Group Project - Marketing Research

24 March, 2021

Prepared by: Andrea Espinosa, Danielle Kotter, Mireia Salagaray Ruiz

Introduction

In the following project we aimed to analyze the effect of different variables on life expectancy by applying various statistical models and techniques. Hereby we have performed a series of regressions and factors models having as dependent variable, life expectancy. The ultimate objective is to observe whether the correlation between different variables within our data set affects, or not, life expectancy.

The data set

The data used during the project is from the open source platform: data.world. The data set is prepared by the Global Health Observatory (GHO) under World Health Organization (WHO) & United Nations (Kaggle, 2021). It consists of data from 183 countries and includes 22 variables that cover different identifiable factors of countries such as the GDP, alcohol consumption, adult mortality and the average number of years in school. Another important aspect to mention is that the data set also includes the variable years, which goes from 2000 to 2015. Therefore, all data and all other variables are included for 15 years for each country.

Preparing the data

Throughout the preliminary research we found due to the 15 years of data, there would be a high intraclass correlation affecting results of the models. In order to perform a static analysis, an initial subset was made for one year. Primarily the most recent year was applied, being 2015. However, due to the high number of missing values, the year 2014 was more applicable.

Moreover, we additionally found that there were two variables specifically with missing values throughout all years and countries. These were GDP and population and from a total of 139 missing values they had the highest missing share.

In order for these variables to be representative, we decided that the best option would be to find other data sets containing the missing data and merge them. Ultimately 3 different data sets were utilized: the original data set, an additional data set containing population, and another containing GDP. The final data set used throughout the project was obtained by merging these three. Whilst the amount of missing values did significantly reduce, the complete data was still not available. However, this was enough to include them in the analysis.

Variables

The final data set used during the project contains a total of 22 different variables. These are:

Life expectancy	Under-five deaths
Country	Polio
Year	Total expenditure
Status	Diphtheria
Adult mortality	HIV/AIDS
Infant deaths	GDP
Alcohol	Population
Percentage expenditure	Thinness 1-19 years
Hepatitis B	Thinness 5-9 years
Measles	Income composition of resources
BMI	Schooling

Please refer to Appendix 1 for the complete definition of each variable.

Data Problems

In the following section we are considering which variables could potentially have an affect on life expectancy. The variable status defines whether a country is “developed” or “developing”. The data set is composed of 17.9% developed countries and 82.51% developing countries.

Please refer to appendix 2. for visual representation of these results.

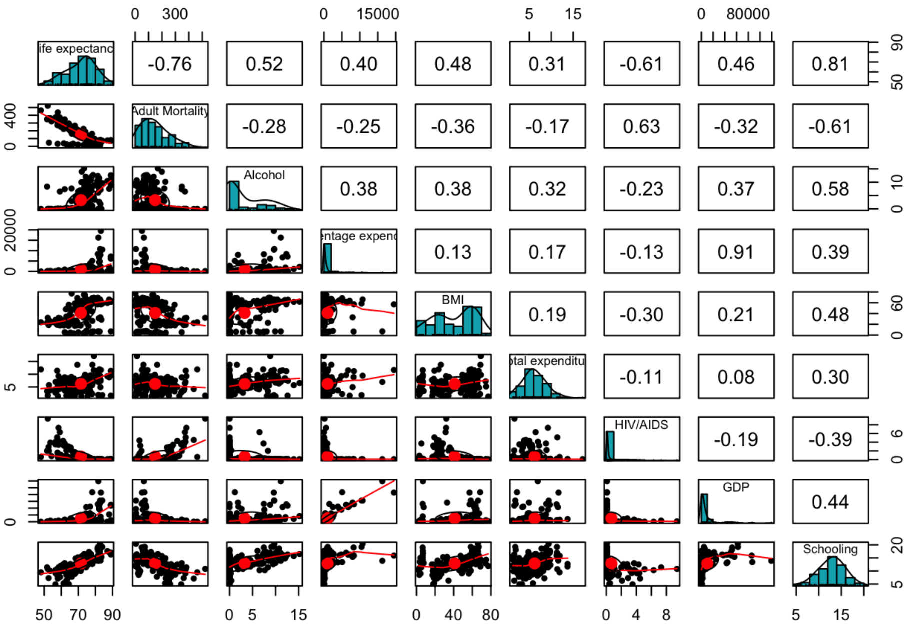
The mean for life expectancy of developed countries is about 81 years old. For developing countries this is equal to approximately 69 years old. Therefore, there is a significant difference of over 11 years.

This fact suggests that the variable status has a clear and strong impact on life expectancy. When performing a non-parametric test, the chi-square test, we set the null hypothesis as there not being a association between the dependent and independent variable. Considering the results of the test had a p-value of < 0.01, we cannot accept the null hypothesis. This implies that indeed there is statistical evidence that the status of the country affects life expectancy.

Moreover, we recognize that this variable is fairly important to the acceptance of models as some variables might affect developed countries that do not necessarily have an effect on developing countries and vice versa. Subsequently, moving forward the analysis will be separated by country status.

Multicollinearity & linearity

The next step was to analyze further the relationship between variables. As we can see in the following figure, we are facing some problems of non linearity. Moreover, we can already identify that there is potential for multicollinearity which will be more thoroughly explained in the next section.



Non Linear Variables

We can observe that there are non linearity problems most evident with variables such as GDP, Percentage Expenditure and Population. To solve this problem we applied logarithms to the three variables. Referring to the example graph in appendix 2., visualizing logging the variables, indeed linearity improved. Therefore, throughout the rest of the analysis, the linearized variables are adopted.

Multicollinearity problems

As demonstrated within the correlation matrix graph showed above, the variables GDP and percentage expenditure are highly correlated with a coefficient of 0.91. This fact generated a problem of multicollinearity which appears because percentage expenditure is the expenditure on health as a percentage of Gross Domestic Product per capita(%).

It is important to be careful when stating that there is a problem of multicollinearity considering when eliminating one of the variables, we can create problems of omitted variables instead. In this case we run a Variance Inflation Factor test and as the results are over five, we can confirm that including the variables jointly in the models would cause a multicollinearity problem.

```
     yBMI   yAlcohol ySchooling    lperexp       lGDP    yIncome 
 1.412383   1.586637   6.340844   8.681341   8.967965   7.485384 
```

Multiple regressions

In order to asses the impact of the variables on life expectancy we performed several multiple regressions. Initially we built 3 different models and adopted those that fit the data as our main model.

In all models Life Expectancy is adopted as the dependent variable, and are formulated as following:

- Model1: Life Expectancy = BMI + Schooling + GDP
- Model2: Life Expectancy = BMI + Schooling + Population + GDP
- Model3: Life Expectancy = BMI + Schooling + Percentage Expenditure

Please refer to appendix 3 for the summary of these regression.

As shown in the results, all of our models are significant and have R-squared coefficients above 68%. However, we are adopting as our main model number 3 using the variable Percentage Expenditure instead of GDP.

According to the World Health Organization, one of the variables that affects life expectancy the most is the total health expenditure made by countries. However, it can vary drastically between countries, for instance the expenditure per capita in some countries can be 300 USD while in others only 30 USD. Subsequently, there are many outliers identified.

Therefore, we believe it is more accurate to use percentage expenditure as it reflects more the reality of investment on health systems. The reality of each country is different, so the percentage expenditure will also depend on the population structure and epidemiological needs of each specific place, not only on GDP. Naturally, model 2 including both GDP and population caused multicollinearity.

The cooks distance test assisted us to identify possible outliers in the accepted model. The results showed that there were multiple outliers that potentially had an affect on the regression line. After excluding the outliers, the model improved slightly, displaying even more reliable results.

The summary of the results of the model before and after removing the outliers and the cooks distance test can be found in Appendix 3.

Interpretation

Body mass index. This variable BMI has a positive impact on life expectancy. We can consider that body mass index is related to a better nutrition. It is important to comment that too high levels of body mass index can imply problems with obesity leading to further complications. In this instance, the relation with life expectancy should be negative. However, on average having good nutrition and a healthy weight especially in the developing countries is logical to have a positive effect on life expectancy.

Schooling. Schooling is a less intuitive variable. However, there are studies that corroborates our results. There exists a correlation between education and health suggesting that education policies can also be seen as indirect health policies and might be correlated with better lifestyle. Education helps individuals to develop health-related resources, for instance eat better quality food.

In order to prove the strength of the variable status on the association with years of schooling, we plotted the variable differentiating between developed and developing countries. It can be easily identified that status separates countries in two clear distributions and that as expected those in developing countries attend school for a shorter amount of time.

Please refer to appendix 2. for the illustration of the difference in years of schooling.

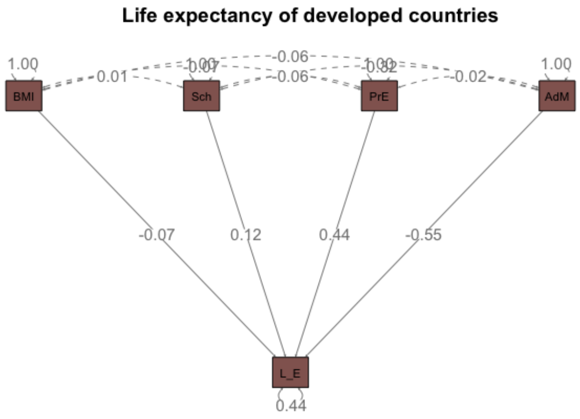
Percentage expenditure. Finally, percentage expenditure on health also has a positive impact on life expectancy which is the results we expected. However according to some studies, at some point it is not only how much money you spend but how effectively organized the system is. Research suggests that it is more important or it has a higher impact investing on social protection, because these are contributions targeted to individuals to provide support during circumstances which adversely affect their welfare.

All the variables included in each regression have a statistically significant impact on life expectancy at a significance level of 1%.

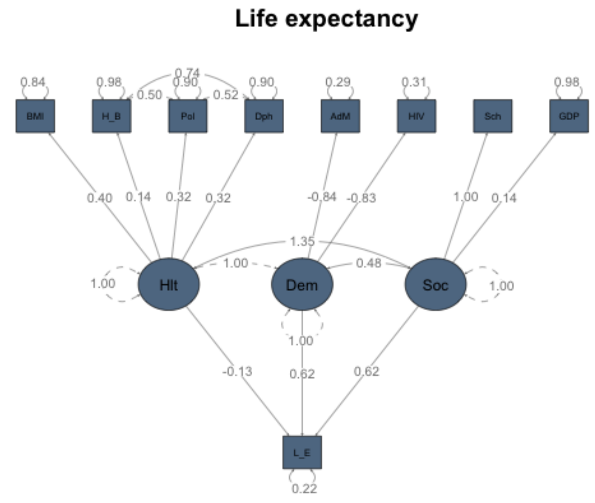
Structural equation model & factor model

Considering the limitations of the multiple regression models and the characteristics of the data set, the theory of structural equation models has been applied. For reasons previously explained, separate models are applied for developed and developing countries.

Starting with developed countries, we have not been able to design a effective model. Whilst some variables significantly influence life expectancy, there is no way to fit the model so that it can be accepted. The variables that do have an significant affect, are percentage spend on health and adult mortality. According to a paper researching the factors impacting life expectancy in western countries some common are: disease of older age such as Alzheimer, Pneumonia and cardiovascular disease, excessive smoking and diabetes (Veena S. Raleigh, 2019). We have to conclude that the variables that we do have in our data set, are not suitable for designing a model for life expectancy in developed countries.



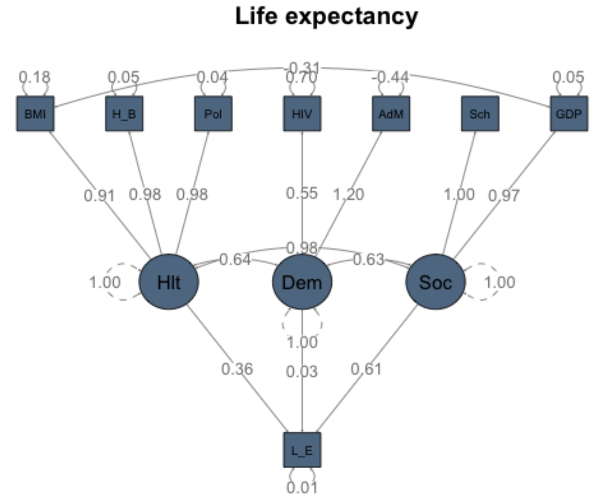
However, for developing countries the selected variables are highly correlated with the dependent. The ultimate model is inspired by a research paper that looks at factors affecting life expectancy in Asian countries (Moon Fai Chan, 2012). It is a complex model that contains three latent variables: health, demographic and socio-economic factors. The latent variable health contains: BMI, Hepatitus B, Polio and Diphtheria immunization. The demographics or so-called mortality includes the variables adult mortality and deaths of HIV/Aids. The socio-economic factor is based on the GDP and the years of schooling.



The result of this model can be found below. All variables included have significant impact at a significance level of 10%. It has 21 degrees of freedom and the null hypothesis can be rejected at a p-value of 0.87. Consequently, we can accept this model.

Chi-square test	DegreesFreedom	P-value	Rmsea
14.03	21	0.868	0.00

Simultaneously, another model similar to the above has been created by including the data of the years 2010-2015 whilst clustering on the variable country to account for the intraclass correlation. Although all variables are significant, at this time the model is not fitted and therefore cannot be accepted. The model has many constraints and is quite complex and the solution might be out of the scope of the research project.



Alternative factors

We also have to consider less continuous quantifiable factors that influence life expectancy. These are some examples that are not included in the data set that we have available that would significantly impact specific moments in time. These are for example: political uncertainty, a pandemic, a war or financial hardship.

To provide an example, in the USA, in February they have learned that the life expectancy in 2020 has fallen by one year alreadyas a consequence of the COVID-19 pandemic. Moreover, there are some things that should be further analysed such as the 7 year drop here in Spain around 2007 because it could be a consequence of the hard financial crisis that the country experienced. Another clear observation can be found in Syria where life expectancy drops drastically by approximately 10 years when the war started in 2011.So, we can intuit that this can be a direct consequence of the war.

Please refer to appendix 4. for the illustration of the evolution of life expectancy in different countries.

Limitations

There are a several limitations that may affect the results of this report.

First of all, we do not have clear knowledge of the criteria used in some variables such as the country status.

Moreover, despite the fact that we improved the final data set, we still had missing values particularly in GDP and population so, this can lead to inaccuracies in some of the results.

Another important aspect is that some of the countries included in the data set may not have the infrastructures to collect data in an accurate way, so data from some countries may not be totally reliable.

Conclusion

Appendixes:

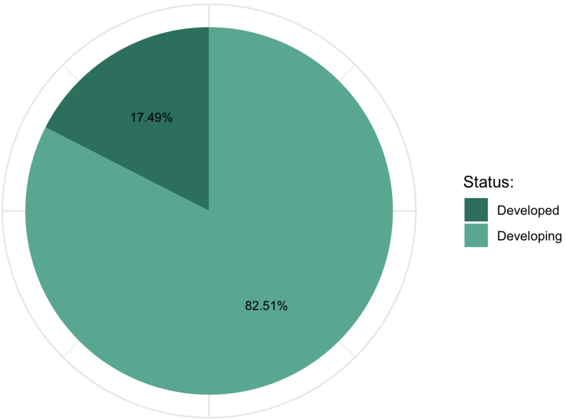
Appendix - 1: Data descriptions

The variables included in the data set are:

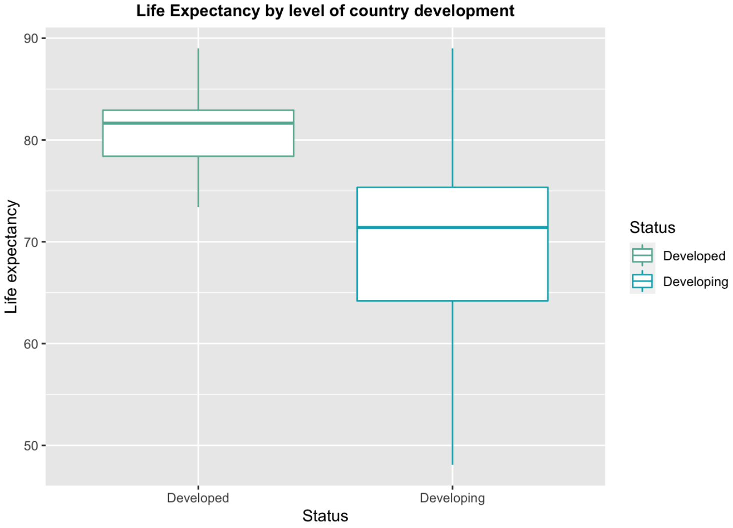
Field	Description
Dependent variable: Life expectancy	Life Expectancy in age
Country	Country
Year	Year
Status	Developed or Developing status
Adult Mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
infant deaths	Number of Infant Deaths per 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	Income composition of resources
Schooling	Number of years of Schooling(years)

Appendix - 2: Plots

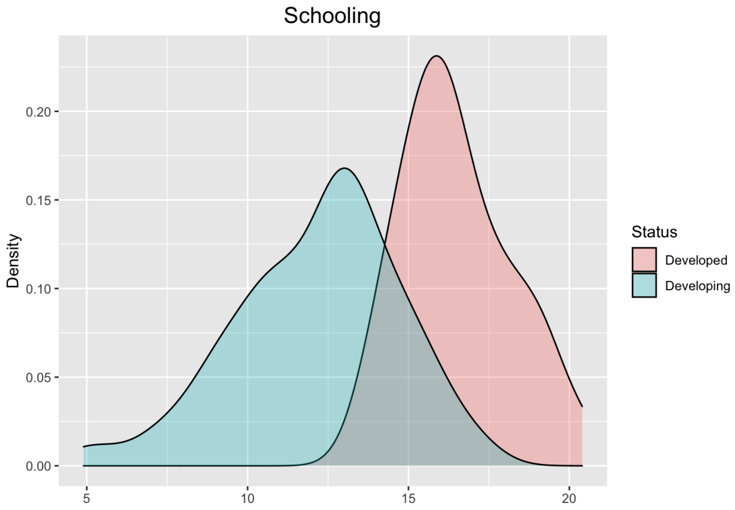
Plot 1: Percentage of developed and developing countries in the data set



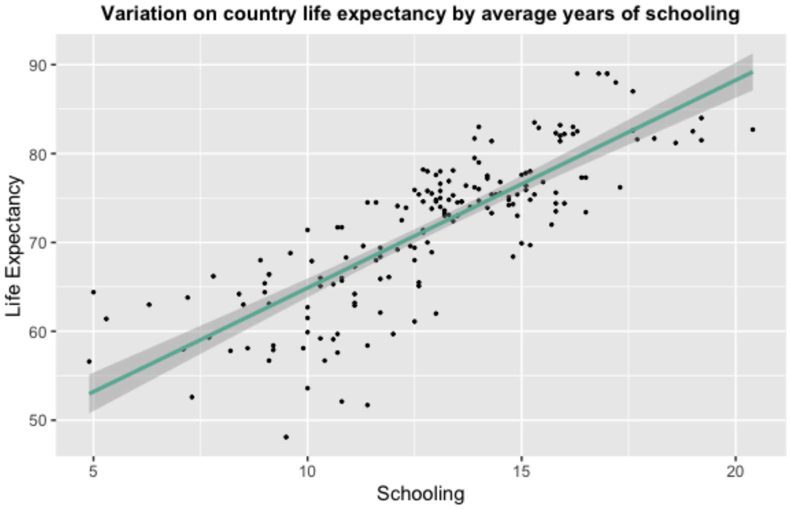
Plot 2: life expectancy by country status



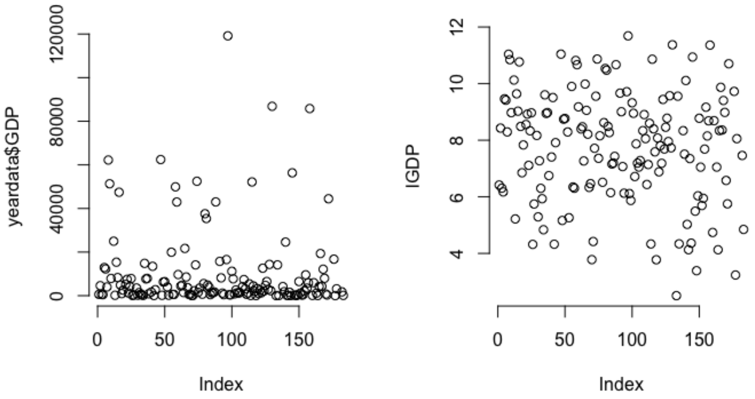
Plot 3: Schooling differentiated by country status



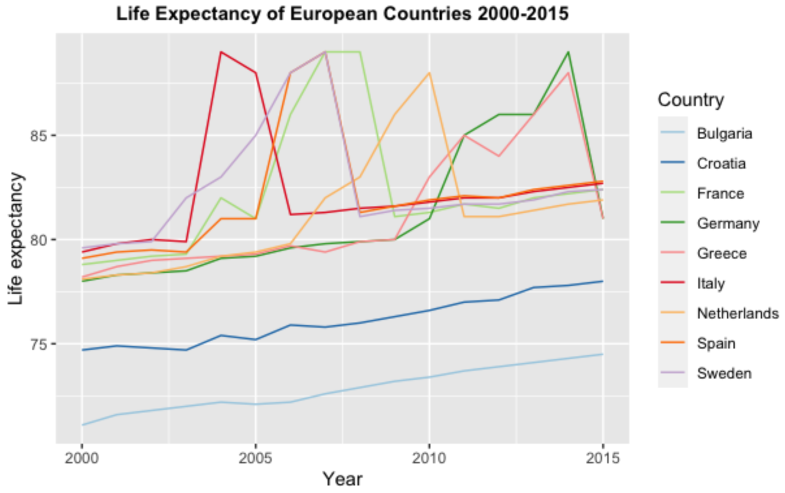
Plot 4: Scatter plot years of schooling - life expectancy.



Plot 5: Example of results from linearizing variables such as GDP.



Plot 6: Variation between European countries

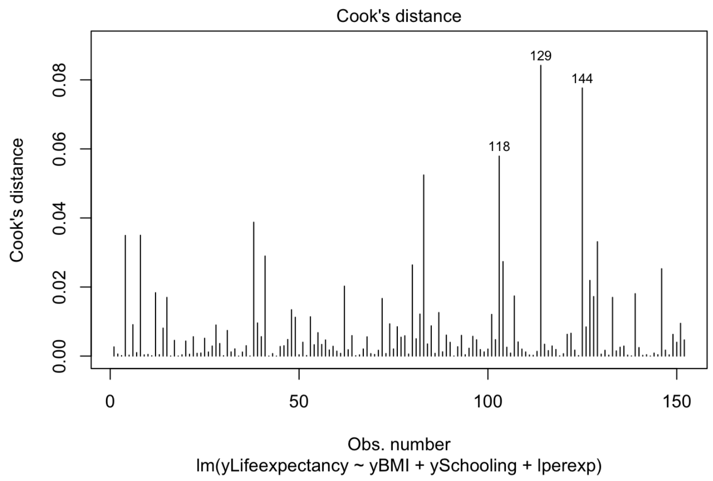


Appendix - 3: Regressions

Creation of 3 different models for multiple regressions to see the impact of different variables on life expectancy

```
=============================================================================================
                                               Dependent variable:
                    -------------------------------------------------------------------------
                                                 yLifeexpectancy
                              (1)                      (2)                     (3)
---------------------------------------------------------------------------------------------
yBMI                        0.049**                  0.042*                  0.044**
                            (0.022)                  (0.022)                 (0.022)

ySchooling                  2.134***                1.719***                 2.124***
                            (0.175)                  (0.235)                 (0.185)

lNYPopulation                                       -1.047**
                                                     (0.413)

lNYGDP                      0.474**                 1.397***
                            (0.202)                  (0.406)

lperexp                                                                      0.503**
                                                                             (0.231)

Constant                   30.119***                29.597***               39.465***
                            (4.367)                  (4.361)                 (1.888)

---------------------------------------------------------------------------------------------
Observations                  151                      150                     152
R2                           0.692                    0.706                   0.688
Adjusted R2                  0.686                    0.698                   0.682
Residual Std. Error     4.871 (df = 147)        4.787 (df = 145)         4.918 (df = 148)
F Statistic         110.143*** (df = 3; 147) 87.107*** (df = 4; 145) 108.781*** (df = 3; 148)
=============================================================================================
Note:                                                             *p<0.1; **p<0.05; ***p<0.01
```

Cook distance test to detect outliers

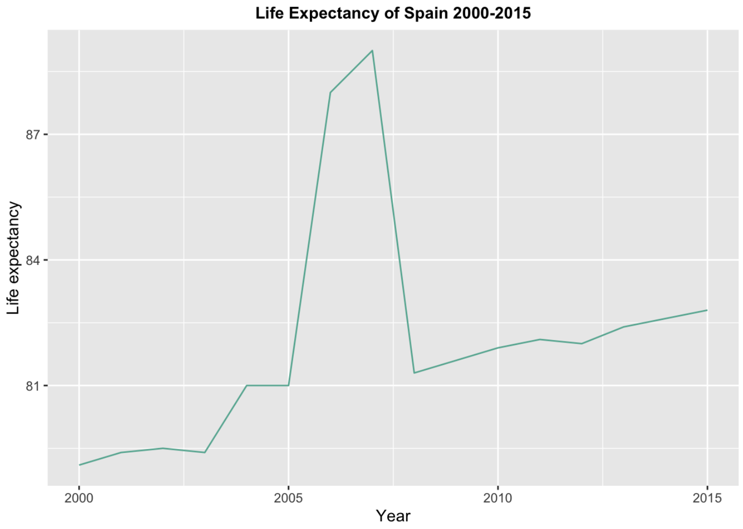


Model 3 multiple regression before and after removing the outliers

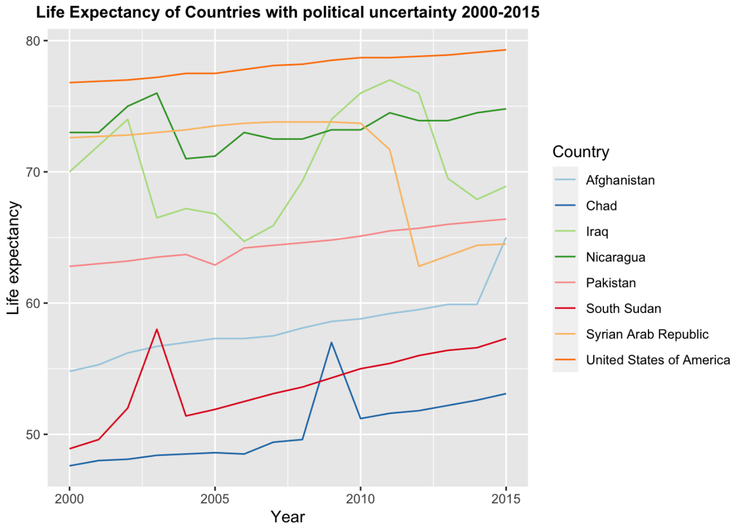
```
=====================================================================
                                   Dependent variable:
                    -------------------------------------------------
                                     yLifeexpectancy
                              (1)                      (2)
---------------------------------------------------------------------
yBMI                        0.044**                  0.053**
                            (0.022)                  (0.021)

ySchooling                  2.124***                 2.105***
                            (0.185)                  (0.182)

lperexp                     0.503**                  0.444**
                            (0.231)                  (0.224)

Constant                   39.465***                39.626***
                            (1.888)                  (1.881)

---------------------------------------------------------------------
Observations                  152                      149
R2                           0.688                    0.694
Adjusted R2                  0.682                    0.688
Residual Std. Error     4.918 (df = 148)         4.706 (df = 145)
F Statistic         108.781*** (df = 3; 148) 109.839*** (df = 3; 145)
=====================================================================
Note:                                     *p<0.1; **p<0.05; ***p<0.01
```

Appendix - 4: Evolution of life expectancy in different countries

Evolution of life expectancy in Spain



Evolution of life expectancy in countries with political uncertainty



References

https://journals.sagepub.com/doi/pdf/10.1177/1010539512454163

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

https://www.oecd-ilibrary.org/social-issues-migration-health/trends-in-life-expectancy-in-eu-and-other-oecd-countries_223159ab-en