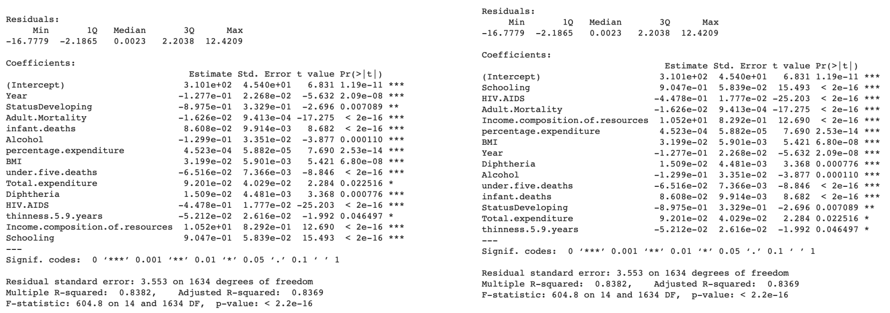
Figure 9: Statistical summary of linear model considering all possible predictors

To tackle this issue we'll undertake a process of stepwise regression in both forward and backward directions to find an optimal set of explanatory variables. The following figures (9 and 10) show the summaries of backward and forward selection respectively:

At a first glance there doesn't seem to be much difference in regards of their $R^2$ values and residuals, not only that but also they have the exact same predictors with the same coefficients. But notice how we went from having 20 predictors in our initial model to 14 in both stepwise models. It is important when constructing a model to keep it as simple as possible while retaining most of the information of the data, that way we may prevent overfitting the model.

Contrary to our expectations, the GDP of a country doesn't play much of a role when predicting life expectancy since it's not present in neither of both stepwise models.

To validate our final models we'll employ bootstrapping techniques (parametric



```
Residuals:
     Min       1Q   Median       3Q      Max 
-16.7779  -2.1865   0.0023   2.2038  12.4209 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      3.101e+02  4.540e+01   6.831 1.19e-11 ***
Year                            -1.277e-01  2.268e-02  -5.632 2.09e-08 ***
StatusDeveloping                -8.975e-01  3.329e-01  -2.696 0.007089 ** 
Adult.Mortality                 -1.626e-02  9.413e-04 -17.275  < 2e-16 ***
infant.deaths                    8.608e-02  9.914e-03   8.682  < 2e-16 ***
Alcohol                         -1.299e-01  3.351e-02  -3.877 0.000110 ***
percentage.expenditure           4.523e-04  5.882e-05   7.690 2.53e-14 ***
BMI                              3.199e-02  5.901e-03   5.421 6.80e-08 ***
under.five.deaths               -6.516e-02  7.366e-03  -8.846  < 2e-16 ***
Total.expenditure                9.201e-02  4.029e-02   2.284 0.022516 *  
Diphtheria                       1.509e-02  4.481e-03   3.368 0.000776 ***
HIV.AIDS                        -4.478e-01  1.777e-02 -25.203  < 2e-16 ***
thinness.5.9.years              -5.212e-02  2.616e-02  -1.992 0.046497 *  
Income.composition.of.resources  1.052e+01  8.292e-01  12.690  < 2e-16 ***
Schooling                        9.047e-01  5.839e-02  15.493  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.553 on 1634 degrees of freedom
Multiple R-squared:  0.8382,	Adjusted R-squared:  0.8369 
F-statistic: 604.8 on 14 and 1634 DF,  p-value: < 2.2e-16
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-16.7779  -2.1865   0.0023   2.2038  12.4209 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      3.101e+02  4.540e+01   6.831 1.19e-11 ***
Schooling                        9.047e-01  5.839e-02  15.493  < 2e-16 ***
HIV.AIDS                        -4.478e-01  1.777e-02 -25.203  < 2e-16 ***
Adult.Mortality                 -1.626e-02  9.413e-04 -17.275  < 2e-16 ***
Income.composition.of.resources  1.052e+01  8.292e-01  12.690  < 2e-16 ***
percentage.expenditure           4.523e-04  5.882e-05   7.690 2.53e-14 ***
BMI                              3.199e-02  5.901e-03   5.421 6.80e-08 ***
Year                            -1.277e-01  2.268e-02  -5.632 2.09e-08 ***
Diphtheria                       1.509e-02  4.481e-03   3.368 0.000776 ***
Alcohol                         -1.299e-01  3.351e-02  -3.877 0.000110 ***
under.five.deaths               -6.516e-02  7.366e-03  -8.846  < 2e-16 ***
infant.deaths                    8.608e-02  9.914e-03   8.682  < 2e-16 ***
StatusDeveloping                -8.975e-01  3.329e-01  -2.696 0.007089 ** 
Total.expenditure                9.201e-02  4.029e-02   2.284 0.022516 *  
thinness.5.9.years              -5.212e-02  2.616e-02  -1.992 0.046497 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.553 on 1634 degrees of freedom
Multiple R-squared:  0.8382,	Adjusted R-squared:  0.8369 
F-statistic: 604.8 on 14 and 1634 DF,  p-value: < 2.2e-16
```

Figure 10: Statistical summary of backward and forward selection models

and non-parametric) by estimating the coefficients with 1000 different resamples of the dataset. After that we compute the mean of the coefficients from the 1000 resamples to get a final bootstrap estimation. from now on we'll be using the backward selection model.

From table 1 we can check that the relative difference of the coefficients are very small, even more for the parametric bootstrap estimation. We have a bit more of evidence that the coefficient estimations of the backward model are robust, but before calling it a day we'll construct 95% confidence intervals for each coefficient estimated by the parametric bootstrap.

| **Variable** | **Non-parametric** | **Parametric** |
|---|---:|---:|
| (Intercept) | 0.0059057802 | 1.353963e-04 |
| Year | -0.0072563433 | -8.691415e-05 |
| StatusDeveloping | -0.0105846408 | -1.855687e-03 |
| Adult.Mortality | -0.0013513237 | -1.363692e-04 |
| infant.deaths | 0.0167377595 | 2.401794e-03 |
| Alcohol | -0.0046502410 | -3.849418e-03 |
| percentage.expenditure | 0.0105526385 | 3.104817e-04 |
| BMI | 0.0021548065 | 3.092455e-03 |
| under.five.deaths | -0.0168569355 | -1.978065e-03 |
| Total.expenditure | 0.0162796442 | 2.065403e-02 |
| Diphtheria | 0.0009473427 | 8.416634e-03 |
| HIV.AIDS | -0.0011706620 | -4.508528e-04 |
| thinness.5.9.years | -0.0202867549 | -2.931623e-02 |
| Income.composition.of.resources | 0.0061773736 | 1.311124e-03 |
| Schooling | 0.0053046788 | 1.789651e-04 |

Table 1: Relative difference between the coefficients of the original backward model and the averaged bootstrap estimated coefficients.

For each coefficient we'll check if its confidence interval contains the value 0. If that is so, we'll accept the null hypothesis $H_0 : \beta_i = 0$ and reject the ith variable as a predictor for our model. None of the confidence intervals contain the number 0 so we preserve all the predictors from the backward model.

| **Variable** | **Lower 95% CI** | **Upper 95% CI** |
|---|---:|---:|
| (Intercept) | 217.8 | 401.5 |
| Year | -0.1737 | -0.0817 |
| StatusDeveloping | -1.512 | -0.270 |
| Adult.Mortality | -0.0182 | -0.0144 |
| infant.deaths | 0.0666 | 0.1066 |
| Alcohol | -0.1960 | -0.0673 |
| percentage.expenditure | 0.0003 | 0.0006 |
| BMI | 0.0190 | 0.0441 |
| under.five.deaths | -0.0802 | -0.0507 |
| Total.expenditure | 0.0113 | 0.1758 |
| Diphtheria | 0.0066 | 0.0235 |
| HIV.AIDS | -0.4833 | -0.4147 |
| thinness.5.9.years | -0.1077 | -0.0032 |
| Income.composition.of.resources | 8.927 | 12.076 |
| Schooling | 1.789651e-04 | 1.789651e-04 |

Table 2: Parametric Bootstrap Confidence Intervals

# 5 Conclusions

A complete statistical study has been done to study the relation between the life expectancy of a country and other socioeconomical factors. An exploratory data analysis has been carried out first in order to get some basic concepts and gain an initial insight from the data on our hands. Some of our assumptions were supported by the results of the model study, like the important role that education and schooling plays in the life expectancy of a country, while some others were rejected, specifically the wrong assumption that the GDP per capita would be a decisive factor.

To make a more precise model we performed a stepwise variable elimination to reduce the number of statistically insignificant variables. To validate these models, bootstrapping techniques were applied to construct confidence intervals for the bootstrap estimated coefficients of the predictors and further eliminate any unnecessary predictor.

This analysis has been made in little time, so there's much more room for improvement and I'm aware that some mistakes have been made, and that's why I'll spend some of my time in the future to review it again and keep correcting it.

## 6 Bibliography

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

## 7 Appendix: R scripts

```
#---Load libraries and datasets---#
#https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
install.packages("dplyr")
library(comprehenr)
library(dplyr)
library(MASS)

dataset <- read.csv(file=’Life_expectancy.csv’)

summary(dataset)

#---histograms of the predictors---#
‘%notin%‘ <- function(x, y) !(x %in% y)

variables = to_list(for(variable in colnames(dataset)) if(variable
    %notin% c(’Country’,’Status’)) variable)
print(variables)

for (variable in variables) {
  hist(dataset[[variable]], col = "lightskyblue",
       main = paste("Histogram of", variable),
       xlab = variable, ylab = "Frequency")
}

#---Plots of Life expectancy vs other variables---#
data_2015 <- subset(dataset, Year == 2015,
    c(Country,BMI,Life.expectancy))

plot(data_2015$BMI, data_2015$Life.expectancy,
     main= "BMI vs. Life expectancy",
     xlab= "BMI",
     ylab= "Life expectancy (in years)",
     col= "lightskyblue", pch = 19, cex = 1, lty = "solid", lwd = 2)
#text(dataset$GDP, dataset$Life.expectancy, labels=data_2015$Country,
    cex= 0.7, pos=3)

#---Trying out models. First Linear regression---#
#Here we remove the Country feature since it’s not useful for predicting
    new values
dataset <- dataset %>% select(-Country)

gdp_log_model <- lm(Life.expectancy ˜ log(GDP), data = dataset )
```

```
summary(gdp_log_model)

schooling_lin_model <- lm(dataset$Life.expectancy ~ Schooling,data =
    dataset)

summary(schooling_lin_model)

adult_lin_model <- lm(dataset$Life.expectancy ~ Adult.Mortality, data =
    dataset)

summary(adult_lin_model)

#---Studying model with all variables---#

full_model <- lm(dataset$Life.expectancy ~ GDP + Schooling +
    Adult.Mortality, data = dataset)
summary(full_model)

#---Performing backward selection---#
full_model <- lm(Life.expectancy ~.,data=dataset)
model_back <- stepAIC(full_model, direction = "backward")
summary(model_back)

#---Performing forward selection---#
null_model <- lm(Life.expectancy ~1, data = dataset)
model_forward <- stepAIC(null_model, direction = "forward", scope
      =list(lower=null_model,upper=full_model),trace=TRUE)
summary(model_forward)


#---USING BOOTSTRAPPING TECHNIQUES---#

library(boot)
#---Non-parametric bootstraping---#
bootstrap_fn <- function(data, indices,predictors) {
  dataset <- data[indices, ]
  #We exclude the variables absent from the stepwise models
  dataset <- dataset %>%
      select(-c(GDP,Polio,Measles,Hepatitis.B,Population,thinness..1.19.years))
  fit <- lm(Life.expectancy ~ . ,data=dataset)
  return(coef(fit))
}
results <- boot(data = dataset, statistic = bootstrap_fn, R = 1000)
print(results)

boot.ci(results, type = "perc")

#Calculate the relative difference between the bootstrap estimation and
    the previous backward model estimation
```

```
bootstrap_means <- apply(results$t, 2, mean)
print(abs((coef(model_back)- bootstrap_means))/coef(model_back))

#---Parametric bootstraping
parametric_bootstrap_fn <- function(data, indices) {
  #We exclude the variables absent from the stepwise models
  dataset <- data %>%
      select(-c(GDP,Polio,Measles,Hepatitis.B,Population,thinness..1.19.years))

  # We extract fitted values and residuals
  fitted_values <- fitted(model_back)
  residuals <- residuals(model_back)

  new_y <- fitted_values + rnorm(length(fitted_values), mean = 0, sd =
      sd(residuals))
  new_data <- dataset
  new_data$Life.expectancy <- new_y
  fit <- lm(Life.expectancy ~ ., data = new_data)
  return(coef(fit))
}

parametric_results <- boot(data = dataset, statistic =
    parametric_bootstrap_fn, R = 1000)
#Calculate the relative difference between the bootstrap estimation and
    the previous backward model estimation
parametric_means <- apply(parametric_results$t, 2, mean)

print(abs((coef(model_back)- parametric_means))/coef(model_back))

ci_list <- lapply(1:length(parametric_results$t0), function(i)
    boot.ci(parametric_results, type = "perc", index = i))
ci_list
```