Global Life Expectancy: Unraveling Health and Economic Determinants (2015-2020)*

Multiple Linear Regression Analyzing Critical Factors Shaping Life Expectancy

Yanfei Huang

December 2, 2024

This paper analyze life expectancy and its determinants through 184 countries and make the prediction of people from different Income Group. Multiple Linear regression is used to deploying the life expectanc with gender, income and region. Predictions of life expectancy of people from different income group is made according to these essential predictors. The finding indicates that the life expectancy is tend to get higher as the economic class grows. And we predict that the average age of person from low, lower-middle, upper-middle and high are 65,70, 75,80. The result of this paper will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

Table of contents

1 Introduction

Life expectancy is a critical measure of a population's overall health and well-being, shaped by various factors such as gender, geographic location and economic group. The index of life expectancy is generally served as a benchmark for income group, indicating the effectiveness of interventions in reducing mortality and improving well-being. Higher life expectancy is believed to linked to stronger physical factor, better living standards, and geographic location. Conversely, low life expectancy often signals systemic challenges such as weaker health, lower economic situation, and inadequate healthcare access of different region. (**socialdeterminantsofhealth?**).

*Code and data are available at: https://github.com/wendyhuan/lifeexpectancy.

The primary goal of this paper is to determine which factors play a statistically significant role in driving lower life expectancy values and to offer actionable insights based on different income group. According to the World Bank Income Group, the countries are classified into low, lower-middle, upper-middle, and high based on the country's Gross National Income. Using Multiple Linear Regression model and Bayesian model, this study focuses on understanding the relationship between Gender, Region and different income group in predicting life expectancy across different countries. By identifying and analyzing these predictors, the study aims to highlight actionable aspects for policymakers to target in their efforts to improve population longevity effectively.

Related Research has shown that there exist difference between men and women related with biological, behavioral, and socioeconomic factors, highlighting that gender-specific health behaviors and societal roles influence longevity (**GenderandLifeExpectancy?**). Higher-income individuals tend to live longer due to better access to healthcare and healthier lifestyles, with notable regional disparities even within similar income groups (**IncomeGroupsandLongevity?**). Additionally, regional socioeconomic differences on health outcomes, especially life expectancy. It emphasizes that addressing regional inequities in wealth and resources is essential for improving population health globally. (**RegionalVariations?**).

Findings from this paper reveal that countries of higher income group tend to achieve higher life expectancy, while factors like lower income and poverty region emerge as significant obstacles. The analysis underscores the importance of targeted interventions in key areas such as healthcare accessibility and economic equity to address disparities in life expectancy. This study further contributes to a deeper understanding of how predictive modeling under different country status can inform public health strategies and policy-making on a global scale.

The ultimate goal is to assist policymakers in developing evidence-based strategies that can enhance population health outcomes. The approach of this paper not only emphasizes the importance of equitable access to healthcare but also contributes to a broader understanding of the multifaceted factors shaping life expectancy globally. By highlighting the interplay between various determinants, this study contributes to a broader understanding of the challenges and opportunities in improving life expectancy on a global scale.

The structure of the paper is as follows: Section **??** outlines the data sources and variables considered, followed by the model setup in Section **??** and justification in **?@sec-modjust**. The results in **?@sec-result** presents the key findings of the analysis, with a discussion on the implications. **?@sec-discussion** then discusses potential limitations and suggestions for future research. **?@sec-appx** provides additional detailed information about the data, model and methodology.

2 Data

2.1 Overview

The data used in this analysis originates from The World Health Organization's (WHO) and Global Health Observatory (GHO) (**lifeexpectancy?**). This data-set related to life expectancy, health factors for 184 countries has been collected from WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those socio-economic factors on the national level were chosen for global scale analysis.

This analysis uses the statistical programming language R (R Core Team 2023) and several libraries, including `tidyverse` (**tidyverse?**), `janitor` (**janitor?**), `knitr` (**knitr?**), `dplyr` (**dplyr?**), `arrow` (**arrow?**), `purrr` (**purrr?**), `sf` (**sf?**), and `here` (**here?**) for data manipulation. `ggplot2` (**ggplot?**), `ggcorrplot` (**ggcorrplot?**) and `kableExtra` (**kableExtra?**) for visualization. The dataset covers various predictors conducted across multiple countries, capturing the support for a country to determine the predicting factor which is contributing to lower value of life expectancy.

2.2 Measurement

The measurement process refers to how real-world factors—such as the Gender, geographic location of a country and income group - are translated into numerical entries representing life expectancy in a dataset. Each entry captures the average life expectancy of individuals in a specific country during a given year.

Life Expectancy (`Life Expectancy`): This variable, life expectancy at birth, represents the number of years a person is expected to live , assuming current mortality conditions persist. It is derived using data from national health records, the World Health Organization (WHO) and Global Health Observatory (GHO). The data on the raw dataset is recorded as a range of age. For simple analysis, we drop the range and take the average of year with one decimal. The values are calculated and represented in age between 10.1 to 87.4.

Income Group(`Income_Group`): This is a categorical variable that classifies countries into four groups by the country's Gross National Income according to the latest index from World Bank Income Group. The countries are classified into low, lower-middle, upper-middle, and high under the standard as follow. Low-income economies is defined as a country with a gross national income less than $1135, lower-middle is between the range to $1136 to $4465, upper-middle in the range of $4465 to $13845 and high is more than $13846.

Gender(`Gender`): The gender of the population of each country is included as a categorical variable (Male/Female/Both Sex).

Region(Region): This is a categorical variable that classifies the geographic location of the countries. It is mapped one by one by the name of the country. The data is stored as the name of the continent('Africa', 'Oceania', 'Asia', 'Europe', 'North America', 'South America').

2.3 Data Cleaning

The raw life expectancy data underwent a several cleaning steps to ensure it was accurate, consistent and ready for analysis. The goal of this cleaning process is to create a table including income group (high, upper middle, lower middle, low), gender (Male, Female, Both Sex), region (Asia, Europe, North America, South America, Africa, Oceania) as rows and life expectancy as column.

To make such table, we first select and rename key variables from raw data to focus on relevant information. To make the subsequent analysis easier, we then convert variables to the proper data types and eliminate the rows that have missing data values. To keep things neat, we organize the decimal for every piece of numerical data and drop the percentage symbol. The income group column and the region column was not given in the raw dataset. We created a mapping over country name to four income group according to the index given by World Bank Income Group and save the data under "Income_Group". We as well made a mapping to the countries according to its continent and saved as "Region". We again merge the table, removing repeated columns and rows. For easier visualization, we created four tables, which are grouped by life expectancy at age 60, life expectancy at birth, life expectancy at birth of Male and life expectancy at birth of Female.

shows the average of life expectancy under male, female, 6 different regions and 4 income groups.

The cleaned dataset was then saved as both CSV and Parquet file for efficient storage and further analysis.

More information on the data cleaning process can be found in **?@sec-appx**.

2.4 Outcome Variables

The outcome variable is **Life Expectancy**. This is the primary dependent variable that the model is designed to predict. It represents the average number of years a person is expected to live, under the condition that current mortality conditions persist. The model seeks to identify the variables affecting this average.

Figure **??** visualizes the distribution of life expectancy from 2015 to 2020 as percentages, combining a histogram and a density plot. A large concentration of observations is observed range of 0% to 25%, reflecting a significant proportion of countries with lower life expectancy relative to others in the dataset. Another visible concentration is around 75% to 100%, indicating a group of countries with relatively high life expectancy. The blue density curve

overlays the histogram, providing a smoothed representation of the distribution. Peaks in the curve correspond to high frequency within the histogram. The density plot emphasized the bimodal nature of the distribution with two distinct peaks: one around 12 as low life expectancy percentages; another peak around 87as high life expectancy percentage.

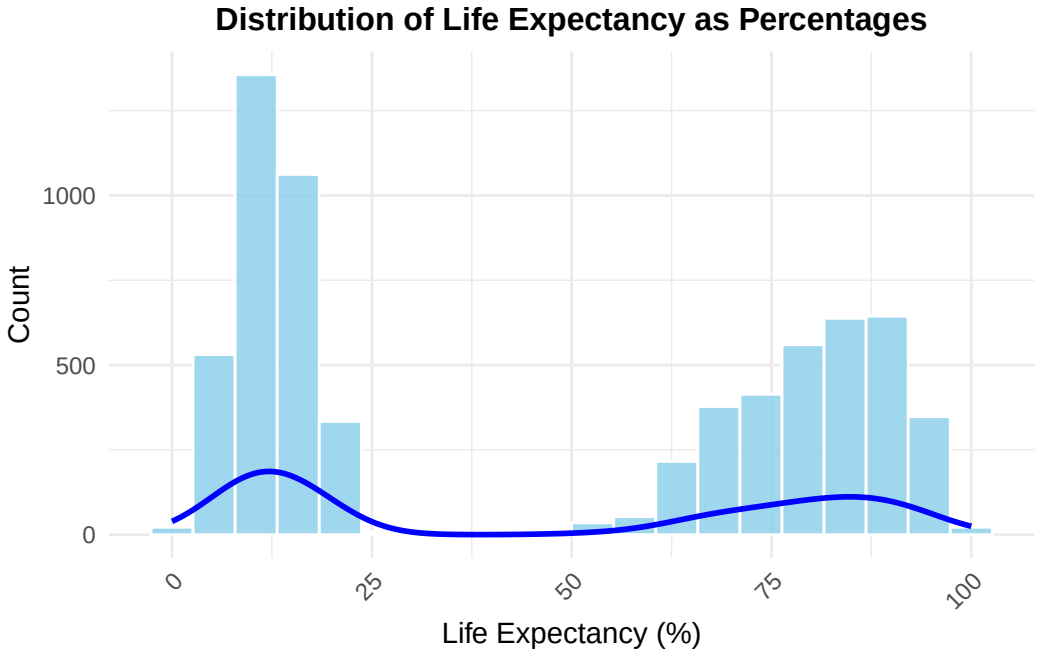

Figure 1: This histogram shows the distribution of the percentage of life expectancy. With a huge gap between 25 and 50, we could indicate that most of the countries has a life expectancy of. There exist large data with low life expectancy less than 25 years.

2.5 Predictor Variables

The **predictor variables** (or independent variables) are the factors believed to influence the life expectancy:

1. **Income Group(`Income_Group`)**: The income group of a country is the key factor influencing the life expectancy. This variable provides context for comparing life expectancy between different levels of country income group. The division of the life expectancy by the income group of a country because income levels often correlate strongly with various factors affecting health and longevity. Higher income countries typically have more resources to invest in robust healthcare systems, with better living conditions while lower income countries often face challenges such as inadequate healthcare infrastructure and malnutrition.

Following by the four income group, we would like to see the distribution of the income group of 184 WHO member countries. (**fid-income?**) shows that there are 55 countries at the high income group, 26 countries at low income group, 54 countries at the lower-middle group while there are 49 countries at the upper-middle group. With significant less low income countries, we expect to see the life expectancy lie in a comparably higher range. On the other hand, there might exist sever outliers dropping the mean.

2. **Gender(`Gender`)**: The gender of the population of each country is included as a categorical variable (Male/Female/Both Sex). This is a key variable in the dataset because gender has been fully recognized connected with the biological physical factor which directly effect the life expectancy. Figure **??** shows that the with an stable average of 74, life expectancy of female is constantly higher than male's average of 70. The line plot reveals a steady increase until 2019, followed by a sharp decline in 2020, which is assumed to be associated with the impact of COVID-19.

3. **Region(Region)**: This is a categorical variable that classifies the geographic location of the countries. It is mapped one by one by the name of the country. The data is stored as the name of the continent('Africa', 'Oceania', 'Asia', 'Europe', 'North America', 'South America'). Figure **??** indicates that with the number of 54, Africa has the most WHO member countries. Both Asia and Europe has 43 countries,listing in the middle while the other three continent has the significant less number of countries. North America has 22 countries, South America has 12 countries and Oceania has the least, 10 countries.

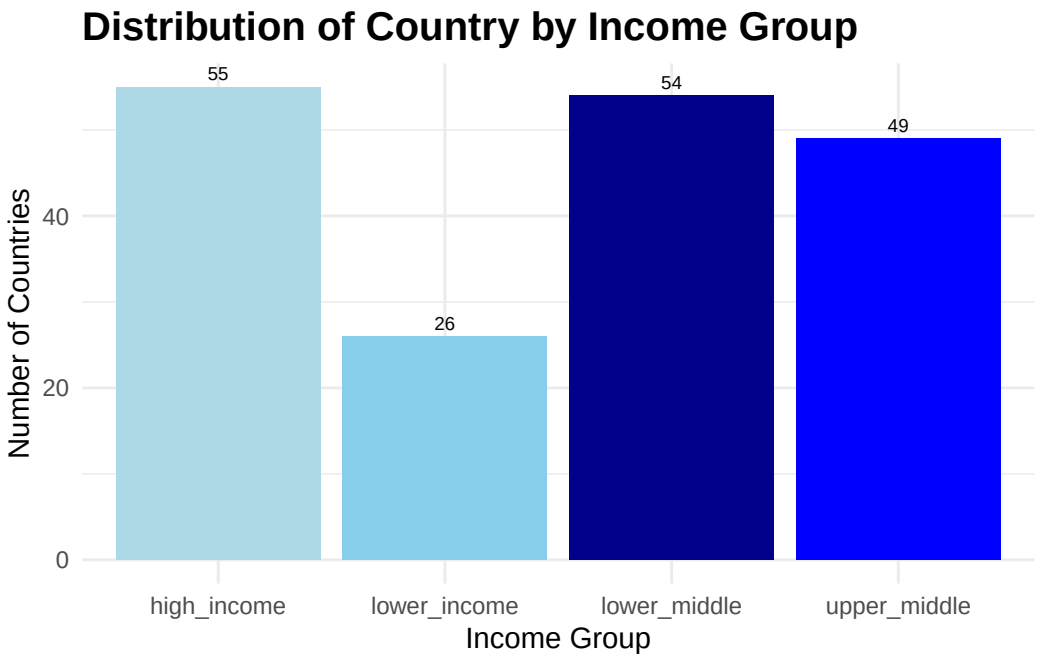

Figure 2: Distribution of country income group.

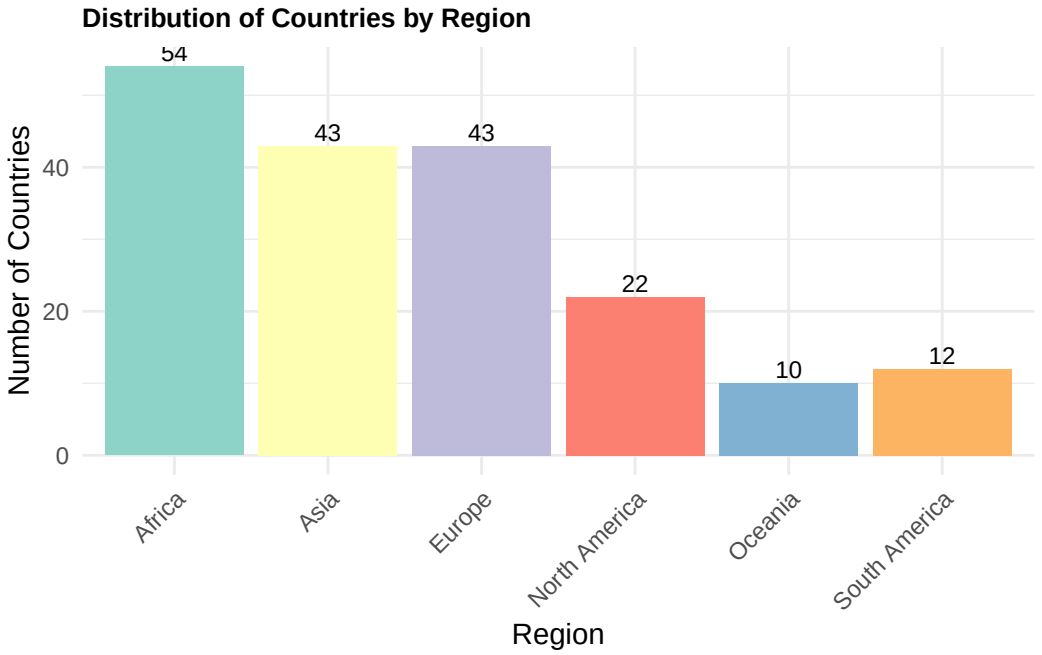

Figure 3: Distribution of different region of the 184 countries. Africa has the most countries with an number of 54. Asia and Europe has the same amount of 43, North America has 22 countries while the other two continent has almost the same amount of countries.

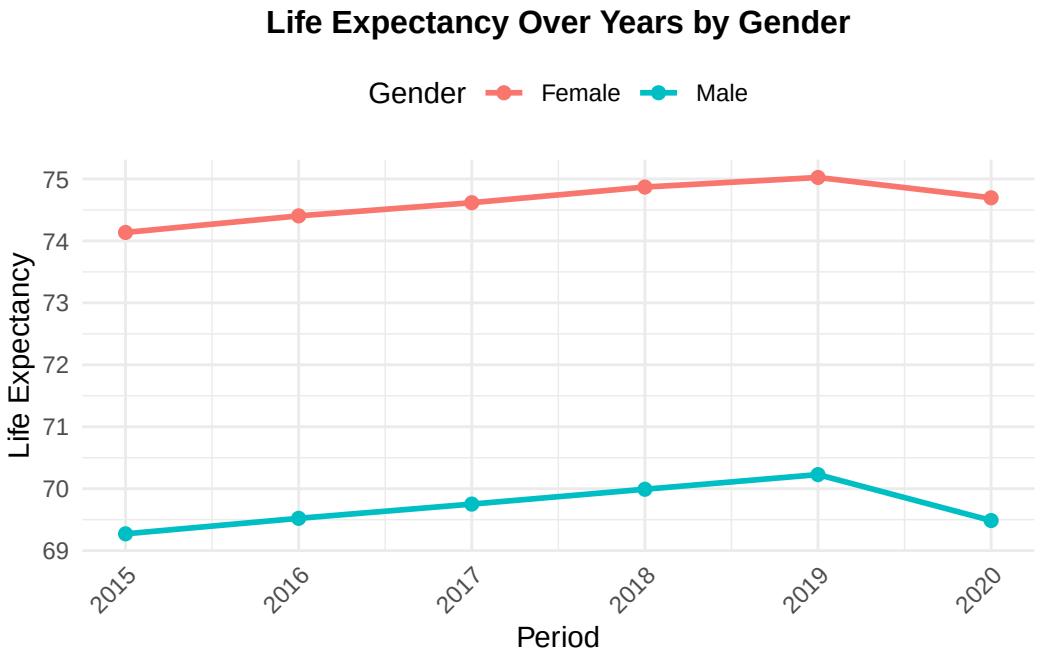

Figure 4: Distribution of gender each year. The average of female life expectancy is approximately 74 around years, constantly higher than male's average of 70. Compared to male, female's life expectancy is marginally more steady.