Life Expectancy Analysis

George Matlis
Department of Computer Science
Aristotle University of Thessaloniki
Thessaloniki, Greece
gmatl@csd.auth.gr

Charalampos Moutafidis
Department of Computer Science
Aristotle University of Thessaloniki
Thessaloniki, Greece
cmoutafid@csd.auth.gr

***Abstract*—Despite an extensive body of scientific research dedicated to discerning the factors impacting life expectancy—exploring demographic variables, income composition, and mortality rates—certain pivotal elements, notably immunization and the Human Development Index, have been relatively understudied. To bridge this gap, the Global Health Observatory (GHO) was established and is currently overseen by the World Health Organization (WHO). The GHO's primary mission is to systematically gather and monitor comprehensive data pertaining to health status and other pertinent indicators across all nations. Through this research, we seek to understand how various indicators affect life expectancy and what their statistical impact is.**

***Index Terms*—Life Expectancy, Data Analysis, Statistics,**

I. Introduction

Life expectancy, a critical metric reflecting the average lifespan of individuals within a population, has been the subject of considerable scientific inquiry. Previous research has extensively delved into demographic variables, income composition, and mortality rates to unravel the intricate web of factors influencing life expectancy. Despite these efforts, certain key determinants, notably immunization and the Human Development Index, have remained relatively underexplored in the existing body of knowledge. Addressing this gap in research, the Global Health Observatory (GHO) emerged as a pivotal initiative, currently under the guidance of the World Health Organization (WHO). The GHO is dedicated to systematically collecting and monitoring comprehensive data on health status and relevant indicators across nations, enhancing our understanding of the multifaceted dynamics that contribute to variations in life expectancy worldwide.

II. Dataset and Data Exploration

A. Dataset

The goal of this study is to investigate the statistical impact of various indicators on the life span of several countries. Specifically, the dataset we are exploring consists of the following indicators/variables:

- Country: The name of the country
- Year: The year corresponding to the life expectancy measurements
- Status: Whether a country is developed or developing
- Life Expectancy: The life expectancy of a country
- 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 1000 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: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- Schooling: Number of years of Schooling. This indicator can be represented by three categories:

  1) Low - Schooling years less than or equal to 8
  2) Medium - Schooling years ranging from 9 to 12
  3) High - Schooling years greater than 12

The first two indicators in the dataset, namely *Country* and *Year*, will not be included in our analysis as they are not meaningful to the target indicator, which is life expectancy.

B. Data Pre-Processing

Having described the dataset and the various indicators, some pre-processing of the data is required. First, we start by transforming the indicators *Status* and *Schooling* into factors, as their values represent categories. Then, we compute their frequency matrix by counting the number of samples for each of their categories. In Table I, it is evident that the majority of countries fall within the developing stage, with only a few classified as developed. Turning to Table II, a notable observation emerges—most countries exhibit low levels of education years, contrasting with the minority of nations that boast a more substantial average of school years.

TABLE I
COUNTS FOR THE STATUS INDICATOR

	No. Samples	Total Sample Rate (%)
Developed	512	17
Developing	2426	83

TABLE II
COUNTS FOR THE SCHOOLING INDICATOR

	No. Samples	Total Sample Rate (%)
Low	323	12
Medium	954	34
High	1498	54

Our logical assumption is that citizens of developed countries would have a longer life span compared to citizens of developing countries and, thus, a higher life expectancy. This assumption is predicated on the determination of such countries to invest in their healthcare, thereby reducing the number of deaths related to general health issues, like the prevalence of thinness among children and teenagers, or reducing the possibility of death from numerous diseases like those mentioned in the dataset. To test this hypothesis, two preliminary steps need to be examined. The initial step would be to examine whether the target indicator, the *Life Expectancy*, closely approximates the normal distribution. In Figure 1, the graphical depiction of the distribution for the *Life Expectancy* reveals a deviation from the normal distribution, despite the close similarity between the mean $x = 69.2$ and median $m = 72.1$ values.

The subsequent step involves assessing whether the variances within the two categories of the *Status* indicator are equal. The Flinger-Killeen test indicated unequal variances $p - value < 0.001$. The Flinger-Killeen test, a non-parametric method for assessing the homogeneity of group variances based on ranks, proves valuable in situations where the data exhibit non-normal distribution or when challenges associated with outliers in the dataset cannot be adequately addressed. Subsequently, we employed the Mann-Whitney test, a widely-used nonparametric (distribution-free) method for comparing outcomes between two independent groups. This test aimed to determine if there exists a statistically significant difference between the *Developed* and *Developing* groups based on the indicator *Status*. Indeed, the Mann Whitney test showed a statistically significant difference with a *p-value* less than 0.001 thereby proving our assumption.

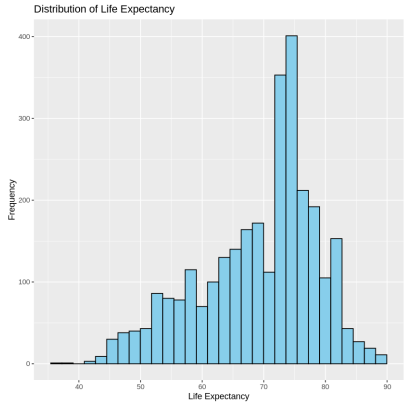

Fig. 1. Distribution of life expectancy

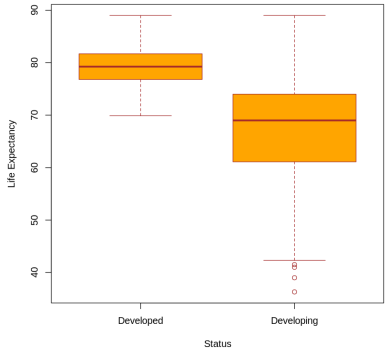

Fig. 2. Box plot of the *Life Expectancy* and *Status* indicators.

Moreover, in Figure 2, it is evident that the interquartile range (IQR) of developing countries is notably broader when compared to the IQR of developed countries. This suggests significant variability or dispersion in the distribution of life expectancy within that category, implying that the range of life expectancies in developing countries is diverse, with significant differences among individual countries. In contrast, a narrower IQR for developed countries indicates relatively less variability, suggesting more consistency in life expectancies across those nations.

A similar assumption can be made considering the years of schooling; that is, a country that dedicates 12 or more years to educating its citizens would have a higher life expectancy than a country that dedicates less than that. Theoretical implications suggest that an increase in years of education may be associated with enhanced career prospects, resulting in greater job opportunities and improved access to healthcare.

Having already examined the distribution of the target indicator, we can skip to the second step and assess whether the variances within the three categories of the *Schooling* indicator

TABLE III
PAIRWISE COMPARISONS USING WILCOXON RANK SUM TEST FOR THE *Schooling* INDICATOR

	Low	Medium
Medium	$p < 0.001$	-
High	$p < 0.001$	$p < 0.001$

are equal. The application of the Kruskal-Wallis test revealed a statistically significant effect ($p - value < 0.001$) of the *Schooling* indicator to the target variable *Life Expectancy*, thereby proving our logical assumption. The Flinger-Killeen test indicated unequal variances $p - value < 0.001$ and the pairwise comparisons using Wilcoxon rank sum test revealed the existence of a statistically significant difference between all pairs under examination. Given our assumption of a country's dedication to its educational system, the results in fig. 3 are expected.

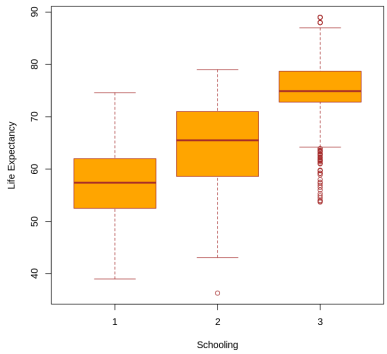

Fig. 3. Box plot of the *Life Expectancy* and *Schooling* indicators.

With the significance of the categorical indicators explored, we can now shift our focus to examining the correlation between the *Life Expectancy* indicator and all the other continuous indicators in the dataset. Prior to exploring the correlations among the indicators, we formulated theoretical assumptions regarding the nature of these correlations. We assume that the following indicators have the highest degree of correlation to *Life Expectancy*:

- Adult Mortality: If the probability of mortality is high for both genders within a given country, it is reasonable to infer that the corresponding life expectancy value for that country would be diminished. However, this indicator remains somewhat abstract as it does not provide specific insights into the underlying causes of mortality. It may encompass a range of factors not explicitly detailed in the dataset.
- Percentage Expenditure: If a significant proportion of a country's Gross Domestic Product (GDP) per capita is allocated to health-related activities, it suggests a potential association with higher life expectancy values.
- Body Mass Index (BMI): If a country's population tends to have a higher average body mass, it indicates an unhealthy lifestyle that may contribute to overall health issues and potentially lead to death.
- Total Expenditure: A substantial investment by a country's government in healthcare is anticipated to positively impact the life expectancy of its citizens, contributing to an overall increase in lifespan.
- GDP: A country's higher Gross Domestic Product (GDP) is associated with improved life expectancy due to increased investments in healthcare, education, and overall living standards.
- Income Composition of Resources: A higher income composition is often associated with improved access to healthcare, better living standards, and increased educational opportunities. These factors contribute to enhanced health outcomes, better disease prevention, and overall well-being, ultimately influencing longer life expectancy.

Our correlation analysis involved the examination of the target indicator with all the other continuous indicators in the dataset. To assess correlation, we opted for the Spearman rank correlation coefficient. This choice was motivated by the non-normal distributions of the independent indicators, for which no transformation could be applied to achieve normality. The findings presented in Table IV indicate that only a limited number of indicators demonstrated a meaningful and strong correlation with life expectancy. Moreover, concerning the indicators anticipated to be correlated with *Life Expectancy*, it is notable that only the *Total Expenditure* indicator is absent from Table IV. Nonetheless, our conjectures regarding the *Adult Mortality* and *Income Composition of Resources* indicators proved accurate. Specifically, a higher income composition emerged as the most significant factor influencing *life expectancy*. Among the newly introduced indicators in Table IV, the *HIV/AIDS* indicator exhibited the strongest correlation with the target variable. This suggests that as life expectancy decreases, the number of deaths per 1000 live births attributed to HIV or AIDS tends to increase.

TABLE IV
CORRELATION TABLE FOR A LIMITED NUMBER OF INDICATORS THAT DEMONSTRATED A MEANINGFUL AND STRONG CORRELATION WITH LIFE EXPECTANCY

Indicator	**r**
Adult Mortality	-0.65
Infant Deaths	-0.60
Percentage Expenditure	0.43
BMI	0.58
Under-five Deaths	-0.62
Polio	0.53
Diphtheria	0.54
HIV/AIDS	-0.75
GDP	0.64
Thinness 1-19	-0.61
Thinenss 5-9	-0.62
Income Composition of Resources	0.86

Knowing which indicators influence *Life Expectancy*, we can proceed to construct the linear regression model.

III. Model Building

A. Linear Regression Model

Among the indicators showing a correlation with *Life Expectancy* in Table IV, only *Thinness 5-9 years* and *GDP* were excluded from the linear regression model using the forward selection algorithm. It is crucial to emphasize at this point that we explored the correlation between every pair of independent variables and opted to exclude variables with high correlation. However, this step did not yield a significant improvement in the models' performance, leading us to disregard this approach. The forward selection algorithm starts with no predictors in the model and adds them iteratively, selecting at each step the feature that most significantly improves the model's performance. This process continues until no additional features significantly enhance the model. Notably, *Income Composition of Resources*, *HIV/AIDS*, *Schooling*, and *Adult Mortality* indicators exhibited substantial reductions in the AIC statistical measure, with the most significant improvement observed for *Income Composition of Resources*, as anticipated. The coefficients of the final linear regression model were all statistically different from zero. However, their estimates were challenging to interpret due to their diverse scales and meanings (see Table V).

TABLE V
Estimates for the Coefficients
Linear Regression Model

Indicator	Estimate
Status(Developing)	-2.239e+00
Adult Mortality	-1.704e-02
Infant Deaths	9.304e-02
Percentage Expenditure	3.891e-04
BMI	3.651e-02
Under-five Deaths	-6.954e-02
Polio	2.465e-02
Diphtheria	2.491e-02
HIV/AIDS	-4.761e-01
Thinness 1-19	-8.200e-02
Income Composition of Resources	8.107e+00
Schooling(Medium)	3.777e+00
Schooling(High)	6.794e+00

Moreover, the model demonstrates a strong explanatory power, accounting for 83% of the variability in the target variable (Multiple $R^2 = 0.83$, Adjusted $R^2 = 0.83$). This is notably robust, considering the complexity inherent in both the target and independent indicators. To assess the model's predictive accuracy, we analyzed its performance on the test set. The closeness of the predicted life expectancy values to the actual values served as an indicator of the model's accuracy, helping us understand its effectiveness in predicting life expectancy. We splitted the dataset into a training set (80%) and a testing set (20%), and utilized the following metrics to evaluate the model:

- Mean/Median Error (ME/MdE)
- Mean/Median Absolute Error (MAE/MdAE)
- Mean/Median Magnitude of Relative Error (MMRE/MdMRE)
- Mean/Median Magnitude of Relative Error to the Estimate (MdMER)

Table VI lists the error values.

TABLE VI
Model Error Evaluation
Linear Regression Model

Metric	Error
ME	0.09
MdE	0.03
MAE	2.79
MdAE	2.22
MMRE	0.04
MdMRE	0.03
MMER	0.041
MdMER	0.03

The error metrics reflect a high degree of accuracy in the linear regression model's predictions. The Mean Error (ME) and Median Error (MdE) are close to zero, indicating minimal bias in the predictions. The Mean Absolute Error (MAE) and Median Absolute Error (MdAE), being relatively low, suggest that the model's predictions are, on average, close to the actual values. Furthermore, the low values of the Mean and Median Magnitude of Relative Error (MMRE and MdMRE) alongside the Magnitude of Error relative to the Estimate (MMER and MdMER) underscore the model's precision in capturing the variance of *Life Expectancy* across the dataset. These error metrics collectively signify that the model performs well not only in fitting the training data but also in generalizing to new, unseen data.

B. Linear Regression Model with Unique Countries

To refine our analysis and account for the uniqueness of each country's data, we implemented a second model. In this approach, we filtered the dataset to retain only a single representative record for each country. This was done to reduce potential biases arising from the multiple entries per country across different years, which may carry similar inherent attributes, thus over-representing certain nations. By utilizing this streamlined dataset, we aimed to capture the overarching trends more accurately, focusing on the nuances that define each country's health landscape. The outcomes from this second model allowed us to reinforce our confidence in the robustness of the variables identified as significant predictors and provide a more nuanced understanding of the determinants of *Life Expectancy*.

TABLE VII
Estimates for the Coefficients
Linear Regression Model with Unique Countries

Indicator	Estimate
Income Composition of Resources	35.930381
HIV/AIDS	-1.267662
Adult Mortality	-0.023041

The implementation of this second, streamlined model, which included only one record per country, yielded a Multiple R-squared of 91.58% demonstrating a high degree of variance in *Life Expectancy* accounted for by our model. The Adjusted R-squared value, sitting at 91.33% also indicated a strong fit while adjusting for the number of predictors in the model. These high R-squared values suggest that our model, with its selected predictors, captures the underlying pattern of the data exceptionally well, reflecting a robust predictive capability when considering each country's unique context in the dataset.

We followed the same training and testing approach as the first model, and calculated the metrics in Table VIII.

TABLE VIII
MODEL ERROR EVALUATION
LINEAR REGRESSION MODEL WITH UNIQUE COUNTRIES

Metric	**Error**
ME	-0.48
MdE	-0.083
MAE	2.31
MdAE	1.63
MMRE	0.04
MdMRE	0.02
MMER	0.034
MdMER	0.023

IV. CONCLUSION

In this study, we conducted a comprehensive analysis of the factors influencing life expectancy, examining various indicators and their characteristics. Our investigation involved exploring the correlation between independent indicators and the target variable, which is life expectancy. Subsequently, we developed two models to elucidate the variability of life expectancy. The first model utilized all instances of the dataset corresponding to the independent variables, while the second model employed unique samples for each country, acknowledging the distinctiveness of their data. Our findings reveal that key indicators, namely the Income Composition of Resources, the HIV/AIDS indicator, and the Adult Mortality indicator, exert significant impact on life expectancy. This underscores the importance of considering these factors in understanding and predicting variations in life expectancy across different countries.