Adult mortality analysis

Pradeepa Ramamurthy, UCID:30270796
Sai P, UCID: 30231852
Gulshan Laskar, UCID: 30256534
Yanwei Yu, UCID: 30249113
Jennifer Stadtfelder, UCID: 30250230

1. INTRODUCTION

1.1. MOTIVATION

1.1.1. Context

The demographic development, which is characterized by increasing life expectancy and a growing number of older people, brings with it new challenges and opportunities. Based on data from the WHO and the United Nations, we intend to identify and analyze the factors that influence health as well as the social and economic factors. On the one hand, it is of great importance for health insurance companies and health institutes that are concerned with the optimal development and implementation of prevention strategies and their cost calculation. On the other hand, it is also highly relevant for the pharmaceutical industry, which relies on research in this area for lucrative, innovative, and curing therapy and drug development for the prevention and/or treatment of diseases as well as risk factors. This is also underpinned by the research of Hoffman (1998), which demonstrated that mortality can be significantly reduced if awareness is created and access to prevention, medication, and clinical treatments is provided based on it.

1.1.2. Problem

Rogers et al. (2005) concluded that other health conditions and chronic diseases such as HIV, hepatitis or the BMI have an influence on adult mortality, since, for example, the BMI predisposes a person to chronic diseases and long-term complications, or HIV weakens the immune system despite life-prolonging measures. In their work, they also emphasized that one of the main influencing factors alongside unfavorable predispositions is age itself, as the risk of chronic diseases and health problems increases significantly with age, and, thus, also for adult mortality (Rogers et al., 2005). Others argue, however, that although age contributes to a higher risk of adult mortality, it is not the main influencing factor.

According to Ronsmans et al., factors at a young age have an influence on adult mortality. First of all, the high rates of infant and child mortality point to inadequate health care and poor living conditions that also affect adults. For example, infant deaths are due to pregnancy complications, which are also often dangerous for adults. Poor medical infrastructure, inadequate nutrition, disease and lack of access to clean water are common causes of high mortality in early life and can have similarly devastating effects on adult health.

Following that, while a study in 1992 has found a positive relationship by age between adult mortality and alcohol consumption (Klatsky et al., 1992), the study by Zhao et al. (1995) suggests that the underlying biological processes in humans indicate that moderate alcohol consumption is beneficial for adult mortality, while excessive consumption is associated with an increased risk of adult mortality (Zhao et al., 1995). However, genetic differences in how different ethnic groups metabolize alcohol imply that these differences are not generalizable and that the combination of country and alcohol consumption should be explored to infer influences on adult mortality (Cahrtier and Caetano, 2010). Consequently, while there are recent case studies on the impact of alcohol on adult mortality in Europe and the US, we will conduct our study using data from around the world (Machenback et al., 2015; Rostron, 2012).

Furthermore, Pozzer et al. (2023) and Kruk et al. (2018) argue that environmental factors have a significant influence on adult mortality. For example, countries with a high GDP have the opportunity to invest in advanced infrastructure, well-distributed and easily accessible health facilities, environmentally friendly technologies (clean air), clean drinking water and healthy diets, resulting in lower adult mortality than developing countries with a low GDP that lack the means to improve these often-inadequate, poor and critical environmental factors and provide preventive measures such as vaccinations against Covid, HPV and cervical cancer that have significantly reduced adult mortality. This is confirmed by Ronsmans et al., according to whom adult mortality can be determined by infant and under-five mortality. First of all, the high rates of infant and under-five mortality indicate inadequate health care and poor living conditions that also affect adults. For example, infant deaths can be caused by pregnancy complications, which are often dangerous for adults as well. Poor medical infrastructure, inadequate nutrition, disease and lack of access to clean water are common causes of high mortality rates in early childhood and can have similarly devastating effects on adult health.

On the other hand, however, Singer et al. (2001) claims in an extensive literature review that over the last hundred-plus years, research has uncovered a clear link between socioeconomic status (SES) and adult mortality. People with higher SES generally live longer and have lower rates of chronic disease, both in industrialized and less developed countries. As a result, the main SES indicators influencing adult mortality were summarized as income, education and occupational status rather than external factors/which country and whether it is a developing country (Singer et al., 2001).

Another controversial dependency is population. Countries with a large population often are considered to have many medical facilities and specialists, especially in urban areas, which can lead to better health infrastructure and, therefore, lower adult mortality. In contrast, a large population can also have a negative impact, as high population growth leads to an overload of the healthcare system, which is then no longer able to meet the demand, resulting in longer waiting times for treatment, inadequate medical care and a lack of preventive measures, which in turn increases adult mortality (Rogers et al., 2005). Supporting that, in the case of Covid, it was also found that the infection often spreads faster in countries with high population density, which increases mortality in adults (Wong and Li, 2020). Accordingly, our assessment and comparison are necessary to capture the contrasting influences of GDP and population on adult mortality.

All in all, there are several factors that influence adult mortality. Many researchers disagree on whether certain factors have a significant influence/ or a stronger influence on mortality than others. As a result, we will solve the contradictions by reviewing and analyzing all these conflicting influences outlined above together to get an accurate insight into the influences on adult mortality.

1.1.3. Challenges

Several aspects can pose a challenge for our research:

First, it will be difficult to draw a clear distinction between the different factors influencing adult mortality, as the factors may also be interdependent. For example, HIV vaccination and education are likely to be interdependent, as educated populations are more likely to be taught about the risks and prevention. Therefore, variable interactions need to be included and measures to test and avoid multicollinearity are required. Statistical methods such as the multicollinearity test (VIF) or visualizations can be used to examine the relationships between these variables in more detail and adjust/remove the model variables accordingly. An interaction model would be another suitable approach to illustrate the interrelationships between, for example, vaccination and GDP/developing country.

In addition, it can be challenging to infer conclusions about adult mortality from the limited influencing factors in the dataset, as not all possible factors are captured. Psychosocial aspects may play a crucial role in relation to mortality, with loneliness, mental illness and loss of social networks also influencing physical well-being. Qualitative data sets through surveys or interviews to capture loneliness, mental illness and social networks could be linked to adult mortality. However, this is beyond the scope of our dataset.

Another important challenge is the possibility of relationships between independent variables, e.g. exponential instead of linear relationships, such as the risk of dying from diseases like HIV that progressively weaken the immune system over time and kill the carrier at an older age. To better grasp such complex relationships, we can test and use higher order models as these models allow us to model not only linear but also non-linear relationships between the variables. A handy tool in R to visually examine these relationships is for example the ggpairs function from the GGally package. With this function, we can create a matrix of diagrams that represent different combinations of the variables in our data set.

All in all, even though our project faces challenges such as limited data, possible relationships and dependencies between independent variables and non-linear relationships, there are still some tools in R that we can utilize to solve these issues.

1.2. OBJECTIVES

1.2.1 Overview

Overall, we intend to study the conflicting variables identified in the existing literature to identify which correlations and influencing factors exist on the mortality rate in adulthood. For this purpose, we have formulated and visualized various hypotheses and correlations that we will examine in detail in our analysis. To analyze the data comprehensively, we will use a range of statistical tools and methods. This approach will allow us to capture the relationships and influences between the variables in detail and develop a deeper understanding of the factors that may influence mortality in adulthood. Our aim is to draw conclusions and give insights through comprehensive data analysis, interpretation and recommendations. This will benefit different parties and open new perspectives for further research.

1.2.2 Goals & Research Questions

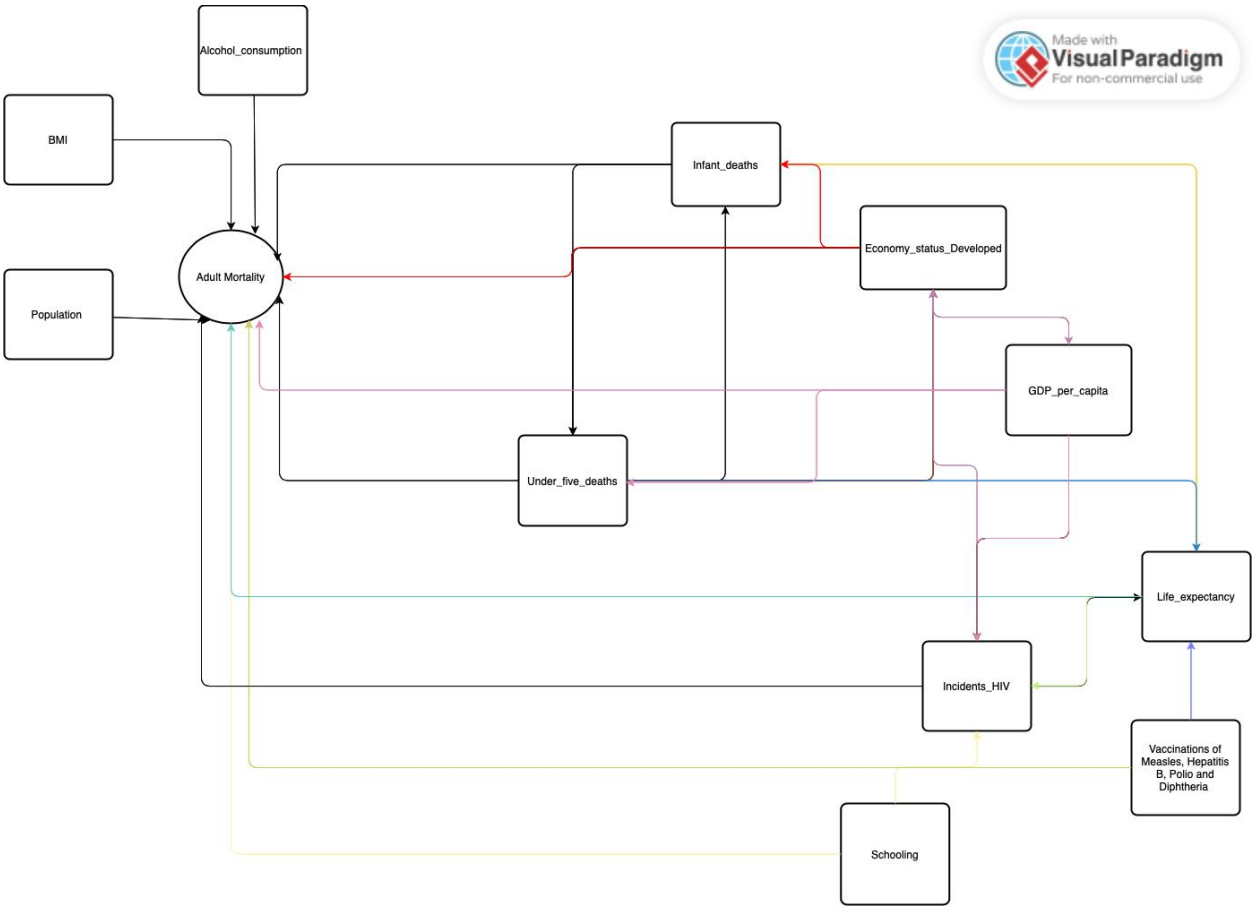

Figure 1: The relationships of the dependent and independent variables

The hypothesis and the analysis test were listed in table 1, and the hypothesis relationship chart was showed as figure 1.

Table 1: The hypothesis of each independent variable and the tests conducted. (Note: The residual analysis includes: 1. Linearity Assumption. 2. Independence Assumption. 3. Equal Variance Assumption. 4. Normality Assumption. 5. Multicollinearity. 6. Outliers.)

Hypothesis	**Tests**
Infant deaths have a significant positive impact on adult mortality (higher adult mortality with higher infant deaths and vice versa).	**Residual analysis** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
Under five deaths have a significant positive impact on adult mortality (higher adult mortality with number of under five deaths and vice versa).	**Residual analysis** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection

	Backward Elimination model selection
Life expectancy has a significant positive impact on the adult mortality (higher adult mortality with higher age and vice versa)	**Residual analysis:** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
Alcohol consumption has a significant positive impact on adult mortality (higher adult mortality with higher alcohol consumption and vice versa).	**Residual analysis:** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
Hepatitis_B has a significant negative impact on the adult mortality (less adult mortality with higher number of Hepatitis_B vaccinations and vice versa).	**Residual analysis:** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
Measles has a significant negative impact on adult mortality (less adult mortality with higher number of Measles vaccinations and vice versa).	**Residual analysis:** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
Diphtheria has a significant negative impact on the adult mortality (less adult mortality with higher number of Diphtheria vaccinations and vice versa).	**Residual analysis:** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
Polio has a significant negative impact on adult mortality (less adult mortality with higher number of Polio vaccinations and vice versa).	**Residual analysis:** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
BMI has a significant negative impact on adult mortality (less adult mortality with higher BMI and vice versa).	**Residual analysis:** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
Incidents_HIV has a significant positive impact on the adult mortality (higher adult mortality	**Residual analysis:** **Model fitting and analysis:**

with higher number of HIV incidents and vice versa).	T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
GDP_per_capita has a significant negative impact on the adult mortality (less adult mortality with higher GDP and vice versa).	**Residual analysis:** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
Population_mln has a significant positive impact on the adult mortality (higher adult mortality with higher Population and vice versa).	**Residual analysis:** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
Schooling has a significant negative impact on adult mortality (less adult mortality with better education and vice versa).	**Residual analysis:** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
Economy_status_Developed has a significant positive impact on adult mortality (higher adult mortality when it is a developing country and vice versa).	**Residual analysis:** **Model fitting and analysis:** T-test for individual regression coefficients Stepwise model selection Forward Selection model selection Backward Elimination model selection
The number of under-five deaths is directly related to infant deaths, as infant deaths in the 0 to 1 year age group are included in the total number of under-five deaths.	**Residual analysis:** **Model fitting and analysis:** T-test for individual and interacting regression coefficients (interaction model)
Economy and GDP interact with each other, as a lower GDP is often a characteristic of a developing country.	**Residual analysis:** **Model fitting and analysis:** T-test for individual and interacting regression coefficients (interaction model)
Infant mortality and under-five mortality interact with the average life expectancy, as high mortality in these age groups reduces overall life expectancy.	**Residual analysis:** **Model fitting and analysis:** T-test for individual and interacting regression coefficients (interaction model)
The risk of HIV infection interacts with the average life expectancy, as despite advanced treatment options, HIV-positive people often have a lower life expectancy, which is exacerbated by good healthcare and adequate	**Residual analysis:** **Model fitting and analysis:** T-test for individual and interacting regression coefficients (interaction model)

treatment. This can also be linked to GDP, which can be taken as an indicator of the economic and social conditions/health care supply.	
Vaccinations also interact with the average life expectancy, as they prevent diseases that can be potentially fatal.	**Residual analysis:** **Model fitting and analysis:** T-test for individual and interacting regression coefficients (interaction model)
The level of education interacts with HIV incidents, as better educated people tend to have a higher awareness of preventive measures and better access to health services.	**Residual analysis:** **Model fitting and analysis:** T-test for individual and interacting regression coefficients (interaction model)

2. METHODOLOGY

2.1 Data

Our variable (Gochiashvili, 2023) shows in the table 2.

Table 2: Independent and dependent variables with description in our dataset.

Variable in model	Variable name	Variable type	Variable description
Dependent variable	Life_expectancy	Quantitative	The average life expectancy of people in that country for that year.
Independent variable	Infant_deaths	Quantitative	Number of infant deaths per 1000 inhabitants.
Independent variable	Under_five_deaths	Quantitative	Number of deaths of children under the age of five per 1000 inhabitants.
Independent variable	Adult_mortality	Quantitative	Number of adult deaths per 1000 inhabitants.
Independent variable	Alcohol_consumption	Quantitative	Alcohol consumption is recorded in liters of pure alcohol per capita with 15+ years old.
Independent variable	Hepatitis_B, Measles, Polio & Diphtheria	Quantitative	Represents the percentage coverage of vaccination against the corresponding diseases (Measles etc.) in 1-year-olds.
Independent variable	BMI	Quantitative	Measure of nutritional status in adults, which means average Body Mass Index.

Independent variable	Incidents_HIV	Quantitative	Incidents of HIV per 1000 population aged 15-49.
Independent variable	GDP_per_capita	Quantitative	GDP per capita (USD).
Independent variable	Population_mln	Quantitative	The total population in that country is in the millions.
Independent variable	Economy_status_Developed & Economy_status_Developing	Qualitative	Developed or developing country (1 means yes, 0 means no). We will combine these two columns into one variable when we do the regression analysis.
Independen variable	Schooling	Quantitative	Average years that people aged 25+ spent in formal education.

The dataset we chose for our analysis was part of a project in which various factors were categorized into immunization-related factors, mortality factors, economic factors and social factors affecting life expectancy as part of a multi-linear regression analysis. The information on life expectancy and health determinants for 193 countries was obtained by Deeksha Russell and Duan Wang from the WHO database, while the corresponding economic data was collected from the United Nations website. As the datasets came from the WHO, there were no obvious errors. Missing/left out data was identified via the "Missmap" function in R and mainly included data on hepatitis B and gross domestic product (GDP) from smaller countries such as Vanuatu, Tonga, Togo and Cabo Verde. The various data files from the open-source databases were then combined into a single dataset and published on the open data source website Kaggle in 2017, indicating that there is no special license to obtain to process and analyze the data further as all the data was based on open sources. We now intend to extend this data analysis/research by focusing on adult mortality and un categorizing the data to identify more precise determinants rather than using broad categories.

2.2 Approach

First, we use model fitting and analysis to identify the relevant relationships within the data. This includes the use of T-tests that examine both individual and interactive regression coefficients in an interaction model. These tests allow us to progressively identify significant relationships between variables and discover the interactions between variables.

Procedures such as stepwise regression, forward stepwise regression and backward stepwise regression also assist in determining the most important influencing factors and reducing model complexity, which leads to better interpretability of the results. In addition, the ANOVA test helps to compare these models and to quantify an improvement of the models.

To confirm the assumption of linearity of our model, we use residual plots as they enable us to identify whether there is a linear relationship between the variables. It is also important to ensure the independence

of the data, so we do not use time series data in our analysis. To test for homoscedasticity, i.e. the assumption of equal variance, we use both residual plots and the Breusch-Pagan test to determine whether the variance of the residuals remains constant.

The normality assumption of the residuals is checked with a histogram and a Q-Q diagram. To determine multicollinearity, we use the Variance Inflation Factor (VIF) with methods such as imcdiag and the pairs() visualization to rule out a linear relationship between individual variables. Furthermore, the analysis of outliers is conducted via residuals vs. leverage plots to identify potentially influential variables and Cook's distance to quantify the influence of individual data points.

If these assumptions could not be fulfilled, applying Box-Cox transformations helps us to stabilize the data and promote the normal distribution, which makes our model more accurate and better for prediction. Another aspect of these adjustments is the careful removal of outliers, as these often distort the overall picture and can therefore jeopardize the validity of our analyses. In addition, we center the data to improve the multicollinearity of variables and optimize the model fit.

2.3 Workflow

This workflow outlines a structured methodology for building and refining predictive models, focusing on hypothesis testing, variable selection, and model optimization. The process emphasizes the iterative evaluation of variables, interactions, and model assumptions to ensure statistical robustness and predictive accuracy. Key steps include exploratory data analysis, regression techniques, residual analysis, and handling outliers, addressing both statistical and practical challenges inherent in model development. Through this approach, we aim to identify the best-fitting model that balances complexity and reliability for prediction.

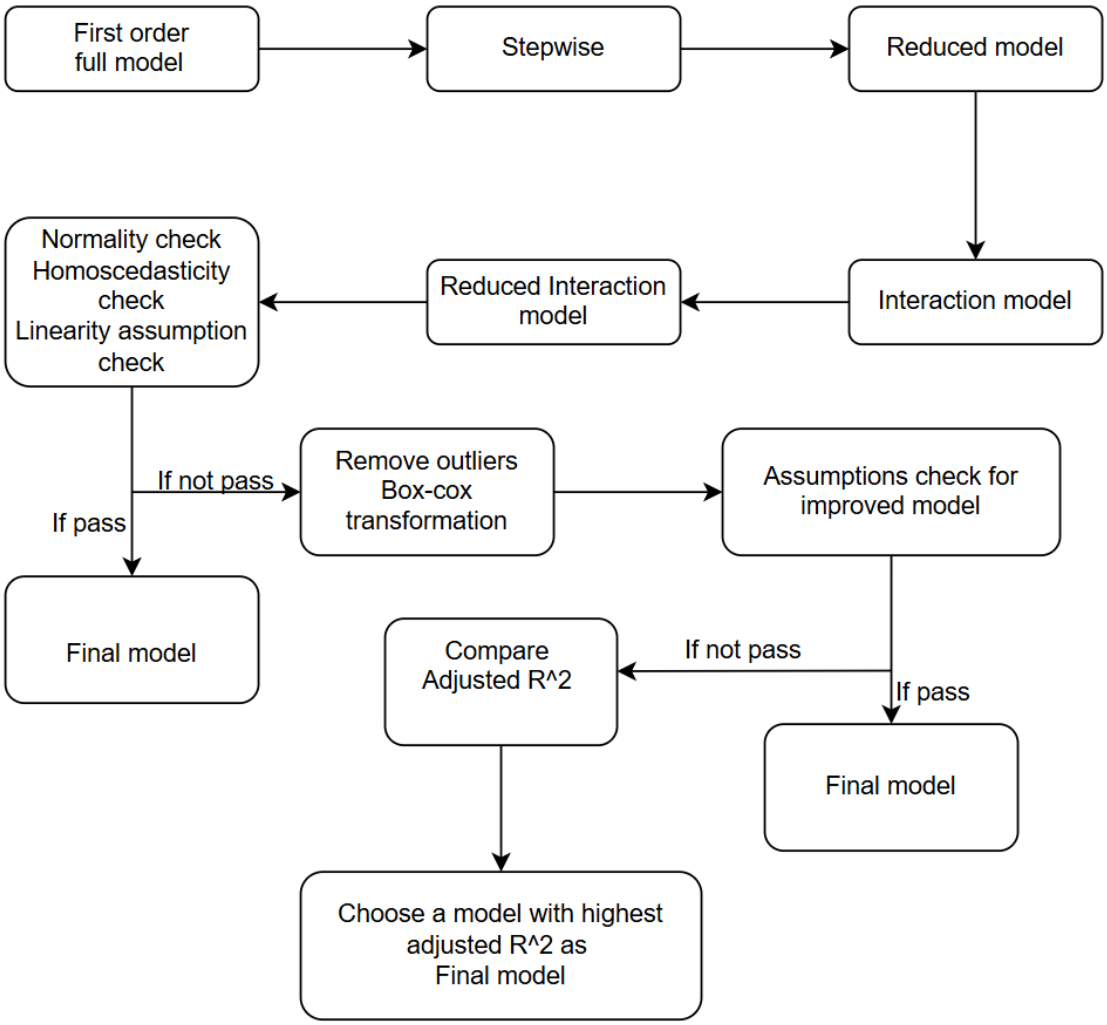

Figure 2: Flow chart of how to get the final best fit model.

In the following, we outline our procedure for investigating and rejecting/confirming the above hypotheses on the variables in detail. First, we developed a model that included all defined relevant variables. At a significant level of 0.05, we conducted individual t-tests to identify the variables that had the strongest relationships with our dependent variable and removed the insignificant variables incrementally. We then analyzed and compared these reduced models using various metrics, such as the residual standard error (RSE), adjusted R^2, and ANOVA test, to check whether the reduced/truncated model had a better fit compared to the previous model. This step is crucial to remove irrelevant predictors from the model and reduce complexity.

In addition, we performed different regression techniques such as stepwise, backward and forward regression. These methods allowed us to compare the previously identified variables and select the variables that were relevant in all tests for the final additive model, ensuring the accuracy and fit of the final model structure.

Based on the significant variables from the best fit additive model, an interaction model was built. Again, we removed manual gradually insignificant interactions and variables at a significance level of 0.05, taking into account a hierarchical order. This hierarchical adjustment is important because it ensures that relevant variables remain in the model even if they are not significant without interactions.

We then conducted the ANOVA test to clarify whether the interaction model or the reduced model provides the better fit and is, therefore, more appropriate for predictions. This allowed us to select the model with the highest prediction accuracy.

In the next step, we applied a residual analysis to test the underlying assumptions of the model. This included normality tests, homoscedasticity tests and the assumption of linearity, supported by visualizations and tests in R (such as VIF, Shapiro-Wilk, etc.). If these tests were positive, we consider the model to be our final model for predictions.

However, if it was found that one or more assumptions were not met, we made further adjustments, for example by gradually removing significant outliers, centering the data and performing a Box-Cox transformation to normalize the distribution and improve the data according to the regression requirements. After these adjustments, we repeated the residual analysis to check whether the model assumptions were now met.

If the model passed all tests, we considered it the final model. Otherwise, we compared the revised model with the previously defined model and selected the model with the highest revised R^2 and the lowest RSE. This iterative process ensures that we get a reliable regression model for our predictions.

Dealing with outliers in data sets can be extremely time-consuming, especially with over a hundred such values. Thus, when having over 100 outliers, the step-by-step process of identifying and removing each outlier is not only laborious but also requires ongoing calculations and analysis to understand the impact of these adjustments on the dataset and the results.

Another challenge is to find the most appropriate model that accurately describes the adjusted data. This requires the evaluation of different models, and the selection of the appropriate evaluation criteria is often a source of debate: Should one be guided by the highest adjusted R^2, the lowest RSE, or the best ACI? The graphs are also usually ambiguous and leave room for controversy: at what point do we refer to an excessive deviation and when are the variances still within the acceptable range? In addition, relevant possible combinations must be considered, as well as rules such as the hierarchical arrangement of the variables which require detailed orientation and patience.

2.4 Contributions

We have distributed the workload among the individual team members. This means that each team member has their own responsibility for certain work but is not limited to that work. As the process progresses, more work can be assigned to everyone, and some initial work can also be removed if it is not needed.

Pradeepa: She verifies the requirements for the regression analysis (for example normal distribution and linearity). In the final description of the code, she explains whether these conditions are satisfied or what steps need to be undertaken to adjust the model accordingly.

Sai: He will visualize the data to show and identify correlations and patterns. He will also review the additive and interactive model fit by applying the forward selection procedure and justifying in the final code description which model fits best and why (e.g. adjusted r2 etc.).

Gulshan: She checks the fit of the additive and interaction model using the backward elimination procedure. In the final code description, she explains which model is best suited and presents the corresponding arguments (e.g. adjusted r2 etc.).

Jennifer: She finds the best additive regression model using the individual t-test, performs data cleaning, such as centering the data and removing outliers, and creates the final code description. She summarizes the results in the "Interpretation and Conclusions" section of the final report and designs the presentation.

Yanwei: He examines the fit of the additive and interaction model using the stepwise procedure and adjusts the model according to the various requirement tests (linearity, normality, etc.). He also summarizes the results in the interpretation and conclusion sections of the final report and supports the design of the presentation.

In the final presentation, the slides and speaking time are divided equally so that everyone is involved.

3. MAIN RESULTS OF THE ANALYSIS

3.1 Results

After conducting T-tests for the individual regression coefficients and a stepwise regression at a significance level of 0.05, in which we gradually eliminated all insignificant variables, we developed an additive model based on the most significant relationships. This model has an adjusted r2 of 0.9587, corresponding to 95.87% of the variation in the dependent variable is explained by this final additive model. Additionally, the RSE of 23.34 of the unexplained variances is explained by the model. Although this model did not have the highest adjusted R-squared (Adjusted R^2) or the lowest residual standard error (RSE), the ANOVA test showed that it provided the best fit to the data. Thus, the variables with the most significant influence on adult mortality are: Infant_deaths, under_five_deaths, life_expectancy, allcohol_consumption, Hepatitis_B, BMI, Incidents_HIV, Population_mln and Schooling.

```
##
## Call:
## lm(formula = Adult_mortality ~ Infant_deaths + Under_five_deaths +
##     Life_expectancy + Alcohol_consumption + Hepatitis_B + BMI +
##     Incidents_HIV + Population_mln + Schooling, data = cleaned_life)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -95.608 -16.155  -0.862  14.429 118.523
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1246.16232   14.64766  85.076  < 2e-16 ***
## Infant_deaths         -0.44733    0.10467  -4.274 1.98e-05 ***
## Under_five_deaths     -0.45452    0.06191  -7.342 2.73e-13 ***
## Life_expectancy      -14.12128    0.15651 -90.229  < 2e-16 ***
## Alcohol_consumption    2.12104    0.14683  14.445  < 2e-16 ***
## Hepatitis_B           -0.09430    0.03259  -2.893  0.00384 **
## BMI                   -3.12498    0.28669 -10.900  < 2e-16 ***
## Incidents_HIV          8.12779    0.25830  31.466  < 2e-16 ***
## Population_mln        -0.01633    0.00331  -4.934 8.54e-07 ***
## Schooling              2.72598    0.26870  10.145  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.34 on 2854 degrees of freedom
## Multiple R-squared:  0.9589, Adjusted R-squared:  0.9587
## F-statistic:  7394 on 9 and 2854 DF,  p-value: < 2.2e-16
```

Figure 3: Result of final best fit additive model.

Final additive model= 1246.16232280 -0.44733348 x Infant_deaths -0.45451495 x Under_five_deaths -14.12127969 x Life_expectancy + 2.12103734 x Alcohol_consumption -0.09429649 x Hepatitis_B -3.12497545 BMI + 8.12779156 x Incidents_HIV -0.01632884 x Population_mln + 2.72597472 x Schooling

Interpretation of ß:
ß0: Number of adult deaths per 1000 inhabitants when all other variables are zero. It is the initial value from which the changes due to the other factors are derived.
ß1: An increase in child mortality/infant deaths by one unit leads to a decrease in adult mortality by about 0.447.
ß2: An increase in deaths under five by one unit leads to an increase in adult mortality by about 0.454.
ß3: An increase in life expectancy by one unit leads to an increase in adult mortality by about 14.12.
ß4: An increase in alcohol consumption by one unit leads to an increase in adult mortality by about 2.12.
ß5: An increase in Hepatitis B vaccinations by one unit leads to a decrease in adult mortality by about 0.094.
ß5: An increase in BMI by one unit leads to a decrease in adult mortality by about 3.12.
ß6: An increase in HIV incidents by one unit leads to an increase in adult mortality by about 8.12.
ß7: An increase in the number of populations by one unit leads to a decrease in adult mortality by about 0.016.
ß8: An increase in the average years that people spend in school by one unit leads to an increase in adult mortality by about 2.73.

Based on the significant variables, we created a full interaction model. We then removed all insignificant interaction terms while retaining the hierarchical model, at a significance level of $\alpha = 0.05$. The resulting reduced interaction model, shown in Figure 4, includes all significant interaction terms at the same level.

The adjusted R-squared is 0.9735, which means that 97.35% of the variation in the dependent variable is explained by this final interaction model. The low residual standard error of 18.69 indicates that the model predictions are close to the actual values on average. Many predictors, including Infant_deaths, Life_expectancy, Alcohol_consumption, BMI, Hepatitis_B, Incidents_HIV, Schooling, and their interactions, have p-values < 0.05, making them statistically significant. These results highlight their strong association with the dependent variable.

```
Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                         2.625e+03  1.862e+02  14.098  < 2e-16 ***
Infant_deaths                      -1.420e+01  1.644e+00  -8.636  < 2e-16 ***
Under_five_deaths                   8.668e+00  1.109e+00   7.816 7.63e-15 ***
Life_expectancy                    -2.869e+01  2.353e+00 -12.193  < 2e-16 ***
Alcohol_consumption                -1.196e+01  2.312e+00  -5.171 2.49e-07 ***
Hepatitis_B                         9.226e-01  8.898e-01   1.037 0.299889
BMI                                -9.501e+01  6.222e+00 -15.272  < 2e-16 ***
Incidents_HIV                       1.309e+01  3.136e+00   4.174 3.08e-05 ***
Population_mln                      3.823e-01  2.102e-01   1.819 0.069068 .
Schooling                           6.843e+01  6.436e+00  10.633  < 2e-16 ***
Infant_deaths:Under_five_deaths    -1.151e-02  1.551e-03  -7.423 1.51e-13 ***
Infant_deaths:Life_expectancy      -1.064e-01  9.131e-03 -11.657  < 2e-16 ***
Infant_deaths:Alcohol_consumption   2.975e-01  3.589e-02   8.290  < 2e-16 ***
Infant_deaths:Hepatitis_B           4.430e-02  5.488e-03   8.071 1.02e-15 ***
Infant_deaths:BMI                   9.192e-01  7.383e-02  12.451  < 2e-16 ***
Infant_deaths:Incidents_HIV        -3.995e-01  3.461e-02 -11.543  < 2e-16 ***
Infant_deaths:Population_mln        4.018e-03  1.961e-03   2.049 0.040533 *
Infant_deaths:Schooling            -5.614e-01  5.481e-02 -10.241  < 2e-16 ***
Under_five_deaths:Alcohol_consumption -1.726e-01  2.280e-02  -7.571 4.98e-14 ***
Under_five_deaths:Hepatitis_B      -3.022e-02  3.268e-03  -9.248  < 2e-16 ***
Under_five_deaths:BMI              -3.341e-01  5.086e-02  -6.568 6.05e-11 ***
Under_five_deaths:Incidents_HIV     2.430e-01  1.967e-02  12.354  < 2e-16 ***
Under_five_deaths:Population_mln   -3.204e-03  1.439e-03  -2.227 0.026036 *
Under_five_deaths:Schooling         1.728e-01  3.520e-02   4.910 9.65e-07 ***
Life_expectancy:Hepatitis_B        -3.213e-02  1.005e-02  -3.198 0.001399 **
Life_expectancy:BMI                 1.073e+00  7.762e-02  13.829  < 2e-16 ***
Life_expectancy:Population_mln     -7.376e-03  2.765e-03  -2.668 0.007678 **
Life_expectancy:Schooling          -7.949e-01  7.807e-02 -10.182  < 2e-16 ***
Alcohol_consumption:Hepatitis_B     6.957e-02  9.309e-03   7.473 1.04e-13 ***
Alcohol_consumption:BMI             2.691e-01  8.285e-02   3.248 0.001176 **
Alcohol_consumption:Incidents_HIV  -3.461e-01  9.748e-02  -3.551 0.000390 ***
Hepatitis_B:BMI                     4.727e-02  1.625e-02   2.908 0.003661 **
Hepatitis_B:Incidents_HIV          -8.028e-02  2.348e-02  -3.419 0.000637 ***
Hepatitis_B:Schooling              -4.001e-02  1.565e-02  -2.557 0.010612 *
Incidents_HIV:Population_mln        9.929e-02  1.420e-02   6.993 3.34e-12 ***
Incidents_HIV:Schooling             4.150e-01  1.843e-01   2.251 0.024450 *
Population_mln:Schooling            1.858e-02  2.878e-03   6.456 1.26e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.69 on 2827 degrees of freedom
Multiple R-squared:  0.9739,    Adjusted R-squared:  0.9735
F-statistic:  2927 on 36 and 2827 DF,  p-value: < 2.2e-16
```

Figure 4: Result of final best fit interaction regression model.

Final interaction model: 2.624765e+03 + 3.822692e-01 Population_mln+ 6.843082e+01 Schooling + 1.309210e+01 Incidents_HIV+ 9.226069e-01 Hepatitis_B -1.195656e+01 Alcohol consumption+ 8.668052e+00 Under five deaths -2.868880e+01 Life_expectancy -9.501348e+01 BMI -1.420109e+01 Infant_deaths+ 1.858167e-02 Population_mln:Schooling+ 4.149899e-01 Schooling:Incidents_HIV+ 9.928750e-02 Population_mln:Incidents_HIV -4.001013e-02 Schooling:Hepatitis_B -8.028339e-02 Incidents_HIV:Hepatitis_B+ 4.727539e-02 Hepatitis_B:BMI -3.461229e-01

Incidents_HIV:Alcohol_consumption+ 2.690985e-01 Alcohol_consumption:BMI+ 6.956495e-02 Hepatitis_B:Alcohol_consumption -7.948872e-01 Schooling:Life_expectancy -7.376268e-03 Population_mln:Life_expectancy+ 1.073460e+00 Life_expectancy:BMI -3.212503e-02 Hepatitis_B:Life_expectancy+ 1.728216e-01 Schooling:Under_five_deaths -3.203992e-03 Population_mln:Under_five_deaths+ 2.429548e-01 Incidents_HIV:Under_five_deaths -3.340532e-01 Under_five_deaths:BMI -3.022240e-02 Hepatitis_B:Under_five_deaths -1.726443e-01 Alcohol_consumption:Under_five_deaths -5.613555e-01 Schooling:Infant_deaths+ 4.017859e-03 Population_mln:Infant_deaths -3.994617e-01 Incidents_HIV:Infant_deaths+ 9.192351e-01 BMI:Infant_deaths+ 4.429934e-02 Hepatitis_B:Infant_deaths + 2.975463e-01 Alcohol_consumption:Infant_deaths -1.064425e-01 Life_expectancy:Infant_deaths -1.151015e-02 Under_five_deaths:Infant_deaths

Interpretation of ß. Note, that the interaction coefficients were added to the indicative variables, as these have either an increasing or decreasing effect that changes the overall influence of the independent variable on the dependent variable:
ß0: Number of adult deaths per 1000 inhabitants when all other variables are zero. It is the initial value from which the changes due to the other factors are derived.
ß1: An increase in child mortality/infant deaths by one unit leads to a decrease in adult mortality by about 14.01476.
ß2: An increase in deaths under five by one unit leads to an increase in adult mortality by about 8.532194.
ß3: An increase in life expectancy by one unit leads to a decrease in adult mortality by about 28.55617.
ß4: An increase in alcohol consumption by one unit leads to a decrease in adult mortality by about 11.90868.
ß5: An increase in Hepatitis B vaccinations by one unit leads to an increase in adult mortality by about 0.9011056.
ß5: An increase in BMI by one unit leads to a decrease in adult mortality by about 92.9689.
ß6: An increase in HIV incidents by one unit leads to an increase in adult mortality by about 13.02346.
ß7: An increase in the number of populations by one unit leads to an increase in adult mortality by about 0.49678.
ß8: An increase in the average years that people spend in school by one unit leads to an increase in adult mortality by about 67.63776.

We then tested various regression assumptions, except for the independence assumption, since independence can be assumed if no time series data are available:

1.Linearity Assumption:
In the following we tested the hypothesis:
H0= Linearity is not present (homoscedasticity

Ha= Linearity is present

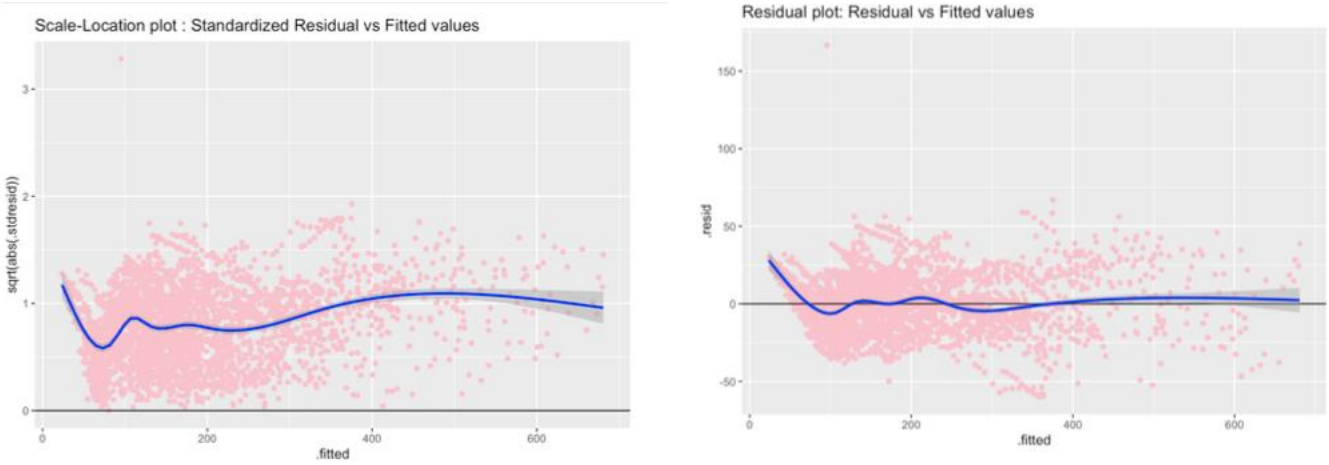

Figure 5: The scale-location plot and the residual plot for final best fit interaction regression model.

The presence of a pattern could indicate a problem with some aspect of the linear model. However, no patterns or structures could be detected in the graph, so we conclude that there is linearity.

2.Equal Variance Assumption:
In the following we tested the hypothesis:
H0= heteroscedasticity is not present (homoscedasticity
Ha= heteroscedasticity is present
At a significance level of 0.05, we discovered that we do have heteroscedasticity.

```
        studentized Breusch-Pagan test

data:  final_interact_reduced_model_life_new
BP = 711.59, df = 36, p-value < 2.2e-16
```

Figure 6: Homoscedasticity test result for final best fit interaction regression model.

3.Normality Assumption:
In the following we tested the hypothesis:
H0= The sample data follows a normal distribution

Ha= The sample data does not follow a normal distribution

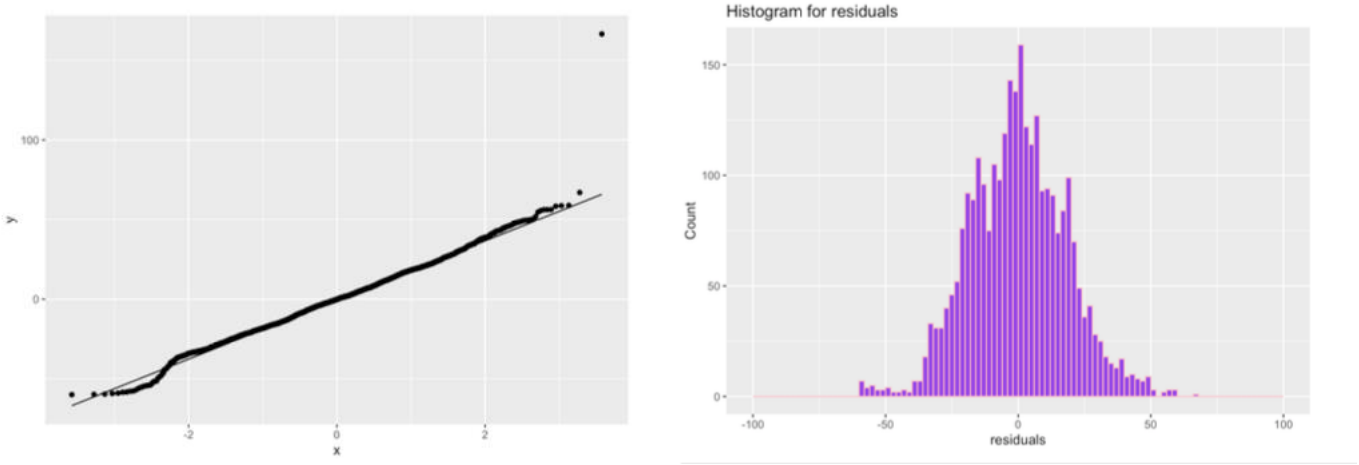

Figure 7: Q-Q plot for final best fit interaction regression model and histogram for residuals for final best fit interaction regression model.

```
        Shapiro-Wilk normality test

data:  residuals(final_interact_reduced_model_life_new)
W = 0.98735, p-value = 2.718e-15
```

Figure 8: Normality test result for final best fit interaction regression model.

The graph revealed a normal distribution. The data points were very close to the Q-Q plot line and the histogram also shows a uniform distribution around the zero value. At a significance level of $\alpha = 0.05$, however, we found that the residuals are not normally distributed since the p-value = 2.718e-15<0.05. Consequently, we used a Box-Cox transformation (transformations for normality and heteroscedasticity) to fulfill the regression requirements and to design a model that can be used for predictions.

4. Multicollinearity

In the following we tested the hypothesis:

1≤VIFS≤5 indicate that there is moderate collinearity, but not severe enough to warrant corrective action. VIFs>5 or 10 represent critical values for multicollinearity, where the coefficients are poorly estimated, and the p-values are questionable.

```
##
## Call:
## imcdiag(mod = interact_recuded_model_life2, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                                        VIF detection
## Population_mln                     6743.2537         1
## Schooling                          3413.7286         1
## Incidents_HIV                       457.0398         1
## Hepatitis_B                        1659.7780         1
## Alcohol_consumption                 694.6291         1
## Under_five_deaths                 20017.7402         1
## Life_expectancy                    4012.8666         1
## BMI                                1526.5037         1
## Infant_deaths                     16802.6868         1
## Population_mln:Schooling             56.7164         1
## Schooling:Incidents_HIV              61.2570         1
## Population_mln:Incidents_HIV          3.5041         0
## Schooling:Hepatitis_B               204.4389         1
## Incidents_HIV:Hepatitis_B           184.9626         1
## Hepatitis_B:BMI                     507.8971         1
## Incidents_HIV:Alcohol_consumption    12.5762         1
## Alcohol_consumption:BMI             609.1650         1
## Hepatitis_B:Alcohol_consumption      93.1673         1
## Schooling:Life_expectancy          3612.0175         1
## Population_mln:Life_expectancy     5801.5294         1
## Life_expectancy:BMI                5645.7021         1
## Hepatitis_B:Life_expectancy        1958.1720         1
## Schooling:Under_five_deaths         285.1158         1
## Population_mln:Under_five_deaths    745.0685         1
## Incidents_HIV:Under_five_deaths     180.2365         1
## Under_five_deaths:BMI             20518.1890         1
## Hepatitis_B:Under_five_deaths       845.6424         1
## Alcohol_consumption:Under_five_deaths   190.5599         1
## Schooling:Infant_deaths             306.7278         1
## Population_mln:Infant_deaths        789.0086         1
## Incidents_HIV:Infant_deaths         216.6011         1
## BMI:Infant_deaths                 16592.7406         1
## Hepatitis_B:Infant_deaths           912.1509         1
## Alcohol_consumption:Infant_deaths   186.9291         1
## Life_expectancy:Infant_deaths      1397.2303         1
## Under_five_deaths:Infant_deaths     351.2771         1
##
## Multicollinearity may be due to Population_mln Schooling Incidents_HIV Hepatitis_B Alcohol_consumption Under_f
ive_deaths Life_expectancy BMI Infant_deaths Population_mln:Schooling Schooling:Incidents_HIV Schooling:Hepatitis
_B Incidents_HIV:Hepatitis_B Hepatitis_B:BMI Incidents_HIV:Alcohol_consumption Alcohol_consumption:BMI Hepatitis_
B:Alcohol_consumption Schooling:Life_expectancy Population_mln:Life_expectancy Life_expectancy:BMI Hepatitis_B:Li
fe_expectancy Schooling:Under_five_deaths Population_mln:Under_five_deaths Incidents_HIV:Under_five_deaths Under_
five_deaths:BMI Hepatitis_B:Under_five_deaths Alcohol_consumption:Under_five_deaths Schooling:Infant_deaths Popul
ation_mln:Infant_deaths Incidents_HIV:Infant_deaths BMI:Infant_deaths Hepatitis_B:Infant_deaths Alcohol_consumpti
on:Infant_deaths Life_expectancy:Infant_deaths Under_five_deaths:Infant_deaths regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ==================================
```

Figure 9: Multicollinearity test with the VIF Multicollinearity Diagnostic.

The VIF test revealed, that we do have multicollinearity of the variables in the model. Therefore, we built a model with a “centering” the dataset.

5. Outliers

First, we look at the residuals vs. leverage plot in Figure 10. Although the data has some extreme values, these are not necessarily crucial for driving the regression line. This means that the results of the analysis

would hardly vary whether we include these extreme values or not. This can be recognized by the fact that there are no data points in the upper right and lower right corners of the plot.

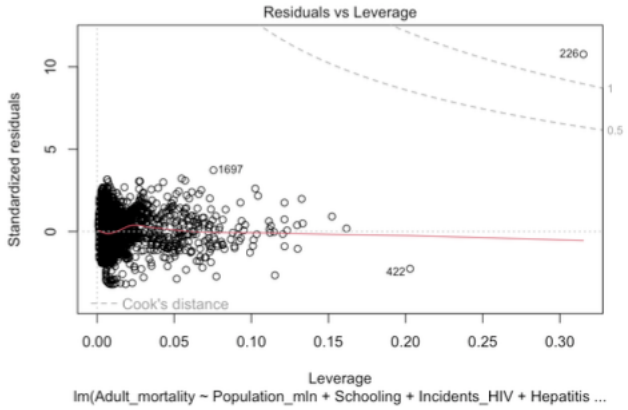

Figure 10: Residuals vs Leverage plot for final best fit interaction regression model.

Data points that are located far away horizontally from a regression line have a high leverage effect. These particular data points have the potential to significantly influence the slope of the calculated least squares line and should therefore be removed. In many cases, this method identified outliers that we removed later in our analysis.

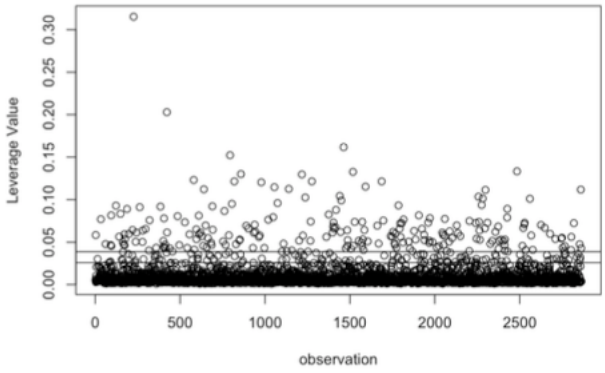

Figure 11: Leverage plot for final best fit interaction regression model.

The above adjustments for the unsuccessful regression assumption tests are carried out below
1.Model after centering the data:

```
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             2.625e+03  1.862e+02  14.098  < 2e-16
## Population_mln                          3.823e-01  2.102e-01   1.819 0.069068
## Schooling                               6.843e+01  6.436e+00  10.633  < 2e-16
## Incidents_HIV                           1.309e+01  3.136e+00   4.174 3.08e-05
## Hepatitis_B                             9.226e-01  8.898e-01   1.037 0.299889
## Alcohol_consumption                    -1.196e+01  2.312e+00  -5.171 2.49e-07
## Under_five_deaths                       8.668e+00  1.109e+00   7.816 7.63e-15
## Life_expectancy                        -2.869e+01  2.353e+00 -12.193  < 2e-16
## BMI                                    -9.501e+01  6.222e+00 -15.272  < 2e-16
## Infant_deaths                          -1.420e+01  1.644e+00  -8.636  < 2e-16
## Population_mln:Schooling                1.858e-02  2.878e-03   6.456 1.26e-10
## Schooling:Incidents_HIV                 4.150e-01  1.843e-01   2.251 0.024450
## Population_mln:Incidents_HIV            9.929e-02  1.420e-02   6.993 3.34e-12
## Schooling:Hepatitis_B                  -4.001e-02  1.565e-02  -2.557 0.010612
## Incidents_HIV:Hepatitis_B              -8.028e-02  2.348e-02  -3.419 0.000637
## Hepatitis_B:BMI                         4.727e-02  1.625e-02   2.908 0.003661
## Incidents_HIV:Alcohol_consumption      -3.461e-01  9.748e-02  -3.551 0.000390
## Alcohol_consumption:BMI                 2.691e-01  8.285e-02   3.248 0.001176
## Hepatitis_B:Alcohol_consumption         6.957e-02  9.309e-03   7.473 1.04e-13
## Schooling:Life_expectancy              -7.949e-01  7.807e-02 -10.182  < 2e-16
## Population_mln:Life_expectancy         -7.376e-03  2.765e-03  -2.668 0.007678
## Life_expectancy:BMI                     1.073e+00  7.762e-02  13.829  < 2e-16
## Hepatitis_B:Life_expectancy            -3.213e-02  1.005e-02  -3.198 0.001399
## Schooling:Under_five_deaths             1.728e-01  3.520e-02   4.910 9.65e-07
## Population_mln:Under_five_deaths       -3.204e-03  1.439e-03  -2.227 0.026036
## Incidents_HIV:Under_five_deaths         2.430e-01  1.967e-02  12.354  < 2e-16
## Under_five_deaths:BMI                  -3.341e-01  5.086e-02  -6.568 6.05e-11
## Hepatitis_B:Under_five_deaths          -3.022e-02  3.268e-03  -9.248  < 2e-16
## Alcohol_consumption:Under_five_deaths -1.726e-01  2.280e-02  -7.571 4.98e-14
## Schooling:Infant_deaths                -5.614e-01  5.481e-02 -10.241  < 2e-16
## Population_mln:Infant_deaths            4.018e-03  1.961e-03   2.049 0.040533
## Incidents_HIV:Infant_deaths            -3.995e-01  3.461e-02 -11.543  < 2e-16
## BMI:Infant_deaths                       9.192e-01  7.383e-02  12.451  < 2e-16
## Hepatitis_B:Infant_deaths               4.430e-02  5.488e-03   8.071 1.02e-15
## Alcohol_consumption:Infant_deaths       2.975e-01  3.589e-02   8.290  < 2e-16
## Life_expectancy:Infant_deaths          -1.064e-01  9.131e-03 -11.657  < 2e-16
## Under_five_deaths:Infant_deaths        -1.151e-02  1.551e-03  -7.423 1.51e-13
##
## (Intercept)                            ***
## Population_mln                         .
## Schooling                              ***
## Incidents_HIV                          ***
## Hepatitis_B
## Alcohol_consumption                    ***
## Under_five_deaths                      ***
## Life_expectancy                        ***
## BMI                                    ***
## Infant_deaths                          ***
## Population_mln:Schooling               ***
## Schooling:Incidents_HIV                *
## Population_mln:Incidents_HIV           ***
## Schooling:Hepatitis_B                  *
## Incidents_HIV:Hepatitis_B              ***
## Hepatitis_B:BMI                        **
## Incidents_HIV:Alcohol_consumption      ***
## Alcohol_consumption:BMI                **
## Hepatitis_B:Alcohol_consumption        ***
## Schooling:Life_expectancy              ***
## Population_mln:Life_expectancy         **
## Life_expectancy:BMI                    ***
## Hepatitis_B:Life_expectancy            **
## Schooling:Under_five_deaths            ***
## Population_mln:Under_five_deaths       *
## Incidents_HIV:Under_five_deaths        ***
## Under_five_deaths:BMI                  ***
## Hepatitis_B:Under_five_deaths          ***
## Alcohol_consumption:Under_five_deaths ***
## Schooling:Infant_deaths                ***
## Population_mln:Infant_deaths           *
## Incidents_HIV:Infant_deaths            ***
## BMI:Infant_deaths                      ***
## Hepatitis_B:Infant_deaths              ***
## Alcohol_consumption:Infant_deaths      ***
## Life_expectancy:Infant_deaths          ***
## Under_five_deaths:Infant_deaths        ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.69 on 2827 degrees of freedom
## Multiple R-squared:  0.9739, Adjusted R-squared:  0.9735
## F-statistic:  2927 on 36 and 2827 DF,  p-value: < 2.2e-16
```

Figure 12: Interaction Model after centering the data.

2.Model and Leverage points as well as Cook's Distance test after removing outliers:

```
##
## Call:
## lm(formula = Adult_mortality ~ Population_mln + Schooling + Incidents_HIV +
##     Hepatitis_B + Alcohol_consumption + Under_five_deaths + Life_expectancy +
##     BMI + Infant_deaths + Population_mln * Schooling + Incidents_HIV *
##     Schooling + Incidents_HIV * Population_mln + Hepatitis_B *
##     Schooling + Hepatitis_B * Incidents_HIV + Hepatitis_B * BMI +
##     Alcohol_consumption * Incidents_HIV + Alcohol_consumption *
##     BMI + Alcohol_consumption * Hepatitis_B + Life_expectancy *
##     Schooling + Life_expectancy * Population_mln + Life_expectancy:BMI +
##     Life_expectancy * Hepatitis_B + Under_five_deaths * Schooling +
##     Under_five_deaths * Population_mln + Under_five_deaths *
##     Incidents_HIV + Under_five_deaths * BMI + Under_five_deaths *
##     Hepatitis_B + Under_five_deaths * Alcohol_consumption + Infant_deaths *
##     Schooling + Infant_deaths * Population_mln + Infant_deaths *
##     Incidents_HIV + Infant_deaths * BMI + Infant_deaths * Hepatitis_B +
##     Infant_deaths * Alcohol_consumption + Infant_deaths * Life_expectancy +
##     Infant_deaths * Under_five_deaths, data = centered_cleaned_life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.5853  -3.3953   0.3434   3.4310  19.8290
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          3.446e+03  1.243e+03   2.772  0.00588
## Population_mln                      -1.447e+00  2.456e+00  -0.589  0.55625
## Schooling                            1.463e+01  3.420e+01   0.428  0.66911
## Incidents_HIV                       -2.478e+03  4.619e+02  -5.365 1.49e-07
## Hepatitis_B                         -4.079e+01  9.405e+00  -4.337 1.90e-05
## Alcohol_consumption                  5.945e+01  1.018e+01   5.839 1.22e-08
## Under_five_deaths                   -5.190e+01  1.960e+02  -0.265  0.79134
## Life_expectancy                     -1.160e+01  1.352e+01  -0.858  0.39129
## BMI                                 -6.439e+01  4.058e+01  -1.587  0.11352
## Infant_deaths                        7.048e+00  2.385e+02   0.030  0.97644
## Population_mln:Schooling            -4.341e-02  2.889e-02  -1.503  0.13380
## Schooling:Incidents_HIV              4.074e+01  2.216e+01   1.838  0.06689
## Population_mln:Incidents_HIV        -6.702e+00  2.431e+00  -2.757  0.00615
## Schooling:Hepatitis_B               -2.154e-02  1.352e-01  -0.159  0.87350
## Incidents_HIV:Hepatitis_B            1.936e+01  4.829e+00   4.010 7.47e-05
## Hepatitis_B:BMI                      8.029e-01  2.209e-01   3.635  0.00032
## Incidents_HIV:Alcohol_consumption    1.883e+01  9.715e+00   1.938  0.05347
## Alcohol_consumption:BMI             -2.873e+00  4.790e-01  -5.997 5.10e-09
## Hepatitis_B:Alcohol_consumption      2.008e-01  7.326e-02   2.741  0.00644
## Schooling:Life_expectancy           -1.322e-01  3.573e-01  -0.370  0.71158
## Population_mln:Life_expectancy       4.134e-02  2.796e-02   1.478  0.14021
## Life_expectancy:BMI                 -7.683e-02  4.318e-01  -0.178  0.85888
## Hepatitis_B:Life_expectancy          1.554e-01  7.254e-02   2.142  0.03289
## Schooling:Under_five_deaths         -8.553e+00  3.933e+00  -2.175  0.03034
## Population_mln:Under_five_deaths    -4.044e-01  4.165e-01  -0.971  0.33227
## Incidents_HIV:Under_five_deaths     -3.829e+02  1.634e+02  -2.343  0.01970
## Under_five_deaths:BMI                5.370e+00  7.336e+00   0.732  0.46467
## Hepatitis_B:Under_five_deaths       -1.517e-01  9.506e-01  -0.160  0.87326
## Alcohol_consumption:Under_five_deaths  6.917e+00  1.573e+00   4.396 1.47e-05
## Schooling:Infant_deaths              9.902e+00  4.073e+00   2.431  0.01558
## Population_mln:Infant_deaths         2.923e-01  4.797e-01   0.609  0.54266
## Incidents_HIV:Infant_deaths          5.405e+02  1.926e+02   2.806  0.00530
## BMI:Infant_deaths                    8.373e-01  8.954e+00   0.094  0.92555
## Hepatitis_B:Infant_deaths            1.270e+00  1.106e+00   1.149  0.25152
## Alcohol_consumption:Infant_deaths   -9.182e+00  1.755e+00  -5.230 2.96e-07
## Life_expectancy:Infant_deaths       -2.678e+00  3.034e-01  -8.825  < 2e-16
## Under_five_deaths:Infant_deaths     -1.354e+00  4.880e-01  -2.774  0.00585
##
## (Intercept)                          **
## Population_mln
## Schooling
## Incidents_HIV                        ***
## Hepatitis_B                          ***
## Alcohol_consumption                  ***
## Under_five_deaths
## Life_expectancy
## BMI
## Infant_deaths
## Population_mln:Schooling
## Schooling:Incidents_HIV              .
## Population_mln:Incidents_HIV         **
## Schooling:Hepatitis_B
## Incidents_HIV:Hepatitis_B            ***
## Hepatitis_B:BMI                      ***
## Incidents_HIV:Alcohol_consumption    .
## Alcohol_consumption:BMI              ***
## Hepatitis_B:Alcohol_consumption      **
## Schooling:Life_expectancy
## Population_mln:Life_expectancy
## Life_expectancy:BMI
## Hepatitis_B:Life_expectancy          *
## Schooling:Under_five_deaths          *
## Population_mln:Under_five_deaths
## Incidents_HIV:Under_five_deaths      *
## Under_five_deaths:BMI
## Hepatitis_B:Under_five_deaths
## Alcohol_consumption:Under_five_deaths ***
## Schooling:Infant_deaths              *
## Population_mln:Infant_deaths
## Incidents_HIV:Infant_deaths          **
## BMI:Infant_deaths
## Hepatitis_B:Infant_deaths
## Alcohol_consumption:Infant_deaths    ***
## Life_expectancy:Infant_deaths        ***
## Under_five_deaths:Infant_deaths      **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.653 on 342 degrees of freedom
## Multiple R-squared:  0.963,  Adjusted R-squared:  0.9591
## F-statistic: 247.5 on 36 and 342 DF,  p-value: < 2.2e-16
```

Figure 13: Interaction Model after removing all the outliers.

2.Model after Cox-box transformation:

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.3290 -1.2111 -0.0315  1.2254 11.5492

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          2.724e+02  1.183e+01  23.019  < 2e-16 ***
Infant_deaths                       -1.923e+00  1.424e-01 -13.509  < 2e-16 ***
Under_five_deaths                    7.387e-01  9.915e-02   7.451 1.23e-13 ***
Life_expectancy                     -2.998e+00  1.546e-01 -19.392  < 2e-16 ***
Alcohol_consumption                 -1.244e+00  2.125e-01  -5.855 5.33e-09 ***
Hepatitis_B                          2.205e-01  5.748e-02   3.836 0.000128 ***
BMI                                 -8.046e+00  5.236e-01 -15.368  < 2e-16 ***
Incidents_HIV                        2.466e-01  6.692e-02   3.685 0.000233 ***
Population_mln                      -2.312e-03  2.949e-04  -7.841 6.28e-15 ***
Schooling                            4.933e+00  3.850e-01  12.816  < 2e-16 ***
Infant_deaths:Alcohol_consumption    3.347e-02  3.369e-03   9.935  < 2e-16 ***
Infant_deaths:BMI                    9.441e-02  6.856e-03  13.771  < 2e-16 ***
Infant_deaths:Incidents_HIV         -2.185e-02  3.118e-03  -7.008 3.01e-12 ***
Infant_deaths:Schooling             -4.871e-02  4.739e-03 -10.279  < 2e-16 ***
Under_five_deaths:Alcohol_consumption -2.028e-02  2.117e-03  -9.580  < 2e-16 ***
Under_five_deaths:Hepatitis_B       -4.731e-04  1.443e-04  -3.279 0.001053 **
Under_five_deaths:BMI               -3.809e-02  4.673e-03  -8.150 5.40e-16 ***
Under_five_deaths:Incidents_HIV      1.397e-02  1.809e-03   7.719 1.62e-14 ***
Under_five_deaths:Schooling          1.630e-02  3.153e-03   5.169 2.51e-07 ***
Life_expectancy:Hepatitis_B         -3.126e-03  7.543e-04  -4.144 3.52e-05 ***
Life_expectancy:BMI                  9.320e-02  6.790e-03  13.727  < 2e-16 ***
Life_expectancy:Schooling           -5.925e-02  4.885e-03 -12.130  < 2e-16 ***
Alcohol_consumption:Hepatitis_B      2.056e-03  7.185e-04   2.862 0.004238 **
Alcohol_consumption:BMI              4.799e-02  7.798e-03   6.155 8.58e-10 ***
Incidents_HIV:Population_mln         1.075e-02  1.050e-03  10.238  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.836 on 2839 degrees of freedom
Multiple R-squared:  0.969,     Adjusted R-squared:  0.9687
F-statistic:  3693 on 24 and 2839 DF,  p-value: < 2.2e-16
```

Figure 14: Result of final best fit interaction regression model with Cox-box transformation.

Retesting the regression assumptions after transforming the model data accordingly:

1.Linearity Assumption
H0= Linearity is not present (homoscedasticity
Ha= Linearity is present

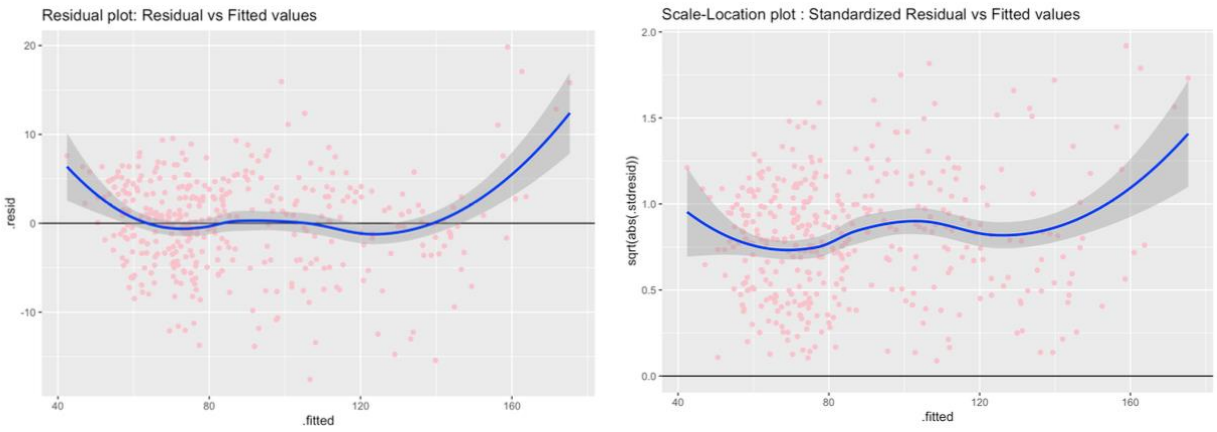

Figure 15: The scale-location plot and the residual plot for final best fit interaction regression model.

After centering the data and removing the outliers, it becomes apparent that a higher order model is more suitable for the analysis, as a curve/pattern can be detected. Following that, the data points are not evenly distributed around the zero line.

2.Equal Variance Assumption
H0= heteroscedasticity is not present (homoscedasticity
Ha= heteroscedasticity is present

```
        studentized Breusch-Pagan test

data:  bcmodel
BP = 299.34, df = 24, p-value < 2.2e-16
```

Figure 16: Homoscedasticity test result for final best fit interaction regression model after Cox-box transformation.

Additionally, we fail to reject the null hypothesis and conclude that this model does meet the equal variance assumption

3.Normality Assumption
H0= The sample data follows a normal distribution
Ha= The sample data does not follow a normal distribution

```
        Shapiro-Wilk normality test

data:  residuals(bcmodel)
W = 0.99532, p-value = 7.433e-08
```

Figure 17: Normality test result for final best fit interaction regression model after Cox-box transformation.

Even after adjustment, the Sharpio-Wilk test requirement at a significance level of 0.05 could not be met. Thus, we have sufficient evidence to conclude that the our dataset is not normally distributed.

4.Multicollinearity

```
##
## Call:
## imcdiag(mod = interact_recuded_model_life4, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                                             VIF detection
## Population_mln                         18906.3515         1
## Schooling                              22290.0878         1
## Incidents_HIV                           1672.1598         1
## Hepatitis_B                            17427.6066         1
## Alcohol_consumption                     7964.3656         1
## Under_five_deaths                    1062500.5286         1
## Life_expectancy                        14027.2637         1
## BMI                                     8837.0650         1
## Infant_deaths                        1220915.0061         1
## Population_mln:Schooling                325.7910         1
## Schooling:Incidents_HIV                 514.4426         1
## Population_mln:Incidents_HIV            129.7795         1
## Schooling:Hepatitis_B                  3074.0861         1
## Incidents_HIV:Hepatitis_B              1635.6659         1
## Hepatitis_B:BMI                       11050.2746         1
## Incidents_HIV:Alcohol_consumption       117.7064         1
## Alcohol_consumption:BMI                12410.4854         1
## Hepatitis_B:Alcohol_consumption         3821.4327         1
## Schooling:Life_expectancy              18424.4741         1
## Population_mln:Life_expectancy         15810.4401         1
## Life_expectancy:BMI                    18036.0285         1
## Hepatitis_B:Life_expectancy             8180.3797         1
## Schooling:Under_five_deaths            43263.2245         1
## Population_mln:Under_five_deaths       14398.8502         1
## Incidents_HIV:Under_five_deaths         6712.0603         1
## Under_five_deaths:BMI                 979234.6886         1
## Hepatitis_B:Under_five_deaths         229613.9216         1
## Alcohol_consumption:Under_five_deaths    9878.4190         1
## Schooling:Infant_deaths                34923.4598         1
## Population_mln:Infant_deaths           13865.1549         1
## Incidents_HIV:Infant_deaths             6914.2342         1
## BMI:Infant_deaths                    1129680.8090         1
## Hepatitis_B:Infant_deaths             239498.5114         1
## Alcohol_consumption:Infant_deaths       8860.9571         1
## Life_expectancy:Infant_deaths          10333.4403         1
## Under_five_deaths:Infant_deaths          618.4819         1
```

Figure 18: Multicollinearity test with the VIF Multicollinearity Diagnostic.

After the adjustments, multicollinearity has increased. Now the variable Population_mln:Incidents_HIV, which was not previously affected by multicollinearity, is also affected.

5.Outliers

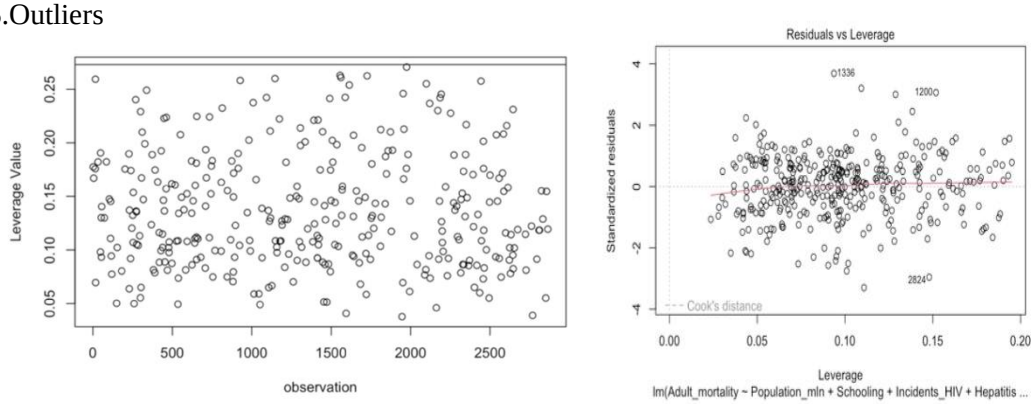

Figure 19: Residuals vs Leverage as well as leverage plot for final interaction regression model after removing outliers.

However, the outliers have improved significantly. The various tests show a dataset that does not contain any influential outliers.

To sum up, we could see from both normality and homoscedasticity check (Figure 16, Figure 17) that we cannot reject the null hypothesis at significant level of $\alpha = 0.05$, which means the residuals are not normally distributed and heteroscedasticity is present. In addition, the multicollinearity tests and linearity tests have deteriorated compared to the raw dataset even though all significant and influential outliers have been removed. Moreover, we found that the adjusted R-square for the model after Cox-box transformation equals to 0.9687 (Figure 14), which is lower than original final model, which is 0.9735 (Figure 4). Therefore, we would like to discard the model after Cox-box transformation and still use the original final model.

4. CONCLUSION AND DISCUSSION

4.1 Discussion

Table 3: Final results of our research.

Hypothesis	**Has the hypothesis proven true?**
Infant deaths have a significant positive impact on adult mortality (higher adult mortality with higher infant deaths and vice versa).	FALSE
Under five deaths have a significant positive impact on adult mortality (higher adult mortality with number of under five deaths and vice versa).	TRUE
Life expectancy has a significant positive impact on the adult mortality (higher adult mortality with higher age and vice versa)	TRUE
Alcohol consumption has a significant positive impact on adult mortality (higher adult mortality with higher alcohol consumption and vice versa).	TRUE

Hepatitis_B has a significant negative impact on the adult mortality (less adult mortality with higher number of Hepatitis_B vaccinations and vice versa).	FALSE
Measles has a significant negative impact on adult mortality (less adult mortality with higher number of Measles vaccinations and vice versa).	FALSE
Diphtheria has a significant negative impact on the adult mortality (less adult mortality with higher number of Diphtheria vaccinations and vice versa).	FALSE
Polio has a significant negative impact on adult mortality (less adult mortality with higher number of Polio vaccinations and vice versa).	FALSE
BMI has a significant negative impact on adult mortality (less adult mortality with higher BMI and vice versa).	FALSE
Incidents_HIV has a significant positive impact on the adult mortality (higher adult mortality with higher number of HIV incidents and vice versa).	TRUE
GDP_per_capita has a significant negative impact on the adult mortality (less adult mortality with higher GDP and vice versa).	FALSE
Population_mln has a significant positive impact on the adult mortality (higher adult mortality with higher Population and vice versa).	TRUE
Schooling has a significant negative impact on adult mortality (less adult mortality with better education and vice versa).	FALSE
Economy_status_Developed has a significant positive impact on adult mortality (higher adult mortality when it is a developing country and vice versa).	FALSE
The number of under-five deaths is directly related to infant deaths, as infant deaths in the 0 to 1 year age group are included in the total number of under-five deaths.	TRUE
Economy and GDP interact with each other, as a lower GDP is often a characteristic of a developing country.	NA
Infant mortality and under-five mortality interact with the average life expectancy, as high	TRUE

mortality in these age groups reduces overall life expectancy.	
The risk of HIV infection interacts with the average life expectancy, as despite advanced treatment options, HIV-positive people often have a lower life expectancy, which is exacerbated by good healthcare and adequate treatment. This can also be linked to GDP, which can be taken as an indicator of the economic and social conditions/health care supply.	FALSE
Vaccinations also interact with the average life expectancy, as they prevent diseases that can be potentially fatal.	TRUE
The level of education interacts with HIV incidents, as better educated people tend to have a higher awareness of preventive measures and better access to health services.	TRUE

Interaction Model

An increase in child/infant mortality is accompanied by a decrease in adult mortality. In societies with high infant mortality, demographic changes could occur that lead to a disproportionate survival of healthier adults, so that selective survival mechanisms are predominant that favor more robust and resilient adults who live longer or have low susceptibility to diseases. However, adult mortality could also be lower in regions with high infant mortality because the population is smaller in rural areas/small villages with high infant mortality.

In contrast, an increase in deaths among children under the age of five increases adult mortality. This could be due to the fact that a high child mortality rate is often associated with poor general health and inadequate medical care, which in turn leads to poor infectious living conditions.

Higher life expectancy leads to lower mortality in adulthood. This could be due to the fact that longer life expectancy is often associated with better living conditions, better access to health services and higher general prosperity which promote the general health of the population and reduce the risk of serious diseases and, therefore, the overall mortality rate in adulthood.

On the other hand, an increase in alcohol consumption correlates with a decrease in adult mortality. This is caused by higher alcohol consumption being associated with greater social interaction or participation in social activities, which can have a positive effect on mental health (as for example stress reduction) and, thus, reduce mortality.

The more hepatitis B vaccinations administered, the higher the mortality rate among adults. This could indicate that in populations where vaccination is recommended, this disease is highly prevalent and leads to a higher mortality rate. At the same time, this could also indicate possible complications of this vaccination or a general imbalance/fail in health care where inadequate treatment prevails despite the availability of vaccinations.

Similarly, an increase in body mass index (BMI) leads to a reduction in mortality in adults. As a result, it appears that in certain populations a higher BMI is associated with a better diet or higher social status, which ultimately has a positive impact on health and therefore mortality.

In contrast, an increase in new HIV cases shows an increase in mortality in adulthood. This is because a higher incidence of HIV is often associated with an increase in associated health risks and complications. People infected with the virus have an increased risk of serious illness and reduced life expectancy, leading to higher mortality.

With a larger population, adult mortality is also increasing. This could be caused by an overburdening of healthcare systems or increased competition for resources such as medical care. In addition, greater population growth can also lead to a greater spread of diseases, which is reflected in higher mortality.

Finally, the increase in average school attendance shows adult mortality. This could indicate that people with higher educational qualifications and academic/higher positions lead an unhealthier lifestyle such as sitting at desk jobs, sedentary behavior, stress, and inadequate self-care and, therefore, a higher mortality rate.

However, the additive model has three deviations from the interactive model:
Firstly, increased alcohol consumption can lead to various health problems, such as liver disease, cardiovascular disease, and an increased risk of accidents. These health problems, therefore, have a significantly increasing effect on adult mortality.

Secondly, an increase in hepatitis B vaccinations has a positive effect on the health of the population. Since hepatitis B can lead to serious complications, a high vaccination rate contributes to a reduction in adult mortality, as it reduces the severe course of the disease and the associated complications.

Thirdly, population size also has a positive influence on mortality. In a growing population, better health care and infrastructure associated with a larger population can reduce adult mortality. In addition, in a large population, younger age groups may be more prevalent, which has a positive impact on adult mortality or adult deaths, as they are less likely to die from fatal diseases such as cancer.

To summarize, although there are several factors that have a decisive influence on mortality in adulthood, three four factors stand out as particularly strong influencing factors in the two models.

One of the most important aspects is HIV prevalence, as a high incidence of HIV infection can lead to a significant increase in mortality. People living with HIV are often more susceptible to other diseases, which reduces their life expectancy.
People who live longer are more susceptible to age-related, deadly and progressive chronic diseases. Access to medical care and preventive measures also becomes more difficult for older people with reduced mobility. I addition, isolation in old age has a negative impact on mental and, therefore. physical health.
Finally, the level of education of the population is also important as high degress/education is often closely linked to a stressful job and unhealthy lifestyles, as the so-called high archievers work longer in often intellectual/non-physical occupations. People with longer education also often have higher financial and social pressure (e.g. student debt) and standards of living, which can put pressure on them to be constantly productive and functional. These demands can in turn have a negative impact on physical and mental health.

4.2 Approach

Overall, the analysis of adult mortality reveals a complex interplay of various factors influencing health outcomes across different populations. As demographic trends shift towards increased life expectancy and a growing elderly population, understanding these factors becomes even more critical. The interplay of age, socioeconomic status, environmental conditions, and health behaviors—such as alcohol consumption—highlights the multifaceted nature of adult mortality. Research indicates that while age is a significant contributor to mortality risk, it is not the sole determinant. Early life conditions, including healthcare access and nutrition, play a pivotal role in shaping adult health. The evidence presented suggests that poor living conditions and high infant mortality rates are indicative of broader systemic issues that ultimately affect adult mortality outcomes. Furthermore, the disparities between developed and developing nations underline the importance of socioeconomic factors and environmental health in influencing life expectancy. Considering these findings, it is essential for policymakers, health institutions, and the pharmaceutical industry to engage in targeted interventions that address these determinants. By focusing on prevention strategies, improving healthcare access, and addressing environmental issues, we can work towards reducing adult mortality rates globally. Continued research and collaboration across disciplines will be crucial to developing effective solutions that enhance health outcomes for all age groups.

To sum up, the approach we chose to investigate the relationships between various variables and adult mortality shows fascinating results as significant and relevant correlations could be identified and interpreted. However, the question arises as to whether a different methodology or a modification of the current approach could possibly provide a more accurate model that could be used for prediction as the data set does not meet the fundamental assumptions of a linear regression model. Therefore, alternative approaches that better address the specific characteristics of the data should be considered. It is possible that the use of non-linear models or other statistical methods would be more effective in order to achieve a more reliable, precise and validated result.

4.3 Future Work

One major aspect that affected the present study is the fact that our data set only includes information up to 2015. This time limitation could mean that important developments and trends that have occurred since then are not taken into account. In potential future research, it would be useful to collect current data and perform a comprehensive regression analysis based on this new information.

Furthermore, expanding the research framework to include additional variables could be interesting. In particular, qualitative factors related to the psychological aspects of adult mortality could be included in the analysis as, for example, our hypothesized factors such as stress levels, social isolation or individual coping mechanisms. Following that, a data set with different quantitative variables that are more suitable for the regression analysis and fulfill the regression assumptions could also be utilized. Lastly, a higher order model could also be tested, as there were indicators of relationships in our linear regression analysis that could be described more accurately by that type of model.

5. REFERENCES

Chartier, K., & Caetano, R. (2010). Ethnicity and health disparities in alcohol research. Alcohol research & health: the journal of the National Institute on Alcohol Abuse and Alcoholism. https://pmc.ncbi.nlm.nih.gov/articles/PMC3887493/
Gochiashvili, L. (2023). Life Expectancy (WHO) Fixed. Kaggle. https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated/data
Hoffman C. 1998. Uninsured in America: A Chart Book. The Kaiser Commission on Medicaid and the Uninsured.
Klatsky, A. L., Armstrong, M. A., & Friedman, G. D. (1992). Alcohol and mortality. Annals of Internal Medicine, 117(8), 646–654. https://doi.org/10.7326/0003-4819-117-8-646
Kumar Rajarshi. (2018, February 10). Life expectancy (WHO). Kaggle. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
Kruk, M. E., Gage, A. D., Joseph, N. T., Danaei, G., García-Saisó, S., & Salomon, J. A. (2018). Mortality due to low-quality health systems in the Universal Health Coverage Era: A systematic analysis of amenable deaths in 137 countries. The Lancet, 392(10160), 2203–2212. https://doi.org/10.1016/s0140-6736(18)31668-4
Mackenbach, J. P., Kulhánová, I., Bopp, M., Borrell, C., Deboosere, P., Kovács, K., Looman, C. W., Leinsalu, M., Mäkelä, P., Martikainen, P., Menvielle, G., Rodríguez-Sanz, M., Rychtaříková, J., & de Gelder, R. (2015). Inequalities in alcohol-related mortality in 17 European countries: A retrospective analysis of mortality registers. PLOS Medicine, 12(12). https://doi.org/10.1371/journal.pmed.1001909
Rogers, R. G., Hummer, R. A., & Krueger, P. M. (2005). Adult mortality. Handbooks of Sociology and Social Research, 283–309. https://doi.org/10.1007/0-387-23106-4_11
Rostron, B. (2012). Alcohol consumption and mortality risks in the USA. Alcohol and Alcoholism, 47(3), 334–339. https://doi.org/10.1093/alcalc/agr171
Roy, D. N., Biswas, M., Islam, E., & Azam, Md. S. (2022). Potential factors influencing COVID-19 vaccine acceptance and hesitancy: A systematic review. PLOS ONE, 17(3). https://doi.org/10.1371/journal.pone.0265496
Topalian, S. L., Drake, C. G., & Pardoll, D. M. (2012). Targeting the PD-1/B7-H1(PD-L1) pathway to activate anti-tumor immunity. Current Opinion in Immunology, 24(2), 207–212. https://doi.org/10.1016/j.coi.2011.12.009
Pozzer, A., Anenberg, S. C., Dey, S., Haines, A., Lelieveld, J., & Chowdhury, S. (2023). Mortality attributable to ambient air pollution: A Review of Global Estimates. GeoHealth, 7(1). https://doi.org/10.1029/2022gh000711
Wong, D. W., & Li, Y. (2020). Spreading of COVID-19: Density matters. PLOS ONE, 15(12). https://doi.org/10.1371/journal.pone.0242398
Singer, B., & Ryff, C. D. (2001). New Horizons in Health: An Integrative Approach. National Academy Press.
Zhao, J., Stockwell, T., Naimi, T., Churchill, S., Clay, J., & Sherk, A. (2023). Association between daily alcohol intake and risk of all-cause mortality. JAMA Network Open, 6(3). https://doi.org/10.1001/jamanetworkopen.2023.6185

APPENDIX

Refer to RMD and xlsx file for more details on the dataset and code used.

End of Project Report