Life Expectancy Statistical Analysis Based on Economic, Health, and Lifestyle Factors

Group L03-22: Jemima Balbastro, Lydia Chien, Janice Zhang, Simran Arora

2024-12-01

1. Introduction

1.1 Motivation

1.1.1 Context

The research topic for this project is to understand how economic, health, and lifestyle factors affects life expectancy. The group has identified a Life Expectancy data set on Kaggle, which contains aggregated data from multiple sources including the World Health Organization, World Bank, and Our World in Data. This data set covers 179 countries from 2000 to 2015 and includes 19 variables representing various factors such as economic conditions, education levels, immunization statistics, lifestyle habits, and mortality rates.

Life expectancy varies by country due to a combination of these factors.By examining life expectancy in different countries, researchers can gain valuable insights into the role each factor plays in influencing life expectancy trends. This topic is critical, as life expectancy serves as a significant indicator of societal well-being and can inform policies aimed at fostering healthier, more equitable societies.

1.1.2 Problem

The primary problem this project seeks to address is identifying which factors have the strongest relationship with life expectancy using multiple linear regression modeling. Specifically, we aim to explore how various socio-economic and health-related variables influence life expectancy. By leveraging techniques such as individual t-tests, additive modeling, automated model selection, and and interaction modeling within a multivariable regression framework, we will determine the key predictors of life expectancy and the statistical significance of its impact. This analysis will provide insights into life expectancy trends and inform strategic policy decisions aimed at improving societal well-being.

1.1.3 Challenges

This problem is challenging due to the complexity and size of the dataset. With 179 countries and 19 variables spanning multiple years , the dataset is large and contains many inter-related factors. Managing this volume of data requires careful processing to handle missing values, with potential multicollinearity among variables, and outliers that could effect the results. Additionally, the relationships between predictors and life expectancy may not be linear or independent, which require advanced modeling techniques to capture interactions and non-linear trends accurately. These factors make it difficult to build an interpretable and accurate model that reliably predicts life expectancy while accounting for these nuances. Another challenge is that the model contains too many significant predictors. Due to this issue, there are 11 x 11 interaction terms, making it difficult for the team to continue the analysis manually. This has led us to reduce the

number of predictors in the additive model and focus only on the predictors that are of interest for further analysis.

1.2 Objective

1.2.1 Overview

The intent of this project is to make use of multiple linear regression model to gather insights into which variables relating to economic, health, and lifestyle factors affect life expectancy. As part of this analysis we check for all six regression assumptions such as Linearity, Multicollinearity and Normality assumptions to test the validity of our data. To find the model that best explains our response variable life expectancy we utilize additive, interaction and higher order models to select the final model for analysis.

1.2.2 Goals & Research Questions

The goal of this project is to identify which variables have the strongest relationship with life expectancy using multiple linear regression modeling, and provide the best model in predicting life expectancy. Understanding the impact of each variable can help forecast life expectancy trends for specific countries and inform strategic policy decisions. For example, if GDP or education levels impact life expectancy significantly, governments can provide make professional training and schooling more accessible for its workforce and boost its countries' productivity. Through this project, the group aims to share findings that deepen our understanding of which variables significantly impact life expectancy and to what degree. We will also determine what interactions, if any, affect life expectancy.

We will be considering three research questions to help us analyze our data:

1. Which factors effect life expectancy?
2. How can we interpret the betas of our final model to assess how these factors affect life expectancy?
3. Does there exist any interaction between these variables?

We will be creating various visualization such as residual and box plots to conduct regression assumption tests and to perform higher order analysis.

2. Methodology

2.1 Data

The data set "Life Expectancy" that we will be using in this research study is a publicly available data set from Kaggle, a website that publicly shares open data for analytical purposes among the data science community. In total, it has 19 variables (aside from our response variable, Life_expectancy) relating to factors such as health, economy and mortality of 179 countries across the years of 2000 to 2015. The source of the data set can be found here:https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated. The Life Expectancy data set is an aggregate of public individual data sets from the World Health Organization data repository, World Bank Open Data, and Our World in Data. There are 2864 rows in the data and each row represents statistics collected from a country in a specific year.

There are 3 qualitative columns in the data set:

1. Country - Identifies the name of the country

2. Region - The region in which a country resides (Region of the country (9 categories – Africa, Asia, Central America and Caribbean, European Union, Middle East, North America, Oceania, Rest of Europe, South America).

3. Economy_status_developed - Indicates if the country is developed or not. 0 = Developing Country and 1 = Developed Country

There are 16 quantitative columns in the data (aside from the response variable, Life_expectancy):

1. Life expectancy - This is the response variable it is quantitative and continuous. It measures the average number of years that a newborn (both genders) can expect to live for a specific year in a given country (from birth).

2. Year - Year of which the information in the row corresponds to.

3. Infant_deaths - Represents the death of infants that are less than one-year-old per 1000 live births.

4. Under_five_deaths - Represents death of children under five per 1000 live births.

5. Adult_mortality - Represents the probability of dying between the ages of 15 to 60 years per 1000 population.

6. Alcohol_consumption - Represents the alcohol per capita consumption for the population over the age of 15 years. The unit of measurement is in liters of pure alcohol per person per year.

7. Hepatitis_B - This is the percentage of coverage of Hepatitis B (HepB3) immunization among one-year-old children.

8. Measles - This is the percentage of coverage of Measles containing vaccine first dose (MCV1) immunization among one-year-old children.

9. BMI - This is the mean body mass index of the adult population calculated using weight and height (kg/m^2).

10. Polio- This is the percentage of coverage of Polio (Pol3) immunization among one-year-old children.

11. Diphtheria - This is the percentage of coverage of Diphtheria tetanus toxoid and pertussis (DTP3) immunization among one-year-old children.

12. Incidents_HIV - This is the number of new HIV infections for populations aged 15-49 per 1000 uninfected population in the year before.

13. GDP_per_capita - The GDP per capita of the country for the corresponding year measured in USD.

14. Population_mln - The total population of the country in the millions.

15. Thinness_ten_nineteen_year - This is based on measured height and weight, measuring the percentage of thinness (with BMI less than 2 standard deviations below the median) among adolescents aged 10-19 years, according to the WHO references for school-age children and adolescents .

16. Thinness_five_nine_years - This is based on measured height and weight, measuring the percentage of thinness among school-aged children aged 5-9 years.

17. Schooling - This measures the average years that individuals over the age of 25 years spent in formal education in each country in each year. Typical duration required to complete primary education usually takes about 6 years, secondary education 4-6 years, and higher education even longer.

2.2 Approach

2.2.1 Assumption Testing

Before proceeding with multiple linear regression analysis, we will first validate the dataset by conducting a series of assumption tests. These tests will ensure the appropriateness of our modeling approach and will be discussed in detail in the Workflow section below. To evaluate the assumptions of linear regression, we will employ various data visualizations, including:

- Residuals vs. Fitted plots: To check for linearity and homoscedasticity.
- Residuals vs. Spatial Variable plots: To identify potential patterns associated with spatial variables.
- Scale-Location plots: To assess the constancy of variance.
- Q-Q plots: To evaluate the normality of residuals.
- Histograms of residuals: To further confirm normality assumptions.

Understanding the behavior of residual errors is critical, as the results of these tests directly influence the reliability of our interpretations and subsequent analyses.

2.2.2 Model Development

The primary objective of this study is to identify the most optimal predictive model for the response variable, Life_expectancy. We will approach this goal using the following steps:

1. Additive Model Creation

We will build an additive linear regression model, including statistically significant predictors at a significance level of $\alpha = 0.05$. Model comparison will involve Full and Partial-F tests, with p-values guiding the inclusion or exclusion of predictors. Individual t-tests will be used to determine the statistical significance of each predictor.

2. Automated Model Selection

We will apply Stepwise Regression, Forward Selection, and Backward Elimination techniques to identify optimal models. Adjusted R-squared will serve as the criterion for comparing the models produced by these methods against the manually created additive model.

3. Exploring Interaction and Higher-Order Terms

To enhance the model further, we will evaluate the inclusion of interaction terms and polynomial (higher-order) terms. Models incorporating these terms will also be assessed using Adjusted R-squared to determine their predictive performance.

2.2.3 Reporting and Interpretation

Throughout this process, the output of each multiple linear regression model will be thoroughly analyzed. Key metrics, including p-values, Adjusted R-squared, and residual errors, will be reported. The significance of predictors and models will be assessed relative to the established $\alpha = 0.05$ threshold, ensuring transparency.

By systematically refining our model through these steps, we aim to derive a robust and best-fitting model that provides meaningful insights into the factors driving variations in life expectancy.

2.3 Workflow

1. Introduction The workflow tasks starts with an introduction of the project and the dataset to provide the background of the study. We use multiple linear regression modelling for the project.

2. Assumptions Build the full model and do the assumption tests first to confirm the validity of the dataset, if any of the test does not work out, explain the reasons and decide if we can proceed to the next steps.

- Linearity Assumption
- Independence Assumption
- Equal Variance Assumption
- Normality Assumption
- Multicollinearity
- Outliers

The assumption tests are hard as we tried to pass each single test, for example, in the Equal Variance Assumption test, we tried to improve the test by transforming the model into log and Box-Cox but they still did not improve. Fortunately, we figured out the reason behind and explain why we stay with the original model. The full model is still valid for the other steps.

3. Additive Model

- Full model test (hypothesis, ANOVA, interpretation, full model equation) to test the revised full model
- Partial model test (individual t-test, hypothesis, interpretation, equation) for the reduced model
- Partial F-test using ANOVA to see if the reduced model is better
- Comparison table for the full and reduced addictive model

4. Model Selection Procedures

- Stepwise/Forward/Backward Regression Procedure
- Comparison of t-test/Stepwise/Forward/Backward

5. Interaction model

- Partial F - test to determine best interaction model
- t-test to determine significant terms

6. Higher order analysis (for each variable in the reduced interaction model)

- Limit to degree 4 terms to prevent over fitting

7. Final model (interpretation and equation)

- Use Adjusted R-squared as the preferred criterion to decide the best model, as it adjusts for the number of predictors and helps reduce over fitting.

8. Conclusion

2.4 Contributions

The group has distributed the workload among our four teammates. Since we only have four members in the group, we assigned two members to complete tasks relating to the first half of the project, including six assumption tests and the additive model. The other two members will focus on tasks relating to the interaction model and higher order analysis. However, we will go through all the code and interpretations as a group to ensure expectations are being met and to move on to the next steps. The division of work and roles of the group members are agreed by all teammates. The significance level we agreed on is 0.05 when conducting tests such as partial F-Test or individual T-test.

The specific contributions are as follows:

Lydia and Janice were responsible for the first half of the project for accuracy and the quality of the model. They have contributed to the introduction of the project and the background of the data sets. They checked the assumption tests to ensure the validity of the data set. They followed the steps provided in class to reduce the model till all variables are significant. The procedure includes all elements mentioned in class, such as stating the hypothesis, testing the p-value and other test statistics, full model test, partial t-test, adjusted R-squared, residual standard error and interpretation of the model. We then did the additive model and reduced model as a group to make sure we are all good to move on to the second half of the project.

Then Jemima and Simran conducted the interaction model analysis, along with the higher order analysis, to further examine the relationships between variables and how they react to the response variable. Like the additive approach, they stated the hypothesis, provide the p-value of the model, the Adjusted R-squared, and the Residual Standard Error. They reduced the interaction model further if needed depending on the p-values of the interactions, and then conduct an F-test to determine if the reduced interaction model is better than the full interaction model. Next, they conducted a higher order analysis to see if any non-linear relationships exist that could affect our response variable. Using the pairwise plots, we observed the curvature of the relationship between Life Expectancy and each independent variable, barring categorical variables. Then We continued the higher order analysis until we found the model with the highest possible Adjusted R-squared value.

Once we have analyzed our additive, interaction, and higher order models as a group we compared the Adjusted R-squared and the Residual Standard Errors between all models to determine which is the best model. We provided the final model equation and interpreted it in the context of life expectancy, making sure our conclusion only included variables and interactions that are the most significant when predicting life expectancy based on economic, lifestyle, and health factors.

3. Main Results of the Analysis

3.1 Preparation

We will start with accessing the necessary packages for the multiple linear regression analysis.

```
## Warning: package 'olsrr' was built under R version 4.4.2
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
## Warning: package 'car' was built under R version 4.4.2
```

```
## Loading required package: carData

## Warning: package 'carData' was built under R version 4.4.2

## Warning: package 'lmtest' was built under R version 4.4.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.4.2

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'MASS'

## The following object is masked from 'package:olsrr':
##
##     cement
```

Table 1: The first 6 rows of the Life Expectancy data set.

	Country	Region	Year	Infant_deaths	Under_five_deaths
1	Turkiye	Middle East	2015	11.1	13.0
2	Spain	European Union	2015	2.7	3.3
3	India	Asia	2007	51.5	67.9
4	Guyana	South America	2006	32.8	40.5
5	Israel	Middle East	2012	3.4	4.3
6	Costa Rica	Central America and Caribbean	2006	9.8	11.2

	Adult_mortality	Alcohol_consumption	Hepatitis_B	Measles	BMI	Polio	Diphtheria
1	105.8240	1.32	97	65	27.8	97	97
2	57.9025	10.35	97	94	26.0	97	97
3	201.0765	1.57	60	35	21.2	67	64
4	222.1965	5.68	93	74	25.3	92	93
5	57.9510	2.89	97	89	27.0	94	94
6	95.2200	4.19	88	86	26.4	89	89

	Incidents_HIV	GDP_per_capita	Population_mln	Thinness_ten_nineteen_years
1	0.08	11006	78.53	4.9
2	0.09	25742	46.44	0.6
3	0.13	1076	1183.21	27.1
4	0.79	4146	0.75	5.7
5	0.08	33995	7.91	1.2
6	0.16	9110	4.35	2.0

	Thinness_five_nine_years	Schooling	Economy_status_Developed
1	4.8	7.8	0
2	0.5	9.7	1
3	28.0	5.0	0

```
## 4                     5.5       7.9                      0
## 5                     1.1      12.8                      1
## 6                     1.9       7.9                      0
##   Economy_status_Developing Life_expectancy
## 1                         1            76.5
## 2                         0            82.8
## 3                         1            65.4
## 4                         1            67.0
## 5                         0            81.7
## 6                         1            78.2
```

In the Life_Expectancy data set, we identified duplicate columns: Economy_status_Developed and Economy_status_Developing. Both columns represent the same information about a country's economic status (developed or developing). To ensure the accuracy of our analysis, we decided to drop the Economy_status_Developing column. Additionally, after removing this column, a new column labeled "Residual" appeared in the data set. To maintain the completeness and integrity of the data, we also decided to drop the "Residual" column.

Table 2: Life Expectancy data set exlcuding the column "Economy_status_developing" as it is a duplicate column.

```
##      Country                       Region Year Infant_deaths Under_five_deaths
## 1    Turkiye                  Middle East 2015          11.1              13.0
## 2      Spain                European Union 2015           2.7               3.3
## 3      India                         Asia 2007          51.5              67.9
## 4     Guyana                South America 2006          32.8              40.5
## 5     Israel                  Middle East 2012           3.4               4.3
## 6 Costa Rica Central America and Caribbean 2006           9.8              11.2
##   Adult_mortality Alcohol_consumption Hepatitis_B Measles  BMI Polio Diphtheria
## 1        105.8240                1.32          97      65 27.8    97         97
## 2         57.9025               10.35          97      94 26.0    97         97
## 3        201.0765                1.57          60      35 21.2    67         64
## 4        222.1965                5.68          93      74 25.3    92         93
## 5         57.9510                2.89          97      89 27.0    94         94
## 6         95.2200                4.19          88      86 26.4    89         89
##   Incidents_HIV GDP_per_capita Population_mln Thinness_ten_nineteen_years
## 1          0.08          11006          78.53                         4.9
## 2          0.09          25742          46.44                         0.6
## 3          0.13           1076        1183.21                        27.1
## 4          0.79           4146           0.75                         5.7
## 5          0.08          33995           7.91                         1.2
## 6          0.16           9110           4.35                         2.0
##   Thinness_five_nine_years Schooling Economy_status_Developed Life_expectancy
## 1                      4.8       7.8                       0            76.5
## 2                      0.5       9.7                       1            82.8
## 3                     28.0       5.0                       0            65.4
## 4                      5.5       7.9                       0            67.0
## 5                      1.1      12.8                       1            81.7
## 6                      1.9       7.9                       0            78.2
```

3.2 Assumptions

Statistical tests and models rely heavily on the assumptions of the underlying data. To ensure the reliability and trustworthiness of our project's results, we conducted an assumption analysis, which includes the

following checks:

1. Linearity Assumption
2. Independence Assumption
3. Equal Variance Assumption
4. Normality Assumption
5. Multicollinearity
6. Outliers

The initial step involves setting up the full life expectancy model with all variables, using Life_expectancy as the response variable. Given the size and complexity of our data set, the output is considerably long. Our data set includes qualitative variables such as Country, Region, and Life_Expectancy_data. Notably, the Country variable encompasses 179 unique categories, while Region contains 9 distinct categories. These categorical variables add complexity to the analysis, requiring careful treatment during assumption testing and model development to ensure accurate and meaningful results.

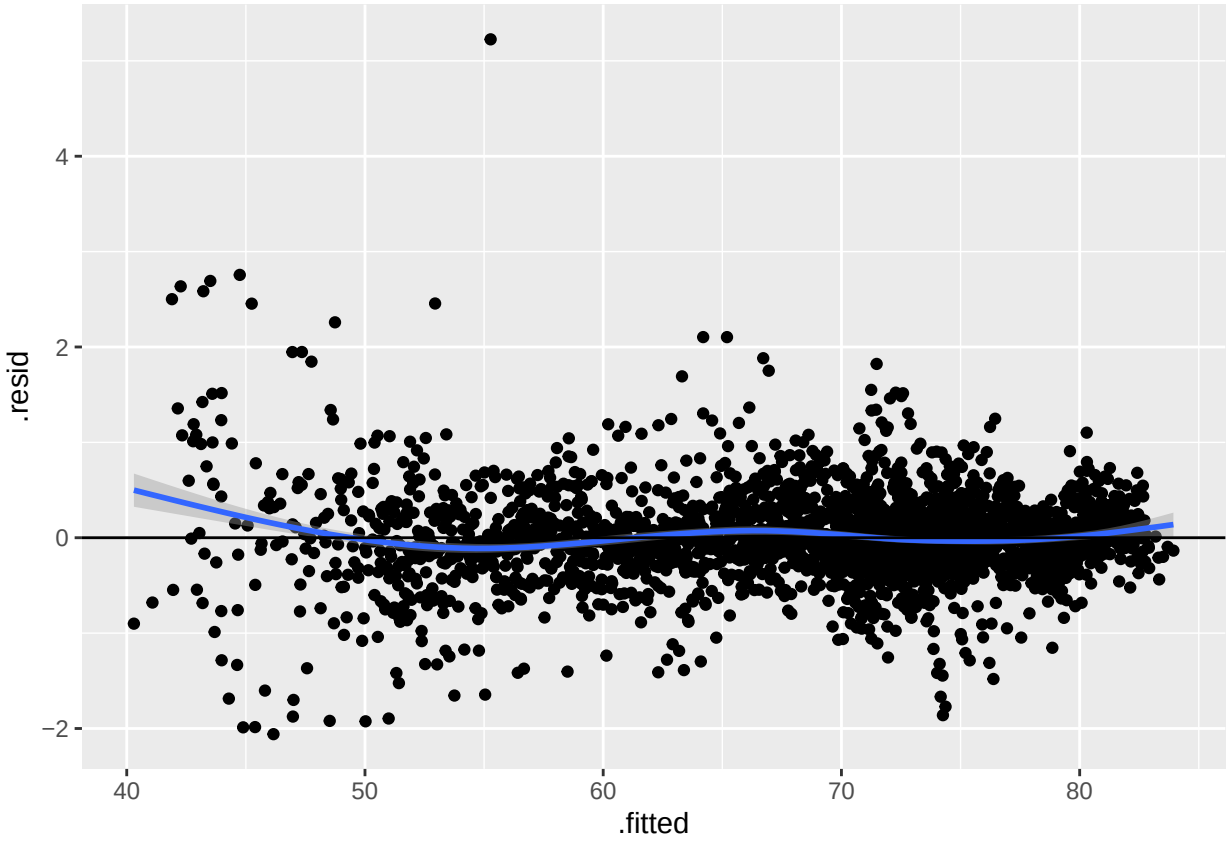
In the output of the full model, the coefficient for Economy_status_Developed is missing (NA). This suggests potential multicollinearity among the variables Country, Region, and Economy_status_Developed. To investigate this, we plan to assess multicollinearity using the Variance Inflation Factor (VIF). However, before proceeding with this analysis, it is essential to verify that the model meets other fundamental assumptions.

3.2.1 Linearity Assumption

The linear regression model assumes a linear relationship between the predictor variables and the response variable. To ensure that our model adheres to this linearity assumption, we will perform appropriate diagnostic checks, such as analyzing residual plots to verify that the residuals are randomly distributed without any discernible patterns. This step is crucial for confirming the validity of the linear relationship and ensuring the reliability of our model's predictions.

Figure 1: Fitted vs. Residual Plot to determine Linearity in the data set

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



The points are mostly scattered randomly around the horizontal line, indicating that the residuals are centered around zero. The smoothed blue line shows a slight curvature, especially in the lower range of the fitted values (40–50) and the higher range (70–80).

Overall, there are some minor non-linear relationship between the predictors and the response variable. However, most of the value have fit in the linear assumption.

We attempted to produce a higher-order graph to further understand where curvatures exist, however we could not output the graph as there is a large number of categories (179 unique countries). Therefore, since the variables closely follows the horizontal line, we will assume that the data set adheres to the linearity assumption and move on to analyzing the data set for independence of residual errors..

3.2.2 Independence Assumption

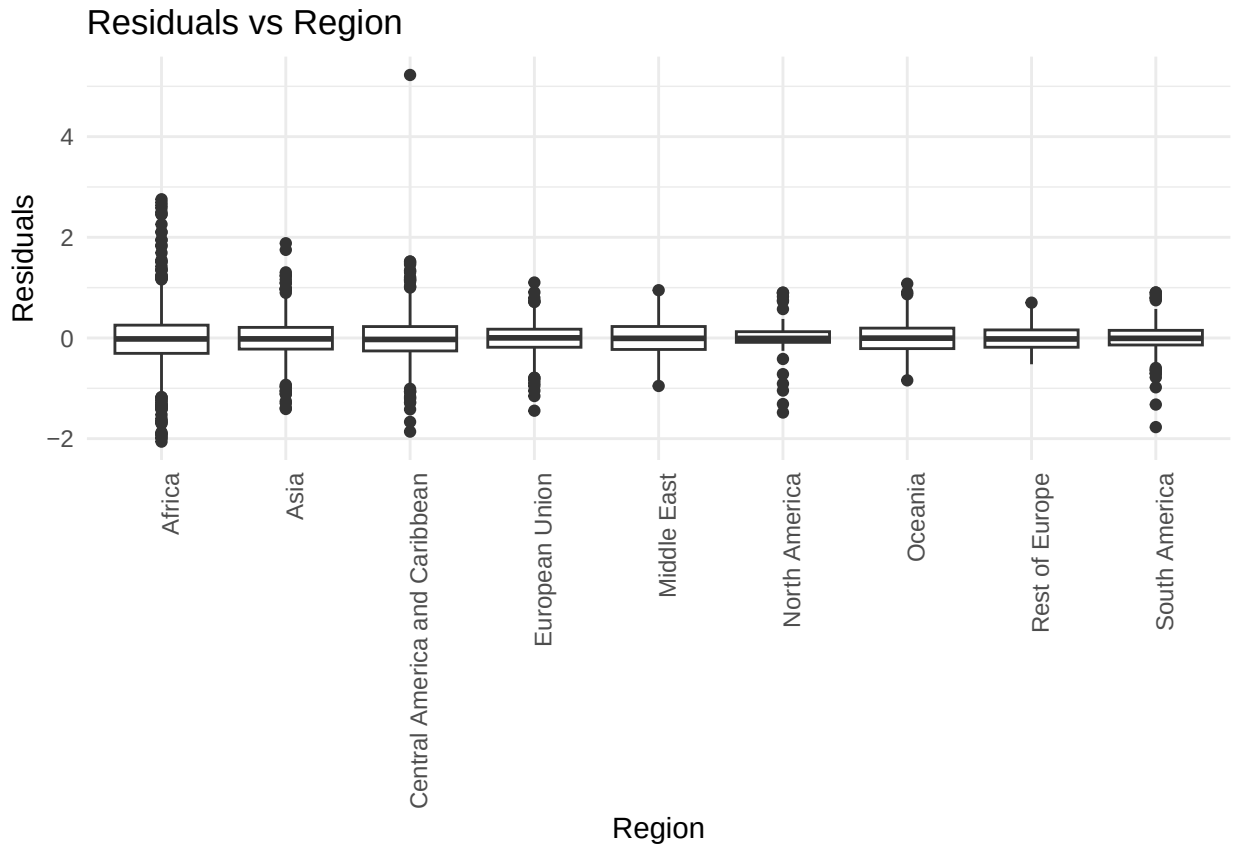
The independence assumption in linear regression requires that the residuals (errors) are independent of each other. Violations of this assumption occur when residuals are correlated, which may arise due to temporal, spatial, or group-based factors. Ensuring independence is critical because correlated residuals can lead to biased estimates and invalid inference.

Ways to check for independence:
- Residuals vs Time (or observation number)
- Residuals vs Spatial variable
- Residuals vs Group (prefer blocking)

Among these checks, we will conduct the Residuals vs Spatial variable analysis and the Residuals vs. Group analysis as these will be the most relevant based on the variables in our data set.

First, as the data set contains the Region variable that represents geographic region, we will perform the Residuals vs Spatial analysis to check for independence:

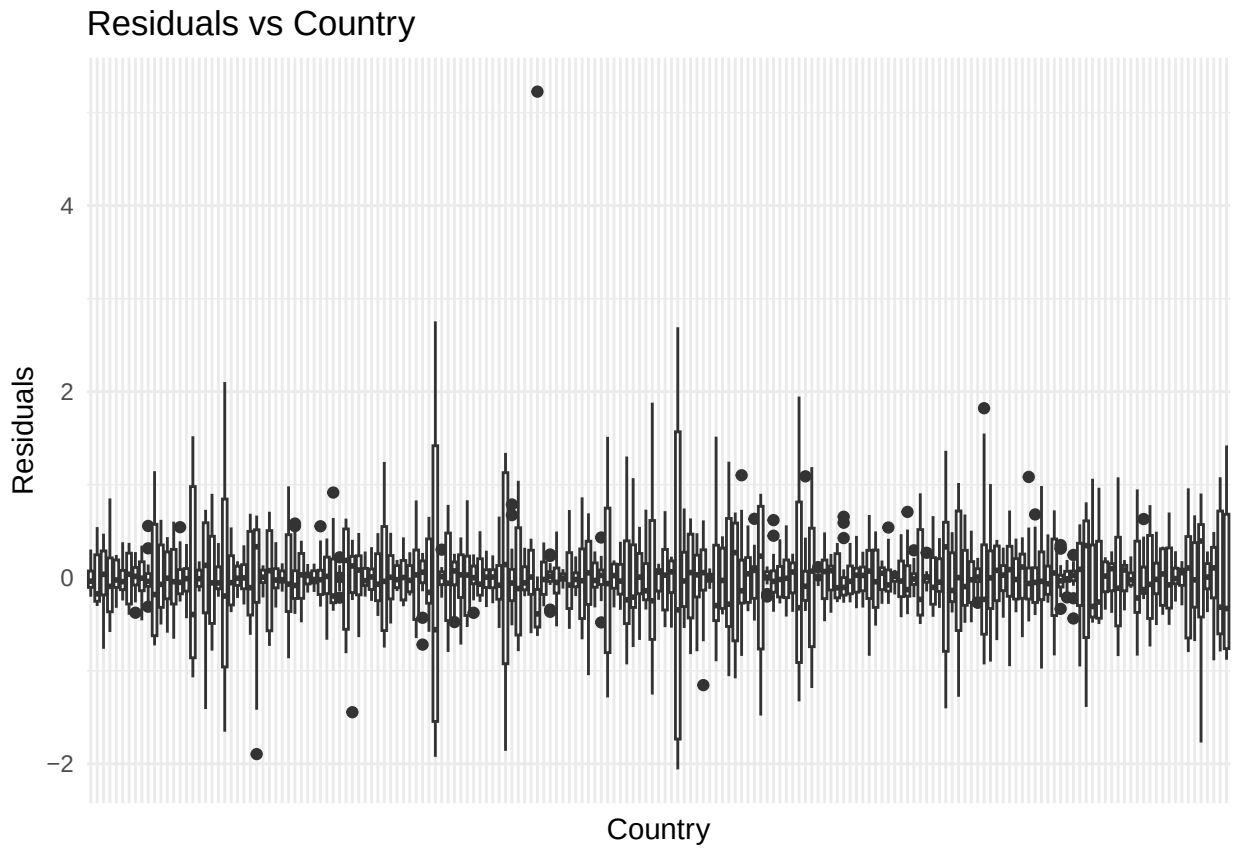
Figure 2: Box plot of Residual Errors by Region. If the interquartile range of the boxplot is centered around 0, we can state that the data set adheres to the independence assumption.



The box plots show slight variation in the central tendency and spread of residuals across regions. Some regions, like "European Union" and "North America," appear to have tighter distributions. However, there are no extreme or systematic deviations. Therefore, the independence assumption across regions is mostly satisfied. There might be minor regional effects influencing the residuals, but they are not severe.

As the data set contains the country variable, we will perform another Residuals vs Spatial variable to check for independence:

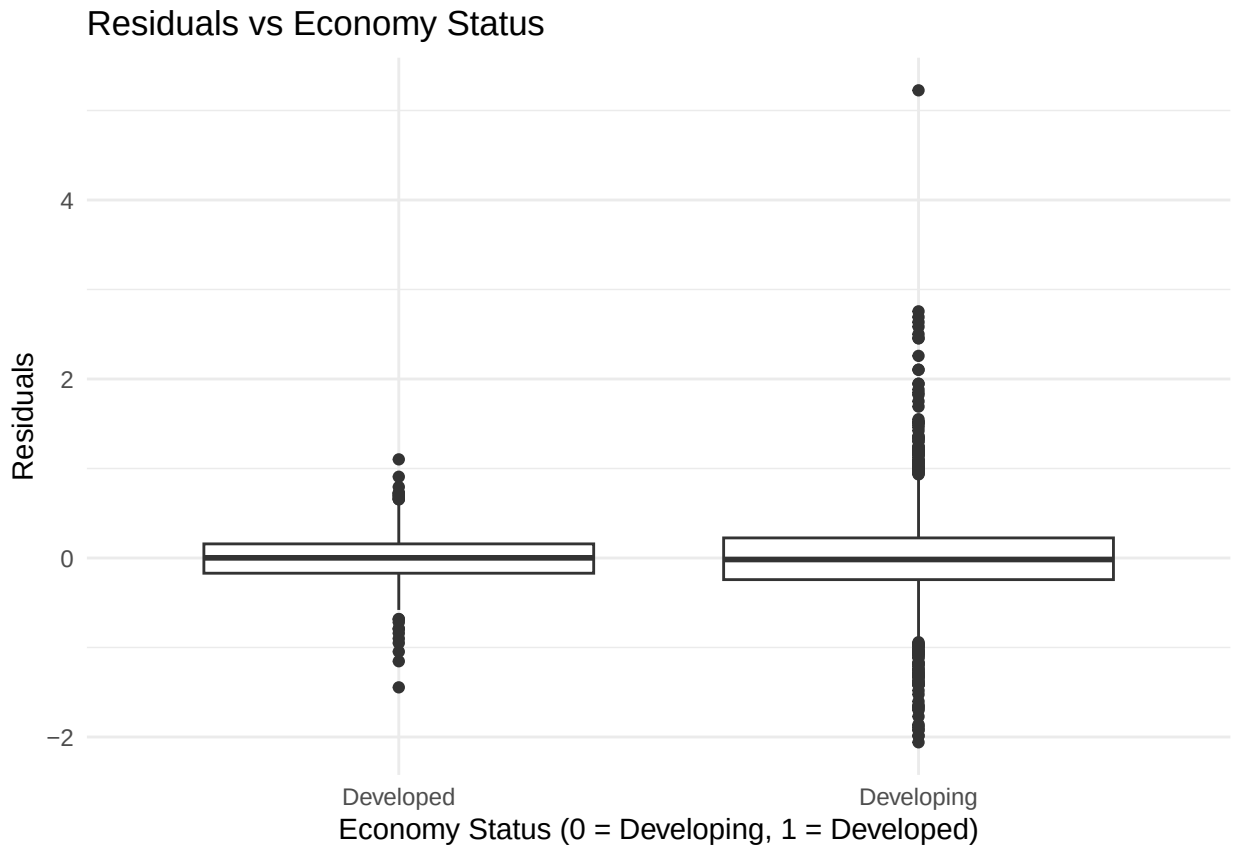
Figure 3: Box plot of Residuals vs. Country



The plot shows residuals distributed around 0 for most countries, but some countries have larger variances. The overall pattern does not suggest a strong violation of independence. Therefore, the independence assumption with respect to Country is satisfied, but certain countries might be outliers or have unique characteristics not well-captured by the model.

As the data set contains a group variable, we do the Residuals vs group (prefer blocking) to check for independence:

Figure 4: Box plot of Residuals by Economy Status. This is a group variable as there are two possible outcomes with this variable: 0 - Developing Country, and 1 - Developed Country.



The boxplots show similar central tendencies around 0 for both groups and the spread of residuals for developing countries is slightly larger than for developed countries. There are a few outliers in both groups, but no systematic pattern. **Therefore, the independence assumption for Economy_status_Developed is satisfied. The model does not show significant bias between the two groups.**

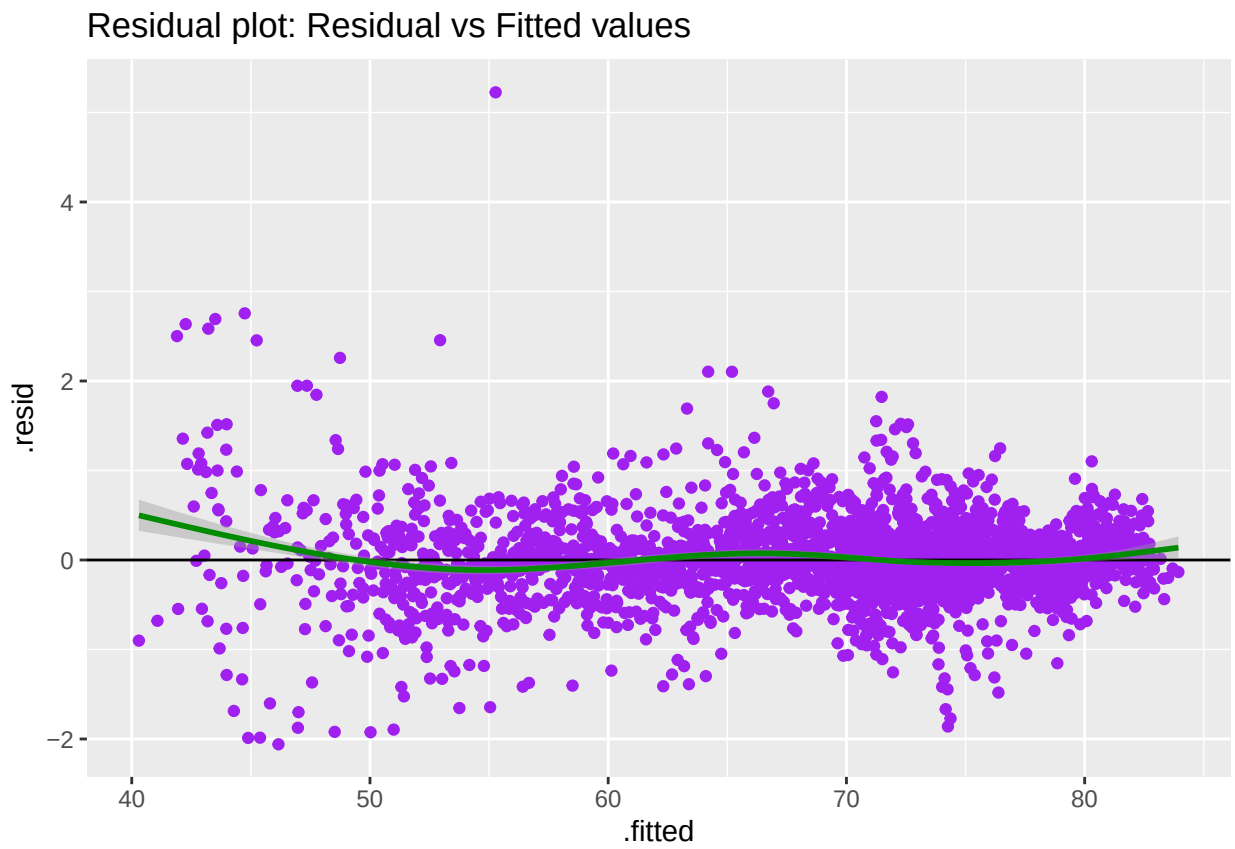
In conclusion, **the residuals for Region, Country and Economy_status_Developed do not show strong evidence of dependence or systematic bias.** Although there are some outliers and minor variations in spread might warrant further investigation, but these are not severe enough to invalidate the model.

3.2.3 Equal Variance Assumption

The Equal Variance Assumption, also known as homoscedasticity, is a key assumption in multiple linear regression. It states that the variance of the residuals should be constant across all levels of the fitted values. This means that the spread of the residuals should not systematically increase or decrease as the predicted values change.

Figure 5: Residuals vs Fitted plot for Homoscedasticity

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

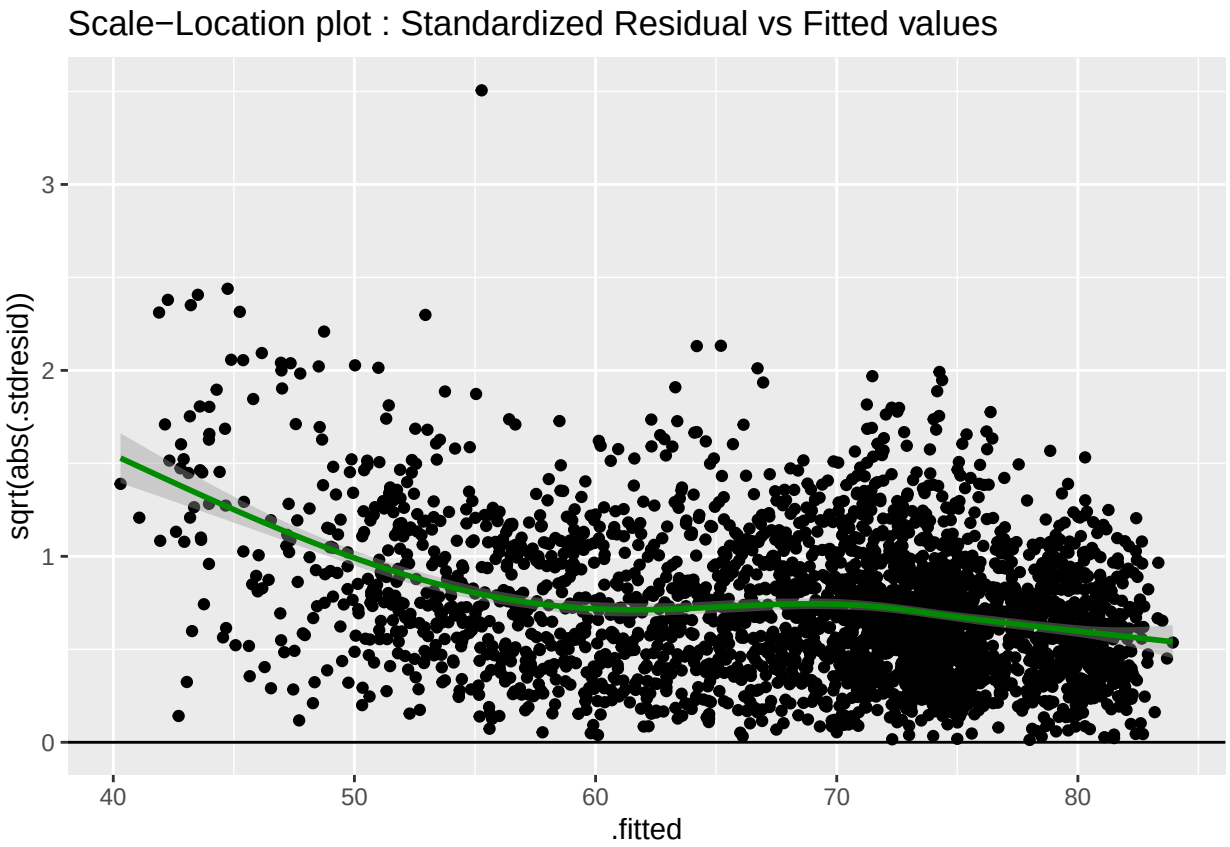


The residuals are scattered around the horizontal line, but the spread of the residuals appears to increase slightly as the fitted values increase. This slight pattern suggests a potential issue with homoscedasticity, as the variance of the residuals is not completely consistent across all fitted values. Therefore, there might be mild heteroscedasticity in the data.

We will conduct another Scale-Location plot to verify our findings:

Figure 6: Scale-Location plot for Homoscedasticity

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



The square root of the absolute standardized residuals decreases slightly as the fitted values increase, forming a slight downward trend. Again, this trend indicates a potential violation of the equal variance assumption. Ideally, the points should be evenly distributed around a horizontal line. The observed decreasing pattern suggests that residual variability is not constant and decreases with higher fitted values.

We then do the Breusch-Pagan Test to confirm homoscedasticity. First, we will state our hypothesis:

Hypothesis:

H_0: Heteroscedasticity is not present (homoscedasticity); all σ are the same

H_a: heteroscedasticity is present; the σ values are not the same.

```
##
##  studentized Breusch-Pagan test
##
## data:  LifeExpectancyFull_old
## BP = 1124.3, df = 194, p-value < 2.2e-16
```

Here, we reject the null hypothesis as we received a p-value of $<$ 2.2e-16 which is less than the significance level of $\alpha = 0.05$, indicating that heteroscedasticity is present in the model.

To address this issue, we need to transform the model in an attempt to adhere to the homoscedasticity assumption. The variable Life_expectancy is transformed using a log-transformation. This transformation is applied to address the issue of heteroscedasticity detected in the original model. Then the next step involves determining the best value for the lambda parameter to apply an appropriate transformation.

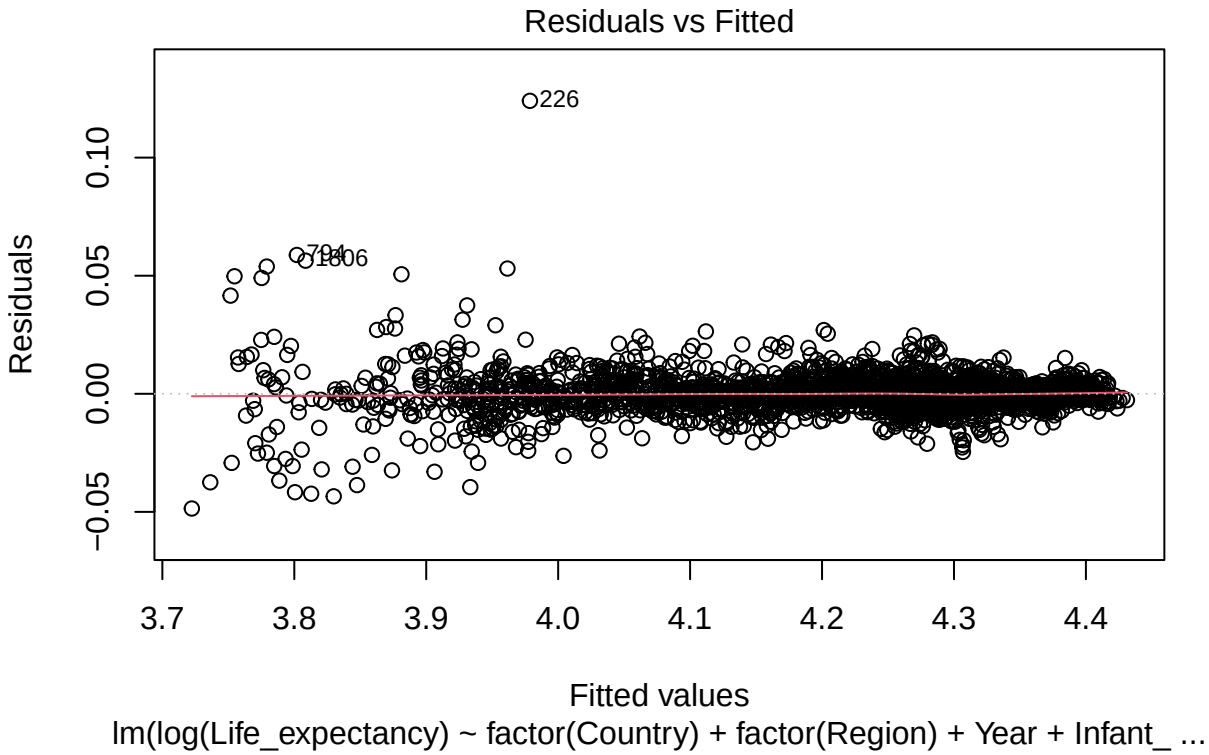
```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  LifeExpectancyFull_log
## BP = 1005.6, df = 194, p-value < 2.2e-16
```

The test statistic (BP = 1005.6) and a very small p-value (p-value < 2.2e-16) indicates that we reject the null hypothesis. This confirms that heteroscedasticity is still present even after the log transformation.

Then we plot to see the residuals vs fitted of the log model:

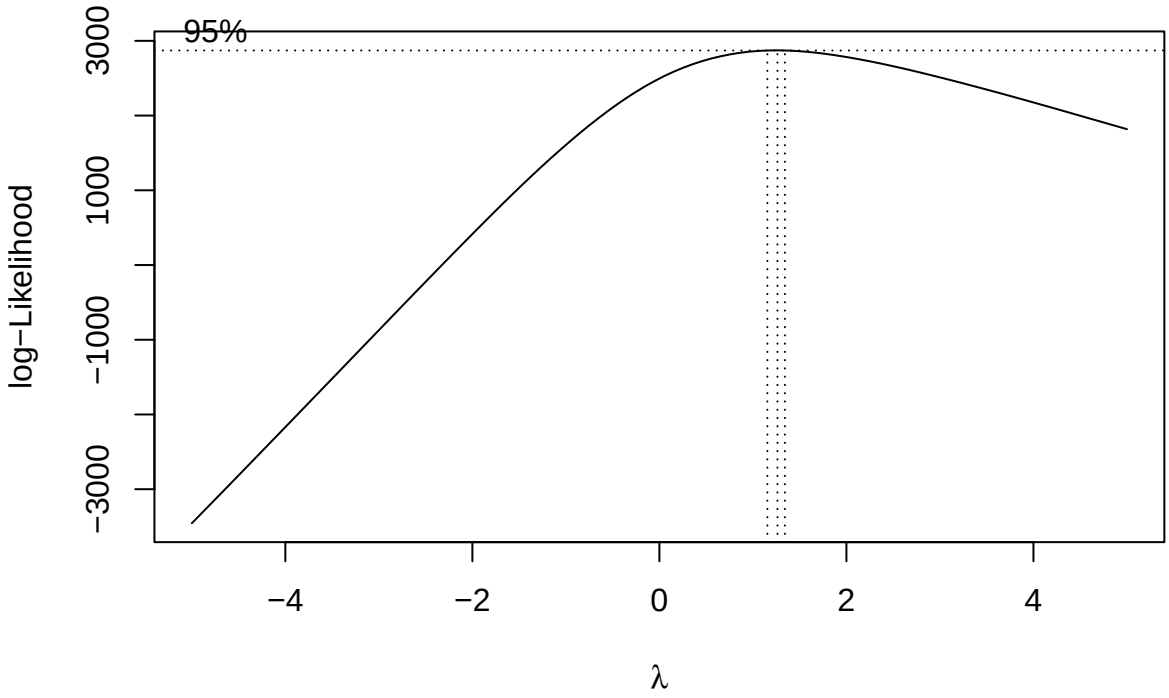
Figure 7: Residuals vs. Fitted plot of the Logarithmic Model to test for Homoscedasticity.



The graph above shows that the log transformation has successfully improved the model as the residuals are now closer the horizontal line. However, we still some funneling in the data as the Residuals are more spread out in the lower range of the Fitted Values, and more concentrated at the higher end of the Fitted Values. The p-value has also not changed as it is still much less than the significance level of $\alpha = 0.05$ which indicates that there is still homoscedasticity.

In an attempt to improve the p-value, we wil use a Box-Cox transformation to see if it will help our data satisfy the homoscedasticity assumption. Before we attempt the Box-Cox transformation, we must first find the best lambda.

Figure 8: Box-Cox Transformation of the Data. This shows the range of where the best lambda may lie.



To understand, where exactly our best lambda resides, we extract it from the Box-Cox transformation above:

```
## [1] 1.262626
```

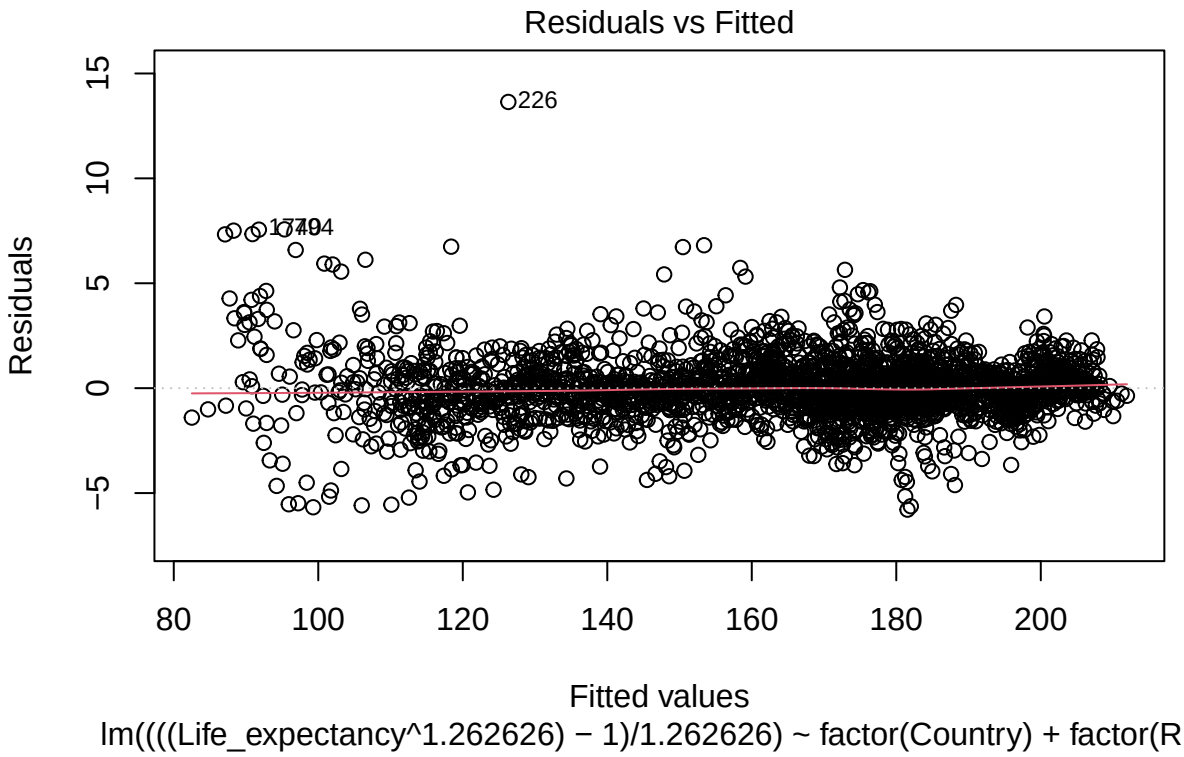
Based on our analysis, we received a best lambda of 1.262626 which we will use in a new model to test homoscedasticity.

```
##
##  studentized Breusch-Pagan test
##
## data:  LifeExpectancyFull_trans
## BP = 1172.7, df = 194, p-value < 2.2e-16
```

After conducting another Breusch-Pagan Test to confirm homoscedasticity in our new model with the best lambda,the test statistic (BP = 1172.7) and a very small p-value (p-value <2.2e-16) indicates that we reject the null hypothesis. This confirms that heteroscedasticity is still present even after the Box-Cox transformation.

Then we plot to see the residuals vs fitted of the transformed Box-Cox model:

Figure 9: Residuals vs. Fitted Plot of the Box-Cox Transformed model



The Box-Cox transformation has slightly improved the residual spread as it is more evenly spread out along the horizontal line but did not fully resolve heteroscedasticity as the p-value is still much lower than the significance level. This suggests that the variance of the errors in our final model may increase with the value of the response variable, Life_expectancy.

3.2.4 Normality Assumption

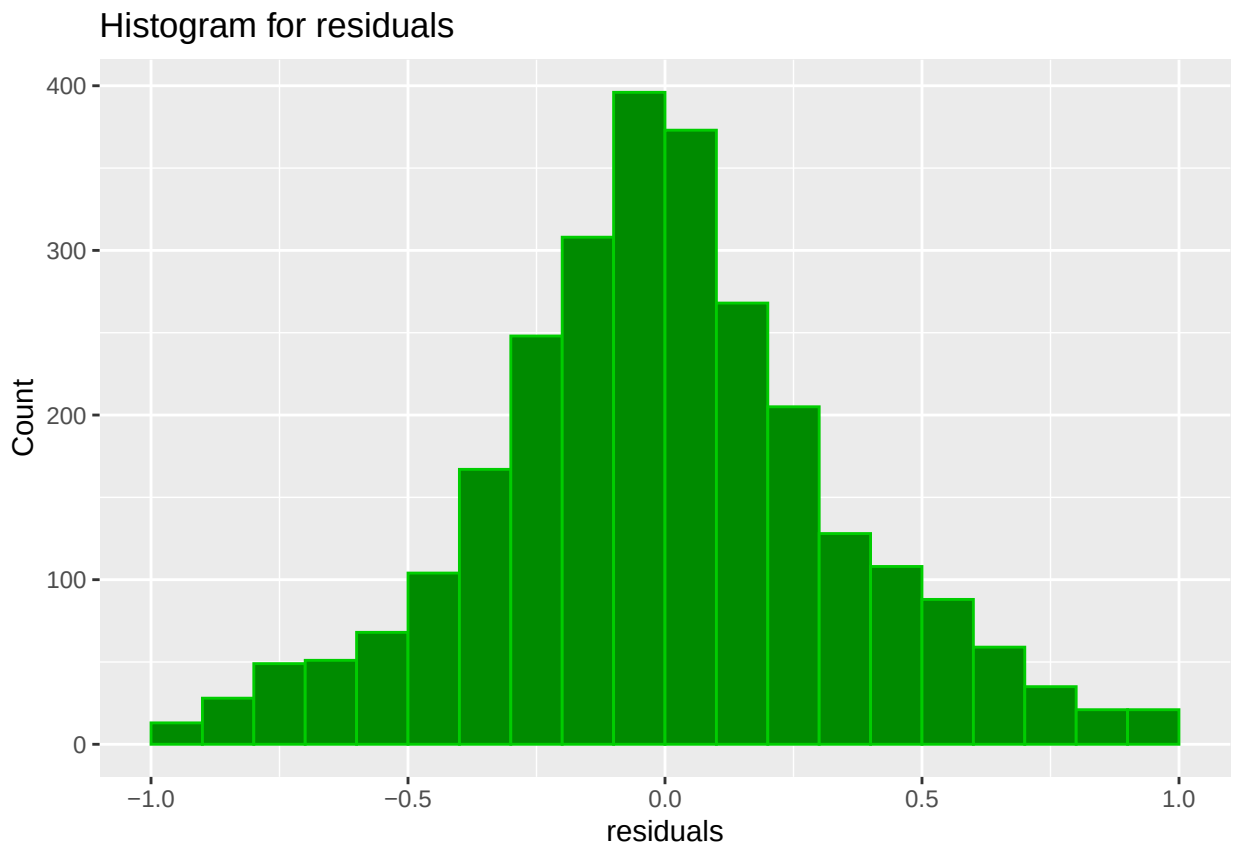
The normality assumption ensures that residuals follow a normal distribution, which is crucial for valid statistical inferences like p-values and confidence intervals in regression. Both a histogram and a Q-Q plot are used to assess this assumption. The histogram shows the overall residual distribution, while the Q-Q plot highlights deviations from normality, such as skewness or heavy tails. Together, they provide a thorough evaluation of residual normality, ensuring the model's reliability.

Hypothesis:

H_0: the sample data is significantly normally distributed

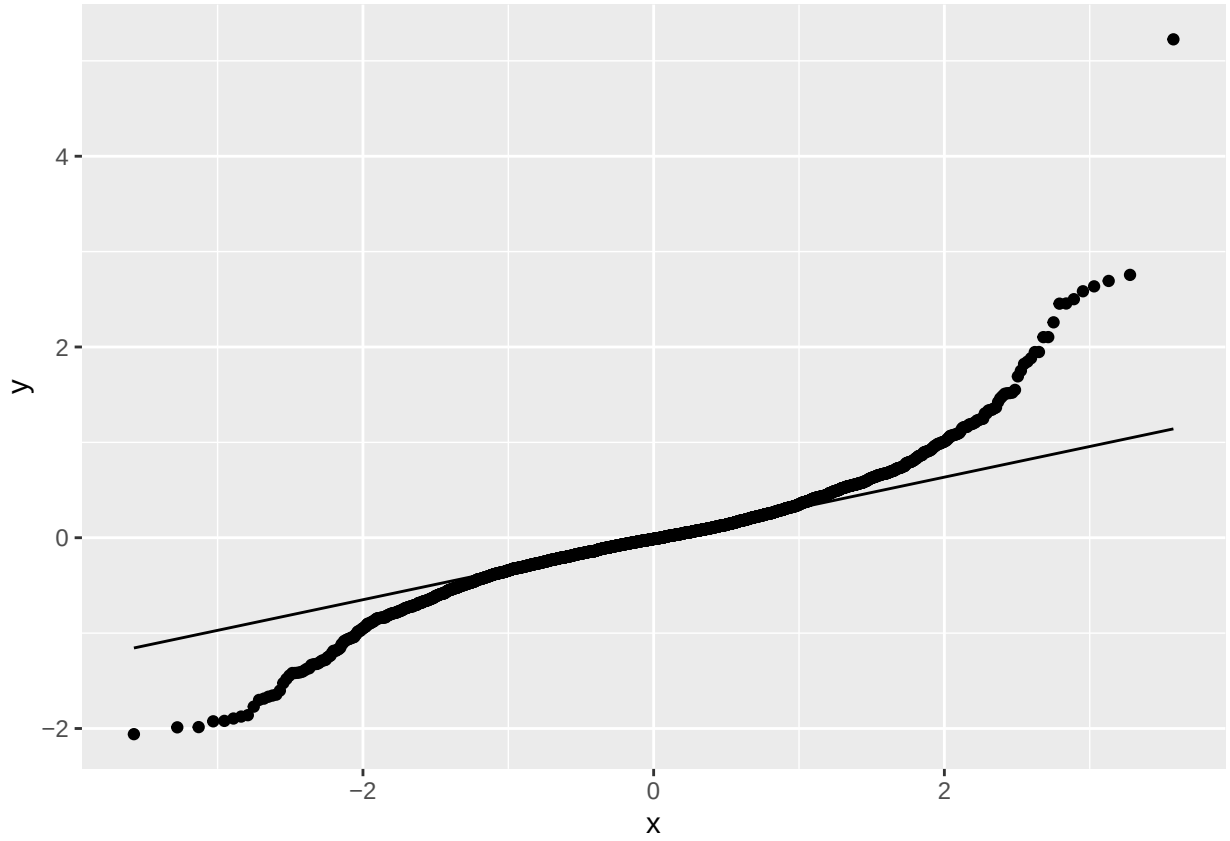
H_a: the sample data is not significantly normally distributed

Figure 10: Histogram of Residuals vs. Fitted for Normality Assumption



The histogram of residuals is approximately symmetric and shows a bell-shaped distribution. However, the tails appear slightly heavy, which could indicate deviations from perfect normality.

Figure 11: Q-Q plot of the Residuals vs. Fitted for Normality Assumption



The Q-Q plot reveals deviations from the straight line, particularly in the tails. This indicates that the residuals do not perfectly follow a normal distribution and may have heavier tails than expected.

We will conduct a formal test for normality using a Shapiro Test and report the p-value.

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(LifeExpectancyFull_old)
## W = 0.91582, p-value < 2.2e-16
```

The test shows a very small p-value (p-value $<$ 2.2e-16) which is less than the significance level $\alpha = 0.05$. Therefore, we reject the null hypothesis, indicating that it does not satisfy the requirement of normality.

We will again do the Shapiro Test, but this time for the log-transformed model and the Box-Cox model

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(LifeExpectancyFull_log)
## W = 0.82641, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(LifeExpectancyFull_trans)
## W = 0.93125, p-value < 2.2e-16
```

The test for log-transformed model shows avery small p-value (p-value < 2.2e-16) which is less than the significance level $\alpha = 0.05$. Therefore, we reject the null hypothesis, indicating that the log-transformed model also does not satisfy the requirement of normality.

Similarly, the test for Box-Cox model shows a a very small p-value (p-value < 2.2e-16) which is less than the significance level $\alpha = 0.05$. Again we reject the null hypothesis, indicating that it does not satisfy the requirement of normality.

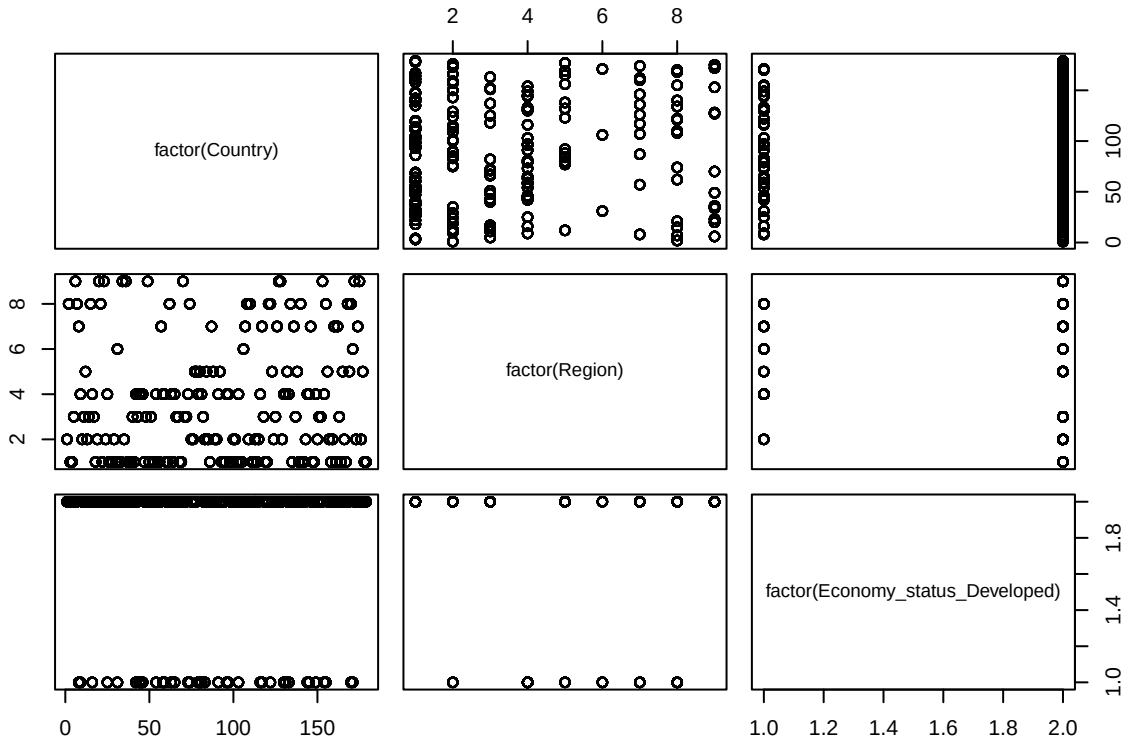
Overall, we reject the null hypothesis for all of the transformed models regarding normality assumption. This means that the residuals errors of the model do not follow a significantly normally distribution. This suggests that there may be a few data points on one or both ends of the model that deviate from the reference line significantly, which can be considered as outliers. Therefore, we should pay closer attention to it.

We will move on to multicollinearity to identify which variables are affected by multicollinearity and the strength of the correlation.

3.2.5 Multicollinearity

Two or more independent variables used in the model usually provide redundant information as the independent variables will be correlated with each other. From our data set, we suspected the variables 'Country','Region' and 'Economy_status_Developed' are correlated with each other and one or more of them are redundant as they all represent region information. When independent variables are linearly correlated, there is multicollinearity exists. We will first test the linearity of the three variables.

Figure 12: Multicollinearity Test between Independent Variables (Country, Region, Economy_status_Developed)



The scatter plot shows multicollinearity between the three variables. If the points in the scatter plots align vertically or horizontally, it suggests that one variable can be predicted by another and indicates

multicollinearity. Based on this, the points in those scatter plots are either aligned vertically or horizontally, which shows clustering. Therefore, we conclude that there is potential multicollinearity between Country and Region, Country and Economy_status_Developed, and Region and Economy_status_Developed.

We will conduct a Variance Inflation Factor (VIF) test to confirm:

```
##
## Call:
## imcdiag(mod = VIFmodel, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                                              VIF detection
## factor(Country)Albania                       Inf         1
## factor(Country)Algeria                       Inf         1
## factor(Country)Angola                        Inf         1
## factor(Country)Antigua and Barbuda           Inf         1
## factor(Country)Argentina                     Inf         1
## factor(Country)Armenia                       Inf         1
## factor(Country)Australia                     Inf         1
## factor(Country)Austria                       Inf         1
## factor(Country)Azerbaijan                 1.9888         0
## factor(Country)Bahamas, The                  Inf         1
## factor(Country)Bahrain                       Inf         1
## factor(Country)Bangladesh                 1.9888         0
## factor(Country)Barbados                      Inf         1
## factor(Country)Belarus                       Inf         1
## factor(Country)Belgium                       Inf         1
## factor(Country)Belize                        Inf         1
## factor(Country)Benin                         Inf         1
## factor(Country)Bhutan                     1.9888         0
## factor(Country)Bolivia                       Inf         1
## factor(Country)Bosnia and Herzegovina        Inf         1
## factor(Country)Botswana                      Inf         1
## factor(Country)Brazil                        Inf         1
## factor(Country)Brunei Darussalam          1.9888         0
## factor(Country)Bulgaria                      Inf         1
## factor(Country)Burkina Faso                  Inf         1
## factor(Country)Burundi                       Inf         1
## factor(Country)Cabo Verde                    Inf         1
## factor(Country)Cambodia                   1.9888         0
## factor(Country)Cameroon                      Inf         1
## factor(Country)Canada                        Inf         1
## factor(Country)Central African Republic      Inf         1
## factor(Country)Chad                          Inf         1
## factor(Country)Chile                         Inf         1
## factor(Country)China                      1.9888         0
## factor(Country)Colombia                      Inf         1
## factor(Country)Comoros                       Inf         1
## factor(Country)Congo, Dem. Rep.              Inf         1
## factor(Country)Congo, Rep.                   Inf         1
## factor(Country)Costa Rica                    Inf         1
## factor(Country)Cote d'Ivoire                 Inf         1
## factor(Country)Croatia                       Inf         1
```

```
## factor(Country)Cuba                          Inf         1
## factor(Country)Cyprus                        Inf         1
## factor(Country)Czechia                       Inf         1
## factor(Country)Denmark                       Inf         1
## factor(Country)Djibouti                      Inf         1
## factor(Country)Dominican Republic            Inf         1
## factor(Country)Ecuador                       Inf         1
## factor(Country)Egypt, Arab Rep.              Inf         1
## factor(Country)El Salvador                   Inf         1
## factor(Country)Equatorial Guinea             Inf         1
## factor(Country)Eritrea                       Inf         1
## factor(Country)Estonia                       Inf         1
## factor(Country)Eswatini                      Inf         1
## factor(Country)Ethiopia                      Inf         1
## factor(Country)Fiji                          Inf         1
## factor(Country)Finland                       Inf         1
## factor(Country)France                        Inf         1
## factor(Country)Gabon                         Inf         1
## factor(Country)Gambia, The                   Inf         1
## factor(Country)Georgia                       Inf         1
## factor(Country)Germany                       Inf         1
## factor(Country)Ghana                         Inf         1
## factor(Country)Greece                        Inf         1
## factor(Country)Grenada                       Inf         1
## factor(Country)Guatemala                     Inf         1
## factor(Country)Guinea                        Inf         1
## factor(Country)Guinea-Bissau                 Inf         1
## factor(Country)Guyana                        Inf         1
## factor(Country)Haiti                         Inf         1
## factor(Country)Honduras                      Inf         1
## factor(Country)Hungary                       Inf         1
## factor(Country)Iceland                       Inf         1
## factor(Country)India                      1.9888         0
## factor(Country)Indonesia                  1.9888         0
## factor(Country)Iran, Islamic Rep.            Inf         1
## factor(Country)Iraq                          Inf         1
## factor(Country)Ireland                       Inf         1
## factor(Country)Israel                        Inf         1
## factor(Country)Italy                         Inf         1
## factor(Country)Jamaica                       Inf         1
## factor(Country)Japan                         Inf         1
## factor(Country)Jordan                        Inf         1
## factor(Country)Kazakhstan                 1.9888         0
## factor(Country)Kenya                         Inf         1
## factor(Country)Kiribati                      Inf         1
## factor(Country)Kuwait                        Inf         1
## factor(Country)Kyrgyz Republic            1.9888         0
## factor(Country)Lao PDR                    1.9888         0
## factor(Country)Latvia                        Inf         1
## factor(Country)Lebanon                       Inf         1
## factor(Country)Lesotho                       Inf         1
## factor(Country)Liberia                       Inf         1
## factor(Country)Libya                         Inf         1
## factor(Country)Lithuania                     Inf         1
```

```
## factor(Country)Luxembourg                    Inf         1
## factor(Country)Madagascar                    Inf         1
## factor(Country)Malawi                        Inf         1
## factor(Country)Malaysia                   1.9888         0
## factor(Country)Maldives                   1.9888         0
## factor(Country)Mali                          Inf         1
## factor(Country)Malta                         Inf         1
## factor(Country)Mauritania                    Inf         1
## factor(Country)Mauritius                     Inf         1
## factor(Country)Mexico                        Inf         1
## factor(Country)Micronesia, Fed. Sts.         Inf         1
## factor(Country)Moldova                       Inf         1
## factor(Country)Mongolia                   1.9888         0
## factor(Country)Montenegro                    Inf         1
## factor(Country)Morocco                       Inf         1
## factor(Country)Mozambique                    Inf         1
## factor(Country)Myanmar                    1.9888         0
## factor(Country)Namibia                       Inf         1
## factor(Country)Nepal                      1.9888         0
## factor(Country)Netherlands                   Inf         1
## factor(Country)New Zealand                   Inf         1
## factor(Country)Nicaragua                     Inf         1
## factor(Country)Niger                         Inf         1
## factor(Country)Nigeria                       Inf         1
## factor(Country)North Macedonia               Inf         1
## factor(Country)Norway                        Inf         1
## factor(Country)Oman                          Inf         1
## factor(Country)Pakistan                   1.9888         0
## factor(Country)Panama                        Inf         1
## factor(Country)Papua New Guinea              Inf         1
## factor(Country)Paraguay                      Inf         1
## factor(Country)Peru                          Inf         1
## factor(Country)Philippines                1.9888         0
## factor(Country)Poland                        Inf         1
## factor(Country)Portugal                      Inf         1
## factor(Country)Qatar                         Inf         1
## factor(Country)Romania                       Inf         1
## factor(Country)Russian Federation            Inf         1
## factor(Country)Rwanda                        Inf         1
## factor(Country)Samoa                         Inf         1
## factor(Country)Sao Tome and Principe         Inf         1
## factor(Country)Saudi Arabia                  Inf         1
## factor(Country)Senegal                       Inf         1
## factor(Country)Serbia                        Inf         1
## factor(Country)Seychelles                    Inf         1
## factor(Country)Sierra Leone                  Inf         1
## factor(Country)Singapore                  1.9888         0
## factor(Country)Slovak Republic               Inf         1
## factor(Country)Slovenia                      Inf         1
## factor(Country)Solomon Islands               Inf         1
## factor(Country)Somalia                       Inf         1
## factor(Country)South Africa                  Inf         1
## factor(Country)Spain                         Inf         1
## factor(Country)Sri Lanka                  1.9888         0
```

```
## factor(Country)St. Lucia                          Inf         1
## factor(Country)St. Vincent and the Grenadines     Inf         1
## factor(Country)Suriname                           Inf         1
## factor(Country)Sweden                             Inf         1
## factor(Country)Switzerland                        Inf         1
## factor(Country)Syrian Arab Republic               Inf         1
## factor(Country)Tajikistan                      1.9888         0
## factor(Country)Tanzania                           Inf         1
## factor(Country)Thailand                        1.9888         0
## factor(Country)Timor-Leste                        Inf         1
## factor(Country)Togo                               Inf         1
## factor(Country)Tonga                              Inf         1
## factor(Country)Trinidad and Tobago                Inf         1
## factor(Country)Tunisia                            Inf         1
## factor(Country)Turkiye                            Inf         1
## factor(Country)Turkmenistan                    1.9888         0
## factor(Country)Uganda                             Inf         1
## factor(Country)Ukraine                            Inf         1
## factor(Country)United Arab Emirates               Inf         1
## factor(Country)United Kingdom                     Inf         1
## factor(Country)United States                      Inf         1
## factor(Country)Uruguay                            Inf         1
## factor(Country)Uzbekistan                      1.9888         0
## factor(Country)Vanuatu                            Inf         1
## factor(Country)Venezuela, RB                      Inf         1
## factor(Country)Vietnam                         1.9888         0
## factor(Country)Yemen, Rep.                        Inf         1
## factor(Country)Zambia                             Inf         1
## factor(Country)Zimbabwe                           Inf         1
## factor(Region)Asia                                Inf         1
## factor(Region)Central America and Caribbean       Inf         1
## factor(Region)European Union                      Inf         1
## factor(Region)Middle East                         Inf         1
## factor(Region)North America                       Inf         1
## factor(Region)Oceania                             Inf         1
## factor(Region)Rest of Europe                      Inf         1
## factor(Region)South America                       Inf         1
## factor(Economy_status_Developed)Developing        Inf         1
##
## Multicollinearity may be due to factor(Country)Albania factor(Country)Algeria factor(Country)Angola :
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

A value of VIF = 1 indicates that there is no collinearity between the independent variable and others. A VIF value between 1 and 5 suggests moderate collinearity, and a VIF value of above 5 represents critical level of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable. As the value of VIF are all above 5 for three variables, this suggest that there are critical levels of multicollinearity in the data. In order to resolve multicollinearity, we will drop 2 out of these 3 variables, and keep only the most relevant one to our research study. Since our research question specifically focuses on evaluating life expectancy between developed and developing economies, we decide to drop 'Country' and 'Region' and keep 'Economy_status_Developed' only, to avoid redundancy.

Therefore we will revise the full model and remove the redundant variables (Country and Region) that we have confirmed have high multicollinearity in the previous steps.

The revised full model is shown below:

```
##
## Call:
## lm(formula = Life_expectancy ~ Year + Infant_deaths + Under_five_deaths +
##     Adult_mortality + Alcohol_consumption + Hepatitis_B + Measles +
##     BMI + Polio + Diphtheria + Incidents_HIV + GDP_per_capita +
##     Population_mln + Thinness_ten_nineteen_years + Thinness_five_nine_years +
##     Schooling + Economy_status_Developed, data = Life_Expectancy_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8770 -0.9210 -0.0464  0.8676  8.0118
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.613e+01  1.141e+01   3.168  0.00155 **
## Year                              2.451e-02  5.693e-03   4.305 1.73e-05 ***
## Infant_deaths                    -4.738e-02  6.162e-03  -7.689 2.03e-14 ***
## Under_five_deaths                -5.634e-02  3.830e-03 -14.711  < 2e-16 ***
## Adult_mortality                  -4.820e-02  6.189e-04 -77.878  < 2e-16 ***
## Alcohol_consumption               6.617e-02  9.895e-03   6.688 2.72e-11 ***
## Hepatitis_B                      -8.807e-03  2.564e-03  -3.435  0.00060 ***
## Measles                           1.598e-03  1.713e-03   0.933  0.35109
## BMI                              -1.494e-01  1.911e-02  -7.818 7.51e-15 ***
## Polio                             2.990e-03  5.830e-03   0.513  0.60809
## Diphtheria                        7.345e-04  5.886e-03   0.125  0.90070
## Incidents_HIV                     9.486e-02  1.813e-02   5.232 1.80e-07 ***
## GDP_per_capita                    2.576e-05  2.283e-06  11.282  < 2e-16 ***
## Population_mln                   -1.443e-04  1.994e-04  -0.724  0.46930
## Thinness_ten_nineteen_years      -3.579e-02  1.710e-02  -2.093  0.03643 *
## Thinness_five_nine_years          2.831e-03  1.677e-02   0.169  0.86597
## Schooling                         8.839e-02  1.684e-02   5.250 1.63e-07 ***
## Economy_status_DevelopedDeveloping -6.622e-01  1.079e-01  -6.135 9.72e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.357 on 2846 degrees of freedom
## Multiple R-squared:  0.9793, Adjusted R-squared:  0.9792
## F-statistic:  7925 on 17 and 2846 DF,  p-value: < 2.2e-16
```

3.2.6 Outliers

In this step we are identifying outliers in the Life_expectancy_data. This step is to ensure the outliers will not affect the model:

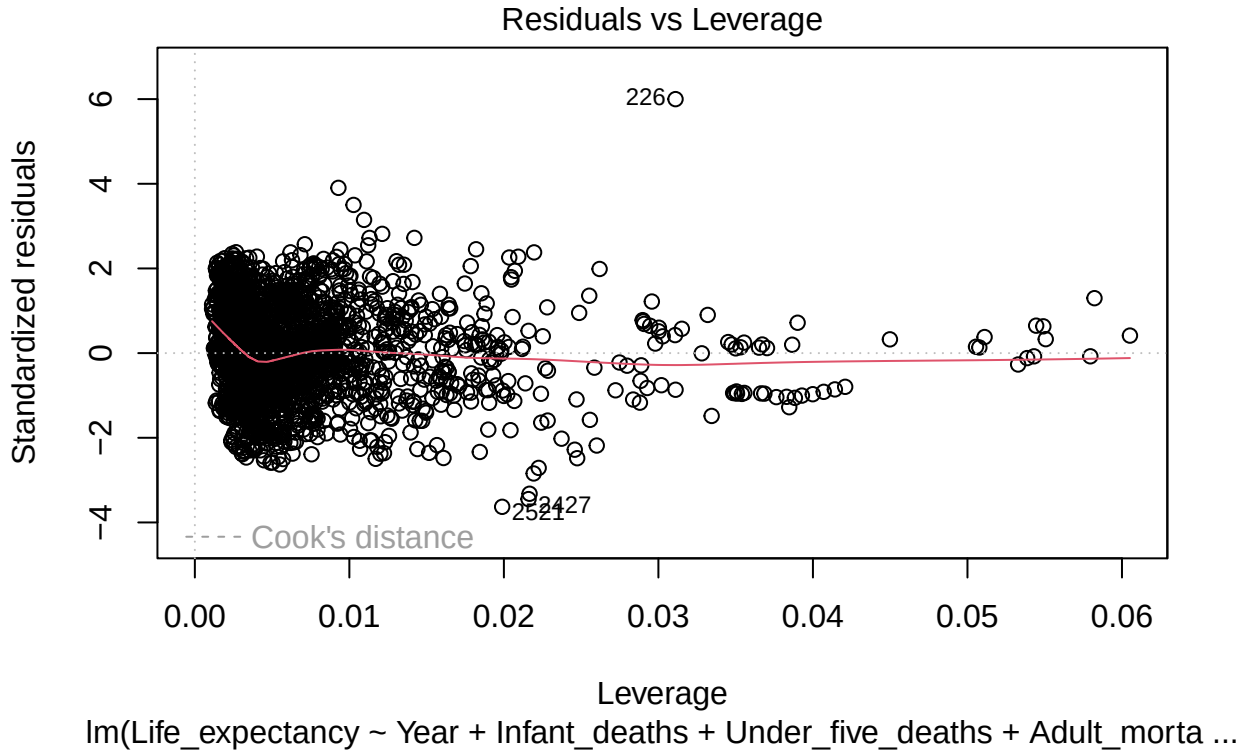
1. Error in recording data

2. Measurement process/tool problem

3. failure of the experimental process

We will use the Residuals vs Leverage plot to help us find influential cases, if any. Outliers that reside within Cook's Distance, meaning the upper right corner or lower right corner of the plot, require special attention is these points can have influence against the regression line.

Figure 13: Residuals vs Leverage Plot of the Revised Full Model



The graph shows a few points with higher leverage, but none of the observations exceeds the Cook's Distance threshold, and no values appear in the upper and lower corner. This suggests that there are no extremely influential outliers that would distort the model excessively.

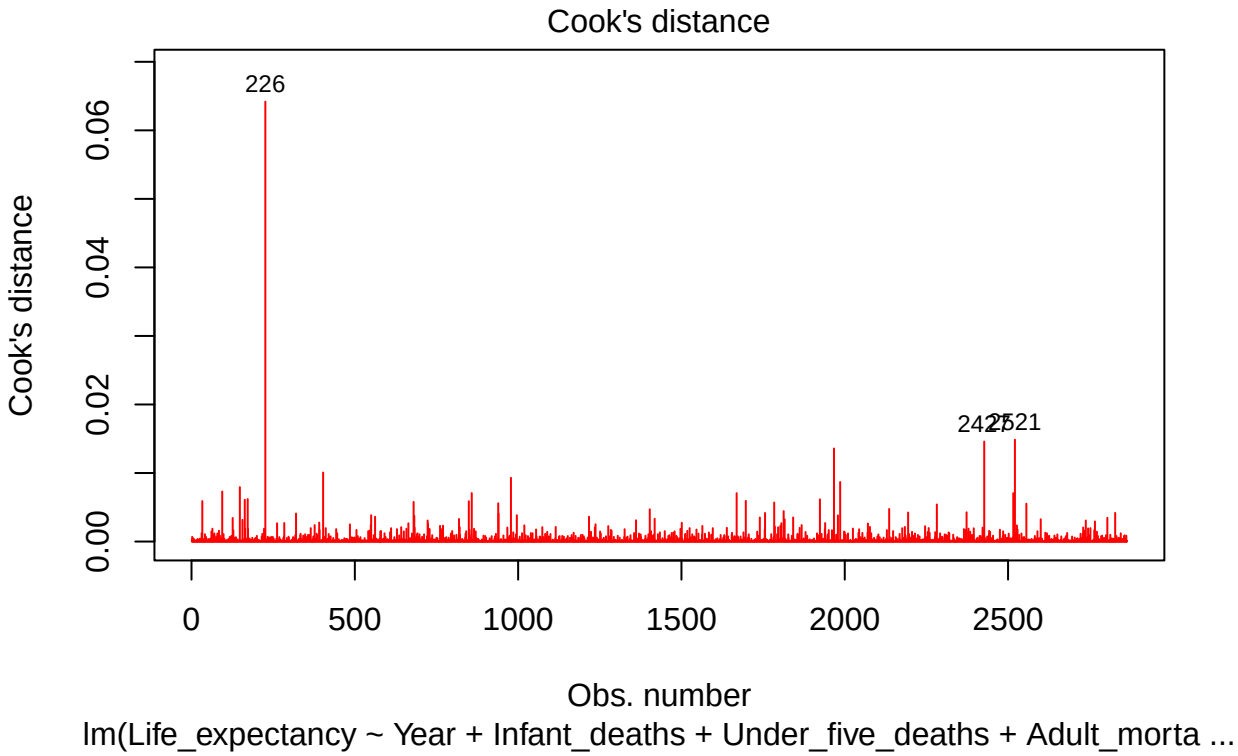
We will further conduct a Cook's Distance formal test to measure the overall influence an outlying observation has on the estimated coefficients. The value that exceeds 0.5 will be considered "too large" and could influence the accuracy of our analysis.

```
##  [1] Country                    Region
##  [3] Year                       Infant_deaths
##  [5] Under_five_deaths          Adult_mortality
##  [7] Alcohol_consumption        Hepatitis_B
##  [9] Measles                    BMI
## [11] Polio                      Diphtheria
## [13] Incidents_HIV              GDP_per_capita
## [15] Population_mln             Thinness_ten_nineteen_years
## [17] Thinness_five_nine_years   Schooling
## [19] Economy_status_Developed   Life_expectancy
```

```
## [21] residuals
## <0 rows> (or 0-length row.names)
```

The Cook's Distance Test did not a show any data points that could possibly influence our model. We will plot the Cook's Distance to observe the behavior of our data points.

Figure 14: Cook's Distance plot to identify outliers in the model.



The plot shows that observations with the highest Cook's distances are well below 0.5. This means that while some observations have higher leverage, therefore, they do not pose a substantial risk to model validity.

Based on the outlier analysis using the Residuals vs Leverage plot, and the Cook's Distance plot, we can conclude that there are no significant outliers or influential points that exceed the commonly accepted thresholds. While a few observations have higher leverage values, they do not pose a substantial risk to the model's validity. Therefore, we consider further processing is not require at this stage.

In summary, our data has passed the Linearity Assumption, Independence Assumption, and the test for outliers. We have also handled multicollinearity among our variables by dropping Region and Country from the model. Unfortunately, we did not pass the Equal Variance and Normality assumptions. This indicates that when we report our final model, it should be interpreted with careful consideration as the analysis may not be reliable due to the presence of errors.

3.3 Additive Modeling

After conducting the assumptions tests, we consider our revised full model (without Country or Region) as a valid model to proceed to the next steps.

As we have many variables in our model, we want to limit it for our study and determine if the following variables are significant predictors of the response variable Life_expectancy:

- Adult Mortality
- Alcohol Consumption
- Hepatitis B
- HIV Incidents
- Population (millions)
- Thinness in children 10-19 years
- GDP per capita
- Schooling
- Economy Status

3.3.1 Full Model Test

We will first check to see if any of the variation in the dependent variable is explained by at least one of the independent variables mentioned above by performing a Full Model Test using ANOVA.

$H_0 : \beta_i = 0 \ \forall i$

$H_a : \beta_i \neq 0$ for at least one i

$\alpha = 0.05$

```
##
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Alcohol_consumption +
##     Hepatitis_B + Incidents_HIV + GDP_per_capita + Population_mln +
##     Thinness_ten_nineteen_years + Schooling + factor(Economy_status_Developed),
##     data = Life_Expectancy_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2937 -1.2751  0.0658  1.2665  9.1230
##
## Coefficients:
##                                             Estimate   Std. Error  t value
## (Intercept)                                75.022149814  0.363594111  206.335
## Adult_mortality                            -0.070980560  0.000665201 -106.705
## Alcohol_consumption                         0.138568109  0.014334054    9.667
## Hepatitis_B                                 0.028983638  0.002660646   10.893
## Incidents_HIV                               0.394838510  0.024657167   16.013
## GDP_per_capita                              0.000007476  0.000003295    2.269
## Population_mln                              0.000066478  0.000291484    0.228
## Thinness_ten_nineteen_years                -0.020644244  0.010993249   -1.878
## Schooling                                   0.534863222  0.020520428   26.065
## factor(Economy_status_Developed)Developing -0.068792430  0.153391072   -0.448
##                                                    Pr(>|t|)
## (Intercept)                                <0.0000000000000002 ***
## Adult_mortality                            <0.0000000000000002 ***
## Alcohol_consumption                        <0.0000000000000002 ***
## Hepatitis_B                                <0.0000000000000002 ***
## Incidents_HIV                              <0.0000000000000002 ***
## GDP_per_capita                                        0.0233 *
```

```
## Population_mln                                          0.8196
## Thinness_ten_nineteen_years                             0.0605 .
## Schooling                                  <0.0000000000000002 ***
## factor(Economy_status_Developed)Developing              0.6538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.007 on 2854 degrees of freedom
## Multiple R-squared:  0.9546, Adjusted R-squared:  0.9545
## F-statistic:  6670 on 9 and 2854 DF,  p-value: < 0.00000000000000022
```

```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ 1
## Model 2: Life_expectancy ~ Adult_mortality + Alcohol_consumption + Hepatitis_B +
##     Incidents_HIV + GDP_per_capita + Population_mln + Thinness_ten_nineteen_years +
##     Schooling + factor(Economy_status_Developed)
##   Res.Df    RSS Df Sum of Sq      F               Pr(>F)
## 1   2863 253277
## 2   2854  11494  9    241782 6670.3 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The significance level we agreed on is 0.05 when analyzing the p-value of the variables. According to the output, the overall significance F-statistic is *6670* and the p-value is *2.2e-16*, which is smaller than $\alpha = 0.05$, so we reject the null hypothesis and we have evidence to conclude that at least one of the independent variables is not equal to 0. In summary the large F stat and small p-value suggests that at least one of the variables must be related to the response variable life expectancy.

The full model equation is as follows:

$\hat{LifeExpectancy} = 75.02215 - 0.07098_{AdultMortality} + 0.13857_{AlcholConsumption} + 0.02898_{HepatitisB} + 0.39484_{IncidentsHIV} + 0.000007_{GDP} + 0.00007_{Population} - 0.02064_{Thinness} + 0.53486_{Schooling} - 0.06879_{EconomyStatus}$

3.3.1 Strategy for Model Selection: Partial Model Test (t-test and F-test)

We will use individual t-tests to determine if the independent variables are significant at $\alpha = 0.05$.

Our hypothesis is:

$H_0 : \beta_i = 0$

$H_a : \beta_i \neq 0$ $(i = 1, 2, ..., p)$

Where p is the number of independent variables and interaction terms in the regression model.

According to the summary of the revised full model, the p-value for Population_mln is 0.8196, Thinness_ten_nineteen_years is 0.0605, and factor(Economy_Status_Developed) is 0.6538, which are all greater than $\alpha = 0.05$. These are considered not significant in the full model at $\alpha = 0.05$ level. Therefore, we fail to reject the null hypothesis and conclude that Population_mln, Thinness_ten_nineteen_years, and factor(Economy_Status_Developed) have no significant impact on Life_Expectancy at $\alpha = 0.05$.

The following is our reduced model:

```
##
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Alcohol_consumption +
```

```
##     Hepatitis_B + Incidents_HIV + GDP_per_capita + Schooling,
##     data = Life_Expectancy_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.134 -1.274  0.052  1.265  9.276
##
## Coefficients:
##                          Estimate   Std. Error  t value             Pr(>|t|)
## (Intercept)         74.756336982  0.315965707  236.596 < 0.0000000000000002 ***
## Adult_mortality     -0.071041085  0.000658908 -107.816 < 0.0000000000000002 ***
## Alcohol_consumption  0.145474510  0.012390886   11.740 < 0.0000000000000002 ***
## Hepatitis_B          0.028868972  0.002626361   10.992 < 0.0000000000000002 ***
## Incidents_HIV        0.391941557  0.024610495   15.926 < 0.0000000000000002 ***
## GDP_per_capita       0.000008248  0.000002961    2.786              0.00537 **
## Schooling            0.547299103  0.019518662   28.040 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.007 on 2857 degrees of freedom
## Multiple R-squared:  0.9546, Adjusted R-squared:  0.9545
## F-statistic: 1e+04 on 6 and 2857 DF,  p-value: < 0.00000000000000022
```

The reduced model is as follows:

$\hat{LifeExpectancy} = 74.75634 - 0.07104X_{AdultMortality} + 0.14547X_{AlcoholConsumption} + 0.02887X_{HepatitisB} + 0.39194X_{IncidentsHIV} + 0.000008_{GDP} + 0.54730Schooling$

According to the output of the reduced model, the overall significance F-statistic is *10000* and the p-value is *2.2e-16*, which is smaller than $\alpha = 0.05$. The adjusted R-squared of the reduced model is 0.9545.

The p-values of all the independent variables in the reduced model are less than 0.05, which suggests that they are all statistically significant and that there are no variables to remove from the model using the t-test.

Partial F-test

To determine whether the reduced additive model is a better fit than the full additive model, we will perform a Partial F-test using the ANOVA function in R with the significance level as $\alpha = 0.05$

Our hypothesis for the Partial F test is:

$H_0 : \beta_{thiness} = \beta_{population} = \beta_{EconomyStatusDeveloped} = 0$

H_a : at least one of $\beta_{thiness}, \beta_{population}, \beta_{EconomyStatusDeveloped}$ does not equal zero

```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ Adult_mortality + Alcohol_consumption + Hepatitis_B +
##     Incidents_HIV + GDP_per_capita + Schooling
## Model 2: Life_expectancy ~ Adult_mortality + Alcohol_consumption + Hepatitis_B +
##     Incidents_HIV + GDP_per_capita + Population_mln + Thinness_ten_nineteen_years +
##     Schooling + factor(Economy_status_Developed)
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1   2857 11510
## 2   2854 11494  3    15.627 1.2933 0.2749
```

Based on the ANOVA test, the p-value of the reduced model is 0.2749. Since it is larger than $\alpha = 0.05$, we fail to reject the null hypothesis which means that we should drop the variables Population, Thinness, and

Economy_status_Developed from the full model. Therefore, we accept the reduced additive model as our final additive model.

3.3.2 Strategy for Model Selection: Stepwise Regression Procedure

We will apply the Stepwise Regression Procedure with p_enter=0.05 and p_remove=0.1 to the data to find the independent variables most suitable for modeling life expectancy. Previously we used individual t-tests to determine which variables to remove from the full model. We will apply the Stepwise Regression Procedure to see if we will yield similar results.

Stepwise Regression Procedure:

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.134 -1.274  0.052  1.265  9.276
##
## Coefficients:
##                          Estimate   Std. Error  t value             Pr(>|t|)
## (Intercept)         74.756336982  0.315965707  236.596 < 0.0000000000000002 ***
## Adult_mortality     -0.071041085  0.000658908 -107.816 < 0.0000000000000002 ***
## Schooling            0.547299103  0.019518662   28.040 < 0.0000000000000002 ***
## Incidents_HIV        0.391941557  0.024610495   15.926 < 0.0000000000000002 ***
## Alcohol_consumption  0.145474510  0.012390886   11.740 < 0.0000000000000002 ***
## Hepatitis_B          0.028868972  0.002626361   10.992 < 0.0000000000000002 ***
## GDP_per_capita       0.000008248  0.000002961    2.786              0.00537 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.007 on 2857 degrees of freedom
## Multiple R-squared:  0.9546, Adjusted R-squared:  0.9545
## F-statistic: 1e+04 on 6 and 2857 DF,  p-value: < 0.00000000000000022
```

Given the parameters p_enter=0.05 and p_remove=0.1, the variables whose p-values are less than 0.05 are:
- Adult_mortality
- Schooling
- Incidents_HIV
- Alcohol_consumption
- Hepatitis_B
- GDP_per_capita

Therefore, these 6 variables are what we will include in the final Stepwise Regression model. The p-value for the stepwise model is <2.2e-16 and the Adjusted R-squared is 0.9545.

The equation of the step wise model is as follows:

$\hat{LifeExpectancy} = 74.75634 - 0.07104X_{AdultMortality} + 0.14547X_{AlcoholConsumption} + 0.02887X_{HepatitisB} + 0.39194X_{IncidentsHIV} + 0.000008_{GDP} + 0.54730Schooling$

Similarly, we will also perform a Forward Regression model to see how it compares to the Stepwise Regression model and the individual t-tests we performed.

3.3.3 Strategy for Model Selection: Forward Regression Procedure

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.134 -1.274  0.052  1.265  9.276
##
## Coefficients:
##                         Estimate   Std. Error  t value             Pr(>|t|)
## (Intercept)         74.756336982  0.315965707  236.596 < 0.0000000000000002 ***
## Adult_mortality     -0.071041085  0.000658908 -107.816 < 0.0000000000000002 ***
## Schooling            0.547299103  0.019518662   28.040 < 0.0000000000000002 ***
## Incidents_HIV        0.391941557  0.024610495   15.926 < 0.0000000000000002 ***
## Alcohol_consumption  0.145474510  0.012390886   11.740 < 0.0000000000000002 ***
## Hepatitis_B          0.028868972  0.002626361   10.992 < 0.0000000000000002 ***
## GDP_per_capita       0.000008248  0.000002961    2.786              0.00537 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.007 on 2857 degrees of freedom
## Multiple R-squared:  0.9546, Adjusted R-squared:  0.9545
## F-statistic: 1e+04 on 6 and 2857 DF,  p-value: < 0.00000000000000022
```

Given the parameters p_enter=0.05, the variables whose p-values are less than 0.05 are:

- Adult_mortality
- Schooling
- Incidents_HIV
- Alcohol_consumption
- Hepatitis_B
- GDP_per_capita

Therefore, these 6 variables are what we will include in the final forward Regression model. The p-value for the forward model is <2.2e-16 and the Adjusted R-squared is 0.9545.

The equation of the forward regression model is as follows:

$\hat{LifeExpectancy} = 74.75634 - 0.07104X_{AdultMortality} + 0.14547X_{AlcoholConsumption} + 0.02887X_{HepatitisB} + 0.39194X_{IncidentsHIV} + 0.000008_{GDP} + 0.54730Schooling$

3.3.4 Strategy for Model Selection: Backward Regression Procedure

```
##
## Call:
```

```
## lm(formula = paste(response, "~", paste(c(include, cterms), collapse = " + ")),
##     data = l)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.134 -1.274  0.052  1.265  9.276
##
## Coefficients:
##                          Estimate   Std. Error  t value             Pr(>|t|)
## (Intercept)         74.756336982  0.315965707  236.596 < 0.0000000000000002 ***
## Adult_mortality     -0.071041085  0.000658908 -107.816 < 0.0000000000000002 ***
## Alcohol_consumption  0.145474510  0.012390886   11.740 < 0.0000000000000002 ***
## Hepatitis_B          0.028868972  0.002626361   10.992 < 0.0000000000000002 ***
## Incidents_HIV        0.391941557  0.024610495   15.926 < 0.0000000000000002 ***
## GDP_per_capita       0.000008248  0.000002961    2.786              0.00537 **
## Schooling            0.547299103  0.019518662   28.040 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.007 on 2857 degrees of freedom
## Multiple R-squared:  0.9546, Adjusted R-squared:  0.9545
## F-statistic: 1e+04 on 6 and 2857 DF,  p-value: < 0.00000000000000022
```

Given the parameters p_enter=0.05, the variables whose p-values are less than 0.05 are:

- Adult_mortality
- Schooling
- Incidents_HIV
- Alcohol_consumption
- Hepatitis_B
- GDP_per_capita

Therefore, these 6 variables are what we will include in the final backward Regression model. The p-value for the forward model is <2.2e-16 and the Adjusted R-squared is 0.9545.

The equation of the backward regression model is as follows:

$\hat{LifeExpectancy} = 74.75634 - 0.07104X_{AdultMortality} + 0.14547X_{AlcoholConsumption} + 0.02887X_{HepatitisB} + 0.39194X_{IncidentsHIV} + 0.000008_{GDP} + 0.54730Schooling$

3.3.4 Selecting Final Additive Model

We yield the same models from t-test, Stepwise, Forward and Backward regression procedure as a result we get the same coefficients, R^2_{adj} and RSE from these procedures. So, our final additive model is:

$\hat{LifeExpectancy} = 74.75634 - 0.07104X_{AdultMortality} + 0.14547X_{AlcoholConsumption} + 0.02887X_{HepatitisB} + 0.39194X_{IncidentsHIV} + 0.000008_{GDP} + 0.54730Schooling$

3.4 Interaction Model

We want to understand if the reduced model will perform better with interactions. First we will add interaction to all *6* variables from the final additive model:

```
##
## Call:
## lm(formula = Life_expectancy ~ (Adult_mortality + Alcohol_consumption +
##     Hepatitis_B + Incidents_HIV + GDP_per_capita + Schooling)^2,
##     data = Life_Expectancy_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3803 -1.0950  0.0224  1.1473  6.3083
##
## Coefficients:
##                                       Estimate    Std. Error t value
## (Intercept)                         69.4723697649  1.0882505517  63.839
## Adult_mortality                     -0.0595604709  0.0029824412 -19.970
## Alcohol_consumption                  0.3639582659  0.1032929046   3.524
## Hepatitis_B                          0.1193236696  0.0129165425   9.238
## Incidents_HIV                       -1.0552400243  0.2016315216  -5.234
## GDP_per_capita                       0.0002004140  0.0000321711   6.230
## Schooling                            0.4779182620  0.1095862423   4.361
## Adult_mortality:Alcohol_consumption -0.0014136828  0.0002076919  -6.807
## Adult_mortality:Hepatitis_B         -0.0002257136  0.0000376857  -5.989
## Adult_mortality:Incidents_HIV        0.0020296698  0.0001646765  12.325
## Adult_mortality:GDP_per_capita      -0.0000001199  0.0000001051  -1.140
## Adult_mortality:Schooling            0.0021832778  0.0002550410   8.560
## Alcohol_consumption:Hepatitis_B      0.0024411712  0.0008504112   2.871
## Alcohol_consumption:Incidents_HIV   -0.0098206840  0.0117788499  -0.834
## Alcohol_consumption:GDP_per_capita  -0.0000044483  0.0000008062  -5.518
## Alcohol_consumption:Schooling       -0.0099100208  0.0044464311  -2.229
## Hepatitis_B:Incidents_HIV            0.0001154692  0.0019075135   0.061
## Hepatitis_B:GDP_per_capita          -0.0000019379  0.0000002629  -7.371
## Hepatitis_B:Schooling               -0.0054357754  0.0011181725  -4.861
## Incidents_HIV:GDP_per_capita         0.0000037767  0.0000113925   0.332
## Incidents_HIV:Schooling              0.0628413689  0.0127130382   4.943
## GDP_per_capita:Schooling             0.0000032710  0.0000017303   1.890
##                                                Pr(>|t|)
## (Intercept)                         < 0.0000000000000002 ***
## Adult_mortality                     < 0.0000000000000002 ***
## Alcohol_consumption                                0.000433 ***
## Hepatitis_B                         < 0.0000000000000002 ***
## Incidents_HIV                          0.00000017843652 ***
## GDP_per_capita                         0.00000000053673 ***
## Schooling                              0.00001340039112 ***
## Adult_mortality:Alcohol_consumption    0.00000000001214 ***
## Adult_mortality:Hepatitis_B            0.00000000237139 ***
## Adult_mortality:Incidents_HIV       < 0.0000000000000002 ***
## Adult_mortality:GDP_per_capita                     0.254258
## Adult_mortality:Schooling           < 0.0000000000000002 ***
## Alcohol_consumption:Hepatitis_B                    0.004128 **
## Alcohol_consumption:Incidents_HIV                  0.404489
```

```
## Alcohol_consumption:GDP_per_capita      0.00000003742488 ***
## Alcohol_consumption:Schooling                  0.025908 *
## Hepatitis_B:Incidents_HIV                      0.951735
## Hepatitis_B:GDP_per_capita              0.00000000000022 ***
## Hepatitis_B:Schooling                   0.00000122954490 ***
## Incidents_HIV:GDP_per_capita                   0.740288
## Incidents_HIV:Schooling                 0.00000081368667 ***
## GDP_per_capita:Schooling                       0.058796 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.816 on 2842 degrees of freedom
## Multiple R-squared:  0.963,  Adjusted R-squared:  0.9627
## F-statistic:  3520 on 21 and 2842 DF,  p-value: < 0.00000000000000022
```

We will use individual t-tests for each predictor and interaction terms to determine which ones are significant at $\alpha = 0.05$.

$H_0 : \beta_i = 0$

$H_a : \beta_i \neq 0$ $(i = 1, 2, ..., p)$

Where p is the number of independent variables and interaction terms in the regression model.

From the above output we observe that the predictors and interaction terms that have a p-value smaller than $\alpha = 0.05$ are the following: Adult_mortality, Alcohol_consumption, Hepatitis_B, Incidents_HIV, GDP_per_capita, Schooling , Adult_mortality:Alcohol_consumption, Adult_mortality:Hepatitis_B, Adult_mortality:Schooling, Adult_mortality:Incidents_HIV Alcohol_consumption:GDP_per_capita, Alcohol_consumption:Schooling, Alcohol_consumption:Hepatitis_B, Hepatitis_B:GDP_per_capita, Hepatitis_B:Schooling, Incidents_HIV:Schooling. For these terms we reject the null hypothesis, and conclude that they are statistically significant and should be kept in the model.

Alternatively, the interaction terms Adult_mortality:GDP_per_capita, Hepatitis_B:Incidents_HIV, Incidents_HIV:GDP_per_capita and GDP_per_capita:Schooling all have a p-value which is greater than $\alpha = 0.05$. Hence, we fail to reject the null hypothesis for these terms, which means that these terms are statistically insignificant. As a result we will remove these terms from the model.

Removing the insignificant interactions the reduced interaction model is as follows:

```
##
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Alcohol_consumption +
##     Hepatitis_B + Incidents_HIV + GDP_per_capita + Schooling +
##     Adult_mortality * Alcohol_consumption + Adult_mortality *
##     Hepatitis_B + Adult_mortality * Schooling + Adult_mortality *
##     Incidents_HIV + Alcohol_consumption * GDP_per_capita + Alcohol_consumption *
##     Schooling + Alcohol_consumption * Hepatitis_B + Hepatitis_B *
##     GDP_per_capita + Hepatitis_B * Schooling + Incidents_HIV *
##     Schooling, data = Life_Expectancy_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4283 -1.0851  0.0127  1.1380  6.2843
##
## Coefficients:
##                                       Estimate    Std. Error t value
## (Intercept)                     69.0018070236  0.9812434767  70.321
```

```
## Adult_mortality                      -0.0582825306  0.0024077267 -24.206
## Alcohol_consumption                   0.3340996862  0.0915303670   3.650
## Hepatitis_B                           0.1181170347  0.0116127381  10.171
## Incidents_HIV                        -1.1333989593  0.1225428127  -9.249
## GDP_per_capita                        0.0002176624  0.0000239675   9.082
## Schooling                             0.5728138081  0.1028102573   5.572
## Adult_mortality:Alcohol_consumption -0.0014556459  0.0001407165 -10.345
## Adult_mortality:Hepatitis_B         -0.0002130287  0.0000297204  -7.168
## Adult_mortality:Schooling             0.0018910404  0.0002030291   9.314
## Adult_mortality:Incidents_HIV         0.0020600583  0.0001500960  13.725
## Alcohol_consumption:GDP_per_capita  -0.0000037978  0.0000006437  -5.900
## Alcohol_consumption:Schooling       -0.0088967464  0.0042432634  -2.097
## Alcohol_consumption:Hepatitis_B       0.0025558890  0.0008346501   3.062
## Hepatitis_B:GDP_per_capita          -0.0000018881  0.0000002545  -7.418
## Hepatitis_B:Schooling               -0.0057124701  0.0010891063  -5.245
## Incidents_HIV:Schooling               0.0663673778  0.0107717147   6.161
##                                                        Pr(>|t|)
## (Intercept)                         < 0.0000000000000002 ***
## Adult_mortality                     < 0.0000000000000002 ***
## Alcohol_consumption                                0.000267 ***
## Hepatitis_B                         < 0.0000000000000002 ***
## Incidents_HIV                       < 0.0000000000000002 ***
## GDP_per_capita                      < 0.0000000000000002 ***
## Schooling                             0.000000027606321 ***
## Adult_mortality:Alcohol_consumption < 0.0000000000000002 ***
## Adult_mortality:Hepatitis_B           0.000000000000967 ***
## Adult_mortality:Schooling           < 0.0000000000000002 ***
## Adult_mortality:Incidents_HIV       < 0.0000000000000002 ***
## Alcohol_consumption:GDP_per_capita    0.000000004053599 ***
## Alcohol_consumption:Schooling                      0.036110 *
## Alcohol_consumption:Hepatitis_B                    0.002217 **
## Hepatitis_B:GDP_per_capita            0.000000000000157 ***
## Hepatitis_B:Schooling                 0.000000167654316 ***
## Incidents_HIV:Schooling               0.000000000823480 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.817 on 2847 degrees of freedom
## Multiple R-squared:  0.9629, Adjusted R-squared:  0.9627
## F-statistic:  4614 on 16 and 2847 DF,  p-value: < 0.00000000000000022
```

Based on the reduced interaction model, all interaction terms are below the significance level of $\alpha = 0.05$. The adjusted R-squared of the reduced interaction model is 0.9627, which is the same as the full model.

Reduced interaction model:

$\hat{LifeExpectancy} = 69.00181 - 0.05828X_{AdultMortality} + 0.33410X_{AlcoholConsumption} + 0.11812X_{HepatitisB} - 1.13334X_{IncidentsHIV} + 0.00022_{GDP} + 0.57281X_{Schooling} - 0.00146X_{AdultMortality}X_{AlcoholConsumption} - 0.00021X_{AdultMortality}X_{HepatitisB} + 0.00189X_{AdultMortality}X_{Schooling} + 0.00206X_{AdultMortality}X_{IncidentsHIV} - 0.000004X_{AlcoholConsumption}X_{GDP} - 0.00890X_{AlcoholConsumption}X_{Schooling} + 0.00256X_{AlcoholConsumption}X_{Hepatitis_B} - 0.000002X_{HepatitisB}X_{GDP} - 0.00571X_{HepatitisB}X_{Schooling} + 0.06637X_{IncidentsHIV}X_{Schooling}$

In addition to the individual t-tests we will conduct a Partial F-test to see which model to accept - reduced interaction model vs full interaction model. First, we state our hypothesis:

H_0 : For all $\beta_q = 0$

H_a : at least one $\beta_q \neq 0$ where q represents subset of terms to be dropped (Adult_mortality:GDP_per_capita, Hepatitis_B:Incidents_HIV, Incidents_HIV:GDP_per_capita and GDP_per_capita:Schooling)

```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ Adult_mortality + Alcohol_consumption + Hepatitis_B +
##     Incidents_HIV + GDP_per_capita + Schooling + Adult_mortality *
##     Alcohol_consumption + Adult_mortality * Hepatitis_B + Adult_mortality *
##     Schooling + Adult_mortality * Incidents_HIV + Alcohol_consumption *
##     GDP_per_capita + Alcohol_consumption * Schooling + Alcohol_consumption *
##     Hepatitis_B + Hepatitis_B * GDP_per_capita + Hepatitis_B *
##     Schooling + Incidents_HIV * Schooling
## Model 2: Life_expectancy ~ (Adult_mortality + Alcohol_consumption + Hepatitis_B +
##     Incidents_HIV + GDP_per_capita + Schooling)^2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   2847 9404.4
## 2   2842 9376.2  5    28.245 1.7123 0.1283
```

When comparing the reduced model and the full interaction model, we receive a p-value of 0.1283 which is greater than the significance level of 0.05. Therefore we drop the insignificant terms and accept the reduced model.

Therefore, our final interaction model is:

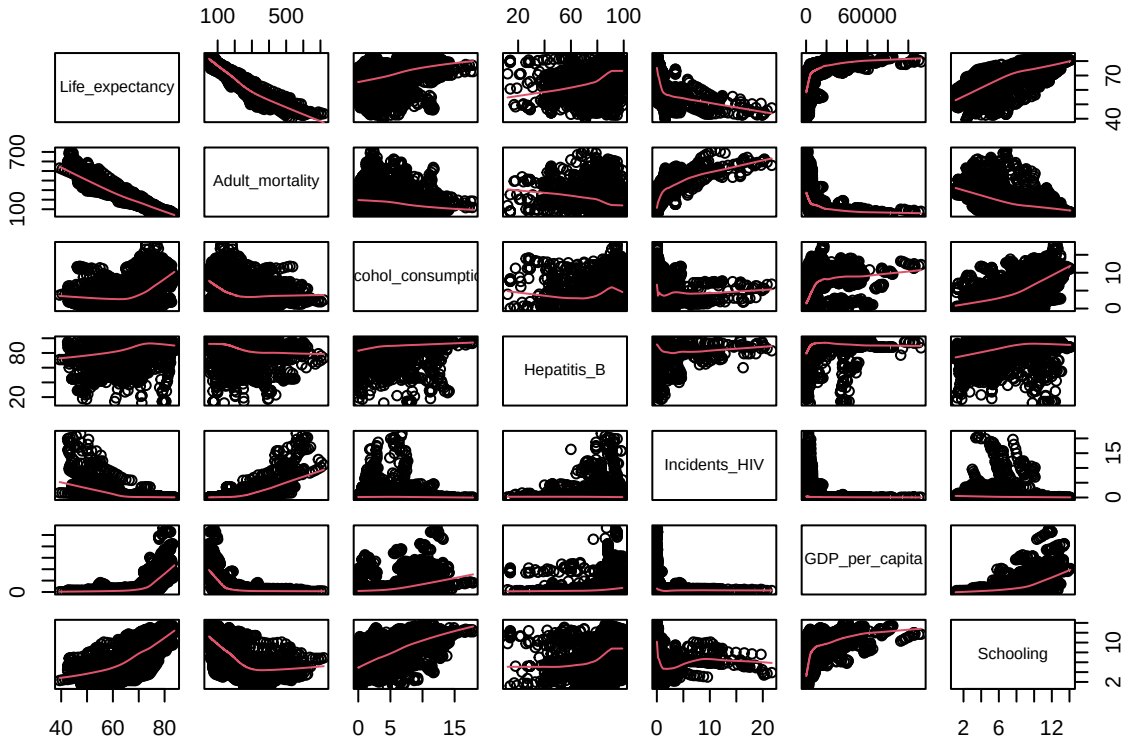
$\hat{LifeExpectancy} = 69.00180 - 0.05828X_{AdultMortality} + 0.33409X_{AlcoholConsumption} + 0.11811X_{HepatitisB} - 1.13339X_{IncidentsHIV} + 0.000217_{GDP} + 0.57281X_{Schooling} - 0.001455X_{AdultMortality}X_{AlcoholConsumption} - 0.000213X_{AdultMortality}X_{HepatitisB} + 0.001891X_{AdultMortality}X_{Schooling} + 0.00206X_{AdultMortality}X_{IncidentsHIV} - 0.000003X_{AlcoholConsumption}X_{GDP} - 0.00889X_{AlcoholConsumption}X_{Schooling} + 0.002555X_{AlcoholConsumption}X_{Hepatitis_B} - 0.0000018X_{HepatitisB}X_{GDP} - 0.00571X_{HepatitisB}X_{Schooling} + 0.06636X_{IncidentsHIV}X_{Schooling}$

3.5 Higher Order Analysis

To see if we can make any further improvements to our model, we will determine if any higher order terms should be included. We will do a higher order analysis for each of our 6 variables in the reduced interactive model. If we identify any variables fit for higher order analysis the maximum degree we will utilize is four to avoid over fitting the model.

We start by visualizing the data to see how the response variable looks with respect to each independent variable.

Figure 15: Pairwise combinations of continuous variables



Based on the plots above we observed that alcohol consumption, schooling and GDP have a presence of curvature in their plots. So, we will test each variable to see if there exists a higher order relationship.

3.5.1 Testing Variables for Higher Order Terms

1) Alcohol Consumption

```
##
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Alcohol_consumption +
##     I(Alcohol_consumption^2) + Hepatitis_B + Incidents_HIV +
##     GDP_per_capita + Schooling + Adult_mortality * Alcohol_consumption +
##     Adult_mortality * Hepatitis_B + Adult_mortality * Schooling +
##     Adult_mortality * Incidents_HIV + Alcohol_consumption * GDP_per_capita +
##     Alcohol_consumption * Schooling + Alcohol_consumption * Hepatitis_B +
##     Hepatitis_B * GDP_per_capita + Hepatitis_B * Schooling +
##     Incidents_HIV * Schooling, data = Life_Expectancy_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4566 -1.0868  0.0177  1.1746  6.6952
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         6.946e+01  9.744e-01  71.279  < 2e-16 ***
## Adult_mortality                    -5.917e-02  2.389e-03 -24.765  < 2e-16 ***
```

```
## Alcohol_consumption                  4.409e-01  9.189e-02   4.798 1.68e-06 ***
## I(Alcohol_consumption^2)            -2.104e-02  2.893e-03  -7.272 4.57e-13 ***
## Hepatitis_B                          1.203e-01  1.151e-02  10.446  < 2e-16 ***
## Incidents_HIV                       -1.162e+00  1.215e-01  -9.564  < 2e-16 ***
## GDP_per_capita                       2.019e-04  2.385e-05   8.467  < 2e-16 ***
## Schooling                            4.385e-01  1.035e-01   4.235 2.36e-05 ***
## Adult_mortality:Alcohol_consumption -1.276e-03  1.416e-04  -9.006  < 2e-16 ***
## Adult_mortality:Hepatitis_B         -2.220e-04  2.948e-05  -7.530 6.79e-14 ***
## Adult_mortality:Schooling            1.999e-03  2.018e-04   9.908  < 2e-16 ***
## Adult_mortality:Incidents_HIV        2.174e-03  1.496e-04  14.534  < 2e-16 ***
## Alcohol_consumption:GDP_per_capita  -3.400e-06  6.402e-07  -5.311 1.18e-07 ***
## Alcohol_consumption:Schooling        1.192e-02  5.087e-03   2.343   0.0192 *
## Alcohol_consumption:Hepatitis_B      1.750e-03  8.345e-04   2.097   0.0360 *
## Hepatitis_B:GDP_per_capita          -1.746e-06  2.530e-07  -6.899 6.45e-12 ***
## Hepatitis_B:Schooling               -5.479e-03  1.080e-03  -5.074 4.14e-07 ***
## Incidents_HIV:Schooling              5.989e-02  1.071e-02   5.591 2.47e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.801 on 2846 degrees of freedom
## Multiple R-squared:  0.9635, Adjusted R-squared:  0.9633
## F-statistic:  4425 on 17 and 2846 DF,  p-value: < 2.2e-16


##
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Alcohol_consumption +
##     I(Alcohol_consumption^2) + I(Alcohol_consumption^3) + Hepatitis_B +
##     Incidents_HIV + GDP_per_capita + Schooling + Adult_mortality *
##     Alcohol_consumption + Adult_mortality * Hepatitis_B + Adult_mortality *
##     Schooling + Adult_mortality * Incidents_HIV + Alcohol_consumption *
##     GDP_per_capita + Alcohol_consumption * Schooling + Alcohol_consumption *
##     Hepatitis_B + Hepatitis_B * GDP_per_capita + Hepatitis_B *
##     Schooling + Incidents_HIV * Schooling, data = Life_Expectancy_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5135 -1.0728  0.0178  1.1766  6.7428
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          6.949e+01  9.740e-01  71.348  < 2e-16 ***
## Adult_mortality                     -5.914e-02  2.388e-03 -24.766  < 2e-16 ***
## Alcohol_consumption                  5.419e-01  1.034e-01   5.239 1.73e-07 ***
## I(Alcohol_consumption^2)            -4.359e-02  1.101e-02  -3.958 7.75e-05 ***
## I(Alcohol_consumption^3)             1.065e-03  5.019e-04   2.122   0.0339 *
## Hepatitis_B                          1.197e-01  1.151e-02  10.401  < 2e-16 ***
## Incidents_HIV                       -1.181e+00  1.217e-01  -9.698  < 2e-16 ***
## GDP_per_capita                       2.022e-04  2.384e-05   8.485  < 2e-16 ***
## Schooling                            4.253e-01  1.037e-01   4.103 4.20e-05 ***
## Adult_mortality:Alcohol_consumption -1.243e-03  1.424e-04  -8.728  < 2e-16 ***
## Adult_mortality:Hepatitis_B         -2.237e-04  2.947e-05  -7.592 4.26e-14 ***
## Adult_mortality:Schooling            1.978e-03  2.019e-04   9.796  < 2e-16 ***
## Adult_mortality:Incidents_HIV        2.191e-03  1.497e-04  14.638  < 2e-16 ***
## Alcohol_consumption:GDP_per_capita  -3.315e-06  6.411e-07  -5.170 2.50e-07 ***
```

```
## Alcohol_consumption:Schooling        1.342e-02  5.133e-03   2.614   0.0090 **
## Alcohol_consumption:Hepatitis_B      1.750e-03  8.340e-04   2.098   0.0360 *
## Hepatitis_B:GDP_per_capita          -1.747e-06  2.529e-07  -6.911 5.92e-12 ***
## Hepatitis_B:Schooling               -5.394e-03  1.080e-03  -4.995 6.22e-07 ***
## Incidents_HIV:Schooling              6.145e-02  1.073e-02   5.727 1.13e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.8 on 2845 degrees of freedom
## Multiple R-squared:  0.9636, Adjusted R-squared:  0.9634
## F-statistic:  4185 on 18 and 2845 DF,  p-value: < 2.2e-16


##
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Alcohol_consumption +
##     I(Alcohol_consumption^2) + I(Alcohol_consumption^3) + I(Alcohol_consumption^4) +
##     Hepatitis_B + Incidents_HIV + GDP_per_capita + Schooling +
##     Adult_mortality * Alcohol_consumption + Adult_mortality *
##     Hepatitis_B + Adult_mortality * Schooling + Adult_mortality *
##     Incidents_HIV + Alcohol_consumption * GDP_per_capita + Alcohol_consumption *
##     Schooling + Alcohol_consumption * Hepatitis_B + Hepatitis_B *
##     GDP_per_capita + Hepatitis_B * Schooling + Incidents_HIV *
##     Schooling, data = Life_Expectancy_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5190 -1.0765  0.0213  1.1791  6.7349
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          6.949e+01  9.746e-01  71.303  < 2e-16 ***
## Adult_mortality                     -5.915e-02  2.389e-03 -24.754  < 2e-16 ***
## Alcohol_consumption                  5.533e-01  1.253e-01   4.415 1.05e-05 ***
## I(Alcohol_consumption^2)            -4.836e-02  3.152e-02  -1.534  0.12503
## I(Alcohol_consumption^3)             1.567e-03  3.149e-03   0.498  0.61881
## I(Alcohol_consumption^4)            -1.653e-05  1.024e-04  -0.161  0.87175
## Hepatitis_B                          1.196e-01  1.152e-02  10.387  < 2e-16 ***
## Incidents_HIV                       -1.181e+00  1.218e-01  -9.695  < 2e-16 ***
## GDP_per_capita                       2.027e-04  2.399e-05   8.449  < 2e-16 ***
## Schooling                            4.250e-01  1.037e-01   4.098 4.29e-05 ***
## Adult_mortality:Alcohol_consumption -1.238e-03  1.450e-04  -8.543  < 2e-16 ***
## Adult_mortality:Hepatitis_B         -2.238e-04  2.948e-05  -7.591 4.28e-14 ***
## Adult_mortality:Schooling            1.975e-03  2.024e-04   9.757  < 2e-16 ***
## Adult_mortality:Incidents_HIV        2.191e-03  1.497e-04  14.633  < 2e-16 ***
## Alcohol_consumption:GDP_per_capita  -3.326e-06  6.452e-07  -5.155 2.71e-07 ***
## Alcohol_consumption:Schooling        1.354e-02  5.195e-03   2.607  0.00918 **
## Alcohol_consumption:Hepatitis_B      1.769e-03  8.429e-04   2.099  0.03590 *
## Hepatitis_B:GDP_per_capita          -1.752e-06  2.542e-07  -6.892 6.77e-12 ***
## Hepatitis_B:Schooling               -5.396e-03  1.080e-03  -4.996 6.22e-07 ***
## Incidents_HIV:Schooling              6.163e-02  1.079e-02   5.712 1.23e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.8 on 2844 degrees of freedom
```

```
## Multiple R-squared:  0.9636, Adjusted R-squared:  0.9634
## F-statistic:  3963 on 19 and 2844 DF,  p-value: < 2.2e-16
```

The above degree 4 term is not significant so we stop testing for terms with a higher degree. The best higher order model when analyzing alcohol consumption is of degree 3 based on the significance of coefficients, R^2_{adj} and RSE.

2) Schooling

```
##
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Alcohol_consumption +
##     Hepatitis_B + Incidents_HIV + GDP_per_capita + Schooling +
##     I(Schooling^2) + Adult_mortality * Alcohol_consumption +
##     Adult_mortality * Hepatitis_B + Adult_mortality * Schooling +
##     Adult_mortality * Incidents_HIV + Alcohol_consumption * GDP_per_capita +
##     Alcohol_consumption * Schooling + Alcohol_consumption * Hepatitis_B +
##     Hepatitis_B * GDP_per_capita + Hepatitis_B * Schooling +
##     Incidents_HIV * Schooling, data = Life_Expectancy_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.715 -1.108  0.026  1.109  5.655
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          6.674e+01  9.902e-01  67.399  < 2e-16 ***
## Adult_mortality                     -5.421e-02  2.401e-03 -22.580  < 2e-16 ***
## Alcohol_consumption                 -9.757e-02  9.964e-02  -0.979 0.327599
## Hepatitis_B                          9.465e-02  1.165e-02   8.126 6.57e-16 ***
## Incidents_HIV                       -1.381e+00  1.229e-01 -11.235  < 2e-16 ***
## GDP_per_capita                       2.128e-04  2.356e-05   9.030  < 2e-16 ***
## Schooling                            1.669e+00  1.485e-01  11.239  < 2e-16 ***
## I(Schooling^2)                      -6.660e-02  6.612e-03 -10.072  < 2e-16 ***
## Adult_mortality:Alcohol_consumption -8.038e-04  1.527e-04  -5.264 1.52e-07 ***
## Adult_mortality:Hepatitis_B         -1.374e-04  3.016e-05  -4.554 5.49e-06 ***
## Adult_mortality:Schooling            1.704e-04  2.627e-04   0.649 0.516610
## Adult_mortality:Incidents_HIV        2.097e-03  1.476e-04  14.211  < 2e-16 ***
## Alcohol_consumption:GDP_per_capita  -3.402e-06  6.338e-07  -5.367 8.64e-08 ***
## Alcohol_consumption:Schooling        2.421e-02  5.310e-03   4.560 5.34e-06 ***
## Alcohol_consumption:Hepatitis_B      3.020e-03  8.216e-04   3.676 0.000241 ***
## Hepatitis_B:GDP_per_capita          -1.875e-06  2.502e-07  -7.496 8.72e-14 ***
## Hepatitis_B:Schooling               -4.830e-03  1.074e-03  -4.498 7.14e-06 ***
## Incidents_HIV:Schooling              9.166e-02  1.088e-02   8.424  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.786 on 2846 degrees of freedom
## Multiple R-squared:  0.9641, Adjusted R-squared:  0.9639
## F-statistic:  4502 on 17 and 2846 DF,  p-value: < 2.2e-16
```

Although the quadratic term is significant, the independent variable Alcohol_consumption becomes insignificant, so we stop testing for higher order terms for schooling.

3) GDP

```
##
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Alcohol_consumption +
##     Hepatitis_B + Incidents_HIV + GDP_per_capita + I(GDP_per_capita^2) +
##     Schooling + Adult_mortality * Alcohol_consumption + Adult_mortality *
##     Hepatitis_B + Adult_mortality * Schooling + Adult_mortality *
##     Incidents_HIV + Alcohol_consumption * GDP_per_capita + Alcohol_consumption *
##     Schooling + Alcohol_consumption * Hepatitis_B + Hepatitis_B *
##     GDP_per_capita + Hepatitis_B * Schooling + Incidents_HIV *
##     Schooling, data = Life_Expectancy_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4531 -1.0952  0.0022  1.1274  6.1820
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          6.914e+01  9.819e-01  70.410  < 2e-16 ***
## Adult_mortality                     -5.883e-02  2.416e-03 -24.351  < 2e-16 ***
## Alcohol_consumption                  2.874e-01  9.340e-02   3.078 0.002107 **
## Hepatitis_B                          1.167e-01  1.162e-02  10.043  < 2e-16 ***
## Incidents_HIV                       -1.130e+00  1.224e-01  -9.228  < 2e-16 ***
## GDP_per_capita                       2.249e-04  2.413e-05   9.322  < 2e-16 ***
## I(GDP_per_capita^2)                 -2.599e-10  1.057e-10  -2.459 0.013989 *
## Schooling                            5.729e-01  1.027e-01   5.578 2.67e-08 ***
## Adult_mortality:Alcohol_consumption -1.366e-03  1.452e-04  -9.409  < 2e-16 ***
## Adult_mortality:Hepatitis_B         -2.095e-04  2.973e-05  -7.046 2.30e-12 ***
## Adult_mortality:Schooling            1.968e-03  2.052e-04   9.588  < 2e-16 ***
## Adult_mortality:Incidents_HIV        2.062e-03  1.500e-04  13.750  < 2e-16 ***
## Alcohol_consumption:GDP_per_capita  -2.625e-06  8.007e-07  -3.278 0.001058 **
## Alcohol_consumption:Schooling       -1.092e-02  4.319e-03  -2.529 0.011500 *
## Alcohol_consumption:Hepatitis_B      2.881e-03  8.444e-04   3.413 0.000652 ***
## Hepatitis_B:GDP_per_capita          -1.833e-06  2.553e-07  -7.180 8.89e-13 ***
## Hepatitis_B:Schooling               -5.822e-03  1.089e-03  -5.346 9.71e-08 ***
## Incidents_HIV:Schooling              6.312e-02  1.084e-02   5.821 6.50e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.816 on 2846 degrees of freedom
## Multiple R-squared:  0.9629, Adjusted R-squared:  0.9627
## F-statistic:  4351 on 17 and 2846 DF,  p-value: < 2.2e-16
```

The coefficients above are significant so we add a term of degree 3.

```
##
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Alcohol_consumption +
##     Hepatitis_B + Incidents_HIV + GDP_per_capita + I(GDP_per_capita^2) +
##     I(GDP_per_capita^3) + Schooling + Adult_mortality * Alcohol_consumption +
##     Adult_mortality * Hepatitis_B + Adult_mortality * Schooling +
##     Adult_mortality * Incidents_HIV + Alcohol_consumption * GDP_per_capita +
##     Alcohol_consumption * Schooling + Alcohol_consumption * Hepatitis_B +
##     Hepatitis_B * GDP_per_capita + Hepatitis_B * Schooling +
##     Incidents_HIV * Schooling, data = Life_Expectancy_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5179 -1.0609 -0.0016  1.1152  6.2472
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          6.913e+01  9.774e-01  70.734  < 2e-16 ***
## Adult_mortality                     -5.883e-02  2.405e-03 -24.467  < 2e-16 ***
## Alcohol_consumption                  2.384e-01  9.343e-02   2.551 0.010789 *
## Hepatitis_B                          1.112e-01  1.161e-02   9.581  < 2e-16 ***
## Incidents_HIV                       -1.172e+00  1.221e-01  -9.597  < 2e-16 ***
## GDP_per_capita                       2.936e-04  2.736e-05  10.734  < 2e-16 ***
## I(GDP_per_capita^2)                 -2.316e-09  4.057e-10  -5.709 1.26e-08 ***
## I(GDP_per_capita^3)                  1.504e-14  2.866e-15   5.247 1.66e-07 ***
## Schooling                            6.071e-01  1.025e-01   5.926 3.48e-09 ***
## Adult_mortality:Alcohol_consumption -1.311e-03  1.449e-04  -9.045  < 2e-16 ***
## Adult_mortality:Hepatitis_B         -1.876e-04  2.988e-05  -6.276 4.00e-10 ***
## Adult_mortality:Schooling            1.808e-03  2.065e-04   8.753  < 2e-16 ***
## Adult_mortality:Incidents_HIV        2.084e-03  1.493e-04  13.957  < 2e-16 ***
## Alcohol_consumption:GDP_per_capita  -2.933e-06  7.991e-07  -3.670 0.000247 ***
## Alcohol_consumption:Schooling       -9.698e-03  4.305e-03  -2.253 0.024353 *
## Alcohol_consumption:Hepatitis_B      3.050e-03  8.411e-04   3.626 0.000293 ***
## Hepatitis_B:GDP_per_capita          -1.828e-06  2.541e-07  -7.194 8.01e-13 ***
## Hepatitis_B:Schooling               -5.974e-03  1.084e-03  -5.509 3.92e-08 ***
## Incidents_HIV:Schooling              6.417e-02  1.079e-02   5.944 3.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.807 on 2845 degrees of freedom
## Multiple R-squared:  0.9633, Adjusted R-squared:  0.9631
## F-statistic:  4149 on 18 and 2845 DF,  p-value: < 2.2e-16
```

The coefficients above are significant so we add a term of degree 4.

```
##
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Alcohol_consumption +
##     Hepatitis_B + Incidents_HIV + GDP_per_capita + I(GDP_per_capita^2) +
##     I(GDP_per_capita^3) + I(GDP_per_capita^4) + Schooling + Adult_mortality *
##     Alcohol_consumption + Adult_mortality * Hepatitis_B + Adult_mortality *
##     Schooling + Adult_mortality * Incidents_HIV + Alcohol_consumption *
##     GDP_per_capita + Alcohol_consumption * Schooling + Alcohol_consumption *
##     Hepatitis_B + Hepatitis_B * GDP_per_capita + Hepatitis_B *
##     Schooling + Incidents_HIV * Schooling, data = Life_Expectancy_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5942 -1.0430 -0.0195  1.1048  6.2936
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          6.902e+01  9.721e-01  71.001  < 2e-16 ***
## Adult_mortality                     -5.900e-02  2.391e-03 -24.673  < 2e-16 ***
```

```
## Alcohol_consumption                  2.338e-01  9.291e-02   2.516  0.01192 *
## Hepatitis_B                          1.069e-01  1.157e-02   9.240  < 2e-16 ***
## Incidents_HIV                       -1.226e+00  1.218e-01 -10.062  < 2e-16 ***
## GDP_per_capita                       3.818e-04  3.122e-05  12.230  < 2e-16 ***
## I(GDP_per_capita^2)                 -8.445e-09  1.139e-09  -7.415 1.59e-13 ***
## I(GDP_per_capita^3)                  1.221e-13  1.882e-14   6.488 1.03e-10 ***
## I(GDP_per_capita^4)                 -5.680e-19  9.869e-20  -5.755 9.59e-09 ***
## Schooling                            6.591e-01  1.023e-01   6.445 1.36e-10 ***
## Adult_mortality:Alcohol_consumption -1.270e-03  1.443e-04  -8.800  < 2e-16 ***
## Adult_mortality:Hepatitis_B         -1.626e-04  3.003e-05  -5.413 6.72e-08 ***
## Adult_mortality:Schooling            1.614e-03  2.081e-04   7.757 1.21e-14 ***
## Adult_mortality:Incidents_HIV        2.124e-03  1.487e-04  14.286  < 2e-16 ***
## Alcohol_consumption:GDP_per_capita  -1.846e-06  8.168e-07  -2.260  0.02391 *
## Alcohol_consumption:Schooling       -9.456e-03  4.281e-03  -2.209  0.02727 *
## Alcohol_consumption:Hepatitis_B      2.720e-03  8.383e-04   3.244  0.00119 **
## Hepatitis_B:GDP_per_capita          -1.616e-06  2.554e-07  -6.328 2.87e-10 ***
## Hepatitis_B:Schooling               -6.315e-03  1.080e-03  -5.848 5.55e-09 ***
## Incidents_HIV:Schooling              6.593e-02  1.074e-02   6.140 9.42e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.797 on 2844 degrees of freedom
## Multiple R-squared:  0.9637, Adjusted R-squared:  0.9635
## F-statistic:  3977 on 19 and 2844 DF,  p-value: < 2.2e-16
```

We stop at the 4th degree to avoid over fitting the model. From the output we observe the best model is the degree 4 model based on the significance of the individual coefficients of the terms, R^2_{adj} and RSE.

3.5.2 Selecting Higher Order Model

From our analysis above we have found that we can add higher order terms to the independent variables alcohol consumption and GDP. Based on the criteria of R^2_{adj} we observe that the higher order model for GDP performs better. This is because it's R^2_{adj}=0.9635 is greater than the $R^2_{adj} = 0.9634$ of the higher order model for alcohol consumption. Therefore, our final higher order model to fit the data is:

$\hat{LifeExpectancy} = 69.01973 - 0.059X_{AdultMortality} + 0.23379X_{AlcoholConsumption} + 0.10691X_{HepatitisB} - 1.22565X_{IncidentsHIV} + 0.00038X_{GDP} + -0.000000008X^2_{GDP} + 0.0000000000001X^3_{GDP} - 0.00000000000000000006X^4_{GDP} + 0.65913X_{Schooling} - 0.00127X_{AdultMortality}X_{AlcoholConsumption} - 0.00016X_{AdultMortality}X_{HepatitisB} + 0.00161X_{AdultMortality}X_{Schooling} + 0.00212X_{AdultMortality}X_{IncidentsHIV} - 0.000002X_{AlcoholConsumption}X_{GDP} - 0.00946X_{AlcoholConsumption}X_{Schooling} + 0.00272X_{AlcoholConsumption}X_{Hepatitis_B} - 0.000002X_{HepatitisB}X_{GDP} - 0.00632X_{HepatitisB}X_{Schooling} + 0.06593X_{IncidentsHIV}X_{Schooling}$

3.6 Results

3.6.1 Selecting the Final Model

After creating an additive model, interaction model, and higher order model, we will use the Adjusted R-Squared criterion to determine the best model in predicting our response variable, life_expectancy. We decided to use Adjusted R-Squared as it is the best way to determine how well a model fits the data. In this case, we will choose the model with the highest Adjusted R-Squared.

Table 3: Comparing Adjusted R-Squared and RSE

```
##              Model    AdjRsq      RSE
## 1    AdditiveModel 0.9544599 2.007170
## 2 InteractionModel 0.9626603 1.817492
## 3 HigherOrderModel 0.9634829 1.797359
```

Based on the table and criteria above the best model to explain life expectancy is the higher order model because it has the highest R^2_{adj} and lowest RSE when compared to other models.

The final model is as follows:

$\hat{LifeExpectancy} = 69.01973 - 0.059X_{AdultMortality} + 0.23379X_{AlcoholConsumption} + 0.10691X_{HepatitisB} - 1.22565X_{IncidentsHIV} + 0.00038X_{GDP} + -0.000000008X^2_{GDP} + 0.0000000000001X^3_{GDP} - 0.00000000000000000006X^4_{GDP} + 0.65913X_{Schooling} - 0.00127X_{AdultMortality}X_{AlcoholConsumption} - 0.00016X_{AdultMortality}X_{HepatitisB} + 0.00161X_{AdultMortality}X_{Schooling} + 0.00212X_{AdultMortality}X_{IncidentsHIV} - 0.000002X_{AlcoholConsumption}X_{GDP} - 0.00946X_{AlcoholConsumption}X_{Schooling} + 0.00272X_{AlcoholConsumption}X_{HepatitisB} - 0.000002X_{HepatitisB}X_{GDP} - 0.00632X_{HepatitisB}X_{Schooling} + 0.06593X_{IncidentsHIV}X_{Schooling}$

The R^2_{adj} of our final model is 0.9634829 which means that 96.34829% of the variation in our response variable life expectancy is explained by the model. The RSE is 1.797359 which means that standard deviation of the unexplained variance by the model is 1.797359 years.

3.6.1 Interpretation of Betas in the Final Model

β_0: Baseline Life-Expectancy when all the variables are 0 is 65.92257 years.

- The effect of Adult Mortality is $-0.059 - 0.00127X_{AlcoholConsumption} - 0.00016X_{HepatitisB} + 0.00161X_{Schooling} + 0.00212X_{IncidentsHIV}$, means that for when other variables are held constant, an increase of 1 year of adult mortality leads to change in life expectancy by $-0.059 - 0.00127X_{AlcoholConsumption} + -0.00016X_{HepatitisB} + 0.00161X_{Schooling} + 0.00212X_{IncidentsHIV}$ years.
- The effect of Alcohol Consumption is $0.23379 - 0.00127X_{AdultMortality} + 0.00946X_{Schooling} + 0.00272X_{Hepatitis_B}$,which means that for when other variables are held constant, an increase of 1 Liters of pure alcohol per person of alcohol consumption leads to change in life expectancy by $0.23379 - 0.00127X_{AdultMortality} + 0.00946X_{Schooling} + 0.00272X_{Hepatitis_B}$ years.
- The effect of Schooling is $-0.65913 + 0.00161X_{AdultMortality} - 0.00946X_{AlcoholConsumption} - 0.00632X_{HepatitisB} + 0.06593X_{IncidentsHIV}$, which means that for when other variables are held constant, an increase of 1 year in schooling leads to change in life expectancy by $-0.65913 + 0.00161X_{AdultMortality} - 0.00946X_{AlcoholConsumption} - 0.00632X_{HepatitisB} + 0.06593X_{IncidentsHIV}$ years.
- The effect of Hepatitis B immunizations is $0.10691 - 0.00016X_{AdultMortality} - 0.000002X_{GDP} - 0.00632X_{Schooling}$, which means that for when other variables are held constant, an increase of 1 percent in coverage of Hepatitis B immunization among one-year-olds leads to change in life expectancy by $0.10691 - 0.00016X_{AdultMortality} - 0.000002X_{GDP} - 0.00632X_{Schooling}$ years.
- The effect of HIV Infections is $-1.22565 + 0.00212X_{AdultMortality} + 0.06593X_{Schooling}$, which means that for when other variables are held constant, an increase of 1000 HIV infections for populations aged 15-49 leads to change in life expectancy by $-1.22565 + 0.00212X_{AdultMortality} + 0.06593X_{Schooling}$ years.
- The effect of GDP is $0.00038 - 0.000000008X^2_{GDP} + 0.0000000000001X^3_{GDP} - 0.00000000000000000006X^4_{GDP} - 0.000002X_{AlcoholConsumption} - 0.000002X_{HepatitisB}$, which means that for when other variables are held constant, an increase of 1 unit in GDP leads to change in life expectancy by $0.00038 - 0.000000008X^2_{GDP} + 0.0000000000001X^3_{GDP} - 0.00000000000000000006X^4_{GDP} - 0.000002X_{AlcoholConsumption} - 0.000002X_{HepatitisB}$ years.

4 Conclusion and Discussion

4.1 Approach

The focus of our research study is to determine which variables have a statistically significant impact in predicting the response variable, "Life_expectancy". Using the Life_Expectancy data set from the Kaggle website, we performed multiple linear regression modelling to arrive at the best model in predicting our response variable which is:

$\hat{LifeExpectancy} = 69.01973 - 0.059X_{AdultMortality} + 0.23379X_{AlcoholConsumption} + 0.10691X_{HepatitisB} - 1.22565X_{IncidentsHIV} + 0.00038X_{GDP} + -0.000000008X_{GDP}^2 + 0.0000000000001X_{GDP}^3 - 0.0000000000000000006X_{GDP}^4 + 0.65913X_{Schooling} - 0.00127X_{AdultMortality}X_{AlcoholConsumption} - 0.00016X_{AdultMortality}X_{HepatitisB} + 0.00161XAdultMortalityX_{Schooling} + 0.00212X_{AdultMortality}X_{IncidentsHIV} - 0.000002X_{AlcoholConsumption}X_{GDP} - 0.00946X_{AlcoholConsumption}X_{Schooling} + 0.00272X_{AlcoholConsumption}X_{HepatitisB} - 0.000002X_{HepatitisB}X_{GDP} - 0.00632X_{HepatitisB}X_{Schooling} + 0.06593X_{IncidentsHIV}X_{Schooling}$

Overall, using an alpha of $\alpha = 0.05$ and Adjusted R-Squared as the primary criterion, we arrived at the optimal model through various statistical tests such as t-tests, full and partial F-tests, automated model selection, testing for interaction terms, and testing for higher order terms. This approach ensured that we used formal p-value tests to assess significant variables an interaction terms when selecting the best model

However, it should be noted that the final model did not pass all assumption tests, namely homoscedasticity and normality. Therefore, the final model may not entirely be reliable for predicting life expectancy due to the high presence of residual errors.

4.2 Future Work

In the future, it would be best if the data set of the research study fulfilled all assumptions tests. This way, the final model and its interpretations becomes more reliable and impactful, especially because the purpose of the study could potentially impact government policies and direction. Additionally, our study limited the variables when we created the additive model. If given more time for research, perhaps the model would be more informative had we included all variables. For example, we had attempted to perform an "all possible regressions" selection procedure, however due to the large number of variables, it was computationally demanding and time-consuming for our research.

In terms of the data set, the data was an aggregate of multiple individual data sets. Another improvement that could be made would be the approach in the collection of the data; it would be better if the data set itself had a clear objective and didn't involve taking information from different sources. The data set was also a compilation of information collected between year 2000-2015. We believe it may be more impactful for our study if the data was limited to the last 5-10 years.

4.3 Conclusion

In conclusion, based on our research study, the variables that impact life expectancy include: Adult Mortality, Alcohol Consumption, Hepatitis B, GDP, HIV Incidents, and Schooling. These variables represent different characteristics of a person and the society they live in, indicating that a mix of socio-economic factors like an economy's GDP and the schooling available have an impact on life expectancy. Additionally, health factors like immunization from Hepatitis B and HIV can be beneficial to predicting life expectancy. Lastly, lifestyle factors like alcohol consumption and the adult death rates are also significant predictors of life expectancy. The interaction of these variables are also statistically significant, indicating that having well-balanced socio-economic surroundings and a healthy lifestyle can potentially extend life expectancy. From the perspective of governments, promoting life expectancy through these factors can help create a healthier and more productive society. On a more personal level, understanding the relationship of the economy, health, and other lifestyle factors can also help individuals live a more satisfying and fulfilling life.

Appendix

References

Life Expectancy (WHO) Fixed. (2022). Www.kaggle.com. https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated