ISYE 6414-Regression Analysis

Final Project Report

Group 8

Global Health Metrics Socioeconomic Correlations Study Via Regression

---

Aravind Rajeev Nair

Ayushi Goyal

Divij Mishra

Krishna Raj

Zhibo Zhang

Contents

1 Introduction

The global landscape of public health has seen significant changes over the past decades, particularly in terms of life expectancy and adult mortality rates. These changes are influenced by a myriad of factors ranging from healthcare advancements to socio-economic developments. Understanding these influences is crucial for shaping effective health policies and interventions.

This study aims to delve into the correlations between life expectancy, adult mortality, and various health, economic, and demographic indicators. By analyzing a dataset that encompasses 179 countries over the period from 2000 to 2015, the project seeks to uncover patterns and relationships that can offer insights into the determinants of life expectancy and adult mortality. The objective is not only to quantify these relationships but also to assess their statistical significance, providing a comprehensive overview of global health trends.

2 Problem Statement

This project aims to unravel the complex correlations between two pivotal health metrics, life expectancy and adult mortality, and a range of other indicators, encompassing immunization rates, economic, and demographic data. Utilizing a database that covers information from 179 countries across the years 2000 - 2015, comprising 2864 data points, this study seeks to offer a comprehensive view of global health dynamics.

2.1 Key Objectives

Establish Relationships: To determine the relationships between the dependent variables, life expectancy, and adult mortality, and a wide array of independent variables. These include demographic factors, health indicators like immunization rates, economic data, and lifestyle factors such as alcohol consumption.

Variable Analysis: The study focuses on a variety of independent or predictor variables such as country, region, infant and under-5 mortality rates, alcohol consumption levels, various immunization coverages (Hepatitis B, Measles, Polio, DTP3), BMI, HIV incidents, GDP per capita, population size, prevalence of thinness in different age groups, and average years of schooling. The analysis will also consider the economic status of countries (developed or developing).

Methodology: Employing multiple linear regression methods to construct a model that effectively illustrates the relationship between the dependent variables (Life Expectancy and Adult Mortality) and the aforementioned independent variables.

Quantitative Insights: The model aims to quantify these relationships, specifically focusing on the impact weights of the independent variables. This involves identifying which factors most significantly influence life expectancy and adult mortality, assessing the statistical significance of each factor, and evaluating the explanatory power of the model.

Residual Analysis: Undertaking a thorough residual analysis to assess the adequacy of the model and identify any potential anomalies or areas of improvement.

2.2 Goals and Significance

Insights into Health Metrics: The overarching goal is to provide deeper insights into the factors that significantly affect life expectancy and adult mortality. By quantifying the impact of various predictors, the study aims to highlight critical areas for public health interventions and policy-making.

Understanding Life-Altering Factors: Another key objective is to elucidate potential life-altering factors and risks. This understanding is crucial for individuals, communities, and policymakers to make informed decisions and implement strategies that can positively impact public health outcomes.

Model Utility and Application: The study seeks to develop a model that is not just statistically robust but also practically relevant, offering actionable insights and applications in the realm of global health.

3 Data Description and Preliminary Analysis

The dataset used in this project is a comprehensive collection of data points time-stamped from 2000 to 2015, encompassing 179 countries, which results in a total of 2864 data points. The primary source of this data is Kaggle which consolidates information from the World Health Organization (WHO) and the World Bank.[1]

3.1 Variable Description

Quantitative Variables:
1. Infant_deaths: Number of infant deaths per 1000 population.
2. Under_five_deaths: Deaths of children under five years per 1000 population.
3. Adult_mortality: Deaths of adults per 1000 population. **4. Alcohol_consumption**: Liters of pure alcohol consumed per capita (age 15+).
5. Hepatitis_B: Percentage coverage of Hepatitis B immunization among 1-year-olds.
6. Measles: Percentage coverage of Measles vaccine first dose among 1-year-olds.
7. BMI: Body Mass Index, a measure of nutritional status.

8. **Polio**: Percentage coverage of Polio immunization among 1-year-olds.
9. **Diphtheria**: Percentage coverage of DTP3 immunization among 1-year-olds.
10. **Incidents_HIV**: Incidents of HIV per 1000 population (aged 15-49).
11. **GDP_per_capita**: GDP in current USD.
12. **Population**: Total population in millions.
13. **Thinness_ten_nineteen_years**: Prevalence of thinness among adolescents.
14. **Thinness_five_nine_years**: Prevalence of thinness among children.
15. **Schooling**: Average years in formal education (aged 25+).
16. **Life expectancy**: Average life expectancy across years.
17. **Year**: Observation years from 2000 to 2015.

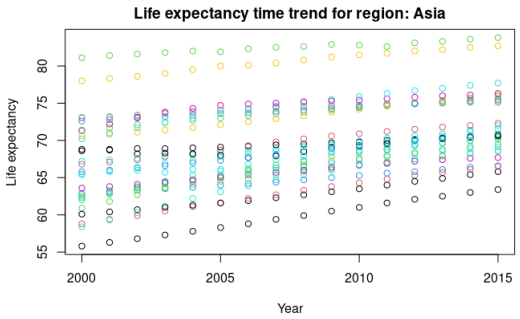
Qualitative Variables:
1. **Country**: Categorical variable representing 179 countries.
2. **Region**: 179 countries categorized into 9 regions.
3. **Economy_status_Developed**: Binary variable indicating developed status.
4. **Economy_status_Developing**: Binary variable indicating developing status.

4 Initial Analyses

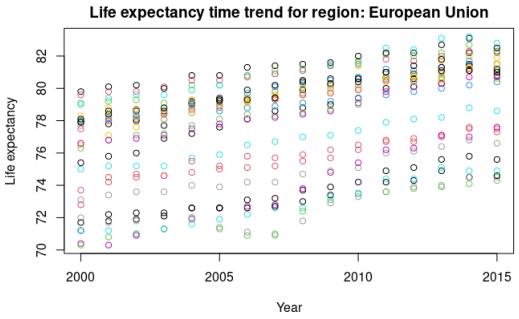
4.1 Time Series Analysis of Life Expectancy

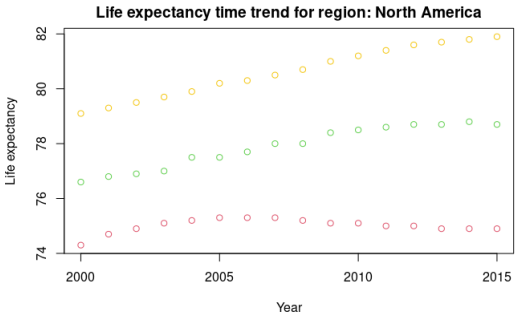
The primary observation from the TSA is that there is a general trend of increasing life expectancy across almost all countries. This trend, however, is not strictly linear, indicating that while the overall direction is towards longer life expectancy, the rate of increase varies significantly across different regions and time periods.

Life expectancy vs. year plots for each region are suggested to visualize these trends. These plots will display how life expectancy has changed over time in different regions.



Almost all countries in Asia and Africa have shown an increasing trend in life expectancy over time. A similar trend is observed in other continents[2] as well, such as Europe and North America, as illustrated in the plots attached below.



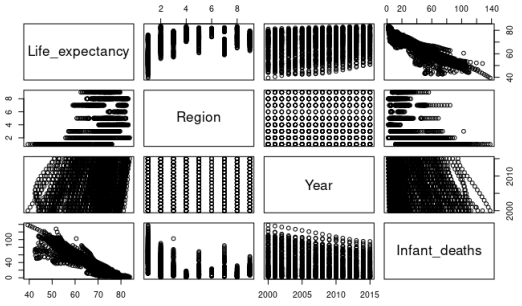


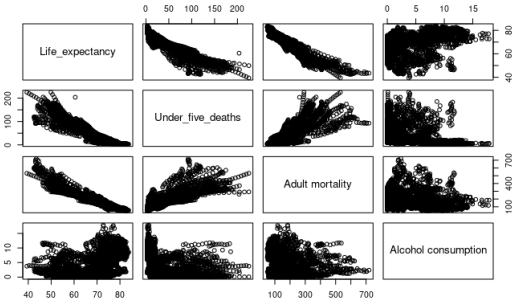
While the time series component reveals valuable insights about the general trend, the study emphasizes the importance of identifying and focusing on non-time variables that are correlated with life expectancy. This approach aims to understand what specific factors, aside from the passage of time, significantly impact life expectancy. The analysis will further delve into various socio-economic, health-related, and demographic variables to determine their correlation and impact on life expectancy. This includes examining factors like GDP per capita, immunization rates, alcohol consumption, and educational attainment.

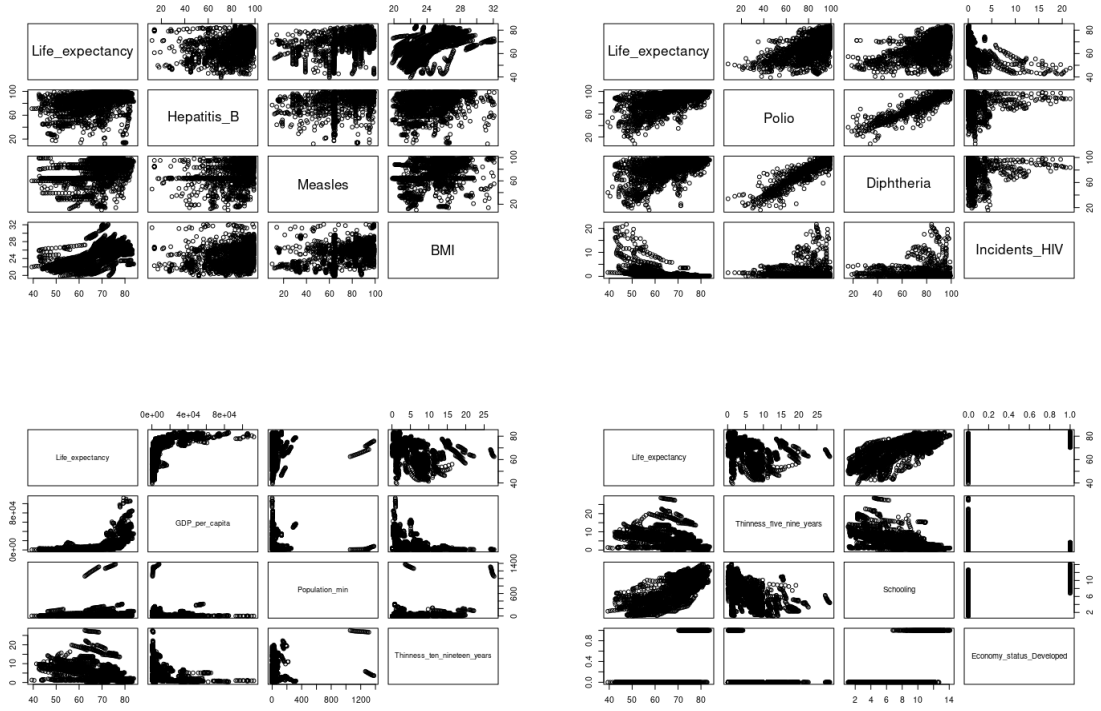
4.2 Exploratory Data Analysis (EDA)

4.2.1 Scatter Plots

The EDA phase includes an extensive examination of scatter plots, which provide insights into the relationships between the dependent variable (life expectancy) and various independent variables. These plots are dense with data, offering a preliminary visual understanding of the variables' interactions.







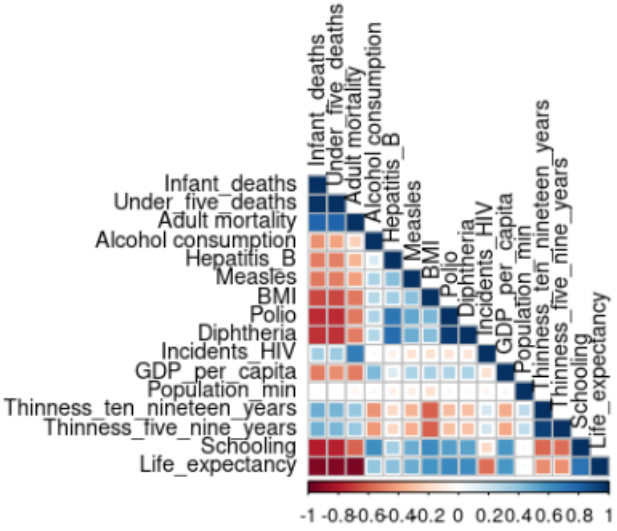
Key Observations from Scatter Plots:
Data Density: Many plots are densely populated, showing a wide range of data points across variables. This can sometimes obscure trends or relationships.
Need for Transformation: Some variables, like GDP_per_capita, may require transformations to better visualize and understand their relationship with life expectancy. For example, using a logarithmic scale could help in flattening and better interpreting these plots.

4.2.2 Correlation Analysis

Correlation analysis is an essential part of EDA, helping to identify how strongly each independent variable is related to the dependent variable.

Key Observations from Correlation Analysis:
Low Correlation Variables: Some variables, such as Population_min, show minimal correlation with life expectancy and might be considered for removal from the analysis.
Potential Removals: Variables like Thinness measures, Alcohol consumption, and

Hepatitis B immunization rates are less correlated compared to others. Their relevance as control variables is questionable, and they may be candidates for exclusion.
Multicollinearity Issues: A significant observation is the presence of multicollinearity, especially between various disease rates and mortality rates. This suggests that improvements in socio-economic factors like GDP and schooling could indirectly impact these disease rates.
Streamlining Variables: To simplify the analysis, it might be prudent to focus on a narrower set of variables that are more strongly correlated with life expectancy. This could include BMI, GDP_per_capita, schooling, Region, and Economy_status_Developed.



5 Model Development and Diagnostics

5.1 Model Development and Iteration Process

We commenced our analysis by filtering the dataset to focus on variables with higher relevance and correlation to life expectancy. This refinement included variables like BMI, GDP_per_capita, Schooling, Region, Economy_status_Developed, Infant_deaths, and Adult_mortality. The aim was to simplify the analysis by excluding variables with low correlation or high correlation with other predictor variables.

Using a 90:10 split for training and validation datasets, we applied both forward and backward stepwise regression techniques to iteratively refine our model. We began with a minimal model and gradually added variables, assessing their impact

based on the Akaike Information Criterion (AIC).The forward stepwise selection started with Adult_mortality as the most significant predictor, progressively adding Infant_deaths, Region, Economy_status_Developed, GDP_per_capita, and Schooling. Similarly, the backward stepwise regression confirmed the significance of these variables.

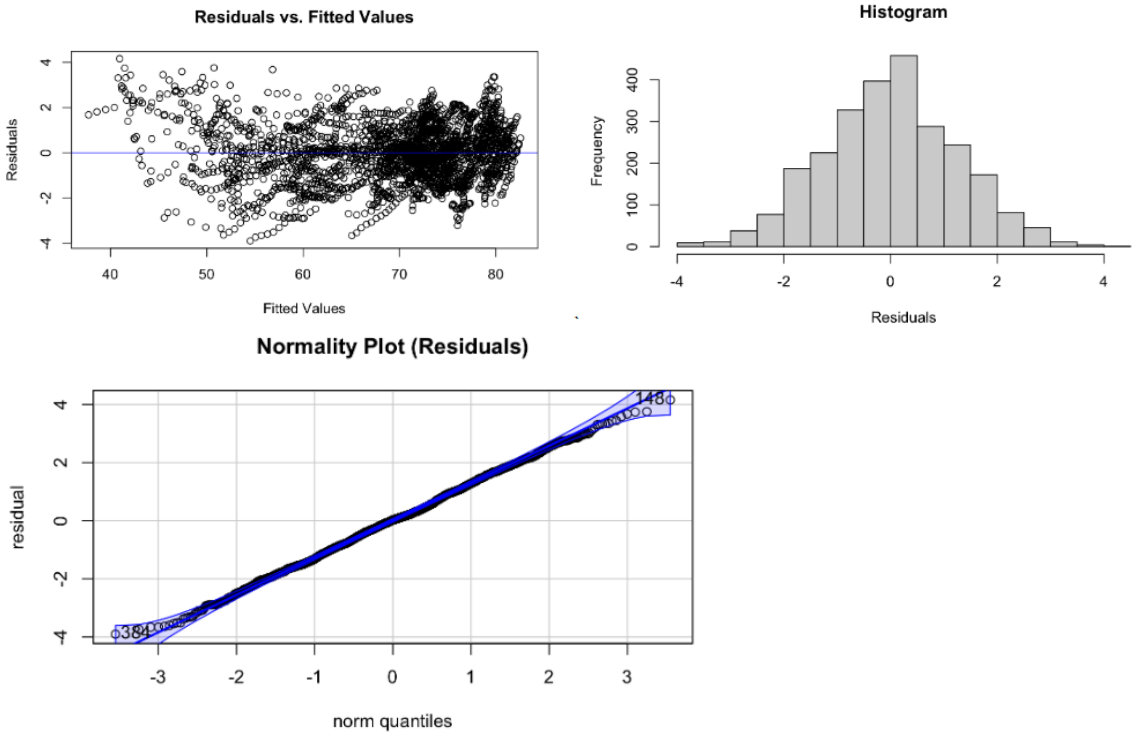
5.2 Model-Fitting, Diagnostics, Performance, and Other Checks

5.2.1 Model Summary

Our final model included the predictors: BMI, GDP_per_capita, Schooling, Region, Economy_status_Developed, Infant_deaths, and Adult_mortality. This model exhibited a high R-squared value of 0.982, indicating a strong fit to the data.

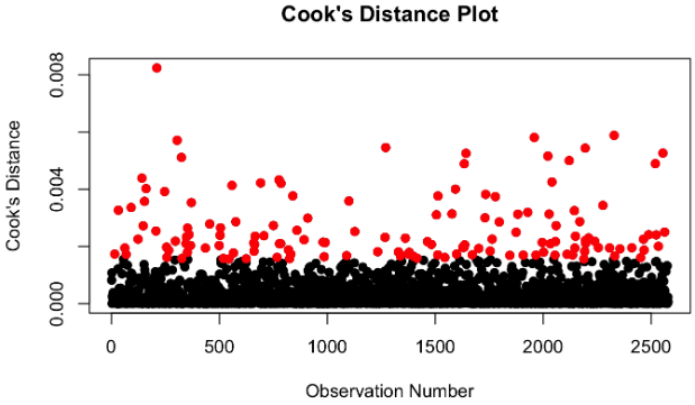
5.2.2 Diagnostics and Residual Analysis

Residuals Analysis: The mean of the residuals is 5.383161e-17 which is close to zero, aligning with model assumptions. The plot of "Residuals vs. Fitted Values" illustrates the absence of patterns, indicating good model fit and constant variance. **Normality Check**: Histogram and QQ plots of residuals suggested normality, which is essential for linear regression assumptions.



5.2.3 Handling of Outliers

Using Cook's Distance, we identified 147 outliers, primarily from the African region.Despite their significant number, these outliers were not removed as they represent crucial aspects of underdeveloped regions and are influential in understanding life expectancy variations.



5.2.4 Multicollinearity Check

Variance Inflation Factor (VIF) Analysis: Our analysis revealed that the VIF values for all predictors were well below the threshold of 55.47, suggesting no multicollinearity concerns in the model.

5.2.5 Model Performance Evaluation

Training and Testing Performance: The model showed low Mean Squared Error (MSE) and high R-squared values on both training and testing datasets, indicating its robustness and predictive accuracy.

6 Conclusions and Recommendations

6.1 Interpretation of Results

Our comprehensive analysis revealed that life expectancy is significantly influenced by a combination of factors including adult mortality, infant deaths, regional differences, economic status, GDP per capita, schooling levels, and BMI. The model's high R-squared value of 0.982 suggests that these variables collectively offer a robust predictive power for life expectancy.

Adult Mortality and Infant Deaths: These emerged as crucial predictors, underscoring the impact of health conditions and healthcare quality on life expectancy.

Economic Factors: GDP per capita and economic status (developed vs. developing) were significant, indicating the profound influence of economic conditions on public health. **Education and Health**: Schooling and BMI were also significant predictors, highlighting the role of education and nutritional health in determining life expectancy.

6.2 Suggestions and Applications

Public Health Policies: The strong correlation between life expectancy and factors like adult mortality, infant deaths, and BMI can inform public health strategies. Policies aimed at improving healthcare access and quality, particularly in developing regions, are crucial.
Educational Initiatives: Given the impact of schooling on life expectancy, investments in education could be a strategic approach to enhance public health outcomes.
Economic Development: The significant role of GDP per capita suggests that economic growth and stability can positively affect life expectancy. This underscores the need for holistic development policies.

Examples of Applications
Regional Health Programs: Tailoring health programs based on regional characteristics and needs, as identified in the analysis, could lead to more effective outcomes.
Targeted Interventions: Interventions focusing on reducing adult and infant mortality rates in regions with lower life expectancy could be prioritized.

6.3 Model Validation and Future Work

The model validation on the test dataset, indicated by low MSE and high R-squared values, confirms the model's predictive accuracy and reliability. However, future studies could explore the incorporation of additional variables, like environmental factors or genetic predispositions, to further refine the understanding of life expectancy determinants.

7 Appendex

[1] Kaggle: Life Expectancy (WHO - Updated).

[2] Other plots from the remaining continents