Global Life Expectancy Trends: A Multivariate Analysis of Influential Indicators*

Dingning Li

2024-04-18

Life expectancy serves as a critical indicator of a nation's overall health and social progress. This study conducts a multivariate analysis of global health indicators to identify key predictors of life expectancy. The analysis combines these variables and builds models that reveaal how eduaction, disease burden, and socioeconomic factors combine to influence life expectancy. The findings provide valuable insights for policymakers and health practitioners, highlighting the importance of strategies to improve life expectancy.

Table of contents

*code and data are avaliable at:https://github.com/iamldn2002/life_expectancy

Introduction

Life expectancy remains one of the most indicators of public health and social progress. While it is influenced by various factors from genetic predispositions, life choices to socioeconomic determinants. These factors have continually contributed to life expectancy. The work by Wilkinson (Wilkinson and Pickett 2006) demonstrated a compelling correlation between income distribution and life expectancy, suggesting that more equitable societies tend to have healthier populations. This relationship underscores the relationship between societal wealth, resource allocation, and health outcomes, making it a subject of ongoing investigation.

The dynamics of life expectancy are not static; they improve as societies progress. (Caselli et al. 2015) provided a study of the shifts in life expectancy over time, discussing the historical trends, current standings, and future projections. Their analysis focus on the factors that have historically underpinned improvements in health and longevity, while also considering the potential challenges that may arise in the future. Despite the advancements in healthcare and medicine, disparities in life expectancy continue to pose challenges globally. The analysis by Mustard (Mustard and Lavis 1999) explored the determinants of national life expectancy, highlighting the role of public policy and the allocation of health care resources . Their investigation highlights the diverse determinants that affect life expectancy across nations underscores the importance of strategies that address both health care systems and broader societal factors.

Building on foundational studies, our research seeks to further explore the determinants of life expectancy. Specifically, the question guiding our study is: How do factors such as a country's socioeconomic status, educational attainment, and health infrastructure collectively influence life expectancy? By adopting a Bayesian approach, this study not only considers the individual contributions of these factors but also addresses the uncertainty inherent in the system. In

doing so, we aim to provide a comprehensive understanding of how these determinants interact to shape life expectancy, offering insights that could inform future research, policy-making, and health care strategies aimed at enhancing longevity and quality of life.

Estimand

The estimand of this study is to quantify the impact of adult mortality rates, alcohol consumption, HIV/AIDS prevalence, schooling years, and country status (developing or developed) on life expectancy, aiming to understand the individual and collective contributions of these factors to variations in future life expectancy.

Data

Data source

The data obtained from the World Health Organization (WHO) provides a view of health outcomes across different countries and is helpful for international health assessments. The WHO gathers this data through a combination of country-reported statistics, health system surveillance, and collaborative international studies. The life expectancy figure, a key metric within this dataset, encapsulates a wide range of factors, including but not limited to, the prevalence of various diseases, the effectiveness of health care systems, and a multitude of socio-economic factors that influence public health.

Data cleaning

The dataset initially includes variousfactors potentially influencing life expectancy, the dataset covers various dimensions, from infectious diseases to lifestyle choices and educational level. To focus the study on the most impactful predictors and to streamline the analysis, the dataset was restricted to the year 2013. Within this temporal scope, particular attention was given to consolidating disease-related variables, selecting only those with the broadest implications on public health—measles and HIV/AIDS. After cleaning, the analysis was focused on seven key factors: 'adult mortality', 'alcohol', 'measles', 'bmi', 'hiv/aids', 'socioeconomic status', and 'schooling'. This refinement was intended to distill the dataset down to the most influential variables, thereby enhancing the clarity and precision of the statistical analysis.

The selected packages, known for their reliability and functionality in the R(R Core Team 2023) community, included tidyverse(Wickham et al. 2019a), dplyr(Wickham et al. 2023), ggplot2(Wickham 2016), kableExtra(Zhu 2024), tidyr(Wickham et al. 2019b), readr(Wickham, Hester, and Bryan 2024). I also utilize 'rstanarm'(Goodrich et al. 2024) to build bayesian model and 'bayesplot' (Gabry et al. 2019) to graph the bayesian model. 'modelsummary'

package(Arel-Bundock 2022) is used to summary models I built in this paper for comparing and analyzing results easily. These tools transform the raw data into a polished dataset ready for multivariate analysis.

Measurement

The methodology of data collection by the WHO involves a rigorous process of validation and cross-reference with other global health databases to ensure accuracy. Data points like alcohol consumption are typically collected from health surveys and sales data, while figures for measles and HIV/AIDS prevalence are often taken from both health reports and epidemiological surveillance. The connection between this data and real-world implications is evident as it informs public health policy, medical research priorities, and international aid distribution. By identifying trends and disparities in life expectancy, stakeholders can devise targeted interventions to address specific health issues and thereby work towards the WHO's overarching goal of improving health worldwide. For instance, a country showing a low life expectancy due to high infectious disease rates may benefit from increased vaccinations and the establishment of better sanitary practices.

Variables Explainations

Here is how we defined and measured key variables:

Country: the nation to which the data row corresponds, usually a categorical variable

Status: denotes the development status of a country, categorized as "Developed" or "Undeveloped"

life_expectancy: a statistical measure of the average time people in a nation is expected to live

adult_mortality: represents the likelihood of dying between the ages of 15 and 60 per 1000 population.

alcohol: represents several different measures related to alcohol, such as a average alcohol consumption per person within a country, measured in liters of pure alcohol consumed per year.

hiv/aids: refers to the prevalence or incidence of HIV/AIDS in a population. It may also represent the number of individuals living with HIV/AIDS or the number of deaths due to HIV/AIDS

Table 1: Sample of the Cleaned Data

Country	Year	Status	Life Expectancy	Adult Mortality	Alcohol	HIV/AIDS	Schooling
Afghanistan	2015	Developing	65.0	263	0.01	0.1	10.1
Afghanistan	2014	Developing	59.9	271	0.01	0.1	10.0
Afghanistan	2013	Developing	59.9	268	0.01	0.1	9.9
Afghanistan	2012	Developing	59.5	272	0.01	0.1	9.8
Afghanistan	2011	Developing	59.2	275	0.01	0.1	9.5

schooling: measures the average number of years of schooling received by people in a country. It can be used as an indicator of a country's educational system and socioeconomic status.

Data Visualization

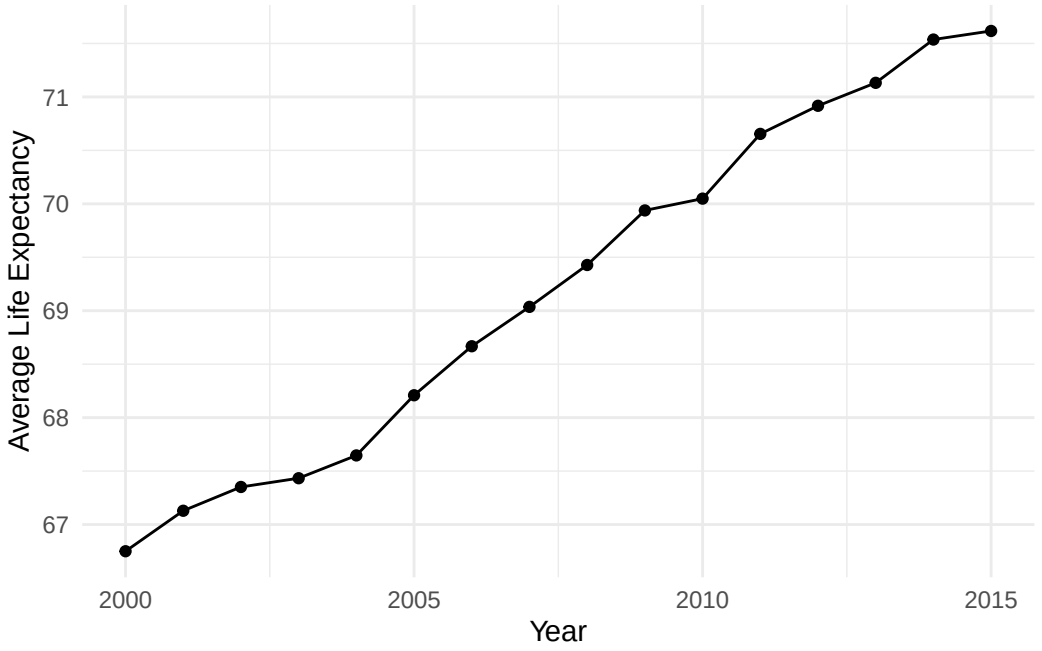


Figure 1: Trend of Yearly Average Life Expectancy

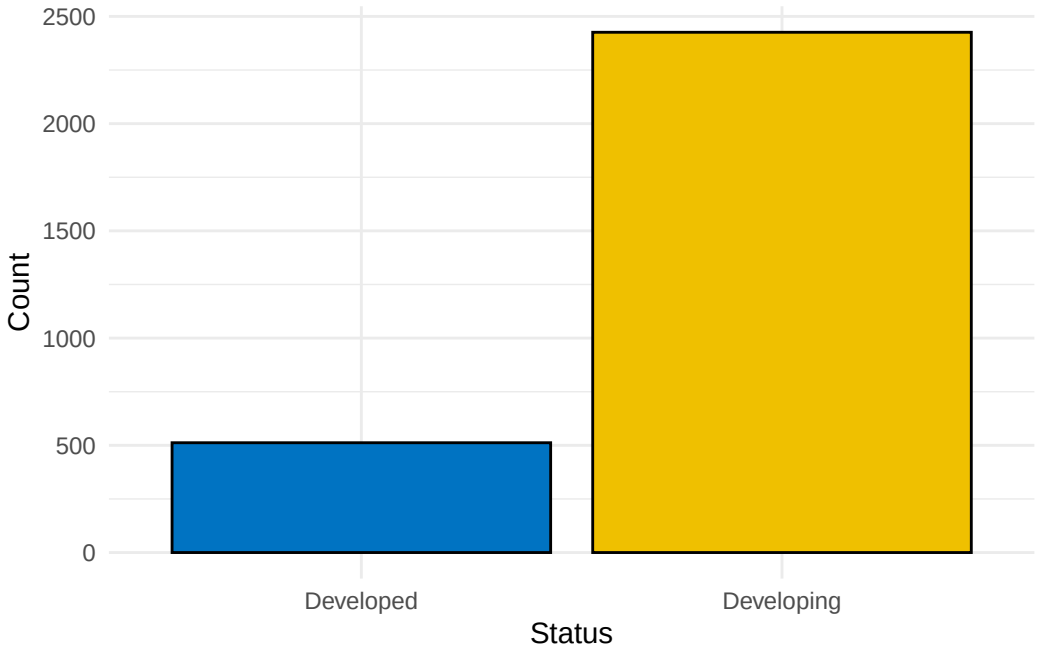


Figure 2: Numbers of Country in different Status

The Figure 2 shows the count of countries classified by their development status, categorized into "Developed" and "Developing". The x-axis represents the two categories of the status variable, and the y-axis represents the count of countries within each category. From what can be observed, the count of developing countries is much higher than that of developed countries based on the provided classification criteria. This type of graph is typically used to visualize the distribution of categorical data, allowing for an easy comparison of the number of occurrences in each category.

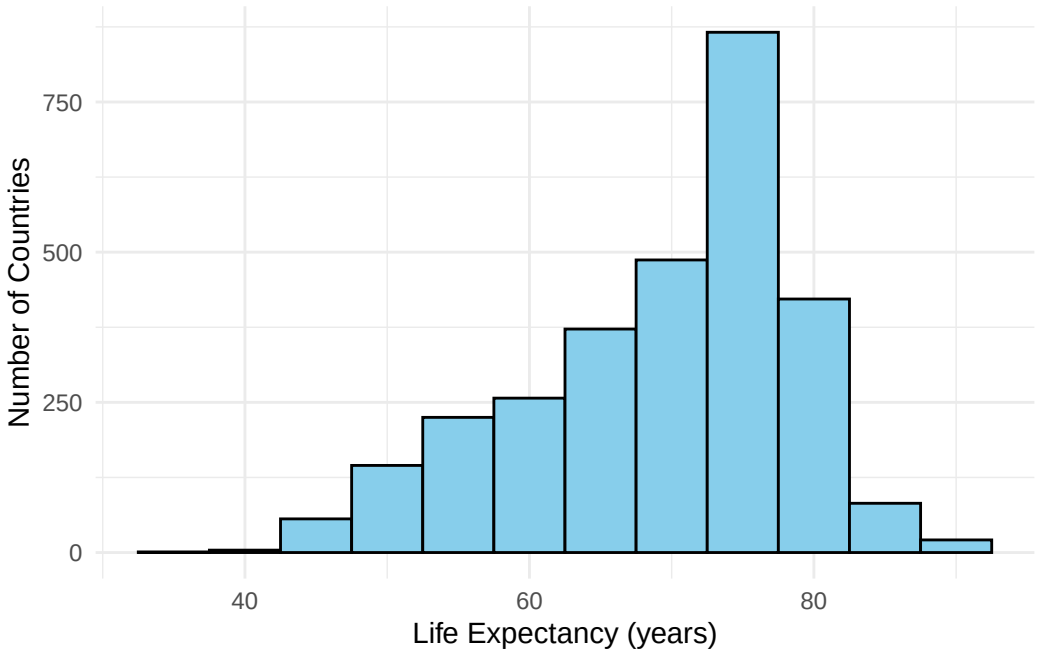


Figure 3: Histogram of Life Expectancy Around the World

The Figure 3 displays the distribution of life expectancy across a set of countries. We can observe that there is a concentration of countries with life expectancy in the 70-75 year range, as indicated by the tallest bar. The distribution is right-skewed, meaning there are fewer countries with very high life expectancies (80 years and above), and the bars to the left (indicating lower life expectancy) are shorter, suggesting fewer countries with lower life expectancies. It suggests that while most countries have a life expectancy between 65 and 75 years, there are relatively fewer countries with both very high and very low life expectancies.

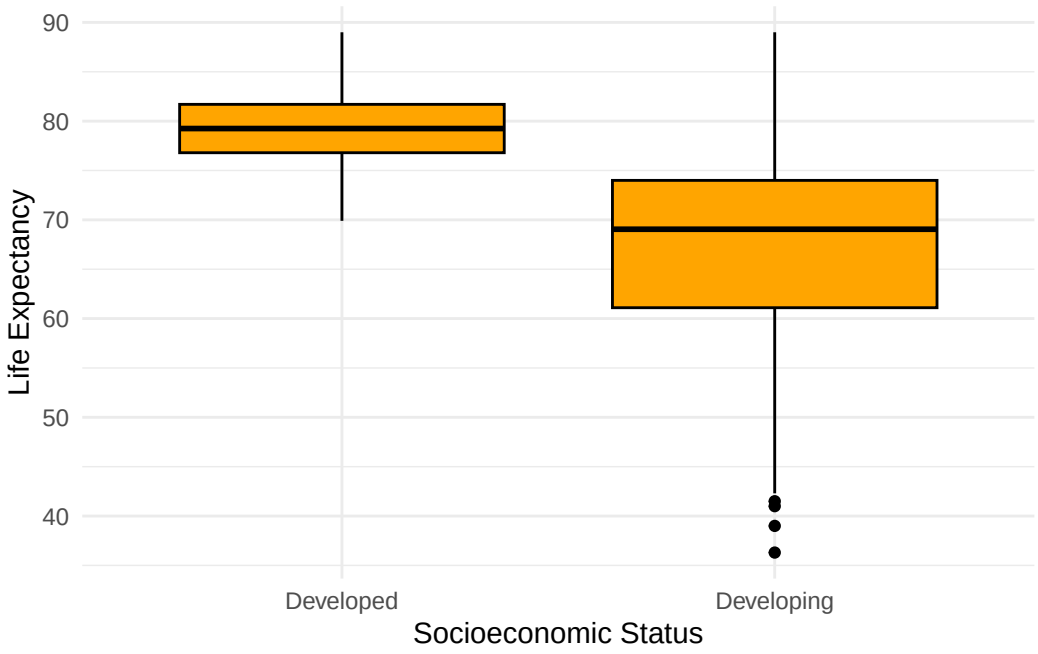


Figure 4: Life Expectancy by Socioeconomic Status

From Figure 4, central line in each box represents the median life expectancy for each category, representing the midpoint of the data.we can interpret that Developed countries have a higher median life expectancy than Developing countries. The range of life expectancy values is also slightly more spread out for Developed countries, indicating more variability within this group compared to Developing countries. The lack of data points outside the whiskers suggests there may not be extreme outliers in the life expectancy data for these two categories.

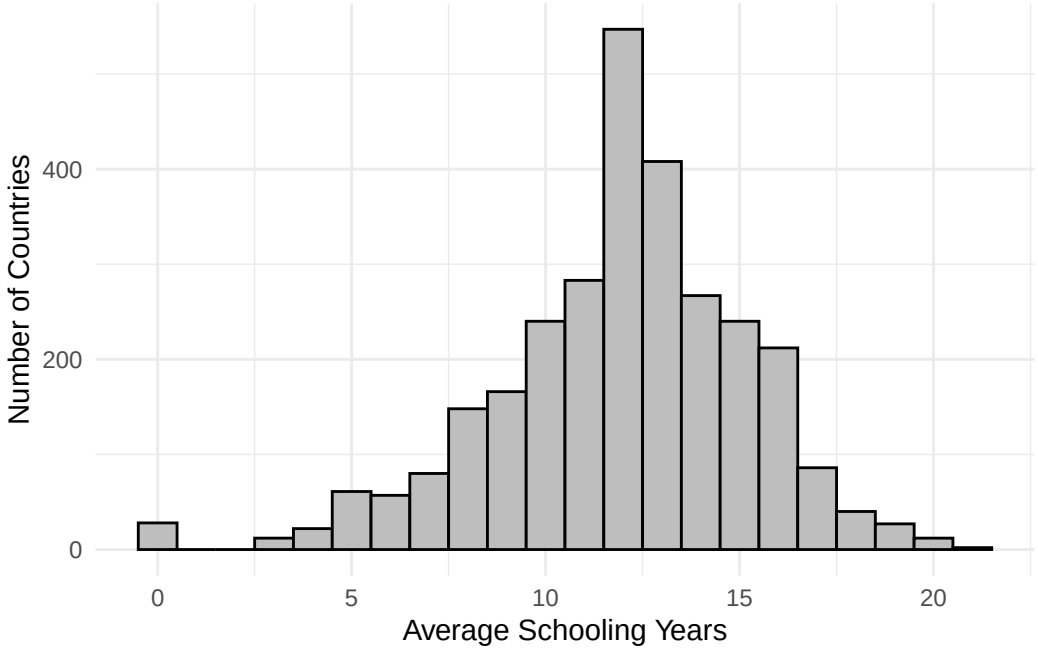


Figure 5: Distribution of Average Schooling Years

Model

Model Setup

Simple Linear Model

Firstly, a simple linear model was built to examine he relationship between just two variables, we can isolate the effect of the predictor on the outcome without the confounding influence of other variables.

$$\mu_i = \beta_0 + \beta_1 A_i$$

However, the simplicity of this model is also a limitation because it does not account for the multifaceted nature of most real-world phenomena. For instance, while adult mortality might influence the response variable, other factors like alcohol consumption, HIV/AIDS prevalence, and socioeconomic status might also play roles. Ignoring these could lead to biased to our expectations or incomplete conclusions.

Multiple Linear Model

A multiple linear model(baayesian linear model) is built for extending the simple linear framework by including multiple predictor variables, allowing us to examine the simultaneous effect of several factors on the response variable.

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \cdot \text{adult\_mortality}_i + \beta_2 \cdot \text{alcohol}_i \\
&\quad + \beta_3 \cdot \text{hiv\_aids}_i + \beta_4 \cdot \text{status}_i \\
&\quad + \beta_5 \cdot \text{schooling}_i \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(50, 10) \\
\beta_2 &\sim \text{Normal}(3, 2.5) \\
\beta_3 &\sim \text{Normal}(0.1, 2.5) \\
\beta_4 &\sim \text{Beta}(4, 1) \\
\beta_5 &\sim \text{Normal}(10, 2) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

We assume all the data of independent variables are normally distributed except status, suggests that when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed. This is especially true for large datasets.

Adjusting the priors to reflect actual conditions, such as setting the mean schooling years to 10, ensures that our model aligns with empirical observations. This practice enhances the realism of our Bayesian analysis, grounding our expectations in observable data. For example, the mean and standard deviation for 'adult mortality' are chosen to reflect typical values but with enough variance to capture differences across populations.

The binary variable (in this case 'status') is that there is an 80% chance of it being 1 (developing) and a 20% chance of it being 0 (developed), reflecting a strong prior belief that the 'status' is more likely to be developing since there is a consensus that the number of developing coiuntries is far more than developed countries. .

Model Justification

In the evaluation of predictive models, the Root Mean Square Error (RMSE) serves as a critical measure of accuracy, quantifying the average magnitude of the errors between predicted and observed outcomes. In my case, the RMSE is particularly relevant as it directly translates to the average number of years the model's life expectancy predictions deviate from the actual

observed figures. So I chose to compare the RMSE of two models to determine the model choice.

As Table 2 shows, the simple linear model yielded an RMSE of 8.33, which indicates that the model missed the actual life expectancy by 8.33 years. While this provides an initial baseline for model performance, this error suggests that the simple model might not be capturing all the relevant factors influencing life expectancy.

In contrast, the multiple linear model, which consider various predictors achieved a lower RMSE of 4.67. This improvement is substantial; the multiple model's predictions are closer to the observed data by approximately 3.66 years compared to the simple model. The reduction in RMSE by nearly half indicates that the additional variables in the multiple linear model contribute essential information that better captures the complexities of life expectancy.

Results

Model Results

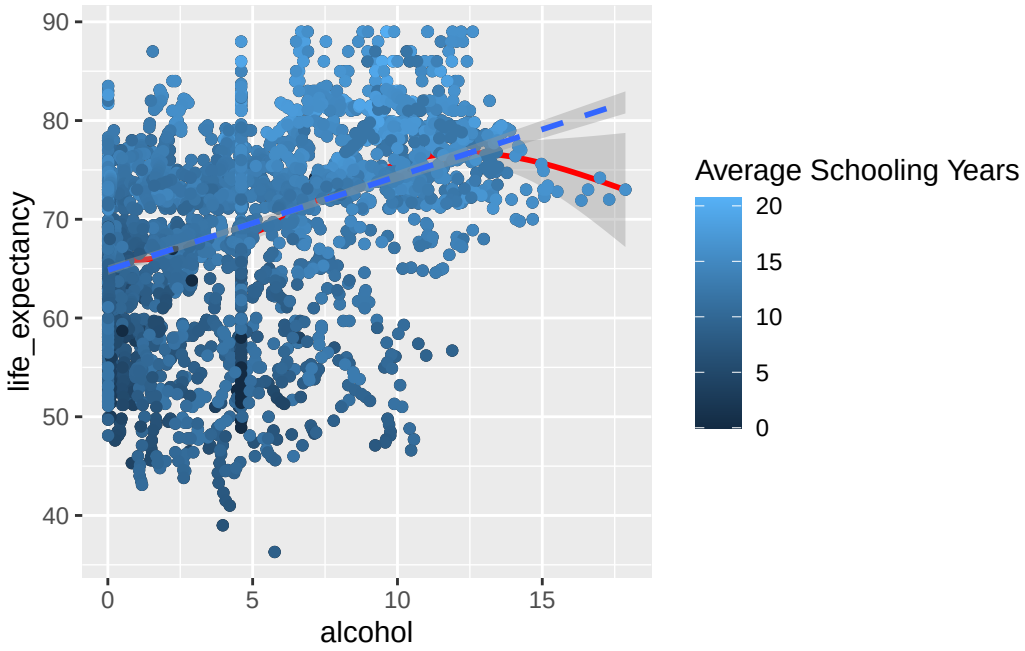


Figure 6: Relationship Between Life Expectancy, Alcohol, and Schooling

Table 2 presents the simple and multiple linear regression models. In the simple linear model, only socioeconomic status for a country is used as independent variable, with an adjusted R-squared value of 0.232. This indicates that approximately 23.2% of the variance in life expectancy is accounted for country's socioeconomic status. The intercept suggest that the

Table 2: Summary for simple and multiple linear models

	Simple Linear Model	Multiple Linear Model
(Intercept)	79.20	59.95
	(0.37)	
statusDeveloping	−12.08	−2.43
	(0.41)	
adult_mortality		−0.02
alcohol		0.09
hiv_aids		−0.51
schooling		1.31
Num.Obs.	2938	2938
R2	0.232	0.758
R2 Adj.	0.232	0.757
AIC	20 799.4	
BIC	20 817.3	
Log.Lik.	−10 396.683	−8700.799
F	888.349	
ELPD		−8709.9
ELPD s.e.		58.8
LOOIC		17 419.7
LOOIC s.e.		117.5
WAIC		17 419.7
RMSE	8.33	4.67

baseline life expectancy, in the indicator of status, is 79.02 years, with implication of being developing country is associated with a reduction in life expectancy by 12.08 years.

The multiple linear model, which considers various additional predictors - alcohol consumption, HIV/AIDS prevalence, average years of schooling, and binary variable for socioeconomic status (Developing or Developed) offers another prediction for life expectancy. The adjusted R-squared value increases significantly from 0.485 to 0.757, indicating that the model accounts for around 75.7% of the variance in life expectancy across countries. This improvement highlights the importance of a multivariate approach when examining global health outcomes.

It is noticeable that estimated life expectancy decreased to 59.95 years. In reality, this might not be a feasible scenario, but it provides a baseline from which the effects of other variables are measured. Adult mortality still remains a large predictor but with a reduced coefficient, indicating a reduced individual impact when other factors are considered. Interestingly, the positive coefficient 0.09 suggests that higher alcohol consumption is associated iwth a slight increase in life expectancy, which may reflect complex social and economic factors rather than direct causality. For example, wealthier countries having higher alcohol consumption and better healthcare, which can lead to longer life expectancy. Negative coefficient -0.51of HIV/AIDS indicates that a higher prevalence of HIV disease in a population is strongly associated with lower life expectancy. Every unit increase in HIV/AIDS prevalence is predicted to decrease life expectancy by 0.51 years, holding all else being equal. For average schooling years, this positive coefficient suggests a strong association between average years of schooling and life expectancy. Each additional year of schooling is associated with an increase in life expectancy by 1.31 years. The coefficient of last indicator, indicates that, on average, developing countries have a lower life expectancy compared to developed countries. The model predicts that being a developing country is associated with a reduction in life expectancy by 2.43 years.

Discussion

Key Findings

The analysis of the impact of various socioeconomic and health-related variables on life expectancy across different countries reveals several critical insights. First and foremost, the number of schooling years a population receives, combined with the overall socioeconomic status of the country, along with the prevalence of HIV/AIDS, emerge as the most important factors influencing life expectancy. This highlights the role of education in improving health outcomes and underscores the profound impact of economic conditions and major health crises like HIV/AIDS on a population's life span.

Conversely, the influence of adult mortality rates and alcohol consumption on life expectancy appears to be relatively minor in comparison. This is particularly interesting because it suggests that while adult mortality is a direct indicator of life expectancy, its relative importance is less than factors like education and socioeconomic status. Moreover, the analysis points to a

counterintuitive finding where increased alcohol consumption is associated with higher life expectancy. However, based on the related studies, such as those pointed out by Nova and other authors, suggest that moderate drinking may have health benefits, thereby extending life expectancy(E.Nova and A.Marcos 2012). This could potentially challenge common perceptions and warrants further investigation to understand the underlying causes, such as variations in drinking patterns and associated lifestyle choices, or possible errors in data collection or model assumptions that could lead to such an unexpected correlation.

Weakness and Bias

There are still many shortcomings in this pediction of life expectancy research, particularly in model building. The first weakness, and also the important weakness I want to point out is that predictor variables are limited in this study. The model's reliance on a limited set of predictor variables may overlook important determinants of life expectancy. Environmental factors, access to healthcare, cultural practices, and genetic predispositions are known to impact life expectancy but are not included in the model. Failing to account for these variables could lead to biased estimates and conclusions. The second weakness is about temporal dynamics. Life expectancy trends are subject to temporal dynamics influenced by evolving socio-economic conditions, public health policies, and medical advancements. However, linear regression models may fail to capture these dynamic relationships. Changes in the determinants of life expectancy over time, as well as their interactions, may be overlooked, limiting the model's ability to provide accurate long-term predictions.

There is also weakness of dataset I chose for the research. The variables represents a national or regional data rather than individuals. Linear regression models often assume homogeneity of effects across different population groups or countries. However, the determinants of life expectancy and the effectiveness of interventions may vary across diverse contexts. Ignoring this heterogeneity could lead to biased estimates and undermine the generalizability of the model's findings. As Fforde pointed out that this assumption, which posits that different countries can be viewed as members of a single homogeneous population, poses challenges to the persuasiveness of economic theory(Fforde 2005). In the same way, homegeneity assumption will also bring trouble to our model prediction

Future Research

Future studies could explore the temporal dynamics of life expectancy trends, focusing on how socio-economic factors, public health interventions, and environmental changes interact to shape life expectancy outcomes over time. Longitudinal analyses incorporating time-series data could elucidate the short-term and long-term effects of policy interventions, economic fluctuations, and societal changes on life expectancy variations. Additionally, investigating the differential impact of temporal trends on various population subgroups, such as different

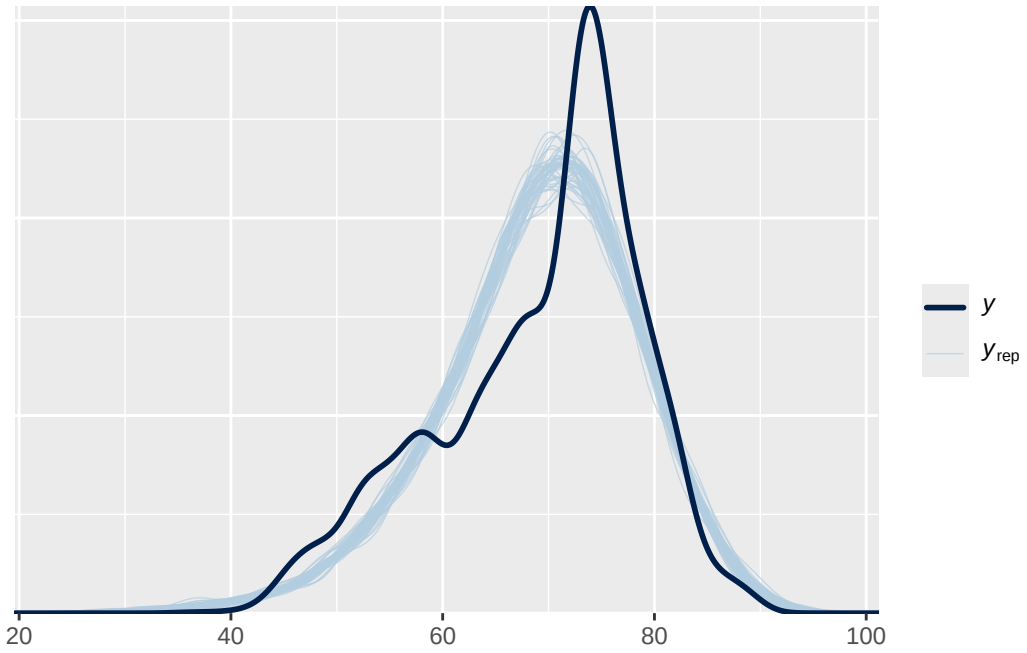
aged groups or socio-economic strata, could provide valuable insights into the mechanisms driving differences in life expectancy trends.

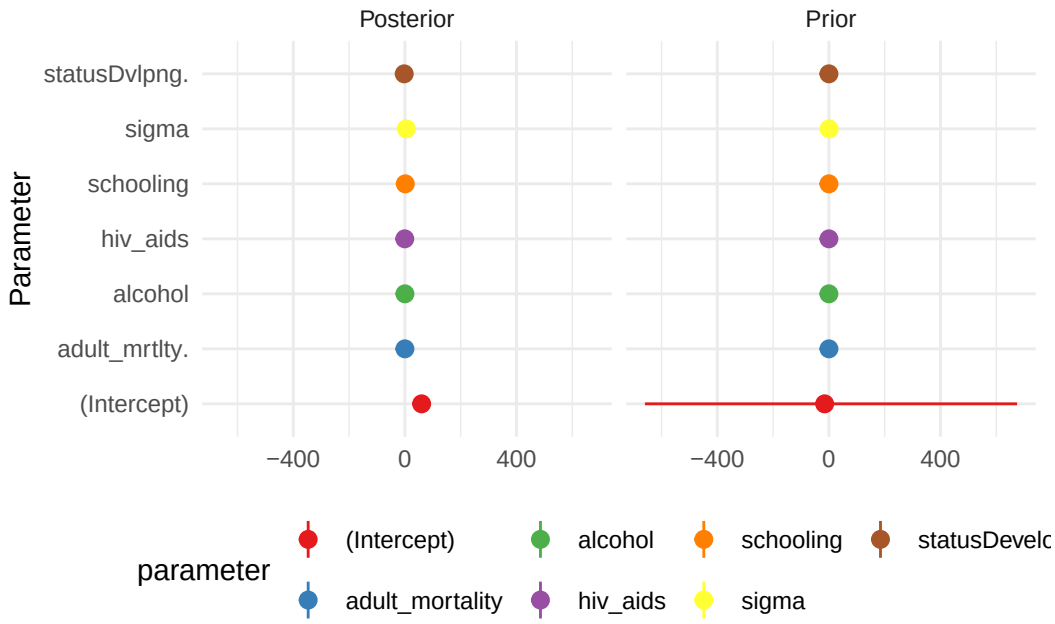
In future studies, researchers could also explore the idea of using a multi-level analysis framework to study the factors that influence life expectancy. This approach involves looking at various levels of influence, such as the individual, community, and societal levels. By doing this, researchers can gain a better understanding of how different factors at each level contribute to life expectancy outcomes. For example, they could examine how individual-level characteristics like lifestyle choices (such as diet and exercise), genetic predispositions, and access to healthcare services impact life expectancy. At the community level, factors like the quality of the neighborhood environment, availability of healthcare facilities, and social support networks could be considered. Additionally, researchers could investigate how broader social factors such as government policies, cultural norms, and economic conditions influence life expectancy trends.

Appendix

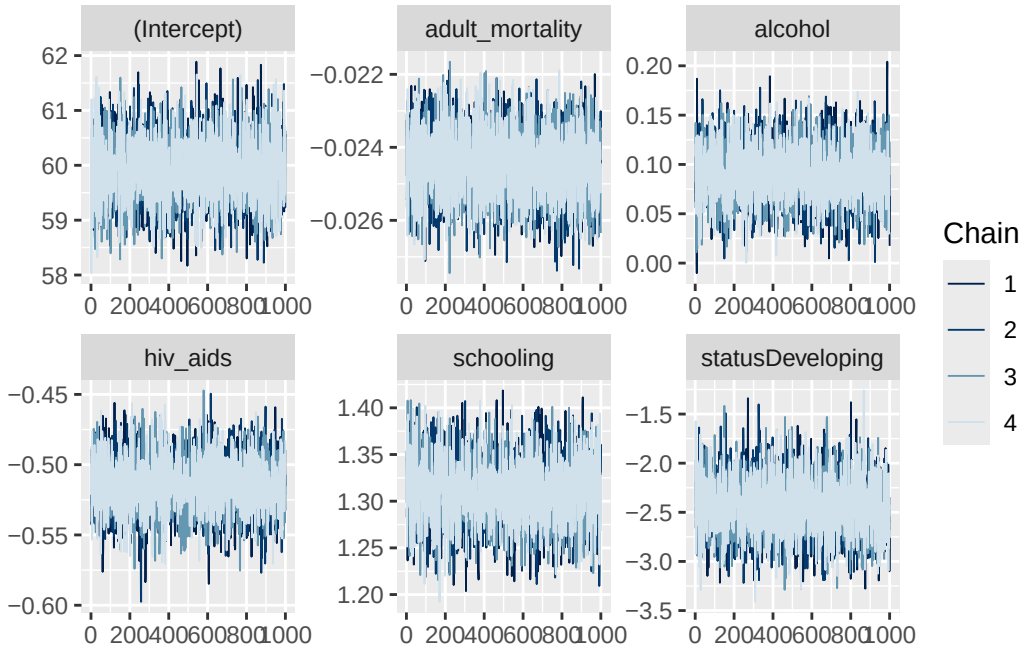
B Model details

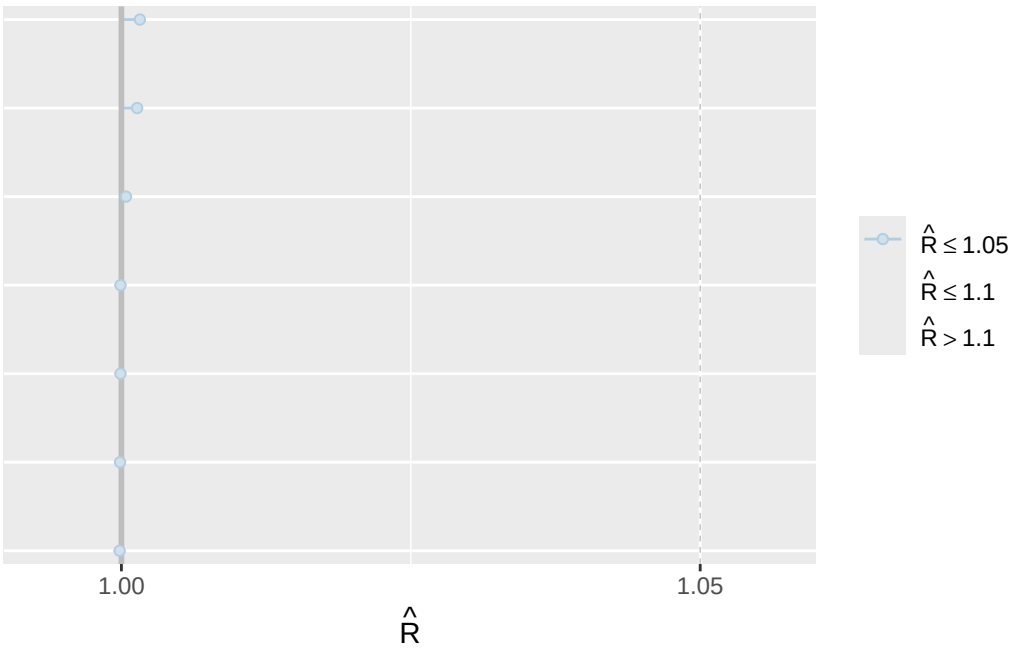
Posterior Predictive Check





MCMC Traceplot




1.00
1.05
$\hat{R}$
$\hat{R} \leq 1.05$
$\hat{R} \leq 1.1$
$\hat{R} > 1.1$

Reference

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software*. https://doi.org/10.18637/jss.v103.i01.

Caselli, Graziella, Jacques Vallin, James W Vaupel, et al. 2015. "Past, Present, and Future of Healthy Life Expectancy." *Cold Spring Harbor Perspectives in Medicine* 5 (11): a025957.

E.Nova, A.Veses, G. C.Baccan, and A.Marcos. 2012. "Potential Health Benefits of Moderate Alcohol Consumption: Current Perspectives in Research." *Proceedings of the Nutrition Society* 71 (2): 307–15. https://doi.org/Unknown DOI.

Fforde, Adam. 2005. "Persuasion: Reflections on Economics, Data, and the 'Homogeneity Assumption'." *Journal of Economic Methodology* 12 (23): 63–91. https://doi.org/https://doi.org/10.1080/1350178042000330904.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. "Visualization in Bayesian Workflow." *J. R. Stat. Soc. A*. https://doi.org/10.1111/rssa.12378.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Mustard, Cameron A, and John N Lavis. 1999. "Determinants of National Life Expectancy." *Canadian Journal of Public Health* 90: S32–36.

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

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019a. *Welcome to the tidyverse. Journal of Open Source Software*. Vol. 4. https://doi.org/10.21105/joss.01686.

———, et al. 2019b. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Wilkinson, Richard G, and Kate E Pickett. 2006. "Income Inequality and Population Health: A Review and Explanation of the Evidence." *Social Science & Medicine* 62 (7): 1768–84.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.