Influence of Different Factors on the Lifespan of a Population

Multivariate Statistics

María Cristina Carmona Fernández

December 2023

Contents

1 Summary

The life expectancy problem is a multifaceted issue influenced by various factors such as immunization and economic factors .

To analyze and address this issue, statistical methods such as Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Principal Component Analysis (PCA), and Factor Analysis (FA) are used.

PCA helps uncover underlying patterns and reduce dimensionality, offering insights into relationships among influencing variables. Meanwhile, FA explores latent factors that explain observed correlations, revealing hidden determinants impacting life expectancy.

The project purpose is to investigate life expectancy and understand the influence of diverse factors. It involves dimensionality reduction by identifying key indicators and essential unobservable variables. Moreover, this project aims is to create a classification model for evaluating population life expectancy levels based on the identified indicators.

2 Introduction

The life expectancy problem is a multifaceted issue influenced by various factors such as demographic variables, income composition, mortality rates, immunization, human development index, social and economic factors. A common issue is, through these factors, predict the average lifespan of a population with a perspective of increasing it.

To adress this problem it is recommended to carry out a preliminary exploratory analysis of the data contained in the data set. To do this, the different numerical and graphical techniques are going to be applied. At first, it is necessary to focus on the analysis of each of the variables independently, performing an univariate analysis.

To analyze and address these complexities, first the assumptions underlying the application of the different multivariate dimension reduction techniques, such as PCA or FA have to be checked . After checking this, said techniques will be applied.

To conduct the analysis statistical methods such as Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Principal Component Analysis (PCA), and Factor Analysis (FA) will be analyzed. And also, LDA and QDA will be performed to assist in discerning how different factors discriminate between groups with varying life expectancies, capturing both linear and non-linear relationships.

PCA helps uncover underlying patterns and reduce dimensionality, offering insights into relationships among influencing variables. Meanwhile, FA explores latent factors that explain observed correlations, revealing hidden determinants impacting life expectancy.

The purpose of this project is to examine life expectancy and analyze how various factors impact it. In this context, the goal is to carry out dimensionality reduction by identifying indicators that best describe the data and the most relevant unobservable variables. Additionally, the objective is to develop a classification model that assesses the level of life expectancy of a population based on these indicators.

3 Materials and Methods

3.1 Materials

The dataset is taken from the Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The data is from year 2000-2015 for 193 countries. It contains the following variables:

- Country: Country Name.
- Year: Year.
- Status: Developed or Developing.
- Life Expectancy: Life expectancy in age.
- Adult Mortality: Probability of dying between 15 and 60 years per 1000 population.
- Infant Deaths: Number of infant deaths per 1000 population.
- Alcohol: Recorded per capita consumption (in litres).
- Percentage Expenditure: Expenditure on health as per GDP (%).
- Hepatitis B: Immunization coverage among 1-year-olds (%).
- Measles: Number of reported cases per 1000 population.
- BMI: Average BMI of the entire population.
- Under-Five Deaths: Number of under-five deaths per 1000 population.
- Polio Immunization: Coverage among 1-year-olds (%).
- Total Expenditure: Government expenditure on health as a percentage of total government expenditure (%).
- Diphtheria: Immunization coverage among 1-year-olds (%).
- HIV/AIDS: Deaths per 1000 population.
- GDP: Gross Domestic Product (GDP) per capita (USD).
- Population: Population of the country.
- Thinness $10-19$ Years: Thinness among children from age $10-19$(%).
- Thinness $5-9$ Years: Thinness among children from age $5-9$ (%).
- Income Composition of Resources: Index ranging from $0-1$.
- Schooling: Number of years of schooling.

The following table of the quantitatives variables provides the main measures of position, dispersion, and shape most relevant in these data:

Now the qualitative variables of the dataset are studied:

Basic Descriptive Statistics						
Variable	Min.	1st Q.	Median	Mean	3rd Q.	Max.
Year	2000	2004	2008	2008	2012	2015
Life expectancy	36.30	63.20	72.00	69.22	75.60	89.00
Adult mortality	1.0	74.0	144.0	164.8	227.0	723.0
Infant deaths	0.0	0.0	3.0	30.3	22.0	1800.0
Alcohol	0.010	1.093	4.160	4.603	7.390	17.870
Percentage expenditure	0.000	4.685	64.913	738.251	441.534	19479.912
Hepatitis B	1.00	80.94	87.00	80.94	96.00	99.00
Measles	0.0	0.0	17.0	2419.6	360.2	212183.0
BMI	1.00	19.40	43.00	38.32	56.10	87.30
Under five deaths	0.00	0.00	4.00	42.04	28.00	2500.00
Polio	3.00	78.00	93.00	82.55	97.00	99.00
Total expenditure	0.370	4.370	5.938	5.938	7.330	17.600
Diphtheria	2.00	78.00	93.00	82.32	97.00	99.00
HIV AIDS	0.100	0.100	0.100	1.742	0.800	50.600
GDP	1.68	580.49	3116.56	7483.16	7483.16	119172.74
Population	$3.400e+01$	$4.189e+05$	$3.676e+06$	$1.275e+07$	$1.275e+07$	$1.294e+09$
Thinness $10-19$ years	0.10	1.60	3.40	4.84	7.10	27.70
Thinness $5-9$ years	0.10	1.60	3.40	4.87	7.20	28.60
Income Composition of Resources	0.0000	0.5042	0.6620	0.6276	0.7720	0.9480
Schooling	0.00	10.30	12.10	11.99	14.10	20.70

Table 1: Table of descriptive statistics for the quantitative variables in the dataset.

In the variable *Country* a smaller sample has been taken for some countries, which are: Cook Islands, Dominica, Marshall Islands, Monaco, Nauru, Nive, Palau, Saint Kitts and Nevis, San Marino and Tuvalu, as they have a smaller population. The basic stadistics for this variable can be seen in the following table:

Basic Descriptive Statistics			
Variable	Frequency	% Valid	Total
Countries with more frequency	16	0.545	0.545
Countries with lesser frequency	1	0.034	0.034

Table 2: Table of descriptive statistics for the *Country* variable.

For *Status* variable:

Basic Descriptive Statistics					
Variable	Frequency	% Valid	% Valid Cum.	Total	Total Cum.
Developed	512	17.43	17.43	17.43	17.43
Developing	2426	82.57	100.00	82.57	100.00

Table 3: Table of descriptive statistics for the *Status* variable.

3.2 Statistical Methods

Firstly, a preliminary exploratory analysis of the data was conducted to identify potential missing values and outliers, which were then addressed. For variables with more than 5% missing values, the random pattern was analyzed by studying homogeneity within groups with other variables without missing data, using a *Student's t-test* since the variables are quantitative. In cases of extreme values identified through boxplots, the decision was made to replace them with the mean, given the quantitative nature of the variables.

Secondly, a classic numerical descriptive analysis was performed, providing the main measures of position, dispersion, and shape to better understand the data.

Thirdly, various multivariate statistical techniques were applied, checking the necessary assumptions for each:

- Verification of correlation between the data: Population-level justification was done using the *Bartlett's sphericity test*, to determine if correlations are significantly different from zero. Sample-level justification involved examining the correlation matrix, the polychoric correlation matrix, and other graphical representations.
- Once the necessary assumptions were verified, different multivariate analysis techniques were applied. In this case, Principal Component Analysis was conducted to reduce dimensionality through observable variables; Factor Analysis was used to identify latent variables (unobservable) highly correlated with a group of observable variables and practically uncorrelated with the rest.
- Univariate normality verification: A graphical exploration of the normality of individual distributions was done using histograms and qqplots. The final conclusion was obtained through a hypothesis test, specifically the *Shapiro-Wilks test.*
- Multivariate normality verification: This was studied using the *Henze-Zirkler hypothesis tests*. It is important to note that multivariate normality can be affected by the presence of multivariate outliers, but an analysis of these outliers has already been conducted.

Lastly, Linear and Quadratic Discriminant Analyses were performed to establish a classification method for new observations.

4 Results

Beginning with the study of the missing values and it's observed that 14 variables have missing values. As 7 of them have more than 5% missing values, for these variables a *Student's t-test* is performed and the result that is obtained is that for the variables Alcohol, Total expenditure, GDP and Population the $p - value > 0.15$, therefore, the hypothesis of homogeneity is accepted, and it is concluded that the pattern is random, it is chosen to replace the NA with the mean.

For the remaining variables that are being studied, which are Hepatitis B, Income composition of resources and Schooling, $p - value < 0.15$, so homogeneity cannot be assumed. In this case,this would have to be discussed with the researcher who poses the problem under analysis because they should neither be eliminated nor replaced, but since in this case it is not feasible, it is decided to act as in the case of a random pattern, so in this case, they are replaced by the median.

The observation of the presence of some outliers is delved into. To address them, it has been decided to replace them with the mean of the other values.

Moving forward, attention shifts to the study of the correlation between the variables. The *Bartlett's test* has provided evidence to reject the hypothesis of independence of the data. Thus, it has been concluded that there is correlation between the data, and therefore, dimensionality reduction can be considered through Principal Component Analysis or Factor Analysis. Additionally, in graphical representations of the correlation, it has been observed that there are 4 to 5 groups of latent variables with high correlation with a group of observable variables and low correlation with the rest.

After all this, through the corresponding tests, it has been concluded that there isn't univariate normality for most variables conditioned to each modality of the qualitative variable studied in Discriminant Analysis. Despite rejecting the null hypothesis of normality for two of them, the analysis continues. Moreover, the lack of multivariate normality has been concluded.

Now, performing a Principal Componente Analysis it is obtained that from the initial 22 variables, a set of 6 observable variables has been reached, which are linear combinations of the originals and accumulate almost 70% of the explained variance.

In the following graph, it can be observed that the first 6 principal components accumulate almost 70% of the explained variance.

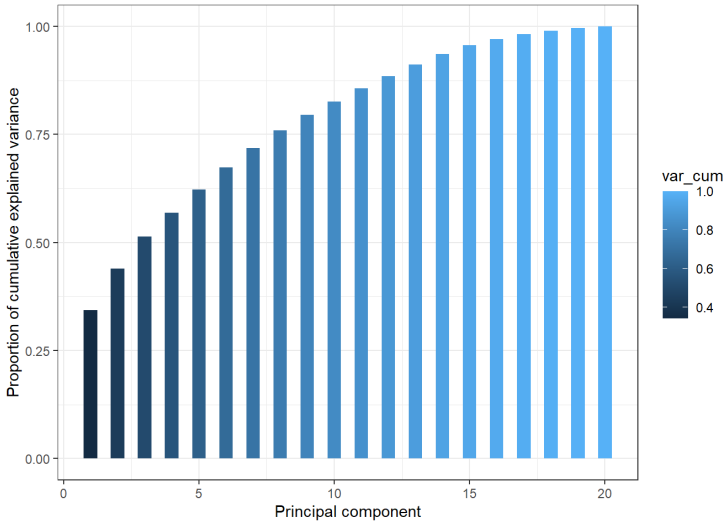

Figure 1: Variance explained by the Principal Components.

Afer PCA, Factor analysis has been conducted, and the analysis, as indicated by a Scree plot, determined that 5 latent factors are sufficient to explain the data. In the subsequent graphs, said Scree plot and the way factors correlate can be seen:

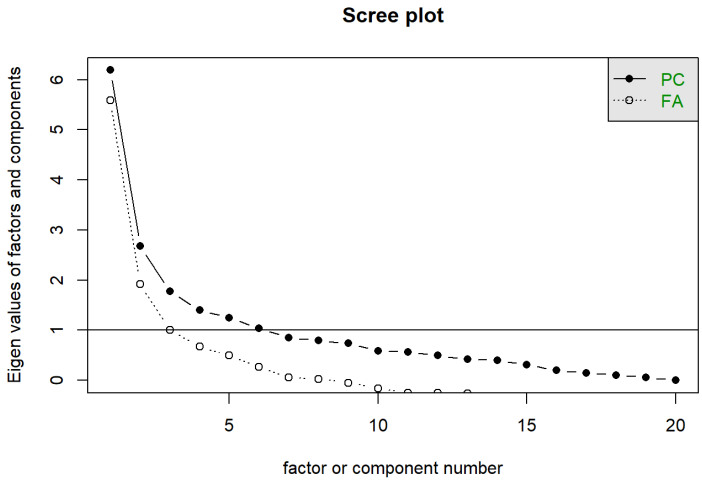

Figure 2: Scree plot to select the latent factors.

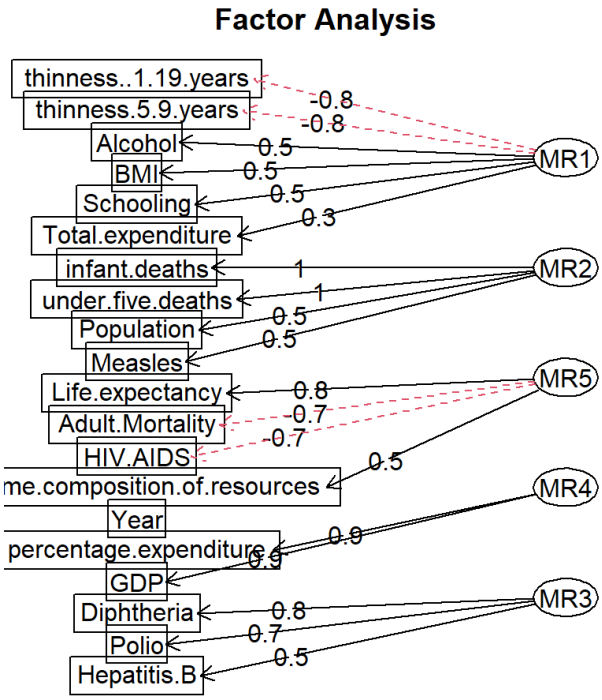

Figure 3: Results of Factor Analysis.

Finally, a categorical variable has been defined, and both Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) have been conducted. The response variable has been defined based on *Life expectancy*, with the other variables being used as predictors. The obtained linear discriminant classification model is given by the following expression:

$$
\begin{aligned}
\text{Odds} =& 1.132738 \times 10^{-2} \times \text{Year} + 4.340997 \times 10^{-3} \times \text{Adult.Mortality} \\
& - 1.262522 \times 10^{-2} \times \text{infant.deaths} - 1.970419 \times 10^{-2} \times \text{Alcohol} \\
& - 9.283862 \times 10^{-4} \times \text{percentage.expenditure} - 1.582141 \times 10^{-2} \times \text{Hepatitis.B} \\
& - 2.458324 \times 10^{-3} \times \text{Measles} - 4.536101 \times 10^{-3} \times \text{BMI} \\
& - 1.064769 \times 10^{-3} \times \text{under.five.deaths} - 3.381613 \times 10^{-3} \times \text{Polio} \\
& - 3.877981 \times 10^{-2} \times \text{Total.expenditure} - 3.881977 \times 10^{-3} \times \text{Diphtheria} \\
& + 5.853773 \times \text{HIV.AIDS} - 1.935186 \times 10^{-5} \times \text{GDP} \\
& - 1.312976 \times 10^{-8} \times \text{Population} - 8.493223 \times 10^{-2} \times \text{thinness..1.19.years} \\
& + 1.001592 \times 10^{-1} \times \text{thinness.5.9.years} - 4.656544 \times \text{Income.composition.of.resources} \\
& - 6.379008 \times 10^{-2} \times \text{Schooling}.
\end{aligned}
$$

In the case of LDA, a classification method with an error rate of 10.892% has been obtained, and in the case of QDA, a classification method with a 10.449% error rate has been achieved.

5 Discussion

This project's aim was to explore life expectancy, analyzing the impact of various factors and subsequently performing dimensionality reduction and classification modeling. Let's assess the achievement of these goals based on the obtained results.

In view of the correlation, where Bartlett's test rejected the hypothesis of independence and graphical representations revealed 4 to 5 groups of latent variables highly correlated, PCA was performed, reducing the initial 22 variables to 6 observable variables, explaining nearly 70% of the original variance.

After PCA, Factor Analysis determined that 5 latent factors were sufficient to explain the data. A Scree plot aided in selecting these factors.

For a classification model a categorical variable was defined, and both Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) were conducted. The linear discriminant classification model provided insights into the influence of various factors on life expectancy with the objective of, given new data, being able to predict the lifespan of a population.

About the overall achievement, the project successfully delved into the multifaceted aspects of life expectancy, addressing missing values, outliers, and exploring intricate correlations. Dimensionality reduction techniques, including PCA and Factor Analysis, provided a comprehensive understanding of the underlying patterns. The development of a classification model showcased the practical application of the identified indicators.

6 Conclusion

In conclusion, this project successfully achieved its goals of exploring life expectancy and analyzing the impact of various factors through an statistical approach.

The strengths of the work lie in the systematic handling of missing values, robust treatment of outliers and the application of advanced techniques such as Principal Component Analysis (PCA), Factor Analysis. These methodologies allowed for an understanding of the relationships within the dataset.

However, certain limitations are presented. The replacement of missing values and outliers with mean and median, while practical, could introduce biases. Additionally, the classification models have been performed without the hypothesis of multivariate normality and have demonstrated error rates that should be addressed.

To make this work better, more advanced methods to deal with missing values and find outliers should be performed. And different models should be explored to make the predictions stronger and more accurate. Since data changes over time, it might be helpful to study trends over a longer period for a better understanding of life expectancy.