Analyzing Life Expectancy

Menghi Michael, Simbari Raffaele

January 2024

1 Abstract

The following project deals with data regarding worldwide life expectancy in years ranging from 2000 to 2015, together with other statistics which describe the socio-economical conditions of such countries. The following sections will employ a variety of statistical tools to try and gain a better insight on how these important data interact.

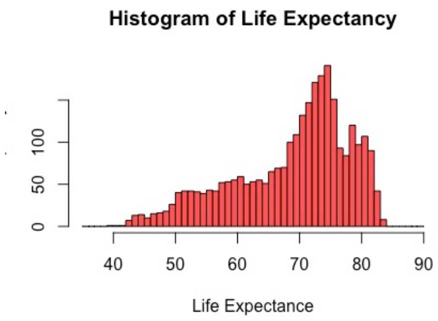
2 Dataset

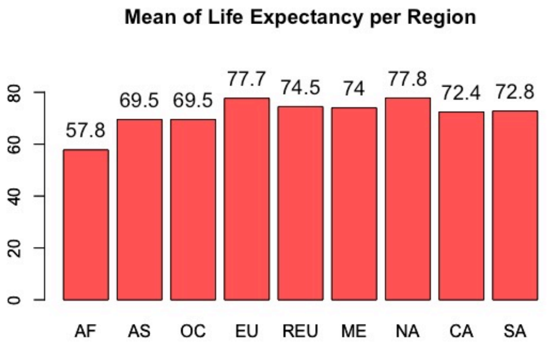
We will start by presenting our dataset and the basic features of its variables; the data are collected from official sources, which we do not mention for brevity (they are however reported in the Kaggle page of the dataset). Our dataset is composed of the following variables:
COUNTRY
REGION
YEAR
INFANT_DEATH: infant deaths per 1000 people;
UNDER_FIVE_DEATH: deaths of children under 5 per 1000 people;
ADULT_MORTALITY: deaths of adults (age 15-60) per 1000 population;
ALCOHOL_CONSUMPTION: liters of pure alcohol per capita with 15+ years old;
MEASLES, POLIO, HEPATITIS B, DIPHTERIA: percentage of immunization against the disease among 1-year-olds;
BMI: Body Mass Index;
INCIDENTS_HIV: occurences of HIV per 1000 people aged 15-49;
GDP_PER_CAPITA: gdp per capita in USD;
POPULATION_MIL: population in millions;
THINNES: Prevalence of thinness among children aged 5-9 years. BMI ¡ -2 standard deviations below the median. Data about age 5-9 and 10-19.
SCHOOLING: average years people aged 25+ spent in school;
ECONOMY_STATUS: categorical variable describing whether a country is developing or developed;
LIFE_EXPECTANCY: average life expectancy (both genders).

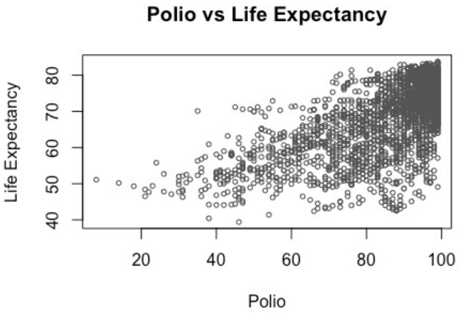
3 Visualization

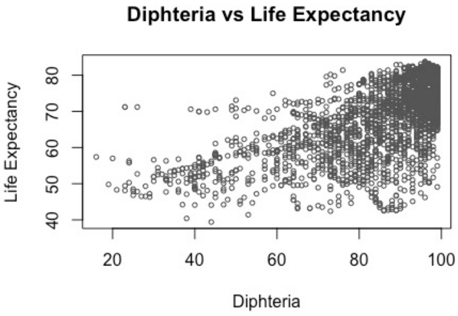
First of all, we proceeded to control whether the dataset needed to be cleaned. We noticed there was not need for cleaning, and then we read on Kaggle that the people who published the dataset already took care of basic cleaning, with methods of statistical imputing, like filling in missing values with averages of ”close” data. Moreover, countries with more than 4 attributes missing were eliminated in order to modify the data as little as possible. Hence, we began our analysis, starting with some plotting, to start visualizing the relationships underlying our data. To help visualizing the distribution of life expectancy, we did an histogram and a barplot (divided by

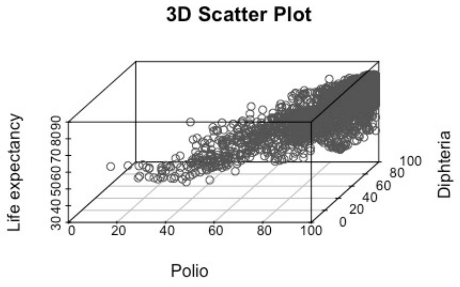
geographic region). Next, we display scatterplots of Life expecatncy compared to some of the predictors; in particular, we plotted Life expectancy vs schooling, polio, diphteria (and then a 3d plot of polio and diphteria together), adult mortality, infant deaths and GDP per capita. These plots immediately show strong correlations between some variables: in particular, we see an important negative correlation between adult mortality, infant deaths and life expectancy, which we will talk about extensively later. On the other hand, the relationship between GDP per capita and Life Expectancy is more complicated than one might think: from the graph, we are able to see that medium and high GDP per capita result in high life expectancy, while at low GDP per capita the relationship becomes troubled (and not very clear from the plot which, maybe, contains too much data to clearly identify a pattern). However, saying that in richer country life expectancy is higher seems supported by our data. These allegations on GDP are supported by empirical evidence and quite a lot of research articles (some information will be linked below, in the bibliography).
Moreover, the plots show a positive correlation between immunization and life expectancy, and between schooling and life expectancy. This was certainly expected: immunization against serious diseases like Polio and Diphteria decreases the probability of dying. Also, it is true that the immunization coverage is generally higher in developed and rich countries (as one may check plotting Economy Status vs Immunization for diseases), where life expectancy is higher. This also explains the correlation between schooling and life expectancy: years spent in school are obviously higher in richer countries.

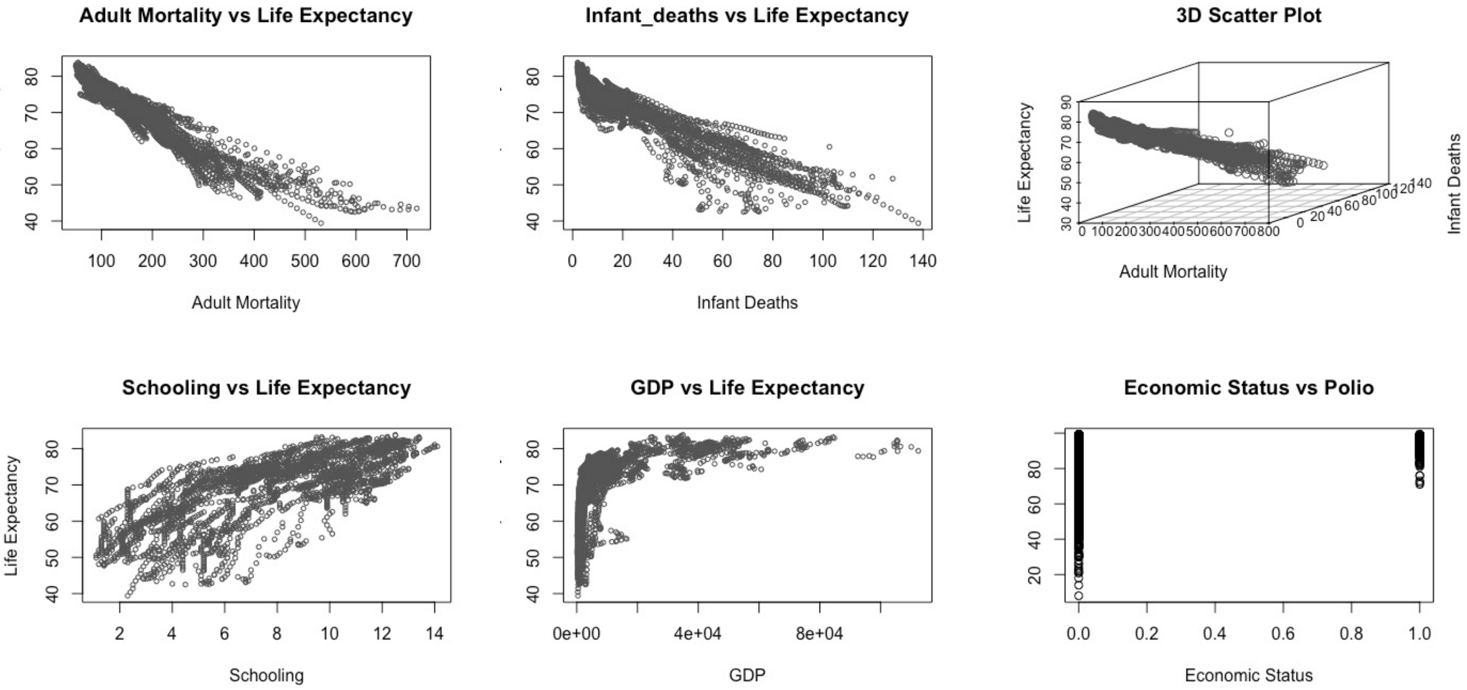

Histogram of Life Expectancy


Mean of Life Expectancy per Region


Polio vs Life Expectancy


Diphteria vs Life Expectancy


3D Scatter Plot



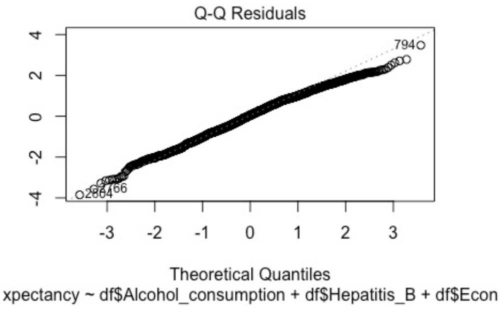
4 Critical Discussion of our Dataset

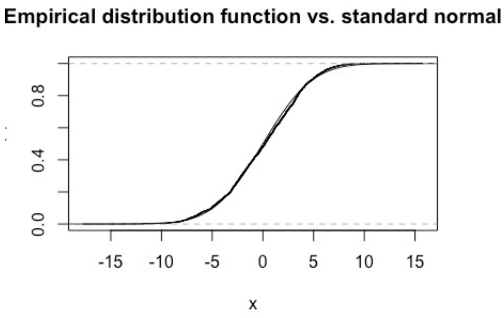
Before moving on to our analysis, we would like to highlight something that will affect our regression analysis. As stated before while displaying the plots, there is evidence of a strong correlation between the predictors Infant Deaths and Adult mortality, and the target variable Life Expectancy. We suspected that these parameters are somehow used in Life Expectancy calculation, and we wanted to confirm these suspects, for explaining a value with a predictor which is used in its calculation is not so meaningful. After a quick research, we found that life expectancy is actually calculated from life tables, i.e. the probability for a person of a certain age to die before his next birthday. These life tables are computed using statistical methods, i.e. using the death rates at any specific age. Even if at first this seem to be quite different from our data, both Infant Deaths and Adult Mortality (and actually even the predictor Under Five Deaths) essentially represent death rates before a certain age (Infant Deaths is the number of child dying before age 1 per 1000 live births, while Adult Mortality essentially is the death rate for the entire age range 15-60). Given this far too strong relations, we decided to exclude this data from our regressions, as it would not be meaningful to analyze them. So, in the next section regarding regression analysis we will just have a quick glance at a model containing these three predictors, to further confirm the discussion we have been carrying out, and then we will move on.

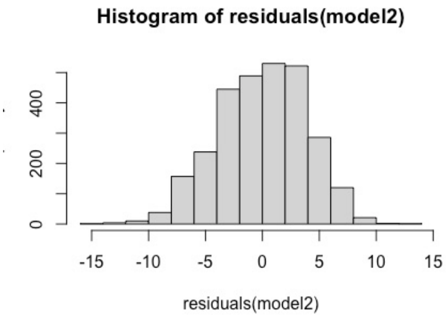
5 Regression Analysis

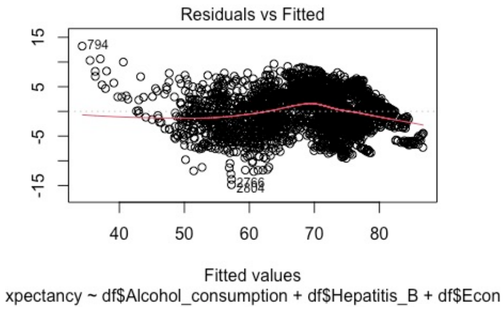
Now it is time for regression analysis. As anticipated before, the first thing will be constructing a model containing all predictors, even the three we were discussing before, just to be sure that they are indeed far too powerful, for the reasons explained above. We will not include summaries of our models in the paper, so we kindly ask to refer to the R code. We will perform a Multivariate Linear Regression on our dataset, Life Expectancy being the target variable, including all predictors. To be precise, we specify we left out right from the start variables such as Year, Country and Region, which wouldn't be of any use in a Multiple linear regression. The

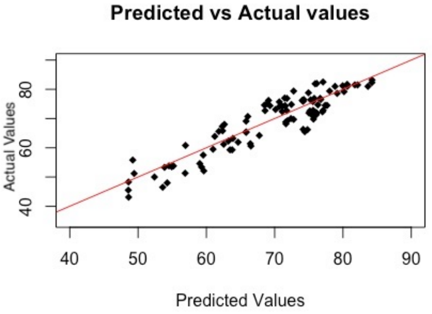
summary of these complete linear model, indeed, yields an R^2 of almost 0.98, meaning the model with all predictors is able to explain 98% of the variance in our data. One could additionally check that a model containing only these three predictors would practically perform the same. This confirms our statement that these three variables yield a not meaningful model.
Hence, we move on, and we try to construct a more significant model. We still carry out a Multivariate Linear Regression, and this time we select all predictors that are meaningful (hence we exclude the three infamous death rates). The resulting model yields an R^2 of 83%, meaning that this fraction of the variance of our data is explained by our model. Next we should check the assumptions of normality and homoscedasticity of the residuals. Analyzing the graphs that summarizes the model, we notice that the distribution of the residuals is a bit skewed, and that the Residual vs Fitted graph shows some problems concerning homoscedascity. On the other hand, linearity seems respected, and the QQ-plot and the Empirical distribution fits more nicely compared to a Normal. Running a Kolomogorov-Smirnov test, we detect a deviation from normality.
We can try to explain some of these ”violations” with the high sensitivity of Kolmogorov-Smirnov, together with the behaviour of some data, in particular the ones having life expectancy smaller than 50 years (refer to the graph Residual vs Fitted). Apart from these (and some other bad-behaving data around 60 years old of life expectancy), homoscedasticity assumption would be much more reasonable. Although not entirely happy, we proceed and try to improve our model.











6 Model Selection

We decided to run model selection algorithms to try and obtain a better model. Specifically, we proceed by a stepwise approach to variable selection. Unfortunately, all the model selection algorithms (Step-up, Step-Down, Step-Both) end up outputting the starting model, i.e. the one containing all the ”meaningful” predictors. This means the best model is the starting one, the one containing all ”meaningful” predictors.

7 Hypothesis Testing

Finally, we shall proceed tackling a research question with hypothesis testing. We decided to perform two hypothesis tests: one, more in tune with the rest of the paper, regards life expectancy, while the other one deals with infant deaths. Indeed, we were performing some calculations on our dataset, and, computing the averages of child mortality in different regions, we noticed fairly big differences (even where we didn't expect them), and so we decided to proceed with a test. We do not include the summary of the test here, so please refer to the R code.

Regarding the first test, we wanted to check whether different regions have significantly different Life Expectancy. We relied on the asymptotic t-test, since we cannot assume same variance between the different data. We focus on the regions "Rest of Europe" and "Middle East" . The t-test yields a p-value of 2.41e-13 hence rejecting the null hypothesis. So there is enough statistical evidence to conclude there is a difference between average life expectancy in these two region. For the second test, we check significant differences in Infant Deaths, still relying on Asymptotic t-test. We focus on the regions "European Union" and "North America". The t-test finds a p-value of 5.276e-7, hence rejecting the null hypothesis: there is enough evidence to conclude there exist a significant difference.

8 Conclusions

In the end, we can say that our results are not entirely satisfying. Even if in the beginning we were very positive about the explanatory power of our model, and the assumption are not "horribly" violated, it is more than clear that our linear model has some limitations, primarily concerning the behaviour of residuals. However, we can say that this limitations were expected: our dataset was composed of a wide variety of very "delicate" data, and expect to fully explain the relationship between with a linear regression model would be optimistic at best (if not pretentious). Indeed, indexes such as Life expectancy are extremely multifactorial, and even where one would expect to find strong correlations, there might not be any, or there might be non-trivial ones, that a linear model is not capable of capturing (GDP is a perfect example in this sense).

There are various strategies one can use to try and address this difficulties. A few example include:

- exploiting more complicated models, to try and capture more subtle relationships between data (returning to the GDP example, some research articles, together with empirical evidence, suggest a logarithmic kind of relationship between life expectancy and GDP per capita) ;

- trying to explore smaller portions of the dataset, to see if a linear correlation becomes stronger only in certain ranges of life expectancy. This idea relies on some of the plots we displayed earlier, since we noticed that observations with low life expectancy (< 50) seem to offset general tendencies;

- enriching the dataset with other predictors, to give a more comprehensive overview. Examples might include indicators of life quality, air pollution data and public health expenditure. Moreover, the diseases contained in the predictors might be updated: for instance, as fatal and serious Polio has been in the past, we can safely say that, thanks to immunization, nowadays it does not really impact on life expectancy. Data regarding leading causes of death of our times (ischaemic heart disease, stroke, lung cancer among the others) may reveal important correlations.

9 Bibliogrphy

1. https://en.wikipedia.org/wiki/Preston_curve

2. https://www.who.int/data/gho/indicator-metadata-registry/imr-details/65

3. ttps://www.researchgate.net/publication/327445254_The_Effects_of_Health_Care_Expenditures_as_a_Percentage_of_GDP_on_Life_Expectancies

4. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death