How does Country Status Influence the Relationship Between Socioeconomic and Healthcare Indicators and Life Expectancy?

Contribution:

Both members contributed to the model decision process.

Tiffany Yang: focused primarily on revising the introduction, methods, ethics, editing summary, and poster design.

Shanjie Jiao: worked on revising the graphs, tables, models, codes, and the results and discussion sections.

1. Introduction (248 words)

This report applies multiple linear regression (MLR) to explore the linear relationship between socioeconomic and healthcare indicators and life expectancy, reflecting causal connections established by research done in recent decades. The analysis specifically investigates how country status influences the effects of adult mortality, alcohol consumption, BMI, Diphtheria, Polio, total expenditure, and HIV/AIDS on life expectancy.

One study using an OLS regression framework highlights the role of age in the relationship between economic development and life expectancy. It finds that, in younger populations, improved health drives economic growth, while in older populations, economic growth extends life expectancy. However, this study overlooks other critical factors like lifestyle or healthcare quality (He & Li, 2020).

Another study explores socioeconomic variables in developing regions, finding fertility rate as the only significant predictor of life expectancy among other variables. However, this report excludes key healthcare factors like health conditions or alcohol consumption, leaving room for deeper analysis in these areas (Kabir, 2008).

Additional research underscores the significant role of obesity and BMI in reducing life expectancy. A study revealed that obesity reduces lifespan of non-smoking men and women by 7.1 and 5.8 years of life, respectively. Although based on late-20th-century data, these findings remain relevant, as adult obesity continues to be a robust predictor of reduced life expectancy, making BMI an important factor for consideration in this report (Peeters et al., 2003).

Overall, these studies highlight the complex interplay between socioeconomic, healthcare indicators, and life expectancy, with varying levels of impact across variables. This report aims to validate these findings using the selected dataset.

2. Methods (468 words)

2.1. Data Collection and Cleaning

The research began by downloading the selected data from Kaggle and importing into R Studio. The dataset was cleaned and sorted: missing data was unnecessary and removed, then 1000 samples were randomly selected from the cleaned dataset. The dataset was then split into training and testing sets.

2.2 Variable Selection and Model Construction

The first step in model development involved variable selection to construct an initial model (**Model 1**) using ten predictors considered relevant to life expectancy. Insignificant variables were removed if p-value fall below 0.05, resulting in **Model 2**. Multicollinearity was assessed using the Variance Inflation Factor (VIF), with predictors exceeding a VIF threshold of 5 considered for removal, if necessary, remain as **Model 2** if no variable is removed. To further evaluate the significance of individual variables, an ANOVA test was conducted to assess whether removing a variable had a significant impact on the original model. Remove a variable and form **Model 3**, compare its p-value with that of Model 2 (full model). Reject the null hypothesis and proceed with **Model 2** if the p-value is below the 0.05 threshold; if not proceed with Model 3. The stepwise selection method was applied to refine the model and minimize the Akaike Information Criterion (AIC), resulting in **Model 4**. Compare Model 4 with Model 2, **Model 2** was ultimately chosen as the final model for providing the best fit for the data.

2.3 Model Validation

Model validation was performed by comparing key metrics, including Adjusted R-squared, R-squared, AIC, and BIC, to assess the performance of the models. Smaller AIC and BIC values indicated better model performance. The finalized model, developed using the training set, was applied to the testing set to evaluate its fit and performance by examining the consistency of results. Significant differences in performance between the training and testing sets could indicate overfitting or underfitting, suggesting the need for further model refinement.

2.4 Model Diagnostics

Diagnostics began by assessing **Condition 1** (Conditional Mean Response) through a scatterplot of response vs. fitted values, checking for visible non-linear trends. Similarly, **Condition 2** (Conditional Mean Predictors) was evaluated using pairwise scatterplots of predictors to ensure the absence of curves or non-linear patterns. Perform Box-Cox transformation if condition is not satisfied, then proceed onward.

To evaluate linearity, residual versus fitted value plots were analyzed, the absence of curves indicating that the assumption was satisfied. Constant variance was checked using the same plot, ensuring residuals were evenly distributed without fanning or contraction patterns. Independence was evaluated through residual versus order plots to confirm the absence of clustering or trends. The normality of residuals was examined using a Q-Q plot, where the points were expected to fall on a straight diagonal line.

Leverage analysis was performed using the Residuals vs. Leverage plot, and influential points were identified by examining Cook's distance. Observations with a Cook's distance greater than 1 were flagged as potentially problematic.

3. Results (764 words)

3.1 Data Description

This dataset, sourced from Kaggle and the WHO's Global Health Observatory (GHO), was uploaded by Kumar Rajarshi (Rajarshi, 2017). It includes data from 193 countries from 2000 to 2015, containing 2938 observations and 20 variables.

To explore differences between developed and developing countries, the variable *Status* was categorized as a dummy variable, assigning values 0 and 1 to developing and developed countries, respectively. After cleaning the data by removing irrelevant variables of this research, a random sample of 1000 observations was selected from the cleaned dataset. The analysis primarily focused on variables related to diseases, economic factors, causes of death, and immunity.

Variable	Description	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Life Expectancy	Life Expectancy in age	44.0	64.68	72.2	69.8	75.3	89.0
Adult Mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)	1.0	73.75	147.0	166.79	225.5	715.0
Alcohol	Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol)	0.01	0.95	3.92	4.6	7.33	17.87
BMI	Average Body Mass Index of the entire population	2.0	21.08	44.3	38.15	55.8	74.6
GDP	Gross Domestic Product	14.14	485.96	1644.19	5859.39	4690.88	119172.74
Diphtheria	An infection caused by the bacterium Corynebacterium diphtheriae	2.0	81.0	93.0	84.08	97.0	99.0
HIV/AIDS	Deaths per 1,000 live births HIV/AIDS (0-4 years)	0.1	0.1	0.1	1.74	0.8	42.1
Status	0 = developing country (total 852) 1 = developed country (total 37)						

Table 1. Statistic Summary of Train and Test Data with Variable Description.

The estimand of this research is life expectancy, influenced by the variables listed in *Table 1*, which also provides a statistical summary of the training and testing datasets along with detailed variable descriptions.

Figure 1 illustrates the distribution of variables in the dataset. Life expectancy is approximately normally distributed, with a slight leftward skew, a mean around 72, and a median that closely matches the mean. Adult mortality and alcohol consumption are highly positively skewed, with adult mortality concentrated around 200, indicating that about one-fifth of the population dies between ages 15-60, while alcohol shows substantial variability. The BMI distribution appears slightly bimodal, peaking in the ranges of 20–30 and 40–50. GDP contains extreme outliers due to the inclusion of developed countries, while most data points are from developing countries. Lastly, HIV/AIDS displays a heavily skewed distribution, with most values near zero and numerous outliers.

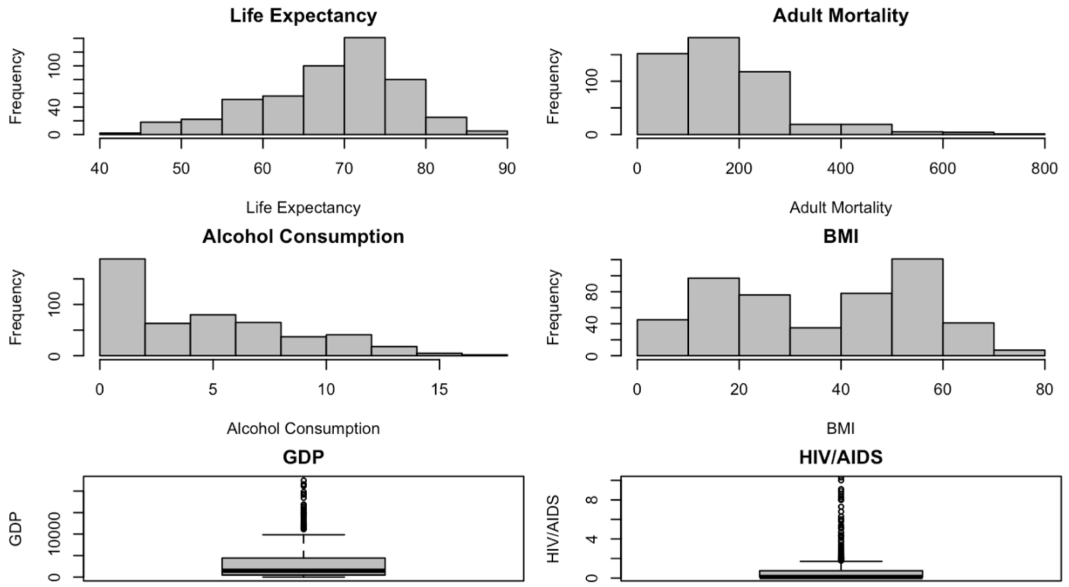

Figure 1: Distribution and Scatter Plot of Variable for Train Data

3.2 Model Selection

The cleaned data was divided into training and testing data with a ratio of 50%. The training data was used to fit a linear model, while the testing data was used to evaluate the final model.

Step 1: Based on the dataset and research topic, life expectancy was chosen as the outcome variable. Combined with other predictor variables of interest, **Model 1** was established.

$$\begin{aligned}\text{Life } \widehat{\text{Expectancy}} = 62.17 + 2.608(\text{Status}) - 0.2453(\text{Adult Mortality}) + 0.02644(\text{Alcohol}) + 0.1260(\text{BMI}) \\ - 0.03403(\text{Hepatitis B}) + 0.02183(\text{Total Expenditure}) + 0.05145(\text{Polio}) \\ - 0.01199(\text{Infant Deaths}) - 0.01676(\text{Measles}) - 0.3914(\text{HIV/AIDS})\end{aligned}$$

Step 2: By examining the P-values, Hepatitis B, Infant Deaths, and Measles were removed as they were insignificant, each having a p-value greater than 0.05. The remaining variables were used to constructed **Model 2**.

$$\begin{aligned}\text{Life } \widehat{\text{Expectancy}} = 61.81 + 2.618(\text{Status}) - 0.2445(\text{Adult Mortality}) + 0.2634(\text{Alcohol}) + 0.1279(\text{BMI}) \\ + 0.2254(\text{Total Expenditure}) + 0.05049(\text{Polio}) - 0.3905(\text{HIV/AIDS})\end{aligned}$$

Step 3: Multicollinearity was assessed using the VIF. Since none of the VIF values exceeded the threshold of 5, it was concluded that no multicollinearity issue exists among the variables in Model 2. However, significance testing and multicollinearity assessment are not enough. To evaluate whether Model 2 could be further simplified, a partial F-test was conducted. The variable HIV.AIDS was randomly selected for removal, resulting in **Model 3**.

$$\begin{aligned}\text{Life } \widehat{\text{Expectancy}} = 63.59 + 2.767(\text{Status}) - 0.03603(\text{Adult Mortality}) + 0.18999(\text{Alcohol}) + 0.13433(\text{BMI}) \\ + 0.05190(\text{Polio})\end{aligned}$$

Step 4: The partial F-test comparing Model 2 and Model 3 resulted in a p-value of less than 2.2e-16, which is significantly below the threshold of 0.05. Therefore, the null hypothesis is rejected, **Model 2** is carried forward for further analysis.

Step 5: To build a model with minimum AIC value, the stepwise method was applied to develop Model 4.

$$\begin{aligned}\text{Life } \widehat{\text{Expectancy}} = 61.81 - 0.0244(\text{Adult Mortality}) + 0.1279(\text{BMI}) - 0.3905(\text{HIV/AIDS}) \\ + 0.2634(\text{Alcohol}) + 0.0505(\text{Polio}) + 2.6178(\text{Status}) + 0.2254(\text{Total Expenditure})\end{aligned}$$

After conducting the partial F-test and removing HIV/AIDS, the stepwise selection resulted in the same model as Model 2, with identical variables and data. This suggests that further simplification does not improve the model. Thus, **Model 2** is selected as the final model. A summary table of the final model including coefficients and p-values, is presented in Table 2.

Predictors	Coefficients	P-value
(Intercept)	61.805688	< 2e-16
Status	2.617777	0.000967
Adult.Mortality	-0.024449	< 2e-16
Alcohol	0.263407	0.000172
BMI	0.127939	< 2e-16
Total.Expenditure	0.22541	0.021138
Polio	0.050493	2.76E-06
HIV/AIDS	-0.390504	< 2e-16

Table 2. Statistic Summary Table of Final Model (Model 2).

3.3 Model Goodness: Assumptions of Linear Regression

3.3.1 Transformation for Model 2

Model 2 exhibits a non-linear pattern, particularly in BMI and Alcohol (illustrated in Appendix A). To address these non-linear relationships observed in the pairwise scatter plot, a Box-Cox transformation was applied with a small constant added to avoid zeros, except Status. The results are presented as **Model 5**.

$$\begin{aligned}\text{Log}(\widehat{\text{Life Expectancy}}) = 4.12934 + 0.041776 \times \text{Status} - 0.014151 \times \text{Log(Adult Mortality)} +\\ 0.027959 \times \text{Log(Alcohol)} + 0.022718 \times \text{Log(BMI)} + 0.028631 \times\\ \text{Log(Total Expenditure)} + 0.010781 \times \text{Log(Polio)} - 0.112766 \times \text{Log(HIV/AIDS)}\end{aligned}$$

3.3.2 Conditions 1 and 2 Check for Model 5

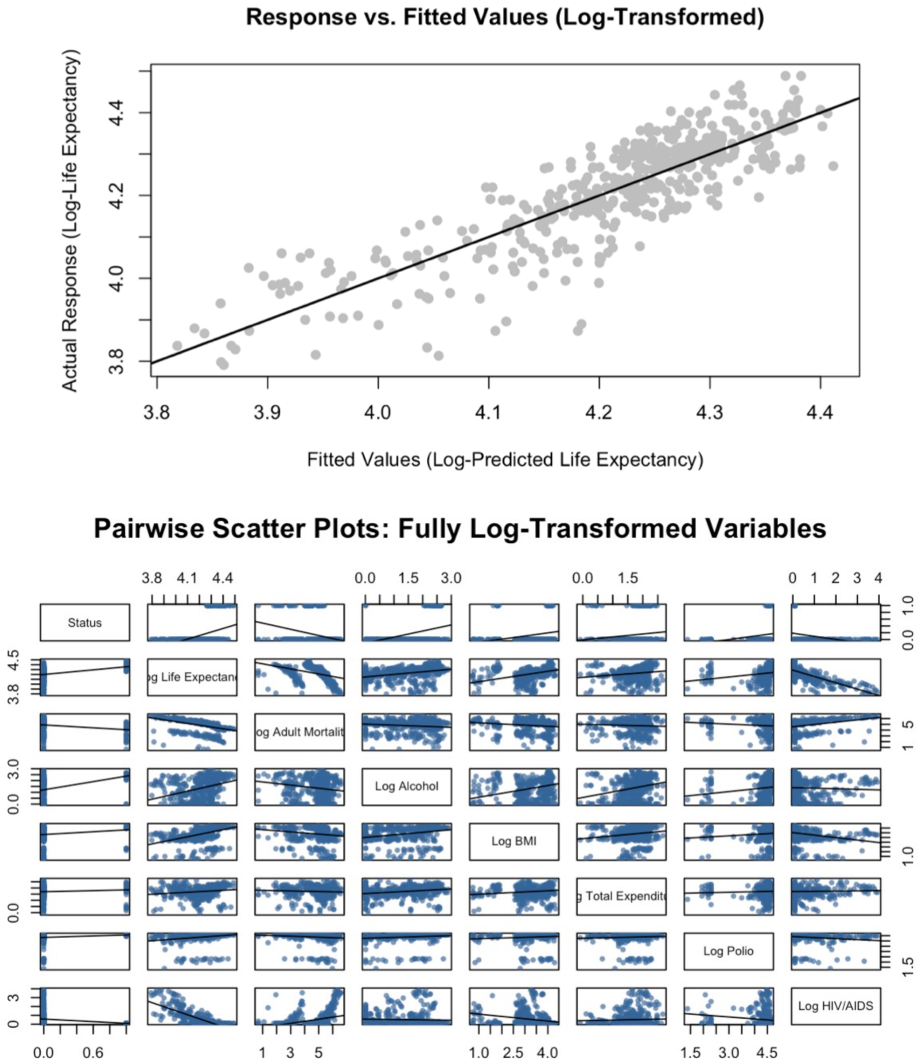

Figure 2. Response and Fitted Value and Pairwise Scatter Plots for Model 5.

Model 5 demonstrates improved linearity and reduced variability, as evidenced by the consistency in the response vs. fit plots and the clearer linear relationships observed in the paired scatter plots.

3.3.3 Assumption Check for Model 5

The residuals vs. fitted values plot shows no distinct patterns, supporting linearity and constant variance. Additionally, the residuals are well-scattered across predictors, indicating uncorrelated errors. The Q-Q plot closely aligns with the diagonal, confirming approximate normality. While Cook's Distance identifies a few influential points, they have minimal impact on the model. However, some leverage points are observed in the Residual vs. Leverage plot.

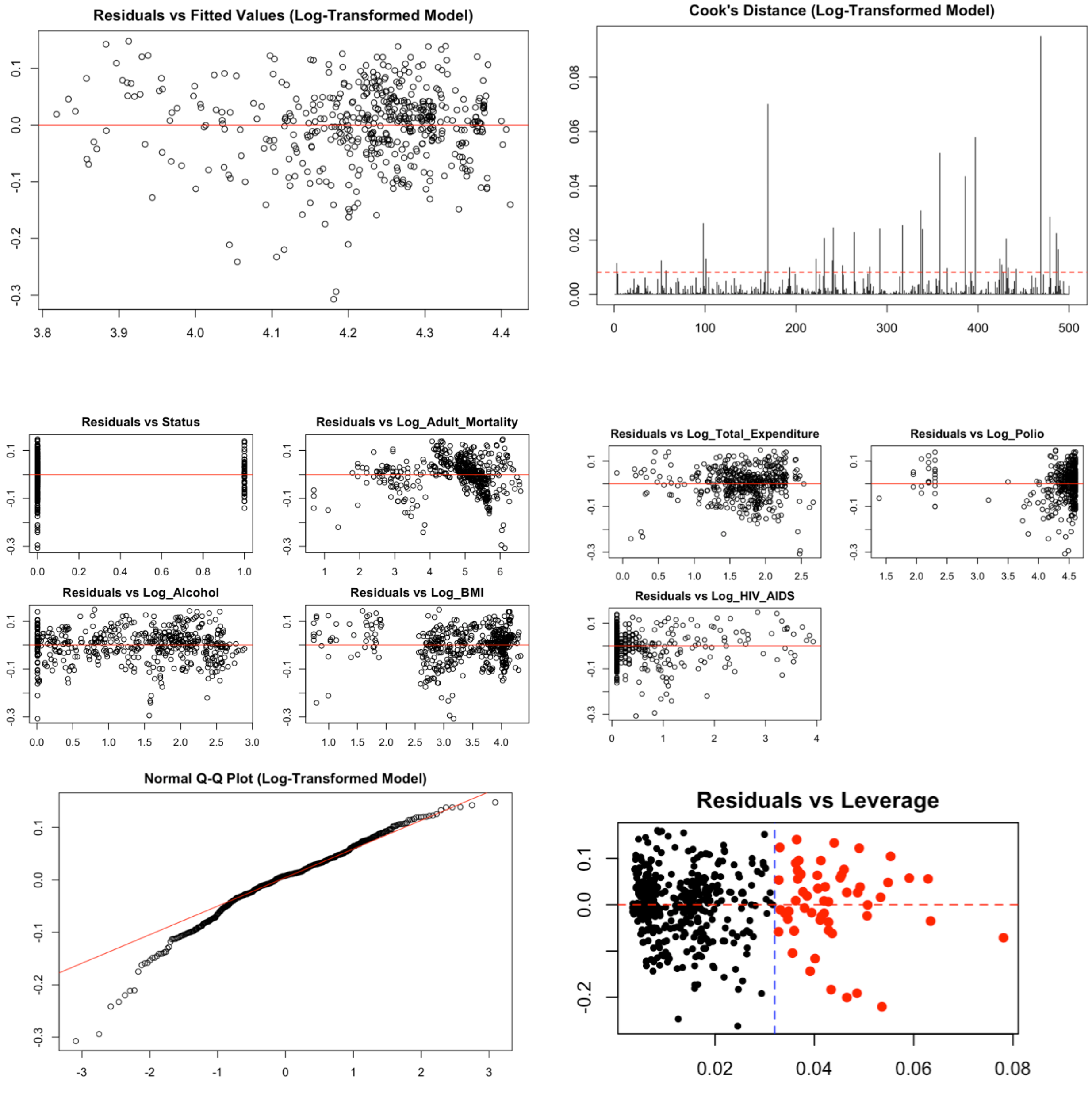

Figure 3. Combination of Residuals vs Variable, Q-Q Plot, and Cook’s Distance of Model 5

3.3.4 Test Data Model for Model 5

Using test data, we validate Model 5 for its accuracy and assess whether it suffers from overfitting, resulting in **Model 6**. The model is as follows:

$$\boxed{\begin{aligned}\text{Log}(\widehat{\text{Life Expectancy}}) = 4.1839 + 0.0519 \times \text{Status} - 0.0122 \times \text{Log}(\text{Adult Mortality}) +\\ 0.0262 \times \text{Log}(\text{Alcohol}) + 0.0120 \times \text{Log}(\text{BMI}) + 0.0224 \times \text{Log}(\text{Total Expenditure}) +\\ 0.0114 \times \text{Log}(\text{Polio}) - 0.1209 \times \text{Log}(\text{HIV/AIDS})\end{aligned}}$$

The detailed results of the test and train data comparison are shown in Appendix B. The R-squared values are 0.7448 for the training set and 0.7729 for the testing set, with a minor difference in RSE (0.06836 to 0.06436, respectively). Additionally, the AIC values for the training and testing datasets are -1254.1419 and -1314.4437, respectively, confirming that the model generalizes well without overfitting while maintaining good performance (detailed table shown in Appendix C).

4. <u>Conclusion and Limitations</u> (345 words)

4.1 Final Model Interpretation

$$\boxed{\begin{aligned}\text{Log}(\widehat{\text{Life Expectancy}}) = 4.12934 + 0.041776 \times \text{Status} - 0.014151 \times \text{Log}(\text{Adult Mortality}) +\\ 0.027959 \times \text{Log}(\text{Alcohol}) + 0.022718 \times \text{Log}(\text{BMI}) + 0.028631 \times\\ \text{Log}(\text{Total Expenditure}) + 0.010781 \times \text{Log}(\text{Polio}) - 0.112766 \times \text{Log}(\text{HIV/AIDS})\end{aligned}}$$

The final model highlights significant differences in life expectancy between developed and developing countries, as represented by the status variable (developed=1, developing=0). For a developing country, the estimated life expectancy is approximately 62.03 years, whereas for a developed country, it increases to 64.78 years. The positive coefficient of 0.0418 for the Status variable suggests that life expectancy in developed countries is approximately 4.28% higher than Furthermore, the impact of log(HIV/AIDS) is more negative in developing countries, reducing life expectancy by approximately 11.28%, compared to 10.65% in developed countries. Similarly, the impact of log(Adult Mortality) is consistently negative, with a 1% increase in adult mortality leading to an approximate 1.41% decrease in life expectancy for developing countries.

In contrast, the positive effects of log(Alcohol), log(BMI), and log(Total Expenditure) are stronger in developed countries, contributing to increases of approximately 2.83%, 2.29%, and 2.91%, respectively. These findings align with prior research discussed in the introduction, which highlights the advantages of better socio-economic, demographic, and public health conditions in developed countries.

4.2 Limitations

Two major limitations of this dataset are the lack of consideration of regional differences and incomplete data. The model relies on datasets that may be incomplete or inaccurate, particularly in countries with poor reporting systems, which can lead to biases and reduced accuracy. Additionally, classifying countries as either developed or developing overlooks significant regional variations. For instance, developing countries in Africa face different challenges than those in Southeast Asia, yet both are grouped as developing. More detailed, region-specific data would better capture these variations and provide a clearer understanding of regional disparities. For instance, the lack of such specificity may result in influential points, as identified in the Cook's Distance plot, distorting the model's fit and reducing its ability to generalize effectively to new datasets.

It is also noticeable that the test data provides better model-fitting performance than the training data, likely caused by the unequal distribution of developed and developing countries in the training and testing datasets. This imbalance may suggest potential overfitting. Further validation using additional datasets would help ensure the model's robustness and mitigate biases.

5. Ethics (238 words)

The final model was selected using the manual selection method, even though both methods resulted in nearly the same outcome. The manually selected model was chosen because both group members believed it to be more ethical compared to the automated generation of the model using the stepwise approach to minimize AIC.

Each step, from selecting the initial variables to deciding on a subset of the full model for conducting a partial F-test, was done with careful consideration, and remained under the group's control. The logic behind selecting the variables was guided by the three research papers reviewed in the introduction, and every step of the model selection process was transparent. This approach allowed the group to gain deeper insights into the dataset and the model. In contrast, the automated method generates the final model using algorithms, which may lack the ability to interpret real-life situations and can overlook relationships between variables. While the project's aim is a relatively general question, it still requires thoughtful consideration of the training data, and the construction of a model based on that.

Regarding the concept of blame discussed in the ethics workshop, choosing an automatically selected model shifts responsibility to the algorithm, removing accountability from the group for the choices made. When responsibility is shifted, blame is also shifted, potentially avoiding negligence. If problems were to arise, the algorithm could be blamed instead of the group taking full ownership of the mistake.

6. References

He, L., & Li, N. (2020). The linkages between life expectancy and economic growth: some new evidence. *Empirical Economics*, *58*(5), 2381-2402.

Kabir, M. (2008). Determinants of Life Expectancy in Developing Countries. *The Journal of Developing Areas*, *41*(2), 185–204. http://www.jstor.org/stable/40376184

Peeters, A., Barendregt, J. J., Willekens, F., Mackenbach, J. P., Mamun, A. A., Bonneux, L., & for NEDCOM, the Netherlands Epidemiology and Demography Compression of Morbidity Research Group*. (2003). Obesity in adulthood and its consequences for life expectancy: a life-table analysis. *Annals of internal medicine*, *138*(1), 24-32.

Rajarshi, K. (2017). *Life Expectancy (WHO)*. Www.kaggle.com. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

World Health Organization. (n.d.). *Data ethics*. World Health Organization. https://wkc.who.int/our-work/health-emergencies/knowledge-hub/health-data-management/data-ethics

7. Appendix

Appendix A: Response and Fitted Value and Pairwise Scatter Plots for Model 2

This appendix provides diagnostic plots for **Model 2** assessing its goodness-of-fit and to check for non-linear relationships among predictors.

Figure A1 shows the linear relationship between the actual response (life expectancy) vs. the fitted values (predicted life expectancy).

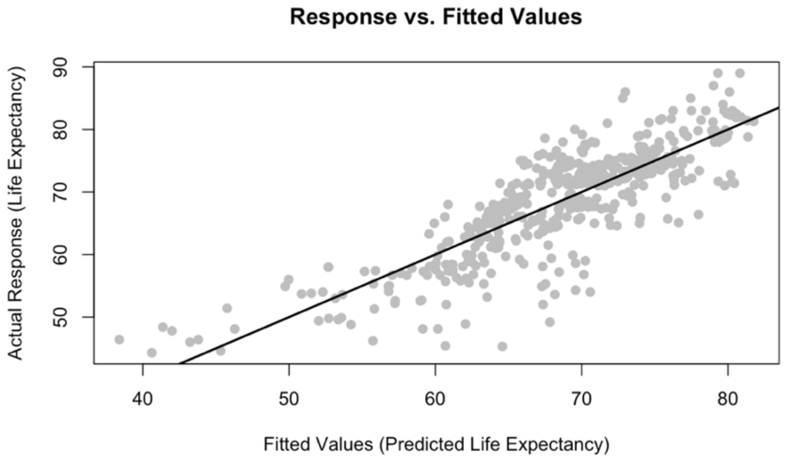

Figure A1. Response and Fitted Values for Model 2

Figure A2 demonstrates a set of pairwise scatter plots showing potential non-linear relationships between predictors, such as BMI and alcohol, as well as the response variables

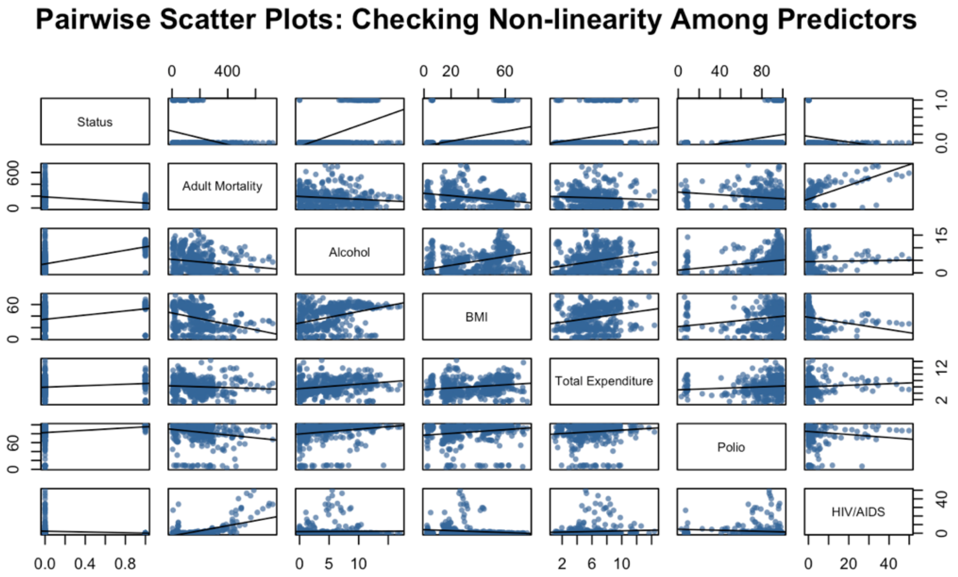

Figure A2. Pairwise Scatter Plots for Non-linearity checks for Model 2

Appendix B: Statistical Summary Table of Model 5 and Model 6

This appendix includes a summary of the statistical metrics for **Model 5** (train) and **Model 6** (test), including model coefficients, residual standard error, and model selection criteria.

Metric	Train_Model_5	Test_Model_6
Intercept	4.1294	4.1839
Status	0.0418	0.0519
Log_Adult_Mortality	-0.0142	-0.0122
Log_Alcohol	0.028	0.0262
Log_BMI	0.0227	0.012
Log_Total_Expenditure	0.0286	0.0224
Log_Polio	0.0108	0.0114
Log_HIV_AIDS	-0.1128	-0.1209
Residual Standard Error (RSE)	0.06836	0.06436
SSres	2.2989	2.0377
R^2	0.7448	0.7729

Adjusted R^2	0.7412	0.7697
AIC	-1254.1419	-1314.4437
AICc	-1253.8486	-1314.1504
BIC	-1216.2104	-1276.5122

Table B1. Statistical Summary Table of Model 5 and Model 6 (Train and Test)