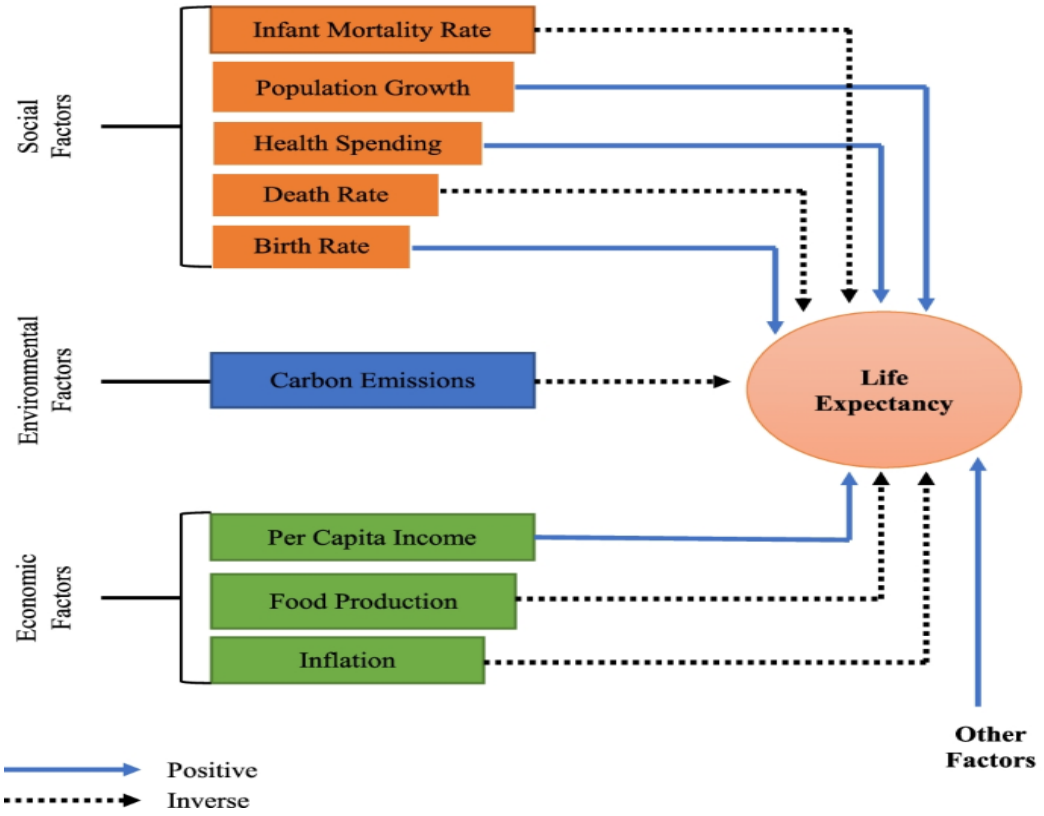
Understanding the Determinants of Life Expectancy



Conducted by:

Names	UCIDS
Adam Virani	30038543
Anthony Obi	30268043
Chris Atta Hawkson	30270419
Maria Aidoo	30234335
Ikeora Ekene	30260948

Date completed: December 1st , 2024

Table Of Content

1. Introduction

Life expectancy ranks among the most important health and quality of life indicators within the population. Understanding trends in life expectancy and those factors that determine it would be crucial to addressing health inequalities globally and improving the well-being of populations, particularly in those regions where disparities in life expectancy are greatest. The present study examines the relative contributions of health, economic, and social factors to life expectancy, identifies the best predictors, and provides actionable insights for policymakers and global health organizations.

We will apply data on health metrics-reducing infant and under-five mortality rates, adult mortality rate, rate of incidence of HIV, rates of vaccination, in addition to socioeconomic data on GDP per capita, overall schooling, and economic status-to a robust multivariate regression model. Using such variables, this study will describe the associations with life expectancy in detailed quantification.

1.1. Motivation

Life expectancy is a very important area of study since it reflects directly on the quality of life and general well-being of populations. Global health disparities persist, with communities in developing countries relatively recording lower life expectancies due to the shortage of healthcare resources and socio-economic problems. This project, therefore, intends to unravel some of the factors that could provide actionable insights into informing policy interventions and resource allocation by looking at underlying drivers of life expectancy. The findings of this study will be very useful to global health organizations and policymakers by providing evidence-based recommendations for addressing disparities and improving health outcomes. This project, in the end, aspires to contribute to the global effort of advancing life expectancy and fostering equity in health, especially in regions where it is most urgently needed.

1.1.1. Context and Applied Domains of the Project

Context:

The project has been contextualized in a setting where global health inequality problems continue to affect the lives of millions of people worldwide. Life expectancy has always been an important yardstick for the general health and quality of life that populations hence an area of vital analysis. The project has dwelled on establishing the socioeconomic

and healthcare-related factors in life expectancy, especially in cases where the gap was most prominent. Therefore, by grasping these divergent reasons, this project hopefully will provide practical insights useful for addressing systemic inequalities to better achieve improved health outcomes in developing economies.

Applied Domains:

- Public Health:

It informs the understanding of disparity in health and improvements in access to healthcare services or lack thereof in under-resourced environments.

- Policy Development:

Analysis will also inform policymakers of proper evidence-based interventions needed in the pursuit of increased life expectancy or reduction in health inequality, on one hand. This project contributes to various global initiatives on equity in health outcomes and SDGs, in particular those targeting health and well-being, by highlighting actionable areas of improvement.

- Socioeconomic Research:

The project demonstrates the linkage between socio-economic factors and health outcomes as a way to expand the research field into the social determinants of health.

1.1.2. Problems to Address

- Identifying Key Predictors in Life Expectancy
- Exploring life expectancy in various Regions
- Figuring out the impact of Socioeconomic and Health Factors on life

1.1.3. Challenges

The problem of analyzing and predicting life expectancy is challenging for several reasons.

- Complex Interactions Between Predictors

Life expectancy results from the interaction of health, economic, and social causes in a nonlinear and complex interaction. Interaction effects between predictors, such as interplay between Schooling and GDP per capita, which could make it difficult to explain the contribution of one variable.

Example: While better health is generally associated with higher GDP per capita, it may be moderated by cultural practice or health care inequality in specific cases.

- Multicollinearity Among Variables:

There is a high degree of inter-relations among many of the predictors, which makes it difficult to isolate the individual effect of each variable.

For example, schooling and GDP per capita often go hand in hand, so it becomes tough to say that improved education or higher income raises life expectancy.

- Violations of Statistical Assumptions:

The assumptions of multiple linear regression, such as linearity, normality of residuals, and homoscedasticity, might get violated with natural data.

Example: Very often, the relationship between life expectancy and GDP per capita is nonlinear, which exhibits diminishing returns at higher levels of income.

- Heteroskedasticity:

Error variance may not be constant across the range of predictors.

For instance, the slope of life expectancy as a function of economic status can strongly vary across different income groups and thus may lead to the violation of regression assumptions.

1.2. Objectives

Develop a Robust Multivariate Regression Model

- Identify and pre-process the relevant predictors to build the regression model.
- Test and validate the model with appropriate metrics in order to ensure that the model is robust and reliable.

Interpret the Model Coefficients

- Look at the coefficients' magnitude and direction of every predictor and determine the level of each predictor's contribution to life expectancy.
- Test the coefficients with regard to their statistical significance and practical importance to yield relevant insights.

Evidence-Based Recommendations

- The findings from this study will be used in developing actionable recommendations for policymakers and identifying from the analysis those predictors that will have the greatest impact.
- Prioritize health, education, and economic development interventions based on the results of the model.

These objectives are specific, actionable, and directly linked to the overall goals of your project.

1.2.1. Overview

The main intention of the project is to analyze health, economic, and social factors related to life expectancy in order to reach a robust multivariate regression model, which will serve to predict life expectancy based on those factors. The interpretations of the findings from the model shall form actionable insights with evidence-based recommendations for the decision-makers and global health organizations for the purpose of tackling health disparities and improvement in health outcomes in resource-constrained low-life expectancy regions.

1.2.2. Research Questions

1. What health metrics - infant and under-five mortality, adult mortality, HIV incidence, and vaccination rates - have the largest impact on life expectancy?
2. How do socioeconomic factors like GDP per capita, schooling, and economic status contribute to the prediction of life expectancy?
3. Which socio-economic factors are the most influential predictors of life expectancy, and how can these be used to guide policy interventions?

2. Methodology

2.1 Data

The dataset used in this project is titled **"Life Expectancy (WHO)"**, sourced from the World Health Organization (WHO) and made available on Kaggle. It is published under the Creative Commons Public Domain Dedication (CC0 1.0 Universal) license, which permits unrestricted use, including copying, modifying, and distributing the data, even for commercial purposes, without requiring permission. The dataset was curated by a third-party contributor and is not proprietary to our group, making it publicly accessible for research purposes.

Overview of the Dataset

- **Source:** World Health Organization (via Kaggle)
- **Time Period:** 2005–2015
- **Size:** 2,864 rows × 21 columns
- **Observations:** Each row represents a single country's data for a specific year.

Variables
The dataset contains the following variables, which are categorized and described as follows:

1. **Country**: Name of the country (qualitative, 179 unique values).
2. **Region**: The region to which the country belongs (qualitative, 9 regions).
3. **Year**: The year of observation (2005–2015, quantitative).

4. **Infant Deaths**: Number of infant deaths per 1,000 population (quantitative).
5. **Under-Five Deaths**: Number of deaths of children under five per 1,000 population (quantitative).
6. **Adult Mortality**: Deaths per 1,000 adults aged 15–60 (quantitative).
7. **Immunization Coverage (%):**
    a. Hepatitis B among 1-year-olds
    b. Measles (MCV1) among 1-year-olds
    c. Polio (Pol3) among 1-year-olds
    d. Diphtheria, tetanus, and pertussis (DTP3) among 1-year-olds
8. **HIV Incidence**: Cases per 1,000 population aged 15–49 (quantitative).
9. **GDP Per Capita**: Current USD (quantitative).
10. **Total Population**: Population size in millions (quantitative).
11. **Mean Years of Schooling**: Average years of formal education for individuals aged 25+ (quantitative).
12. **Economic Status:**

- Developed country (binary: 1 = yes, 0 = no).
- Developing country (binary: 1 = yes, 0 = no).

13. **Alcohol Consumption**: Liters of pure alcohol consumed per capita among individuals aged 15+ (quantitative).
14. **BMI**: Average Body Mass Index of the population (quantitative).
15. **Thinness Prevalence (%):**

- Adolescents aged 10–19 years
- Children aged 5–9 years

16. **Life Expectancy**: Average life expectancy of both genders (quantitative, continuous).

The **response variable** for this analysis is **Life Expectancy**, a continuous, quantitative variable. All other variables are considered predictors. Given their potential relevance to life expectancy, all predictors will initially be included in the model, except for "Country," "Region," and "Year," which are excluded to avoid overly complex models. Residuals for these excluded variables will still be checked to ensure assumptions are met.

2.2 Approach

We employ **multiple linear regression modelling** to investigate relationships between predictors and the response variable, life expectancy. This approach is suitable because it enables analysis of the linear effects of multiple independent variables on a continuous dependent variable.

Key steps include:

1. **Baseline Model**: Construct an initial model with all predictors.
2. **Stepwise Model Selection**: Employ forward and backward stepwise approaches to identify the best additive model.
3. **Interaction and Polynomial Terms**: Explore interaction effects and non-linear relationships using higher-order terms.

4. **Assumption Checks**: Evaluate assumptions such as linearity, normality, homoscedasticity, and independence of residuals.

Statistical Criteria:

- **Alpha Level**: 0.05 for hypothesis testing.
- Model performance will be assessed using metrics like adjusted R-squared, and residual standard error.

2.3 Workflow

1. **Clean, Import, and Inspect Data:**
   Prepare the data for analysis by ensuring its quality and integrity. In this stage a general data inspection was carried out to understand the variables and its corresponding data types
2. **Initial Multilinear Regression Model:**
   Establish a baseline regression model using all relevant predictors. Fit a multiple linear regression model to understand initial relationships between predictors and the target variable. Assess performance metrics like R-squared, adjusted R-squared, and residual standard error.
3. **Multi-collinearity Assumption Check:**
   Detect and address multicollinearity among predictors by calculating the Variance Inflation Factor (VIF) for each predictor. Remove highly correlated predictors to avoid instability in coefficient estimates.
4. **Best Additive Model:**
   Focus on additive effects of predictors by keeping predictors that contribute significantly to explaining variance in the target variable. Achieved by using both forward and backward stepwise approaches. The variables that are consistently selected as the best predictors are kept in the final additive model.

   Note: p_enter=0.05, p_remove=0.1
5. **Best Model with Interactions:**
   Explore interaction effects between predictors. Add interaction terms to the model to capture synergistic relationships.
6. **Best Model with High-Order Terms:**
   Account for non-linear relationships between predictors and the target variable. Add polynomial or other high-order terms to the model based on observed nonlinear curves from pairwise plot.
7. **Assumption Checks:**

Ensure the model meets the key assumptions of linear regression.
Check for normality, homoscedasticity, independence of residuals, and linearity.
Address any violations through model rebuilds or response variable transformation.

8. **Result Interpretation:**
   Derive actionable insights from the final model. Examine coefficient estimates to understand the effect of predictors. Communicate results in a way that aligns with the study's objectives.

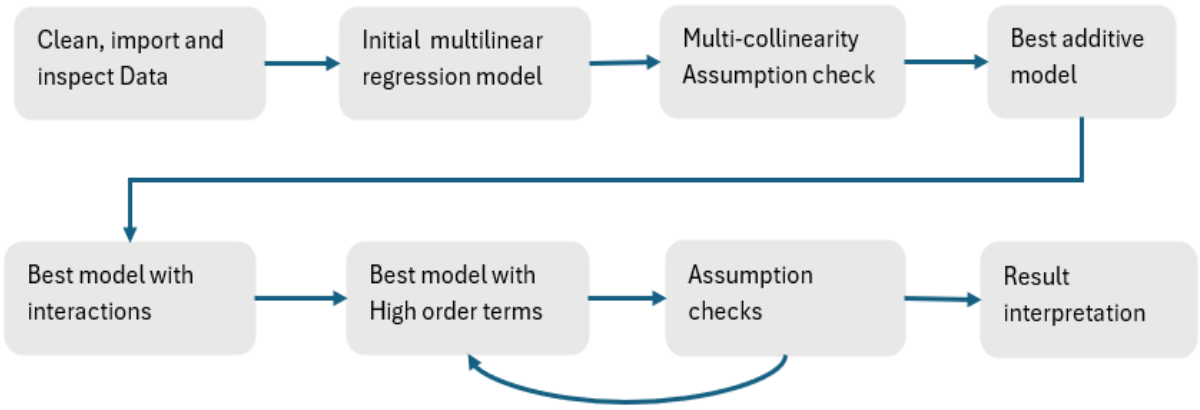


Figure 1. Workflow diagram showing project approach and steps

Challenges
Key challenges include handling polynomial terms and addressing violated assumptions. If the data or corrections fail to meet assumptions, we may need to conclude that multiple linear regression is not appropriate for this dataset, acknowledging the limitations within this project's scope.

2.4 Contributions

The project team divided responsibilities equitably based on individual strengths:

- **Ikeora Ekene**: Project management, timeline oversight, and facilitation of team collaboration.
- **Adam Virani**: Development and refinement of regression models.
- **Chris Atta Hawkson**: Data analysis and identification of trends and patterns.
- **Anthony Obi**: Validation of regression models and enhancement of data-driven decision-making.
- **Maria Aidoo**: Exploratory data analysis and preparation for statistical modeling.

This equitable distribution ensured all members contributed effectively while leveraging their expertise.

3. Main results of the analysis

3 .1 Initial Model

- **Model Description**: A full linear regression model with all predictors from the dataset.
- **Outcomes**:
  - 0.9791 high adjusted r squared, RSE of 1.361
  - Several insignificant variables such as Measles, Polio, Diphtheria, Population, Thinness 5-10

3.2 Multicollinearity Check

- **Method**:
  - VIF calculation for all predictors.
  - Variables with VIF > 5 were considered for exclusion or transformation.
- **Findings**:
  - High multicollinearity among Infant deaths and under five deaths and polio and diphtheria (see table 1 below)
  - Solution: Removed infant deaths, polio and diphtheria were already not significant anyways.

Table 1. Multicollinearity test results for initial model

Variables	VIF	Detection
Infant_deaths	44.7615	1
Under_five_deaths	45.2903	1
Adult_mortality	7.856	0
Alcohol_consumption	2.4089	0

Hepatitis_B	2.5885	0
Measles	1.5878	0
BMI	2.7274	0
Polio	12.0104	1
Diphtheria	13.001	1
Incidents_HIV	2.8988	0
GDP_per_capita	2.3232	0
Population_mln	1.1519	0
Thinness_ten_nineteen_years	8.9502	0
Thinness_five_nine_years	8.9586	0
Schooling	4.3943	0
Economy_status_Developed	2.9678	0

3.3 Best Additive Model

- **Description**: A refined linear regression model with only significant non multicollinear predictors.
- Used forward and backwards stepwise regression model to identified 9 main predictors as shown in table 2
- **The regression line is given by**

$$\begin{aligned}\text{Life Expectancy} = {} & 83.94 - 0.04865 Adult_{m}ortality - 0.08277 Under_{f}ive_{d}eaths \\ & + 0.7465 Economy_{s}tatus_{D}eveloped + 0.00002808 GDP_{p}er_{c}apita \\ & + 0.07456 Alcohol_{c}onsumption + 0.09724 Schooling - 0.1465 BMI \\ & + 0.1076 Incidents_{H}IV - 0.0379 Thinness_{t}en_{n}ineteen_{y}ears\end{aligned}$$

Figure 2. Model equation for best additive model

- The model Adjusted R Square is 0.9786 and RSE is **1.375**

Table 2. Results with coefficients and p values for best additive model

Predictors	Coefficient	P- Value
(Intercept)	83.94	< 2e-16
Adult_mortality	-0.04865	< 2e-16
Under_five_deaths	-0.08277	< 2e-16
Economy_status_Developed	0.7465	5.97e-12
GDP_per_capita	0.00002808	< 2e-16
Alcohol_consumption	0.07456	6.40e-14

Schooling	0.09724	7.43e-09
BMI	-0.1465	6.92e-15
Incidents_HIV	0.1076	2.90e-09
Thinness_ten_nineteen_years	-0.0379	2.17e-06

3.4 Interaction Model

- **Approach**:
  - Constructed the interaction model using the 9 main predictors in section 3.3
- **Findings**:
  - 25 significant interactions
  - Adjusted r squared of 0.9868 and RSE of 1.082

3.5 Higher Order Terms

- **Approach**:
  - gg pairs used to find potential nonlinear looking relations (see figure 3)

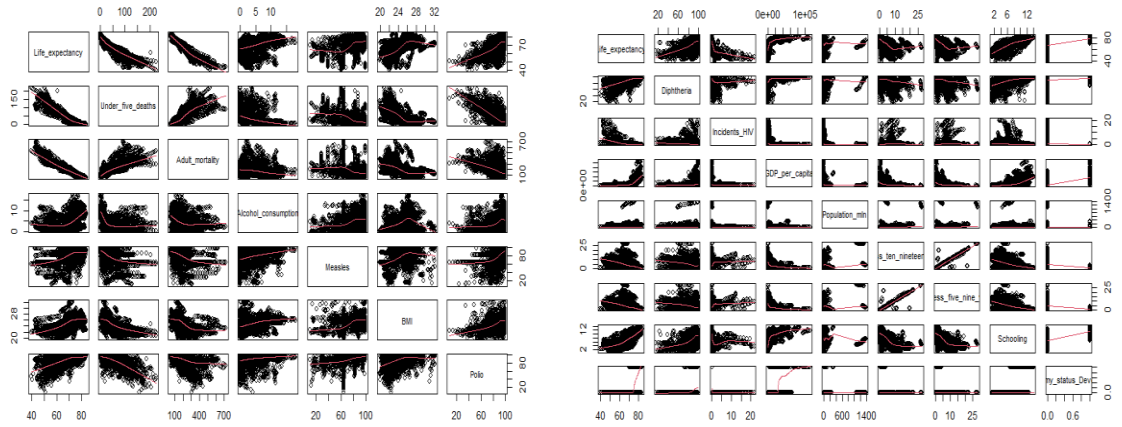


Figure 3. GGpair plots for initial predictors to assess potential higher order transformations

- **Findings**:
  - Table 3 below shows the significant higher order predictors

Table 3. Results of higher order transformations with the highest power the variable was taken to that was significant and that powers p value

Variables	Max Higher Power	P- Value
Alcohol consumption	5	3.31E-08
thinness for 10-19 years of age	5	3.98E-12
BMI	4	1.72e-12
Incidence of HIV	6	0.000137
GDP_per_capita	6	0.014101
Schooling	2	4.74e-06

- Adjusted r squared of 0.9891 and RSE of 0.9811

3.6 Assumption Check

3.6.1 Tests Performed

- Linearity: Scatterplots of residuals.
  The residual plot (Figure 4) indicates no pattern in the data. Therefore, the linearity assumption of our model is satisfied.

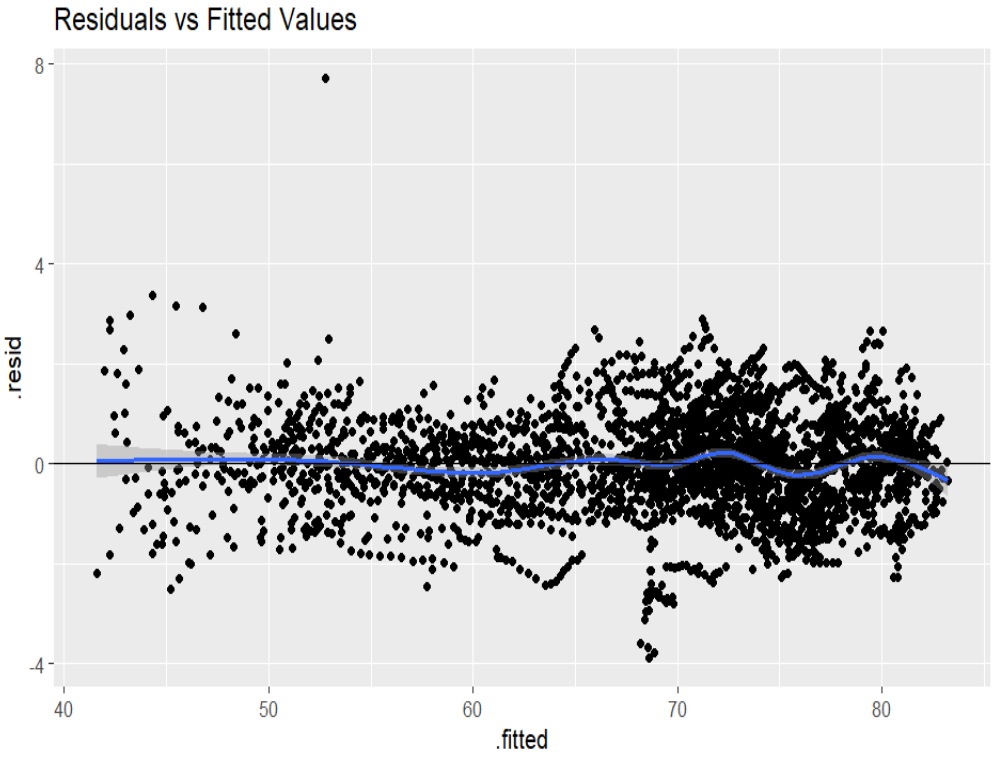


Figure 4. Residual Fitted Values plots to assess linearity assumptions

- Independence, graph of residuals for country, year and region

From the figure 5, 6 and 7 below, the residual plot does not show any obvious pattern. Since the residual points are randomly scatter around the 0 value for the Country, Region and Year groupings, this suggests that the model variables are independent

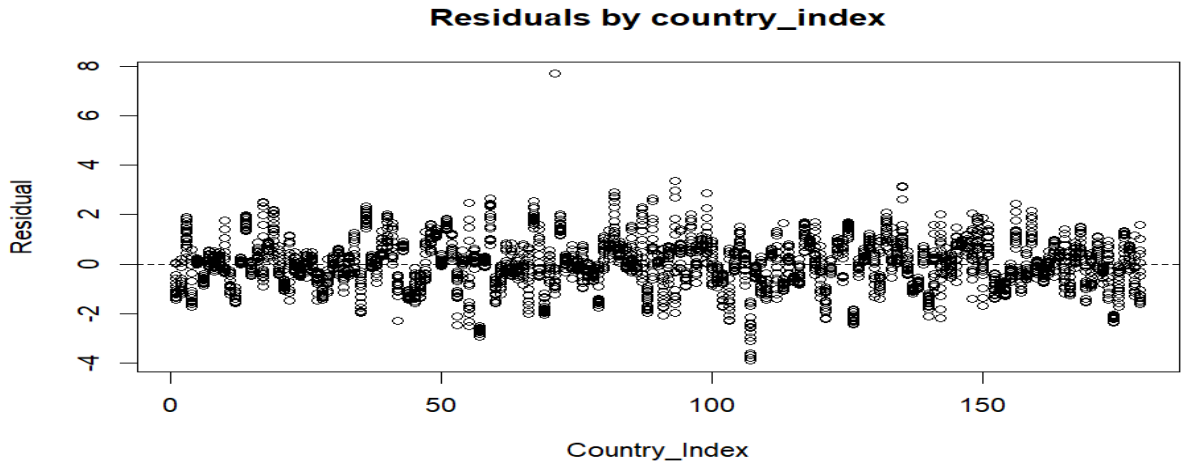


Figure 5. Residual by Country to assess Independence assumption

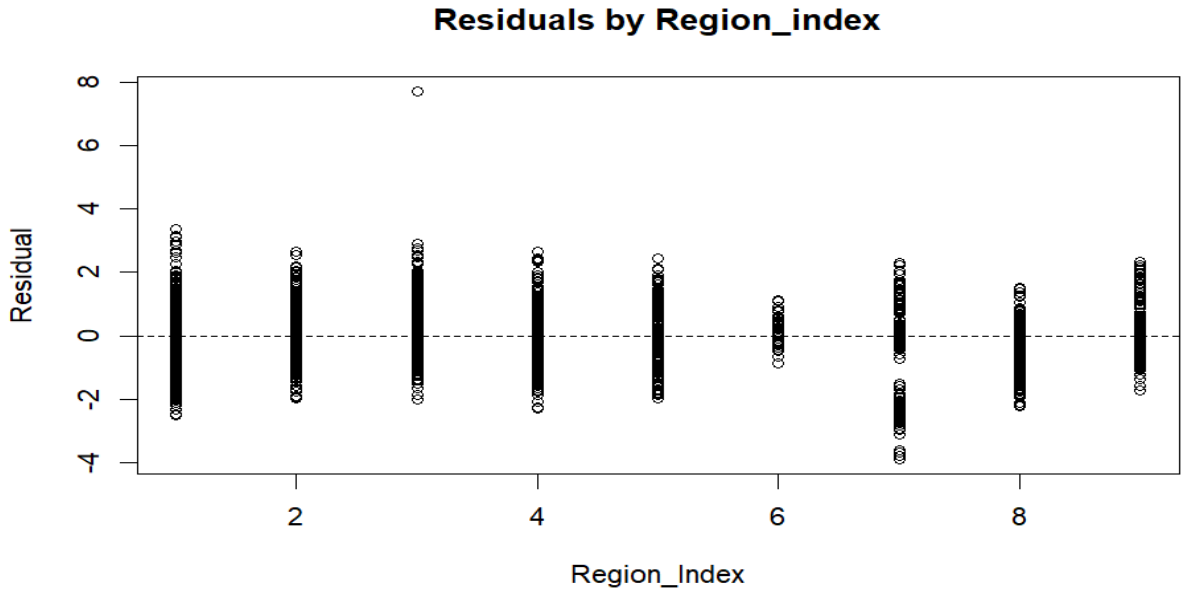


Figure 6. Residual by region plots to assess independence assumptions

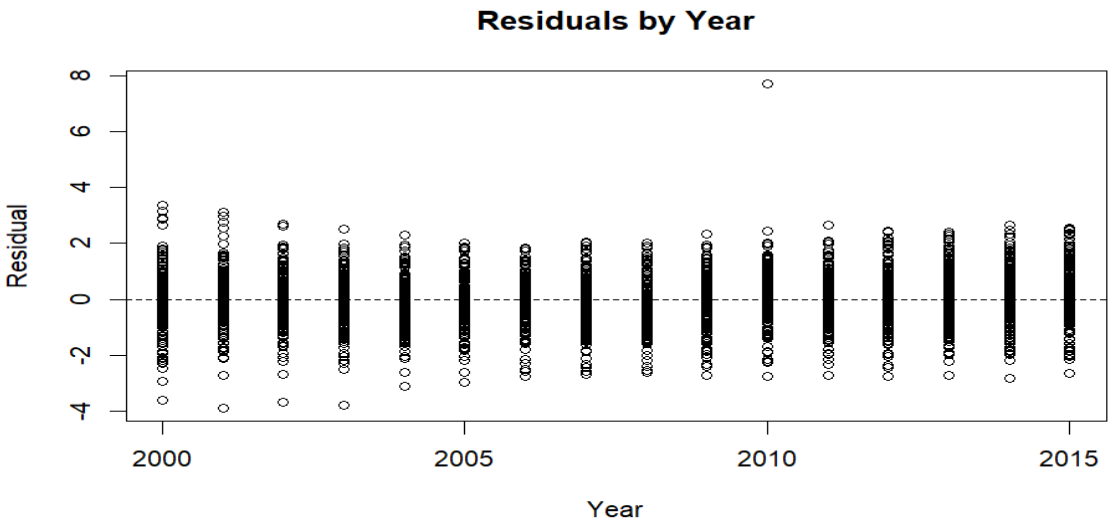


Figure 7. Residual by Year plots to assess independence assumptions

- Homoscedasticity:
  - Performed Scale Plot and Breusch Pagan Test ((p<0.05).
  - The Scale Location plot (figure 8) shows no significant signs of funneling. We performed the Breusch Pagan to further validate the findings

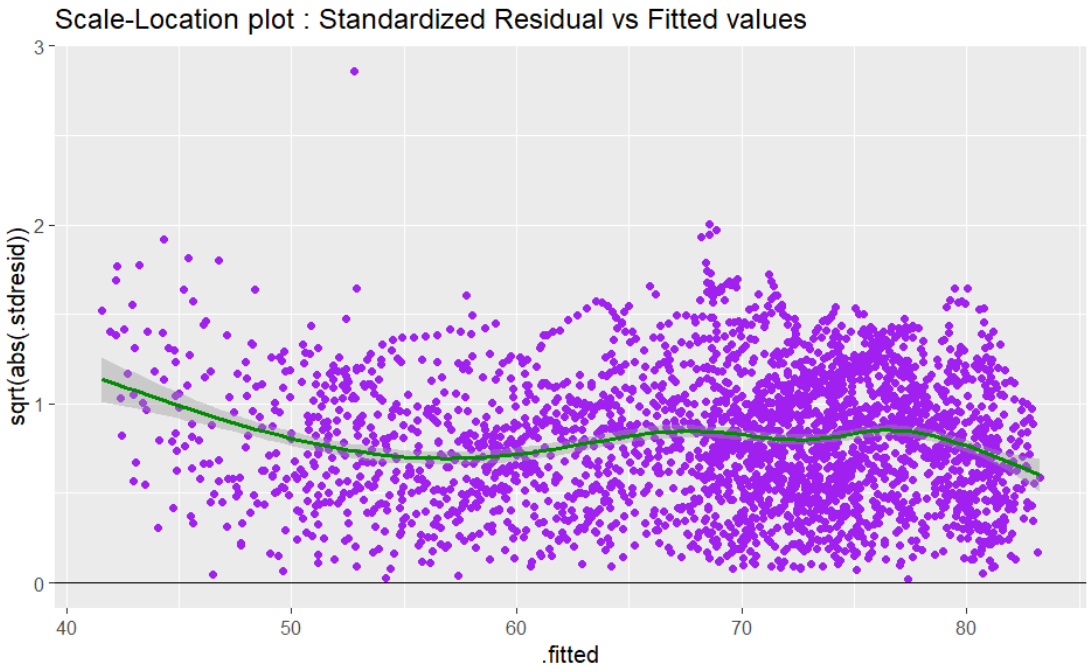


Figure 8. Scale Location plot to assess Heteroscedacity assumptions

- Breusch Pagan Test
  - Hypothesis:
    - Ho: heteroscedasticity is not present
    - Ha: heteroscedasticity is present
  - Significance level is at 0.05

  - Since the P-value is < alpha, we reject the Null and conclude that heteroscedasticity is present

- Normality:
  - Performed Q-Q plot and Shapiro-Wilk test ($p<0.05$)
  - The QQ plot (figure 9) indicates that the data points are close to the normality line except for few points that are further away
  - Further test was done using Shapiro Wilk to validate normality
    - Null hypothesis Ho: There is normality in the residual data
    - Alternative hypothesis Ha: There is no normality in the residual data
    - Significance level is set at 0.05
    - Since the p- value is less than 0.05, we reject the null hypothesis and conclude that there is no normality.

- Therefore, the normality assumption is not met

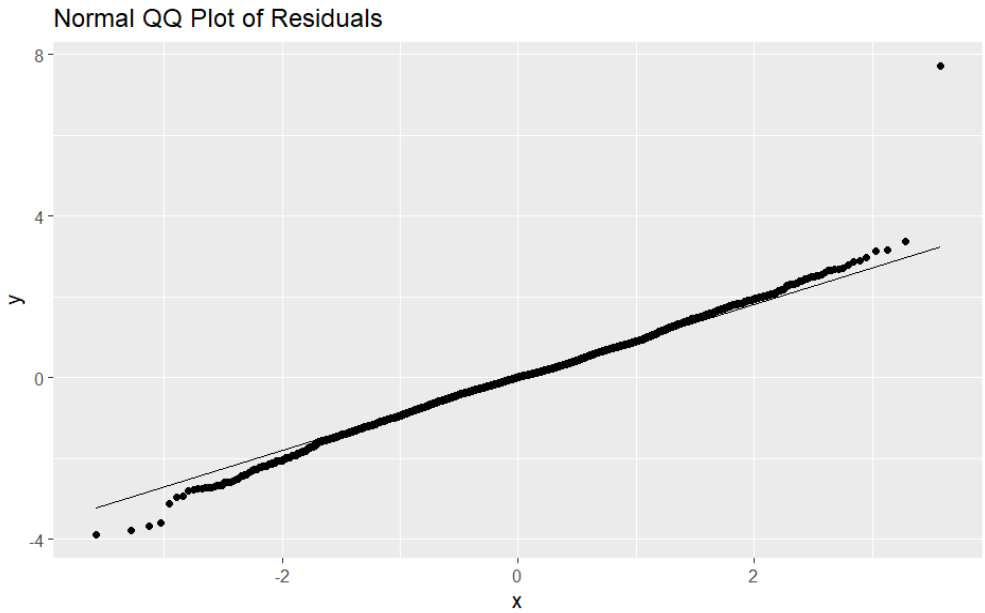


Figure 9. Q-Q Plot to assess normality assumptions

- Outliers and Leverage points
  - Residual Leverage plot to detect outliers
  - From the Residual Leverage plot in figure (10), there are no outliers because all the data points lie within the cook distance.
  - The cook distance plot (figure 11) also further underscores the absence of outliers as all the distances are less than 0.5.
  - The Leverage plot (figure 12) at 3p/n indicates there are 135 influential points in the data.

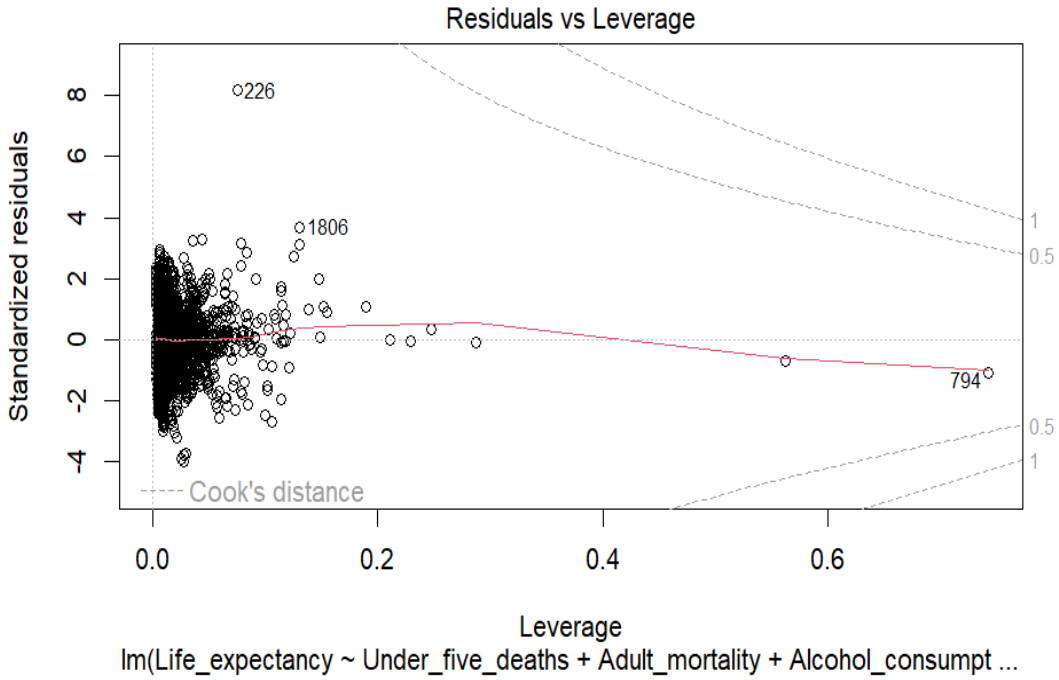


Figure 10. Residual Leverage plot to detect Outliers

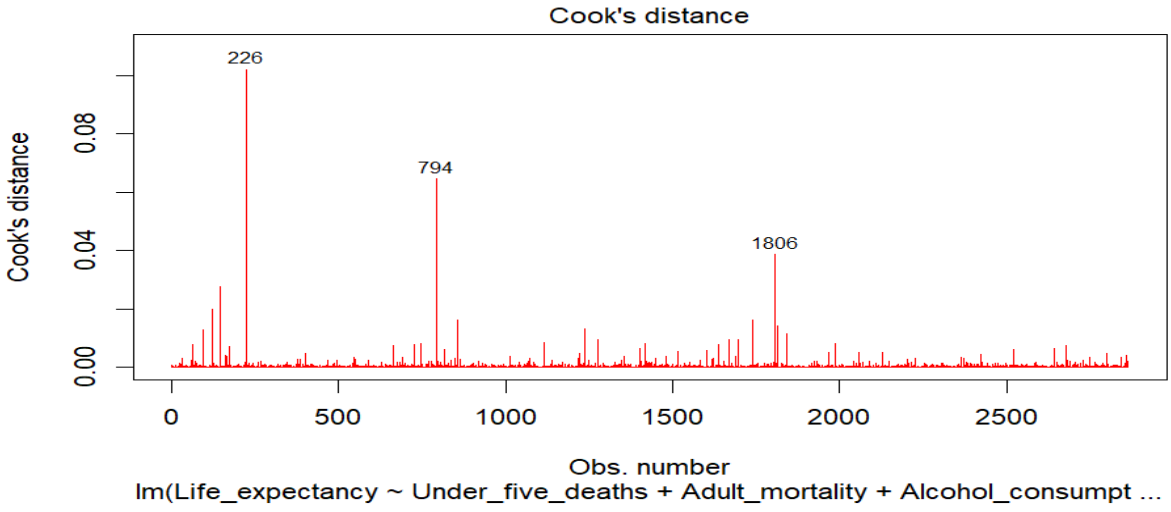


Figure 11. Cook distance plot to detect Outliers

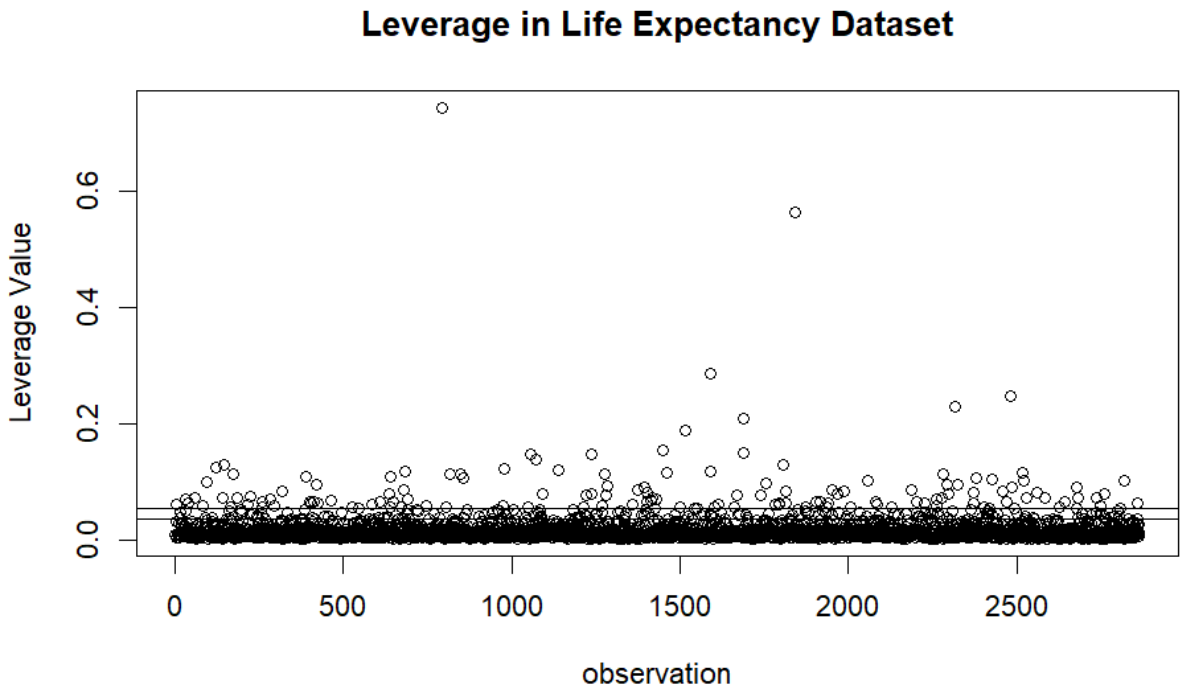


Figure 12. Leverage plot to detect influential leverage points

3.6.2 Findings from Assumption Check:

- Normality and heteroscedasticity assumptions were violated
- Attempt to fix through box cox transformation with log and best lamda, but violations were not fixed. For future analysis, we suggest using the Weight Least Squares Regression to address the normality and heteroscedasticity issues.
- Attempt to remove the influential points significantly reduced model predictive power. We therefore retained the influential leverage points.

3.7 Final Model

- The final linear model consists of 52 variables (main predictors, Interaction and higher order variables.

- **Adjusted R square is** 0.9891. This shows that 98.91% of the variation in Life expectancy can be explained by the model.
- Residual Square error (RSE) is 0.9891. This shows that on the average, the predicted Life expectancy value will deviate by 0.9891.

Key Findings:

- The effect developed countries have on the life expectancy is influenced by under five deaths, Alcohol consumption, HIV cases, and GDP per capita.
- Based on the final multilinear regression model, the Life expectancy difference between developed and developing countries is difficult to quantify due to the complex interactions developed countries have with under five deaths, Alcohol consumption, HIV cases, and GDP per Capita. The coefficient is given by: **(3.478 -0.09539 Under_five_deaths – 0.2825 Alcohol_consumption + 10.83 Incidents_HIV – 0.00008785 GDP_per_capita)**
- The effect mortality rate has on life expectancy is influenced by under five deaths, BMI, HIV cases, and GDP per capita. Due to the complexity of its coefficient in the model, it is hard to precisely state how mortality rate affects life expectancy. The coefficient is given by: **(0.005279 +0.00006012 Under_five_deaths – 0.002131 BMI+ 0.0007749 Incidents_HIV – 0.0000006439 GDP_per_capita)**

Sub Model

The final model equation has been divided into two sub-models (developed countries and developing countries sub-models as shown in figures 13 and 14.

$$
\begin{aligned}
\hat{LifeExpectancy} = {} & 1199 - 0.2801_{\text{Under\_five\_deaths}} + 0.005279_{\text{Adult\_mortality}} - 0.8051_{\text{Alcohol\_consumption}} \\
& + 0.2535_{\text{Alcohol\_consumption}^2} - 0.04733_{\text{Alcohol\_consumption}^3} + 0.003228_{\text{Alcohol\_consumption}^4} \\
& - 0.00007369_{\text{Alcohol\_consumption}^5} - 171.8_{\text{BMI}} + 9.787_{\text{BMI}^2} \\
& - 0.246_{\text{BMI}^3} + 0.002303_{\text{BMI}^4} - 2.656_{\text{Incidents\_HIV}} + 0.6867_{\text{Incidents\_HIV}^2} \\
& - 0.1296_{\text{Incidents\_HIV}^3} + 0.01073_{\text{Incidents\_HIV}^4} - 0.0004181_{\text{Incidents\_HIV}^5} \\
& + 0.00000632_{\text{Incidents\_HIV}^6} + 0.0002914_{\text{GDP\_per\_capita}} - 0.00000002194_{\text{GDP\_per\_capita}^2} \\
& + 0.0000000006612_{\text{GDP\_per\_capita}^3} - 0.00000000000009642_{\text{GDP\_per\_capita}^4} + 0.0000000000000000006603_{\text{GDP\_per\_capita}^5} \\
& - 0.000000000000000000001716_{\text{GDP\_per\_capita}^6} + 0.1788_{\text{Thinness\_ten\_nineteen\_years}} \\
& - 0.2080_{\text{Thinness\_ten\_nineteen\_years}^2} + 0.02363_{\text{Thinness\_ten\_nineteen\_years}^3} - 0.001065_{\text{Thinness\_ten\_nineteen\_years}^4} \\
& + 0.00001647_{\text{Thinness\_ten\_nineteen\_years}^5} + 1.305_{\text{Schooling}} - 0.02936_{\text{Schooling}^2} \\
& + (3.478 - 0.09539_{\text{Under\_five\_deaths}} - 0.2825_{\text{Alcohol\_consumption}} + 10.83_{\text{Incidents\_HIV}} \\
& - 0.00008785_{\text{GDP\_per\_capita}})_{\text{Economy\_status\_Developed}} \\
& + 0.00006012_{\text{Under\_five\_deaths} \times \text{Adult\_mortality}} - 0.001076_{\text{Under\_five\_deaths} \times \text{Alcohol\_consumption}} \\
& + 0.008992_{\text{Under\_five\_deaths} \times \text{BMI}} + 0.003744_{\text{Under\_five\_deaths} \times \text{Incidents\_HIV}} \\
& - 0.005619_{\text{Under\_five\_deaths} \times \text{Schooling}} - 0.002131_{\text{Adult\_mortality} \times \text{BMI}} \\
& + 0.0007749_{\text{Adult\_mortality} \times \text{Incidents\_HIV}} - 0.00000006439_{\text{Adult\_mortality} \times \text{GDP\_per\_capita}} \\
& + 0.02679_{\text{Alcohol\_consumption} \times \text{BMI}} + 0.02003_{\text{Alcohol\_consumption} \times \text{Incidents\_HIV}} \\
& + 0.00000847_{\text{Alcohol\_consumption} \times \text{GDP\_per\_capita}} - 0.01104_{\text{Alcohol\_consumption} \times \text{Schooling}} \\
& + 0.02128_{\text{BMI} \times \text{Thinness\_ten\_nineteen\_years}} - 0.02941_{\text{BMI} \times \text{Schooling}} \\
& + 0.1104_{\text{Incidents\_HIV} \times \text{Schooling}} - 0.000007590_{\text{GDP\_per\_capita} \times \text{Thinness\_ten\_nineteen\_years}} \\
& + 0.00001029_{\text{GDP\_per\_capita} \times \text{Schooling}}
\end{aligned}
$$

Figure 13. Final model equation for developed countries

$$
\begin{aligned}
\hat{LifeExpectancy} = {} & 1199 - 0.2801_{\text{Under\_five\_deaths}} + 0.005279_{\text{Adult\_mortality}} - 0.8051_{\text{Alcohol\_consumption}} \\
& + 0.2535_{\text{Alcohol\_consumption}^2} - 0.04733_{\text{Alcohol\_consumption}^3} + 0.003228_{\text{Alcohol\_consumption}^4} \\
& - 0.00007369_{\text{Alcohol\_consumption}^5} - 171.8_{\text{BMI}} + 9.787_{\text{BMI}^2} \\
& - 0.246_{\text{BMI}^3} + 0.002303_{\text{BMI}^4} - 2.656_{\text{Incidents\_HIV}} + 0.6867_{\text{Incidents\_HIV}^2} \\
& - 0.1296_{\text{Incidents\_HIV}^3} + 0.01073_{\text{Incidents\_HIV}^4} - 0.0004181_{\text{Incidents\_HIV}^5} \\
& + 0.00000632_{\text{Incidents\_HIV}^6} + 0.0002914_{\text{GDP\_per\_capita}} - 0.00000002194_{\text{GDP\_per\_capita}^2} \\
& + 0.0000000006612_{\text{GDP\_per\_capita}^3} - 0.00000000000009642_{\text{GDP\_per\_capita}^4} + 0.0000000000000000006603_{\text{GDP\_per\_capita}^5} \\
& - 0.000000000000000000001716_{\text{GDP\_per\_capita}^6} + 0.1788_{\text{Thinness\_ten\_nineteen\_years}} \\
& - 0.2080_{\text{Thinness\_ten\_nineteen\_years}^2} + 0.02363_{\text{Thinness\_ten\_nineteen\_years}^3} - 0.001065_{\text{Thinness\_ten\_nineteen\_years}^4} \\
& + 0.00001647_{\text{Thinness\_ten\_nineteen\_years}^5} + 1.305_{\text{Schooling}} - 0.02936_{\text{Schooling}^2} \\
& + 0.00006012_{\text{Under\_five\_deaths} \times \text{Adult\_mortality}} - 0.001076_{\text{Under\_five\_deaths} \times \text{Alcohol\_consumption}} \\
& + 0.008992_{\text{Under\_five\_deaths} \times \text{BMI}} + 0.003744_{\text{Under\_five\_deaths} \times \text{Incidents\_HIV}} \\
& - 0.005619_{\text{Under\_five\_deaths} \times \text{Schooling}} - 0.002131_{\text{Adult\_mortality} \times \text{BMI}} \\
& + 0.0007749_{\text{Adult\_mortality} \times \text{Incidents\_HIV}} - 0.00000006439_{\text{Adult\_mortality} \times \text{GDP\_per\_capita}} \\
& + 0.02679_{\text{Alcohol\_consumption} \times \text{BMI}} + 0.02003_{\text{Alcohol\_consumption} \times \text{Incidents\_HIV}} \\
& + 0.00000847_{\text{Alcohol\_consumption} \times \text{GDP\_per\_capita}} - 0.01104_{\text{Alcohol\_consumption} \times \text{Schooling}} \\
& + 0.02128_{\text{BMI} \times \text{Thinness\_ten\_nineteen\_years}} - 0.02941_{\text{BMI} \times \text{Schooling}} \\
& + 0.1104_{\text{Incidents\_HIV} \times \text{Schooling}} - 0.000007590_{\text{GDP\_per\_capita} \times \text{Thinness\_ten\_nineteen\_years}} \\
& + 0.00001029_{\text{GDP\_per\_capita} \times \text{Schooling}}
\end{aligned}
$$

Figure 14. Final model equation for non-developed countries

4. Conclusion and Discussion

4.1 Conclusion

Living in a developed country generally improves life expectancy, but the effect is not straightforward.

The difference in life expectancy between developed and developing countries depends on other factors such as:
Under-Five Deaths: High rates of child deaths under five years old reduce the positive impact of living in a developed country on life expectancy.
Alcohol Consumption: Increased alcohol consumption in developed countries slightly diminishes the life expectancy advantage.
HIV Cases: Surprisingly, higher HIV cases in developed countries seem to amplify the life expectancy difference. This could be because developed countries might provide better access to healthcare and treatments for people with HIV.
GDP per Capita: Economic prosperity (measured as GDP per 1000 persons) plays a role, but its effect is relatively small.

Simply being in a developed country doesn't guarantee a higher life expectancy. Health and economic factors interact with development status to shape outcomes. For example, reducing child deaths and alcohol abuse while leveraging healthcare infrastructure could lead to further improvements in life expectancy.

Mortality rates (deaths in the adult population) affect life expectancy in ways that are influenced by other health and economic factors. Mortality rates are not an isolated factor. Addressing related issues like child deaths, improving public health (BMI), and increasing access to healthcare can mitigate the negative effects of mortality rates on life expectancy.

Life expectancy is not determined by a single factor. It’s the result of a complex interplay of health, economic, and social conditions. Improving life expectancy requires a multi-faceted approach. By addressing the key issues affecting child deaths, alcohol use, body weight, and economic conditions, countries can work toward healthier, longer lives for their populations.

4.2. Approach Discussion

The approach taken in this analysis is conceptually sound, as multiple linear regression is a widely used technique to model relationships between a response variable and its

predictors. However, its application in this context is limited because the data violates some of the key assumptions necessary for multiple linear regression. These assumptions include linearity, normality of residuals, homoscedasticity (constant variance of residuals), and the absence of multicollinearity. When these assumptions are not met, the reliability and validity of the results become questionable, as the model may no longer accurately represent the underlying relationships in the data.

To overcome these limitations, it is worth considering alternative modeling approaches that are less sensitive to assumption violations. Generalized Linear Models (GLMs) provide a flexible framework by allowing error distributions other than normal, making them suitable for datasets that do not meet the normality assumption. Similarly, regularization techniques such as Ridge or Lasso regression can address multicollinearity issues, ensuring more stable coefficient estimates. Additionally, incorporating methods that capture non-linear relationships, such as polynomial regression or splines, can better model the inherent curvature and complex interactions in the data. These alternatives offer a more robust and adaptable approach to understanding the factors affecting life expectancy.

4.3. Future Work

To address the limitations identified in the current analysis, several steps can be undertaken to enhance the robustness and applicability of the model. Exploring nonlinear regression models or generalized additive models (GAMs) can help account for complex relationships between predictors and the response variable. These methods provide greater flexibility in model specification and are better suited for datasets that deviate from strict linearity assumptions.

Investigating alternative distributions for the residuals is also critical, given the violation of the normality assumption. Generalized linear models (GLMs) can provide a more accurate representation of the data by accommodating non-normal error distributions.

Model validation through k-fold cross-validation is another important step. This approach ensures that the model's predictive accuracy is evaluated on unseen data, reducing the risk of overfitting and improving generalizability. Additionally, expanding the dataset by incorporating more recent data points or relevant predictors can enhance the representativeness of the analysis. This will help address potential sampling biases and increase the reliability of the results.

Lastly, incorporating updated datasets for recent years and investigating additional variables that influence life expectancy could provide more comprehensive insights. By

broadening the scope of the analysis, future studies can better capture the evolving factors that shape life expectancy trends across different countries and contexts.

This structured approach ensures that future analyses will be more robust, reliable, and reflective of the real-world complexities underlying life expectancy determinants.

5. References

1. Lasha. (2023). Life expectancy (WHO) fixed. Retrieved from https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated/data

2. Gareth James & Daniela Witten & Trevor Hastie Robert Tibshirani, An Introduction to Statistical Learning with Applications in R: Springer New York Heidelberg Dordrecht London

6. Appendix

Data sample:

	Country	Region	Year	Infant_dea	Under_five	Adult_mort	Alcohol_cc	Hepatitis_I	Measles	BMI	Polio	Diphtheria	Incidents_I	GDP_per_c	Population	Thinness_t	Thinness_f	Schooling	Economy_s	Economy_s	Life_expectanc
2	Turkiye	Middle Eas	2015	11.1	13	105.824	1.32	97	65	27.8	97	97	0.08	11006	78.53	4.9	4.8	7.8	0	1	76.5
3	Spain	European l	2015	2.7	3.3	57.9025	10.35	97	94	26	97	97	0.09	25742	46.44	0.6	0.5	9.7	1	0	82.8
4	India	Asia	2007	51.5	67.9	201.0765	1.57	60	35	21.2	67	64	0.13	1076	1183.21	27.1	28	5	0	1	65.4
5	Guyana	South Ame	2006	32.8	40.5	222.1965	5.68	93	74	25.3	92	93	0.79	4146	0.75	5.7	5.5	7.9	0	1	67
6	Israel	Middle Eas	2012	3.4	4.3	57.951	2.89	97	89	27	94	94	0.08	33995	7.91	1.2	1.1	12.8	1	0	81.7
7	Costa Rica	Central Am	2006	9.8	11.2	95.22	4.19	88	86	26.4	89	89	0.16	9110	4.35	2	1.9	7.9	0	1	78.2
8	Russian Fe	Rest of Eur	2015	6.6	8.2	223	8.06	97	97	26.2	97	97	0.08	9313	144.1	2.3	2.3	12	0	1	71.2
9	Hungary	European l	2000	8.7	10.1	192.969	12.23	88	99	25.9	99	99	0.08	8971	10.21	2.3	2.3	10.2	1	0	71.2
10	Jordan	Middle Eas	2001	22	26.1	129.764	0.52	97	87	27.9	97	99	0.13	3708	5.22	4	3.9	9.6	0	1	71.9
11	Moldova	Rest of Eur	2008	15.3	17.8	217.857	7.72	97	92	26.5	96	90	0.43	2235	2.87	2.9	3.1	10.9	0	1	68.7
12	Brazil	South Ame	2012	15.4	17.2	150.2245	7.12	96	70	26.1	96	95	0.24	9057	199.29	2.8	2.8	7.3	0	1	74.2
13	Malta	European l	2007	6	6.8	58.0185	7.47	82	77	27.1	76	74	0.08	19338	0.41	0.7	0.7	9.9	1	0	79.8
14	Bahamas,	Central Am	2011	13	15.2	165.538	9.23	95	83	27.6	97	98	0.46	32027	0.36	2.5	2.5	11	0	1	72.3
15	Ukraine	Rest of Eur	2002	14.3	16.6	261.6095	7.13	48	98	25.8	99	99	0.55	1660	48.2	2.9	3	10.5	0	1	68.3
16	Switzerlan	Rest of Eur	2006	4.2	4.9	63.2435	10.24	88	71	25	94	94	0.08	79368	7.48	0.6	0.4	12.3	1	0	81.5
17	Norway	Rest of Eur	2001	3.8	4.7	80.4955	5.49	88	90	25.7	91	91	0.03	68489	4.51	0.8	0.7	12.1	1	0	78.8
18	Finland	European l	2013	2.2	2.7	76.567	9.08	88	92	26	98	98	0.08	43045	5.44	0.9	0.8	12.3	1	0	81
19	Comoros	Africa	2007	66.8	91.9	255.8815	0.15	75	64	23.5	75	75	0.02	1166	0.64	7.3	7.2	3.5	0	1	60.7
20	Japan	Asia	2005	2.8	3.7	68.768	7.98	83	84	22.6	95	98	0.17	33099	127.77	1.7	1.5	11.2	1	0	81.9