Exploring Global Life Expectancy through the use of Socioeconomic Data to perform Factor Analysis and Multiple Linear Regression

La Mar Holmes
Binghamton University, lholmes6@binghamton.edu

Abstract - **This research paper presents an analysis of WHO-collected data on life expectancy and other socioeconomic indicators for 179 countries between 2000 and 2015. The techniques used for the exploration of the socioeconomic data were Factor Analysis and Multiple Linear Regression. With Factor Analysis, the data was narrowed down to two latent factors: Public Health Infrastructure and Nutrition/National Production. Using the solutions found with Factor Analysis, a Regression model was constructed with the computed factor scores. The regression model demonstrated that improvements to Public Health Infrastructure and reductions in disease had a substantial impact on Life Expectancy. Examples of policy improvements for vaccinations and youth nutrition are discussed, but not necessarily recommended without further study. Further analysis is recommended to explore how changes to the latent factors could improve the length of life.**

Keywords- Life Expectancy, Factor Analysis, Multiple Linear Regression, Socioeconomic

Introduction

Life expectancy refers to the average number of years an individual or group of individuals can expect to live within a city, nation, or specific region [5]. Life expectancy serves as both an indicator of the socioeconomic status of the population. It is dependent on a multitude of other indicators that have implications for increasing or decreasing the number of years of life. Many nations and studies break down these indicators and attempt to learn which specific factors have the most significant impact [1,2,5]. In studies like Miladinov's, researchers examined GDP per capita and infant mortality as indicators to determine how socioeconomic changes in those areas affect life expectancy [1]. In contrast, Kim and Kim's study identified indicators such as population, national income, schooling, and the number of internet users [2]. So far, there doesn't seem to be a set requirement for which indicators are necessary to determine what to include or not include when measuring impact on Life Expectancy (LE).

In this study, the data is pulled from the WHO and aggregated in Kaggle by the content provider Lasha [3]. While Kaggle is a data science competition platform, the dataset sources are published for review, allowing for secondary evaluation, which was conducted before use within this paper's analysis. Within the dataset, there is a diverse range of data, including, but not limited to, socioeconomic indicators involving GDP, disease infection rates, youth mortality rates, and nutrition-related measures. This data also includes these socioeconomic performance indicators for fifteen years, 2000 to 2015. The sheer amount of data and variability of formats within the set allows for multiple ways of analysis and interpretation, which could lead to different conclusions. In this study, Factor Analysis and Multiple Linear Regression were employed to provide a simplified view of the impact of various indicators and to predict outcomes based on changes to those variables.

According to Sharma, "in the behavioral and social sciences, researchers need to develop scales for the various unobservable constructs such as attitudes, image, intelligence, personality, and patriotism.[4]" In this research paper, developing scales to determine latent factors affecting life expectancy is necessary. Factor Analysis is a technique used to describe the underlying correlations among socioeconomic variables.

To supplement the discovery of latent factors, a linear regression model was built upon the Factor Analysis to assist in predicting future Life Expectancy numbers. This method enables the reduction of variables into uncorrelated factors, resulting in a more simplified model that should facilitate easier interpretation of the prediction.

After simplification and model creation, the next step is to analyze the driving factors that impact the length of life, how these factors can be used to predict changes, and what measures could be taken to improve those numbers. This paper is not intended to propose socioeconomic improvements, but rather to identify potential areas for future analysis.

Methods

Data Review and Preparation

To begin the process of analyzing the Life Expectancy and Socioeconomic data, it was necessary to review the data that came from the Kaggle Life Expectancy (WHO) Fixed dataset. The data within the dataset were aggregated from

multiple sources, including the WHO database and the Global Health Observatory, a platform operated by the WHO [3]. Within the data, twenty-one columns were available for analysis, but a review was necessary to determine which ones to include. The columns were initially categorized by their respective data types, as detailed in Table 1, to facilitate clarity in the initial decision-making process.

Table 1.
Data Categories and Columns

Data Category	**Column(s) in WHO dataset**
Grouping Information	Country, Economy_status_Developed, Economy_status_Developing
Time Information	Year
Dependent Variables	Life Expectancy
Independent Variables	Infant_deaths, Under_five_deaths, Adult_mortality, Alcohol_consumption, Hepatitis_B, Measles, BMIPolio, Diphtheria, Incidents_HIV, GDP_per_capita, Population_mln, Thinness_ten_nineteen_years, Thinness_five_nine_years, Schooling

The first step was to review the data for unnecessary duplicates. It was found that there were duplicate entries, specifically in the numbers, but these were left in place in case there were no changes from year to year. The next task done was to review the data for columns with more than five percent of the data missing. Due to the dataset being revised by the author, any missing data had previously been filled in from multiple sources. With the revisions done, it is assumed that the data has been cleaned enough for a certain level of analysis.

After reviewing the data, the formats of each column were considered to determine their applicability and necessity for Factor Analysis, the first step in the actual analysis. Based on the idea of not moving beyond the standard method of performing Factor analysis, grouping or categorical data would not be included: Country, Year, Region, Economy_status_Developed, and Econonomy_status_Developing. Due to some of those columns being necessary for a portion of the analysis, only Region, Economy_status_Developed, and Economy_status_Developed were initially removed.

After preparing the analysis-ready file, the data was passed to a Jupyter notebook, where Python was used for averaging, Factor analysis, and multiple linear regression.

Libraries Used, Averaging, and Scaling the Country Health Indicators

Once the data was ready, the Python libraries pandas, numpy, sklearn, factor_analyzer, matplotlib, statsmodel, and seaborn were imported for the different processes in the analysis. Pandas was used to group the data where necessary and manipulate different columns. Sklearn was utilized for its StandardScaler functionality to scale the column values into a more manageable range, as well as for obtaining the multiple linear regression models. Factor_analyzer was used for performing the necessary factor analysis techniques. Matplotlib was used for plotting heatmaps, scree, and scatter plots. Lastly, the statsmodels library was used to do a deeper review of the regression results.

The analysis-ready file includes fifteen years' worth of data for each column previously mentioned. An evaluation was conducted, and it was decided that each country's non-categorical data would be averaged. This would allow for a better overall understanding of what each country's indicators were like over a considerable period. There was also an evaluation of the idea of doing a regression analysis on each country, then taking the model to map the observations to predictive points, and then taking the average of each country's predicted results. The expected and averaged results would then be used for Factor Analysis, which would later be used for creating a regression model. This approach was rejected because it could lead to overfitting the data to the observations, ultimately making the data harder to interpret.

After averaging the data, it was scaled because each column had different units of measurement; for instance, some data were monetary metrics, while others were percentages or population multiples. With the scaling done, it ensures that no single indicator has a disproportionate influence on the derived factor solution.

Factor Analysis and Multiple Linear Regression

To initiate the Factor Analysis process, a correlation matrix was created as a heatmap to examine the partial correlations between the indicators. Also, the Kaiser-Meyer-Olkin (KMO) was measured to determine if the numbers were sufficient to proceed with factor analysis. If the KMO value is high enough, proceed with starting the Factor Analysis and choose a temporary number of factors for preliminary computations. After reviewing the data, the formats of each column were considered to determine their applicability and necessity for Factor Analysis, the first step in the actual analysis. Based on the idea of not moving beyond the standard method of performing Factor analysis, grouping or categorical data would not be included: Country, Year, Region, Economy_status_Developed, and Econonomy_status_Developing. Due to some of those columns being necessary for a portion of the analysis, only Region, Economy_status_Developed, and Economy_status_Developed were initially removed. Factor Analysis enables the plotting of the Scree Plot, allowing for a visual determination of the elbow and selection of the appropriate number of factors. With the factor count finalized, the factor loadings can be computed, and analysis of the latent factors can begin. These factor loadings can be used later to calculate the factor scores of observations.

To perform a regression analysis, observation points are needed, so the factor scores were calculated based on the respective countries within the original data. Once the

scores were calculated, the regression analysis was performed to provide the intercept and coefficients to create a model. Lastly, to determine if the model was viable, the R^2 and Adjusted R^2 were computed. Both Factor Analysis and Multiple Linear Regression were performed using Python and the previously mentioned data analysis libraries.

Results

Factor Analysis

Factor Analysis of the data started with plotting a heatmap of the Correlation Matrix. What was found was that for the most part, the partial correlation was decently low across most of the variables, as can be seen in Figure 1. Some indicators showed high correlation, which would speak to the probability of being able to group variables into homogeneous sets or factors[4].

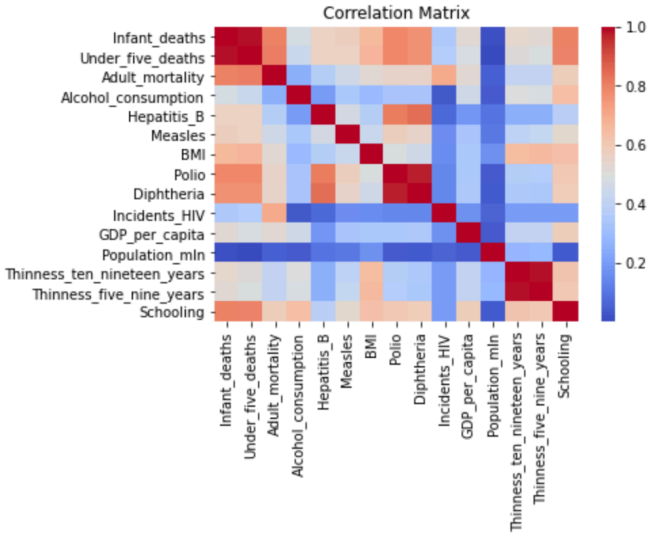

Figure 1.
Correlation Matrix

The subsequent analysis done before running the Factor Analysis was Kaiser's measure of overall sampling adequacy. When Keasure-Meyer-Olkin (KMO) was measured, it returned an overall value of 0.84, which, according to the suggestion of Kaiser and Rice, would be considered a "Meritorious" value [4]. With a value reaching a satisfactory level, analysis proceeded.

When running the Factor Analysis, the necessary number of factors was unknown, so a Scree Plot was created to find the elbow of the line and choose the number of factors based on that. As can be seen in Figure 2, the scree plot indicates that the elbow occurs at the second factor.

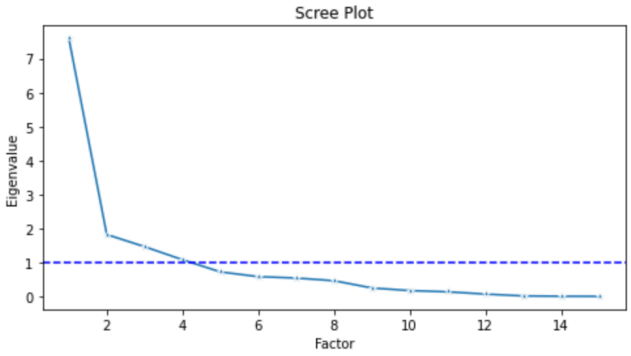

Figure 2.
Scree Plot

Taking insight from the plot, two factors were chosen for the factor solution, which would mean that the variables were reduced from fifteen variables to two. The loadings, as presented in Table 2, enable one to gain an understanding of the basic underlying descriptions of the factors of Factor 1 and Factor 2.

Table 2.
Factor Loadings

Indicators	**Factor 1 (F1)**	**Factor 2 (F2)**
Infant deaths	-0.853	0.422
Under Five Deaths	-0.859	0.390
Adult Mortality	-.0659	0.349
Alcohol Consumption	0.304	-0.479
Hepatitis B	0.740	-0.003
Measles	0.562	-0.304
BMI	0.470	-0.565
Polio	0.935	-0.102
Diphtheria	0.933	-0.076
Incidents HIV	-0.257	0.172
GDP per Capita	0.373	-0.429
Population mln	0.035	0.234
Thinness_ten_nineteen_years	-0.238	0.910
Thinness_five_nine_years	-0.233	0.903
Schooling	0.611	-0.588

From what can be seen with Factor 1, the indicators that are having the most significant impact are as follows:

Infant deaths, Under five deaths, Adult mortality, Hepatitis B, BMI, Polio, Diphtheria, and Schooling. In Factor 2, the indicators with the largest values are Thinness_ten_nineteen_years, Thinness_five_nine_years, Schooling, BMI, Infant deaths, Alcohol Consumption, and GDP per Capita. For both factors, a cutoff of 0.40 was used to determine what variables were having a pattern of impact; this was based on one of the cutoff suggestions in Sharma's text [4].

Between both of the factors, it was calculated to have a cumulative variance of 0.592 or about 60% of the variance. This number isn't particularly high, but there could be many unseen elements that are affecting this score, such as other necessary indicators. It should also be noted that this analysis focuses on socioeconomic indicators, which may make it more challenging to achieve higher results in accounting for variability.

The final step in the factor analysis process was to plot the observations based on their new axes. Returning to the averaged observations per country, a factor score was derived for each and plotted in the chart in Figure 3.

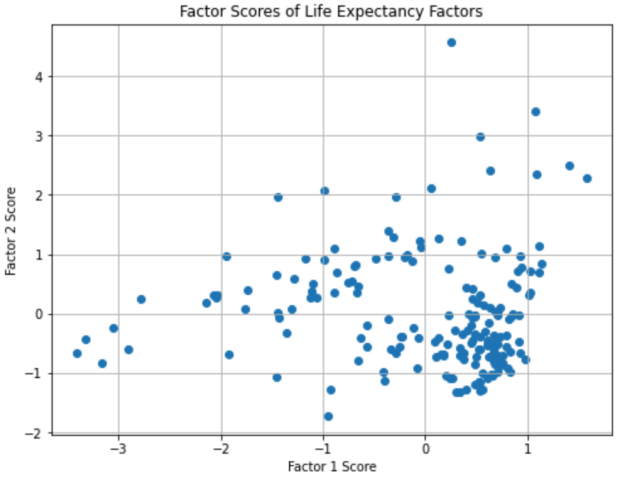

Figure 3
Factor Scores of Life Expectancy Factors

Multiple Regression Model

Taking the factor scores that were found from applying the country specific observations, a linear regression model was created. The resulting multiple regression model seen in Eq. 1 shows the intercept at 68.86, the Factor 1 coefficient at 7.17, and the Factor 2 coefficient at -3.92.

$$LE = 68.86 + (7.17*F1) + (-3.92*F2) + \epsilon \quad (1)$$

Based on the coefficients, it was found that Factor 1 has a larger impact on the outcome of Life Expectancy. Looking at the individual effect of each predictor variable (F1 and F2), both had p-values less than 0.05, which means that both of the new variables have a statistically significant impact on the number of years included in the Life Expectancy prediction.

The last portion of the linear regression process involved checking both the R^2 and Adjusted R^2 values. The analysis revealed that both values were approximately 0.78, indicating that the regression model accounts for 78% of the variance in predicting a nation's average length of life..

Discussion

Upon reviewing the factor loadings of the indicators within each factor, it became clear that some indicators have more significant impacts than others. These would be the areas to look at specifically when deciding on socioeconomic changes that are necessary to improve the length of life. For instance, polio and Diphtheria are life-threatening infections that significantly impact life expectancy. Therefore, it is recommended to conduct further research on the causes of these infection rates and the necessary measures to reduce them. Examining these types of diseases wouldn't be the end of that type of investigation, but rather a starting point. In Wirayuda and Chan's writings, they found that the kind of disease that dominates varies by nation, depending on their socioeconomic status [5]. The goal is to identify patterns that reveal the impacts of the data points.

It should also be reiterated that, in the process of reviewing and analyzing data, averaging globally would be the best step forward for an overall analysis. The Factor Analysis was based on global data, not nation specific data, which leaves room for error when predicting how a nation's particular indicators would impact latent factors. In turn, that nation's specific factors might have factor scores that lead to different coefficients and intercepts than what was derived in this study. This being said, with methods used for analysis, the solutions will not be overfit to any specific region in the world, so they should allow for an overall understanding of what nations around the world should be striving towards, agnostic to the particular internal or external circumstances that cause the variance with their specific data.

Given that the factor analysis is generalized globally, the two latent factors can be described, and future analysis can then be conducted to identify potential systemic changes that could improve Life Expectancy numbers. Based on the highest impacting indicators for Factor 1, the analysis leans towards a latent factor that encompasses public health infrastructure, including vaccination rates for various diseases, and the associated educational aspects. For the second Factor, the highest impacting indicators describe an underlying factor relating to the general nutrition, healthy habits, and GDP of the nation. For clarity purposes, Factor 1 could be named Public Health Infrastructure, and Factor 2 would be named Nutrition and National Production.

Once the factor scores were found and the multiple regression model was derived, it should allow for an easier prediction of what Life Expectancy should approximately be based on the two latent factors previously mentioned. Upon further examination of the model, it becomes clear that for Factor 1, countries' efforts to enhance their public health infrastructure, through policies like increasing vaccination rates, lead to improved life expectancy for their citizens. On

the other hand, for Factor 2, there is a negative 3.92 year multiplier, which would decrease the expected length of life. This would point to the idea that countries with high levels of malnutrition are shrinking life expectancy. The model, with an adjusted R2 of 78%, indicates a reasonably accurate prediction model. The goal of this prediction model is not to guarantee easy decision making, but to facilitate quick evaluations of simple changes.

Conclusion

To solidify the goal of this research paper, we aimed to examine socioeconomic factors that could serve as indicators of life expectancy in countries. In the case of the analysis, the choice to reduce the dimensions of variables to a multi-factor model would allow for further understanding of the correlation between indicators to determine if there are underlying policy connections or a lack thereof that lead to specific results. From what could be gathered from this first round of analysis, the first place for nations to start looking for areas of improvement is in their Public Health Infrastructure, as well as in the nutrition of the general public, but focusing on the youth of the nation.

Another key takeaway from the Factor analysis was identifying the specific socioeconomic indicators with the most considerable impact, thereby informing the initial steps to be taken towards improvement. For example, in Factor 1, specific disease infection rates have the largest effect, so it would be beneficial for countries to take measures to lessen the rates of infection through vaccination campaigns. Another example of high impact socioeconomic indicators is shown in Factor 2, where thinness in the youth indicators is having a significant effect. Policy makers could take this as a suggestion to make sure that the younger citizens are provided with the proper nutrition to maintain their health.

The regression model was created to provide a starting point for predicting how small changes may have a significant or negligible impact on the average lifespan. Future studies on the subject could involve analyzing minor adjustments to policies in different areas and determining their impact early by plugging the expected numbers into the regression model.

References

[1] Miladinov, G. (2020). Socioeconomic development and life expectancy relationship: evidence from the EU accession candidate countries. Genus, 76(1), 2.

[2] Kim, J. I., & Kim, G. (2016). Country-level socioeconomic indicators associated with healthy life expectancy: income, urbanization, schooling, and internet users: 2000–2012. Social Indicators Research, 129(1), 391-402.

[3] Life Expectancy (WHO) Fixed. 2023. Lasha. Kaggle. https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated

[4] Sharma, S. (1995). Applied multivariate techniques. John Wiley & Sons, Inc..

[5] Wirayuda, A. A. B., & Chan, M. F. (2021). A systematic review of sociodemographic, macroeconomic, and health resources factors on life expectancy. Asia Pacific Journal of Public Health, 33(4), 335-356..

Author Information

La Mar Holmes is a Product Manager with over twelve years of experience in web development, solutions engineering, product management, and analytics. He holds a B.S. in Graphic Information Technology, an M.B.A. in Information Technology Management, and an M.S. in Engineering Management. He is currently pursuing a Ph.D. in Systems Science at Binghamton University