Predicting Life Expectancy of Different Countries from Gross Domestic Product (GDP)

Lindsey Hornberger

Department of Biostatistics
University of Kansas, USA
December 15, 2023

Aims

The purpose of the study is to investigate the relationship between GDP and life expectancy across difference countries between 2000 and 2015 while controlling for other potential sources of variation.

I. Abstract

To investigate the impact of outside factors on life expectancy in various countries, a dataset is considered. The data within the dataset is from World Bank Data, World Health, and 'Our World in Data' (a project of the University of Oxford). There are many different factors to consider when discussing life expectancy and a few of those variables are considered in this study. The objective of this study is to determine if there is an association between the GDP of a country and life expectancy of the residents of that country and 5 different predictor variables. In this study we used linear regression models to determine if an association between variables was present. The best linear regression model indicates there is a correlation between life expectancy and GDP, schooling, BMI, incidents of HIV, and an interaction term of GDP and BMI. These different variables have a significant correlation at the 95% level.

II. Introduction

The worldwide average life expectancy is 72.27 years. However, when that is split between men and women, the average life expectancy for a man is 68.9 years and for a woman is 73.9 years. The average life expectancy is measured through surveying countries and government reporting. Life expectancy has increased dramatically since 1900 when the life expectancy for a person was 47 years. Life expectancy has been increasing steadily worldwide for some time and the driving force behind this is due to advances in technology and healthcare. In 1950, the life expectancy for men was 66.5 and 71.8 years for women. There are many different factors that can increase or decrease a person's life expectancy. One major factor in life expectancy is the country one lives in due to the access they will have to resources and healthcare. The difference in life expectancy across different countries can vary drastically especially in countries that are developed compared to countries that are not. Life expectancy tends to increase when residents of a country have proper access to healthcare and nutrition. However, there are many other factors than can impact life expectancy. In this study, we will consider a few potential factors that lead to a country having a higher or lower life expectancy.

III. Materials and Methods

Data Sources

The dataset was obtained online from Kaggle. The data within the dataset regarding population, GDP, and life expectancy was updated according to World Bank Data. Information about vaccinations for measles, Hepatitis B, polio, and diphtheria, alcohol consumption, BMI, HIV incidents, mortality rates, and thinness were collected from World Health Organization public datasets. Information about schooling was collected from 'Our World in Data' which is a project of the University of Oxford. The data was originally collected from research experiments implemented by each of the specific organizations and the data was compiled into one dataset available on Kaggle. The variables: **GDP**, the gross domestic product (GDP) per capita in current USD of a country; **schooling**, average years that people aged 25+ spent in formal education; **alcohol consumption**, alcohol consumption that is recorded in liters of pure alcohol per capita

with 15+ years old; **BMI**, average BMI measurement recorded; incidents of **HIV**, incidents of HIV per 1000 population, and **economy status**, which is a dummy variable that is equal to 0 if a country is developed and is equal to 1 if a country is developing. The objective of this study is to determine if there is an association between life expectancy (the response variable) and GDP while controlling for four other sources of variation.

Statistical Analysis

The data was available in .xlsx (excel) format. The data analysis is done using the statistical software R version 4.3.1 (2023-06-16). This project focuses mainly on multiple linear regression. All of the predictor variables are explored individually. In this dataset the sample size is 97 and there are no missing values. The smaller sample size could assume less predictability and a larger sampling variability.

The Preliminary Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \varepsilon_i$$

A multiple linear regression model is considered. Let

Y_i = , life expectancy, the life expectancy for residents for the i^{th} country,

X_{i1}= GDP, gross domestic product per capita in current USD for the i^{th} country,

X_{i2}= Schooling, average years that people aged 25+ spent in formal education for the i^{th} country,

X_{i3} = Alcohol consumption, alcohol consumption that is recorded in liters of pure alcohol per capita with 15+ years old for the i^{th} country,

X_{i4}= BMI, average BMI measurement recorded for the i^{th} country,

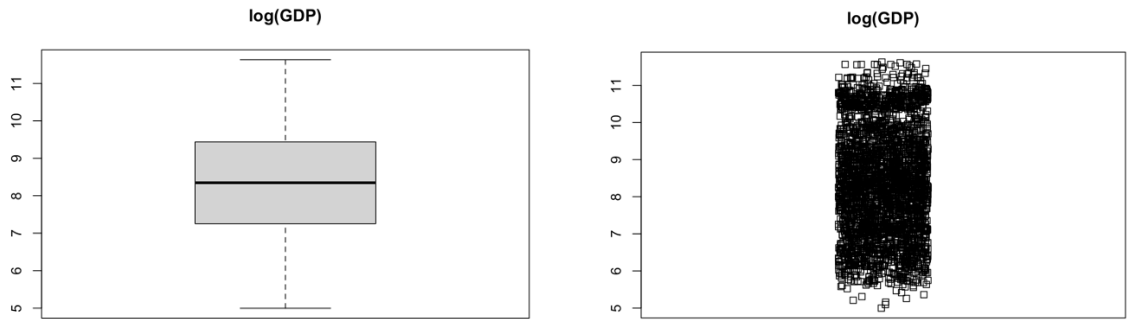
X_{i5}= HIV, incidents of HIV per 1000 population aged 15-49 for the i^{th} country.

X_{i6}= economy status, if variable = 0, the country is developed, if variable = 1 the country is developing.

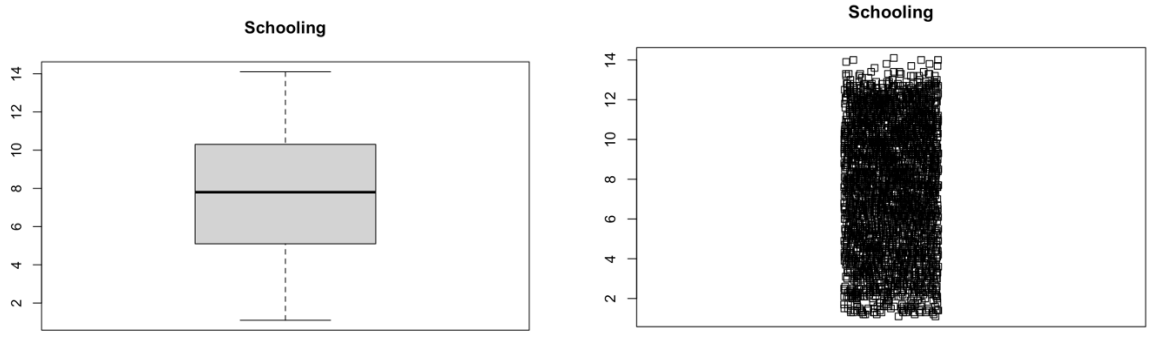
Based on automatic variable selection methods in combination with criterion-based statistics, income was dropped from the model. Partial residual plots, residual-versus-fitted plots, and measures of influence were investigated and no issues with high influence points, linearity, constant variance, independence, or normality were identified. Details are included in the Appendix.

IV. Results and Discussion

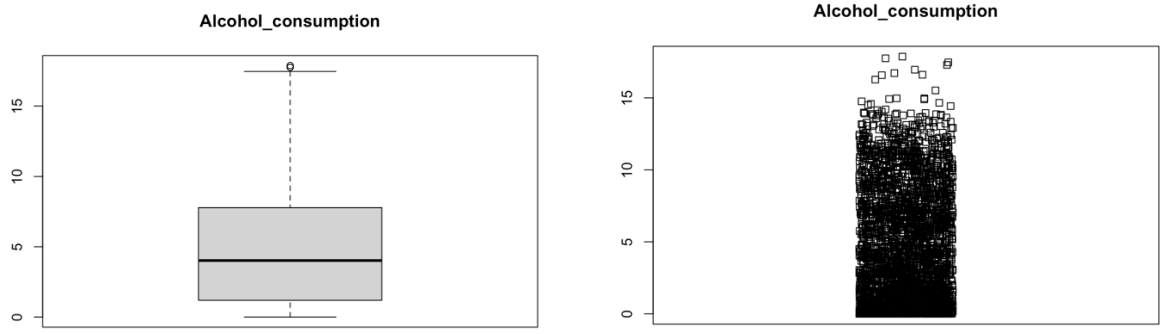
A. Distribution of Predictor Variables



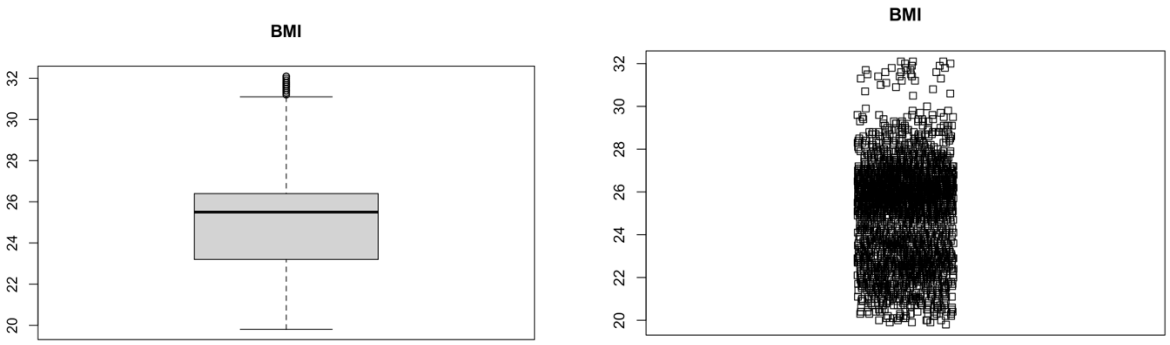
Figures 1 (a) and (b): Boxplot and strip chart for log(GDP)



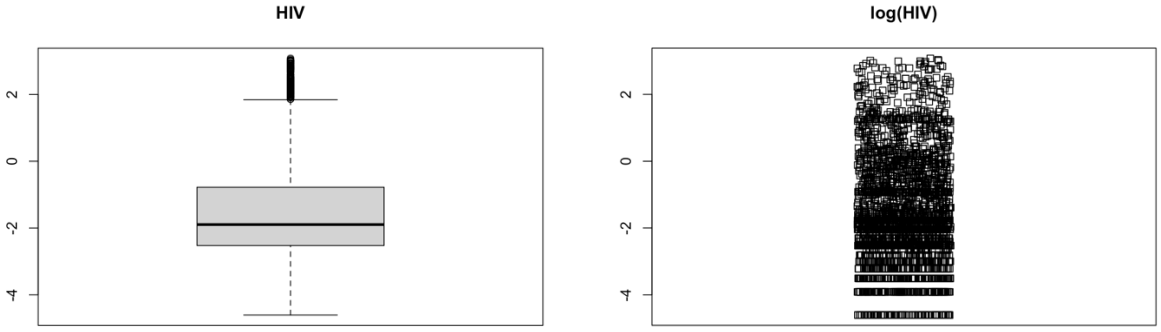
Figures 2 (a) and (b): Boxplot and strip chart for schooling.



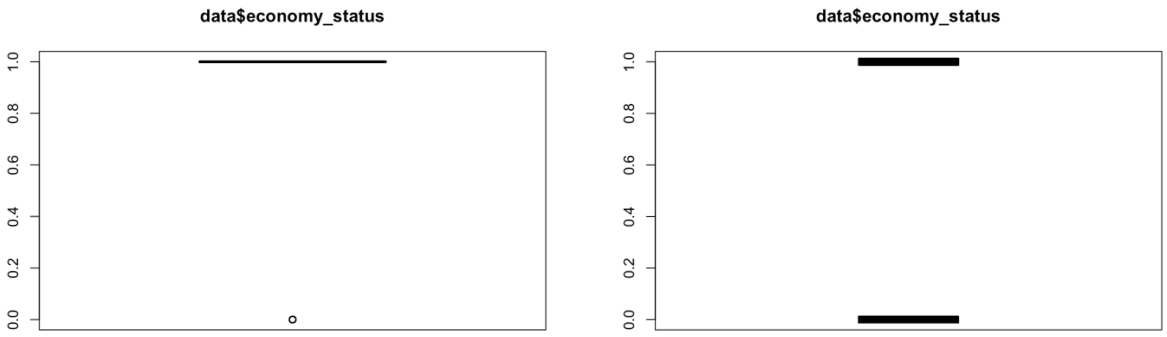
Figures 3 (a) and (b): Boxplot and strip chart for alcohol consumption.



Figures 4 (a) and (b): Boxplot and strip chart for BMI.



Figures 5 (a) and (b): Boxplot and strip chart for alcohol log(HIV).



Figures 6 (a) and (b): Boxplot and strip chart for alcohol economy status.

Strip plots for all predictors and the dependent variable (jittered) are shown next to boxplots of the same data. First, it should be acknowledged that a log transformation of both GDP and HIV were taken. GDP, as expected, is primarily positively skewed with the majority of observations clustered together and a few observations at much higher GDP levels. Similarly, HIV was positively skewed with the majority of the observations clustered together around zero and fewer

HIV values at might higher levels. Due the skewed nature of both of there variables, a natural-log transformation is appropriate for both. The boxplot and strip chart of economy status show that the majority of countries are not developed compared to those that are developed.

Other features of note: there is a wide range of values for BMI. There is also skewness visible in the distributions of alcohol consumption, BMI and log(HIV). For the variable alcohol consumption with observations clustered close to zero and a few data points with large values, we may need to apply transformations if the model diagnostics and assumption checks indicate it. Even after a log transformation, there is still some positively skewed values present in the distribution of log(HIV).

B. Diagnostics for Predictors

In this section, we will examine the distribution of predictors, identify any unusually large or small values, and examine bivariate associations to identify multicollinearity between variables. A scatterplot matrix indicates positive linear associations between life expectancy and GDP, schooling, alcohol consumption, and BMI and a negative linear association between life expectancy and HIV.

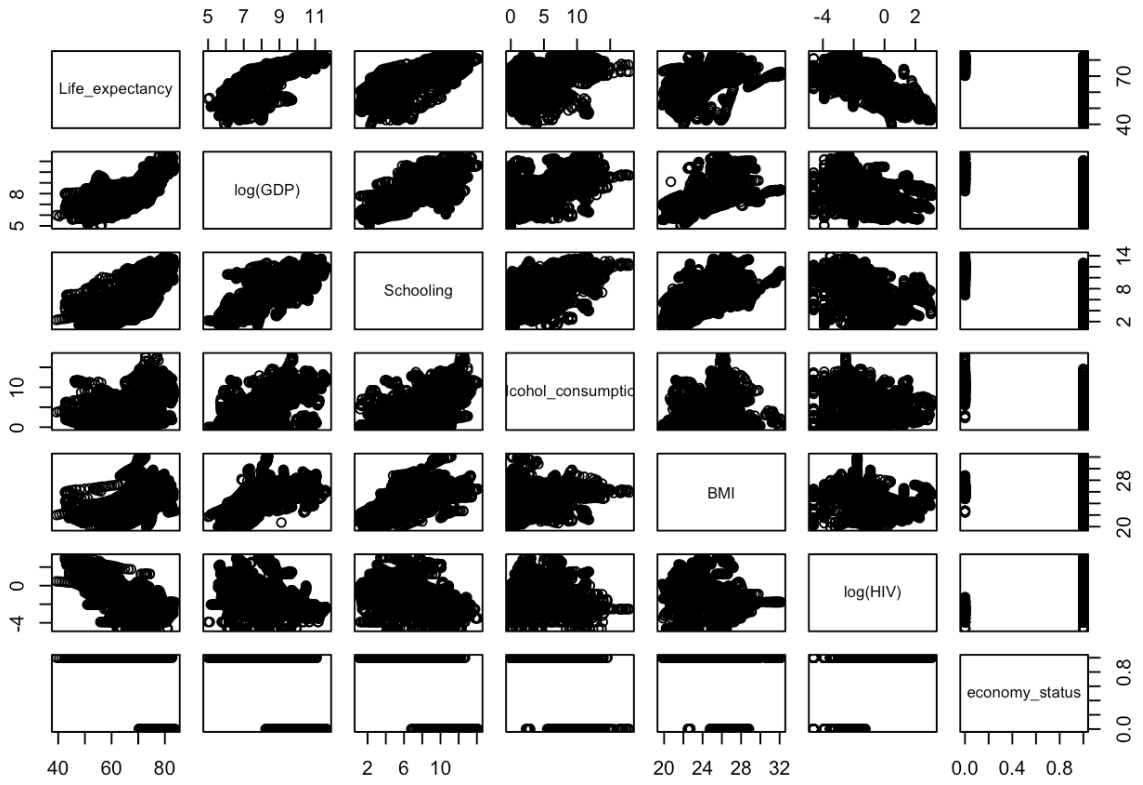


Figure 7: Scatterplot matrix for all variables.

Multicollinearity occurs when two or more predictor variables are highly correlated. This can cause problems when it comes to linear regression analysis. The multicollinearity is typically measured by analyzing the coefficient of correlation or Pearson correlation r. Values of r close to ±1 are considered to be highly correlated. Schooling appears to be highly associated with the dependent variable (life expectancy) with a correlation coefficient of 0.7324845. While this value is not too close to ±1, we can check for collinearity by calculating the VIF (variation inflation factor). If VIF exceeds 10, multicollinearity is present.

	LE	GDP	Schooling	Alcohol	BMI	HIV	economy
LE	1.0000000	0.5830897	0.7324845	0.39915911	0.5984233	-0.55302746	-0.5237910
GDP	0.5830897	1.0000000	0.5806259	0.44396595	0.3361796	-0.16958972	-0.6675469
Schooling	0.7324845	0.5806259	1.0000000	0.61572804	0.6354752	-0.20124620	-0.5994394
Alcohol	0.3991591	0.4439660	0.6157280	1.00000000	0.2840319	-0.03411801.	-0.6703661
BMI	0.5984233	0.3361796	0.6354752	0.28403195	1.0000000	-0.16114208.	-0.2432870
HIV	-0.5530275	-0.1695897	-0.2012462	-0.03411801	-0.1611421	1.00000000.	0.1756352
economy	-0.5237910	-0.6675469	-0.5994394	-0.67036609	-0.2432870	0.17563524	1.0000000

LE = Life_expectancy

Alcohol = Alcohol_consumption

economy_= economy_status

Figure 8: Correlation coefficients for all variables.

log(GDP)	Schooling	Alcohol	BMI	log(HIV)	economy
2.899910	3.590868	1.832045	1.914816	1.306446	2.622406

Figure 9: VIF values for all variables.

Due to the variance inflation factors (VIFs) of the response variables, each value is in the range 1-5 meaning that there is moderate correlation between log(GDP) and any of the predictor variables. However, this correlation is not significant enough to require attention. A maximum VIF in excess of 10 is a good rule of thumb for multicollinearity problems. Based on the maximum VIF, 3.590868, there do not appear to be any issues that need remediation. However, the VIF value for schooling much larger than the others, which indicates schooling may be redundant in the model

Partial residual plots for each of the covariates are displayed. Partial residual plots provide evidence of the importance of a covariate given the other covariates already in the model. They also display the nature of the relationship between the covariate and the outcome (i.e., linear, curvilinear, transformation necessary, etc.) and any problematic data points with respect to the

predictor. It is important to note that a log transformation of GDP and HIV was done prior due to the significant skewedness data points within the variable.

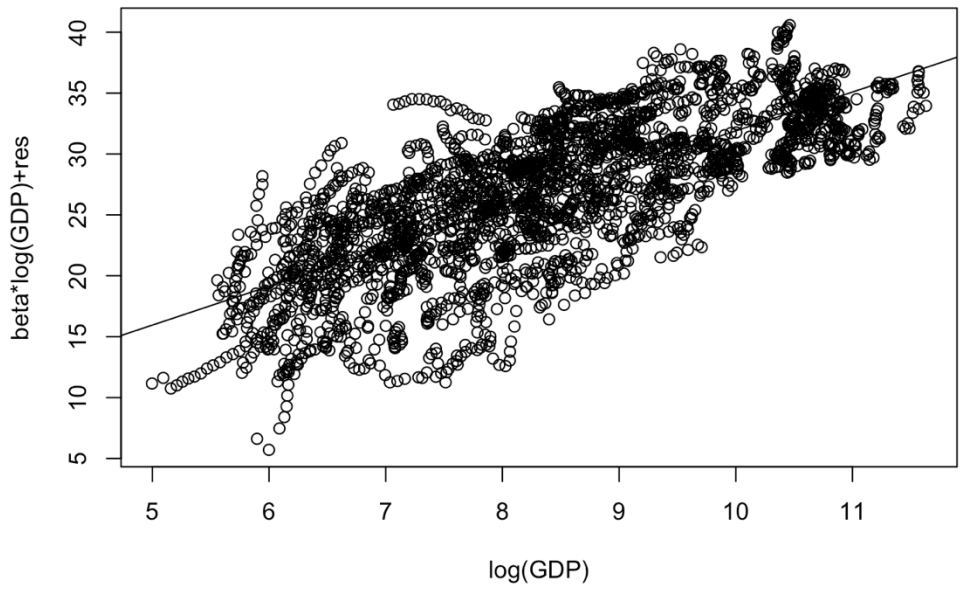


Figure 10: Partial residual plot for log(GDP).

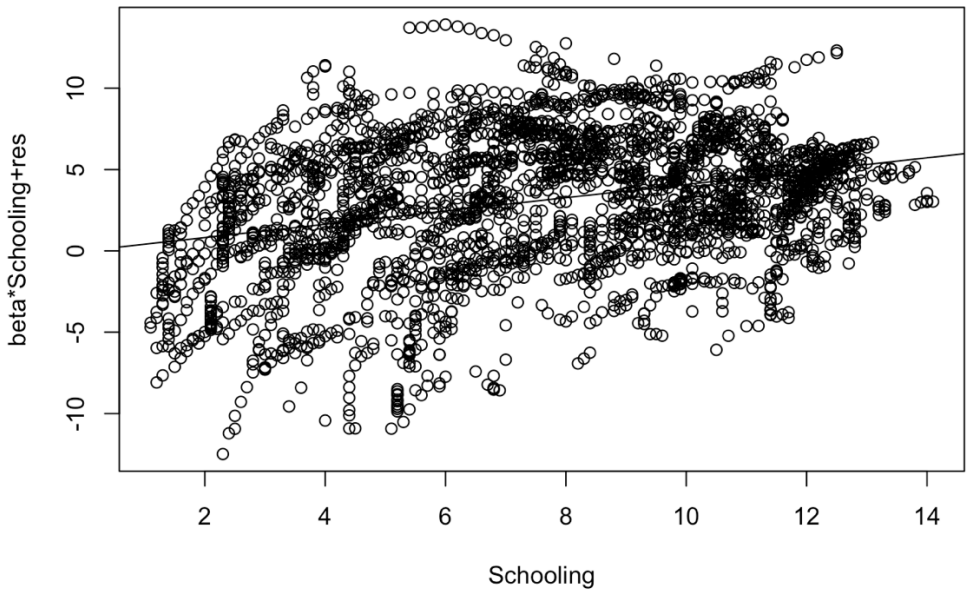


Figure 11: Partial residual plot for schooling

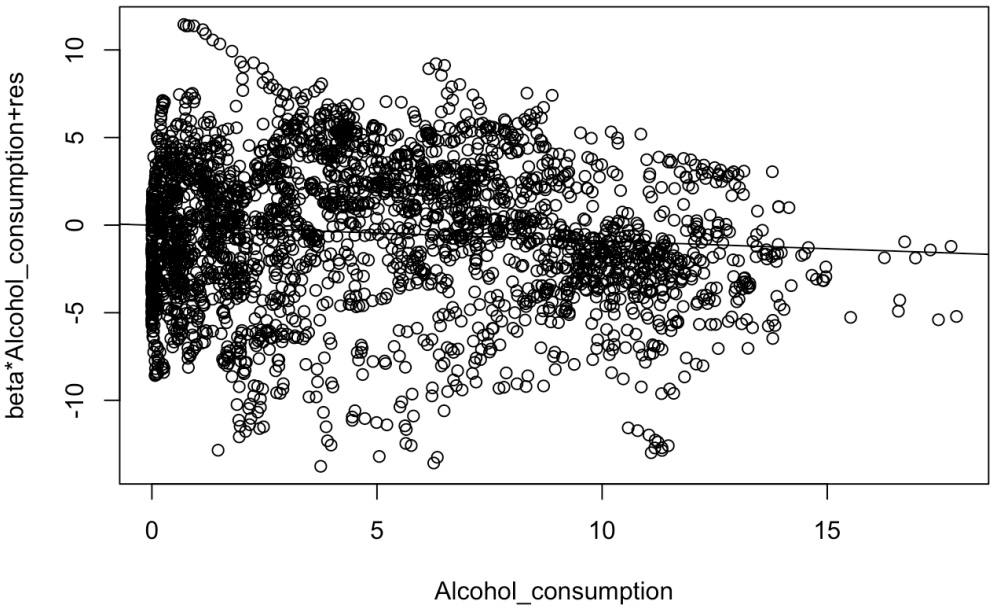


Figure 12: Partial residual plot for alcohol consumption.

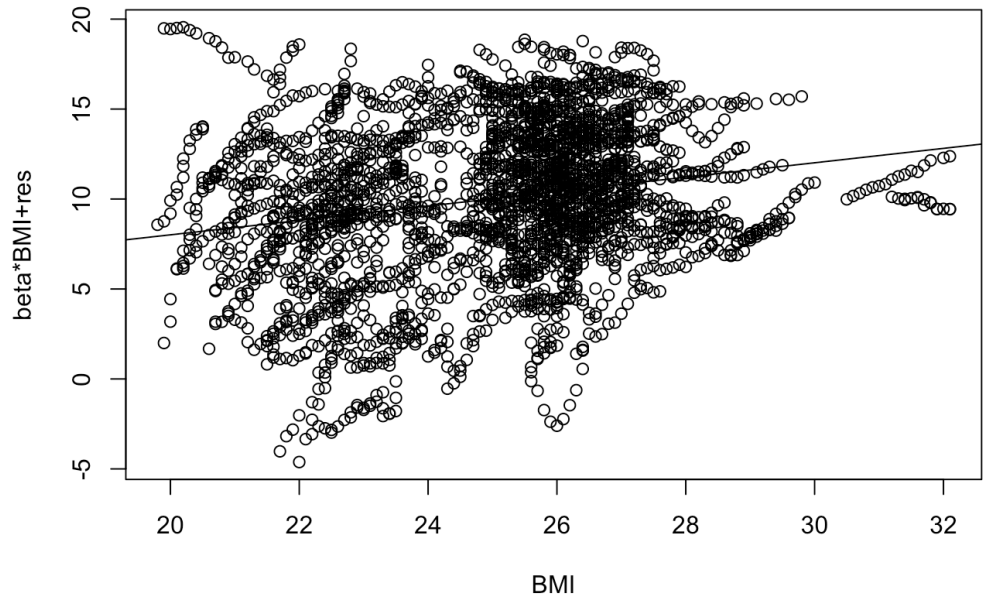


Figure 13: Partial residual plot for BMI.

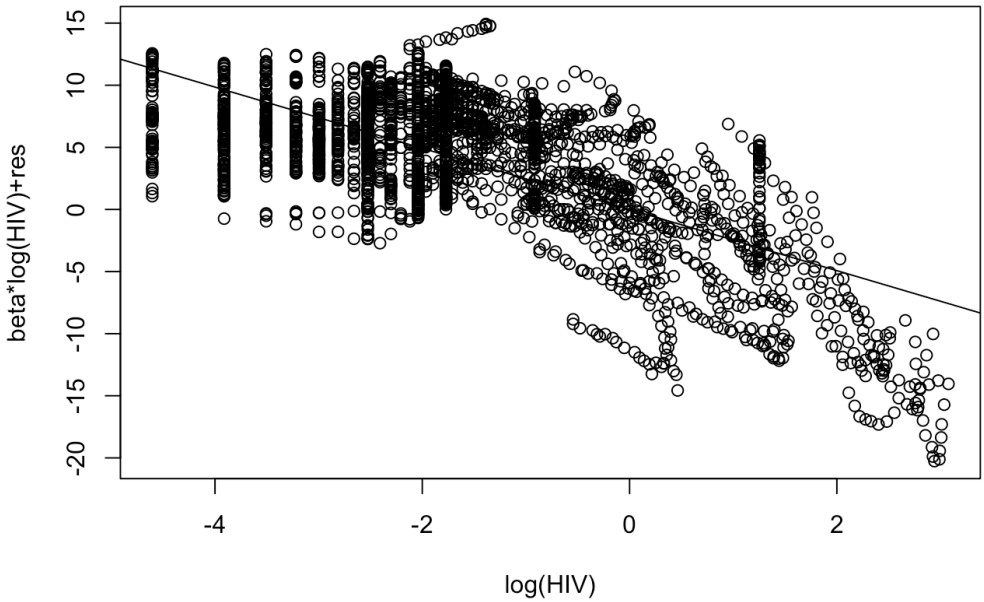


Figure 14: Partial residual plot for log(HIV).

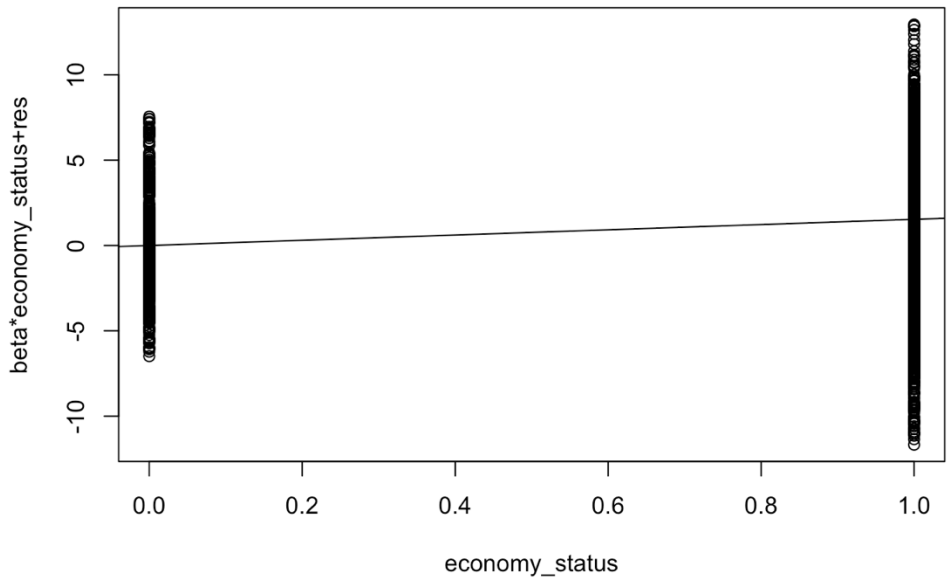


Figure 15: Partial residual plot for economy status

The partial residual plots for each predictor variable appear to be evenly distributed. These plots also indicate that each predictor variable provides some added value to a model that already includes all other covariates because the slopes of the linear relationships are all appear to be non-zero.

These partial residual plots are useful in detecting the significance of a variable in the presence of the other variables. This can also be helpful to detect possible outliers. This also serves as a tool to determine if any transformations are necessary in the predictor models. In this case, Figure 8 shows us that no transformations are required for the current model.

Automatic Variable Selection Methods helps to eliminate redundant variables from the dataset. For this, the "leap" package is used and more specifically the "regsubsets" function from the package. The best model is selected based on Mallow's C_p, BIC and adjusted R^2 or R_a^2. The selection method selects the model which has least value for Mallow's C_p and BIC. The model with the largest R_a^2 value is considered the "best" or most accurate. In this case, the best model includes all the five predictor variables. Refer to ??? below for a summary of automatic selection method.

```
## Subset selection object
## Call: regsubsets.formula(Life_expectancy ~ log(GDP) + Schooling + Alcohol_
consumption +
##     BMI + log(HIV) + economy_status, data, force.in = 1, method = "seqrep"
)
## 6 Variables  (and intercept)
##                      Forced in Forced out
## log(GDP)                  TRUE      FALSE
## Schooling                FALSE      FALSE
## Alcohol_consumption      FALSE      FALSE
## BMI                      FALSE      FALSE
## log(HIV)                 FALSE      FALSE
## economy_status           FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: 'sequential replacement'
##          log(GDP) Schooling Alcohol_consumption BMI log(HIV) economy_statu
s
## 2  ( 1 ) "*"      " "       " "                 " " "*"      " "
## 3  ( 1 ) "*"      " "       " "                 "*" "*"      " "
## 4  ( 1 ) "*"      "*"       "*"                 "*" " "      " "
## 5  ( 1 ) "*"      "*"       " "                 "*" "*"      "*"
```

```
## 6  ( 1 ) "*"       "*"        "*"                       "*" "*"       "*"
```

Figure 16: Automatic variable selection summary() output.

The summary output includes a matrix indicating which predictors are included in each of the 4 candidate models. The only predictor variable "forced in" was log(GDP). In the first model (first row of the matrix, indicated by a '2' for the number of predictors) with two predictors, only log(GDP) and log(HIV) are included. In the second model (row 2) with three (indicated by a '3') predictors, log(GDP), BMI, and log(HIV) are included. In the third model (row 3) with four (indicated by a '4') predictors, log(GDP), Schooling, Alcohol_consumption, and BMI are included. In the fourth model (row 4) with five (indicated by a '5') predictors, log(GDP), Schooling, Alcohol_consumption, BMI, and log(HIV) are included. In the fourth model (row 5) with six (indicated by a '6') predictors, log(GDP), Schooling, Alcohol_consumption, BMI, log(HIV), and economy status are included (this is the full model).

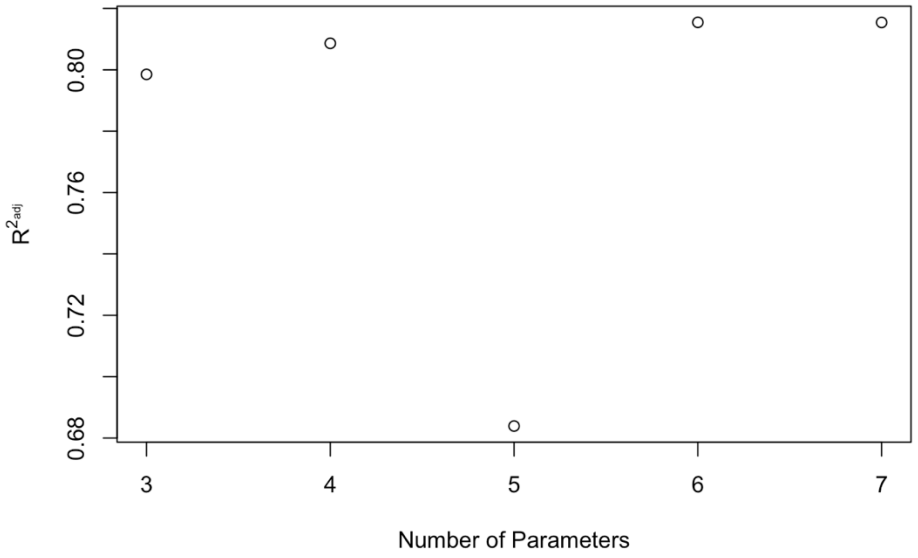


Figure 17: Plot of R_a^2 for each model.

```
## [1] 0.7985050 0.8086289 0.6839124 0.8154611 0.8154313
```

Figure 18: R_a^2 for each model.

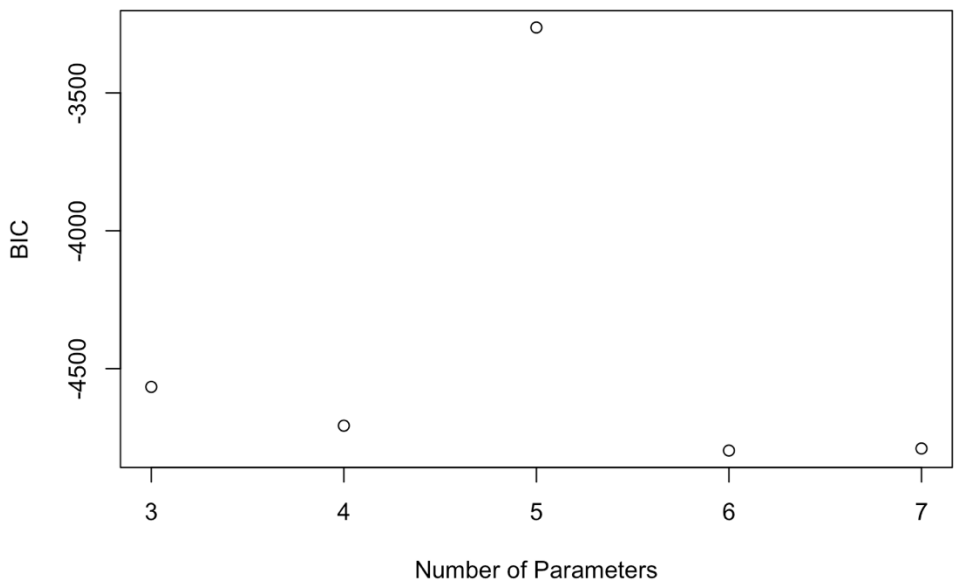


Figure x: Plot of BIC for each model.

```
## [1] -4566.222 -4706.904 -3262.776 -4797.105 -4789.686
```

Figure 19: BIC for each model

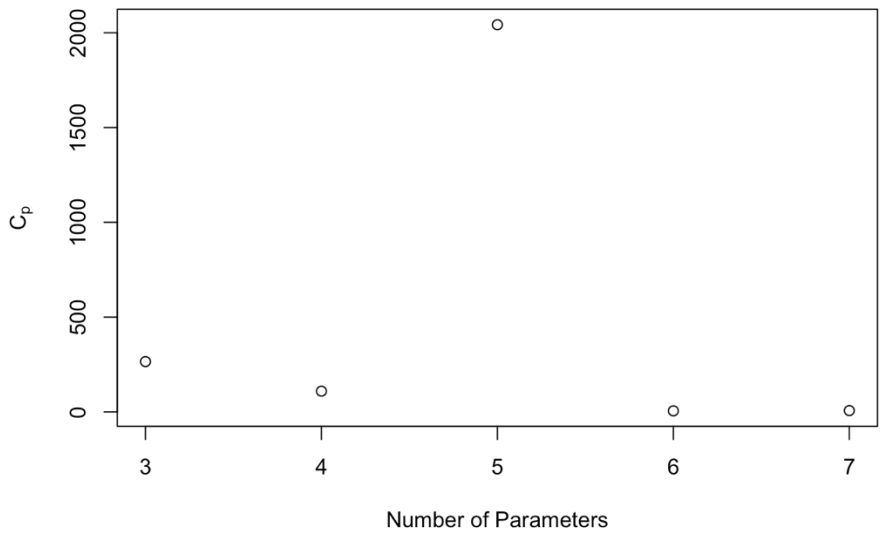


Figure 20: Plot of Mallow's C_p for each model

```
## [1]  265.375236  109.407169   2042.249700    5.539093    7.000000
```

Figure 21: Mallow's C_p for each model

```
## 46890.94
```

Figure 22: PRESS statistic for full model.

Several criteria for selecting the best model are produced, including R-adj-squared (large values are better), Bayes Information Criterion BIC (smaller values are better), and Mallow's C_p statistic (values of C_p close to p (number of beta coefficients) are better). Other criteria not produced by the regsubsets() function are AIC and $PRESS$. We will calculate these statistics for the two potential final models based on the results of automatic variable selection. Here, all statistics indicate that the best model is the reduced model including GDP, schooling, BMI, HIV and economy status: $R_a^2 = 0.8154611$, BIC = -4797.105, $C_p = 5.539093$, and PRESS = 46890.94. The second best is the full model.

C. Residual Diagnostics

1. Model Completeness

It's a good idea to also check for possible interactions (though we wouldn't hypothesize any for this analysis). The fitted-versus-residual plot looks like noise. This plot supports normality and constant variance of the residuals.

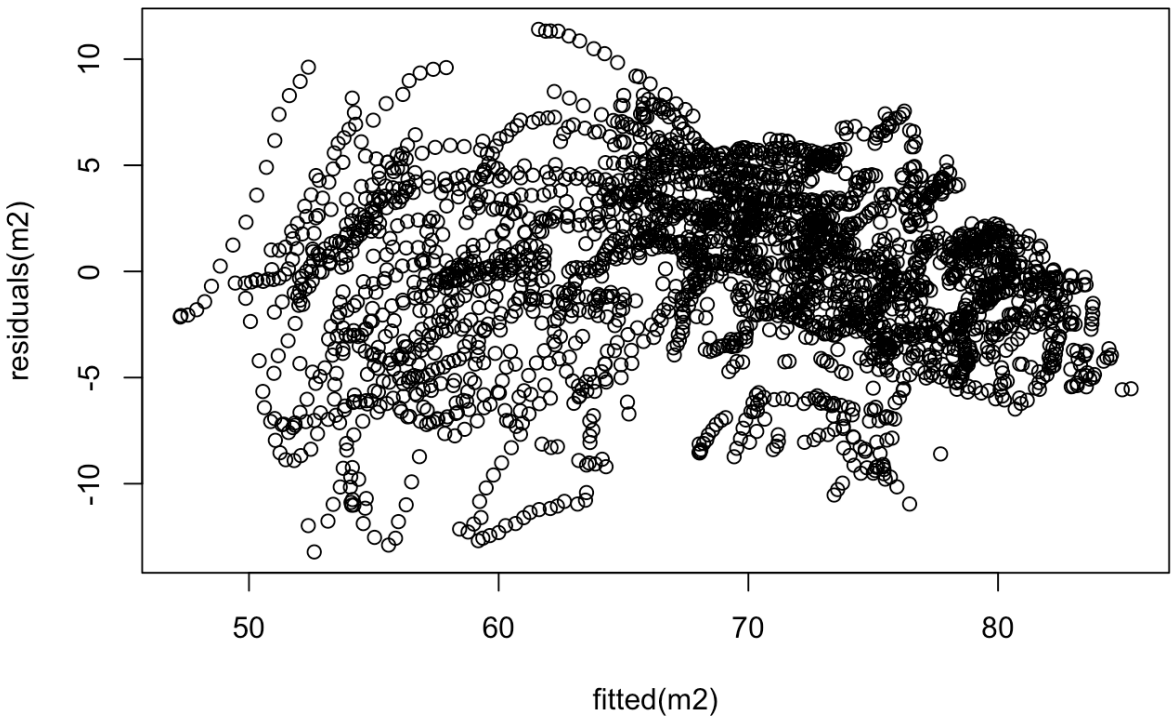


Figure 23: Plot of the fitted values vs. the residuals of the full model.

The **interaction plots** test for significant interaction between GDP and the other predictor variables using the general linear f-test. None of the interaction plots suggest the consideration of the addition of an interaction term in the model.

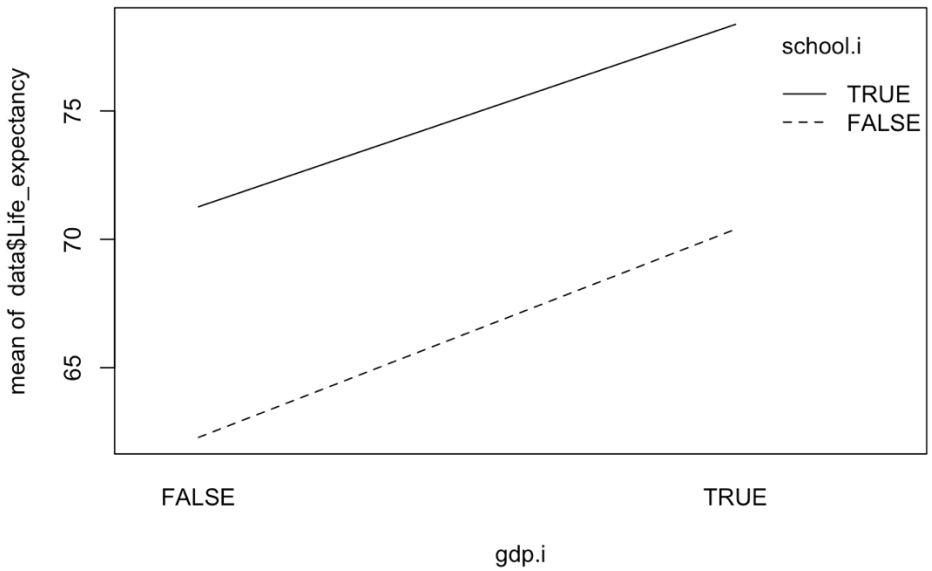


Figure 24: Interaction plot of GDP and schooling.

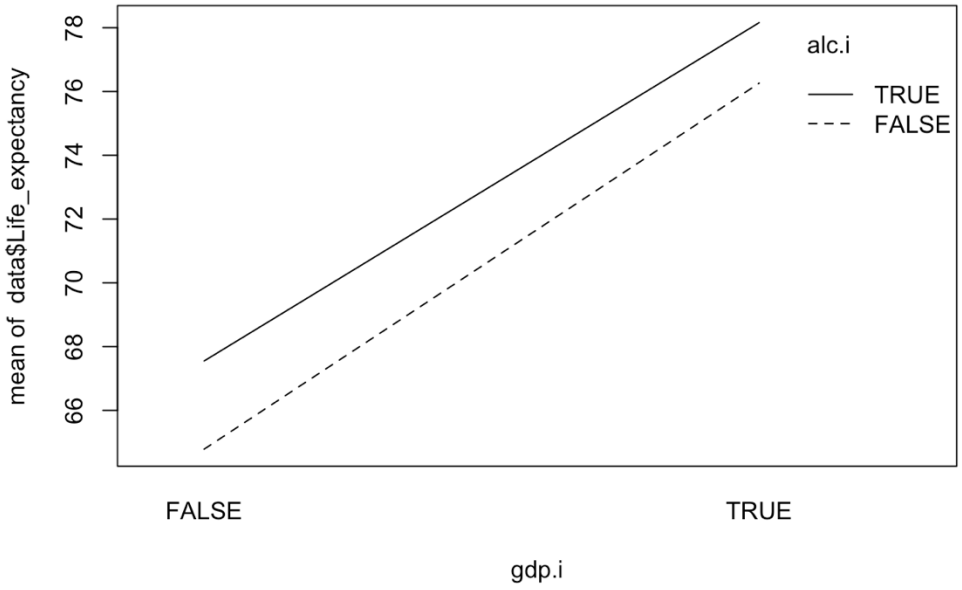


Figure 25: Interaction plot of GDP and alcohol consumption.

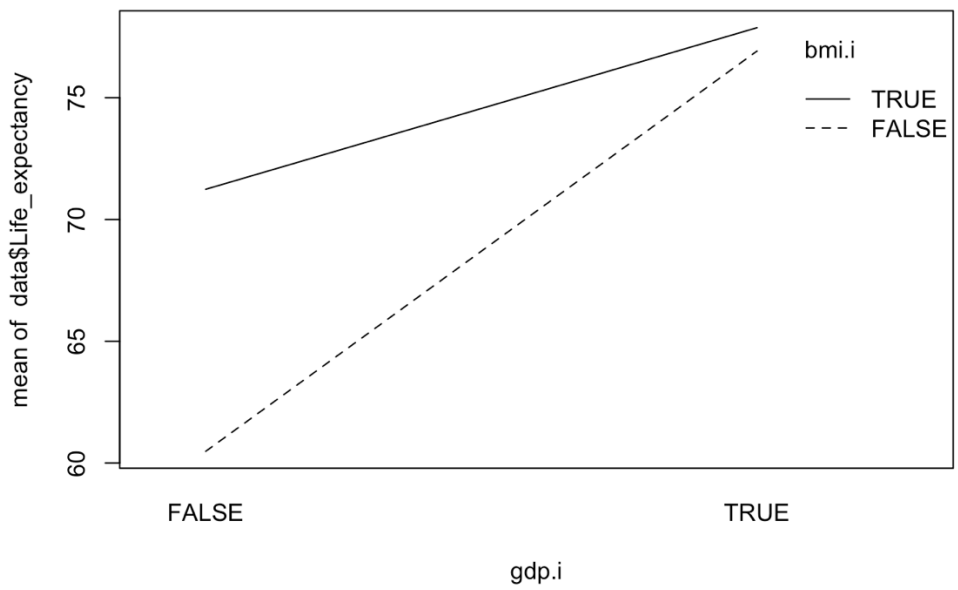


Figure 26: Interaction plot of GDP and BMI.

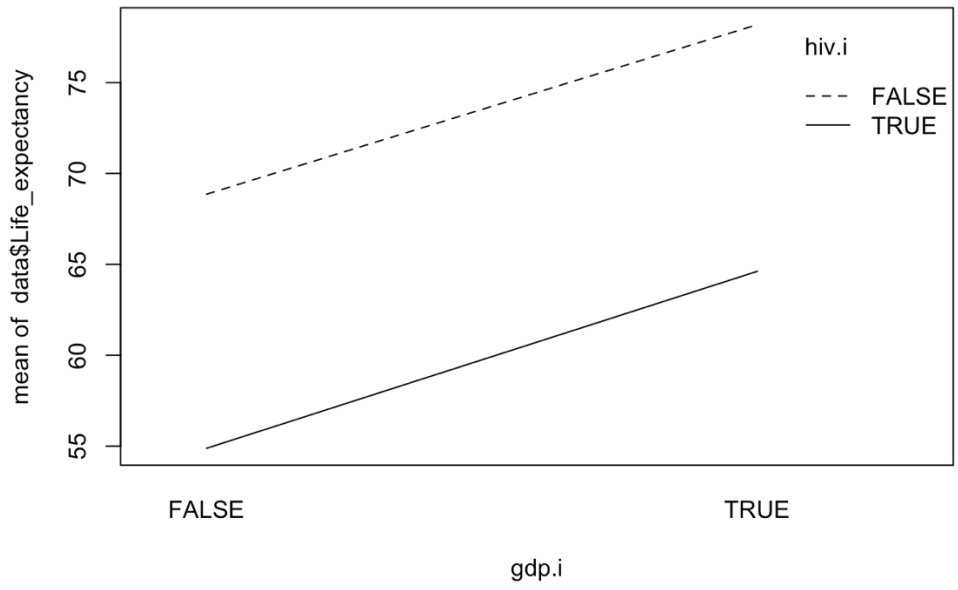


Figure 27: Interaction plot of GDP and incidents of HIV.

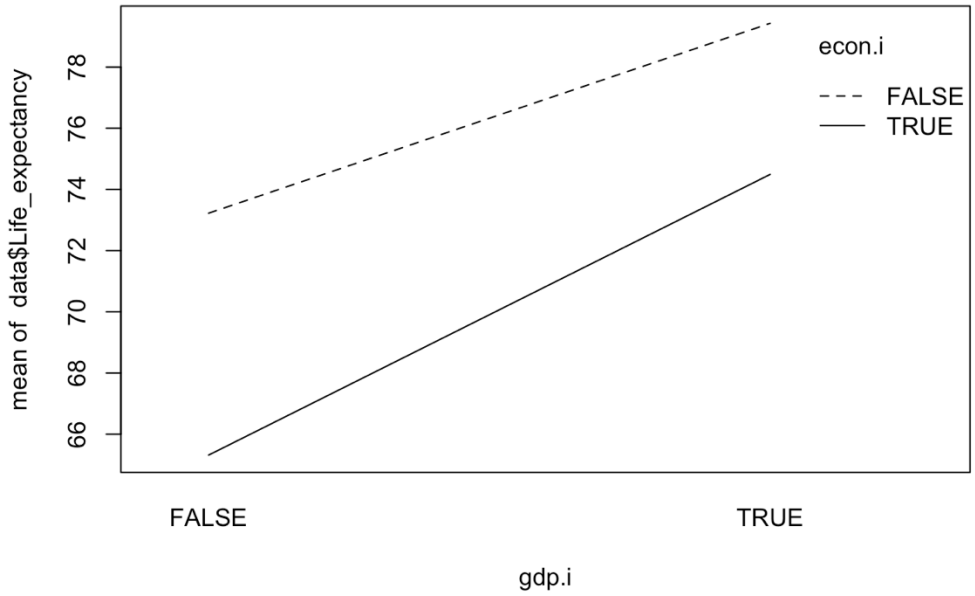


Figure 28: Interaction plot of GDP and incidents of economy status.

Based on the interaction plots for each of the predictor variables against GDP, it appears that there could be some interaction between the two variables and the addition of an interaction term could benefit the fit of the model.

D. Model Validation

Model validation can help us select the model that has the best predictive performance in a hold-out sample. There are several approaches to model validation, two of which are **Leave-one-out cross validation** specifically for smaller datasets and **K-fold cross validation** which is for larger datasets. We will use the **K-fold cross validation** meant for larger datasets.

K-fold cross validation is useful for larger datasets where training and testing data are available/feasible. This method involves:

1. Randomly split the data into k subsets. Reserve one of the subsets for testing.
2. Build (train) the model on the remaining $k-1$subsets.
3. Test the model on the reserved subset and record the mean squared prediction error.
4. Repeat the process, changing the testing subset each time, until all k subsets have served as the testing set.
5. Calculate the average of the $k$mean squared prediction errors.
6. If comparing models, the model with the lowest MSPE should be chosen.

This dataset has n = 2864 observations which is a large dataset, and we will use **K-fold cross validation** for our model validation ($k = 10$). This test is done for three models: the full model (6 predictor variables), the reduced model that includes GDP, schooling, BMI, HIV,

and economy but not alcohol consumption (5 predictor variables), and a full model plus an addition interaction term between log(GDP) and BMI.

Full model:

```
## Linear Regression
##
## 2864 samples
##    6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2577, 2579, 2579, 2577, 2577, 2578, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.044276  0.8158247  3.213774
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Reduced Model:

```
## Linear Regression
##
## 2864 samples
##    5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2578, 2577, 2578, 2576, 2578, 2579, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.044089  0.8157803  3.215835
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Interaction model:

```
## Linear Regression
##
## 2864 samples
##    7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2578, 2578, 2578, 2576, 2578, 2578, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.014525  0.8184591  3.153004
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Figure 29 (a,b,c): Summary output for K-fold cross validation for the full, reduced, and interaction models.

Using the **K-fold cross validation** method to check assumptions and outliers, we found there there were no apparent issues. K = 10 and the sample size for each fold was between 2577-2579. In this output, we are given three different values: RMSE, R-squared, and MAE. The model with the smallest RMSE and MAE and the larger R-squared is the best model. From this analysis, we can see that the interaction model with predictors, log(GDP), Schooling, Alcohol_consumption, BMI, log(HIV), economy status, and the interaction term (log(GDP)*BMI) is the best model.

2. Outliers

To detect outliers in this study, the studentized residual plot is used. Based on this plot, at least 90% of the data points should be within the range of ±3 and this is found to be true.

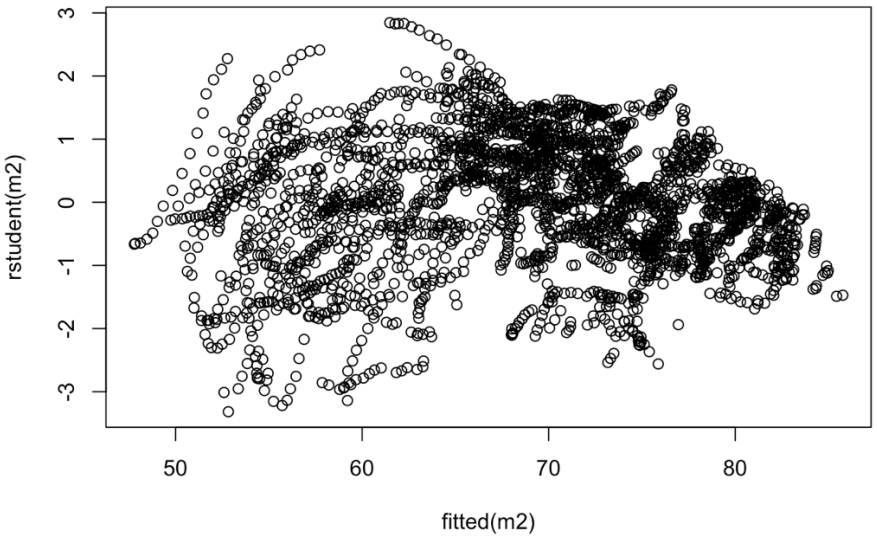


Figure 30: The fitted values plotted against the R student values.

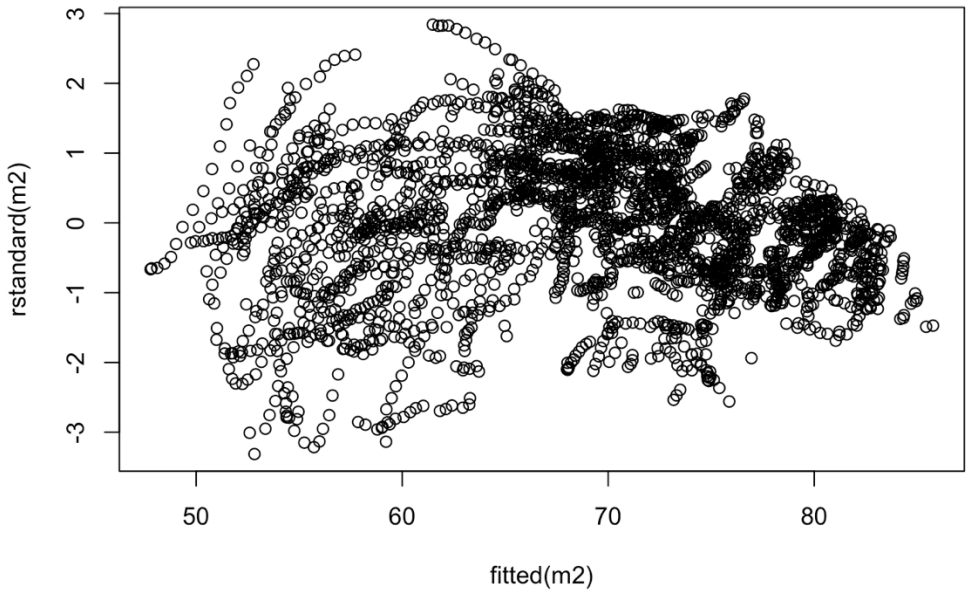


Figure 31: The fitted values plotted against the R standard values.

These datapoints have an abs(rstandard(m2)) > 3 and are considered to be outliers:

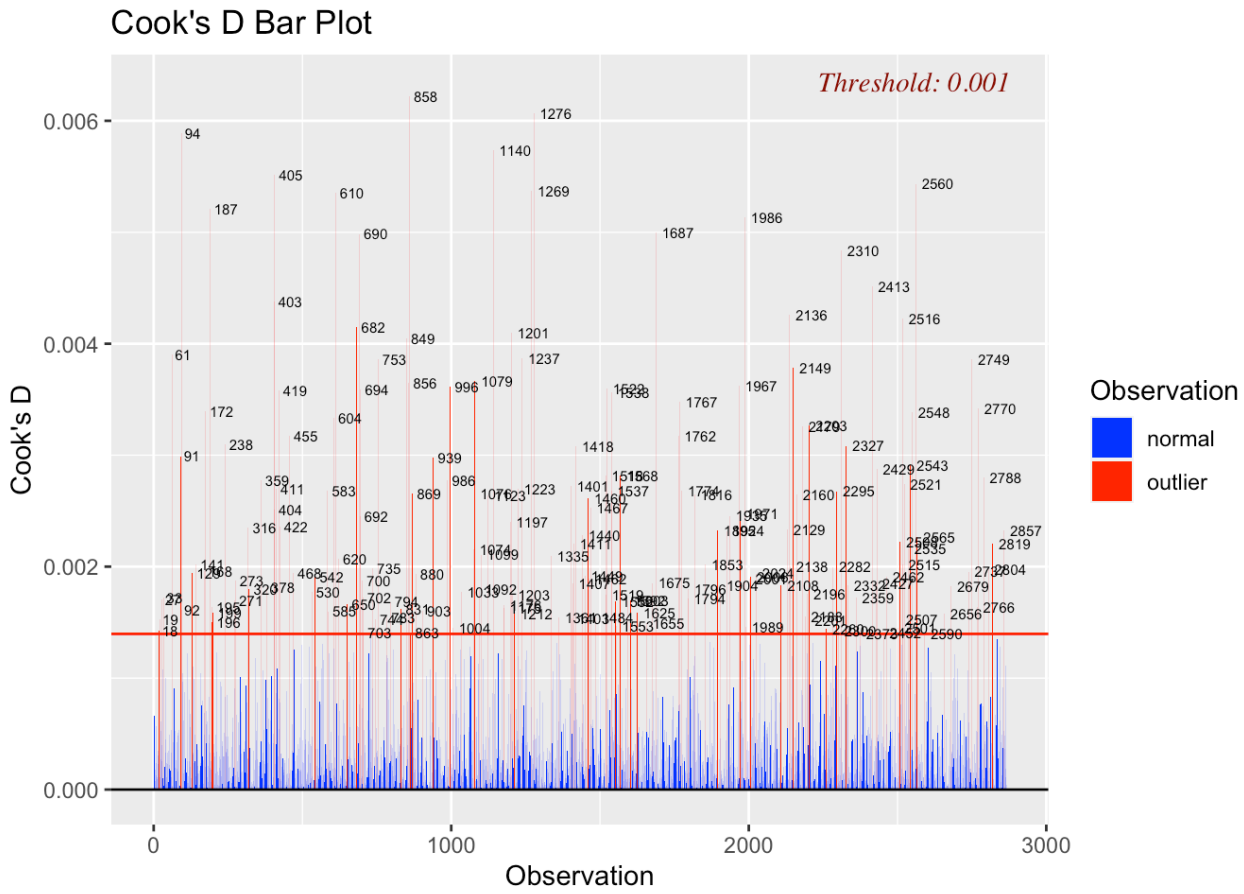
```
##   94  187  405  610  858 1269 1276 1418 1971 2560
##   94  187  405  610  858 1269 1276 1418 1971 2560
```

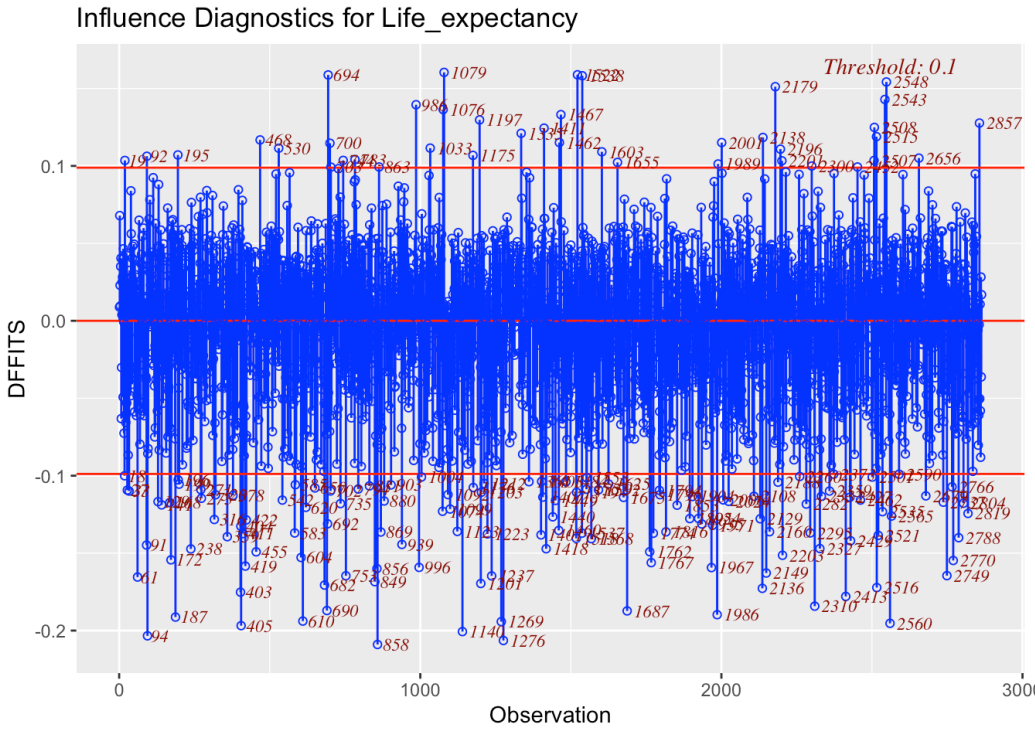
A model using the subset of the data was made excluding the above data points and the R-squared value was 0.8208. This does slightly improve the fit of the model, however, it is not enough to exclude the data points from the model.

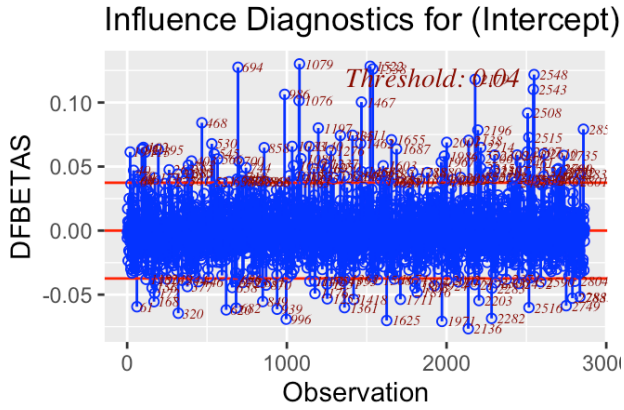
Another way to detect outliers is to analyze the studentized residual plot (above) and plot of Cook's distance (below).

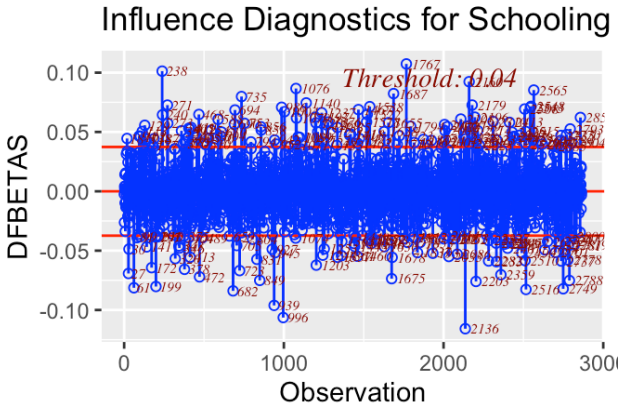
Figure 32: Plot of Cook's distance.

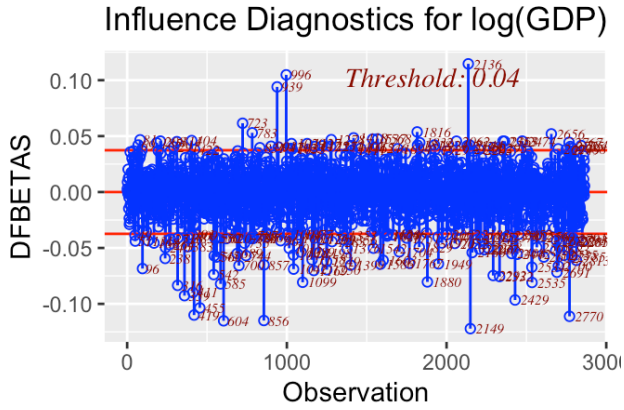
As shows by the plot, the threshold is 0.001 and there appears to be many datapoints above this threshold. This indicates that there are quite a few outliers in this dataset. We can further break up this plot into specific variable-based influential diagnostics plots focusing on DFBETAS for each predictor variable and DFITTS for the dependent variable (life expectancy).

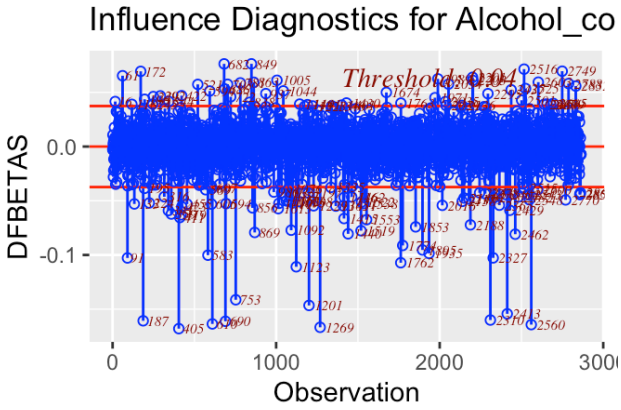


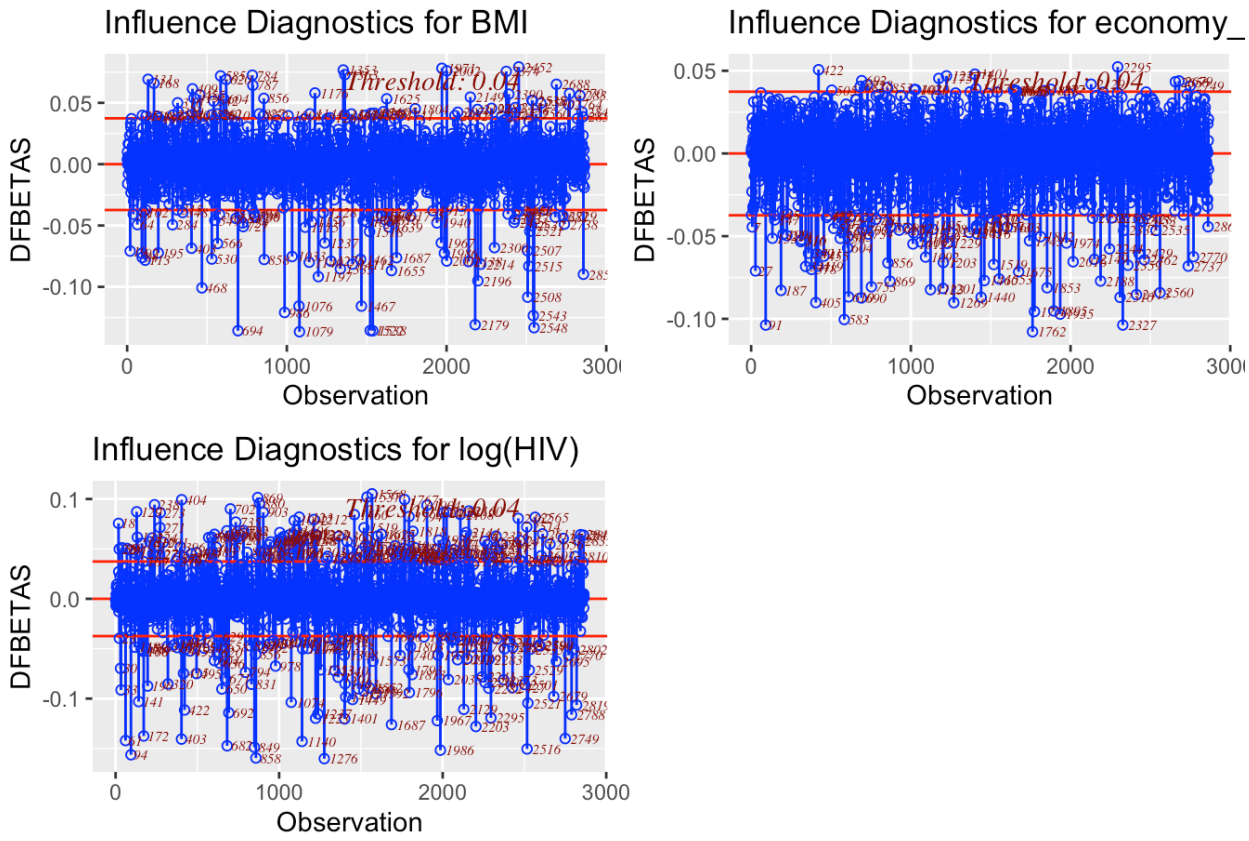

Influence Diagnostics for Life_expectancy
Threshold: 0.1
DFFITS
Observation


Influence Diagnostics for (Intercept)
Threshold: 0.04
DFBETAS
Observation


Influence Diagnostics for Schooling
Threshold: 0.04
DFBETAS
Observation


Influence Diagnostics for log(GDP)
Threshold: 0.04
DFBETAS
Observation


Influence Diagnostics for Alcohol_co
Threshold: 0.04
DFBETAS
Observation



Figures 33 (a-g): DFFITS and DFBETAS plots to determine outliers.

There do appear to be many outliers with values above the threshold. However, all of these values are real and therefore will not be dropped from the dataset.

3. Constant Variance

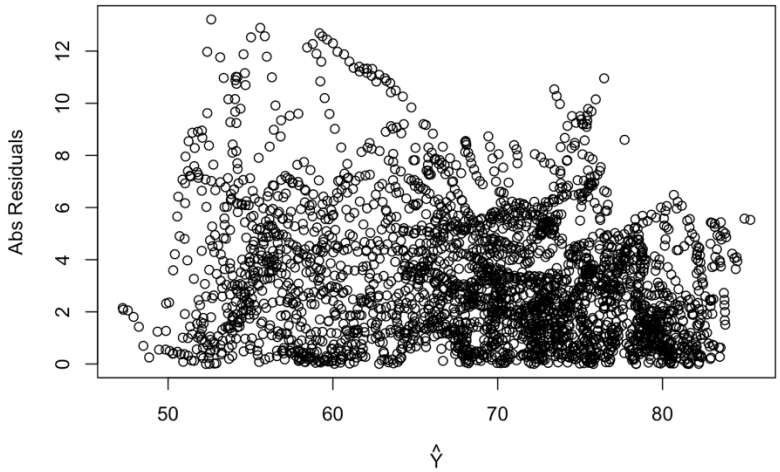


Figure 34: Abs residuals vs. fitted values of Y.

There are no apparent issues with non-constant variance. The absolute value of the residuals vs fitted values of Y appear to be normally distributed.

4. Normality

A Q-Q plot supports approximate normality.

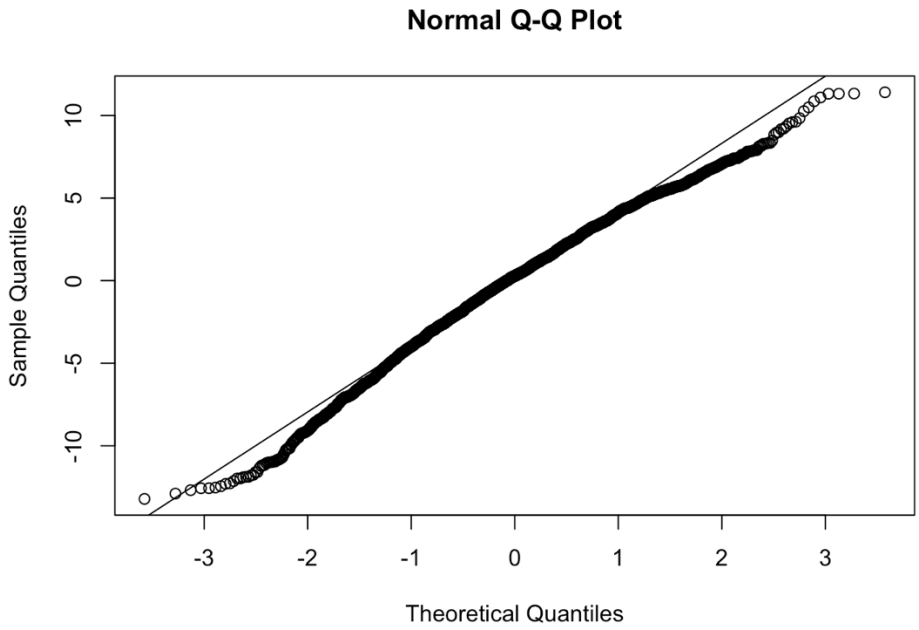


Figure 35: Normal Q-Q plot for final model.

Both tails do appear to depart from the fitted line slightly, but overall, the distribution does appear to be normal and symmetric and shows a reasonably linear pattern.

E. Testing hypotheses

For each of the predictor variables, we test if there is a linear association between them and life expectancy. The t test is run under the following conditions:

* **Null hypothesis** :: $H_0 : \beta_x = 0$
* **Alternative hypothesis** :: $H_1 : \beta_x \neq 0$.

(where x = 1, 2, 3, 4, 5, 6 ,7 for each predictor variable)

The decision is taken considering $t^* = (b_x - \beta_x) / (SE(b_x))$,
where,
t^* is the test-statistics for the t test
b_x is the observed slope coefficient
β_x is the expected slope coefficient of the fitted regression model
$SE(b_x)$ is the sampling variability of b_x

The t* is tested against $t(1 - \alpha/2, df)$, where, α is the level of significance = 0.05, and *df* is the degrees of freedom, i.e, *df = no. of observations - no. of estimate parameters = (n − 2).* If $t^* > t(1 - \alpha/2, df)$, H_0 is rejected. Otherwise, we fail to reject H_0.
The decision rule also considers the p-value and the R^2. If the p-value $\leq \alpha$, then we reject H_0. Otherwise, we fail to reject H_0. While considering the coefficient of determination (R^2), if this value is close to 1, then the association between variables is considered strong and the proportion of explained variation within Y is significantly higher than the unexplained variation. If the value is closer to 0, the model is not considered a "good fit" indicating a weak association between the variables and the unexplained variation of Y is significantly high. The t statistic = 1.64.
The predictor variables: log(GDP), schooling, BMI, log(HIV), economy, and the interaction term all are significant $p < 0.001$ level. However, the abs(t-value) for the predictor variable alcohol consumption is 1.088 which means we fail to reject H_0 and the predictor variable is not significant to the model. This concludes our final model which is described in the next section.

F. The Final Model

The **final model** is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \varepsilon_i$$

where $\varepsilon_i \sim iidN(0, \sigma^2)$, $i = 1,2,3, \cdots, 2864$ and $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7,$ and σ^2 are the unknown parameters to be estimated.

V. Conclusions

In real terms, the final model can be expressed as:

$$Y_i = -13.29700 + 8.15489X_{i1} + 0.35855X_{i2} + 1.85970\,X_{i3} - 2.55546X_{i4} + 1.27243X_{i5} - 0.19185X_{i6} + \varepsilon_i$$

where

Y_i = , life expectancy, the life expectancy for residents for the i^{th} country,
X_{i1}= GDP, gross domestic product per capita in current USD for the i^{th} country,
X_{i2}= Schooling, average years that people aged 25+ spent in formal education for the i^{th} country,
X_{i3} = BMI, average BMI measurement recorded for the i^{th} country,
X_{i4}= HIV, incidents of HIV per 1000 population aged 15-49 for the i^{th} country.
X_{i5}= economy, if variable = 0, the country is developed, if variable = 1 the country is developing,
X_{i6}= interaction term, $X_{i1} * X_{i4}$ (log(GDP) * BMI).

ε_i is the error term; $\varepsilon_i \sim iidN(0, \sigma^2)$
$i = 1,2,3, \cdots, 2864$.
$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6,$ and σ^2 are the unknown parameters to be estimated.

```
Call:
lm(formula = Life_expectancy ~ log(GDP) + Schooling + BMI + log(HIV) +
    economy_status + interaction, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-13.3567  -2.2835   0.2905   2.8170  11.8976

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -13.29700    5.96498  -2.229   0.0259 *
log(GDP)         8.15489    0.76603  10.646  < 2e-16 ***
Schooling        0.35855    0.04300   8.339  < 2e-16 ***
BMI              1.85970    0.24713   7.525 7.02e-14 ***
log(HIV)        -2.55546    0.05385 -47.456  < 2e-16 ***
economy_status   1.27243    0.27234   4.672 3.12e-06 ***
interaction     -0.19185    0.03042  -6.306 3.31e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.013 on 2857 degrees of freedom
Multiple R-squared:  0.8183,   Adjusted R-squared:  0.8179
F-statistic:  2145 on 6 and 2857 DF,  p-value: < 2.2e-16
```

Figure 36: Summary table of final model.

```
Analysis of Variance Table

Response: Life_expectancy
                Df Sum Sq Mean Sq  F value    Pr(>F)
log(GDP)         1 160301  160301 9952.359 < 2.2e-16 ***
Schooling        1   8463    8463  525.447 < 2.2e-16 ***
BMI              1   1534    1534   95.266 < 2.2e-16 ***
```

```
log(HIV)           1  35704   35704 2216.671 < 2.2e-16 ***
economy_status     1    616     616   38.260 7.078e-10 ***
interaction        1    640     640   39.764 3.310e-10 ***
Residuals       2857  46017      16
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 37: Anova table of final model.

This study shows that a country's average GDP, years of schooling, BMI, HIV, and an interaction term between GDP and BMI can predict the life expectancy of residents within that country. All statistical analysis was conducted at a 95% confidence interval and at 0.05 significance level (α). A linear association was found between 5 of the 6 original predictor variables with an additional linear association from an interaction term.

Figure 36 above show the estimated regression coefficient, the standard error, t value, p value associated with each of the predictors, R^2, adjusted R^2, MSE and F statistics of the final model. Figure 37 shows the ANOVA table for the final model having SSE, SSR, MSE, MSR, F values and the corresponding p-value. Through analyzing these test statistics, we can conclude that the strongest association is between log(GDP) and life expectancy.

There was a strong presence of outliers within the dataset. However, while trimming the dataset of these outliers would have likely increased the adjusted value, it would have compromised the integrity of the model even if the fit would have been "better". All of the original datapoints were kept in the dataset and were used in the analysis of the model.

VI. Code Appendix

```{r setup, include=FALSE}

# Load packages
library(tidyverse)
library(caret)
library(asbio)
library(olsrr)
library(xtable)
library(shiny)
library(knitr)
library(DT)
require(scatterplot3d)
require(Hmisc)
require(rgl)
require(faraway)
```

```
library(car)
data(chredlin)
attach(chredlin)


# declare global chunk options
knitr::opts_chunk$set(echo = FALSE) # Turn off print for all code chunks simultaneously

# determine output format dynamically
out_type <- knitr::opts_knit$get("rmarkdown.pandoc.to")

# define custom function for data label outputs
# The DT::datatable function is great for producing tables for HTML docs
# Otherwise, use the knitr::kable function to produce tables
# You should use the R help to learn about these two functions as they
# will need to be used to produce visually appealing tables for your
# report

display_output <- function(dataset, out_type, filter_opt = 'none') {

  if (out_type == "html") {
    out_table <- DT::datatable(dataset, filter = filter_opt)
  } else {
    out_table <- knitr::kable(dataset)
  }

  out_table
}

# Function to calculate predicted sum of squares (PRESS)
PRESS <- function(linear.model) {
  #' calculate the predictive residuals
  pr <- residuals(linear.model)/(1-lm.influence(linear.model)$hat)
  #' calculate the PRESS
  PRESS <- sum(pr^2)

  return(PRESS)
}

data <- read_csv("Life-Expectancy-Data-Updated.csv",
    col_types = cols(Infant_deaths = col_skip(),
        Under_five_deaths = col_skip(), Adult_mortality = col_skip(),
        Hepatitis_B = col_skip(), Measles = col_skip(),
        Polio = col_skip(), Diphtheria = col_skip(),
        Population_mln = col_skip(), Thinness_ten_nineteen_years = col_skip(),
        Thinness_five_nine_years = col_skip(),
```

```
    Schooling = col_double(), Economy_status_Developed = col_skip(),
    Life_expectancy = col_double()))
data <- rename(data, GDP = GDP_per_capita, HIV = Incidents_HIV, economy_status =
Economy_status_Developing)
```

```{r describe}
# the display_output function was defined above, it's producing a table
# for each of the calls below
#display_output(data, out_type)
head(data)
```

```{r}
# I didn't even get this entered until I did the Appendix steps!
# Fit the final model in order to describe it
data$interaction <- log(data$GDP)*data$BMI
m2 <- lm(Life_expectancy ~ log(GDP) + Schooling + BMI + log(HIV) + economy_status +
interaction, data)
```

```{r}
summary(m2)
anova(m2)
confint(m2)
```

```{r}
# Fit the initial model in order to refine it
m1 <- lm(Life_expectancy ~ log(GDP) + Schooling + Alcohol_consumption + BMI +
log(HIV) + economy_status, data)
```

A scatterplot matrix indicates positive linear associations between all variables.

```{r}
pairs(Life_expectancy ~ log(GDP) + Schooling + Alcohol_consumption + BMI + log(HIV)
+ economy_status, data)
```

```{r}
cor(data[,c(10,7,8,4,5,6, 9)])
```

```{r}
boxplot(log(data$GDP), main = "log(GDP)")
stripchart(log(data$GDP), vertical = T, method = "jitter", main = "log(GDP)")

boxplot(data$Schooling, main = "Schooling")
stripchart(data$Schooling, vertical = T, method = "jitter", main = "Schooling")

boxplot(data$Alcohol_consumption, main = "Alcohol_consumption")
stripchart(data$Alcohol_consumption, vertical = T, method = "jitter", main =
"Alcohol_consumption")

boxplot(data$BMI, main = "BMI")
stripchart(data$BMI, vertical = T, method = "jitter", main = "BMI")

boxplot(log(data$HIV), main = "HIV")
stripchart(log(data$HIV), vertical = T, method = "jitter", main = "log(HIV)")

boxplot(data$economy_status, main = "data$economy_status")
stripchart(data$economy_status, vertical = T, method = "jitter", main = "economy_status")
```

```{r}
prplot(m1,1)
prplot(m1,2)
prplot(m1,3)
prplot(m1,4)
prplot(m1,5)
prplot(m1,6)
```

```{r}
vif(m1)
vif(m2)
```

3. **Automatic variable selection methods**

```{r}
library(leaps)
ma <- regsubsets(Life_expectancy~log(GDP) + Schooling + Alcohol_consumption + BMI + log(HIV) + economy_status, data, force.in = 1, method = "seqrep")
sma <- summary(ma)
sma
```

```{r}
sma$adj # Adjusted R2 big
plot(3:7,sma$adj, xlab = "Number of Parameters", ylab = expression(R^2[adj]))
sma$bic # BIC small
plot(3:7, sma$bic, xlab = "Number of Parameters", ylab = expression(BIC))
sma$cp # Cp = p
plot(3:7, sma$cp, xlab = "Number of Parameters", ylab = expression(C[p]))
```

```
sma$bic

# Extract PRESS
PRESS(m2)
```

C. Model Validation

Model validation can help us select the model that has the best predictive performance in a hold-out sample. There are several approaches to model validation, two of which are **Leave-one-out cross validation** specifically for smaller datasets and **K-fold cross validation** which is for larger datasets. We will use the **K-fold cross validation** meant for larger datasets.

K-fold cross validation is useful for larger datasets where training and testing data are available/feasible. This method involves:

1. Randomly split the data into \(k\) subsets. Reserve one of the subsets for testing.
2. Build (train) the model on the remaining \(k-1\) subsets.
3. Test the model on the reserved subset and record the mean squared prediction error.

4. Repeat the process, changing the testing subset each time, until all \(k\) subsets have served as the testing set.
5. Calculate the average of the \(k\) mean squared prediction errors.
6. If comparing models, the model with the lowest MSPE should be chosen.

```{r}
# Define training control
set.seed(123)
train.control <- trainControl(method = "cv", number = 10)
# Train the model
fullmodel <- train(Life_expectancy~log(GDP) + Schooling + Alcohol_consumption + BMI + log(HIV) + economy_status, data, method = "lm",
               trControl = train.control)

reducedmodel <- train(Life_expectancy~log(GDP) + Schooling + BMI + log(HIV) + economy_status, data, method = "lm",
               trControl = train.control)

interactionmodel <- train(Life_expectancy~log(GDP) + Schooling + Alcohol_consumption + BMI + log(HIV) + economy_status+ interaction, data, method = "lm",trControl = train.control)

# Summarize the results
print(fullmodel)
print(reducedmodel)
print(interactionmodel)
```

D. Residual Diagnostics

1. **Model Completeness**

```{r}
plot(residuals(m2)~fitted(m2)) # Model looks appropriate

gdp.i <- data$GDP > mean(data$GDP)
school.i <- data$Schooling > mean(data$Schooling)
alc.i <- data$Alcohol_consumption > mean(data$Alcohol_consumption)
bmi.i <- data$BMI > mean(data$BMI)
hiv.i <- data$HIV > mean(data$HIV)
econ.i <- data$economy_status > mean(data$economy_status)
interaction.plot(gdp.i, school.i, data$Life_expectancy)
interaction.plot(gdp.i,alc.i,data$Life_expectancy)
interaction.plot(gdp.i,bmi.i,data$Life_expectancy)
```

interaction.plot(gdp.i,hiv.i,data$Life_expectancy)
interaction.plot(gdp.i,econ.i,data$Life_expectancy)

Test for significant interaction using general linear f-test
#m3 <- lm(involact~race+fire+theft+age+race*theft)
#anova(m3) # Doesn't look like it is important but should probably consider for model validation, just in case

```

#### 2. **Outliers**

```{r}
plot(residuals(m2)~fitted(m2))
plot(rstudent(m2)~fitted(m2)) #Studentized residual
identify(rstudent(m2)~fitted(m2))
plot(rstandard(m2)~fitted(m2)) #Deleted studentized residual
```

These datapoints have an abs(rstandard(m2)) > 3:
```{r}
which(abs(rstandard(m2)) > 3)
m3 <- lm(Life_expectancy ~ log(GDP) + Schooling + BMI + log(HIV) + economy_status + interaction, data, subset = -c(94,187,405,610,858,1269,1276, 1418,1971,2560))
summary(m3)
```

Difference in fits, difference in betas, and Cooks distance plots:
```{r}
plot(dffits(m2)) # Compare to 2sqrt(p/n) (0.09) for large datasets and 1 for small
which(dffits(m2)>0.08)
which(dfbetas(m2)>0.04) # Compare to 2/sqrt(n) for large datasets and 1 for small
plot(cooks.distance(m2)) # Compare percentile F(p,n-p) to 10th or 20th
q <- pf(cooks.distance(m2),7,2864-7)
which(q>.1)
which(q>.2)
```

OLS Cook's D plot, dfbetas, and dffits plots are shown below.
```{r}
ols_plot_cooksd_bar(m2) # One way to visualize Cook's distance
```
```

```
ols_plot_dfbetas(m2) # Visualize influence on estimation of betas
ols_plot_dffits(m2) # Visualize influence on estimation of Y
```

3. Constant Variance

```{r}
plot(abs(residuals(m2))~predict(m2), xlab = expression(hat(Y)), ylab = "Abs Residuals")
```

4. Normality

A Q-Q plot supports approximate normality.

```{r}
# with outliers
qqnorm(residuals(m2))
qqline(residuals(m2))
```

Both tails do appear to depart from the fitted line slightly, but overall the distribution does appear to be normal and symmetric and shows a reasonably linear pattern.

VII. References

1. B;, Crimmins EM; Preston SH; Cohen. “Explaining Divergent Levels of Longevity in High-Income Countries.” *National Center for Biotechnology Information*, U.S. National
2. Kaggle, www.kaggle.com/datasets. Accessed 14 Dec. 2023.