Life Expectancy: An Analysis and Prediction Based on Key Influencing Factors

November 30, 2024

Contents

1 INTRODUCTION

1.1 MOTIVATION

1.1.1 Context

Life expectancy is a crucial indicator of a country's overall health and quality of life. It is used by governments, health organizations, and policymakers to measure progress in public health, identify disparities, and allocate resources effectively. By analyzing health, economic, and social indicators, this project aims to uncover the strongest predictors of life expectancy across countries and regions. The applied domain of this research is public health and global development, focusing on the interplay between critical variables such as mortality rates, immunization coverage, and economic conditions. This analysis holds the potential to inform targeted strategies for improving health outcomes worldwide.

1.1.2 Problem

The central research question we address is: What are the most significant factors influencing life expectancy across countries and regions? While numerous indicators are believed to impact life expectancy, determining their relative importance and understanding their interactions remain complex tasks. We aim to construct a multiple linear regression model to explore these relationships and provide actionable insights. Specifically, this model will help answer research questions such as:

- How do immunization rates and healthcare access correlate with life expectancy?
- What role do socio-economic factors like GDP per capita and schooling play in influencing life expectancy?
- Are there significant differences in life expectancy predictors across regions?

1.1.3 Challenges

This problem presents several challenges:

1. Diverse Predictors: Life expectancy is influenced by a combination of health metrics, socio-economic factors, and categorical variables (e.g., region). Balancing these diverse predictors in a single model can be complex.

2. Multicollinearity: Many predictors, such as infant mortality, under-five mortality, and immunization rates for Diphtheria and Polio, are highly correlated with each other. Although removing the countries column and replacing it with Region has helped reduce multicollinearity, caution is still needed in addressing relationships between these variables.

3. Exclusion of Country-Specific Effects: By removing the countries column and relying on Region instead, we lose the ability to capture country-specific nuances. While regions share common characteristics, individual countries within a region can have significantly different healthcare systems, policies, and socio-economic conditions, which might influence life expectancy in ways that regional-level aggregation cannot fully explain.

4. Limited Scope: Focusing on regions rather than countries simplifies the model and addresses multicollinearity but may overlook important local factors that could affect life expectancy.

1.2 OBJECTIVES

1.2.1 Overview

The overall intent of this project is to identify the key factors influencing life expectancy across different countries and regions. By leveraging multiple linear regression, we aim to uncover relationships between life expectancy and various health, economic, and social indicators. The goal is to build a predictive model that highlights which factors have the strongest impact on life expectancy, thereby providing insights that can guide policy decisions and interventions aimed at improving health outcomes worldwide.

1.2.2 Goals & Research Questions

The primary goals of this project are to:

- Develop a multiple linear regression model that identifies and quantifies the relationship between life expectancy and several predictor variables.
- Investigate how health indicators (e.g., infant mortality, immunization rates), socio-economic factors (e.g., GDP, schooling), and lifestyle factors (e.g., alcohol consumption, BMI) influence life expectancy at a regional level.
- Provide actionable insights for policymakers by identifying which factors have the most significant impact on life expectancy.

The specific research questions guiding this project are:

1. What are the strongest predictors of life expectancy across different regions?

2. How do those predictors correlate with life expectancy?

3. Are there significant differences in life expectancy predictors across different regions?

These objectives aim to enhance understanding of life expectancy trends and offer insights into areas for improvement in global health policy.

2 METHODOLOGY

2.1 Data

The dataset [Figure 2.1] is publicly available under the CC0: Public Domain and can be accessed via Kaggle. The CC0 license allows us to freely use, share and modify it without any legal restrictions, including for any commercial purposes.

The data was compiled by the World Health Organization (WHO) and United Nations. It includes information on global life expectancy and its related factors. The dataset was collected under the conditions of global health surveys conducted by the WHO and related health organizations. The sample size spans 193 countries and includes data from the years 2000 to 2015. The sampling method is based on the available health and demographic data.

Each row in the dataset represents a country's life expectancy, along with several demographic and health-related indicators such as income, education and healthcare factors.

Response Variable

The focal response variable for this analysis is Life Expectancy. Life expectancy is a quantitative, continuous variable representing the average number of years a person is expected to live, based on statistical averages. It is measured in years and is recorded with whole values.

Predictor Variables

1. Country

Country is a qualitative variable that represents the name of the country for each observation. While not used directly in the analysis, this variable helps to identify the data for each country.

2.Region

Region is a qualitative variable representing the geographic region to which each country belongs. This variable is crucial for grouping countries based on regional similarities

3. Year

Year is a qualitative variable that represents the specific year in which the data was collected. It spans from 2000 to 2015 and serves as a categorical identifier for different time periods. The inclusion of this variable allows us to analyze trends in life expectancy over time and examine how health, socio-economic, and environmental factors evolve from year to year.

4. Infant deaths

Infant Deaths is a quantitative variable that measures the number of infant deaths per 1,000 live births. Infant mortality is a critical indicator of a country's healthcare quality, particularly in terms of prenatal and neonatal care.

5. Under five deaths

Under 5 deaths is a quantitative variable representing the number of deaths of children under 5 per 1000 live births. This indicator is a key measure of healthcare quality and child health, reflecting the effectiveness of healthcare systems in preventing early childhood deaths.

6. Adult mortality

Adult mortality is a quantitative variable measures the number of deaths pers 1000 adults aged 15-60 years. This indicator is often used as a proxy for healthcare quality and public health, as higher mortality rates in this age group typically reflect poor healthcare infrastructure or insufficient public health systems.

7. Alcohol consumption

Alcohol consumption is a quantitative variable measured in litres per capita. This variable represents the average amount of alcohol consumed per person in a country. High levels of alcohol consumption are linked

to various health issues, including liver disease, accidents, and chronic illnesses, all of which can reduce life expectancy.

8. Hepatitis B

Hepatitis B is a quantitative variable that represents the percentage of the population with chronic hepatitis B infection. Hepatitis B is a significant risk factor for liver disease and can reduce life expectancy, particularly in countries with high prevalence rates.

9. Measles

Measles is a quantitative variable representing the number of reported measles cases per 1,000 people. Measles is a preventable disease, and high rates of measles are often indicative of poor vaccination coverage or inadequate healthcare systems, both of which can negatively impact life expectancy.

10. BMI

BMI is a quantitative variable that measures the average BMI of the population. High BMI values, particularly those indicating obesity, are associated with numerous health risks, including cardiovascular disease and diabetes, which can lower life expectancy.

11. Polio

Polio is a quantitative variable that measures the percentage of children who have received the polio vaccine. High vaccination rates are typically associated with better public health outcomes and lower rates of infectious diseases, which contribute to higher life expectancy.

12. Diphtheria*

Diphtheria is a quantitative variable that measures the percentage of children vaccinated against diphtheria. High vaccination coverage for diphtheria indicates better healthcare systems and disease prevention efforts, which are correlated with longer life expectancy.

13. Incidents HIV

Incidents HIV is a quantitative variable that measures the percentage of the adult population living with HIV. The disease has a significant impact on mortality rates, especially in countries with high infection rates.

14. GDP per capita

GDP is a quantitative variable that measures a country's total economic output, divide by its population. It is reported in USD. GDP per capita serves as an indicator of a country's economic prosperity, with higher GDP typically associated with better healthcare access, improved infrastructure, and overall higher living standards.

15. Population

Population is a quantitative variable that represents the total population of each country in the dataset. This variable, while not directly influencing life expectancy in a causal sense, helps contextualize the data, especially when comparing countries of different sizes.

16. Thinness 10-19 years

Thinness 10-19 is a quantitative variable that represents the percentage of the population aged 10-19 years with a BMI below the healthy range. High rates of thinness in this age group may indicate malnutrition or insufficient healthcare, both of which can reduce life expectancy.

17. Thinness 5-9 years

Thinness 5-9 is a quantitative variable that represents the percentage of children aged 5-9 years who are underweight (based on BMI). High thinness rates in children are indicative of poor nutrition, which can adversely affect child development and overall life expectancy in the population.

18. Schooling

Schooling is a quantitative variable that measures the average number of years of schooling in a country. This indicator is important because higher educational attainment is associated with better health literacy, healthier behaviors, and greater access to healthcare resources.

19. Economy Status

Economic Status is a qualitative variable that classifies countries based on their development status, typically categorized as developed or developing. Developed countries tend to have higher levels of healthcare infrastructure, better education systems, and more robust social services, which all contribute to longer life expectancy.

2.1.1 Data Distribution

We analyzed the distribution of key variables, including Region, Year, GDP, economic status, infant mortality, adult mortality, Polio, BMI, Schooling, Incidents HIV, Hepatitis B, Diphtheria, Measles and alcohol consumption. Below are the histograms illustrating the spread and skewness of each variable in our dataset.

[Figure 2.2]

[Figure 2.3]

[Figure 2.4]

[Figure 2.5]

[Figure 2.6]

[Figure 2.7]

[Figure 2.8]

[Figure 2.9]

[Figure 2.10]

[Figure 2.11]

[Figure 2.12]

[Figure 2.13]

[Figure 2.14]

[Figure 2.15]

2.2 Approach

For our data analytics solution, we utilized a dataset sourced from Kaggle. This dataset provides information on various factors related to life expectancy across 173 countries, with one column specifically labeled "Country." Since we have multiple independent variables influencing the dependent variable, Life Expectancy, we will employ Multiple Linear Regression to analyze and model these relationships effectively. Due to the complexity of handling individual countries and creating sub-models for each, as advised by our instructor, we used the Region column. This approach helped simplify the analysis and provided meaningful regional insights.

We started by creating a first-order model by eliminating the Country column. To ensure the robustness of this model, we applied the multicollinearity assumption by using statistical measures like Variance Inflation Factor (VIF). Based on the results of the multicollinearity check, we refined the model by removing highly correlated variables and built a second version of the model. To assess the significance of the predictors, we performed the F-test for overall model significance and t-tests for individual predictors. Using these results, we created a refined model that only included statistically significant terms. Next, we implemented stepwise

regression using stepmod and evaluated the model using criteria such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Mallows' Cp to verify the model's validity and select the best fit.

Additionally, we create an interaction model. From this interaction model, we extended our analysis to a Higher-order model to capture any non-linear relationships among predictors. The Higher-order model was tested to determine if it was a better fit for the data. If the model contained significant terms, it was considered a valid model and was retained for further analysis. Finally, we performed model evaluation and checked for adherence to residual assumptions, including normality, independence, and homoscedasticity, to ensure the model's reliability. This systematic approach allowed us to refine the model step by step, ensuring it was both statistically sound and predictive.

By following this method, we ensured that our model accounted for the complexity of the data while maintaining interpretability. The use of statistical tests and validation techniques made the approach robust, and we are confident it will provide meaningful and reliable insights for the data analytics solution.

2.3 Workflow

For this project, we followed a structured approach to model development, which can be broken down into several key tasks. Below is a detailed workflow [Figure 2.1] outlining the steps we undertook, the challenges faced, and how we dealt with any difficulties.

1. Getting the Data:

We started by acquiring a dataset from Kaggle, which contained various variables related to life expectancy across world. One of the columns, Country, had data for 173 countries. Due to the high number of unique values in this column, creating separate sub-models for each country would have been impractical. Following the advice of our instructor, we decided to focus on the Region column instead, which making the modeling process more manageable.

2. Data Cleaning and Model Initialization:

After not considering the Country column, we created a first-order model using the remaining variables. This model provided a baseline for further analysis.

3. Multicollinearity Testing:

We applied the assumption of multicollinearity by calculating the Variance Inflation Factor (VIF) for each predictor in the model and we plot a graph for highly correlated terms. The goal was to identify any highly correlated variables that might distort the model's performance. Based on the results, we eliminated some variables that showed high multicollinearity, ensuring that our model remained stable and reliable.

4. Performing F-test and Individual t-tests:

Once we cleaned the model by addressing multicollinearity, we performed an F-test to check the overall significance of the model and individual t-tests to assess the significance of each predictor. By evaluating these tests, we were able to refine our model and eliminate any non-significant terms.

5. Stepwise Model Refinement:

To further improve the model, we applied stepwise regression using the stepmod function. This iterative process helped us identify and retain only the most significant terms. Additionally, we calculated evaluation criteria such as AIC, BIC, and Mallows' Cp to verify the adequacy of the model.

6. Creating Interaction Terms:

After refining the model, we created interaction terms to capture the relationships between different predictors. This step was crucial for identifying potential synergies between variables. If any interaction terms proved to be non-significant, we removed them and repeated this process until we were left with only significant terms. This approach helped ensure that our model was both interpretable and valid.

7. Exploring High-Correlation Terms and Higher-Order Models:

Using the GGally package, we created a plot to identify the terms with the highest correlations. Based on these findings, we selected the most highly correlated terms and tested higher-order models incorporating them. If the addition of a higher-order term was statistically significant and the adjusted R-squared (R^2_adj) value increased, we proceeded to include higher-order terms. If the additional term did not improve the model, we reverted to the previous model with one less term. We continued this process by moving to the next higher-order term and repeating the evaluation until all terms with higher correlation were tested. This iterative approach helped us identify the best-fitting model for the data.

8. Building and Evaluating Higher-Order Models:

As we iterated through different model orders, we built higher-order model and checked the performance. If the models had significant terms and improved upon previous models, we retained them as the final model. If not, we returned to the previous steps, selected the next highest correlated term, and continued building higher-order models until we arrived at the best model.

9. Model Evaluation and Residual Assumption Check:

Once the final model was determined, we performed a Residual Assumption Check. This step involved evaluating key assumptions, such as homoscedasticity, normality of residuals, and the absence of autocorrelation. By verifying these assumptions, we ensured that our model met the necessary criteria for reliable predictions.

3 MAIN RESULTS OF THE ANALYSIS

3.1 Research Questions Result:

Bases on the Correlation matrix [Figure 3.1]

The 3 most correlated terms are:

- Infant deaths
- Adult mortality
- Economy status Developed

Result (Infant deaths): Based on the graphical analysis [Figure 3.2], there is a clear linear relationship between life expectancy and infant deaths. Life expectancy decreases as the number of infant deaths per 1,000 population increases. This indicates a significant negative correlation, where higher infant mortality is associated with shorter life expectancy.

Result (Adult mortality): The plot of life expectancy versus adult mortality [Figure 3.3] shows a strong inverse relationship. As adult mortality rates increase, life expectancy decreases significantly. This highlights the critical impact of adult mortality on life expectancy.

Result(Economy status): From graph [Figure 3.4], life expectancy is notably higher in developed countries compared to developing countries. The distinction between the two groups is evident, with developed countries consistently achieving higher life expectancy values. This underscores the role of economic status in influencing life expectancy.

3.2 Multicollinearity Result:

3.2.1 Initial Full Model: [Figure 3.5]

$$LifeExpectancy = 83.301550588 + 0.373540986 * RegionAsia + 1.945042477 * RegionCentralAmericaandCaribbean - 0.657631937 * RegionEuropeanUnion + 0.267344658 * RegionMiddleEast + 0.664637877 * RegionNorthAmerica -$$

$0.813834987 * RegionOceania + 0.280601144 * RegionRestofEurope + 1.734598696 * RegionSouthAmerica + 0.071750017 * Year_{Y2001} + 0.083610190 * Year_{Y2002} + 0.004729251 * Year_{Y2003} + 0.035900270 * Year_{Y2004} - 0.001230519 * Year_{Y2005} + 0.014358678 * Year_{Y2006} + 0.043131386 * Year_{Y2007} + 0.097431267 * Year_{Y2008} + 0.155014184 * Year_{Y2009} + 0.238632483 * Year_{Y2010} + 0.268800877 * Year_{Y2011} + 0.295888104 * Year_{Y2012} + 0.410287395 * Year_{Y2013} + 0.505466388 * Year_{Y2014} + 0.528827324 * Year_{Y2015} - 0.052710586 * Infant_deaths - 0.051038167 * Under_five_deaths - 0.046575680 * Adult_mortality - 0.007583318 * Hepatitis_B - 0.004477760 * Alcohol_consumption + 0.001943927 * Measles - 0.132934140 * BMI + 0.009792741 * Polio + 0.008138577 * Diphtheria + 0.090954743 * Incidents_HIV + 0.000020266 * GDP_per_capita - 0.000227516 * Population_mln - 0.036645320 * Thinness_ten_nineteen_years + 0.025003027 * Thinness_five_nine_years + 0.100686332 * Schooling + 2.501124616 * EconomyStatusDeveloped_Yes$

Outcome: The initial VIF analysis detected multicollinearity for four predictors: Infant_deaths, Under_five_deaths, Polio, and Diphtheria.

Reason for Detection: High VIF values were observed, exceeding the commonly accepted threshold of 10 for Infant_deaths (48.20), Under_five_deaths (50.05), Polio (12.20), and Diphtheria (13.26), indicating strong collinearity among these variables.

Resolution: To address this issue, one variable was removed from each highly correlated pair. Specifically:

- Under_five_deaths was removed, as it was closely related to Infant_deaths.
- Diphtheria was removed, as it was closely related to Polio.

3.2.2 Revised Model: [Figure 3.6]

Outcome: After excluding Under_five_deaths and Diphtheria, the VIF values were recalculated for the remaining predictors.

Results: No multicollinearity was detected in the revised model. The highest VIF values in the revised model (Infant_deaths: 8.64 and Thinness_five_nine_years: 9.20) were below the critical threshold, confirming that multicollinearity was sufficiently reduced.

Significance: The removal of collinear predictors improved model stability, ensuring that coefficients are more interpretable and estimates more reliable.

3.3 F-Test Result Result:

The F-test was conducted to determine if the revised model contains at least one predictor that significantly explains variation in life expectancy.

Based on image [Figure 3.7], the p-value for the F-test is extremely low, significantly below 0.05, which means that at least one predictor in the revised model significantly varies with life expectancy. This indicates that the revised model contains a predictor that is statistically significant in explaining the variation in life expectancy.

3.4 Individual T-test Result:

In the refined model, we removed the following insignificant variables: Alcohol_consumption, Measles, and Population_mln, as they had high p-values indicating no significant effect on life expectancy. The Year variable was kept because it is categorical, and if any year within the range is significant, the entire term is retained. The remaining predictors, including Region, Infant_deaths, Adult_mortality, Hepatitis_B, BMI, Polio, Incidents_HIV, GDP_per_capita, Thinness_ten_nineteen_years, Thinness_five_nine_years, Schooling, and Economy_status_Developed, were retained due to their strong statistical significance.

Output for those iteration are shown below in Figures,

Model based on Full F-Test result, [Figure 3.8] Model After performing Individual t-test, [Figure 3.9]

3.5 Stepwise Result:

The stepwise selection procedure resulted in a model with an adjusted R-squared of 0.9826, indicating a strong fit. Notably, the variables "Thickness_ten_nineteen_years" and "Thickness_five_nine_years" were not included, likely due to their minimal contribution to increasing the adjusted R-squared. However, we chose to proceed with the additive model that retains these thickness terms for comparison.

Result of Step-wise Model Selection as image shown here, [Figure 3.10]

3.6 All-Possible-Regressions Selection Procedure:

Based on output show in the image [Figure 3.24], The models in your output show improvement in fit as you progress from model 1 to model 13. Specifically, model with all variables provides the best balance of fit (high R-squared, low AIC, and BIC) with the smallest Mallows' Cp value, indicating that it has a good selection of predictor variables without excessive over-fitting. Therefore, model with all variables would likely be the most optimal model in this context.

3.7 Interaction model Result:

For the interaction model, we began by including all pairwise interaction terms and iteratively removed insignificant terms until all remaining interaction terms were statistically significant. However, we decided to proceed with the additive model. Although the adjusted R squared of the final interaction model (0.9921) was slightly higher than that of the additive model (0.9826), the improvement was minimal. Considering the risk of over-fitting with the interaction model, we chose the additive model for its simplicity and robustness.

Result of Interaction Model Summary as image shown here,

[Figure 3.11]

[Figure 3.12]

3.8 Higher order Result:

For the higher-order model, we examined the correlation matrix and identified highly correlated terms, specifically Adult_mortality and Infant_deaths. Incorporating second-order terms for these variables yielded adjusted R squared values of 0.9844 and 0.9835 , respectively. While these values show a slight improvement compared to the additive model (adjusted R squared = 0.9826), the increase is minimal. To avoid over fitting and maintain model simplicity, we decided not to proceed with the higher-order model and retained the additive model.

Result of Higher Order Model Summary as image shown here, [Figure 3.13]

3.9 Residual Checking:

3.9.1 Linearity Assumption Result:

The linearity assumption was assessed using Residuals vs Fitted Values plots as show below,

[Figure 3.14] [Figure 3.15] [Figure 3.16] [Figure 3.17] [Figure 3.18]

Adjustments were made to address initial non-linear patterns by integrating second-order and third-order terms for highly correlated predictors, Adult_mortality and Infant_deaths, respectively. After these modifications, the residuals from the plot showed no discernible pattern, indicating the linearity assumption was satisfied. The final model, incorporating these higher-order terms, achieved an adjusted R squared value of 0.9846 and an RMSE of 1.167, demonstrating strong predictive accuracy.

3.9.2 Independence Assumption Result:

The independence assumption is satisfied as our data is not related to time, space, or group, ensuring that the residuals are independent of each other.

3.9.3 Equal Variance Assumption Result:

A significant result, p-value < 2.2e-16 which is less than 0.05 from the Breusch-Pagan test suggests heteroscedasticity, meaning the variance of the residuals is not constant across levels of the independent variables. We reject null hypothesis.

[Figure 3.19]

After applying the Box-Cox transformation in an attempt to resolve the issue of heteroscedasticity, the Breusch-Pagan test confirmed that heteroscedasticity still persists in the residuals, indicating that the transformation was not sufficient to address the variance instability.

[Figure 3.19]

[Figure 3.20]

3.10 Normality Assumption:

Plotted Q-Q plot [Figure 3.21], indicate that the residuals follow a normal distribution. This suggests that the normality assumption of the regression model is satisfied.

The Shapiro-Wilk normality test further supports that the residuals are normally distributed, as the p-value (0.06246) is greater than 0.05. Therefore, we fail to reject the null hypothesis, indicating that the residuals follow a normal distribution.

[Figure 3.22]

3.11 Outliers Result:

Based on the graph, there are no outliers identified using Cook's Distance with a threshold of 0.5. This indicates that no observations exceed this threshold and are deemed influential in the context of the model.

[Figure 3.23]

Final Model Equation Result:

$$\widehat{Life_expectancy} = 83.846 + 0.5985 * Region_{Asia} + 2.134 * Region_{CentralAmericaandCaribbean} - 0.5379 * Region_{EuropeanUnion} + 0.07193 * Region_{MiddleEast} + 0.7183 * Region_{NorthAmerica} - 0.4468 * Region_{Oceania} + 0.3035 * Region_{RestofEurope} - 1.885 * Region_{SouthAmerica} + 0.06593 * Year_{Y2001} + 0.06393 * Year_{Y2002} + 0.004886 * Year_{Y2003} + 0.05015 * Year_{Y2004} + 0.01471 * Year_{Y2005} + 0.04960 * Year_{Y2006} + 0.08880 * Year_{Y2007} + 0.1522 * Year_{Y2008} + 0.2178 * Year_{Y2009} + 0.2926 * Year_{Y2010} + 0.3463 * Year_{Y2011} + 0.3679 * Year_{Y2012} + 0.4794 * Year_{Y2013} + 0.5691 * Year_{Y2014} + 0.5864 * Year_{Y2015} - 0.1189 * InfantDeaths - 0.0003900 * InfantDeaths^2 + 0.000003453 * InfantDeaths^3 - 0.06159 * AdultMortality + 0.00002649 * AdultMortality^2 + 0.01159 * HepatitisB - 0.1046 * BMI + 0.01180 * Polio - 0.03977 * IncidentsHIV +$$

$0.000007030 * GDP_per_capita - 0.04017 * ThinnessTenNineteenYears + 0.04967 * ThinnessFiveNineYears + 0.1097 * Schooling + 2.084 * EconomyStatusDeveloped_{Yes}$

Interpretation:

Intercept: This is the baseline life expectancy when all the other variables are zero life expentancy would be 8.385 years. It's the starting point of the model.

Region: Each coefficient for a specific region (e.g., Region_{Asia}, Region_{Central America and Caribbean}) represents the change in life expectancy relative to the baseline region (which is Africa). A positive coefficient indicates that life expectancy is higher in that region compared to the baseline region, while a negative coefficient means life expectancy is lower. For example, Region_{Asia} = 0.5985 indicates life expectancy in Asia is 0.5985 years higher than the baseline region. And Region_{Oceania} = - 0.4468 indicates life expectancy in Oceania is 0.5985 years lower than the baseline region.

Year: Each coefficient for a year (e.g., Year_{Y2001}, Year_{Y2002}) represents the change in life expectancy relative to the baseline year (usually the year not included in the model, such as Y2000). A positive coefficient indicates that life expectancy has increased in that particular year compared to the baseline year, while a negative coefficient suggests a decrease. For instance, Year_{Y2001} = 0.06593 means life expectancy in 2001 is 0.06593 years higher than in the baseline year.

InfantDeaths: The negative coefficient for InfantDeaths (-0.1189) suggests that an increase in infant deaths decreases life expectancy.If InfantDeaths increases by one unit, the life expectancy decreases by 0.1189 years, assuming all other factors are constant. This suggests that higher infant death rates are negatively associated with life expectancy, meaning that as infant deaths increase, the overall life expectancy in the population decreases.

InfantDeaths^2: The negative coefficient for the square of The Quadratic term (-0.0003900 * InfantDeaths^2) increase in infant deaths increase life expectancy. If InfantDeaths^2 increases by one unit, the life expectancy decreases by 0.0003900 years, assuming all other factors are constant. This suggests that as infant deaths increase, the negative impact on life expectancy accelerates in a non-linear way. The quadratic term captures the increasing severity of infant mortality's effect on life expectancy as the number of infant deaths rises.

InfantDeaths^3: The negative coefficient for the square of The Cubic term increase in infant deaths decreases life expectancy. If InfantDeaths^3 increases by one unit, the life expectancy increases by 0.000003453 years, assuming all other factors are constant. Although this term represents a non-linear relationship, the small positive coefficient suggests that while the impact of infant deaths accelerates as the number of deaths increases, the overall effect on life expectancy remains relatively minimal compared to the quadratic and linear terms.

AdultMortality: The negative coefficient for AdultMortality suggests that higher adult mortality decreases life expectancy. If AdultMortality increases by one unit, the life expectancy decreases by 0.06159 years, assuming all other factors are constant. This suggests that higher adult mortality rates are negatively associated with life expectancy, meaning as more adults die, the overall life expectancy in the population decreases.

AdultMortality^2: The positive coefficient for the square of AdultMortality indicates that as adult mortality increases, its impact on life expectancy accelerates. If AdultMortality^2 increases by one unit, life expectancy increases by 0.00002649 years, assuming all other factors are constant. This suggests that the relationship between adult mortality and life expectancy is non-linear. In other words, as adult mortality rises, its impact on life expectancy accelerates, but at a diminishing rate due to the small coefficient.

Hepatitis B: The coefficient 0.01159 suggests a positive relationship, meaning that higher levels of Hepatitis B increase life expectancy, If Hepatitis B increases by one unit, the interpretation based on the coefficient would be that life expectancy increases by 0.01159 years, assuming all other factors are constant which is counterintuitive. Hepatitis B typically reduces life expectancy, so this could be a model result artifact.

BMI: The negative coefficient for BMI (-0.1046) suggests that higher BMI (indicating obesity) decreases life expectancy. If BMI increases by one unit, the interpretation based on the coefficient -0.1046 would be

that life expectancy decreases by 0.1046 years, assuming all other factors are constant.

Polio: The positive coefficient (+0.01180) means that higher polio vaccination rates increase life expectancy. If Polio increases by one unit, life expectancy increases by 0.01180 years, assuming all other factors are constant. This suggests that higher rates of polio vaccination or the presence of polio control measures are positively associated with life expectancy, likely due to improved public health outcomes resulting from vaccination efforts.

IncidentsHIV: The negative coefficient (-0.03977) indicates that higher rates of HIV incidents are associated with lower life expectancy. If IncidentsHIV increases by one unit, life expectancy decreases by 0.03977 years, assuming all other factors are constant. This suggests that higher rates of HIV incidents are negatively associated with life expectancy.

GDP_per_capita: A positive coefficient here suggests that increases in GDP per capita are associated with higher life expectancy. If GDP_per_capita increases by one unit, the life expectancy increases by 0.000007030 years, assuming all other factors are constant. This suggests that higher GDP per capita, which typically indicates a stronger economy and better access to healthcare, is positively associated with life expectancy.

ThinnessTenNineteenYears: These coefficients show the relationship between malnutrition in certain age groups and life expectancy. For the terms -0.04017 * ThinnessTenNineteenYears a negative coefficient for ThinnessTenNineteenYears indicates that higher thinness rates in this age group (10-19 years) are associated with a decrease in life expectancy. If ThinnessTenNineteenYears increases by one unit, the life expectancy decreases by 0.04017 years, assuming all other factors are constant. This suggests that higher thinness (indicating malnutrition or poor health) in the 10-19 age group is negatively associated with life expectancy.

ThinnessFiveNineYears: These coefficients show the relationship between malnutrition in certain age groups and life expectancy. For the terms 0.04967 * ThinnessFiveNineYears A positive coefficient for ThinnessFiveNineYears suggests that higher thinness rates in the 5-9 age group are linked to an increase in life expectancy. If ThinnessFiveNineYears increases by one unit, the life expectancy increases by 0.04967 years, assuming all other factors are constant. This suggests that higher thinness rates (indicating malnutrition) in the 5-9 age group are positively associated with life expectancy, which could be counterintuitive.

Schooling: More years of schooling are associated with higher life expectancy, as expected from the positive coefficient (0.1097). If Schooling increases by one unit, the life expectancy increases by 0.1097 years, assuming all other factors are constant. This suggests that more years of schooling are positively associated with life expectancy

EconomyStatusDeveloped: This binary variable indicates whether a country is considered developed. If developed, life expectancy is increased by 2.084 years. If EconomyStatusDeveloped increases by one unit, the life expectancy increases by 2.084 years, assuming all other factors are constant. This suggests that being in a developed country, characterized by higher economic status, is positively associated with life expectancy.

Lets derive some sub-models:

1. For Region = Asia, Year = 2002 and developed countries Life expectancy,

$$Lifeexpectancy = 86.59243 - 0.1189 * InfantDeaths - 0.0003900 * InfantDeaths^2 + 0.000003453 * InfantDeaths^3 - 0.06159 * AdultMortality + 0.00002649 * AdultMortality^2 + 0.01159 * HepatitisB - 0.1046 * BMI + 0.01180 * Polio - 0.03977 * IncidentsHIV + 0.000007030 * GDP_per_capita - 0.04017 * ThinnessTenNineteenYears + 0.04967 * ThinnessFiveNineYears + 0.1097 * Schooling$$

2. For Region = North America, Year = 2010 and developed countries Life expectancy,

$$Lifeexpectancy = 86.9409 - 0.1189 * InfantDeaths - 0.0003900 * InfantDeaths^2 + 0.000003453 * InfantDeaths^3 - 0.06159 * AdultMortality + 0.00002649 * AdultMortality^2 + 0.01159 * HepatitisB - 0.1046 * BMI + 0.01180 * Polio - 0.03977 * IncidentsHIV + 0.000007030 * GDP_per_capita - 0.04017 * ThinnessTenNineteenYears + 0.04967 * ThinnessFiveNineYears + 0.1097 * Schooling$$

3. For Region = Africa, Year = 2007 and developing countries Life expectancy,

$\widehat{Life_expectancy} = 83.9348 - 0.1189 * InfantDeaths - 0.0003900 * InfantDeaths^2 + 0.000003453 * InfantDeaths^3 - 0.06159 * AdultMortality + 0.00002649 * AdultMortality^2 + 0.01159 * HepatitisB - 0.1046 * BMI + 0.01180 * Polio - 0.03977 * IncidentsHIV + 0.000007030 * GDP_per_capita - 0.04017 * ThinnessTenNineteenYears + 0.04967 * ThinnessFiveNineYears + 0.1097 * Schooling$

Prediction:

Based on the model, we observe that economic status and GDP per capita have higher coefficients, indicating a strong influence on life expectancy. Using the model's predict() function, we can quantify the impact of these predictors and analyze how changes in their values affect life expectancy.

The model predicts that Canada, with its developed economy, strong healthcare system, and higher economic indicators, has a significantly higher life expectancy of 81.35 years, compared to Mexico, which has a life expectancy of 73.75 years. Mexico's status as a developing country, with greater healthcare challenges and lower economic performance, contributes to this disparity.

This analysis underscores the critical role of economic development, healthcare access, and education in shaping a country's overall life expectancy. Specifically, Canada's life expectancy is approximately 7.6 years higher than Mexico's, highlighting the importance of addressing systemic disparities to improve global health outcomes.

4 CONCLUSION AND DISCUSSION

4.1 Approach

The approach used in this analysis demonstrates a systematic and effective framework for understanding the determinants of life expectancy globally. The key steps, from initial data cleaning to higher-order model refinement, ensured both the robustness and interpretability of the findings.

4.1.1 Effectiveness of the Approach:

- **Data Simplification:** The decision to use the Region variable instead of the Country variable reduced the complexity of handling 193 unique country entries while still allowing for meaningful regional insights.
- **Multicollinearity Resolution:** The rigorous application of VIF to identify and address multicollinearity enhanced the model's stability and interpretability. Removing highly collinear variables, such as Under_five_deaths and Diphtheria, allowed for a more robust model.
- **Model Refinement:** The use of stepwise regression and evaluation criteria like AIC, BIC, and Mallows' Cp ensured the selection of the most parsimonious model.
- **Higher-order and Interaction Models:** These advanced modeling techniques captured non-linear relationships and synergies between predictors, providing a more nuanced understanding of the factors influencing life expectancy.
- **Validation of Residual Assumptions:** The adherence to key statistical assumptions, including normality and homoscedasticity, further bolstered the reliability of the findings.

4.1.2 Challenges and Trade-offs:

- The exclusion of country-specific details, while necessary to simplify the analysis, limited the ability to identify unique country-level characteristics national trends or outliers.
- The refinement process, especially testing interaction and higher-order terms, was resource-intensive and required careful attention to ensure statistical validity.
- Despite these challenges, the rigorous validation of assumptions (normality, homoscedasticity, and independence) strengthened the model's reliability.

4.2 Future Work

1. Granular Data Analysis:

- Include additional country-specific variables, such as healthcare expenditure, literacy rates, or environmental indices, to uncover country-level insights.
- Extend the dataset to include more recent years or incorporate data from specific events, such as the COVID-19 pandemic, to analyze temporal changes in life expectancy determinants.

2. Advanced Modeling Techniques:

- Explore machine learning approaches, such as Random Forest, XGBoost, or Neural Networks, to handle interactions and non-linear relationships more efficiently.

2. Policy Simulation and Impact Assessment:

- Use the refined models to simulate the impact of targeted interventions, such as reducing infant mortality or increasing vaccination rates, on life expectancy.
- Compare life expectancy determinants between developed and developing regions to identify specific strategies that could improve outcomes in each context.

3. Interactive Reporting and Visualization:

- Develop interactive dashboards using tools like R Shiny to visualize results and allow stakeholders to explore data dynamically.
- Create regional or country-specific visualizations to highlight disparities and suggest actionable recommendations.

4. Validation and Generalization:

- Validate the model using external datasets to assess its robustness and generalizability.

By addressing these areas, future studies can deepen the understanding of life expectancy determinants and further enhance the impact of this work on global health policies and interventions.

5 REFERENCES

[1] L. Gochia, "Life expectancy (WHO) fixed dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated. [Accessed: Nov, 30, 2024].

6 APPENDIX

6.1 Images

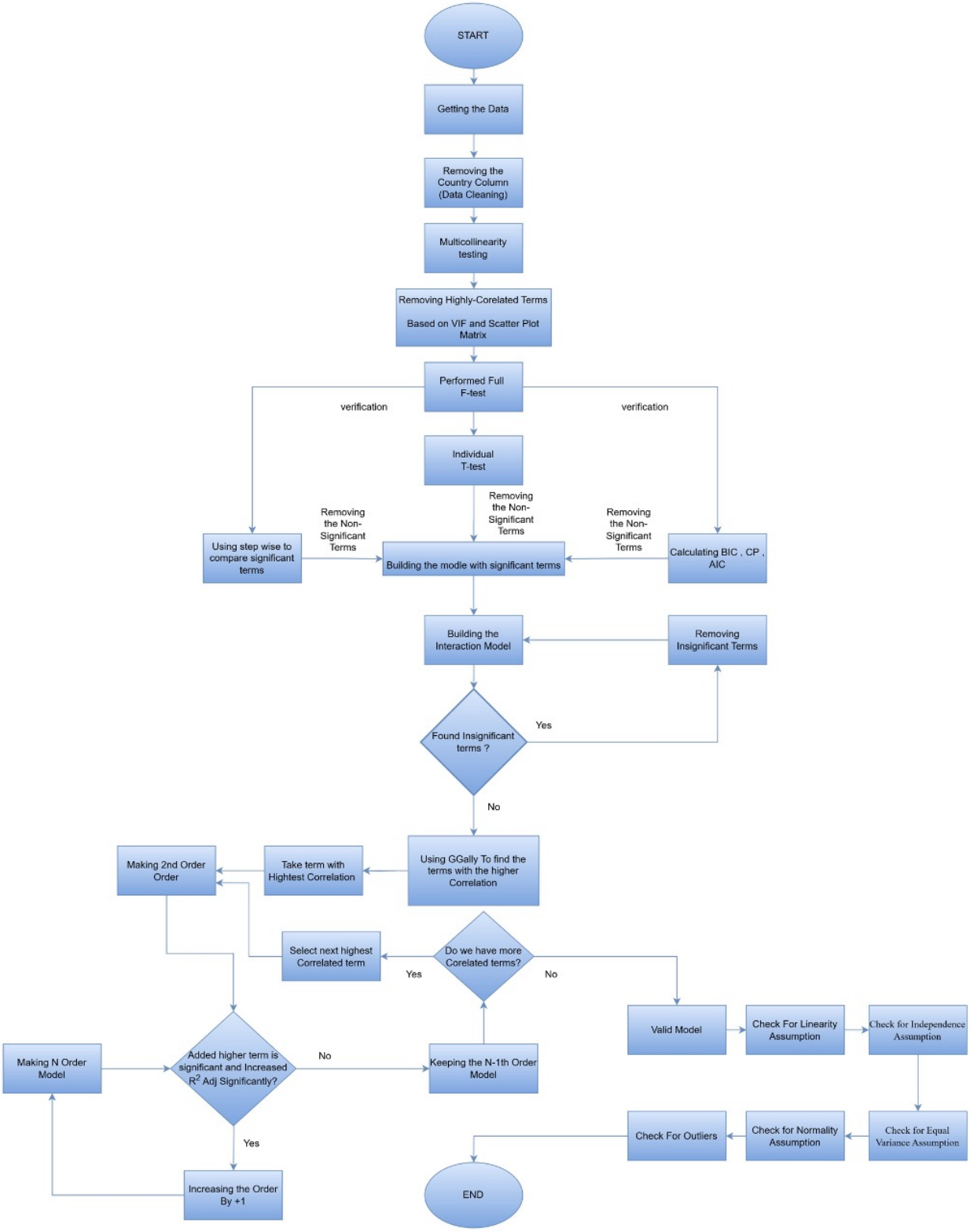


*Figure*2.1*WorkFlowDiagram*

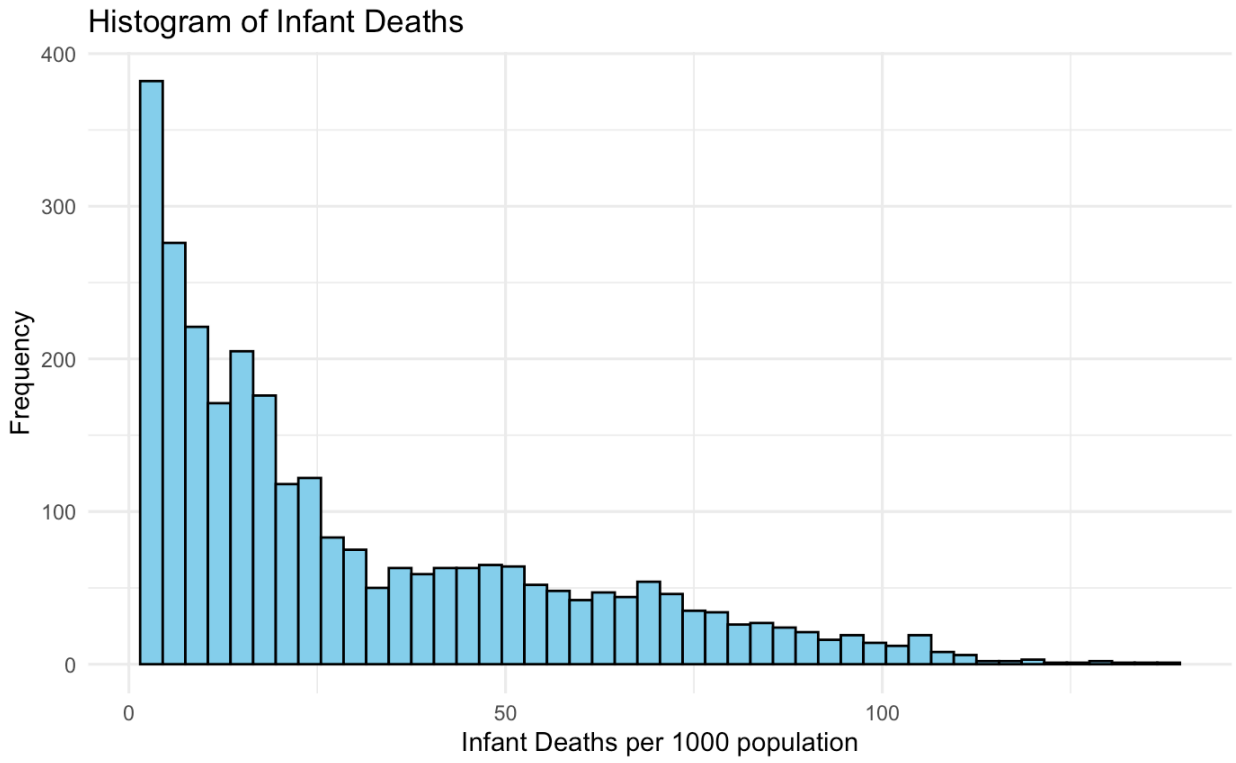


Figure 2.2 Infant vs Frequancy

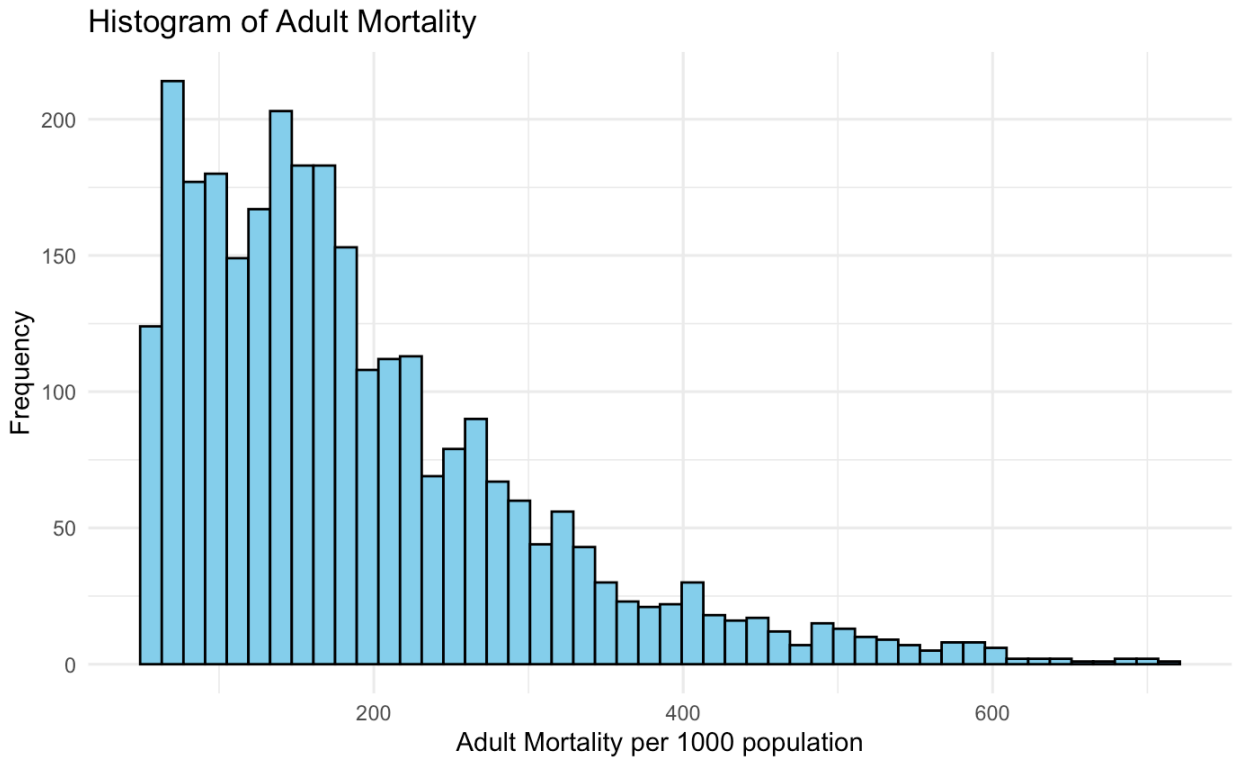


Figure 2.3 Adult Mortality vs Frequency

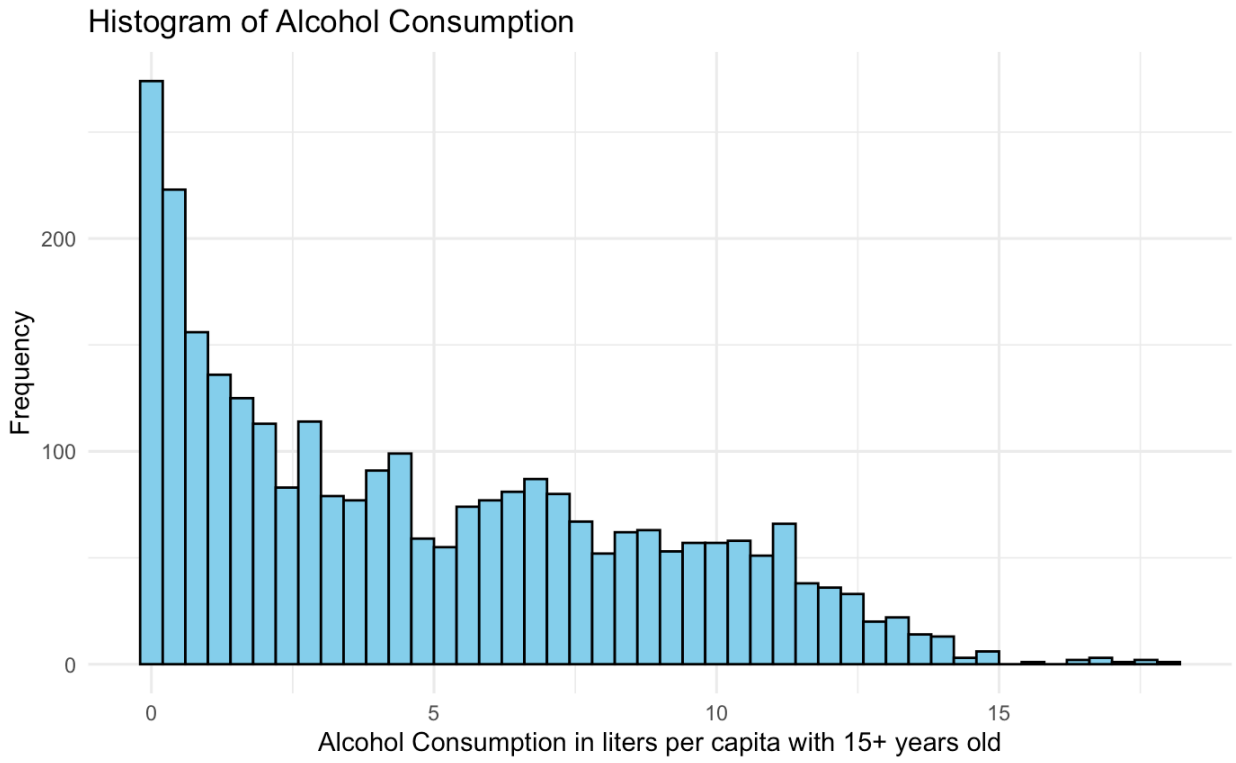


Figure 2.4 Alcohol consumptions Vs Frequancy

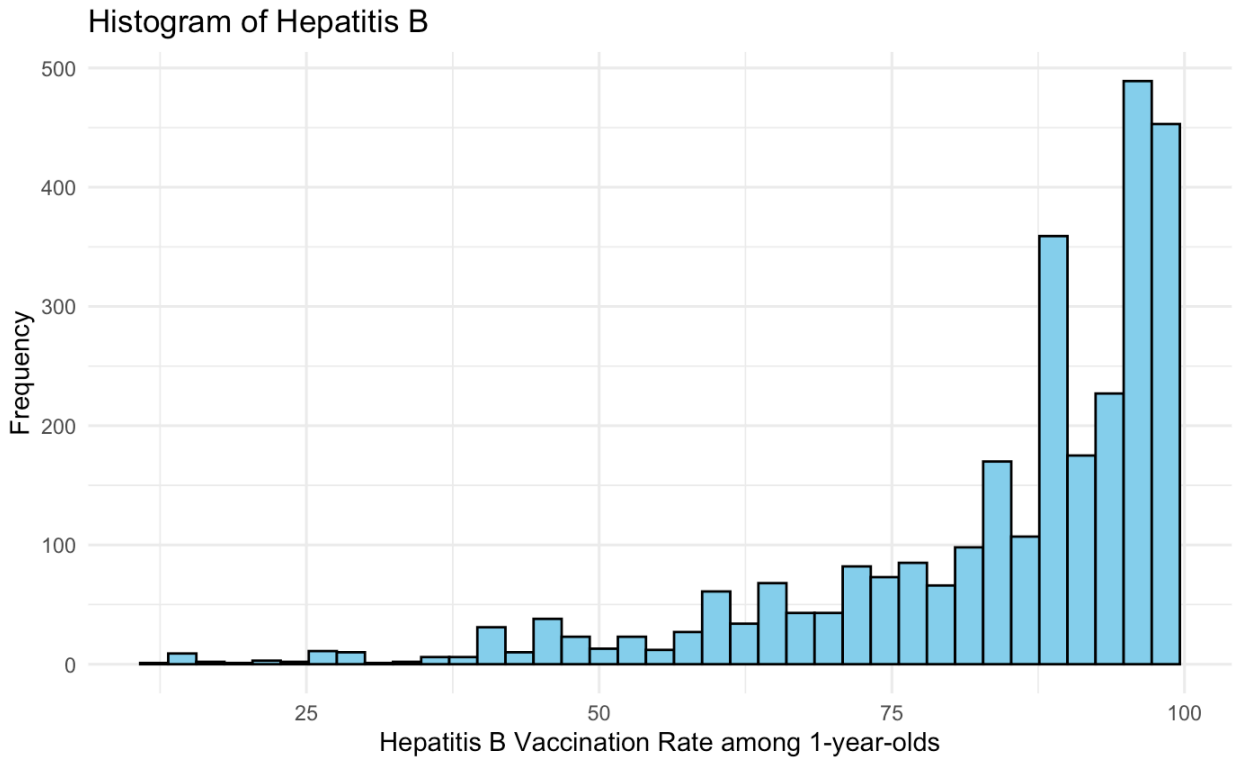


Figure 2.5 Hepatitis B vs Frequancy

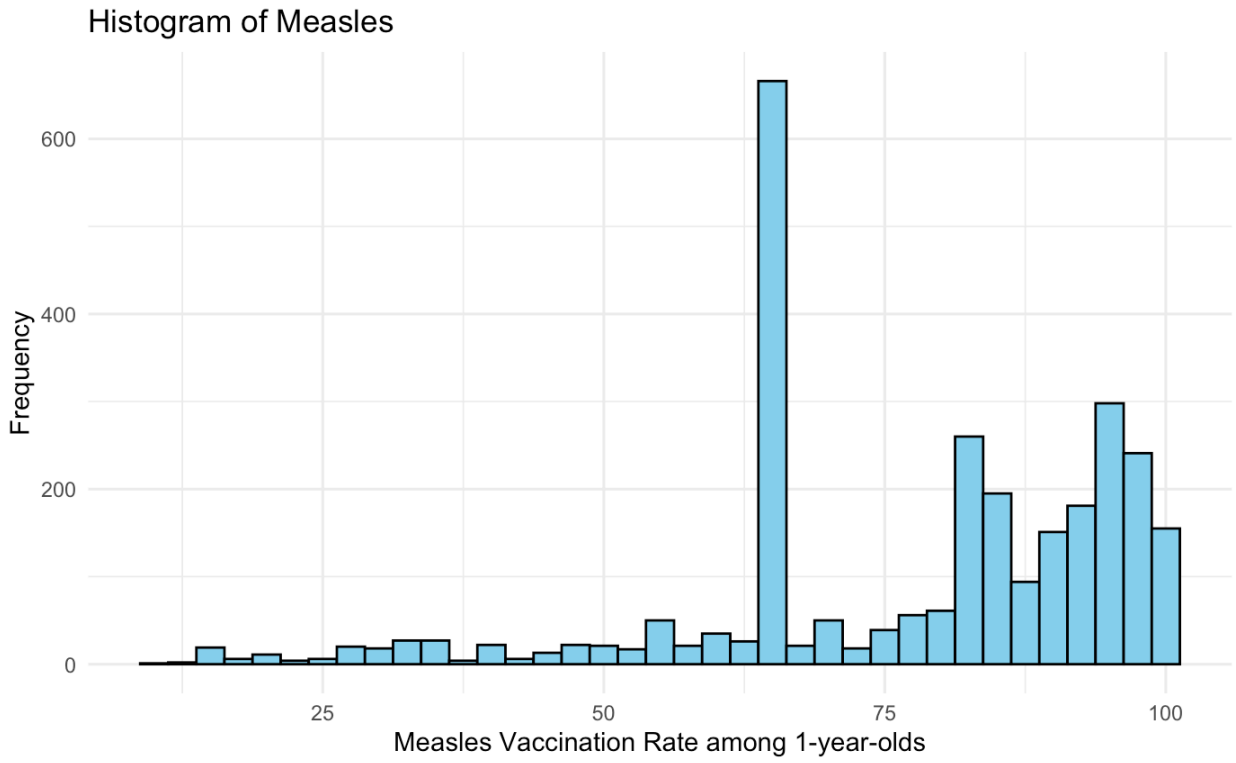


Figure 2.6 Measles vs Frequancy

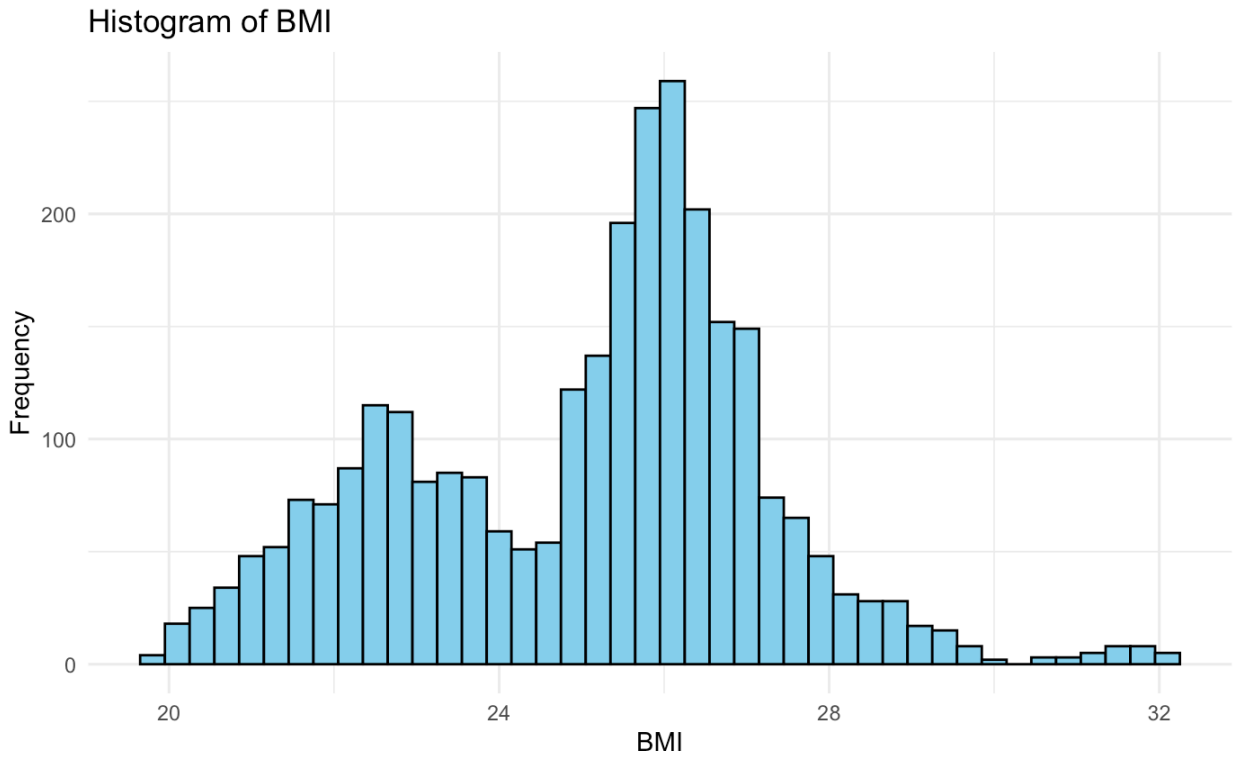


Figure 2.7 BMI vs Frequancy

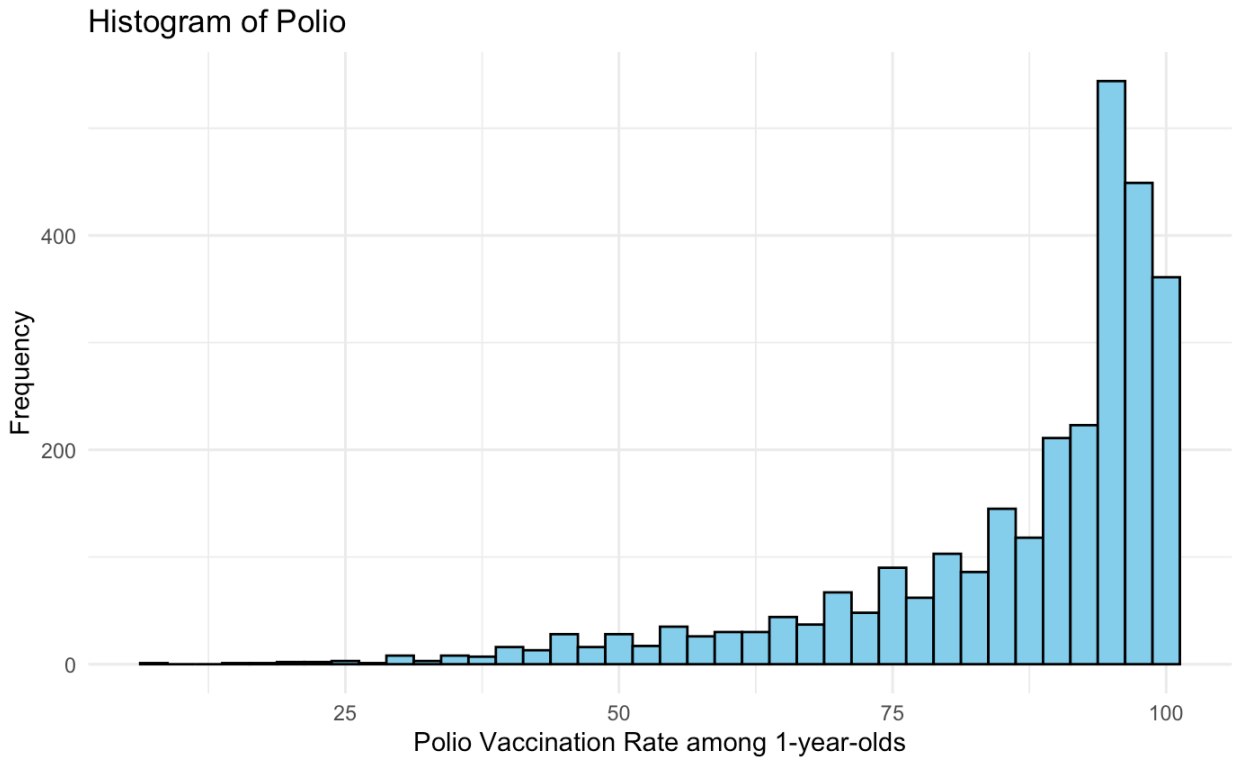


Figure 2.8 Polio vs Frequancy

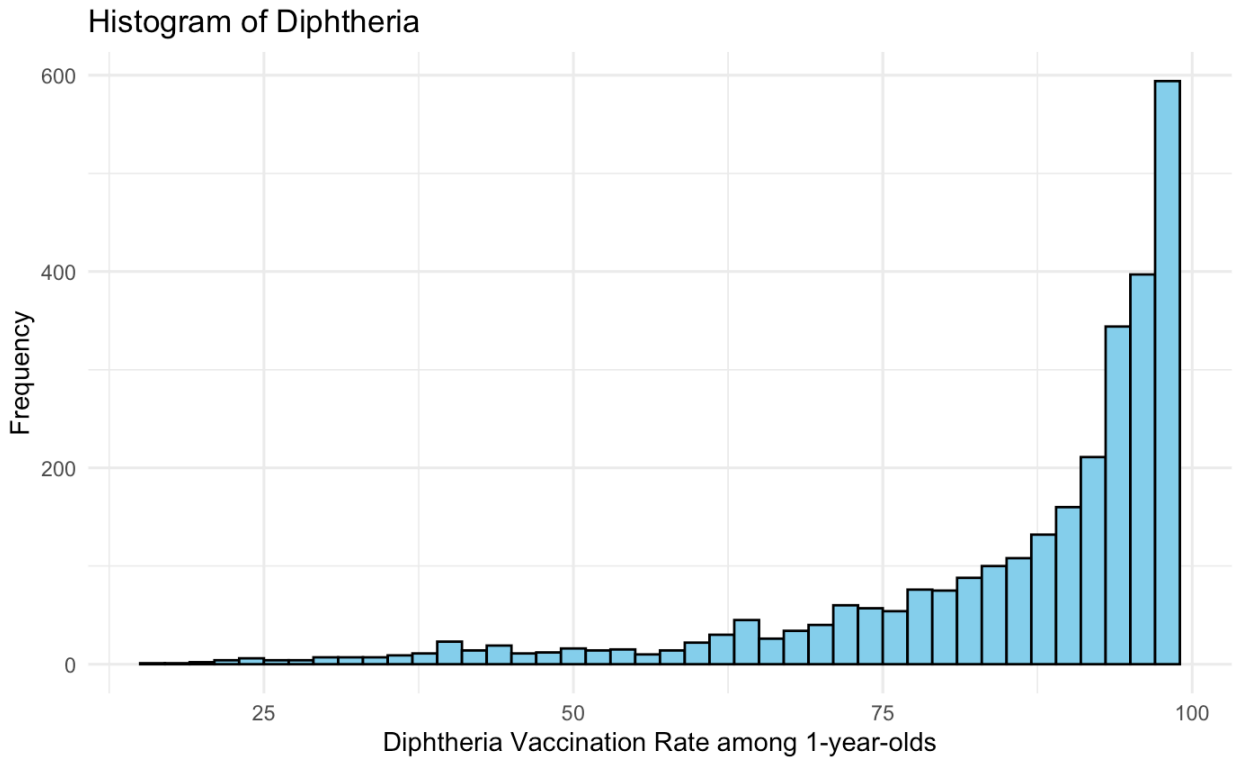
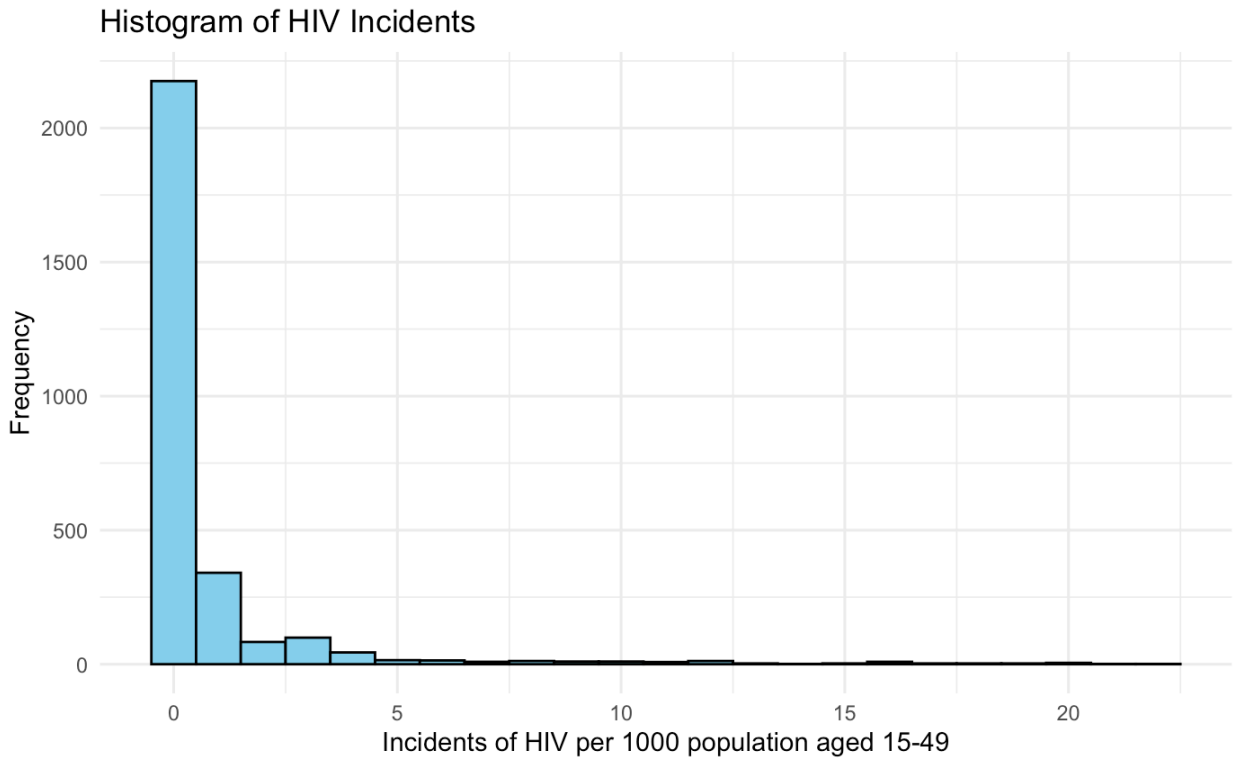


Figure 2.9 Diphtheria vs Frequancy



$Figure 2.10 HIV vs Frequacy$

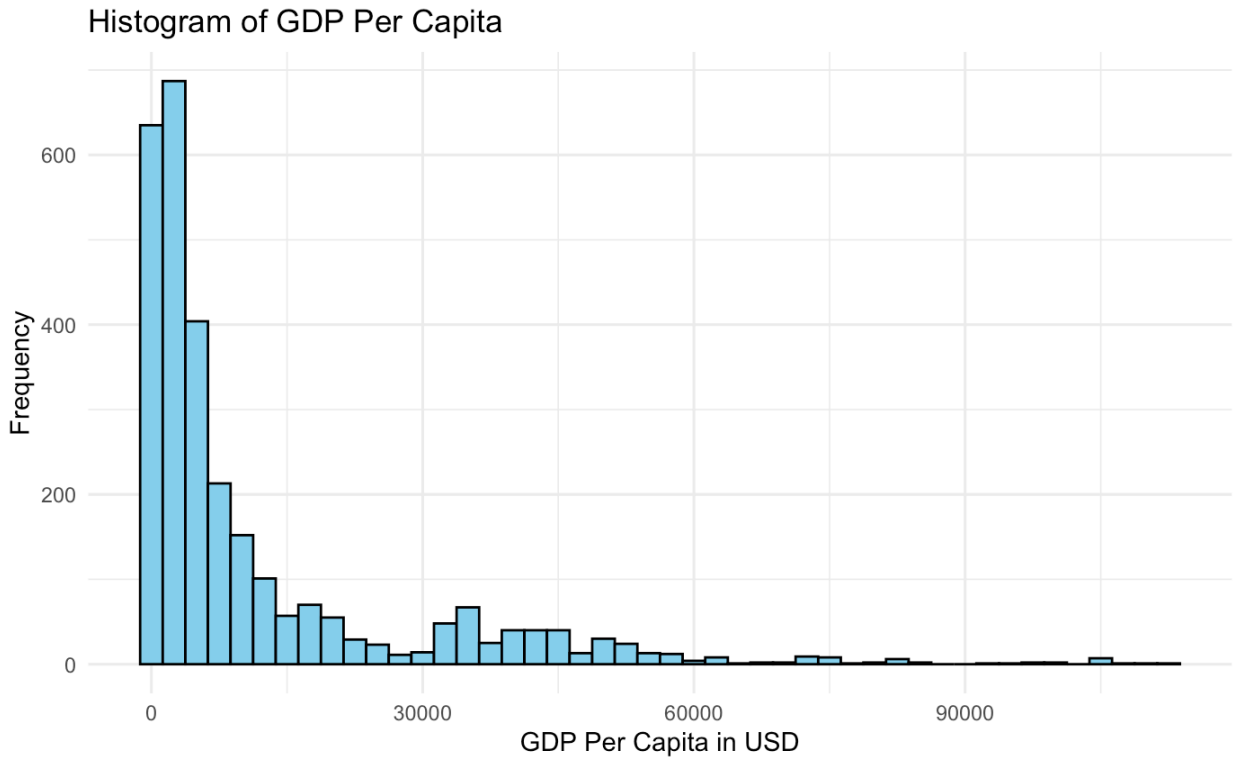


Figure 2.11 *GDP vs Frequancy*

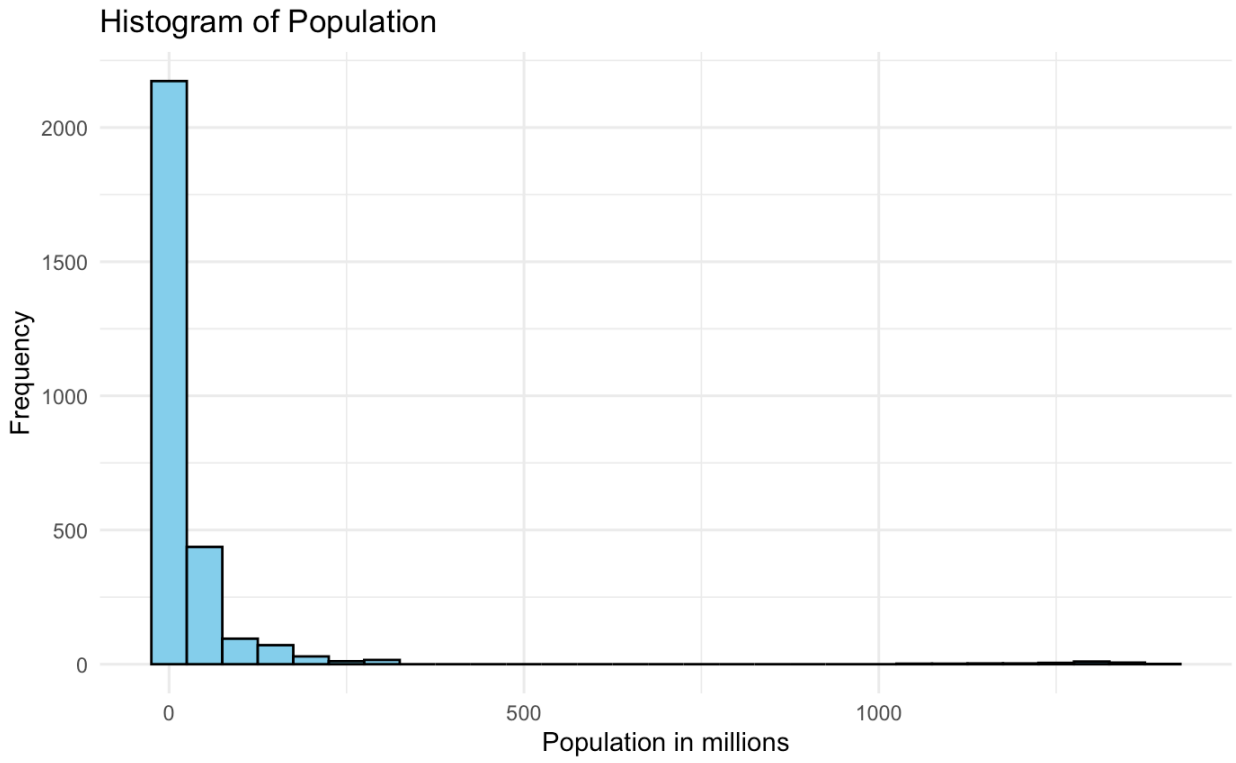


Figure 2.12 *Population vs Frequncy*

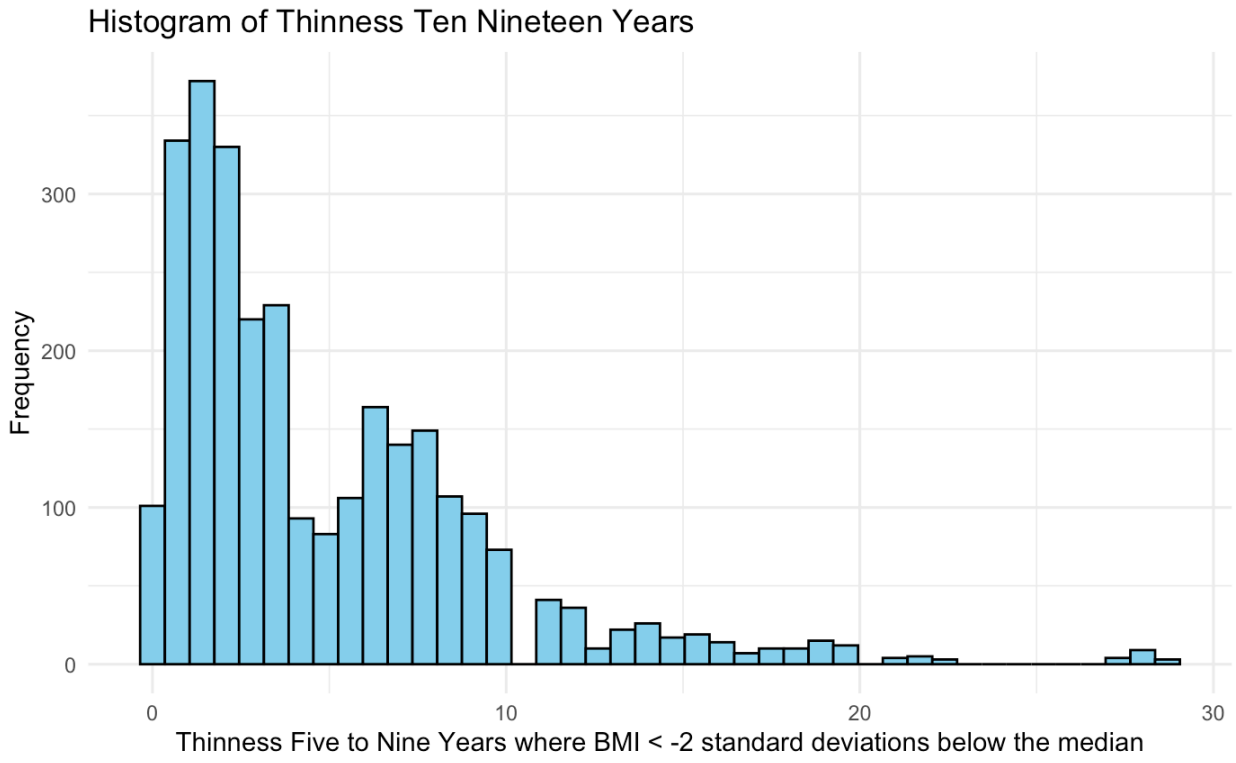


Figure 2.13 *ThinnessFiveninevsFrequancy*

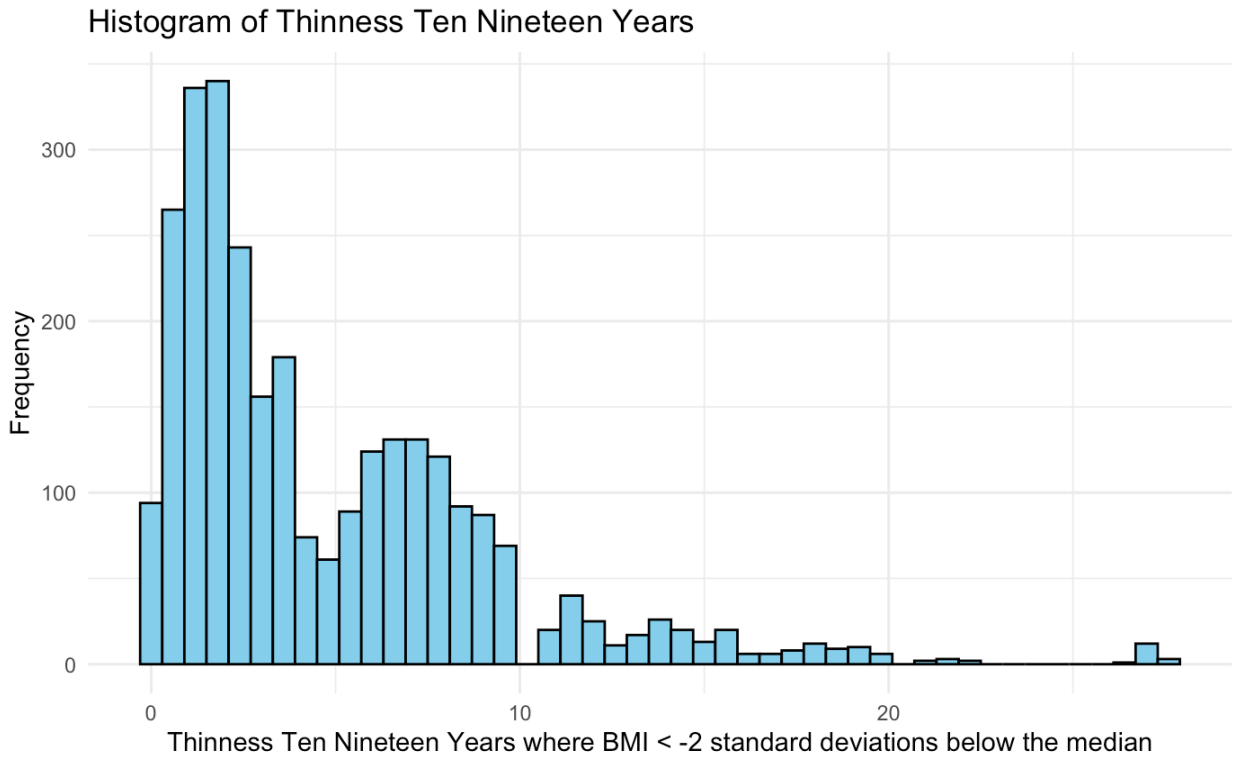
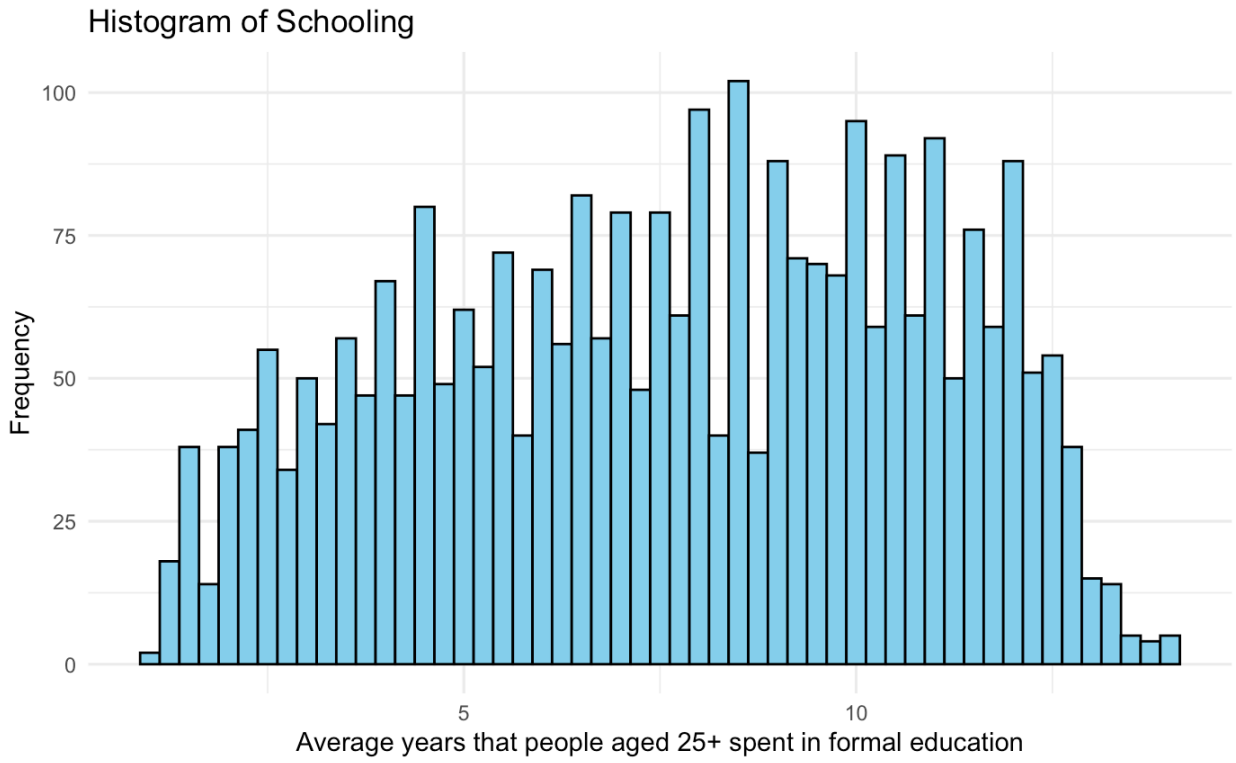


Figure 2.14 *Thinnesstennineteenvs Frequancy*



$Figure 2.15 schoolingvsFrequancy$

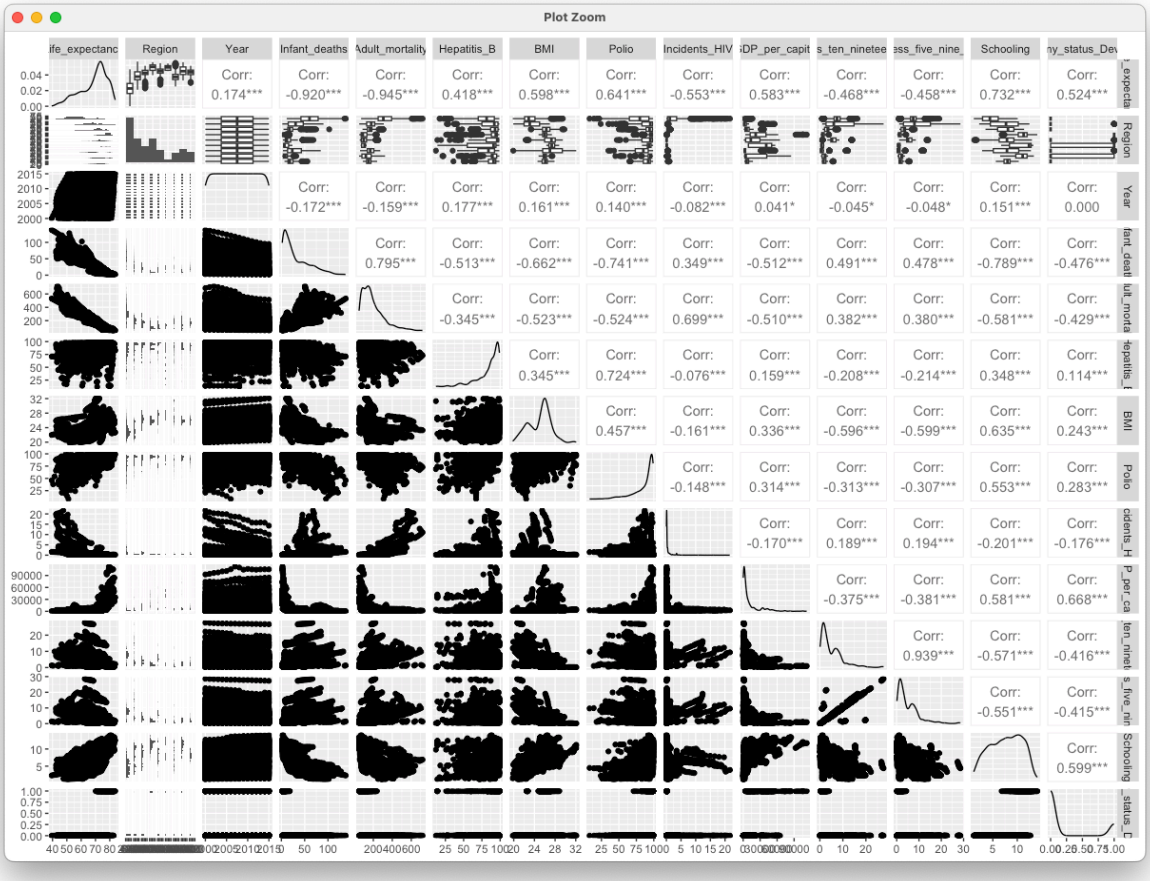


Figure 3.1 Correlation Matrix

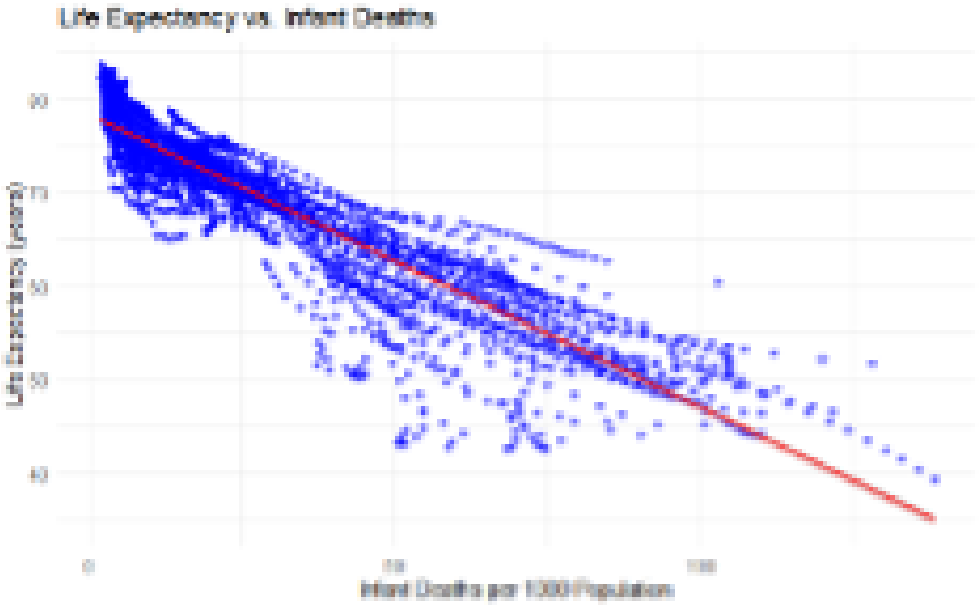


Figure 3.2 EDA Infants Death

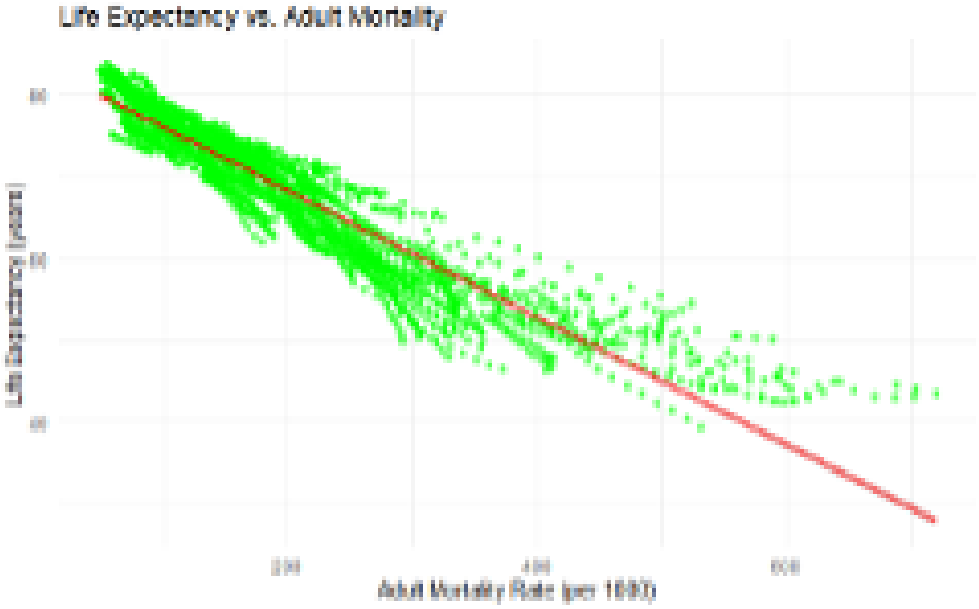


Figure 3.3 EDA Adult Mortality Rate

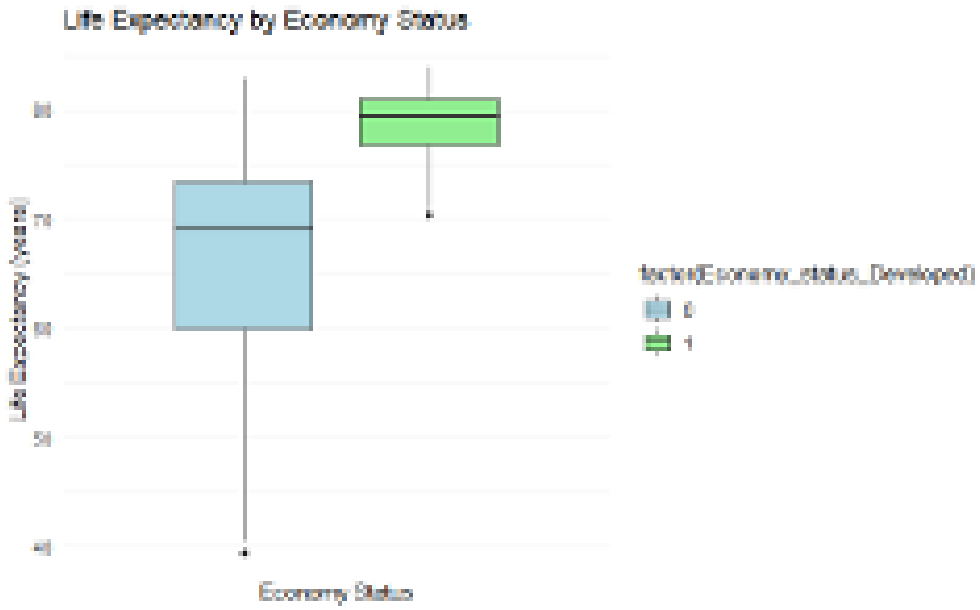


Figure 3.4 EDA Economy Status

VIF Multicollinearity Diagnostics		
	VIF	detection
RegionAsia	2.6230	0
RegionCentral America and Caribbean	2.2181	0
RegionEuropean Union	6.6204	0
RegionMiddle East	2.2448	0
RegionNorth America	1.5346	0
RegionOceania	1.9782	0
RegionRest of Europe	2.4437	0
RegionSouth America	1.9035	0
YearY2001	1.8761	0
YearY2002	1.8801	0
YearY2003	1.8815	0
YearY2004	1.8852	0
YearY2005	1.8911	0
YearY2006	1.8967	0
YearY2007	1.9024	0
YearY2008	1.9066	0
YearY2009	1.9137	0
YearY2010	1.9186	0
YearY2011	1.9262	0
YearY2012	1.9340	0
YearY2013	1.9437	0
YearY2014	1.9519	0
YearY2015	1.9588	0
Infant_deaths	48.2026	1
Under_five_deaths	50.0525	1
Adult_mortality	8.1421	0
Hepatitis_B	2.7445	0
Alcohol_consumption	3.3884	0
Measles	1.6634	0
BMI	3.9027	0
Polio	12.2016	1
Diphtheria	13.2644	1
Incidents_HIV	3.0411	0
GDP_per_capita	2.6502	0
Population_mln	1.2131	0
Thinness_ten_nineteen_years	9.0115	0
Thinness_five_nine_years	9.2297	0
Schooling	5.5301	0
Economy_status_DevelopedYes	7.2835	0

Figure 3.5 Multicollinearity for Full Model

```
 VIF Multicollinearity Diagnostics

                                    VIF detection
RegionAsia                        2.4534         0
RegionCentral America and Caribbean 2.1513       0
RegionEuropean Union              6.5419         0
RegionMiddle East                 2.2283         0
RegionNorth America               1.5147         0
RegionOceania                     1.9019         0
RegionRest of Europe              2.4140         0
RegionSouth America               1.8480         0
YearY2001                         1.8761         0
YearY2002                         1.8794         0
YearY2003                         1.8813         0
YearY2004                         1.8851         0
YearY2005                         1.8909         0
YearY2006                         1.8967         0
YearY2007                         1.9020         0
YearY2008                         1.9064         0
YearY2009                         1.9135         0
YearY2010                         1.9185         0
YearY2011                         1.9251         0
YearY2012                         1.9334         0
YearY2013                         1.9430         0
YearY2014                         1.9510         0
YearY2015                         1.9580         0
Infant_deaths                     8.6365         0
Adult_mortality                   8.0123         0
Hepatitis_B                       2.3251         0
Alcohol_consumption               3.3720         0
Measles                           1.6604         0
BMI                               3.8322         0
Polio                             3.7884         0
Incidents_HIV                     3.0380         0
GDP_per_capita                    2.6070         0
Population_mln                    1.2126         0
Thinness_ten_nineteen_years       8.9921         0
Thinness_five_nine_years          9.2034         0
Schooling                         5.4809         0
Economy_status_DevelopedYes       7.2274         0
```

Figure 3.6 Multicollinearity for Revised Model

```
Analysis of Variance Table

Model 1: Life_expectancy ~ 1
Model 2: Life_expectancy ~ Region + Year + Infant_deaths + Adult_mortality +
    Hepatitis_B + Alcohol_consumption + Measles + BMI + Polio +
    Incidents_HIV + GDP_per_capita + Population_mln + Thinness_ten_nineteen_years +
    Thinness_five_nine_years + Schooling + Economy_status_Developed
  Res.Df    RSS Df Sum of Sq      F               Pr(>F)
1   2863 253277
2   2826   4336 37    248940 4384.9 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.7 Anova Table

```
Coefficients:
                                  Estimate   Std. Error t value             Pr(>|t|)
(Intercept)                     82.104978482  0.592335707 138.612  < 0.0000000000000002 ***
RegionAsia                       0.743246105  0.101301040   7.337   0.000000000000284 ***
RegionCentral America and Caribbean  2.209487890  0.110215061  20.047  < 0.0000000000000002 ***
RegionEuropean Union            -0.408489184  0.165417105  -2.469             0.013591 *
RegionMiddle East                0.417905797  0.128681655   3.248             0.001177 **
RegionNorth America              0.839824436  0.221909059   3.785             0.000157 ***
RegionOceania                   -0.446217763  0.132915975  -3.357             0.000798 ***
RegionRest of Europe             0.477875335  0.129788445   3.682             0.000236 ***
RegionSouth America              2.030134268  0.125816860  16.136  < 0.0000000000000002 ***
YearY2001                        0.076533155  0.130973777   0.584             0.559039
YearY2002                        0.091212838  0.131089634   0.696             0.486608
YearY2003                        0.013439055  0.131154123   0.102             0.918393
YearY2004                        0.042126997  0.131286565   0.321             0.748326
YearY2005                        0.001204410  0.131489300   0.009             0.992692
YearY2006                        0.024415205  0.131688382   0.185             0.853928
YearY2007                        0.057544475  0.131874824   0.436             0.662611
YearY2008                        0.113774122  0.132025995   0.862             0.388895
YearY2009                        0.173733656  0.132270164   1.313             0.189129
YearY2010                        0.249156780  0.132443264   1.881             0.060043 .
YearY2011                        0.298265159  0.132672946   2.248             0.024645 *
YearY2012                        0.325882484  0.132957816   2.451             0.014305 *
YearY2013                        0.443183971  0.133286420   3.325             0.000895 ***
YearY2014                        0.541712333  0.133560220   4.056    0.000051283649973 ***
YearY2015                        0.563851933  0.133799385   4.214    0.000025856670242 ***
Infant_deaths                   -0.126142209  0.002470528 -51.059  < 0.0000000000000002 ***
Adult_mortality                 -0.047582684  0.000570263 -83.440  < 0.0000000000000002 ***
Hepatitis_B                     -0.011618685  0.002206868  -5.265    0.000000150882495 ***
Alcohol_consumption             -0.013998097  0.010675858  -1.311             0.189900
Measles                          0.001778407  0.001598650   1.112             0.266042
BMI                             -0.104259993  0.020656788  -5.047    0.000000476636691 ***
Polio                            0.010454329  0.002987948   3.499             0.000475 ***
Incidents_HIV                    0.096399867  0.016944166   5.689    0.000000014066735 ***
GDP_per_capita                   0.000016317  0.000002207   7.393    0.000000000000188 ***
Population_mln                  -0.000265264  0.000186777  -1.420             0.155654
Thinness_ten_nineteen_years     -0.036657176  0.015641383  -2.344             0.019167 *
Thinness_five_nine_years         0.033840966  0.015519944   2.180             0.029304 *
Schooling                        0.123170779  0.017088638   7.208    0.000000000000727 ***
Economy_status_DevelopedYes      2.335417458  0.153665907  15.198  < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.239 on 2826 degrees of freedom
Multiple R-squared:  0.9829,    Adjusted R-squared:  0.9827
F-statistic:  4385 on 37 and 2826 DF,  p-value: < 0.00000000000000022
```

Figure 3.8 Model based on Full F − Test Output

```
Coefficients:
                                  Estimate   Std. Error t value             Pr(>|t|)
(Intercept)                     82.066222240  0.587989314 139.571  < 0.0000000000000002 ***
RegionAsia                       0.734062617  0.099347408   7.389    0.000000000000194 ***
RegionCentral America and Caribbean  2.177344688  0.107674866  20.221  < 0.0000000000000002 ***
RegionEuropean Union            -0.455863848  0.160787845  -2.835             0.004612 **
RegionMiddle East                0.455190549  0.126107923   3.610             0.000312 ***
RegionNorth America              0.793349667  0.220032322   3.606             0.000317 ***
RegionOceania                   -0.452897448  0.132076369  -3.429             0.000614 ***
RegionRest of Europe             0.450803583  0.127535967   3.535             0.000415 ***
RegionSouth America              1.972877201  0.122508164  16.104  < 0.0000000000000002 ***
YearY2001                        0.076320202  0.131025854   0.582             0.560288
YearY2002                        0.091497598  0.131139484   0.698             0.485415
YearY2003                        0.013558460  0.131196761   0.103             0.917697
YearY2004                        0.042216826  0.131325032   0.321             0.747879
YearY2005                       -0.000861437  0.131520171  -0.007             0.994774
YearY2006                        0.023134335  0.131735331   0.176             0.860611
YearY2007                        0.053914612  0.131914435   0.409             0.682784
YearY2008                        0.109759593  0.132052920   0.831             0.405943
YearY2009                        0.171316429  0.132264166   1.295             0.195336
YearY2010                        0.246157554  0.132417236   1.859             0.063137 .
YearY2011                        0.296414618  0.132674818   2.234             0.025551 *
YearY2012                        0.322993856  0.132940109   2.430             0.015177 *
YearY2013                        0.441606717  0.133255710   3.314             0.000931 ***
YearY2014                        0.542028872  0.133535733   4.059    0.000050607058674 ***
YearY2015                        0.565357295  0.133768060   4.226    0.000024497685880 ***
Infant_deaths                   -0.125707354  0.002460972 -51.080  < 0.0000000000000002 ***
Adult_mortality                 -0.047707246  0.000558212 -85.464  < 0.0000000000000002 ***
Hepatitis_B                     -0.011201606  0.002186708  -5.123    0.000000321615890 ***
BMI                             -0.100427770  0.020597420  -4.876    0.000001143881240 ***
Polio                            0.011082724  0.002966152   3.736             0.000190 ***
Incidents_HIV                    0.096466797  0.016924461   5.700    0.000000013229075 ***
GDP_per_capita                   0.000016251  0.000002201   7.382    0.000000000000204 ***
Thinness_ten_nineteen_years     -0.036282729  0.015544778  -2.334             0.019661 *
Thinness_five_nine_years         0.030349029  0.015390212   1.972             0.048710 *
Schooling                        0.118805537  0.016434978   7.229    0.000000000000624 ***
Economy_status_DevelopedYes      2.293441075  0.151476407  15.141  < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.239 on 2829 degrees of freedom
Multiple R-squared:  0.9828,    Adjusted R-squared:  0.9826
F-statistic:  4768 on 34 and 2829 DF,  p-value: < 0.00000000000000022
```

Figure 3.9 Model After performing Individual t − test and Removing Insignificant terms

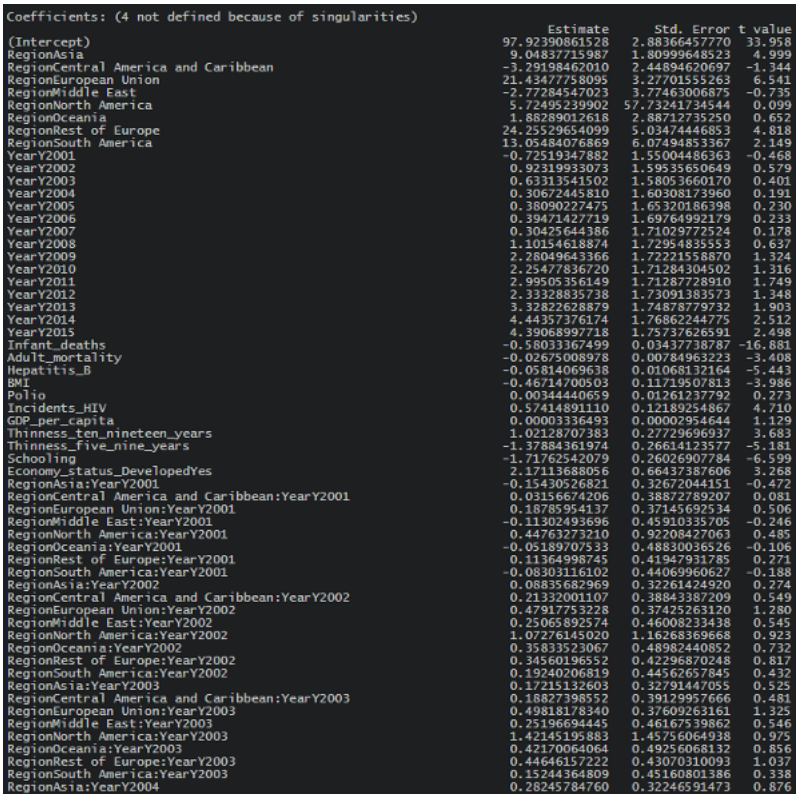
```
Coefficients:
                                  Estimate   Std. Error t value             Pr(>|t|)
(Intercept)                     81.966649189  0.545806962 150.175 < 0.0000000000000002 ***
Adult_mortality                 -0.047681271  0.000558089 -85.437 < 0.0000000000000002 ***
Infant_deaths                   -0.125821225  0.002459639 -51.154 < 0.0000000000000002 ***
Economy_status_DevelopedYes      2.294812967  0.151054334  15.192 < 0.0000000000000002 ***
GDP_per_capita                   0.000016173  0.000002198   7.356   0.000000000002464 ***
RegionAsia                       0.730038623  0.095710838   7.628   0.000000000000325 ***
RegionCentral America and Caribbean  2.183309642  0.107658774  20.280 < 0.0000000000000002 ***
RegionEuropean Union            -0.449540879  0.160839436  -2.795              0.005225 **
RegionMiddle East                0.443489301  0.124399945   3.565              0.000370 ***
RegionNorth America              0.795967204  0.220344618   3.613              0.000305 ***
RegionOceania                   -0.438750738  0.131908712  -3.326              0.000892 ***
RegionRest of Europe             0.455491023  0.127556544   3.571              0.000362 ***
RegionSouth America              1.983318339  0.121751435  16.290 < 0.0000000000000002 ***
Schooling                        0.122563547  0.016161342   7.584   0.000000000000452 ***
Incidents_HIV                    0.095846620  0.016758409   5.719   0.000000011813653 2 ***
BMI                             -0.098181094  0.019743029  -4.973   0.000000698778589 5 ***
Hepatitis_B                     -0.011327512  0.002186768  -5.180   0.000002374270972 ***
Polio                            0.011074379  0.002967132   3.732              0.000193 ***
YearY2001                        0.076832106  0.131104131   0.586              0.557896
YearY2002                        0.085618731  0.131194098   0.653              0.514060
YearY2003                        0.013191022  0.131274729   0.100              0.919967
YearY2004                        0.041028789  0.131394131   0.312              0.754868
YearY2005                       -0.003393630  0.131543086  -0.026              0.979420
YearY2006                        0.018341541  0.131732616   0.139              0.889276
YearY2007                        0.046028884  0.131904883   0.349              0.727149
YearY2008                        0.104907586  0.132055475   0.794              0.427017
YearY2009                        0.163549129  0.132271832   1.236              0.216389
YearY2010                        0.240217085  0.132441765   1.814              0.069821 .
YearY2011                        0.290065647  0.132694095   2.186              0.028899 *
YearY2012                        0.317558071  0.132943962   2.389              0.016975 *
YearY2013                        0.433527440  0.133221179   3.254              0.001151 **
YearY2014                        0.539750374  0.133488173   4.043   0.000054080000735 ***
YearY2015                        0.559486474  0.133711955   4.184   0.000029479518371 4 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.24 on 2831 degrees of freedom
Multiple R-squared:  0.9828,    Adjusted R-squared:  0.9826
F-statistic:  5059 on 32 and 2831 DF,  p-value: < 0.00000000000000022
```

*Figure*3.10*Resultof Step − wiseModelSelection*

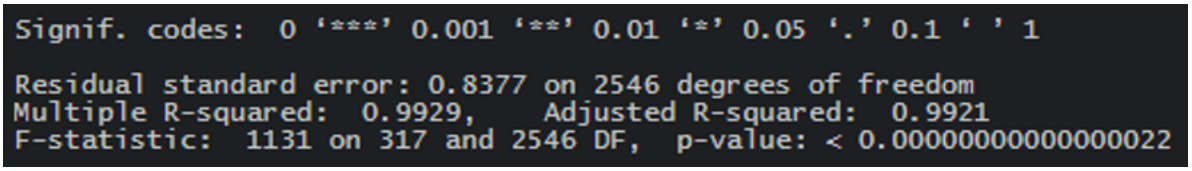
```
        rsquare AdjustedR          cp       AIC        BIC
 [1,] 0.8937062 0.8936691 14671.03495 14551.086  -6403.875
 [2,] 0.9710178 0.9709976  1922.03201 10831.245 -10117.756
 [3,] 0.9751387 0.9751126  1244.38250 10287.362 -10549.039
 [4,] 0.9772792 0.9772474   893.33732  9619.966 -10798.939
 [5,] 0.9787456 0.9787084   653.49103  9530.695 -10982.050
 [6,] 0.9797976 0.9797552   481.97571  9500.560 -11119.482
 [7,] 0.9807825 0.9807354   321.54992  9464.634 -11254.652
 [8,] 0.9812176 0.9811650   251.77972  9443.244 -11312.289
 [9,] 0.9816077 0.9815497   189.44460  9418.419 -11364.436
[10,] 0.9819913 0.9819281   128.18101  9406.586 -11416.837
[11,] 0.9821105 0.9820415   110.50941  9394.528 -11427.909
[12,] 0.9822325 0.9821577    92.39530  9394.943 -11439.540
[13,] 0.9823265 0.9822459    78.88395  9393.009 -11446.780
```

*Figure*3.24*All − Possible − RegressionsSelectionProcedure*



```
Coefficients: (4 not defined because of singularities)
                                                  Estimate      Std. Error t value
(Intercept)                                   97.92390861528  2.88366457770  33.958
RegionAsia                                     9.04837715987  1.80999648523   4.999
RegionCentral America and Caribbean           -3.29198462010  2.44894620697  -1.344
RegionEuropean Union                          21.43477758095  3.27701555263   6.541
RegionMiddle East                             -2.77284547023  3.77463006875  -0.735
RegionNorth America                            5.72495239902 57.73241734544   0.099
RegionOceania                                  1.88289012618  2.88712735250   0.652
RegionRest of Europe                          24.25529654099  5.03474446853   4.818
RegionSouth America                           13.05484076869  6.07494853367   2.149
YearY2001                                     -0.72519347882  1.55004486363  -0.468
YearY2002                                      0.92319933073  1.59535650649   0.579
YearY2003                                      0.63313541502  1.58053660170   0.401
YearY2004                                      0.30672445810  1.60308173960   0.191
YearY2005                                      0.38090227475  1.65320186398   0.230
YearY2006                                      0.39471427719  1.69764992179   0.233
YearY2007                                      0.30425644386  1.71029772524   0.178
YearY2008                                      1.10154618874  1.72954835553   0.637
YearY2009                                      2.28049643366  1.72221558870   1.324
YearY2010                                      2.25477836720  1.71284304502   1.316
YearY2011                                      2.99505356149  1.71287728910   1.749
YearY2012                                      2.33328835738  1.73091383573   1.348
YearY2013                                      3.32822628879  1.74878779732   1.903
YearY2014                                      4.44357376174  1.76862244775   2.512
YearY2015                                      4.39068997718  1.75737626591   2.498
Infant_deaths                                 -0.58033367499  0.03437738787 -16.881
Adult_mortality                               -0.02675008978  0.00784963223  -3.408
Hepatitis_B                                   -0.05814069638  0.01068132164  -5.443
BMI                                           -0.46714700503  0.11719507813  -3.986
Polio                                          0.00344440659  0.01261237792   0.273
Incidents_HIV                                  0.57414891110  0.12189254867   4.710
GDP_per_capita                                 0.00003336493  0.00002954644   1.129
Thinness_ten_nineteen_years                    1.02128707383  0.27729696937   3.683
Thinness_five_nine_years                      -1.37884361974  0.26614123577  -5.181
Schooling                                     -1.71762542079  0.26026907784  -6.599
Economy_status_DevelopedYes                    2.17113688056  0.66437387606   3.268
RegionAsia:YearY2001                          -0.15430526821  0.32672044151  -0.472
RegionCentral America and Caribbean:YearY2001  0.03156674206  0.38872789207   0.081
RegionEuropean Union:YearY2001                 0.18785954137  0.37145692534   0.506
RegionMiddle East:YearY2001                   -0.11302493696  0.45910335705  -0.246
RegionNorth America:YearY2001                  0.44763273210  0.92208427063   0.485
RegionOceania:YearY2001                       -0.05189707533  0.48830036526  -0.106
RegionRest of Europe:YearY2001                 0.11364998745  0.41947931785   0.271
RegionSouth America:YearY2001                 -0.08303116102  0.44069960627  -0.188
RegionAsia:YearY2002                           0.08835682969  0.32261424920   0.274
RegionCentral America and Caribbean:YearY2002  0.21332001107  0.38843387209   0.549
RegionEuropean Union:YearY2002                 0.47917753228  0.37425263120   1.280
RegionMiddle East:YearY2002                    0.25065892574  0.46008233428   0.545
RegionNorth America:YearY2002                  1.07276145020  1.16268369668   0.923
RegionOceania:YearY2002                        0.35833523067  0.48982440852   0.732
RegionRest of Europe:YearY2002                 0.34560196552  0.42296870248   0.817
RegionSouth America:YearY2002                  0.19240206819  0.44562657845   0.432
RegionAsia:YearY2003                           0.17215132603  0.32791447055   0.525
RegionCentral America and Caribbean:YearY2003  0.18827398552  0.39129957666   0.481
RegionEuropean Union:YearY2003                 0.49818178340  0.37609263161   1.325
RegionMiddle East:YearY2003                    0.25196694445  0.46167539862   0.546
RegionNorth America:YearY2003                  1.42145195883  1.45756064938   0.975
RegionOceania:YearY2003                        0.42170064064  0.49256068132   0.856
RegionRest of Europe:YearY2003                 0.44646157222  0.43070310093   1.037
RegionSouth America:YearY2003                  0.15244364809  0.45160801386   0.338
RegionAsia:YearY2004                           0.28245784760  0.32246591473   0.876
```

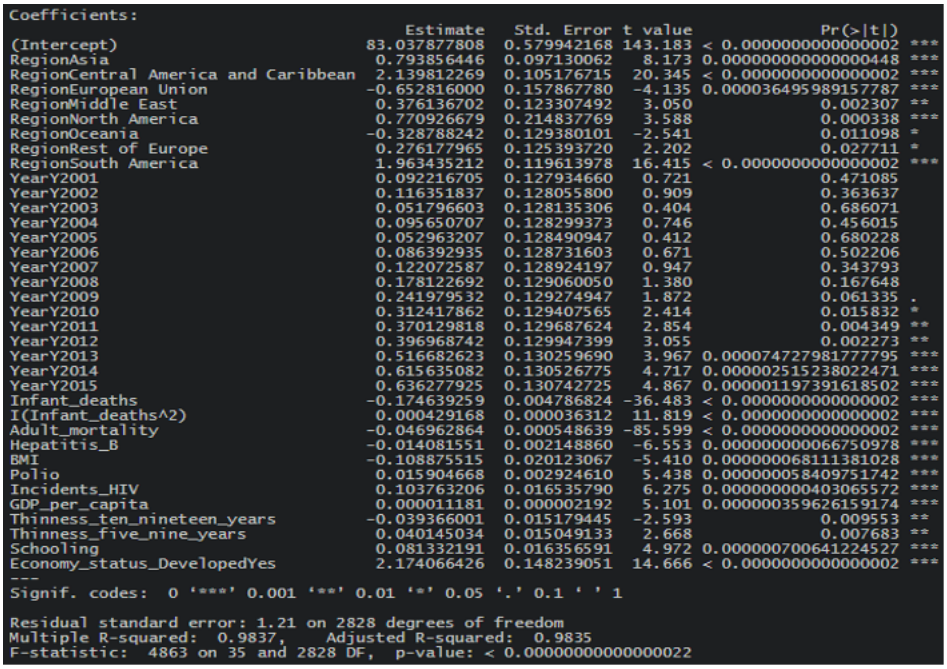
Figure 3.11 *Result of Interaction Model Summary – I*



```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8377 on 2546 degrees of freedom
Multiple R-squared:  0.9929,    Adjusted R-squared:  0.9921
F-statistic:  1131 on 317 and 2546 DF,  p-value: < 0.00000000000000022
```

Figure 3.12 *Result of Interaction Model Summary – II*



```
Coefficients:
                                    Estimate   Std. Error t value             Pr(>|t|)
(Intercept)                       83.037877808  0.579942168 143.183 < 0.0000000000000002 ***
RegionAsia                         0.793856446  0.097130062   8.173 0.000000000000000448 ***
RegionCentral America and Caribbean  2.139812269  0.105176715  20.345 < 0.0000000000000002 ***
RegionEuropean Union              -0.652816000  0.157867780  -4.135 0.000036495989157787 ***
RegionMiddle East                  0.376136702  0.123307492   3.050             0.002307 **
RegionNorth America                0.770926679  0.214837769   3.588             0.000338 ***
RegionOceania                     -0.328788242  0.129380101  -2.541             0.011098 *
RegionRest of Europe               0.276177965  0.125393720   2.202             0.027711 *
RegionSouth America                1.963435212  0.119613978  16.415 < 0.0000000000000002 ***
YearY2001                          0.092216705  0.127934660   0.721             0.471085
YearY2002                          0.116351837  0.128055800   0.909             0.363637
YearY2003                          0.051796603  0.128135306   0.404             0.686071
YearY2004                          0.095650707  0.128299373   0.746             0.456015
YearY2005                          0.052963207  0.128490947   0.412             0.680228
YearY2006                          0.086392935  0.128731603   0.671             0.502206
YearY2007                          0.122072587  0.128924197   0.947             0.343793
YearY2008                          0.178122692  0.129060050   1.380             0.167648
YearY2009                          0.241979532  0.129274947   1.872             0.061335 .
YearY2010                          0.312417862  0.129407565   2.414             0.015832 *
YearY2011                          0.370129818  0.129687624   2.854             0.004349 **
YearY2012                          0.396968742  0.129947399   3.055             0.002273 **
YearY2013                          0.516682623  0.130259690   3.967 0.000074727981777795 ***
YearY2014                          0.615635082  0.130526775   4.717 0.000002515238022471 ***
YearY2015                          0.636277925  0.130742725   4.867 0.000001197391618502 ***
Infant_deaths                     -0.174639259  0.004786824 -36.483 < 0.0000000000000002 ***
I(Infant_deaths^2)                 0.000429168  0.000036312  11.819 < 0.0000000000000002 ***
Adult_mortality                   -0.046962864  0.000548639 -85.599 < 0.0000000000000002 ***
Hepatitis_B                       -0.014081551  0.002148860  -6.553 0.000000000066750978 ***
BMI                               -0.108875515  0.020123067  -5.410 0.000000068111381028 ***
Polio                              0.015904668  0.002924610   5.438 0.000000058409751742 ***
Incidents_HIV                      0.103763206  0.016535790   6.275 0.000000000403065572 ***
GDP_per_capita                     0.000011181  0.000002192   5.101 0.000000359626159174 ***
Thinness_ten_nineteen_years       -0.039366001  0.015179445  -2.593             0.009553 **
Thinness_five_nine_years           0.040145034  0.015049133   2.668             0.007683 **
Schooling                          0.081332191  0.016356591   4.972 0.000000700641224527 ***
Economy_status_DevelopedYes        2.174066426  0.148239051  14.666 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.21 on 2828 degrees of freedom
Multiple R-squared:  0.9837,    Adjusted R-squared:  0.9835
F-statistic:  4863 on 35 and 2828 DF,  p-value: < 0.00000000000000022
```

*Figure*3.13*ResultofHigherOrderModelSummary*

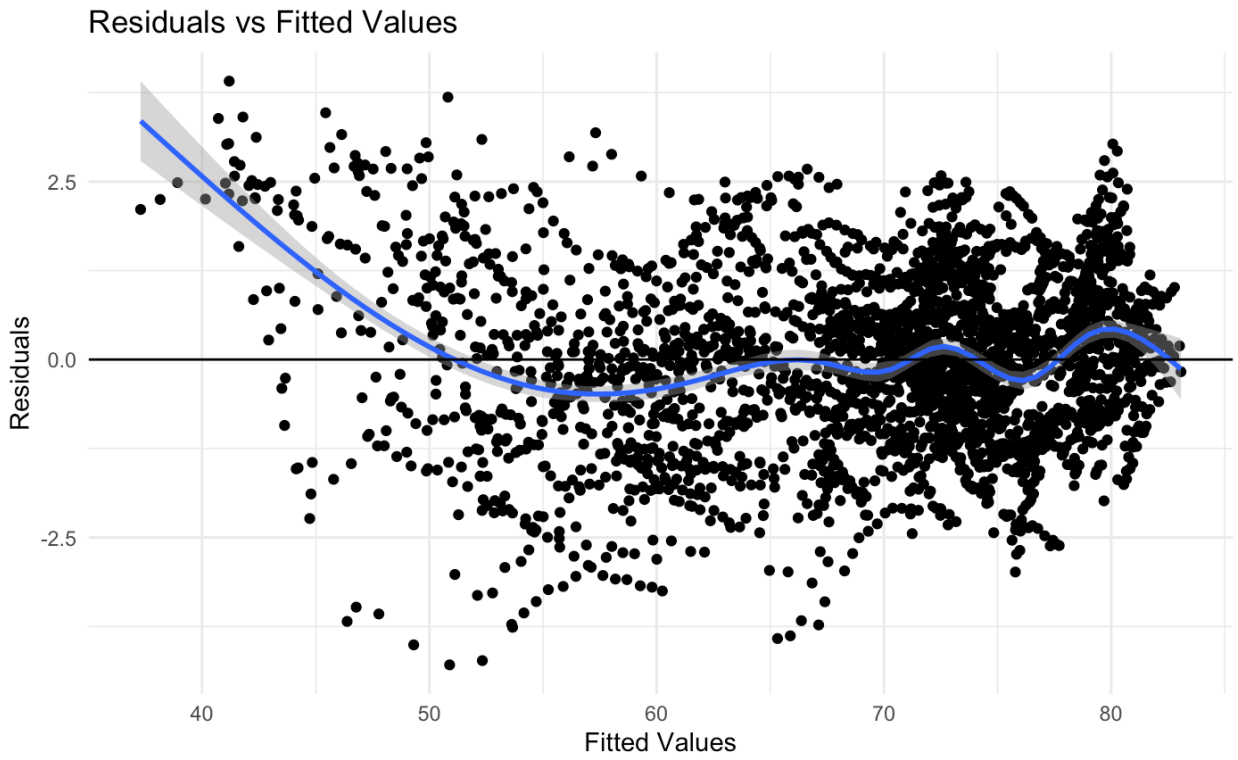


Figure 3.14 *Result of Residuals vs Fitted Values – I*

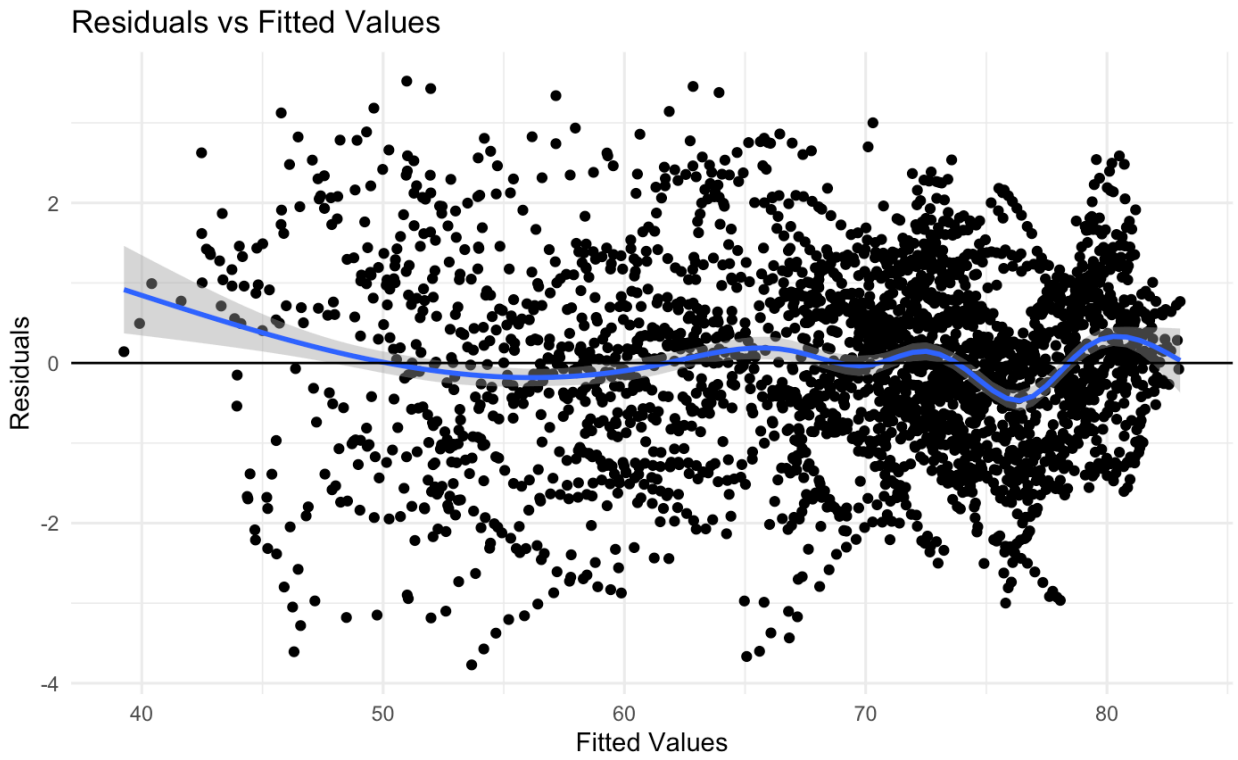


Figure 3.15 Result of Residuals vs Fitted Values – II

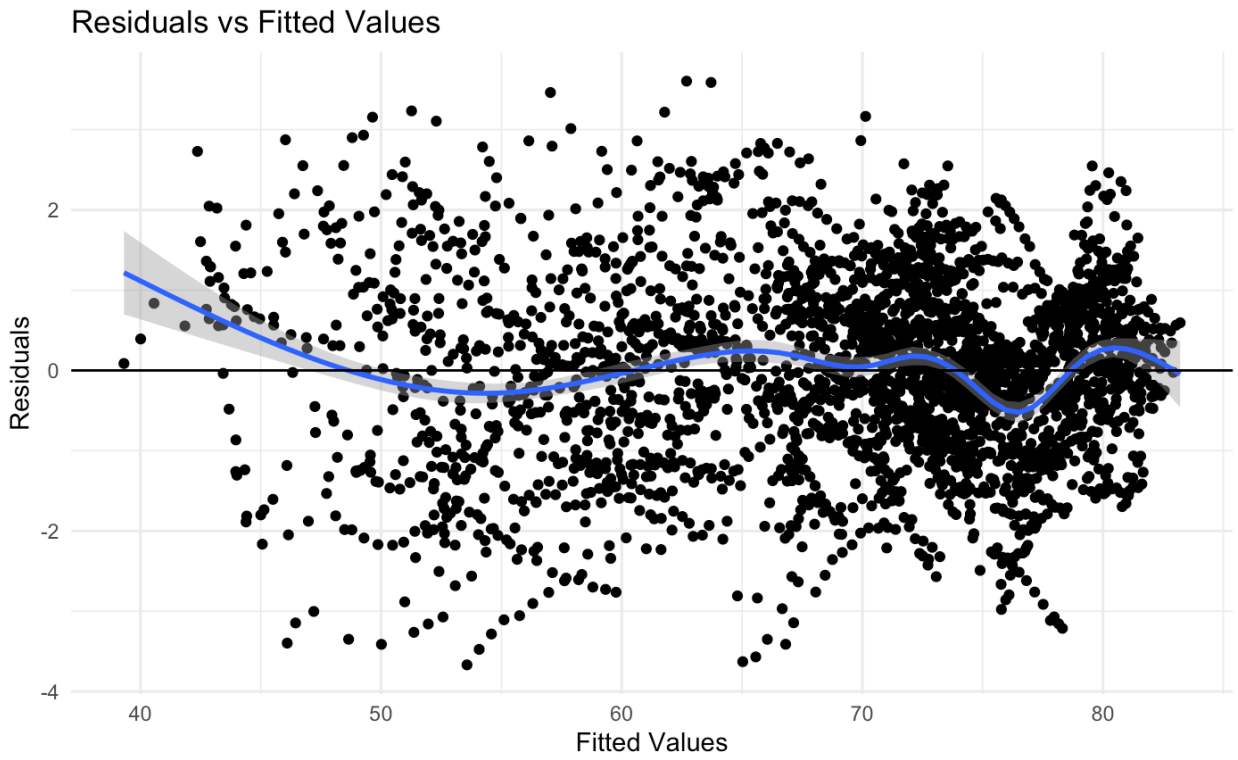


Figure 3.16 *Result of Residuals vs Fitted Values – III*

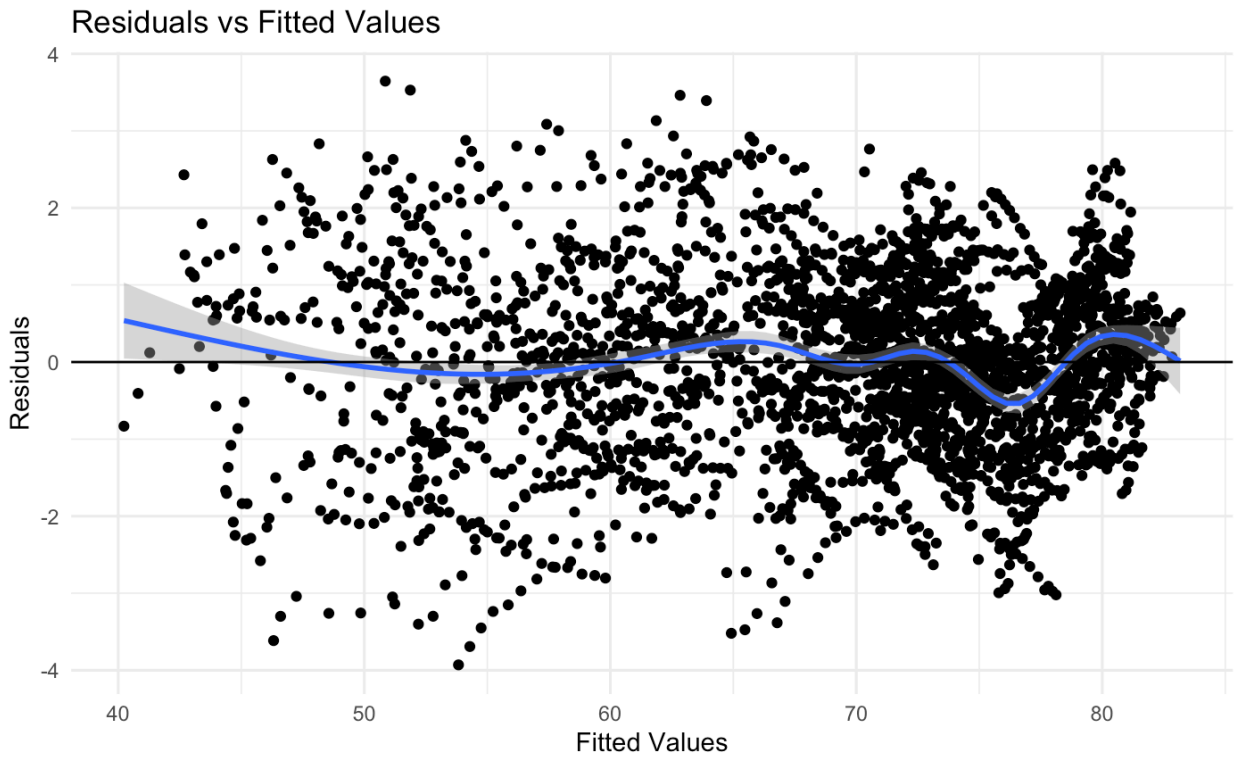


Figure 3.17 Result of Residuals vs Fitted Values – IV

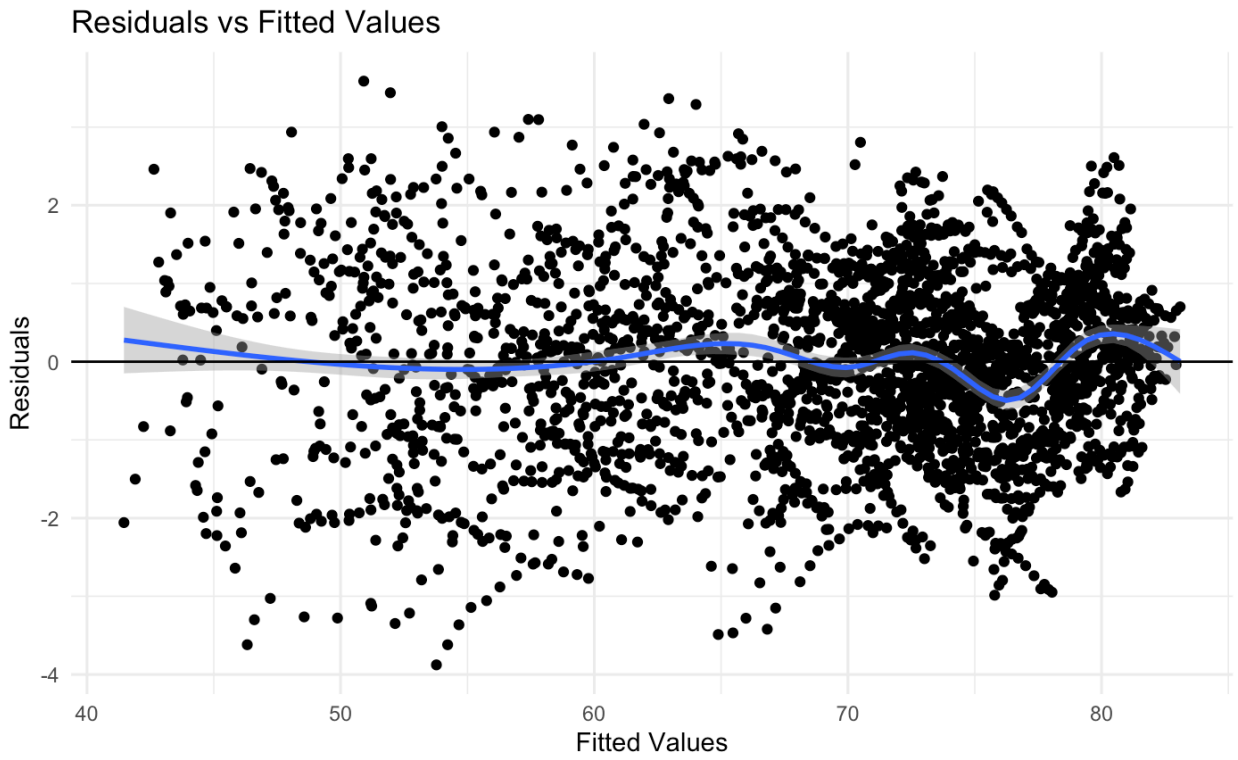


Figure 3.18 *Result of Residuals vs Fitted Values − V*

```
        studentized Breusch-Pagan test

data:  valid_model
BP = 530.63, df = 37, p-value < 0.00000000000000022
```

Figure 3.19 *Breusch − Pagan Test* 1

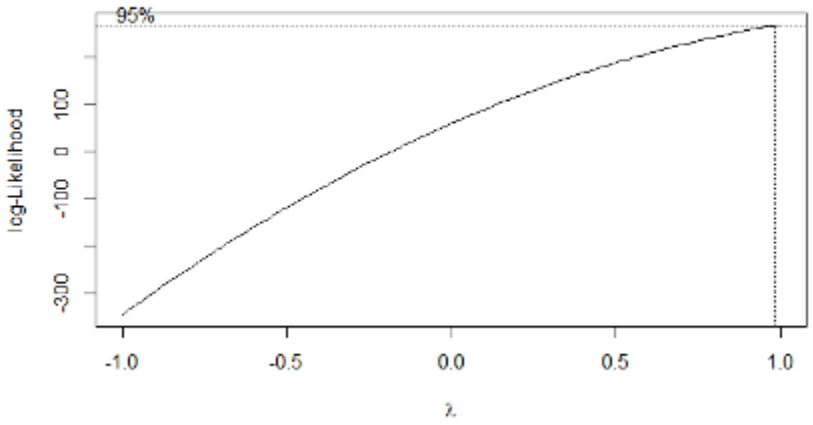


*Figure*3.20*BoxCoxPlot*

```
        studentized Breusch-Pagan test

data:  model_transformed
BP = 679.85, df = 37, p-value < 0.00000000000000022
```

*Figure*3.21*Breusch* − *PaganTest*2

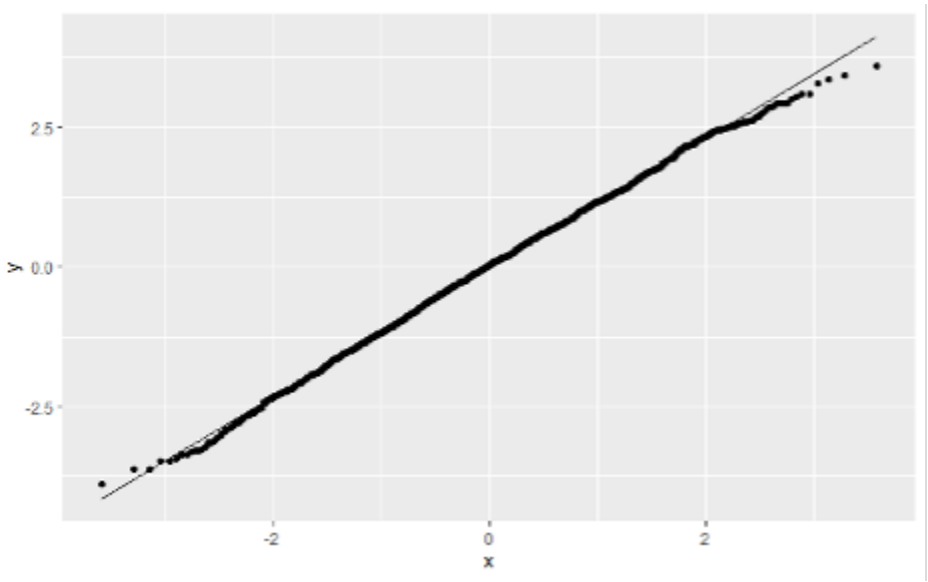


Figure 3.22 *QQ Plot* 2

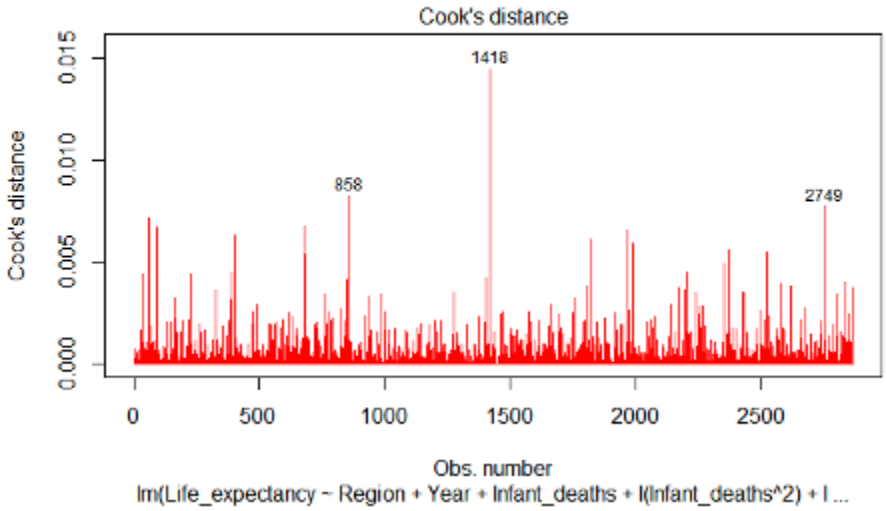


Figure 3.23 *Cooks Distance Plot*