Multiple Regression Analysis:

Measuring Life Expectancy by Predictive Health Factors

DATA 603, L01

Connor Horemans, Ethan Scott, (Adrian) James Huvenaars

Table of Contents

Introduction

In today's world, everyone wants to live a longer and healthier life. It is, therefore, essential to understand what factors determine life expectancy. According to the Harvard School of Public Health, lifestyle choices such as maintaining a healthy diet, exercising regularly, and avoiding harmful habits are crucial in enhancing life expectancy (2022). Various studies have also highlighted the complex relationship between life expectancy and factors such as socioeconomic conditions and health-related elements (Miladinov, 2020).

Given the importance of life expectancy as an indicator of a population's well-being, our research project aims to investigate the impact of critical health-related factors on life expectancy in developing countries. Our primary research question is to determine these factors' specific influence and implications for life expectancy in diverse developing contexts. By addressing this gap in the current understanding, we hope to provide valuable insights that can inform health policies, guide targeted funding, and empower nations to optimize their efforts to improve their citizens' longevity and well-being. This research is not just an academic pursuit but a quest to contribute actionable knowledge that can make a real difference in people's lives.

Methodology

Data

Our study utilizes a comprehensive life expectancy dataset from Kaggle, originating from the World Health Organization (WHO) and the United Nations. Health data from the WHO’s Global Health Observatory (GHO) and economic data from the United Nations website were amalgamated, resulting in a dataset with 22 columns and 2938 rows, covering a range of economic and health variables linked to life expectancy (Life Expectancy(WHO), n.d.).

After obtaining the dataset, we conducted preliminary cleaning, addressing missing values. We selectively retained " Developing " rows in the Status column to narrow our focus to developing countries.

With a specific interest in proactive health-related factors influencing life expectancy, we strategically selected intervention-related health variables. For instance, variables such as alcohol consumption, immunization coverage, and body mass index were considered, while retroactive measures like mortality rates were intentionally omitted.

Selected Variables for Multiple Linear Regression:

- **Life expectancy:** The life expectancy of a country (Unit: Age) - Quantitative Dependent variable
- **Alcohol:** Alcohol Consumption per capita (15+) (Unit: litres of pure alcohol) - Quantitative independent variable
- **Hepatitis B:** HepB immunization coverage among 1-year-olds (Unit: Percentage) - Quantitative independent variable
- **Measles:** Number of measles cases per 1000 (Unit: Rate of per 1000) - Quantitative independent variable
- **BMI:** Average body mass index of the entire country population (Unit: BMI) - Quantitative independent variable
- **Polio:** Polio immunization coverage among 1-year-olds (Unit: Percentage) - Quantitative independent variable
- **Diphtheria:** Diphtheria, Tetanus and Pertussis immunization coverage among 1 year olds (Unit: percentage) - Quantitative independent variable
- **HIV/AIDS:** HIV deaths in 0-4 year olds per 1000 live births (Unit: Rate of per 1000 live births) - Quantitative independent variable

These variables, meticulously chosen for their relevance to health outcomes, underpin our multiple linear regression analysis, ensuring a nuanced understanding of the intricate relationship between proactive health factors and life expectancy in developing countries.

Approach

Approaching the exploration of life expectancy in developing countries, we'll employ the methods from Data 603, specifically, multiple linear regression. The order of steps to conduct our analysis involves building an initial full additive model containing all of our desired variables of interest (health-related predictors and response variable life expectancy). An individual t-test will then determine if any of the variables can be removed from our model, reducing our initial full additive model.

After confirming the main effects, we will build a full interaction model. The interaction term model will then be reduced through stepwise selection. After reducing the interaction model, the individual t-test will assess if higher-order terms should be added to the model. Once those steps have been conducted and we have selected the best model, the model will then undergo scrutiny for six key assumptions:

1. Linearity - Examined through residual plots.
2. Independence - Assessed by reviewing residuals against one another
3. Normality - Checked using the Shapiro-Wilk normality test.
4. Equal Variance (heteroscedasticity) - Verified through the Breusch-Pagan test.
5. Multicollinearity - Examined via VIF.
6. Outliers - Detected through Cook’s distance and leverage.

If any assumptions are not met, appropriate steps will be taken to try and further improve the model. These steps may include box cox transformations and removal of outliers. Once we have done all we can within the scope of the techniques learned in DATA 603, a final best-fit model will be chosen, and the beta coefficients will be interpreted.

Workload Distribution

Our team has collaboratively structured the workload, recognizing individual strengths and ensuring a balanced contribution from each member. Responsibilities are tailored to leverage specific skills, resulting in equitable efforts across the team.

Ethan: Ethan undertakes the crucial tasks of preliminary data cleaning, preparation, and the foundational construction of both the initial additive and interaction models.

Connor: Connor assumes responsibility for the subsequent phases of model building, employing techniques such as Box-Cox transformations and outlier removal to refine the predictive accuracy of the models.

James: James plays a pivotal role in the later stages, overseeing the verification of model assumptions, creating visualizations, and compiling the initial draft of the project report.

This distribution aligns with individual strengths, ensuring each member contributes meaningfully to the project's success. The allocation is not purely equal in task volume but is equitable, considering the inherent complexity of different tasks. We justified this tailored approach by the diverse skill sets within the team, optimizing efficiency and expertise.

Results

Model Selection Procedures:

Our starting model for our selection process is:

$$\widehat{Y_{Life\ Expectancy}} = \beta_0 + \beta_1 X_{Alcohol} + \beta_2 X_{Hepatitis\ B} + \beta_3 X_{BMI} + \beta_4 X_{Polio} + \beta_5 X_{Diphtheria} + \beta_6 X_{HIV.AIDS} + \beta_7 X_{Measles}$$

We then reduce the model using individual T-tests. The hypothesis for the individual t-test is as follows:

H(0): $\beta_i = 0$
H(A): $\beta_i \neq 0$
i = Alcohol, Hepatitis B, Measles, BMI, Polio, Diphtheria, and HIV.AIDS

Main Effects Individual T-test Values:

- *Alcohol:* t = 5.788, $p < 0.001$
- *Hepatitis B:* t = -2.490, p = 0.0129
- *Measles :* t = 0.694, p = 0.4875
- *BMI:* t = 19.986, $p < 0.001$
- *Polio:* t = 4.457, $p < 0.001$
- *Diphtheria:* t = 6.526, $p < 0.001$
- *HIV.AIDS:* t = -30.539, $p < 0.001$

Individual T-tests were used in our variable selection to determine the best predictors based on a significance level of α = 0.05. From the results of these tests, we would reject the null hypothesis in favour of the alternative for all our variables except Measles (since its p-value is greater than 0.05). This suggests that Alcohol, HepatitisB, BMI, Polio, Diphtheria, and HIV/AIDS are all significant predictors of life expectancy. For this reason, these variables will be added to our model for further comparison between interaction and higher-order terms. The variable Measles will not be added to our model since it was found to be an insignificant predictor. Our main effect model is shown below:

$$\widehat{Y_{Life\ Expectancy}} = \beta_0 + \beta_1 X_{Alcohol} + \beta_2 X_{Hepatitis\ B} + \beta_3 X_{BMI} + \beta_4 X_{Polio} + \beta_5 X_{Diphtheria} + \beta_6 X_{HIV.AIDS}$$

Adjusted R^2 = 0.6006

RMSE = 5.28

Following our main effects model, we created a full interaction model with all variables and reduced the model using stepwise regression.

Hypothesis Statement for Individual T-tests (Interaction Terms):

H(0): $\beta_i = 0$

H(A): $\beta_i \neq 0$

i = Polio: Diphtheria, BMI: Polio, Hepatitis. B: Alcohol, etc.

Interactions Term T-tests:

- *Polio:Diphtheria:* t = *2.439 0.01487 **
- *BMI:Polio:* t = *0.0003086*, p < 0.001
- *Hepatitis B:Alcohol:* t = *0.0018239*, p < 0.001
- *HIV.AIDS:Alcohol:* t = *0.0096871*, p < 0.001
- *Diphtheria:Alcohol:* t = *3.001, p = 0.00274*
- *Hepatitis B:Polio:* t = *1.810, p = 0.07058*
- *HIV.AIDS:Diphtheria:* t = *-1.784, p = 0.07471*

$$\widehat{Y_{Life\ Expectancy}} = \beta_0 + \beta_1 X_{Alcohol} + \beta_2 X_{Hepatitis\ B} + \beta_3 X_{BMI} + \beta_4 X_{Polio} + \beta_5 X_{Diphtheria} + \beta_6 X_{HIV.AIDS} + \beta_7 X_{Polio:Diphtheria} + \beta_8 X_{BMI:Polio} + \beta_9 X_{Hepatitis\ B:Alcohol} + \beta_{10} X_{HIV.AIDS:Alcohol} + \beta_{11} X_{Diphtheria:Alcohol} + \beta_{12} X_{Hepatitis\ B:Polio} + \beta_{13} X_{HIV.AIDS:Diphtheria}$$

Adjusted R^2 = 0.6259

$$\textit{RMSE} = 5.11$$

We then used a pairs plot (shown in Appendix A) to observe which predictor variables have a higher-order (non-linear) relationship with the response variable life expectancy. The only variable that appeared to have a higher-order relationship was HIV.AIDS. In adding a quadratic term to the model, the model significantly improved its adjusted R^2 and RMSE values. Therefore, we kept it in our model. We also checked individual t-test values to ensure that the higher-order variable was to be kept in the model. The t-test hypothesis is as follows:

Hypothesis Statement for Individual T-tests (Higher Order Terms):

$$H(0): \beta_i = 0$$

$$H(A): \beta_i \neq 0$$

$$i = \text{Alcohol}^2, \text{Hepatitis B}^2, \text{BMI}^2, \text{Polio}^2, \text{Diphtheria}^2, \text{and HIV.AIDS}^2$$

Higher Order Individual T-tests:

- *HIV.AIDS^2:* t = 15.310, $p < 0.001$

We have now reduced our model, added interactions, reduced it yet again and added higher power terms. Below is the model after doing all these steps. This is our best-fitted model with and without coefficient values:

$$\widehat{Y_{Life\ Expectancy}} = \beta_0 + \beta_1 X_{Alcohol} + \beta_2 X_{Hepatitis\ B} + \beta_3 X_{BMI} + \beta_4 X_{Polio} + \beta_5 X_{Diphtheria} + \beta_6 X_{HIV.AIDS} + \beta_7 X^2_{HIV.AIDS} + \beta_8 X_{Polio:Diphtheria} + \beta_9 X_{BMI:Polio} + \beta_{10} X_{Hepatitis\ B:Alcohol:} + \beta_{11} X_{HIV.AIDS:Alcohol} + \beta_{12} X_{Diphtheria:Alcohol} + \beta_{13} X_{Hepatitis\ B:Polio} + \beta_{14} X_{HIV.AIDS:Diphtheria}$$

$$\widehat{Y_{Life\ Expectancy}} = 56.0141191 + 0.7211983X_{\text{Alcohol}} - 0.0057522X_{\text{HepatitisB}} + 0.2779569X_{\text{BMI}} + 0.0297612X_{\text{Polio}} + 0.0119486X_{\text{Diphtheria}} - 1.3219465X_{\text{HIV.AIDS}} + 0.0242920X^2_{\text{HIV.AIDS}} + 0.0004922X_{\text{polio:Diphtheria}} - 0.0018225X_{\text{BMI:Polio}} - 0.0080679X_{\text{HepatitisB:Alcohol}} + 0.0026948X_{\text{HIV.AIDS:Alcohol}} + 0.0025360X_{\text{Diphtheria:Alcohol}} + 0.0003109X_{\text{HepatitisB:Polio}} - 0.0028802X_{\text{HIV.AIDS:Diphtheria}}$$

$$\textit{Adjusted } R^2 = 0.6796$$

$$\textit{RMSE} = 4.729$$

The adjusted R2 of 0.6796 indicated that approximately 67.96% of the variability in life expectancy is explained by the independent variables in our model. An RMSE of 4.729 means that the standard deviation of unexplained variance in the model is 4.729.

Multiple Linear Regression Assumptions

The sections below will be where the assumptions behind our model are tested. Multiple Linear Regression has various assumptions behind it, so we will test our model to ensure it meets those assumptions and check if it is valid.

Linearity Assumption

The linearity assumption assumes a straight-line (linear) relationship between the response variables and the predictor variable. To check for linear or non-linear patterns, we plotted a residuals vs fitted values plot (FIGURE 1). From the plot, we can see a distinct non-linear pattern, suggesting that we do not pass the assumption for linearity.

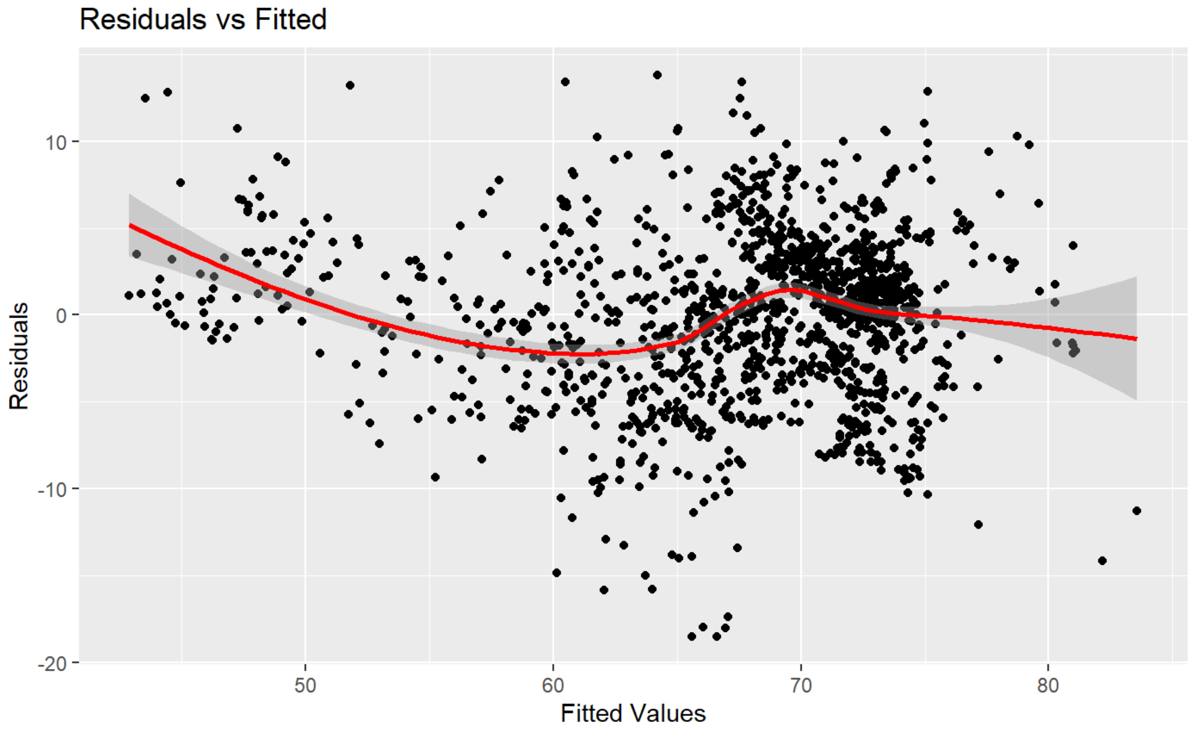


Figure 1: Plot of Residuals vs Fitted Values

Independence Assumption

The assumption of independence assumes that the error terms are independent. In the context of DATA 603, we assume this is the case, so the assumption of independence holds.

Normality Assumption

The normality assumption considers the residuals of the regression to be normally distributed. To verify this, we examined a normal Q-Q plot and a histogram of the residuals (Figure 2). The normal Q-Q plot indicates that the residuals follow the line closely, but a few points flare outwards at the lower tail, indicating the possible presence of outliers.

Additionally, there is a noticeable bump away from the line in the middle of the distribution. Upon looking at the histogram of residuals, we can see a generally normal pattern. However, some data points near the left tail suggest the distribution of residuals may not be normal.

To further confirm the normality assumption, we conducted a Shapiro-Wilk test. The null hypothesis for the Shapiro-Wilk test is:

$$H_o: The\ Residuals\ are\ Normally\ Distributed$$

$$H_A: The\ Residuals\ are\ not\ Normally\ Distributed$$

Based on a significant level of $\alpha = 0.05$, the results of the Shapiro-Wilk test are (W = 0.98798, p = 2.2443-09). Since the p-value is less than our significant level of 0.05, we reject the null hypothesis. From this, we can infer that the residuals are not normally distributed, so the normality assumption does not hold. It should be noted that the Shapiro-Wilk test typically fails with larger datasets and cannot always be relied upon in such scenarios. In our analysis, though, we observed that the graphs showed indications of non-normality, which was then confirmed by the findings of the Shapiro-Wilk test.

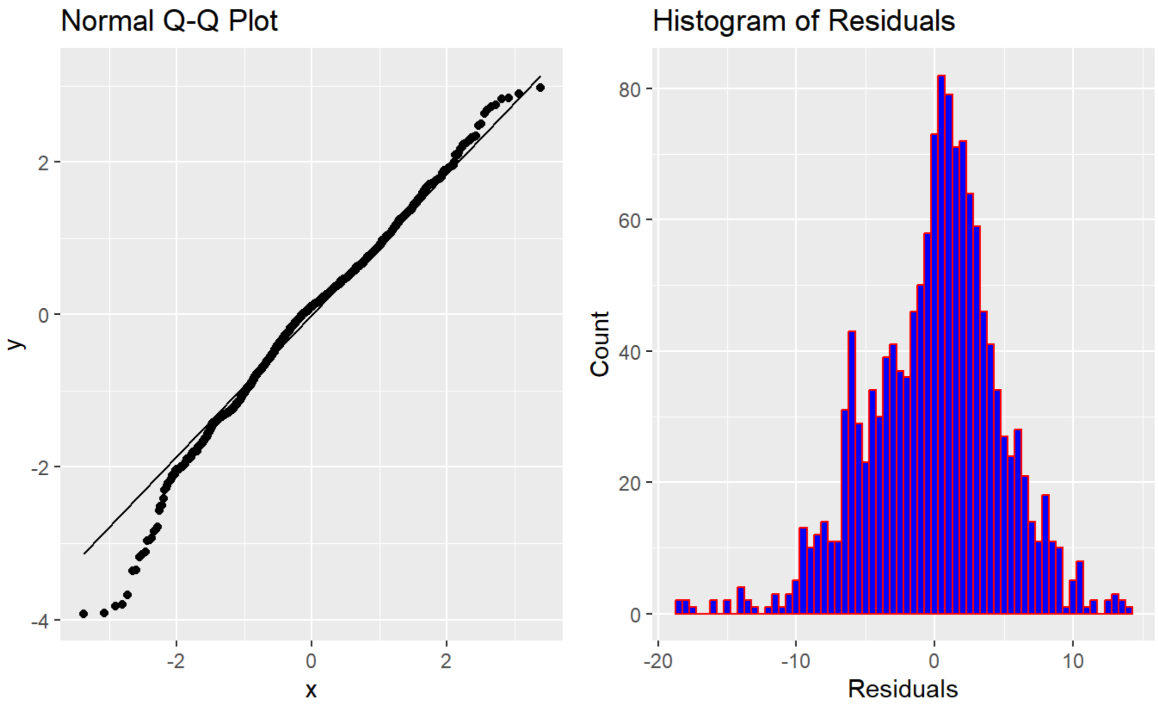


Figure 2: Plots for Normality Assumption

Equal Variance Assumption

The assumption of equal variance assumes that the error terms have a constant variance. To check this assumption, we plotted residuals vs fitted (FIGURE 3). Upon observing the plot, the points on the left side of the graph become narrower, while they tend to get wider towards the middle of the graph. This change in spread indicates that the assumption of equal variance is invalid.

To further confirm this, we conducted a Breusch-Pagan test. The hypothesis for this test is as follows:

$$H_o: \ Heteroscedascity\ is\ not\ present$$

$$H_A: \ Heteroscedascity\ is\ present$$

The results from the Breusch-Pagan test are BP = 108.16, p = 2.2e-16. When using a significance level of 0.05, we reject the null hypothesis. From this, we can conclude that heteroscedasticity is present, so the assumption of equal variance does not hold.

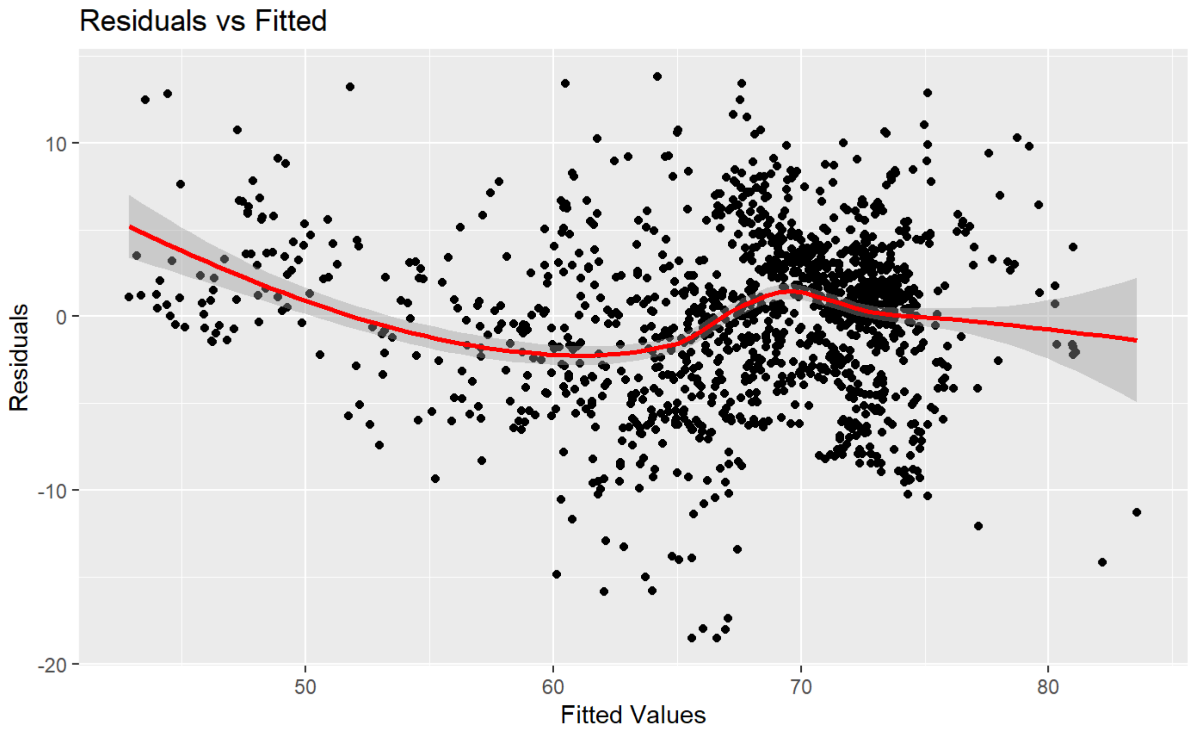


Figure 3: Plot for Equal Variances Assumption

Multicollinearity Tests

The assumption of multicollinearity assumes that the independent variables in the model are not highly correlated with each other. We conducted a multicollinearity analysis by examining the variance inflation factors (VIF) to decide which variables should be retained in our model. After conducting the test, we found that none of our variables had a high VIF value, including Alcohol (1.135326), Hepatitis B (1.670110), Measles (1.041770), BMI (1.172335), Polio (1.617654), Diphtheria (2.004144), and HIV/AIDS (1.069017). Therefore, we didn't need to drop any variables since none of the VIF values exceeded 5. From this, we can conclude that multicollinearity is not present, and this assumption holds.

Influential Points and Outliers

To assess influential points, we depict the residuals against Cook's distance, represented by a dashed line (FIGURE 4). The residuals vs. leverage plot reveals that no points surpass Cook's distance, indicating the absence of influential points disproportionately affecting our regression results.

To further determine if outliers are present in our data, we plotted a Cook’s Distance plot and a Leverage plot (FIGURE 5). The left plot illustrates Cook's distance for each observation, offering insight into the overall influence of outlier points on our regression. Noteworthy observations include numbers 119, 421, and 1187, which exhibit the highest Cook's distance. However, their Cook's Distance values are well below 0.5, rendering them non-influential. Subsequently, we employed the leverage plot on the right to eliminate outliers beyond 2p/n and 3p/n thresholds. Despite refitting our model for both thresholds, our model had a higher RMSE and Lower Adjusted R^2. Therefore, we elected to keep all points in our dataset in the final model.

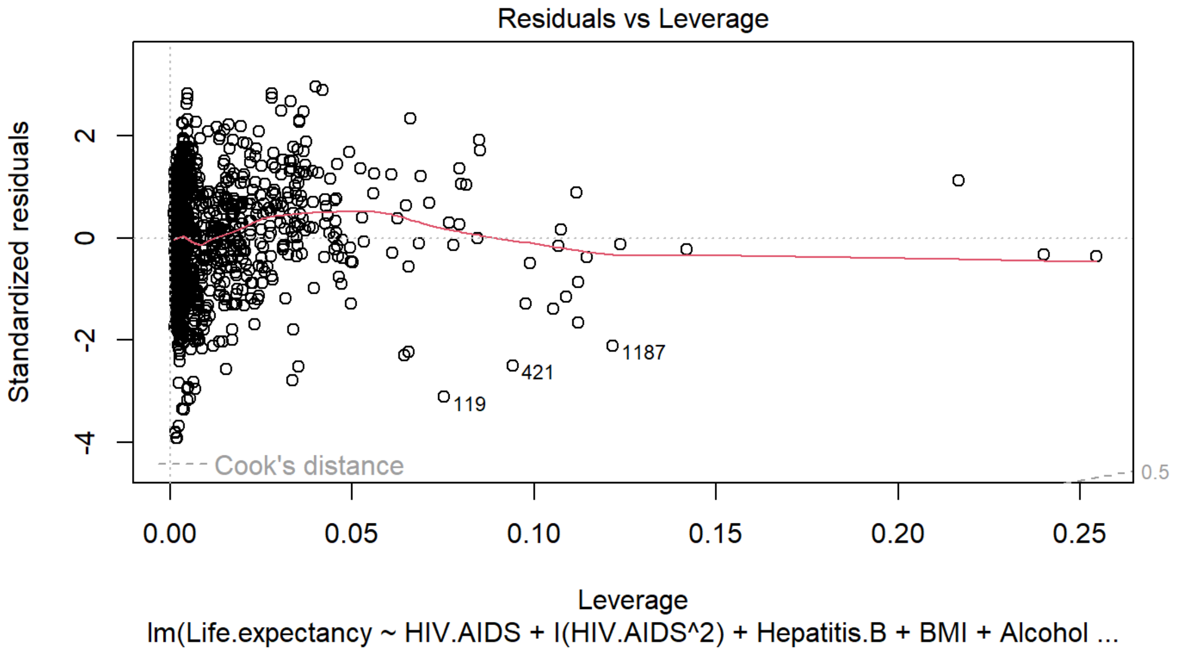


Figure 4: Plot of Residuals vs. Leverage

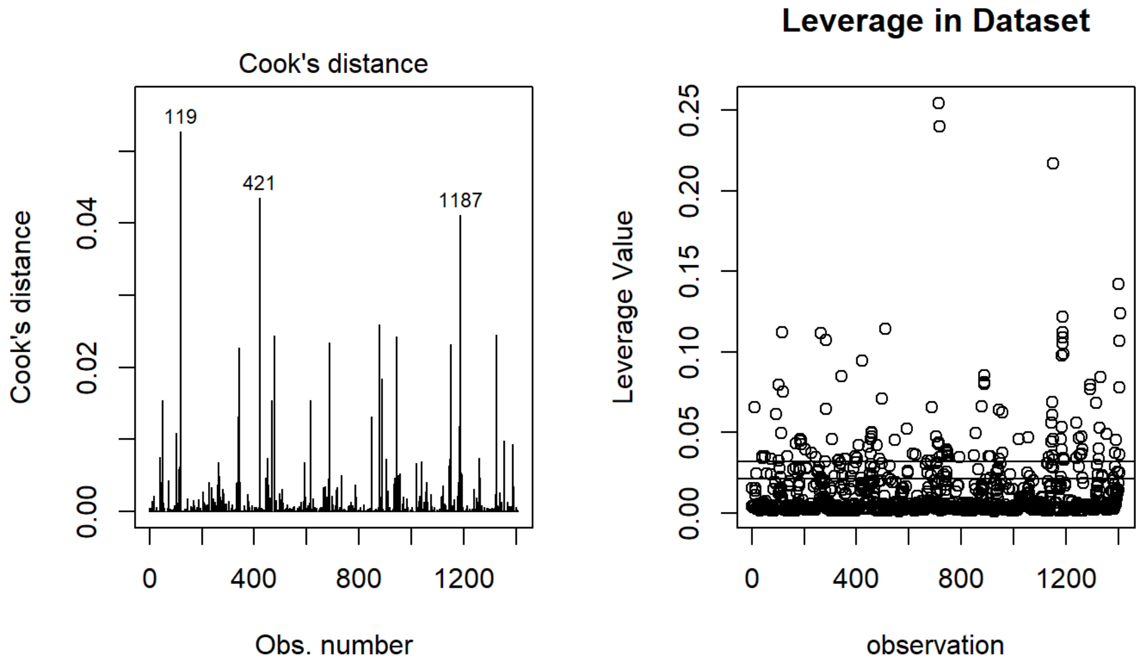


Figure 5: Plots of Cook's Distance and Leverage in the Dataset

Boxcox Transformation

Since our model didn't meet the assumptions of linearity, normality and equal variance, we tried a Boxcox transformation to see if it could improve the model's fit and assumptions. We began by running a Boxcox test to determine the lambda value to impose on our response variable, and the test returned a lambda value of 1.151515. This value is nearest to 1, indicating that no transformations were necessary for Y to improve the model.

Despite the Boxcox test results, we still attempted a Boxcox transformation of 1.151515 to see if it could improve the assumptions and the model's predictive ability. However, the transformed model yielded a lower adjusted R^2 (0.6758) and higher RMSE (8.942319). Additionally, the transformation introduced less normality and more heteroscedasticity into the model.

Based on these findings, we concluded that no transformations should be made to the model. From this, we concluded that our best model required no transformations to Y.

Removal of Outliers

Since our model didn't meet the assumptions of linearity, normality and equal variance, we attempted to remove any outliers that were above the leverage points 3p/n and 2p/n. After removing the outliers, we built two new models and fitted them.

However, we observed that the adjusted R^2 values for both new models (3p/n: 0.6628, 2p/n: 0.6537) were lower than the original fitted model. This implies that the original fitted model could better explain the variance in the data, resulting in us sticking with the original fitted model.

We also checked the assumptions for the new models, but they still did not hold. This suggests that removing outliers did not significantly change the model's assumptions. Furthermore, we noticed in the two new models without 3p/n or 2p/n outliers that when checking the outlier assumption in the residuals versus leverage plot, it got outliers that had higher Cook's Distance values to the point where some points were influential in the model without 2p/n outliers (one point had a cooks distance greater than 1) (FIGURE 6). This caused us to disregard these models and not use them since we did not want to keep on going through the process of removing more and more outliers since this created the issue of overfitting for our model.

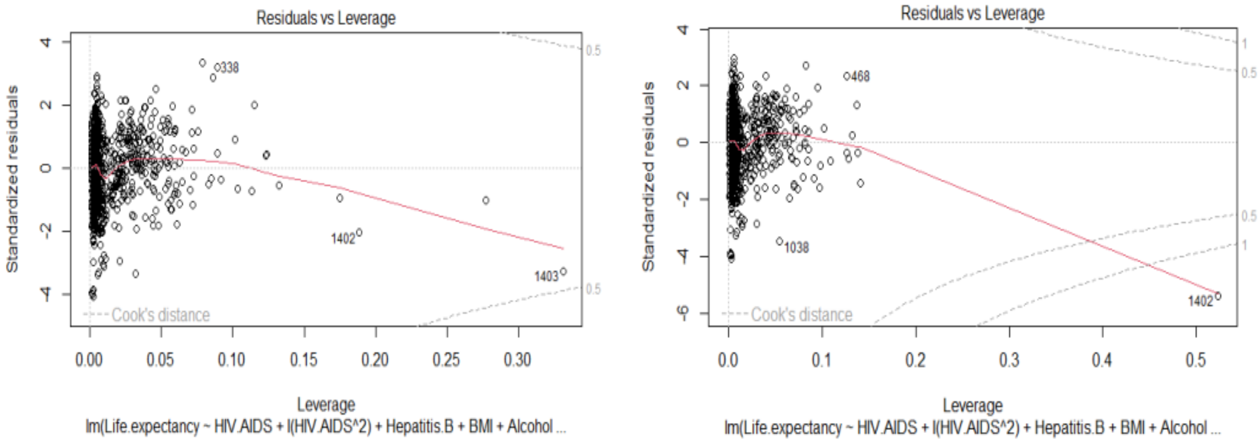


Figure 6. Plots of Residuals vs Leverage for Model Without 3p/n Outliers (LEFT) and Model Without 2p/n Outliers (RIGHT)

In conclusion, despite removing outliers, we could not improve the model. This led us to conclude that our original model is our best-fit model. Below is the model with the appropriate coefficients:

$$\widehat{Y_{Life\ Expectancy}} = 56.0141191 + 0.7211983X_{\text{Alcohol}} - 0.0057522X_{\text{HepatitisB}} + 0.2779569X_{\text{BMI}} + 0.0297612X_{\text{Polio}} + 0.0119486X_{\text{Diphtheria}} - 1.3219465X_{\text{HIV.AIDS}} + 0.0242920X^{2}_{\text{HIV.AIDS}} + 0.0004922X_{\text{polio:Diphtheria}} - 0.0018225X_{\text{BMI:Polio}} - 0.0080679X_{\text{HepatitisB:Alcohol}} + 0.0026948X_{\text{HIV.AIDS:Alcohol}} + 0.0025360X_{\text{Diphtheria:Alcohol}} + 0.0003109X_{\text{HepatitisB:Polio}} - 0.0028802X_{\text{HIV.AIDS:Diphtheria}}$$

Adjusted R^2 = 0.6796
RMSE = 4.729

Interpreting Coefficients

Below are the beta coefficient interpretations. Keep in mind that interpreting the beta coefficient for HIV.AIDS is quite complicated due to the presence of a higher power term and outside the scope of DATA 603.

- β_0**(y-intercept):** This value is 56.0141191. This is the predicted life expectancy when all predictor variables = 0.
- β_1**(Alcohol):** Life expectancy changes by approximately ($0.7211983 - 0.0080679X_{\text{HepatitisB}} + 0.0026948X_{\text{HIV.AIDS}} + 0.0025360X_{\text{Diphtheria}}$) for every 1 unit increase in alcohol consumption per litre of pure alcohol when all other predictor variables are held constant.
- β_2**(HepatitisB):** Life expectancy changes by approximately ($-0.0057522 - 0.0080679X_{\text{Alcohol}} + 0.0003109X_{\text{Polio}}$) for every 1 unit increase in HepatitisB immunization coverage among 1-year-olds in percentage when all other predictor variables are held constant.
- β_3**(BMI):** Life expectancy changes by approximately ($0.2779569 - 0.0018225X_{\text{Polio}}$) for every 1 unit increase in the Average Body Mass Index of the entire population when all other predictor variables are held constant.
- β_4**(Polio):** Life expectancy changes by approximately ($0.0297612 + 0.0004922X_{\text{Diphtheria}} - 0.0018225X_{\text{BMI}} + 0.0003109X_{\text{HepatitisB}}$) for every 1 unit increase in Polio immunization coverage among 1-year-olds in percentage when all other predictor variables are held constant.
- β_5**(Diphtheria):** Life expectancy changes by approximately ($0.0119486 + 0.0004922X_{\text{Polio}} + 0.0025360X_{\text{Alcohol}} - 0.0028802X_{\text{HIV.AIDS}}$) for every 1 unit increase in

diphtheria-tetanus toxoid and pertussis immunization coverage among 1-year-olds in percentage when all other predictor variables are held constant

- β_6**(HIV.AIDS):** Life expectancy changes by approximately (-1.3219 + 0.0242920$X_{\text{HIV.AIDS}}$ + 0.0026948X_{Alcohol} - 0.0028802$X_{\text{Diphtheria}}$) for 1 unit increase in HIV/AIDS deaths per 1000 live births when all other predictor variables are held constant

Conclusion and Discussion

Our approach yielded intriguing results, but their reliability is questionable. The model fell short of meeting crucial assumptions such as normality, linearity, and homoscedasticity, undermining the accuracy of our findings despite our diligent efforts.

Upon analysis, we identified key areas for developing countries to consider for improving life expectancy:

1. Reducing the prevalence of HIV/AIDS through community education programs and enhanced access to contraceptives.
2. Addressing and increasing per capita alcohol consumption to achieve a positive impact.
3. Exploring strategies to increase the average BMI of the population.
4. Increasing Polio and Diphtheria vaccination rates.
5. Decreasing Hepatitis B vaccination rates.

It is essential to note that these suggestions should be treated with caution and not implemented without rigorous testing and validation. Conventional wisdom questions the empirical support for recommendations involving increased alcohol consumption and average BMI, as well as decreasing hepatitis B vaccination rates.

Life expectancy is an incredibly complex and nuanced topic that can be challenging to narrow down to specific factors. While our examination focused on seven health factors, numerous other variables could influence a country's life expectancy. Our analysis highlighted two critical insights: Multiple Linear Regression may not be the optimal method for estimating life expectancy, and a more comprehensive approach considering various factors is necessary.

For future analyses, overcoming the challenges encountered with our multiple linear regression model suggests the need for alternative modelling approaches to predict a developing country's life expectancy accurately. Potential models to explore include Decision Trees, Random Forest, or Ridge and Lasso Regression.

It is important to consider additional factors when measuring life expectancy beyond the ones we have used in our model. Although we have done our best to predict life expectancy based on predictive health factors, many other factors should be considered. Our research has shown that our study's limited scope has yielded results proving our approach was insufficient in the context of multiple linear regression.

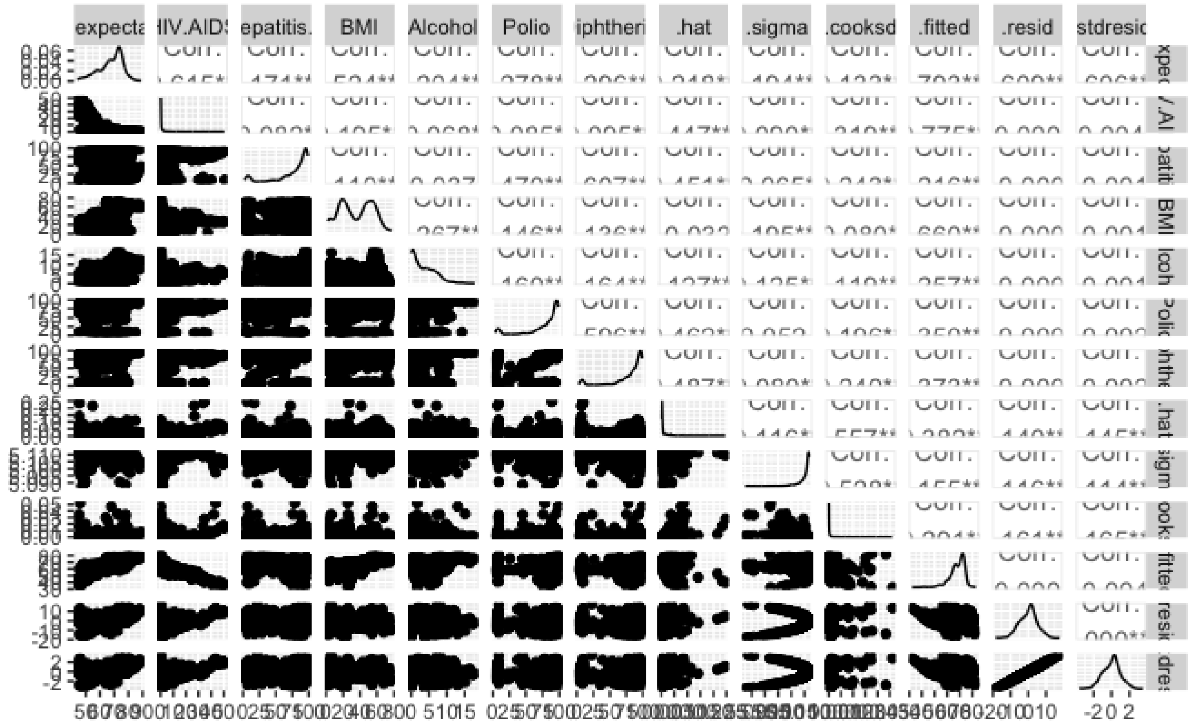
References

Harvard School of Public Health. (2022, November 18). *Healthy longevity*. The Nutrition Source. https://www.hsph.harvard.edu/nutritionsource/healthy-longevity/

Life expectancy(Who). (n.d.). Retrieved December 7, 2023, from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Miladinov, G. (2020). Socioeconomic development and life expectancy relationship: Evidence from the EU accession candidate countries. *Genus*, *76*(1), 2. https://doi.org/10.1186/s41118-019-0071-0

Appendix A - Pairs Plot



Appendix B - Code

# DATA 603 Final Project R CODE

This file contains all of our R code for the project. It contains our model building process in the order of these steps:

1. Building our first/initial model and doing a full F test.

2. Reducing the model using the individual t-test. Use partial F-test to confirm the reduced model is significant

3. Adding interactions to the model and using partial F-test to confirm the interaction model is significant

4. Reducing the interaction model via stepwise selection

5. Checking plots between response variable and predictor variables to determine if higher powers need to be present

6. Adding higher powers and using partial F-test to confirm the higher power model is significant

7. Checking the assumptions of the final model (linearity, normality, equal variance, multicollinearity and outliers)

8. Doing various techniques if the model does not meet assumptions (boxcox transformation and removal of outliers)

9. rebuilding the models from the techniques done above and check if the new model meets assumptions better vs the other model.

10. Choosing the best final model within the scope of this course

First we loaded in our dataset, removed na values and filtered for only rows that were considered "developing" countries based off of the column Status (states if a row is a devoloped or non-developed country). We filtered for these rows since our project only focuses on developing countries. We also loaded in any libraries we planned on using.

~~~
```{r}
library(car)
library(lmtest)
library(GGally)
library(olsrr)
library(MASS)
library(dplyr)
```
~~~

~~~
```{r}
#read in original csv file and remove na values
LE =
read.csv("https://raw.githubusercontent.com/ethan2411/Data-603-604/main/Life%20Expectancy%20Data.csv")
LE = na.omit(LE)
~~~

head(LE)

```{r}
#select for "developing" rows based on Status column
developing_LE <- LE %>%
  filter(Status == "Developing")
nrow(LE)
nrow(developing_LE)
head(developing_LE)
```

Building our first/initial model and doing a full F test.

```{r}
options(scipen = 999)
#make first model
develop_firstlm = lm(Life.expectancy ~  Alcohol + Hepatitis.B  + Measles + BMI + Polio + Diphtheria + HIV.AIDS, data = developing_LE)
summary(develop_firstlm)
```

This is our first initial model above. You can also see from the full f test the p-value is <2.2e-16 meaning that at least one of the predictor variables is related to the response variable.

##Reducing the model using the individual t-test. Use partial F-test to confirm the reduced model is significant

Looking at individual t-test p-values for our initial model above, we see that the p-value for measles is 0.4875. So we fail to reject the null hypothesis and from this we can infer that the variable Measles is not related to the response variable and so can be dropped from the model

We build the reduced model and check to see if we need to drop any other variables.
```{r}
#reduce model via t-test
develop_lmred = lm(Life.expectancy ~  Alcohol + Hepatitis.B + BMI + Polio + Diphtheria + HIV.AIDS, data = developing_LE)
summary(develop_lmred)
```

~~~
#check if reduced model is accepted (it is) via partial f test
anova(develop_firstlm,develop_lmred)
```
~~~

We see that all our p-values are less than 0.05, so no more variables need to be dropped from the model. The partial f-test p-value is also 0.4875. From this we can infer that the reduced model is accepted (further confirming measles can be dropped).

Adding interactions to the model and using partial F-test to confirm the interaction model is significant

```{r}
#make full interaction model
develop_lmintfull = lm(Life.expectancy ~  (Alcohol + Hepatitis.B + BMI + Polio + Diphtheria + HIV.AIDS)^2, data = developing_LE)
summary(develop_lmintfull)

#run anova to see if full interaction model is accepted
anova(develop_lmintfull, develop_lmred)
```

From the anova partial f-test we see a p-value < 0.05 so we can reject the null hypothesis. From this we can infer that the full interaction model is the preferred/accepted model.

##Reducing the interaction model via stepwise selection

```{r}

#reduce interaction model via ols stepwise because there are too many interactions to do it manually
develop_stepmod_intred = ols_step_both_p(develop_lmintfull, pent= 0.1, prem = 0.3, details = FALSE)
summary(develop_stepmod_intred$model)

```

##Checking plots between response variable and predictor variables to determine if higher powers need to be present

```{r}
#code in the reduced interaction model from stepwise
developing_intred2 = lm(Life.expectancy ~ HIV.AIDS + Hepatitis.B + BMI + Alcohol + Polio
+ Diphtheria + Polio:Diphtheria +BMI:Polio + Hepatitis.B:Alcohol + HIV.AIDS:Alcohol +
Diphtheria:Alcohol +  Hepatitis.B:Polio + HIV.AIDS:Diphtheria, data = developing_LE)

#look for graphs to see if higher order models are required
ggpairs(developing_intred2)
```

We see that there seems to be some sort of curved relationship with our response variable (life expectancy) and the predictor variable HIV.AIDS. So we now will add a quadratic power to our model and see if the higher power is accepted from the partial F-test

###Adding higher powers and using partial F-test to confirm the higher power model is significant

```{r}
developing_highermod = lm(Life.expectancy ~ HIV.AIDS + I(HIV.AIDS^2) + Hepatitis.B +
BMI + Alcohol + Polio + Diphtheria + Polio:Diphtheria +BMI:Polio + Hepatitis.B:Alcohol +
HIV.AIDS:Alcohol + Diphtheria:Alcohol +  Hepatitis.B:Polio + HIV.AIDS:Diphtheria, data =
developing_LE)

summary(developing_highermod)
#ANOVA TO TEST IF HIGHER ORDER MODEL IS ACCEPTED. IT IS ACCEPTED
anova(developing_highermod, developing_intred2)

#check R sq adj and RMSE of our fitted model
summary(developing_highermod)$adj.r.squared
sigma(developing_highermod)
```

We see that adding the quadratic power into the model for HIV.AIDS increases the models R squared value from 0.6259 to 0.6796. The partial f-test also accepts the higher order model (since p-value is < 0.05). For these reasons we have decided to keep the higher order term in our model. We also decided to keep it in our model since we noticed it helps with our linearity assumption (which you will see below). Now we have our final model, now we check the assumptions of the model.

Our final model is:

We also see the R squared adjusted value for our final fitted model is 0.6795512 and the RMSE is 4.729039

###Checking the assumptions of the final model

~~~
```{r}

#CHECK ASSUMPTIONS OF OUR FINAL MODEL
summary(developing_highermod)

#equal variance assumption
bptest(developing_highermod)

#normality assumption
shapiro.test(residuals(developing_highermod))

#plots to show linearity, equal variance, normality and outlier assumptions
plot(developing_highermod)

```
~~~

~~~
```{r}
#check vif for multicolinnearity to see if we need to drop any variables with VIF > 5
vif(develop_firstlm)
vif(develop_lmred)

```
~~~

We see from vif that they are all below 5 so we do not have any multicollinearity in our model and so none of the variables need to be removed. (Included the full additive model and reduced additive model (without measles) for vif tests and see there is still no difference (none greater than 5))

~~~
```{r}
#more plots to show the outlier assumption
lev=hatvalues(developing_highermod)
p = length(coef(developing_highermod))
n = nrow(developing_LE)


par(mfrow = c(1,2))
plot(developing_highermod, which = 4)
plot(rownames(developing_LE),lev, main = "Leverage in Dataset", xlab="observation",
ylab = "Leverage Value")
abline(h = 2 *p/n, lty = 1)
abline(h = 3 *p/n, lty = 1)
```

```{r}
#more plots to show the linearity, equal variance and outliers assumptions

# Residuals vs Fitted Values Plot
ggplot(data.frame(residuals = residuals(developing_highermod), fitted =
fitted(developing_highermod)), aes(x = fitted, y = residuals)) +
  geom_point() +
  geom_smooth(method = "loess", se = TRUE, color = "red") +
  labs(title = "Residuals vs Fitted",
       x = "Fitted Values",
       y = "Residuals")

# Scale-Location (Spread-Location) Plot
ggplot(data.frame(fitted = fitted(developing_highermod), std_resid =
sqrt(abs(rstandard(developing_highermod)))), aes(x = fitted, y = std_resid)) +
  geom_point() +
  geom_smooth(method = "loess", se = TRUE, color = "red") +
  labs(title = "Scale-Location Plot",
       x = "Fitted Values",
       y = "Square Root of Standardized Residuals")

# Residuals vs Leverage Plot
ggplot(data.frame(hat_values = hatvalues(developing_highermod), std_resid =
rstandard(developing_highermod)), aes(x = hat_values, y = std_resid)) +
~~~

~~~
  geom_point() +
  geom_smooth(method = "loess", se = TRUE, color = "red") +
  labs(title = "Residuals vs Leverage",
       x = "Leverage",
       y = "Standardized Residuals")
```
~~~

~~~
```{r}
library(patchwork)

# Normal Q-Q Plot
qq_plot <- ggplot(data.frame(std_resid = rstandard(developing_highermod)), aes(sample
= std_resid)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "Normal Q-Q Plot")

# Histogram of Residuals
hist_plot <- ggplot(data = developing_highermod, aes(x =
developing_highermod$residuals)) +
  geom_histogram(binwidth = 0.5, fill = 'blue', col = 'red') +
  labs(title = "Histogram of Residuals")+
  xlab("Residuals") +
  ylab("Count")

# Combine plots using patchwork
combined_plot <- qq_plot + hist_plot

# Display the combined plot
combined_plot
```
~~~

For the linearity assumption, we notice for the residuals vs fitted values plot that there seems to be some sort of non linear pattern occuring. So from this we say the linearity assumption does not hold.

For the normality assumption, we notice from the QQ plot that the tail ends flare out quite alot possibly hinting towards heteroscedasticity. We also see from the histogram a generally normal pattern however there is some data points near the left tail suggesting the distribution may not be normal. We also ran the shapiro wilk test with the hypothesis:

H_0: The Residuals are Normally Distributed
H_a: The Residuals are not Normally Distributed

The p-value from the shapiro wilk test was 0.000000002244. Since 0.000000002244 < 0.05, we reject the null hypothesis. So we can infer from this that the residuals are not normally distributed. So the normality assumption does not hold.

For the equal variance assumption, we notice from the residuals vs fitted plot that there is some funneling present (narrower on the left side vs wider as you move to the right) which shows that heteroscedasticity may be present. WE also ran the breusch pagan test with the hypothesis:

H_0: Heteroscedascity is not present
H_a: Heteroscedascity is present

The p-value from the bp test was 0.00000000000000022. Since 0.00000000000000022 < 0.05, we reject the null hypothesis. So we can infer from this that their is heteroscedasticity present. So the equal variance assumption does not hold.

For the multicollinearity assumption, we used the VIF function and found that no variables were over a VIF of 5, so we do not need to remove any variables from the model. So the multicollinearity assumption holds.

For the outliers, from the residuals vs leverage plot we notice no outliers with a high cooks distance. For the cooks distance plot, we notice 3 points that stand out, however they still have an incredibly small cooks distance (around 0.04). From the leverage in dataset plot, we do notice points over a leverage of 3p/n and 2p/n, these will be addressed later when we look for ways to make our model meet the normality, linearity and equal variance assumption. Overall we see that there are no influential outliers with a high cooks distance.

Doing various techniques if the model does not meet assumptions (boxcox transformation and removal of outliers)

Since we do not meet the assumption of equal variance, normality and linearity, we will do a boxcox transformation to see if it helps meet the assumption.

=======

FROM ASSUMPTION CHECKS WE DO BOXCOX

~~~~
```{r}
bc = boxcox(developing_highermod, lambda=seq(-2,2))

bestlambda = bc$x[which(bc$y==max(bc$y))]
bestlambda
```
~~~~

We see from the boxcox transformation that the lambda value it gave us is 1.151515. When rounding to the nearest value it tells us that the model doesnt need any transformations done, we however did use the lambda value of 1.151515 to test if this transformation will fix our assumptions

~~~~
```{r}
bcmodel = lm((((Life.expectancy^bestlambda)-1)/bestlambda) ~ HIV.AIDS + I(HIV.AIDS^2)
+ Hepatitis.B + BMI + Alcohol + Polio + Diphtheria + Polio:Diphtheria +BMI:Polio +
Hepatitis.B:Alcohol + HIV.AIDS:Alcohol + Diphtheria:Alcohol +  Hepatitis.B:Polio +
HIV.AIDS:Diphtheria, data = developing_LE)

#check r squared adjusted value and RMSE of boxcox model
summary(bcmodel)

#check assumptions of model
plot(bcmodel)

#normality test
shapiro.test(residuals(bcmodel))

#equal variance test
bptest(bcmodel)
```
~~~~

From the new model with the boxcox transformation of 1.151515 we notice the model has a worse R squared adjusted value of 0.6758 compared to the original fitted model. We also notice from the residuals vs fitted plot and the breusch pagan test that the assumption of equal variance still does not hold (since the null hypothesis is rejected).

We also notice from the QQ plot no difference comapred to the original QQ plot showing no change in normality. Furthermore, we see from the shapiro test that the assumption of normality still does not hold (since the null hypothesis is rejected).

So overall the boxcox transformation model did not help us meet our assumptions and is a worse model compared to our original fitted model.

Since we also noticed there were points with leverage values over 3p/n, we removed these points and refit the model to see if any improvements occurred for our assumptions.

~~~
```{r}

lev=hatvalues(developing_highermod)
p = length(coef(developing_highermod))
n = nrow(developing_LE)
outlier3p = lev[lev>(3*p/n)]
print("h_I>3p/n, outliers are")
print(outlier3p)

rows_remove3p = which(lev>(3*p/n))
print(rows_remove3p)

developing_LE_no_outliersSTEP = developing_LE[-rows_remove3p, ]
head(developing_LE_no_outliersSTEP)

nrow(developing_LE)
nrow(developing_LE_no_outliersSTEP)
```
~~~

~~~
```{r}
model3pgone = lm(Life.expectancy ~ HIV.AIDS + I(HIV.AIDS^2) + Hepatitis.B + BMI +
Alcohol + Polio + Diphtheria + Polio:Diphtheria +BMI:Polio + Hepatitis.B:Alcohol +
HIV.AIDS:Alcohol + Diphtheria:Alcohol +  Hepatitis.B:Polio + HIV.AIDS:Diphtheria, data =
developing_LE_no_outliersSTEP)
summary(model3pgone)

#looking at plots for assumptions
plot(model3pgone)
~~~

~~~
#looking at normality assumption
shapiro.test(residuals(model3pgone))
#looking at equal variance assumption
bptest(model3pgone)
```
~~~

From the fitted model without 3p/n outliers, we notice this model has a worse R squared adjusted value of 0.6628 compared to the original fitted model. From looking at the residuals vs fitted plot we also see that there is still a non-linear pattern present. We also notice from the residuals vs fitted plot and the breusch pagan test conducted that there is still heteroscedasticity present (since the null hypothesis is rejected) so the assumption of equal variance still does not hold. We also notice no improvement in the QQ plot for tailing of the residuals. Furthermore we see from the shapiro test conducted that the assumption of normality still does not hold (since the null hypothesis is rejected). Another interesting thing to note as well is that when looking at the residuals vs leverage plot, we see that there are points that have a higher leverage now comapred to the original fitted model. One point (1403) is even reaching close to a cooks distance of 0.5, showing it is almost influential.

So overall, the removal of 3p/n outliers did not help us meet our assumptions better and it also made the outliers present in its model worse.

Next we attempt the same model refitting but this time with the removal of outliers with a leverage above 2p/n

~~~
```{r}
#removing 2p/n outliers and making a new DF without them
lev=hatvalues(developing_highermod)
p = length(coef(developing_highermod))
n = nrow(developing_LE)
outlier2p = lev[lev>(2*p/n)]
print("h_I>2p/n, outliers are")
print(outlier2p)

rows_remove2p = which(lev>(2*p/n))
print(rows_remove2p)

developing_LE_no_outliersSTEP2P = developing_LE[-rows_remove2p, ]
head(developing_LE_no_outliersSTEP2P)
~~~

~~~
nrow(developing_LE)
nrow(developing_LE_no_outliersSTEP2P)
```
~~~

~~~
```{r}
#fitting the previous model with the new data that had outliers
model2pgone = lm(Life.expectancy ~ HIV.AIDS + I(HIV.AIDS^2) + Hepatitis.B + BMI +
Alcohol + Polio + Diphtheria + Polio:Diphtheria +BMI:Polio + Hepatitis.B:Alcohol +
HIV.AIDS:Alcohol + Diphtheria:Alcohol +  Hepatitis.B:Polio + HIV.AIDS:Diphtheria, data =
developing_LE_no_outliersSTEP2P)
summary(model2pgone)

#looking at plots for assumptions
plot(model2pgone)
#looking at normality assumption
shapiro.test(residuals(model2pgone))
#looking at equal variance assumption
bptest(model2pgone)


```
~~~

From the fitted model without 2p/n outliers, we notice this model has a worse R squared adjusted value of 0.6537 compared to the original fitted model. From looking at the residuals vs fitted plot we also see that there is still a non-linear pattern present. We also notice from the residuals vs fitted plot and the breusch pagan test conducted that there is still heteroscedasticity present (since the null hypothesis is rejected) so the assumption of equal variance still does not hold. We also notice no improvement in the QQ plot for tailing of the residuals (if anything, the tailing of the residuals is worse in this QQ plot). Furthermore we see from the shapiro test conducted that the assumption of normality still does not hold (since the null hypothesis is rejected). Another interesting thing to note as well is that when looking at the residuals vs leverage plot, we see that there is a point (1402) that is influential since it has a cooks distance greater than 1, so it made the outliers present worse compared to the original fitted model.

So overall, even after attempting to make the model better fit to possibly meet the assumptions we were unable to do so in the scope of DATA 603. SO the best fit model that we are sticking with is the original fitted model which is:

YLife Expectancy = 56.0141191 + 0.7211983XAlcohol - 0.0057522XHepatitisB + 0.2779569XBMI + 0.0297612XPolio + 0.0119486XDiphtheria - 1.3219465XHIV.AIDS + 0.0242920X2 HIV.AIDS + 0.0004922Xpolio:Diphtheria - 0.0018225XBMI:Polio - 0.0080679XHepatitisB:Alcohol + 0.0026948XHIV.AIDS:Alcohol + 0.0025360XDiphtheria:Alcohol + 0.0003109XHepatitisB:Polio - 0.0028802XHIV.AIDS:Diphtheria

and its Adjusted R squared value is:0.6796

and its RMSE value is: 4.729

Note: all interpretations are in our final report and will not be here

Below here we have extra code that we wanted to keep in if you were curious about some of the other stuff that we did that we did not include in our report. It is all commented out so that it does not get ran. This code includes the rebuilding of our model entirely for the the model without 3p/n outliers and the model without 2p/n outliers. We found that even attempting it this way that we still did not find better fit models because their assumptions were just as bad or worse (made outliers present with higher leverage and cooks distance).
```{r}

#we also made the model without outliers from scratch and checked the assumptions but they were not better at all and stil worse than the original fitted model. wE have kept this code here but it is commented out since we did not use it.

#make initial model without 3p/n outliers

#first3plm= lm(Life.expectancy ~  Alcohol + Hepatitis.B  + Measles + BMI + Polio + Diphtheria + HIV.AIDS, data = developing_LE_no_outliersSTEP)
#summary(first3plm)
```

```
#reduced model
#red3plm = lm(Life.expectancy ~  Alcohol + Hepatitis.B + BMI + Polio + Diphtheria +
HIV.AIDS, data = developing_LE_no_outliersSTEP)
#summary(red3plm)

#anova accepts the reduced model
#anova(first3plm, red3plm)

#int model
#int3plm = lm(Life.expectancy ~  (Alcohol + Hepatitis.B + BMI + Polio + Diphtheria +
HIV.AIDS)^2, data = developing_LE_no_outliersSTEP)
#summary(int3plm)

#anova accepts the interaction model
#anova(int3plm, red3plm)

#intred3plm = ols_step_both_p(int3plm, pent = 0.1, prem = 0.3, details = FALSE)
#summary(intred3plm$model)

#MAKING REDUCED INT MODEL AFTER STEPWISE
#step_red_3p = lm(Life.expectancy ~  HIV.AIDS + Hepatitis.B + Polio + Diphtheria +
Alcohol + BMI + BMI:Diphtheria + Diphtheria:Polio + Hepatitis.B:Alcohol +
#Diphtheria:Alcohol + BMI:Hepatitis.B + BMI:Polio + HIV.AIDS:Polio +
Diphtheria:Hepatitis.B, data = developing_LE_no_outliersSTEP)


#CHECK GGPAIRS TO SEE IF WE ADD HIGHER ORDER TO MODEL
#ggpairs(step_red_3p)

#LOOKS LIKE WE DO IT FOR HIV
#make higher order model
#higher3plm = lm(Life.expectancy ~ HIV.AIDS + I(HIV.AIDS^2) + Hepatitis.B + Polio +
Diphtheria + Alcohol + BMI + BMI:Diphtheria + Diphtheria:Polio + #Hepatitis.B:Alcohol +
Diphtheria:Alcohol + BMI:Hepatitis.B + BMI:Polio + HIV.AIDS:Polio +
Diphtheria:Hepatitis.B, data = developing_LE_no_outliersSTEP)
#stopped at 2. Dont want to create an overfitted model.


#anova accepts higher order model
#anova(higher3plm,step_red_3p)
```

```
#quick check of assumptions for model without 3p/n outliers
#summary(higher3plm)

#equal variance assumption
#bptest(higher3plm)

#normality assumption
#shapiro.test(residuals(higher3plm))

#check linearity, equal variance, normality and outlier assumptions via plots
#plot(higher3plm)




#building initial model with new dataset without 2p outliers


#first2plm= lm(Life.expectancy ~  Alcohol + Hepatitis.B  + Measles + BMI + Polio +
Diphtheria + HIV.AIDS, data = developing_LE_no_outliersSTEP2P)
#summary(first2plm)
#everything signficiant, so no reductions. make int model next

#int model
#int2plm = lm(Life.expectancy ~  (Alcohol + Hepatitis.B  + Measles + BMI + Polio +
Diphtheria + HIV.AIDS)^2, data = developing_LE_no_outliersSTEP2P)
#summary(int2plm)

#reduced int model via stepwise
#intred2plm = ols_step_both_p(int2plm, pent = 0.1, prem = 0.3, details = FALSE)
#summary(intred2plm$model)


#MAKE NEW MODEL HERE AFTER DOING STEPWISE OLS
#step_red_2p = lm(Life.expectancy ~  Alcohol + Hepatitis.B  + Measles + BMI + Polio +
Diphtheria + HIV.AIDS, data = developing_LE_no_outliersSTEP2P)


#ggpairs(intred2plm)
```

~~~
#make higher order model
#higher2plm = lm(Life.expectancy ~ Alcohol + Hepatitis.B + Measles + BMI + Diphtheria
+ HIV.AIDS + I(HIV.AIDS^2) + Alcohol:Hepatitis.B + Alcohol:Measles + #Alcohol:BMI +
Alcohol:Diphtheria + Alcohol:HIV.AIDS + Hepatitis.B:Diphtheria + Measles:Diphtheria +
Measles:HIV.AIDS + BMI:Diphtheria + Diphtheria:HIV.AIDS, #data =
developing_LE_no_outliersSTEP2P)

#summary(higher2plm)
#bptest(higher2plm)
#plot(higher2plm)

#shapiro.test(residuals(higher2plm))

```
~~~