Global Life Expectancy: Unraveling Health and Economic Determinants (2000-2015)*

Analyzing Critical Factors Shaping Mortality and Well-Being Across 155 Nations

Yanfei Huang

November 30, 2024

This paper analyze life expectancy and its determinants through 155 countries and make the prediction of people from both developed and developing country. Multiple Linear regression is used to deploying the health and socio-economic factors. Predictions of life expectancy of people from developed or developing country is made according to the essential predictors. The finding indicates that the developed countries tend to live longer life as assumed. And we predict that the average age of persons in developed country is 80, while those in underdeveloped countries are 75. 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

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

1 Introduction

Life expectancy is a critical measure of a population's overall health and well-being, shaped by various factors such as personal expenditure on health, government healthcare access, economic stability, and social structures. The index of life expectancy is generally served as a benchmark for development, indicating the effectiveness of interventions in reducing mortality and improving well-being. Higher life expectancy is believed to linked to better living standards, access to education, and equitable healthcare. Conversely, low life expectancy often signals systemic challenges such as poverty, disease burden, and inadequate healthcare access (**socialdeterminantsofhealth?**).

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 country status. Using Multiple Linear Regression model, this study focuses on understanding the relationship between country status(developed or developing), personal average expenditure on health based on GDP, percentage of government expenditure on health, income composition of resources and year of schooling 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 higher healthcare expenditure of government positively impact life expectancy (**valueofvaccination?**). Additionally, addressing social determinants

such as education and income inequality has been identified as a vital pathway to improving longevity (**healthequity?**). Studies also highlight the role of global collaboration in tackling health crises, which can disproportionately affect lower-income nations, mostly developing countries under certain continents, further influencing life expectancy disparities (**World_Health_Organization?**).

Findings from this paper reveal that countries with greater healthcare spending tend to achieve higher life expectancy, while factors like income inequality and absence of education 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 2 outlines the data sources and variables considered, followed by the model setup in Section 3.1 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 155 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 status of a country, average personal health care investment, average year of education and social determinants (country's healthcare investment, human development index of income) — 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 (`LifeExpectancy`): This variable represents the average 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 values are calculated and represented in age between 36.3 to 89.

Percentage Expenditure on Health (`PercentageExpenditure`): This variable reflects the expenditure allocated to health-related expenditures as a percentage of GDP per capita. The data is typically collected from government reports, economic surveys, and global health databases. It indicates the level of personal financial investment in the health sector and is expressed as a percentage.

Schooling (`Schooling`): This variable captures the average number of years of education received by individuals in a country. Data is obtained from educational surveys and censuses conducted by international and national organizations. The values are recorded in years and used to assess the link between education and health outcomes like life expectancy. The data range from 0 to 20.7

Status (`Status`): This is a categorical variable that classifies countries as "Developed" or "Developing," based on various socio-economic and health indicators accepted globally. It provides context for comparing life expectancy across different stages of national development.

Total expenditure (`TotalExpenditure`): This is a numerical variable that shows the general government expenditure on health as a percentage of total government expenditure. The number drop of the percentage and shows as the number between 0 to 100, where higher values indicate a greater share of government spending allocated to health.

Income composition of resource (`IncomeComposition`): This is a numerical variable representing the income component of the Human Development Index (HDI). It reflects the contribution of income to human development, with values ranging from 0 to 1, where higher values signify a greater income contribution to overall development.

2.3 Data Cleaning

The raw life expectancy data underwent a several cleaning steps to ensure it was accurate, consistent and ready for analysis. We first select and rename key variables from raw data to focus on relevant information. The key variables of interest in our analysis include Country, Year, Status of the country, Life expectancy, Percentage expenditure, Total expenditure, Income composition of resources and the year of education. 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. The cleaned dataset was then saved as a 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 `LifeExpendency`. 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. To examine the differences in life expectancy under the same variables, we separate the data into developed and developing countries, where developed country are generally believed to have higher life expectancy. Figure 1 displays density of the historical data of life expectancy across 195 countries from 2000 to 2015, comparing developing to developed countries.The average age of most developed country is approximately 80, while the majority developing countries typically have average life expectancy around mid-seventies.

2.5 Predictor Variables

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

1. **Status(`Status`)**: The status of a country is the key factor influencing the life expectancy. We divide the life expectancy by the status of a country as it is believed that developed country generally have higher life expectancy due to better healthcare access, sanitation, and living standards, while developing countries face greater health challenges. This variable provides context for comparing life expectancy between different levels of country development. Figure 2 includes 28 developed countries and 127 developing countries. The graph on left shows that we've got much greater representation of developing nations across the dataset. Comparing both graph, we still could tell that the dataset contains significantly more data for developing countries than for developed countries, even relative to the difference in the number of countries themselves.

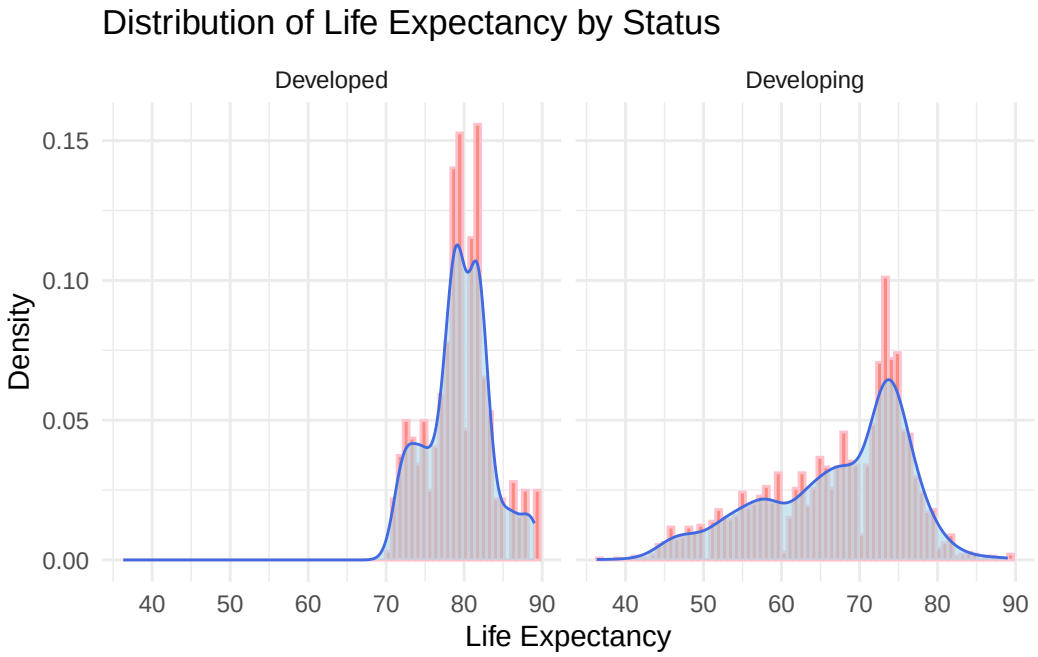


Figure 1: This histogram shows the distribution of the percentage of life expectancy dividing by the status of countries. The graph of developed countries shows narrower spread with density curve located on the rightside of the x-axis. The graph of developing countries performs a wider spread while the highest density position at mid-seventies.

2. **Percentage Expenditure on Health (`PercentageExpenditure`)**: This variable reflects the expenditure allocated to health-related expenditures as a percentage of GDP per capita.
3. **Total expenditure (`TotalExpenditure`)**: This variable shows the general government expenditure on health as a percentage of total government expenditure. This is considered as a reflection of economic situation of a country. A higher percentage indicates that a country places significant emphasis on healthcare as a key area of public investment.
4. **Income composition of resource (`IncomeComposition`)**: his is a numerical variable representing the income component of the Human Development Index (HDI).
5. **Schooling (`Schooling`)**: This is a numerical variable showing the number of education per person.

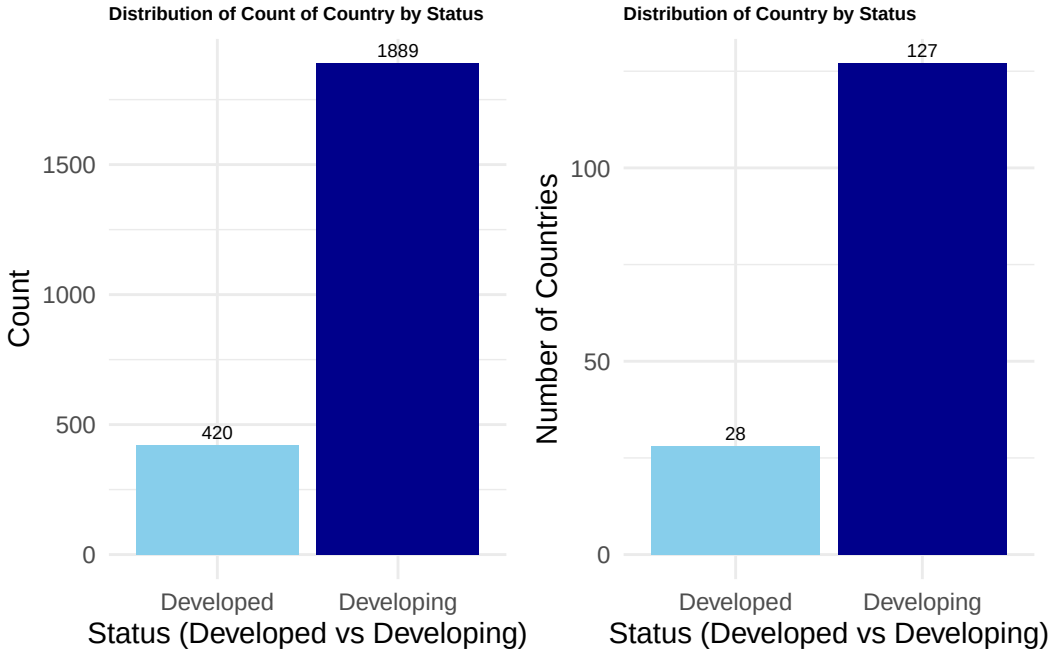
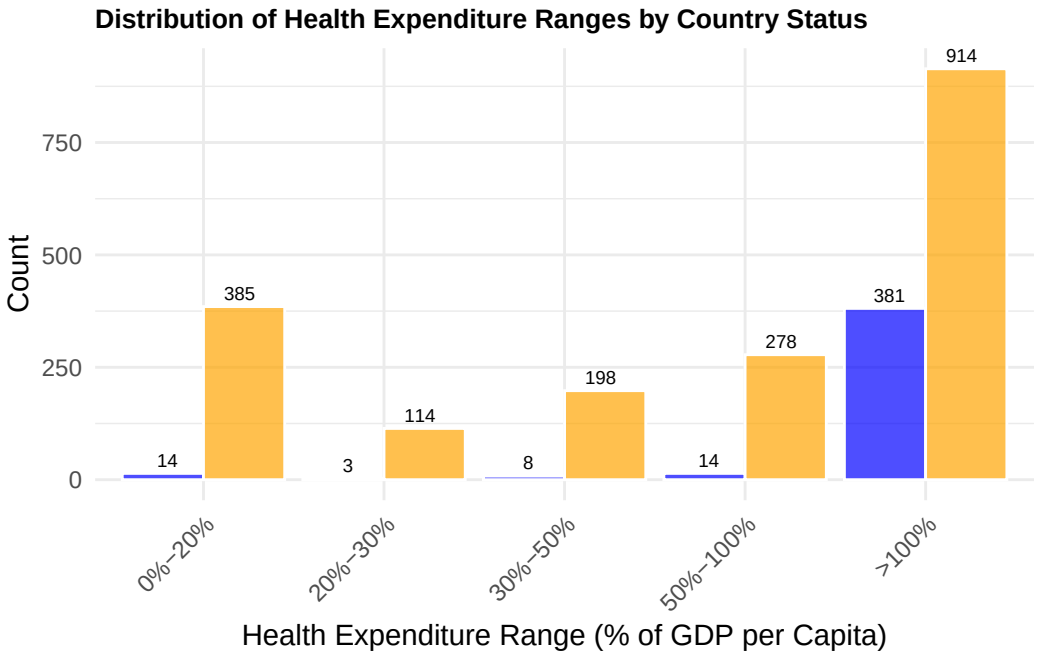


Figure 2: Distribution of country status, showing 28 dveloped countries and 127 developing countries, with a substantial difference in the number of data points. From 2000 to 2015, notably, the number of data entries for developing countries far exceeds that of developed countries, reflecting the greater representation of developing nations across the dataset.

```
# Define percentage ranges
ranges <- c("0%-20%", "20%-30%", "30%-50%", "50%-100%", ">100%")
```

```r
# Create a new variable to classify the ranges
cleaned_expectancy <- cleaned_expectancy %>%
  mutate(
    ExpenditureRange = case_when(
      PercentageExpenditure <= 20 ~ "0%-20%",
      PercentageExpenditure > 20 & PercentageExpenditure <= 30 ~ "20%-30%",
      PercentageExpenditure > 30 & PercentageExpenditure <= 50 ~ "30%-50%",
      PercentageExpenditure > 50 & PercentageExpenditure <= 100 ~ "50%-100%",
      PercentageExpenditure > 100 ~ ">100%"
    )
  ) %>%
  # Ensure the right order
  mutate(ExpenditureRange = factor(ExpenditureRange, levels = ranges))

# Plot the histogram of expenditure ranges separated by status
ggplot(cleaned_expectancy, aes(x = ExpenditureRange, fill = Status)) +
  geom_bar(stat = "count", position = "dodge", color = "white", alpha = 0.7) +
  # Adds count on top of each bin
  geom_text(stat='count', aes(label = ..count..), position = position_dodge(width = 0.8), vj
  scale_fill_manual(values = c("Developed" = "blue", "Developing" = "orange")) +
  labs(
    title = "Distribution of Health Expenditure Ranges by Country Status",
    x = "Health Expenditure Range (% of GDP per Capita)",
    y = "Count",
    fill = "Status"
  ) +
  theme_minimal() +
  theme(
    legend.position = "none",
    axis.text.x = element_text(angle = 45, hjust = 1), # Adjusts the x-axis labels for clari
    plot.title = element_text(size = 10,face = "bold"))
```



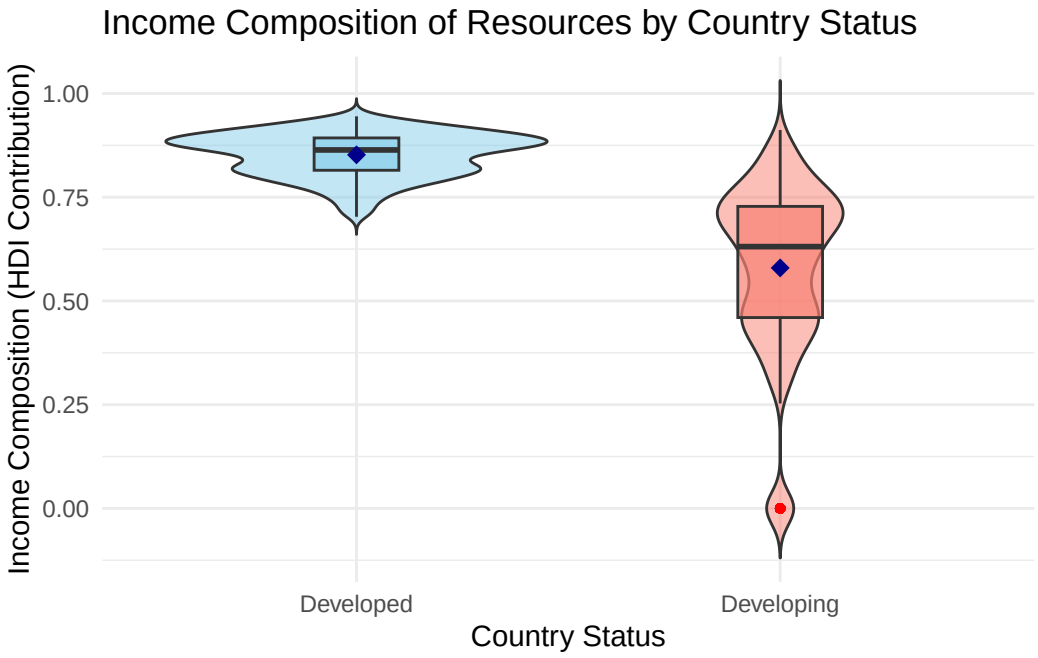


Figure 3: Combined Box plot and Violin plot of income component of the Human Development Index (HDI) showing the variance and the average of the income component under status of different statuses.

3 Model

3.1 Model set-up

To predict the outcome of the life expectancy of people from different regions, we developed two linear regression models using R (R Core Team 2023): one for the developed country and one for developing country. The outcome variable, `LifeExpectancy`, is a continuous and represents the average number of years a person is expected to live, assuming all related resources remain constant throughout their lifetime. These models aim to estimate the life expectancy of people under different environment. The GAM allows for capturing potential non-linear relationships between the predictors and the outcome.

The model is specified as follows:

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

Where:

- $Lifeexpectancy\_developed$ is the life expectancy of developed country.
- $Lifeexpectancy\_developing$ is the life expectancy of developing country.
- $PercentageExpenditure$ is a smooth function of the runner's age at the time of the race.
- $Continent$ is
- $TotalExpenditure$ is the runner's gender (treated as a categorical variable).
- $IncomeComposition$ is a smooth function of the total number of NYRR races the runner has participated in.
- $Schooling$ is the country the runner is from under IAAF standards.
- ϵ_i is the error term, assumed to be normally distributed with a mean of 0 and constant variance $\epsilon_i \sim N(0, \sigma^2)$

This GAM model allows for smooth, non-linear effects of continuous predictors (**age** and **races_count**) while keeping the categorical predictors (**gender** and **iaaf_category**) in a linear framework.

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix C.

We run the model in R (R Core Team 2023) using the **rstanarm** package of Goodrich et al. (2022). We use the default priors from **rstanarm**.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table **??**.

1. check different continent has its different life expectancy.
2. compare the develop country with
3. check the shcooling year of develop countries to developing country, showing a relationship between shcooling with the life expenctancy.
4.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

A Appendix

A.1 Data Cleaning Notes

We began by importing the raw dataset using the read_csv function from the tidyverse package. To focus our analysis on more relevant variables, we selected specific columns, such as percentage of expenditure of the country, total health expenditure by government .etc., omitting any unnecessary columns.

We filter out the rows containing NA values in any of the selected columns for reducing the noise and simpler further analysis..

We then renaming columns for clarity. For example, we changed 'Income.composition.of.resources' into 'IncomeComposition' , making it easier for anyone working with the data to read and understand what each variable represents.

Each column is rounded using the round function, specifying the desired number of decimal places for each. Columns not mentioned in mutate remain unchanged. This ensures a flexible and precise cleaning process tailored to further compairation and graphing.

We also modified the country name of 'Republic of Korea' under Country variable to 'South Korea' to avoid misunderstanding of North Korea.

We also created a new variable called Continent, which indicates which continent does the country comes from in order to provide a geographical context to the analysis. Life expectancy may be higher in developed regions like Europe or North America compared to regions like Sub-Saharan Africa due to differences in healthcare, living standards, and economic development.

We also wrapped the cleaning process in a tryCatch block in order to mitigate any errors that arose throughout the cleaning process.

After completing the cleaning, we saved the final dataset in both Parquet and CSV formats for later analysis.

A.2 Idealized Survey

Survey: Understanding Life Expectancy and Influencing Factor

Thank you for participating in this survey. This survey aims to gather insights into the factors influencing life expectancy, including health expenditure, education, ethnicity, government spending on healthcare, and income levels. Your responses will help us understand individual perspectives and experiences towards national policy. Participation is voluntary, and your answers will remain anonymous.

Contact Information: If you have any questions about the survey or the data collection process, please contact

Survey Coordinator: Yanfei Huang
Email: yanfei.huang@mail.utoronto.ca

Section 1: Personal Health Expenditure

1.What percentage of your monthly income do you spend on healthcare (e.g., insurance, medications, doctor visits)?

- Less than 5%
- 5%–10%
- 10%–20%
- More than 20%

2.Do you or your household have health insurance coverage?

- Yes
- No

3. How frequently do you visit a healthcare professional(eg.doctors, nurse) in a year?

- 0-1 times
- 2-5 times
- 6-10 times
- More than 10 times

Section 2: Education Background

4.What is the highest level of education you have completed?

- No formal education
- Primary school
- Secondary school
- College or university degree
- Postgraduate degree

5.How many years of formal schooling have you completed?

Please write the number: _____

Section 3: Ethnicity and Geographic Factors

6.Which of the following best describes your ethnicity?

- African
- Asian
- European
- Hispanic or Latino

- Middle Eastern
- Indigenous
- Other (please specify): _____

7.What's the distance to the nearest healthcare facility from your residence?

- Less than 1 km
- 1-5 km
- More than 5 km

Section 4: Government Expenditure on Healthcare

8.Are health services in your country subsidized by the govenment?

- Yes, fully subsidized
- Yes, partially subsidized
- No

9. What is the approximate cost of your most recent healthcare visit (in local currency)?

Please write the amount _____

10.Do public healthcare facilities in your area provide all the services you need?

- Yes
- No
- Not applicable

Section 5: Income and Development Factor

11.What is your approximate monthly household income after taxes?(in local currency)

- Less than 1,000
- 1,000 - 2,999
- 3,000 - 4,999
- 5,000 - 7,999
- 8,000 - 10,999
- More than 11,000
- Prefer not to say

12.How would you describe your current financial situation?

- Struggling to make ends meet
- Just getting by
- Comfortable, but not wealthy

- Financially secure
- Wealthy
- Prefer not to say

13.How often do you have difficulty affording basic healthcare (medical visits, medications, etc.)?

- Never
- Rarely
- Sometimes
- Often
- Always
- Prefer not to say

Final Section

Thank you for completing this survey! Your responses will help capture information about economic conditions, access to healthcare and education, and basic living standards, which are critical for understanding life expectancy predictors and the impact of socioeconomic factors on health outcomes.

B Additional data details

C Model details

C.1 Additional graph for analysis

C.2 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

C.3 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

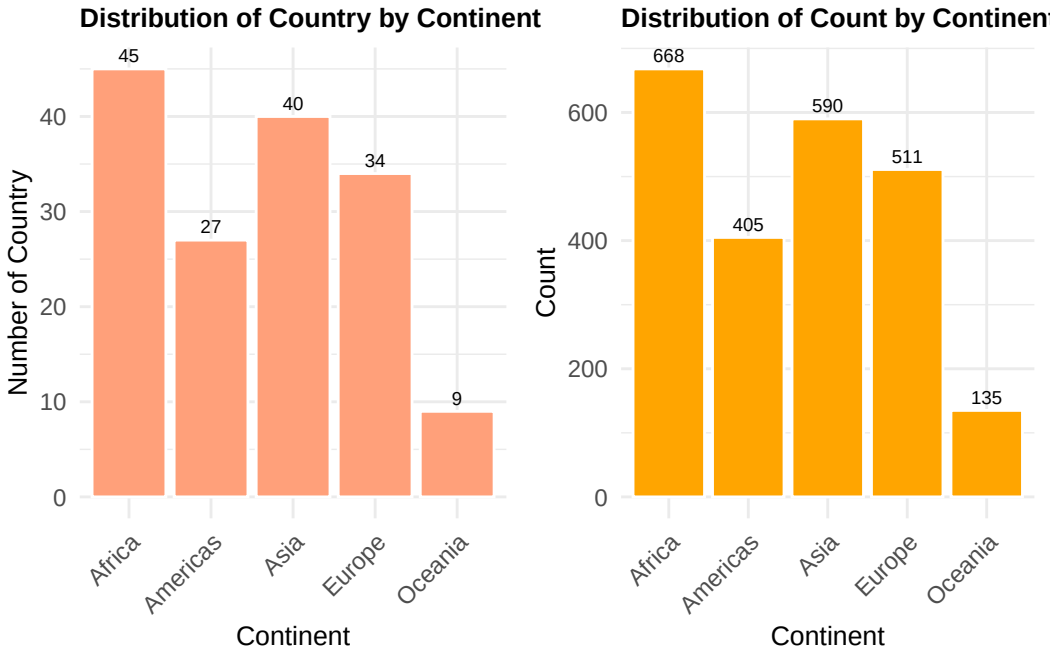


Figure 4: Distribution of contient. Africa has the highest number of countries represented, followed by Asia among all 5 continents. The distribution of count of data aslo shows that a significant portion of the data is focused on African and Asian nations.

Examining how the model fits, and is affected by, the data

Checking the convergence of the MCMC algorithm

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.