What to Expect with Life Expectancy?*

Examining How Status, Healthcare Expenditure, Education, and Medical Infrastructure Influence Life Expectancy Across Developed and Developing Regions

Manjun Zhu

November 27, 2024

Life expectancy has seen unprecedented growth in recent decades, driven by advancements in healthcare, improved access to medical services, and reduced child mortality. However, this growth has not been evenly distributed across the globe, with wealthier nations benefiting disproportionately. This research examines the impact of a nation's economic standing and health conditions on life expectancy, exploring nuanced variables beyond those traditionally studied. By developing a linear model, the analysis identifies key relationships between economic and healthcare indicators and their influence on population longevity. The findings aim to inform health policy and economic planning, providing actionable insights to address disparities and promote equitable growth in life expectancy worldwide.

1 Introduction

Life expectancy has witnessed remarkable growth over the past century, driven by advancements in healthcare, improved access to medical services, and reduced child mortality rates. However, this progress has not been distributed equitably across the globe. While nations with robust economies and advanced healthcare systems have achieved significant gains, others, particularly in underdeveloped regions, continue to face barriers such as inadequate infrastructure, persistent diseases, and economic instability. Despite existing research on life expectancy determinants, gaps remain in understanding the nuanced relationships between economic standing, healthcare access, and other critical variables.

This study aims to address these gaps by developing a linear model to analyze the impact of a nation's status, healthcare factors, and economic indicators on life expectancy. The response variable is Life Expectancy, while the predictors include Status (Developed vs. Developing),

*Code and data are available at: https://github.com/Karrrrmen/Life_Expectancy.

Diphtheria Immunization Coverage, Government Total Expenditure, Numbers of Doctors and Nurses per 10000, and Schooling. The analysis seeks to quantify the influence of these variables, providing insights into the disparities in life expectancy and identifying factors that contribute to these differences.

Previous studies provide valuable context for this research. One study (Garon et al. 2015) examined the long-term effects of diphtheria on survivors, revealing that paralyzed individuals experience greater reductions in life expectancy, which underscores the importance of healthcare infrastructure in mitigating preventable diseases in underdeveloped regions. Another study extended this understanding by exploring the role of the Human Development Index (HDI), emphasizing the influence of public health initiatives, economic conditions, and education on life expectancy in low and medium-HDI nations (Girum et al. 2018). Besides that, advanced econometric techniques, including panel cointegration analysis, have demonstrated a positive relationship between health care expenditure per capita and life expectancy, providing a robust foundation for integrating economic variables into this study (He and Li 2020). Additionally, a finding (Hanushek and Woessmann 2016) underscored the importance of combining both educational investments and government support in improving life expectancy, particularly in developing regions. The strong correlation between education and health suggests that higher levels of education improve overall health, reducing mortality rates and extending life expectancy.

The analysis revealed that life expectancy in developed countries is generally higher and more stable, attributed to stronger healthcare systems, consistent access to medical services, and robust economic stability. These factors are typically supported by well-established public health policies. In contrast, life expectancy in developing countries demonstrates a broader range of variation, though with a noticeable upward trend. This improvement is driven by diverse factors like better healthcare access, economic development, and increasing resource availability, but the inconsistency across regions highlights disparities in infrastructure. The faster growth trajectory in these nations indicates progress, yet challenges remain in achieving stable health outcomes. These findings suggest that policymakers in developing countries should prioritize healthcare system strengthening, resource access, and reducing regional inequalities, while in developed countries, maintaining consistent healthcare access and supporting sustained economic growth should remain priorities.

The remainder of this paper is structured as follows: Section Section 2 discusses the data sources and cleaning methods, followed by an analysis of key predictors in Section Section 3. Section Section 4 presents the findings from the linear regression model, and Section Section 5 explores the implications of these findings, concluding with recommendations for policy interventions and areas for future research.

2 Data

2.1 Overview

This study utilizes two primary datasets related to global health and life expectancy. The first dataset, World Health Statistics 2020 Complete, sourced from Kaggle (Utkarsh 2020), provides comprehensive global health data, including variables such as GDP, life expectancy, mortality rates, and access to healthcare. The second dataset, Life Expectancy (WHO), also from Kaggle (Kumar 2017), focuses specifically on life expectancy trends and factors, offering a detailed breakdown by country, year, and socioeconomic variables.

Together, these datasets enable an in-depth exploration of the relationships between life expectancy and various predictors, such as economic indicators, healthcare access, and immunization coverage. Key variables include Life Expectancy, Status (Developed vs. Developing), Diphtheria Immunization Coverage, Government Health Expenditure as percentage of Total Expenditure, Numbers of Doctors and Nurses per 10000, and Schooling.

We performed data cleaning and analysis using R (R Core Team 2023), incorporating a range of packages for efficient data manipulation, visualization, and statistical modeling, leveraging packages such as `tidyverse` for data manipulation and visualization (Wickham et al. 2019), `dplyr` for streamlined data wrangling (Wickham et al. 2023), `ggplot2` for creating high-quality plots (Wickham 2016), `ggridges` for displaying the distribution of variables across different categories with ridge plots (Wilke 2024), `here` for managing file paths (Müller 2020), `lubridate` for handling date-related variables (Grolemund and Wickham 2011), `kableExtra` for generating polished tables (Zhu 2024), `modelsummary` for summarizing model results (Arel-Bundock 2022).

2.2 Measurement

The measurement process refers to how real-world phenomena—such as healthcare access, economic indicators, or demographic factors—are translated into numerical entries in the dataset. Each entry corresponds to a specific country and year, capturing variables related to life expectancy and associated determinants. Below is a description of key variables in the dataset:

Life Expectancy (`LifeExpectancy`): This is the dependent variable, representing the average number of years a person is expected to live from birth, assuming current mortality rates. The data is reported by international organizations like the World Health Organization (WHO) and is recorded in numeric format as age in years.

Status (`Status`): This categorical variable identifies whether a country is classified as "Developed" or "Developing." The classification is based on economic and social indicators defined by the World Health Organization (WHO) or related agencies.

Health Expenditure (`TotalExpenditure`): This represents the general government expenditure on health as a percentage of total government expenditure. It is reported by national governments or international organizations and recorded as a percentage in the dataset.

Diphtheria Immunization Coverage (`Diphtheria`): This variable reflects the percentage of one-year-olds who have received the DTP3 vaccine (diphtheria, tetanus toxoid, and pertussis). Immunization data is typically gathered through health system reports or household surveys and is recorded as a numeric percentage.

Schooling (`Schooling`): This measures the average number of years of schooling for individuals in the population. Data on schooling is sourced from education ministry reports or international databases like UNESCO, and it is recorded in years as a numeric value.

Medical Doctors (`pct_doctor`): This variable represents the number of medical doctors per 10,000 people in the population. It is typically reported by national health ministries or equivalent institutions. The dataset standardizes this figure as a numeric value to facilitate cross-country comparisons.

Nursing and Midwifery Personnel (`pct_nursing`): This measures the number of nursing and midwifery personnel per 10,000 people in the population. It is collected through similar national reporting systems and is recorded as a numeric value.

2.3 Data Cleaning

The raw Life Expectancy dataset, along with two supplementary datasets for medical doctors and nursing personnel, underwent several cleaning steps to ensure it was accurate, consistent, and ready for analysis. Initially, the columns in the supplementary datasets were renamed for clarity and compatibility with the main dataset. A left join was then performed to merge information on `pct_doctor` and `pct_nursing` into the main dataset. Subsequently, the data was filtered to include only entries from 2004 to 2015, and rows with missing values were removed to maintain data integrity. Column names were standardized, and data types were adjusted as necessary to ensure compatibility with analytical processes. The cleaned dataset was then saved as a Parquet file for efficient storage and further analysis.

2.4 Outcome Variables

The outcome variable is Life Expectancy. It represents the average number of years a person is expected to live, based on current health conditions and mortality rates. The model aims to understand how various factors, such as healthcare availability, education, and economic status, influence life expectancy across different countries. Figure 1 displays the historical data of life expectancy across 195 countries from 2004 to 2015, comparing developed and developing countries. The data reveals distinct trends in life expectancy, with developed countries generally experiencing higher life expectancy compared to their developing counterparts.

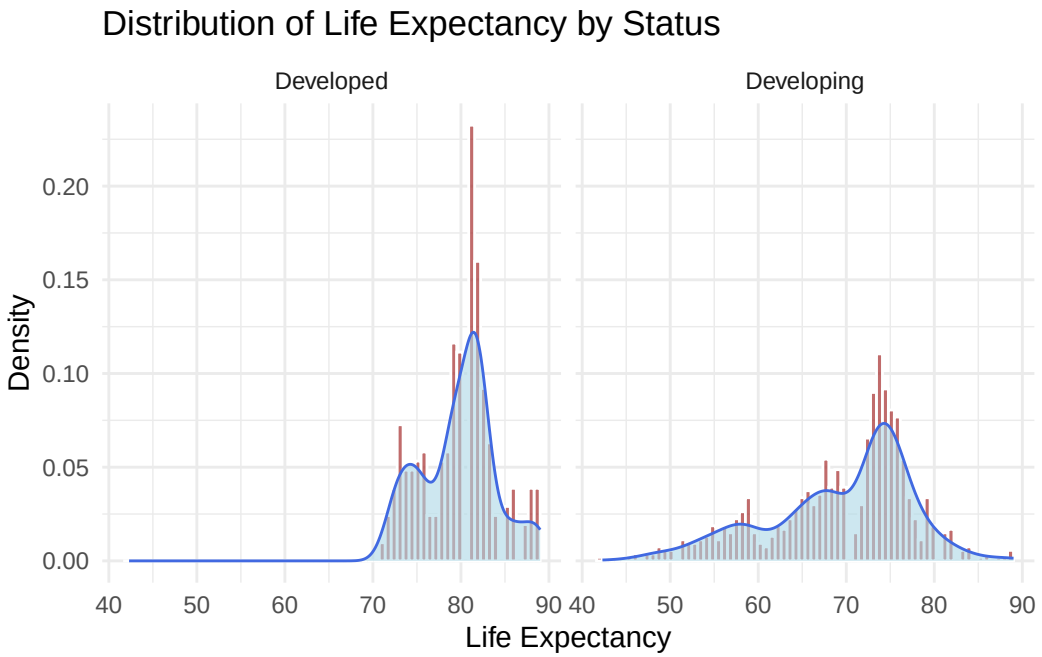

Figure 1: Distribution of life expectancy across different Status groups (Developed vs. Developing), highlighting the distinct patterns in their density. The mode of the density for life expectancy in developed countries tends to be around 81 years with a narrower variance, whereas in developing countries, the mode is approximately 74 years with a wider variance, reflecting disparities in health outcomes.

2.5 Predictor Variables

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

1. **Status (`Status`)**: This variable categorizes countries into "Developed" and "Developing" groups. The development status of a country plays a significant role in determining life expectancy, with developed countries generally having higher life expectancies due to better healthcare, sanitation, and infrastructure, as depicted in Figure 2.
2. **Diphtheria Immunization Coverage (`Diphtheria`)**: This variable represents the percentage of one-year-olds who have received the Diphtheria, Tetanus, and Pertussis (DTP3) immunization. Immunization coverage(Figure 3) is an important health indicator, as it reflects the effectiveness of a country's healthcare system in preventing infectious diseases, which in turn affects life expectancy.
3. **Schooling (`Schooling`)**: The number of years of schooling in a country is used as a proxy for education levels. Higher education levels typically correlate with better health outcomes and, consequently, higher life expectancy, as educated individuals are more likely to engage in healthier behaviors and have access to better healthcare. The Educated condition is shown in Figure 4.
4. **Medical Doctors (`pct_doctor`)**: This variable indicates the number of medical doctors per 10,000 people (Figure 6). Access to healthcare professionals is a key determinant of health outcomes, and a higher number of doctors generally corresponds with better healthcare services and longer life expectancy.
5. **Nursing and Midwifery Personnel (`pct_nursing`)**: Similar to the number of doctors, this variable represents the number of nursing and midwifery personnel per 10,000 population (Figure 7). Nurses play a crucial role in delivering healthcare, and higher availability of nursing staff can contribute to improved health outcomes and increased life expectancy.
6. **Health Expenditure (`TotalExpenditure`)**: This variable reflects the general government expenditure on health as a percentage of total government expenditure, as shown in Figure 5. Higher health spending is often associated with better healthcare infrastructure, which can lead to longer life expectancy due to better disease prevention and treatment options.

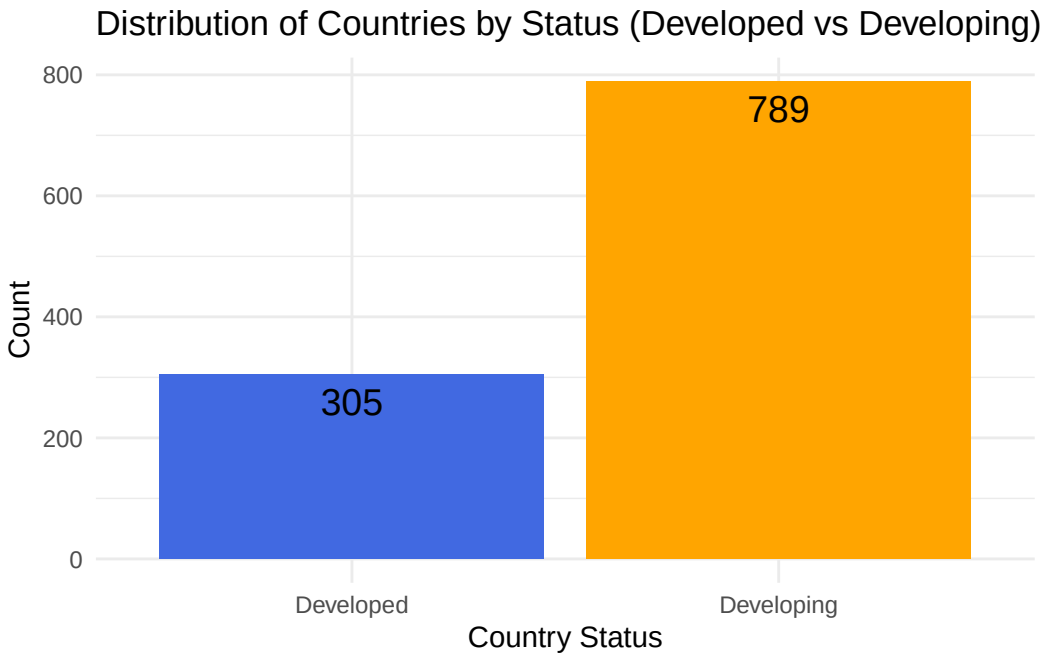

Figure 2: Histogram of Status, showing the counts of developed and developing countries in the dataset. There are 305 data points from developed countries and 788 from developing countries.

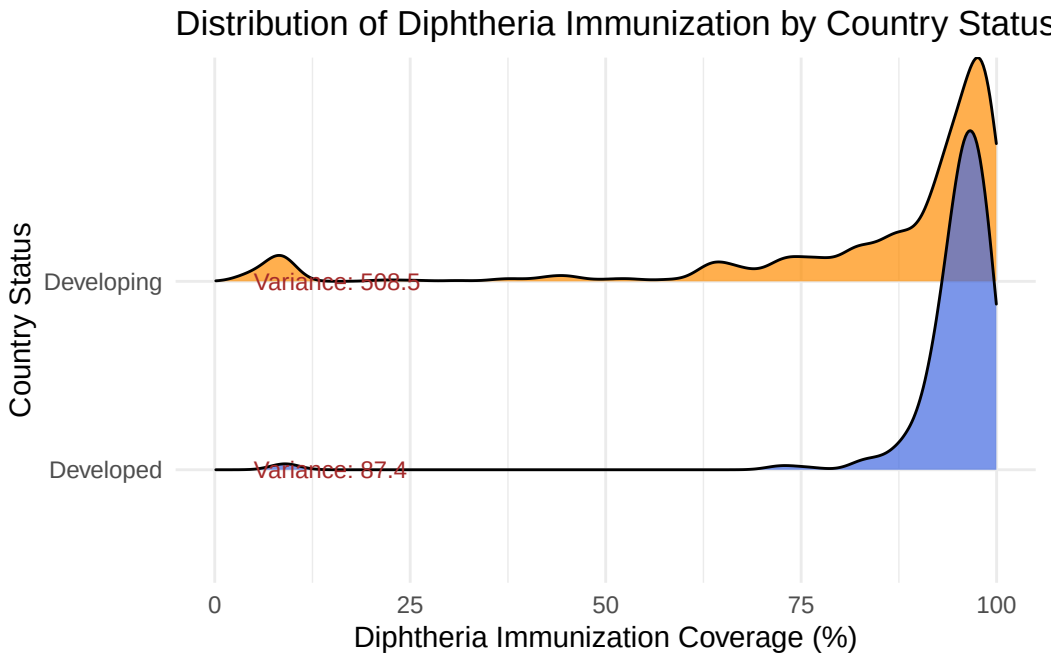

Figure 3: The plot shows similar immunization coverage conditions across developed and developing countries, with a noticeable increase in variance for developing countries due to some countries exhibiting low diphtheria immunization rates.

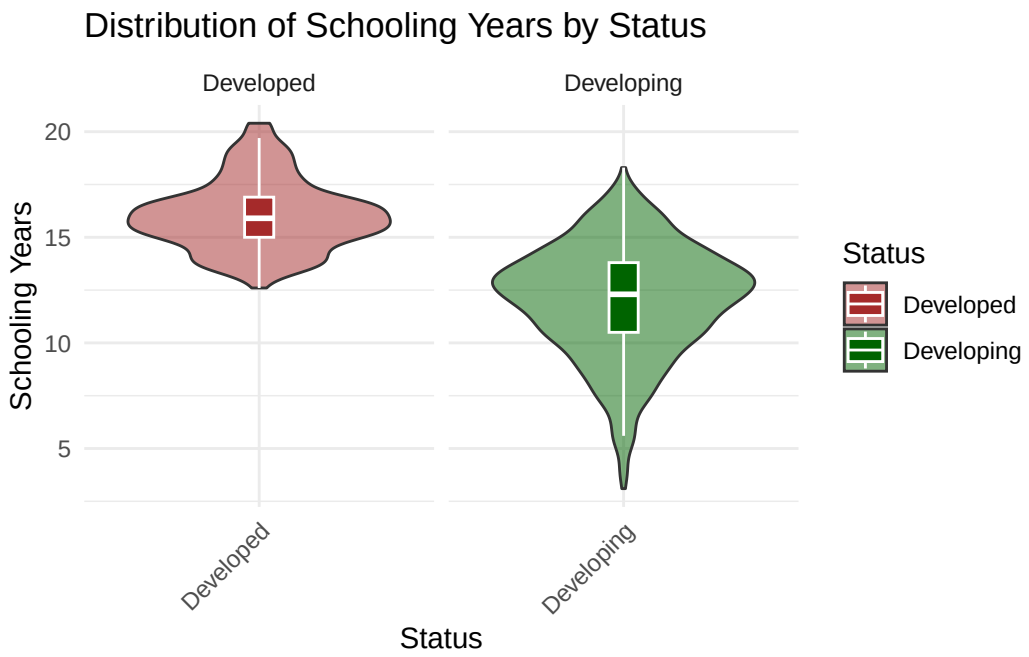

Figure 4: Violin plot with boxplot for Schooling by Status, highlighting a narrow distribution for developed regions with a higher median of 16 years of schooling. In contrast, developing countries show a wider variance and a relatively lower median of around 12.5 years, indicating significant disparities in educational access and attainment, with developed countries having more consistent and higher levels of education compared to the varied and generally lower levels in developing countries.

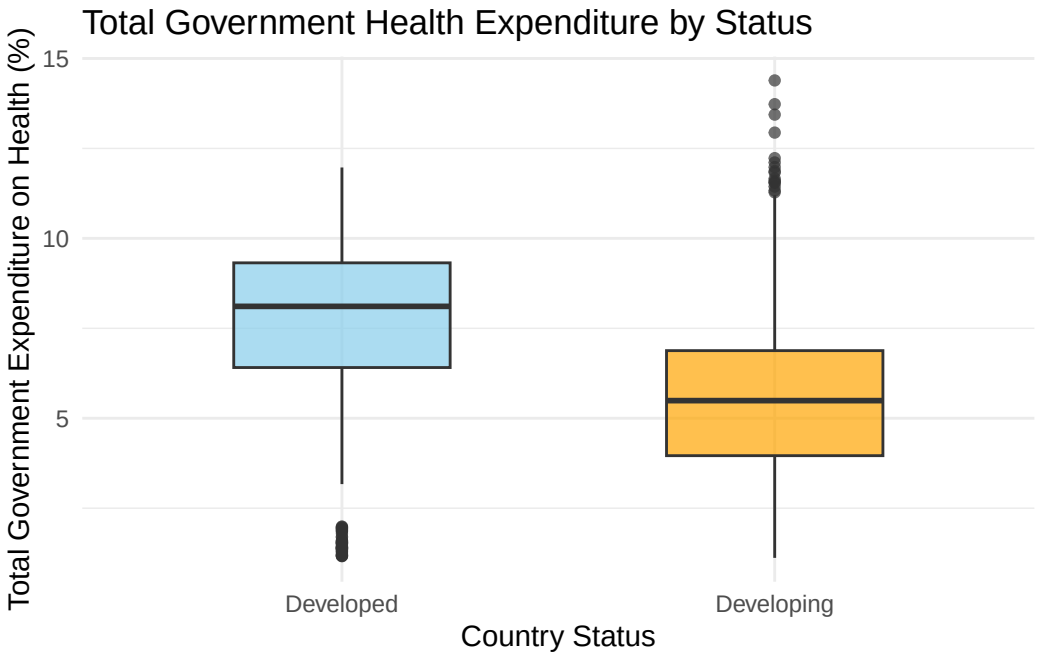

Figure 5: Boxplot for government expenditure on health as a percentage of total expenditure, showing relatively high percentages for developed countries compared to developing countries, highlighting that developed countries' governments typically allocate more resources to health, though there are exceptions.

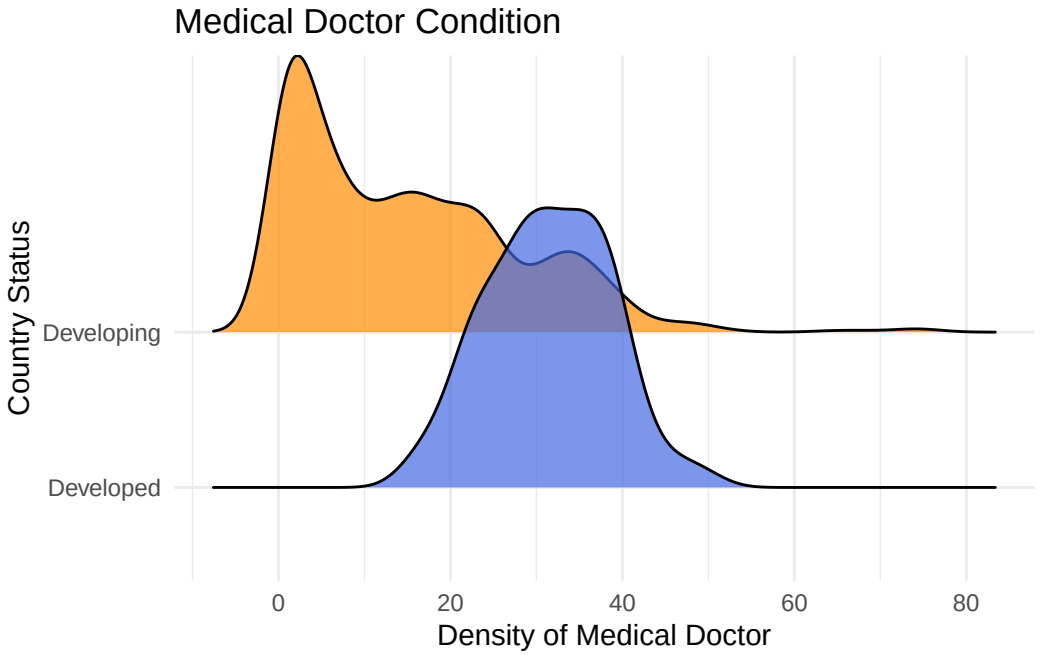

Figure 6: Density plot for doctors per 10,000 population by Status, showing that the density of medical doctors in developed regions is relatively higher compared to developing regions. In developing countries, the density is concentrated at lower levels, suggesting that the healthcare system is generally stronger and more accessible in developed countries.

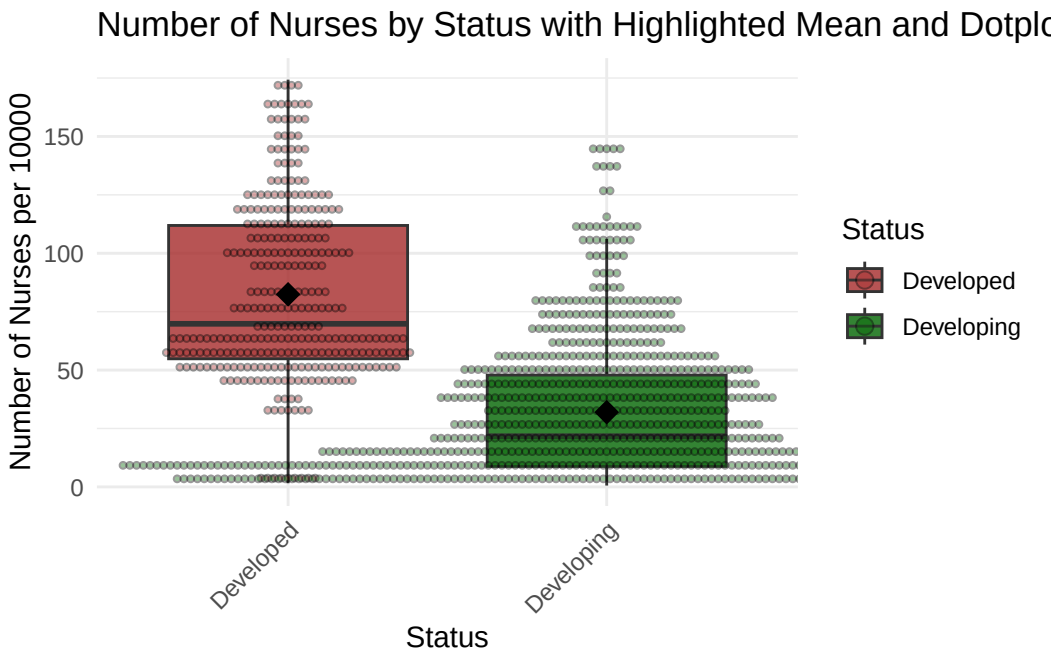

Figure 7: Number of Nurses by Status with Highlighted Mean and Dotplot. The mean for developed countries is around 80, while the mean for developing countries is around 35, with many data points showing low nurse density in developing countries. Both status groups have medians lower than their means, suggesting a right-skewed distribution with a concentration of countries with fewer nurses.

3 Model

The goal of this analysis is to construct a linear model to predict life expectancy (`LifeExpectancy`), which represents the average number of years an individual is expected to live. This model incorporates several predictor variables that capture economic and social factors influencing life expectancy.

3.1 Model set-up

The models assume a linear relationship between the predictors and life expectancy, where each predictor contributes independently or interactively to the variation in the outcome variable. Define y_i as the Life Expectancy for country i. The factors x_1, x_2,...represent socioeconomic predictors such as Diphtheria, Schooling, TotalExpenditure, pct_doctor, pct_nursing, and interaction term.

The linear regression model for developed and developing regions are specified as follows:

$$\begin{aligned}\text{LifeExpectancy}_{\text{developed}} = \beta_0 &+ \beta_1 \text{Diphtheria Immunization} + \beta_2 \text{Schooling} + \\ &\beta_3 \text{Health Expenditure} + \beta_4 \text{Number of Doctors per 10000} + \\ &\beta_5 \text{Number of Nurses per 10000} + \\ &\beta_6 \text{Health Expenditure} \times \text{Number of Doctors per 10000} \times \text{Number of Nurses per 1000}\end{aligned}$$

$$\begin{aligned}\text{LifeExpectancy}_{\text{developing}} = \beta_0 &+ \beta_1 \text{Diphtheria Immunization} + \beta_2 \text{Schooling} + \\ &\beta_3 \text{Health Expenditure} + \beta_4 \text{Number of Doctors per 10000} + \\ &\beta_5 \text{Number of Nurses per 10000} + \\ &\beta_6 \text{Health Expenditure} \times \text{Number of Doctors per 10000} \times \text{Number of Nurses per 100}\end{aligned}$$

Where:

- `LifeExpectancy`: Represents the average number of years a person is expected to live in a specific country, based on current mortality trends.
- `Diphtheria Immunization`: The percentage of children who received the diphtheria-tetanus-pertussis (DTP3) vaccine, serving as a proxy for the quality and reach of immunization programs.
- `Schooling`: Average years of schooling for individuals aged 25 and above, reflecting the level of education in the population.

- `Health Expenditure`: The percentage of government expenditure allocated to healthcare, indicating the priority given to public health.
- `Number of Doctors per 10,000`: The availability of medical doctors per 10,000 people, capturing the accessibility of professional healthcare services.
- `Number of Nurses per 10,000`: The availability of nursing staff per 10,000 people, further representing healthcare capacity.
- Interaction Term: The combined effect of health expenditure, doctors, and nurses on life expectancy, considering how these variables may amplify or mitigate each other's impact.
- epsilon is the error term, assumed to be normally distributed with a mean of 0 and constant variance $\epsilon_i \sim N(0, \sigma^2)$

3.2 Model Justification

The models were built using linear regression (LM), which is appropriate for predicting life expectancy as a continuous variable. This method captures the relationship between predictors and the response variable with interpretable coefficients. For the Developed model, predictors such as Diphtheria Immunization, Schooling, Total Expenditure, and the number of Doctors per 10,000 were selected as they are widely recognized to influence life expectancy in high-income countries. In the Developing model, we added Schooling, and both Doctors and Nurses per 10,000, reflecting healthcare access and lifestyle factors relevant to developing nations. Linear regression was chosen because it is effective for normally distributed continuous outcomes, which aligns with our dataset.

3.3 Model Weaknesses and Limitations

Linear regression assumes that the residuals errors of the model are normally distributed. If this assumption is violated, it could lead to biased estimates and incorrect conclusions. This issue is especially important in cross-country data, where life expectancy may be influenced by diverse factors that do not follow a normal distribution.

Predictors such as the number of doctors per 10,000 and the number of nurses per 10,000 might be correlated with each other. High correlation between predictors can cause multicollinearity, making it difficult to assess the individual contribution of each predictor.

3.4 Model Validation

The models for both Developed and Developing countries were implemented using R, with applying out-of-sample testing, the data was randomly split into training and test sets, with 80% used for training the model and 20% reserved for testing. This allows us to assess how well the model performs on data that was not used during training. The Root Mean Square Error (RMSE) was calculated for each model to evaluate the performance and generalization to unseen data. A lower RMSE indicates better predictive performance.

In Table 1, the Developed model is more efficient and better fitted primarily due to its higher R^2 (64.5% vs 51.5%), indicating that it explains a greater proportion of variance in life expectancy. Additionally, its RMSE value is lower (3.78 vs 5.06), showing more precise predictions. The Developing model has a larger sample size and variability, which might have diluted the effect of individual predictors. The better fit of the Developed model is partly due to the smaller, more homogeneous sample, allowing it to provide more reliable estimates with fewer outliers.

3.4.1 Model for Developed Countries

For the Developed model, the predictors considered included Diphtheria Immunization, Schooling, Total Expenditure, and Number of Doctors per 10000. The coefficients and standard errors for these predictors were as follows:

Diphtheria Immunization: Coefficient = 0.022, Standard Error = 0.025

Schooling: Coefficient = 0.801, Standard Error = 0.151

Health Expenditure: Coefficient = 0.056, Standard Error = 0.084

Number of Medical Doctors per 10000: Coefficient = 0.018, Standard Error = 0.034

These coefficients indicate that `Schooling` has the highest contribution to explaining life expectancy, followed by `Diphtheria Immunization`. The relatively small contribution of `#Doctors per 10000` suggests that doctor practices may have a less direct impact on life expectancy in developed countries, but this could be nuanced by other health factors not captured in the model.

3.4.2 Model for Developing Countries

For the Developing model, the same predictors were considered, but their relationships with life expectancy differed. The coefficients and standard errors for each predictor were as follows:

Diphtheria Immunization: Coefficient = 0.036, Standard Error = 0.010

Schooling: Coefficient = 1.727, Standard Error = 0.113

Number of Medical Doctors per 10000: Coefficient = 0.090, Standard Error = 0.021

Table 1: Summary of key model estimates for Developed and Developing counties, including coefficients for predictors like Diphtheria, Schooling, and TotalExpenditure, pct_doctor, pct_nursing, with standard errors for each estimate. Model performance statistics, such as sample size, R^2, and adjusted R^2, are also displayed.

	Model for Developed	Model for Developing
Diphtheria Immunization	0.023	0.031
	(0.024)	(0.009)
Schooling	0.645	1.703
	(0.161)	(0.106)
Health Expenditure	−1.740	−0.364
	(1.025)	(0.162)
Number of Medical Doctors per 10000	−0.446	0.441
	(0.232)	(0.084)
Number of Nurses per 10000	−0.120	−0.009
	(0.106)	(0.033)
Num.Obs.	305	788
R2	0.191	0.635
R2 Adj.	0.166	0.630
AIC	1695.5	4774.5
BIC	1736.4	4825.8
Log.Lik.	−836.754	−2376.237
F	7.739	
RMSE	3.76	4.94

Number of Nurses per 10000: Coefficient = -0.003, Standard Error = 0.003

In the Developing model, `Schooling` again has the most significant impact, and `#Doctors per 10000` is also a substantial predictor. This highlights the importance of healthcare workforce availability in improving life expectancy in developing countries, unlike the developed context where Schooling was more prominent. The negative relationship of `Total Expenditure` suggests that in developing countries, spending in certain healthcare sectors could be misallocated or inefficient.

4 Results

Table 2: The prediction table indicates that the average life expectancy in developed countries is 78.53 years, while in developing countries, it is slightly lower at 71.68 years. This difference of approximately 5.1 years in life expectancy suggests a significant health and infrastructure disparity between developed and developing countries. These results align with broader global trends where developed countries typically benefit from better healthcare systems, higher standards of living, and advanced medical technology, contributing to longer life expectancies.

Average Life Expectancy in Developed Countries	Average Life Expectancy in Developing Countries
78.52	71.67

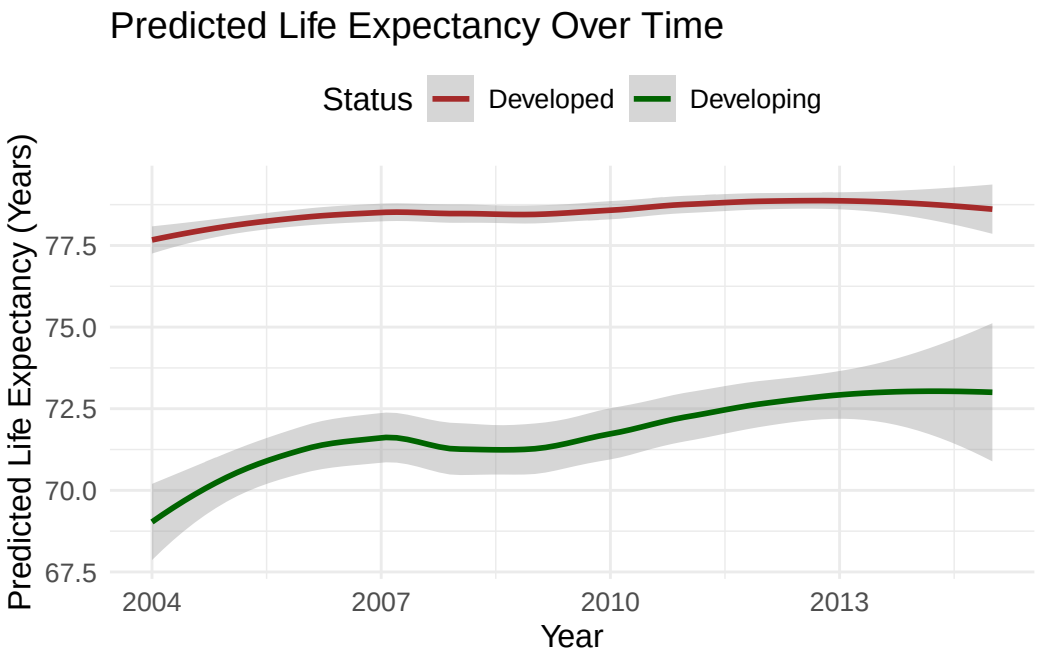

Figure 8: The prediction plots demonstrate that life expectancy in developed countries tends to remain higher with less variance, indicating stability in health outcomes due to better healthcare systems, more consistent access to medical services, and overall economic stability with less variance. In contrast, the life expectancy in developing countries exhibits a broader range of variation, with a more rapid growth trajectory. This suggests that while life expectancy is increasing in these countries, it is influenced by a variety of factors such as improvements in healthcare, economic development, and access to resources, all of which vary more significantly across different regions.

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.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

B.2 Diagnostics

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Garon, Julie R. et al. 2015. "The Challenge of Global Diphtheria Eradication." *Global Health* 29 (4).

Girum, Tadele et al. 2018. "Determinants of Life Expectancy in Low and Medium Human Development Index Countries." *Medical Studies* 34: 218–25. https://www.researchgate.net/publication/328140185_Determinants_of_life_expectancy_in_low_and_medium_human_development_index_countries.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Hanushek, Eric A., and Ludger Woessmann. 2016. "The Role of Education Quality for Economic Growth." *Economics of Education Review* 55: 13–27. https://hdl.handle.net/10986/7154.

He, L., and N. Li. 2020. "The Linkages Between Life Expectancy and Economic Growth: Some New Evidence." *Empirical Economics* 58: 2381–2402. https://doi.org/10.1007/s00181-018-1612-7.

Kumar, Rajarshi. 2017. "Life Expectancy (WHO)." *Kaggle*. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?resource=download.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

Utkarsh, XY. 2020. "World Health Statistics 2020 Complete." *Kaggle*. https://www.kaggle.com/datasets/utkarshxy/who-worldhealth-statistics-2020-complete/data.

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. 2019. "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://CRAN.R-project.org/package=dplyr.

Wilke, Claus O. 2024. *Ggridges: Ridgeline Plots in 'Ggplot2'*. https://wilkelab.org/ggridges/.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra/.