WHO Life Expectancy Data Analysis

STAT 4355.001 Team Project

Team 1: Alt + F4

Authors:
Luke Robbins, Hamza Dahshi, Rachel Holley

May 13th, 2021

Introduction

Health indicators are one of the most important variables a country can focus on for the longevity of its people. As one of these variables, life expectancy is widely used as an indicator to the status of a country, and the livelihood of their population. Since life expectancy can help describe a country in this way, we chose to analyze the collected WHO and UNESCO data to try to answer the following question: How can a country improve their life expectancy?

The CDC shows that in the United States, life expectancy at birth has remained fairly stagnant at an average of 78 years (Centers for Disease Control and Prevention, n.d.), and while the world life expectancy has continued to increase, the percentage increase by year is getting smaller. (*MacroTrends, n.d.*) This begs the question, what relationships can we find that explain the most efficient way to raise the life expectancy of a country, given popular metrics tracked by world organizations or the country themselves.

Data Description

Our data set was taken from Kaggle. The author chose several variables that they thought were important to life expectancy reported from both the GHO and UNESCO. These variables include economic data, health data, educational indicators, and demographic information. The life expectancy variable has been reported by the WHO for many years, and was matched with the corresponding GHO (Global Health Organization) and UNESCO (United Nations Educational, Scientific and Cultural Organisation) data for that year. In total, the data set has 3,111 observations with 32 variables. These variables include data for 183 countries over the years 2000 to 2016.

The list of variables is as follows:

country: Country name
country_code: 3 letter shorthand for the country
region: Continent/region of earth where country is located
year: Year of data, ranges from 2000 to 2016
life_expectancy: Life Expectancy at birth in years
life_exp_at_60: Life Expectancy at the age of 60 in years

adult_mort: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population), values between 0 and 1000

infant_mort: Death rate up to age 1 years old, values between 0 and 1

age1_4mort: Death rate between ages 1 and 4, values between 0 and 1

alcohol: Alcohol, recorded per capita (age 15+) consumption (in litres of pure alcohol)

Bmi: Mean BMI (kg/mˆ2) (18+) (age-standardized estimate)

age5_19thinness: Prevalence of thinness among children and adolescents, BMI < (median - 2 s.d.) (crude estimate) (% values between 0 and 100)

age5_19obesity: Prevalence of obesity among children and adolescents, BMI > (median + 2 s.d.) (crude estimate) (% values between 0 and 100)

hepatitis: Hepatitis B (HepB) immunization coverage among 1-year-olds (% values between 0 and 100)

measles: Measles-containing-vaccine first-dose (MCV1) immunization coverage among 1-year-olds (% values between 0 and 100)

polio: Polio (Pol3) immunization coverage among 1-year-olds (% values between 0 and 100)

diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (% values between 0 and 100)

basic_water: Population using at least basic drinking-water services (% values between 0 and 100)

doctors: Medical doctors (per 10,000)

hospitals: Total density of Hospitales per 100 000 population

gni_capita: Gross national income per capita, purchasing power parity international U.S. dollar (PPP int.$)

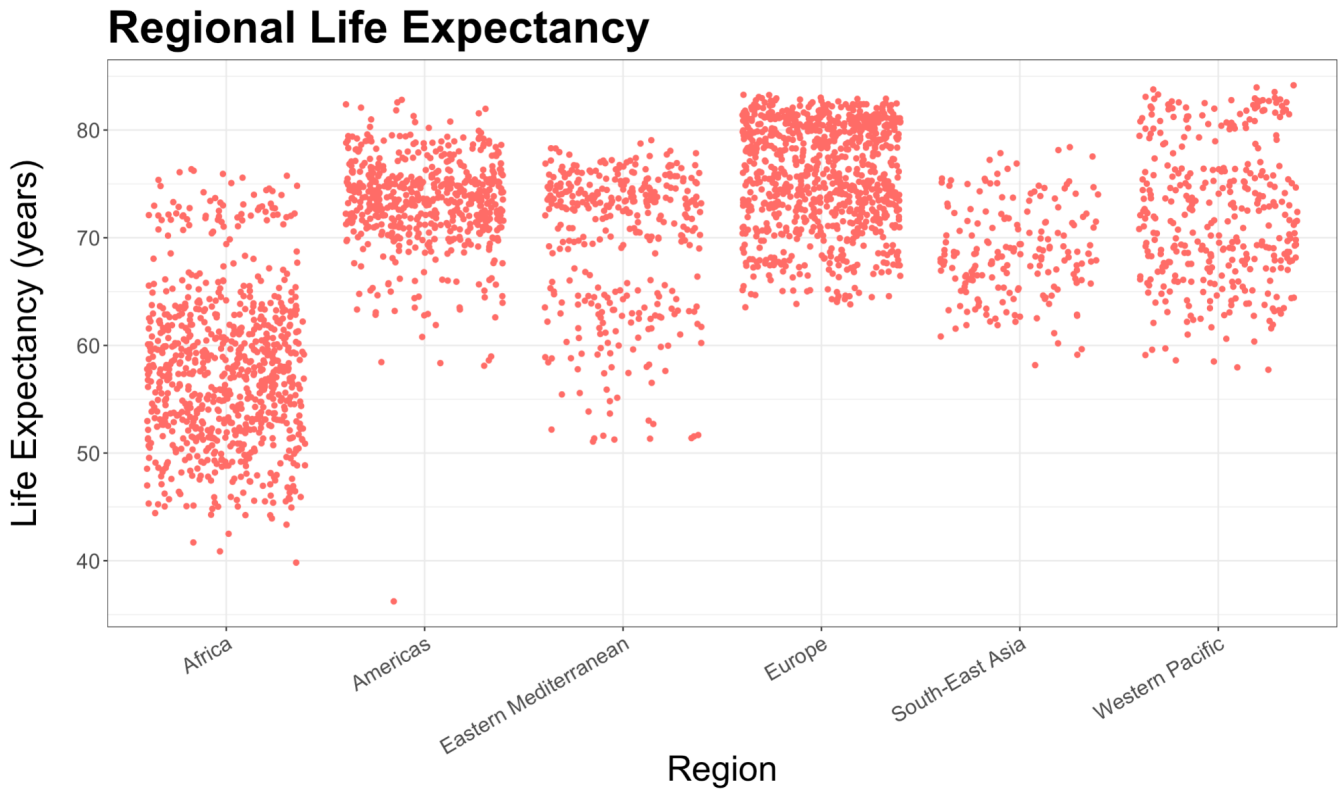
gghe_d: Domestic general government health expenditure (GGHE-D) as percentage of gross domestic product (GDP) (% values between 0 and 100). Share of current health expenditures funded from general government sources, social health insurance and compulsory prepayment

che_gdp: Current health expenditure (CHE) as percentage of gross domestic product (GDP) (% values between 0 and 100). Level of Current Health Expenditure expressed as a percentage of GDP
une_pop: Population (thousands)
une_infant: Mortality rate, infant (per 1,000 live births)
une_life: Life expectancy at birth, total (years)
une_hiv:Prevalence of HIV, total (% of population ages 15-49)
une_gni: GNI per capita, (PPP int. $)
une_poverty: Poverty headcount ratio at $1.90 a day (PPP) (% of population)
une_edu_spend: Government expenditure on education as a percentage of GDP (% values between 0 and 100)
une_literacy: Adult literacy rate, population 15+ years, both sexes (% values between 0 and 100)
une_school: Mean years of schooling (ISCED 1 or higher), population 25+ years, both sexes

Using these variables we can look at the relationships they have with life expectancy. All of the following graphs will be plotting our potential regressors against life expectancy (the response variable on the y-axis). This will include all data (that is available) from the years 2000 to 2016.

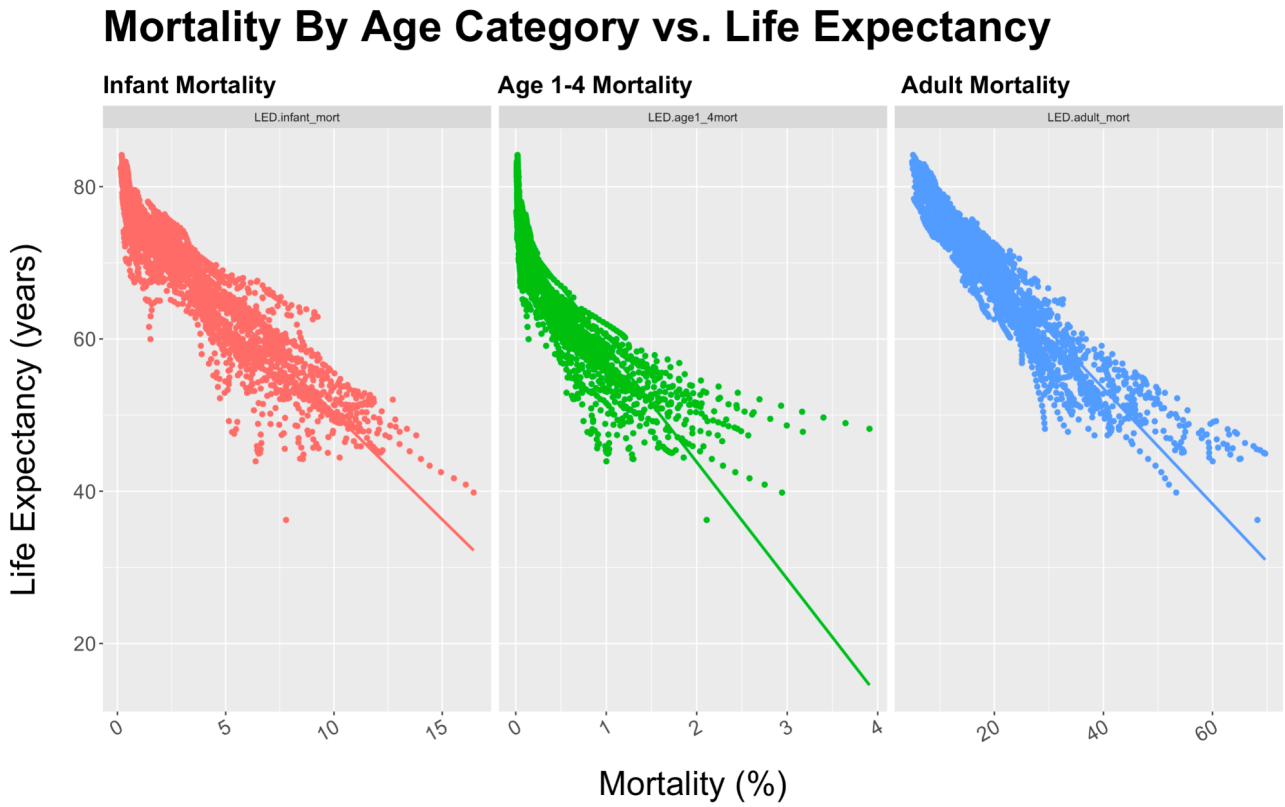
Demographic regressors:

Note that regions that are generally considered developed have a higher life expectancy. Although our analysis deals with the world as a whole, it is interesting to see how each region of the world has its own mean and variance for life expectancy values.



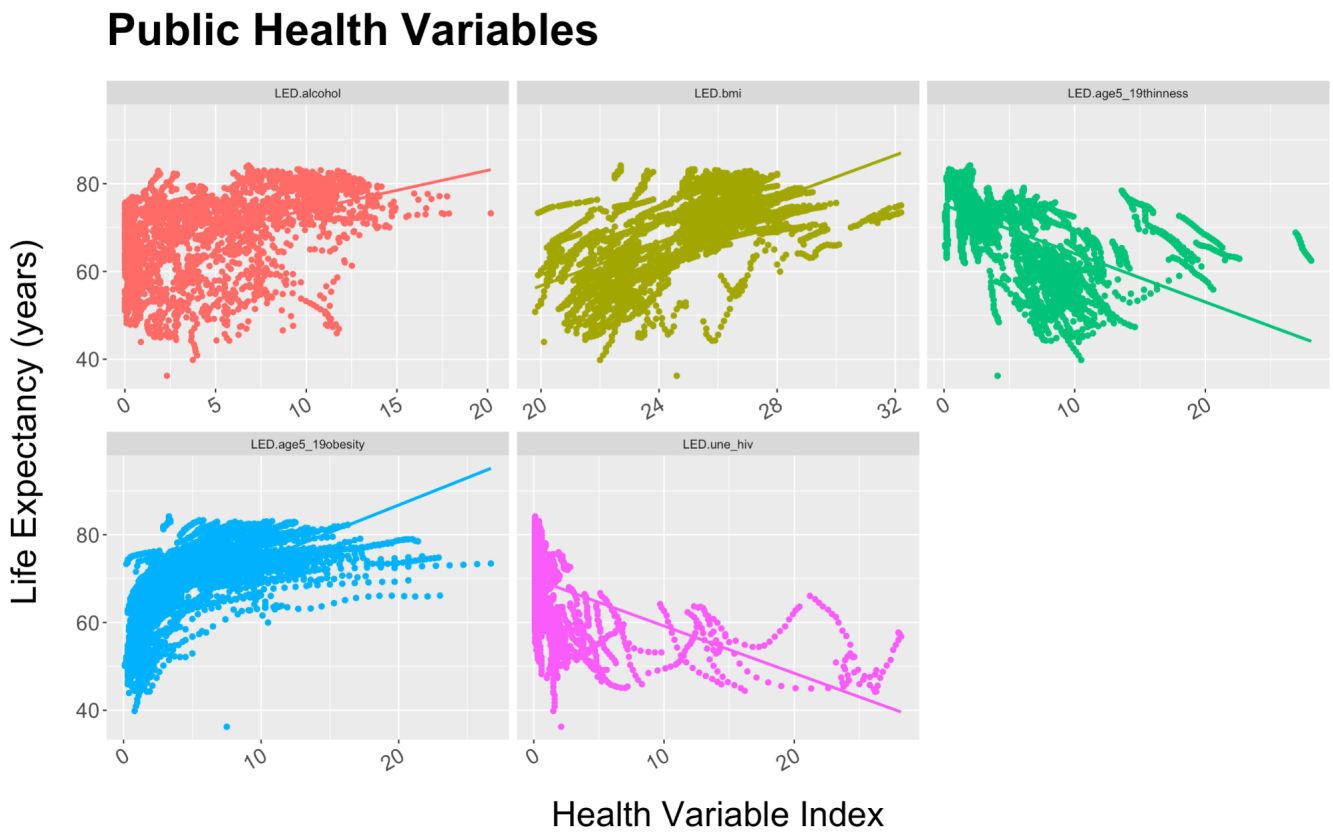
Mortality regressors:

The variables, from left to right, are infant mortality, age 1-4 mortality, and adult mortality. The index is the rate of mortality in percentage. Life expectancy is partially calculated by each of these variables, so it was expected to see a strong correlation with each.



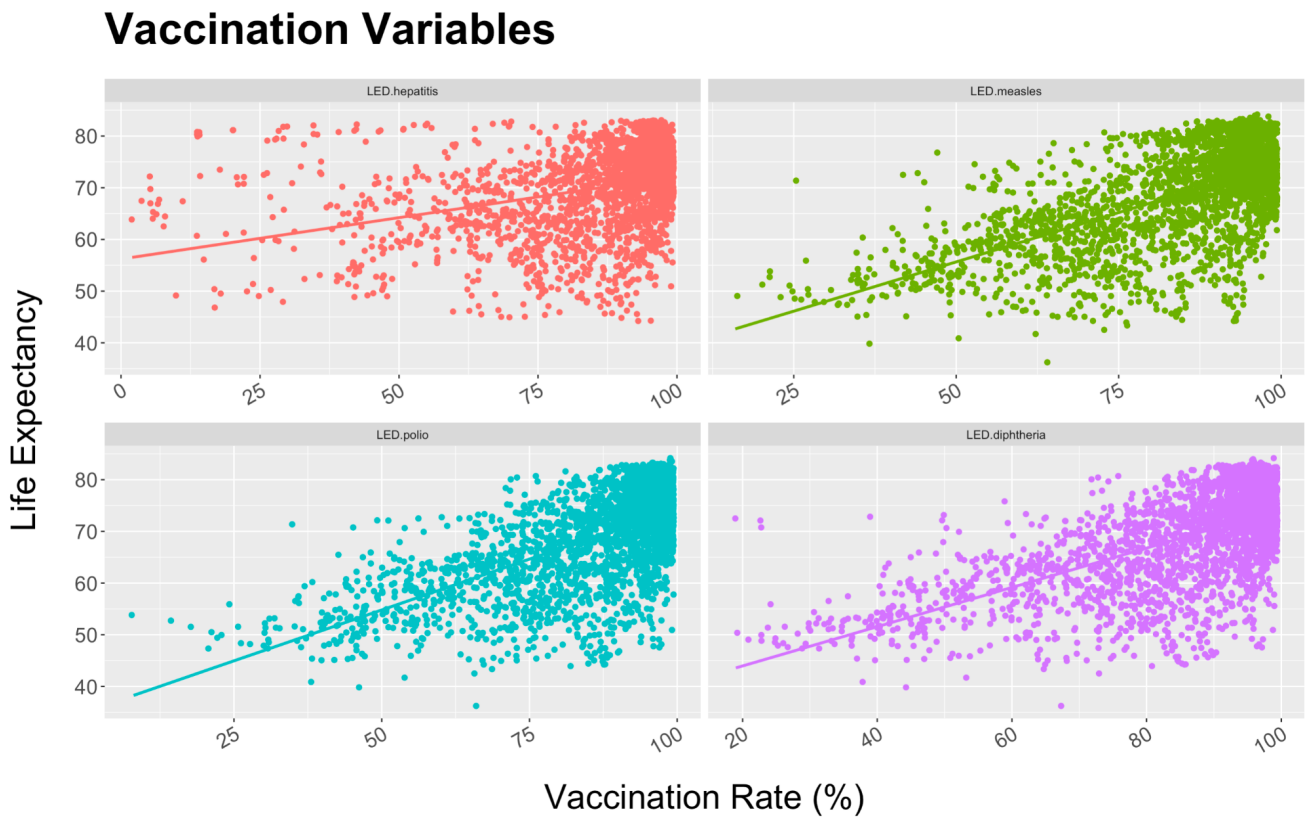
Public health regressors:

The five public health variables in our data set are alcohol consumption (top left), average BMI (top middle), age 5-19 thinness (top right), and age 5-19 obesity (bottom left), and prevalence of HIV (bottom middle). The distribution of the points varies. Alcohol and obesity both seem to have a slight logarithmic shape to it, although alcohol has many points that deviate from that shape. We can see that thinness has a negative relationship with life expectancy, but obesity and BMI has a positive relationship. Unexpectedly, alcohol consumption has a positive relationship. HIV seems to have a negative logarithmic shape. Further analysis is required to determine the true nature of each of these regressors and their relationship with life expectancy.



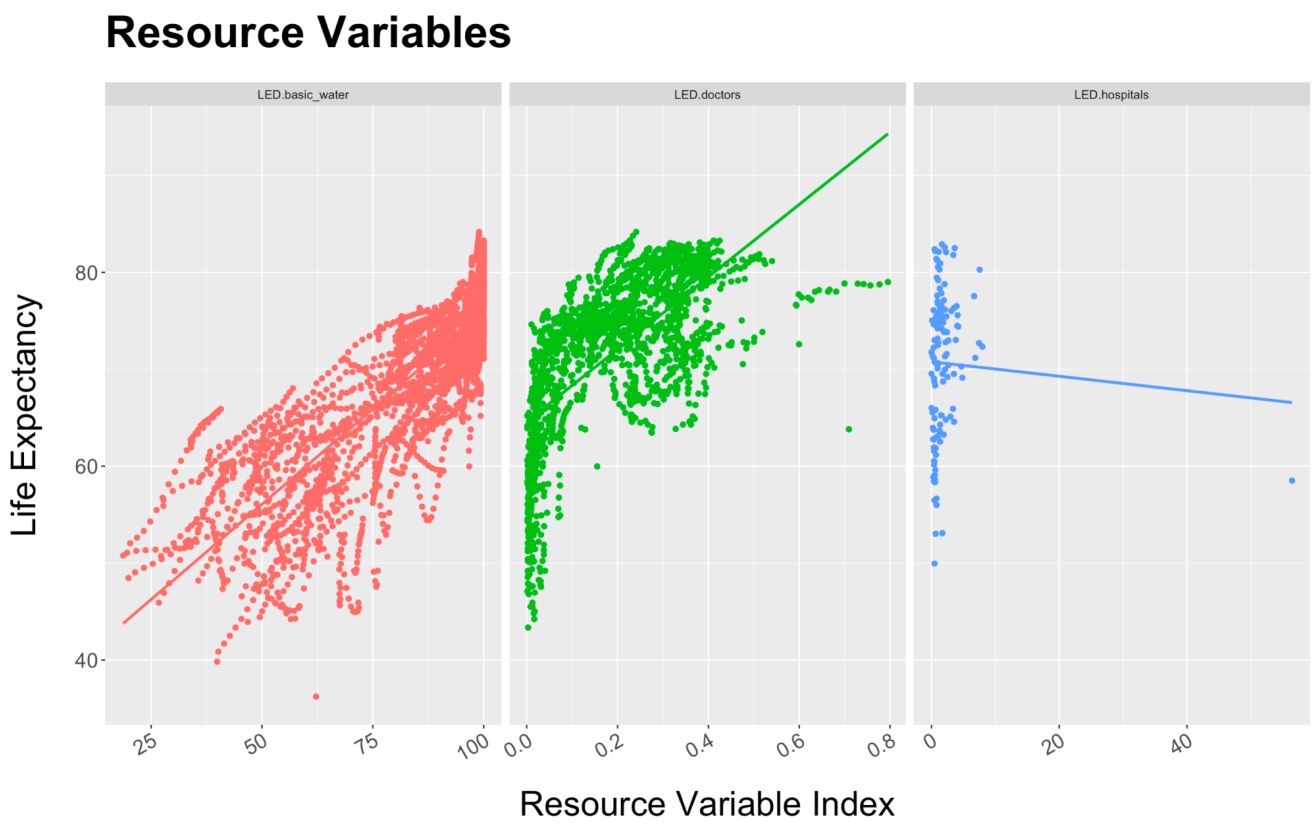
Vaccination regressors:

The four vaccination variables are Hepatitis B (top left), Measles (top right), Polio (bottom left), and Diphtheria (bottom right). Their scales are in percent coverage. We notice that Measles, Polio, and Diphtheria seem to look very much alike in their distribution, making it likely that they are highly correlated. Hepatitis B vaccinations seem to deviate a bit with a lot of points showing higher life expectancies with lower vaccination rates. In general, all 4 plots show increasing variance in distribution as vaccination rates increase.



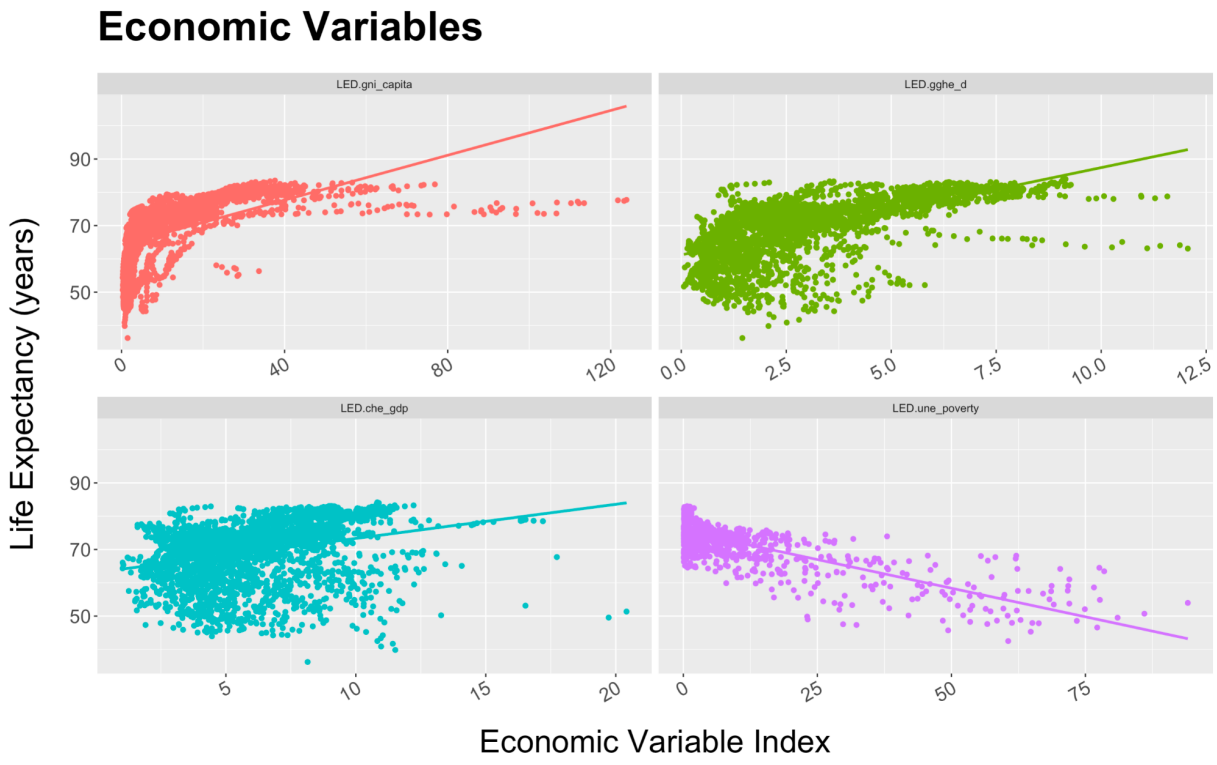
Resource regressors:

Our three variables are basic water access (left), prevalence of doctors (middle), and density of hospitals (right). It is clear that while basic water access maintains a lot of its samples, doctor prevalence and especially hospital density have much data missing. Basic water access and doctor prevalence seem positively correlated with life expectancy, meanwhile the relationship is unidentifiable for hospital density.



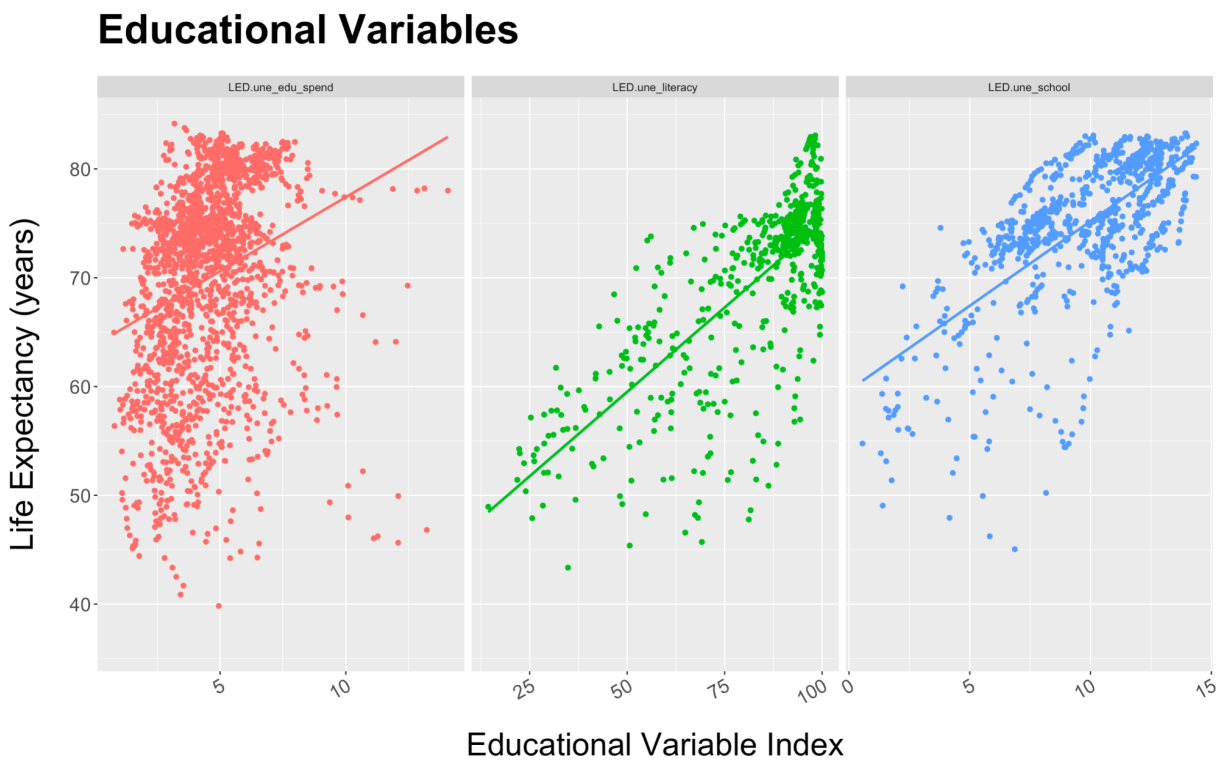
Economic regressors:

Our 4 economic variables are GNI per capita (top left), Domestic General Government Health Expenditure (GGHE-D) (top right), Current Health Expenditure (CHE) (bottom left), and poverty rates (bottom right). GNI per capita, GGHE-D, and CHE all have positive relationships but with slightly different distributions. GNI per capita has a clean and clear logarithmic shape to it while GGHE-D and CHE follow the logarithmic shape more loosely (just like our variable alcohol from public health). It's not yet clear if a relationship exists for those two variables. Poverty seems to be directly and negatively correlated with life expectancy, although it is important to note there are a lot of data points missing for that variable.



Educational regressors:

Our three educational variables are government spending on education (left), literacy rates (middle), and mean years of schooling for adults (right). Educational spending seems to be distributed in a vertical line, showing little to no relationship with life expectancy. Although literacy rates and mean years of schooling show a positive relationship, there are plenty of data points missing (around 80% of data missing).



Data Analysis

Data Cleaning

Fortunately, there was not a lot of cleaning needed for our group to complete our analysis. The first step of our cleaning process was to change the variable names to a uniform nomenclature, remove any extra characters, and make sure the names were properly descriptive. This included things like changing separators from "." to "_", and renaming life.exp60 to life_expect_at_60.

Lastly, we scaled some variables so that their values were more consistent and easier to work with. Not all variables were scaled based on the judgement of whether we were going to use them for analysis or not. Variables that deal with rates were scaled to represent a percent value (if not already). Other variables like GNI per capita were scaled to represent a thousand for each unit. The following scaling occured:

- Infant mortality * 100 (from decimal to percent)
- Age 1-4 mortality * 100 (from decimal to percent)
- Adult mortality / 10 (from 1:1,000 to percent)
- Doctors / 100 (from 1:10,000 to percent)
- GNI per capita / 1000 (scaled down by 1000 for 1:$1000)

This allowed for more feasible interpretation and graphing of these variables.

Removing Faulty Variables

In order to find the variables that would be helpful to use in the model, we wanted to remove variables that were likely to be unhelpful based on reasons not used in the general variable selection process. In general, we wanted to include as many continuous variables as we could and leave out indicator variables. This was because when cleaning and creating the model, we had not yet learned the use of indicator variables in class and wanted to keep our model as accurate as possible with the instruction we had been given. The filtering results are as follows:

Region: we decided not to include the *region* variable since our model was focused on providing advice to any country, given indicators, and not based on location or region. However, future studies should include this variable as it may make models more accurate or use subsets of data based on *region*.

Mortality: All mortality variables were left out based on the given advice from the professor that life expectancy is partially calculated from mortality and their existence would dominate the model.

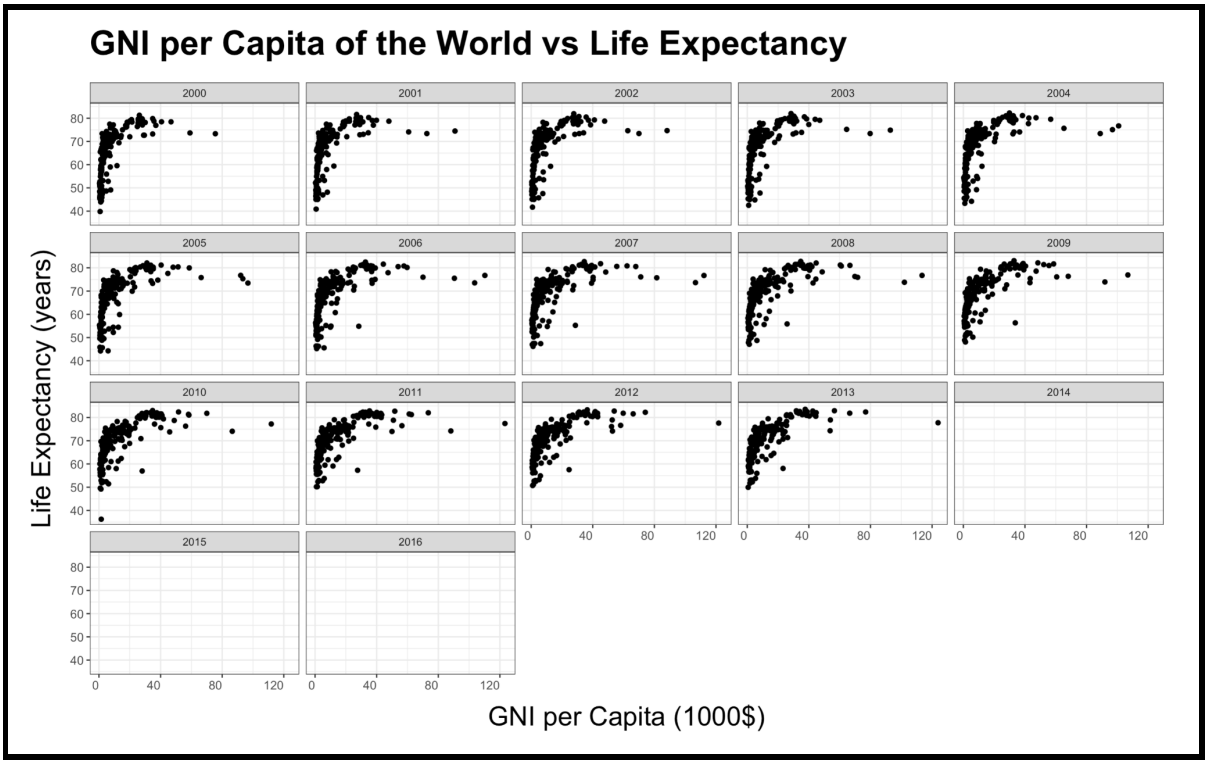
Other life expectancy variables: We did not include the variables *life_exp_at_60* and *une_life_expectancy* since they have either the same definition or are too closely related to the response variable: *life_expectancy*. It is important to note that these variables could be helpful response variables in another model.

Missing data: The variables *doctors*, *hospitals*, *une_hiv*, une_*poverty*, *une_edu_spend*, *literacy*, and *une_school* had too many missing data points and needed to be removed from the model (all had more than 20% of their data missing, HIV being the lowest at 23% and hospitals the highest at 100%). Most of these variables did not benefit from popular missing value strategies such as calculating the mean and filling in missing values.

Similar variables: *gni_capita* (GHO variable) and *une_gni* (UNESCO variable) both tell the GNI per capita of a sample. Since *gni_capita* had less missing data and had similar values to *une_gni,* we chose to remove *une_gni.*

Influential Decisions

In order to create a model with the least amount of missing data and the most impactful baseline result, we chose to only include the data from the year 2012. The major benefits of this decision boil down to eliminating possible autocorrelation, missing data, and maintaining recent & relevant information. We know that a country's life expectancy and other statistics can be dependent on the previous measured years. This can create imbalances in the model, so it is helpful to only choose one year.



An example of a variable with missing data past 2012 is *gni_capita*. GNI per capita no longer has any data available after the year 2013 (and there are many data points missing in the year 2013 as well). Since the year 2012 is the last year where GNI per capita was readily available, ***we chose to base our analysis on the year 2012 alone***. GNI per capita (as shown later) is also an important predictor, so it would be wasteful to choose a year (or include years) where it has data missing. Years previous to 2012 (2000-2011) could have been used as well, but we wanted to use the most recent data to maintain relevance.

However, these decisions are not without their negatives. By only choosing to analyze only the year 2012, we reduce the number of samples from 3,111 to 183. Although 183 is plenty of data points for all our statistical techniques, we are in fact only using a small portion of our data. It is important to point out that due to an overall lesser amount of data points, it is possible that some relationships could be missed or altered. We also cannot use the *year* variable as a regressor, which could show a meaningful relationship over time. Thus, the model is not perfect. These are things that should be heavily considered in our future works and any future analysis done on this data set.

Variable Selection

In order to select our variables, we decided to use a more rigorous process, but similar to stepwise variable selection. We used four major steps and 1 precursor step. The precursor step is to eliminate variables for reasons that do not apply to our major steps. This was already described in the "Removing Faulty Variables" section, so here is the summary:

- Infant mortality (from both GHO and UNESCO), age 1-4 mortality, and adult mortality will not be used.
- GNI per capita from UNESCO (une_gni) will not be used.
- Region is not used since it is an indicator variable.
- Doctor prevalence, hospital density, HIV prevalence, poverty rates, educational spending, literacy rates, and average schooling will not be used due to too much missing data.

As for our four major steps, they are as follows:

- Univariate Analysis
- Removal of similar variables
- Elimination of multicollinearity
- Testing Significance of Regression

These four steps are used every single time we choose to create a linear model, whether it is our first model or the model after performing transformations and outlier removal.

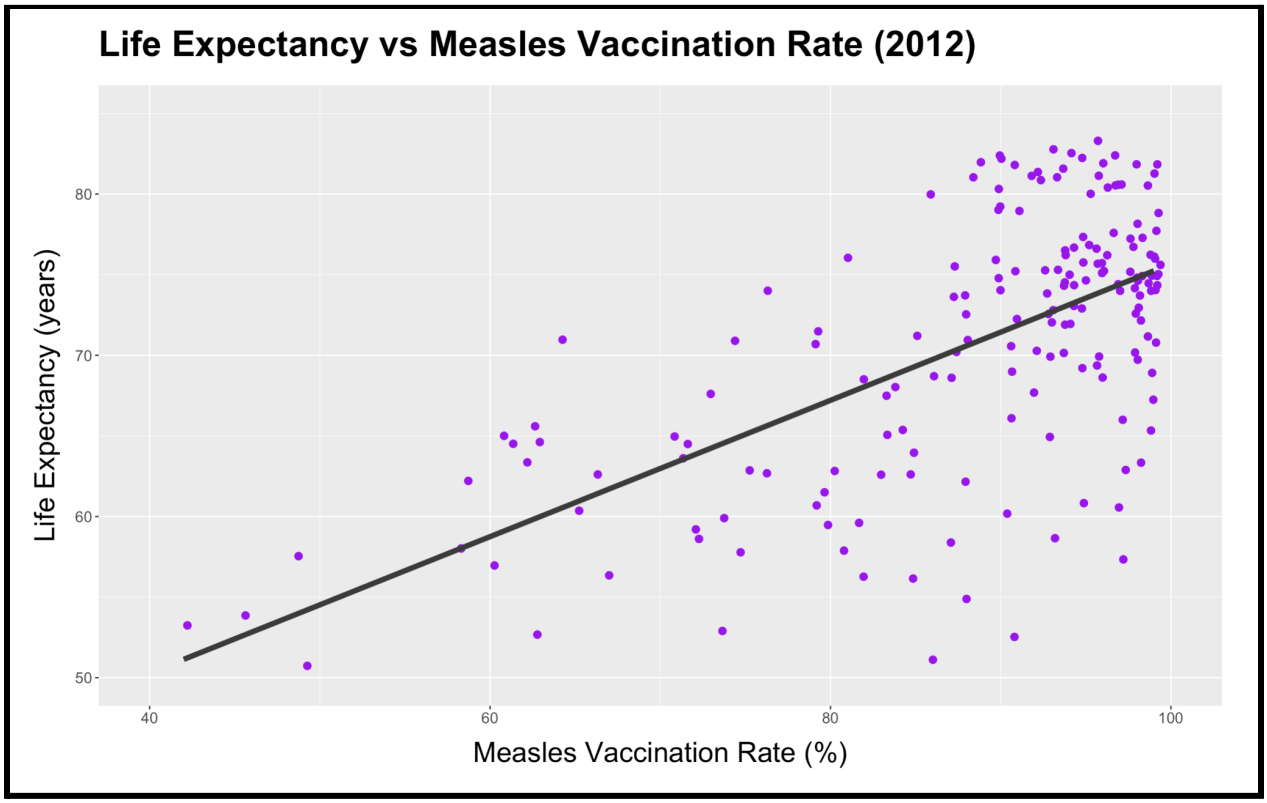
Univariate Analysis

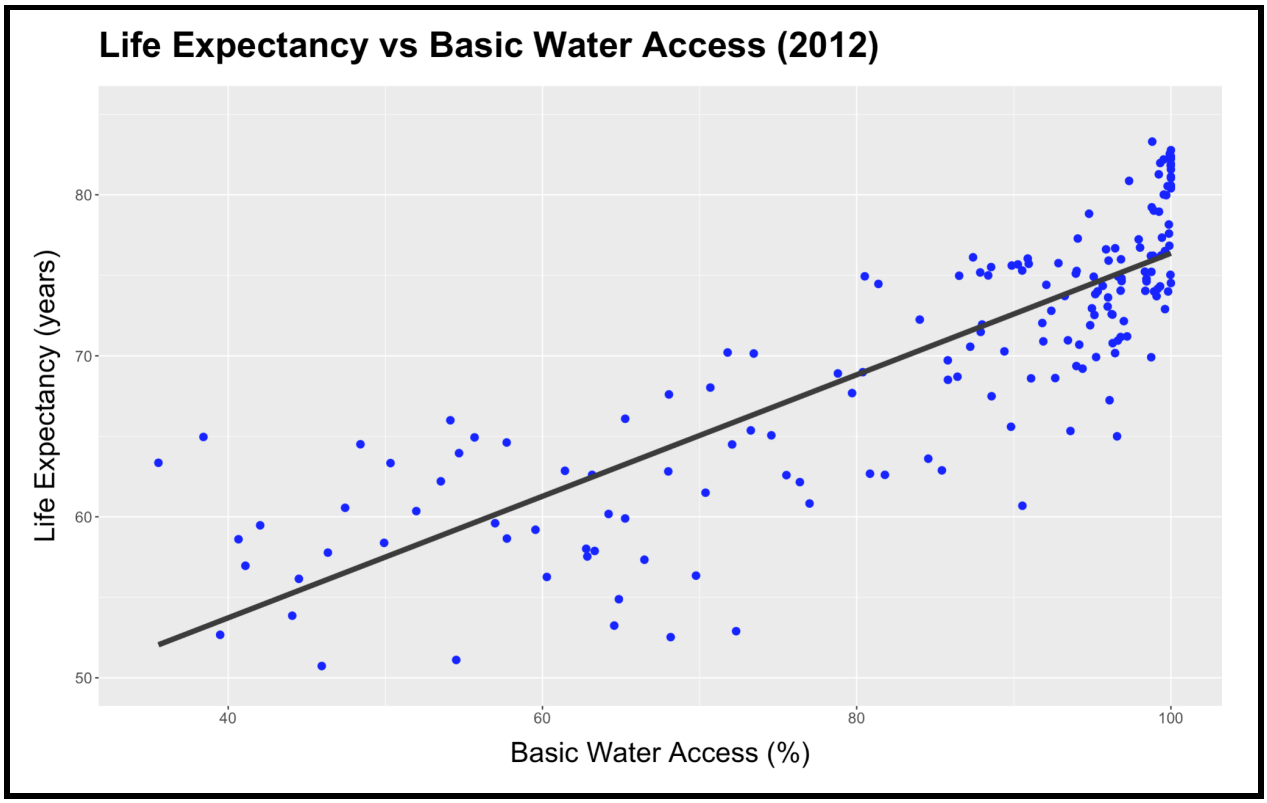
In this step, we took any of the remaining regressors open for analysis and tested them in a simple linear regression model against life expectancy. The following list shows which variables were tested and what their significance of regression results were:

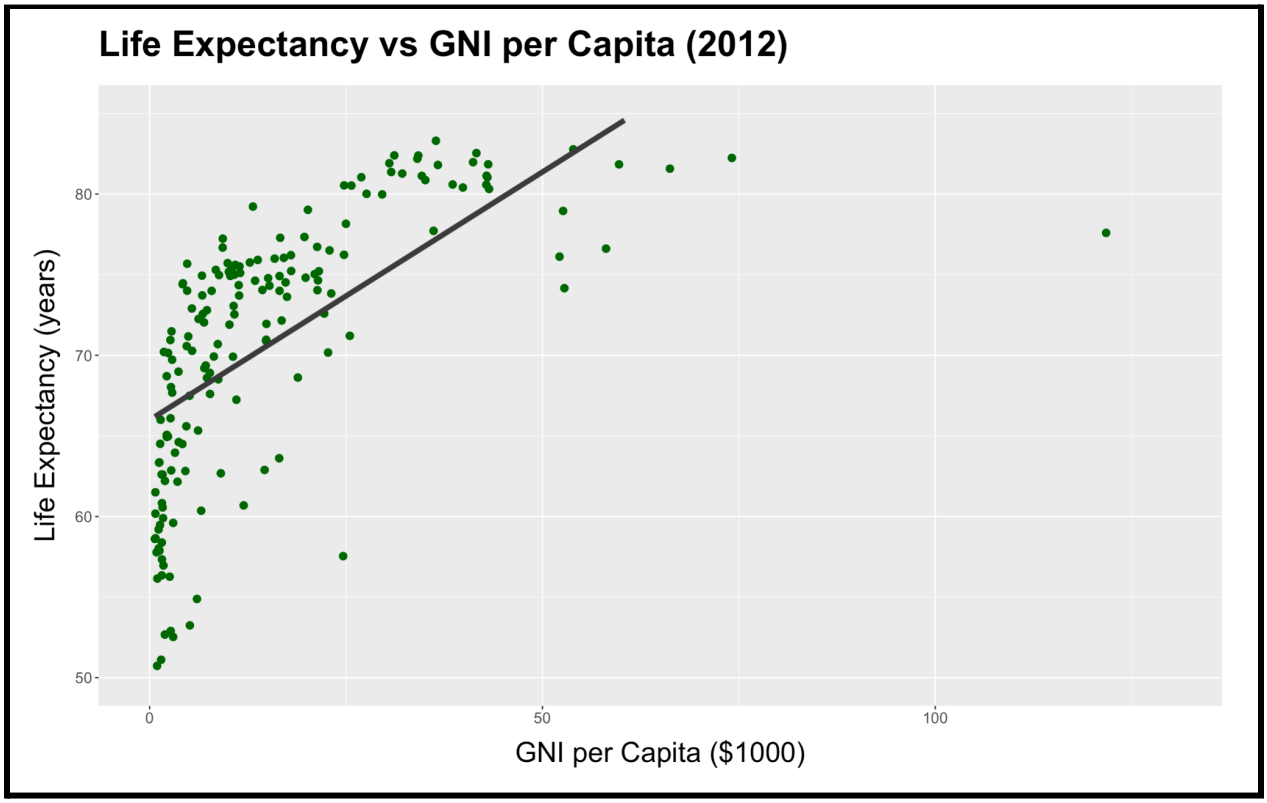
- Measles vaccination(measles), PASSES TEST
- Basic water access (basic_water), PASSES TEST
- GNI per capita (gni_capita), PASSES TEST
- GGHE-D (gghe_d), PASSES TEST
- Alcohol consumption (alcohol), PASSES TEST
- Average BMI (bmi), PASSES TEST
- Age 5-19 thinness (age5_19thinness), PASSES TEST
- Age 5-19 obesity (age5_19obesity), PASSES TEST
- Hepatitis B vaccination (hepatitis), PASSES TEST
- Polio vaccination (polio), PASSES TEST

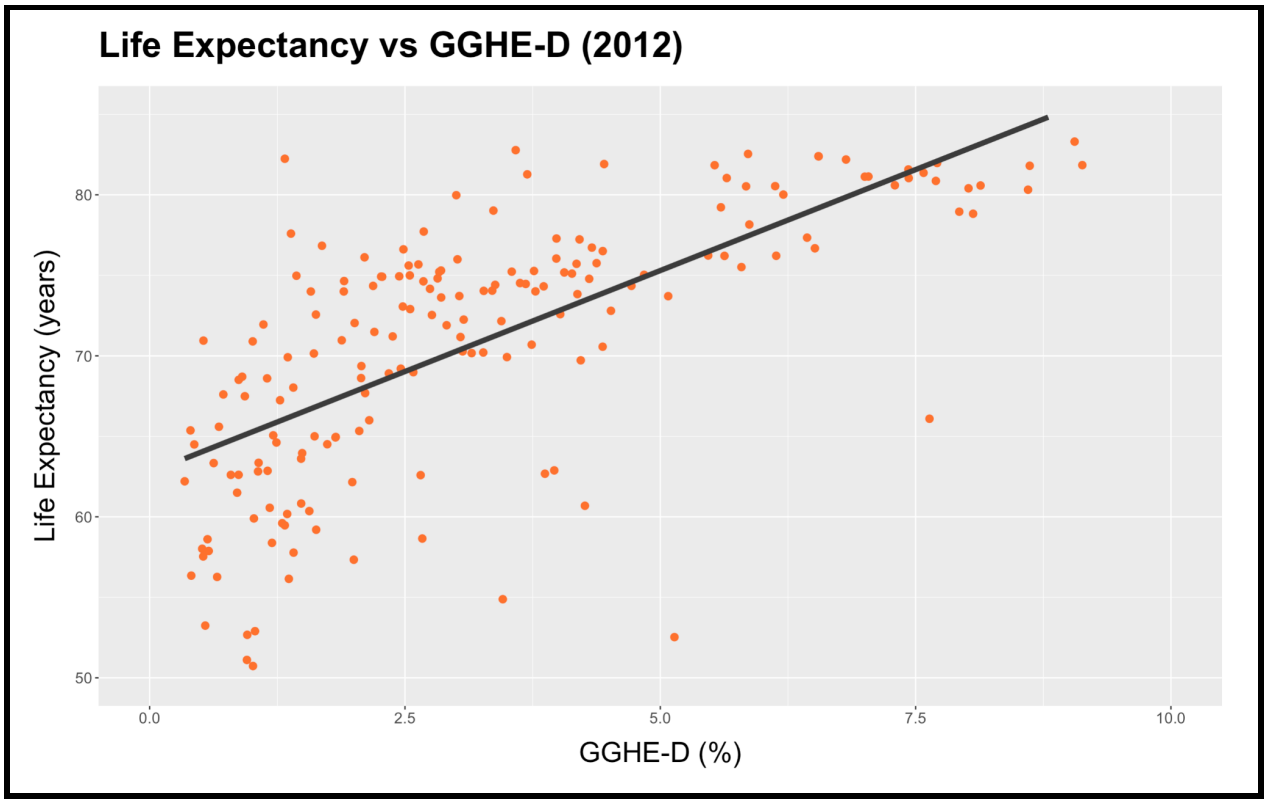
- Diphtheria vaccination (diphtheria), PASSES TEST
- Current Health Expenditure (che_gdp), PASSES TEST
- Population in thousands (une_pop), FAILS TEST

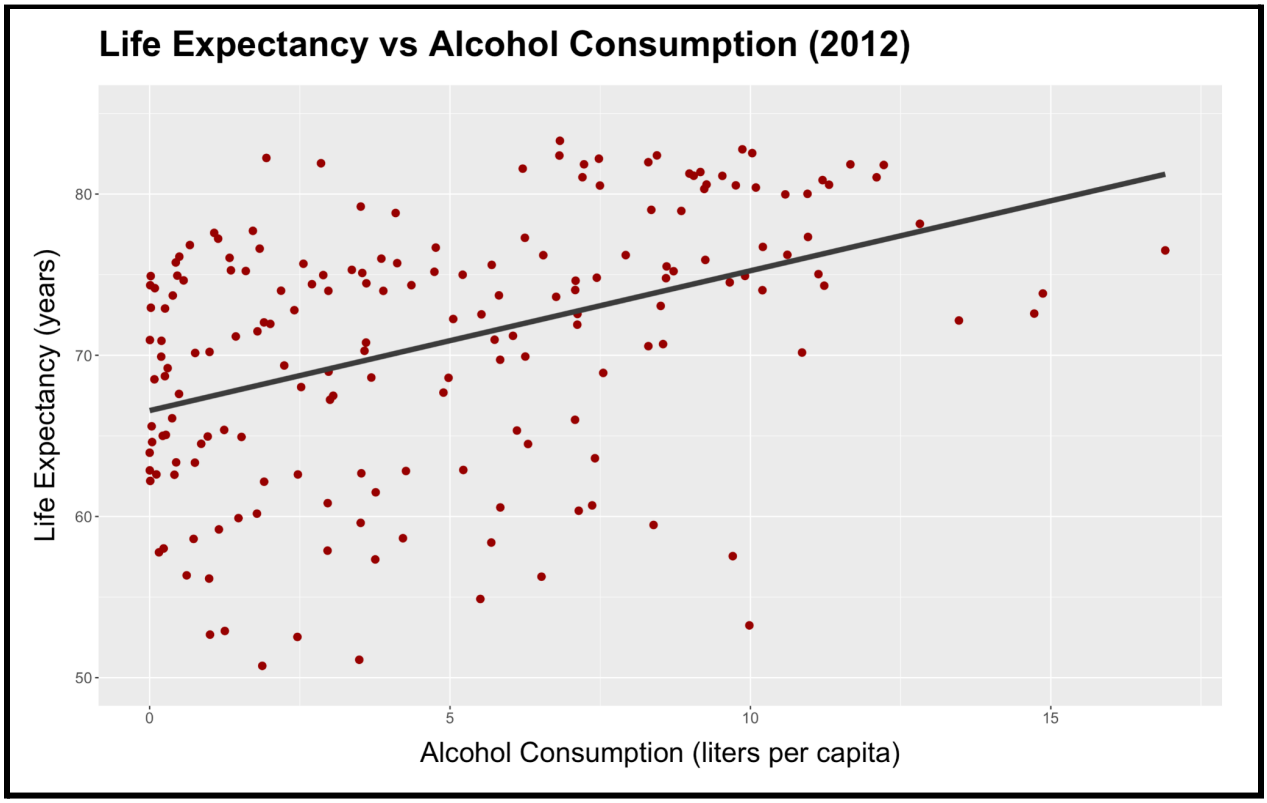
The plots for each of the variables is shown below. A red line indicates that the variable failed the significance of regression test.

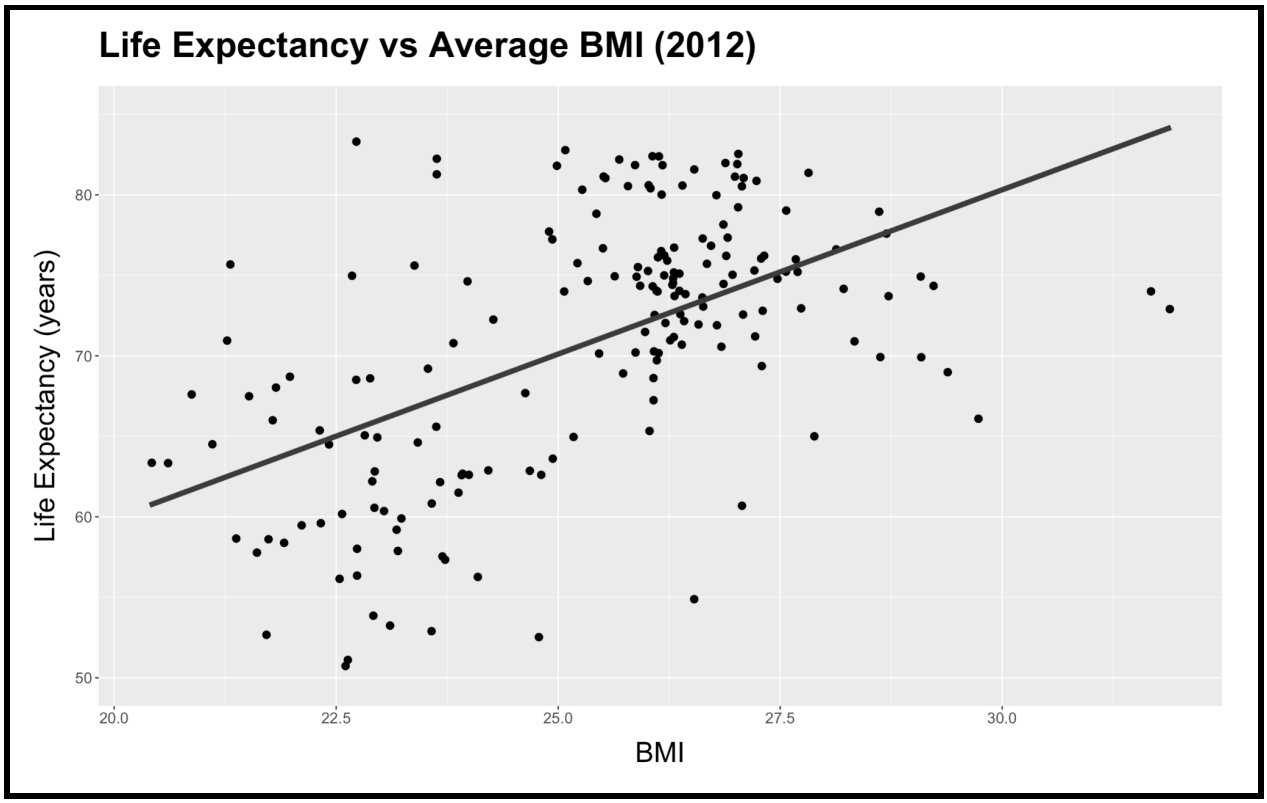


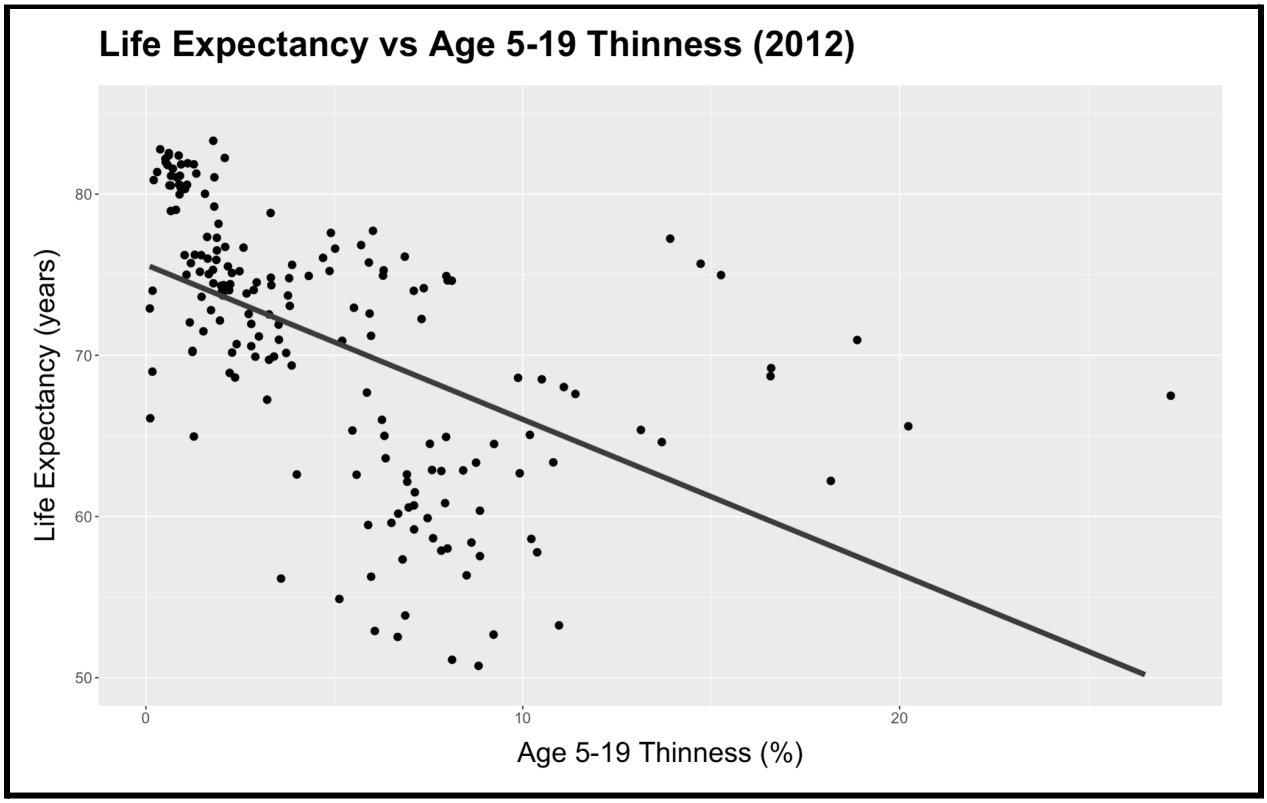

Life Expectancy vs Basic Water Access (2012)
Life Expectancy (years)
80
70
60
50
40
60
80
100
Basic Water Access (%)

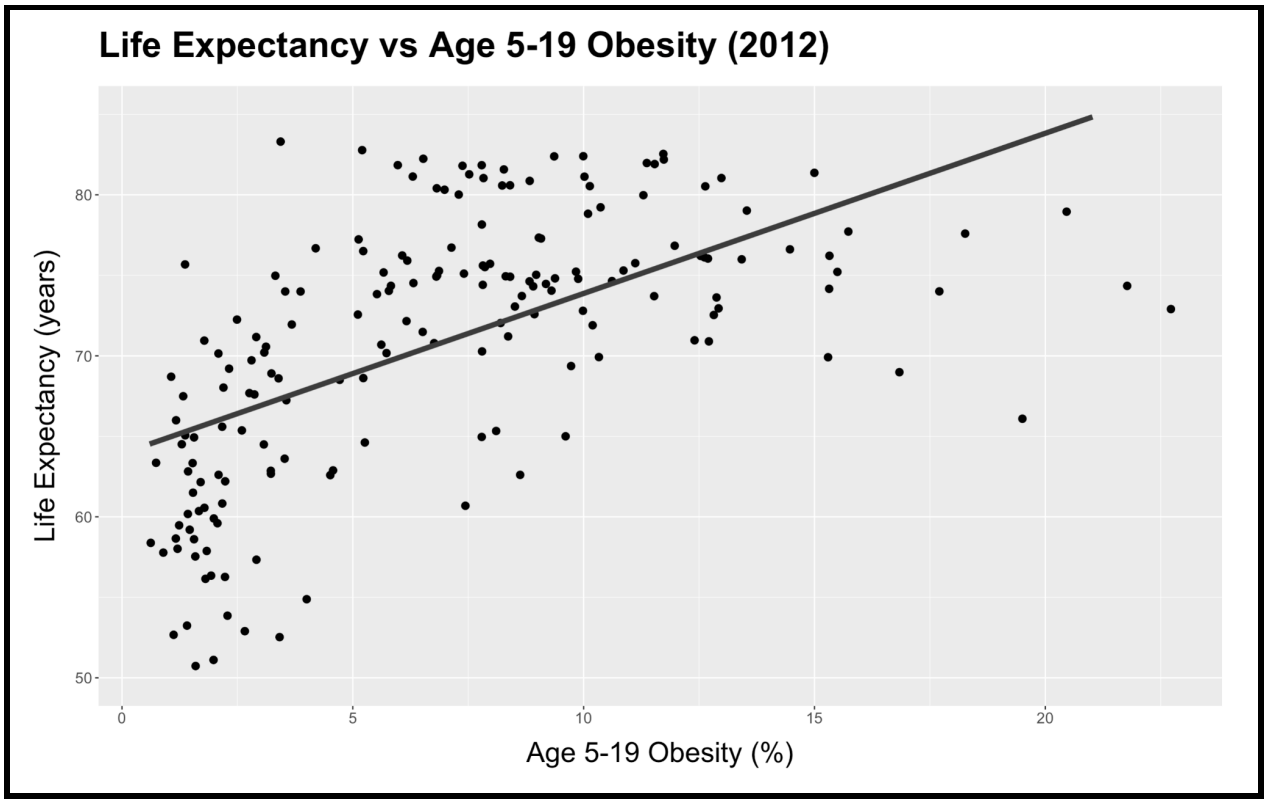

Life Expectancy vs GNI per Capita (2012)
Life Expectancy (years)
80
70
60
50
0
50
100
GNI per Capita ($1000)

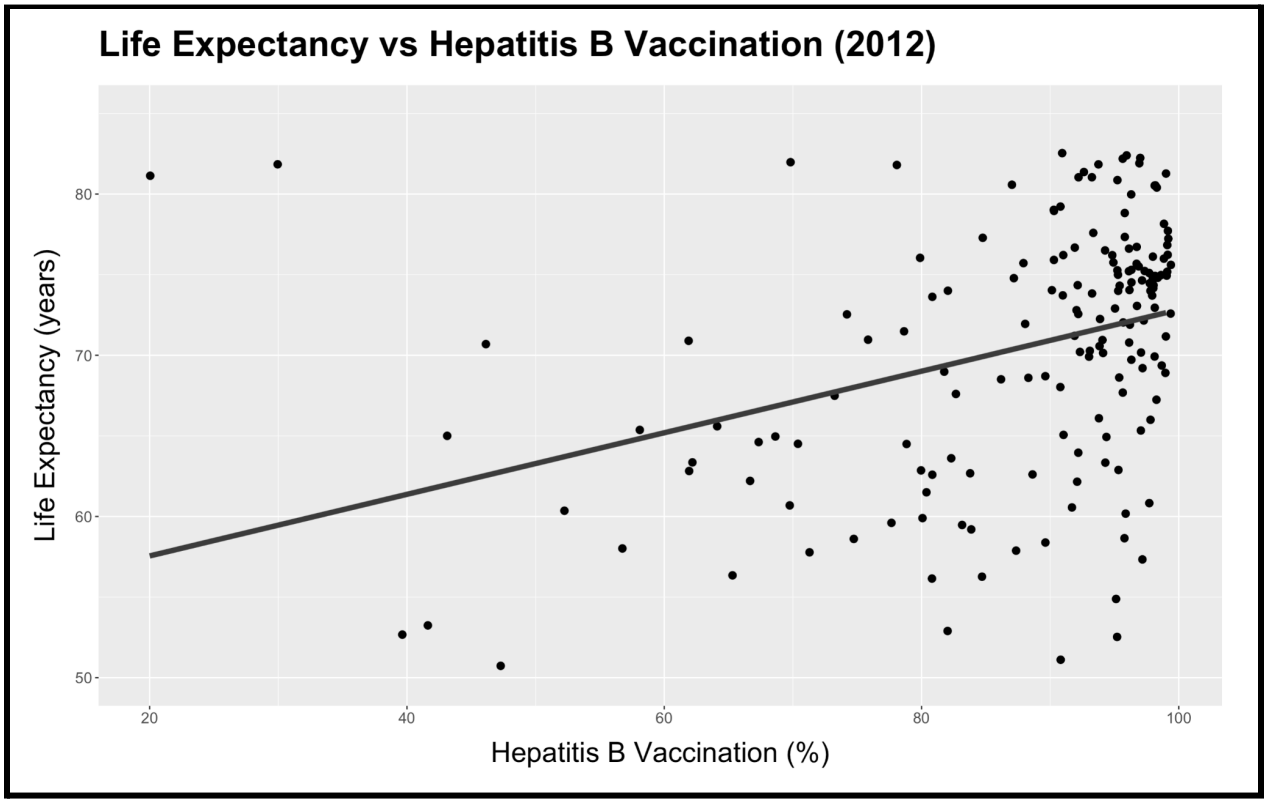

Life Expectancy vs GGHE-D (2012)
Life Expectancy (years)
80
70
60
50
0.0
2.5
5.0
7.5
10.0
GGHE-D (%)

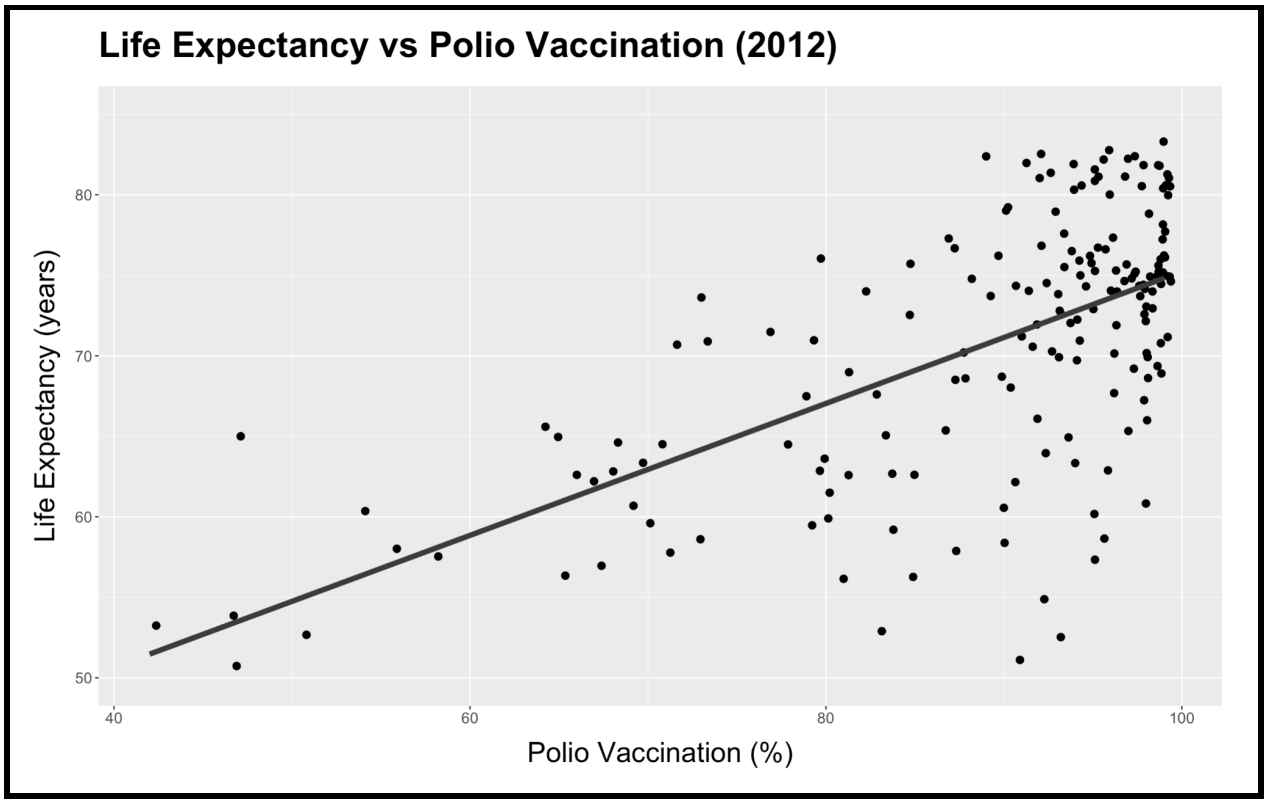

Life Expectancy vs Alcohol Consumption (2012)
Life Expectancy (years)
80
70
60
50
0
5
10
15
Alcohol Consumption (liters per capita)

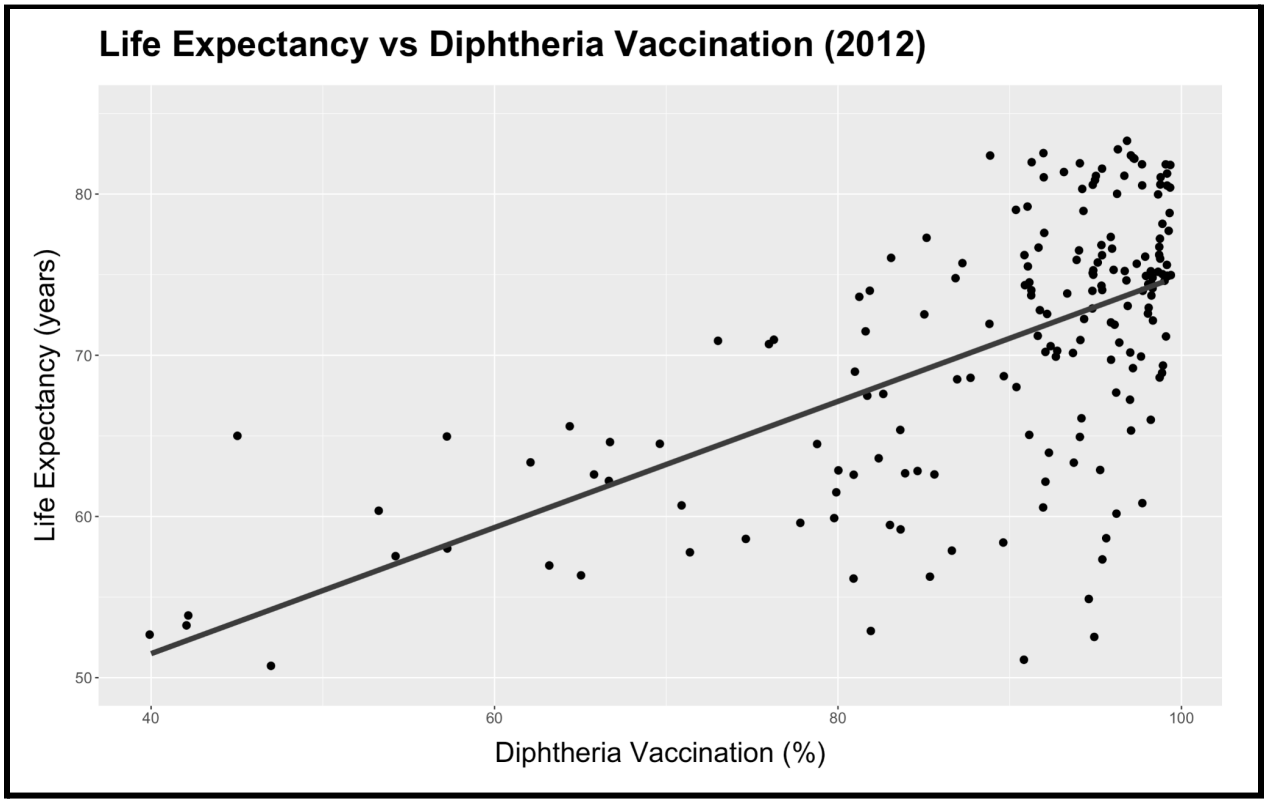

Life Expectancy vs Average BMI (2012)
Life Expectancy (years)
80
70
60
50
20.0
22.5
25.0
27.5
30.0
BMI

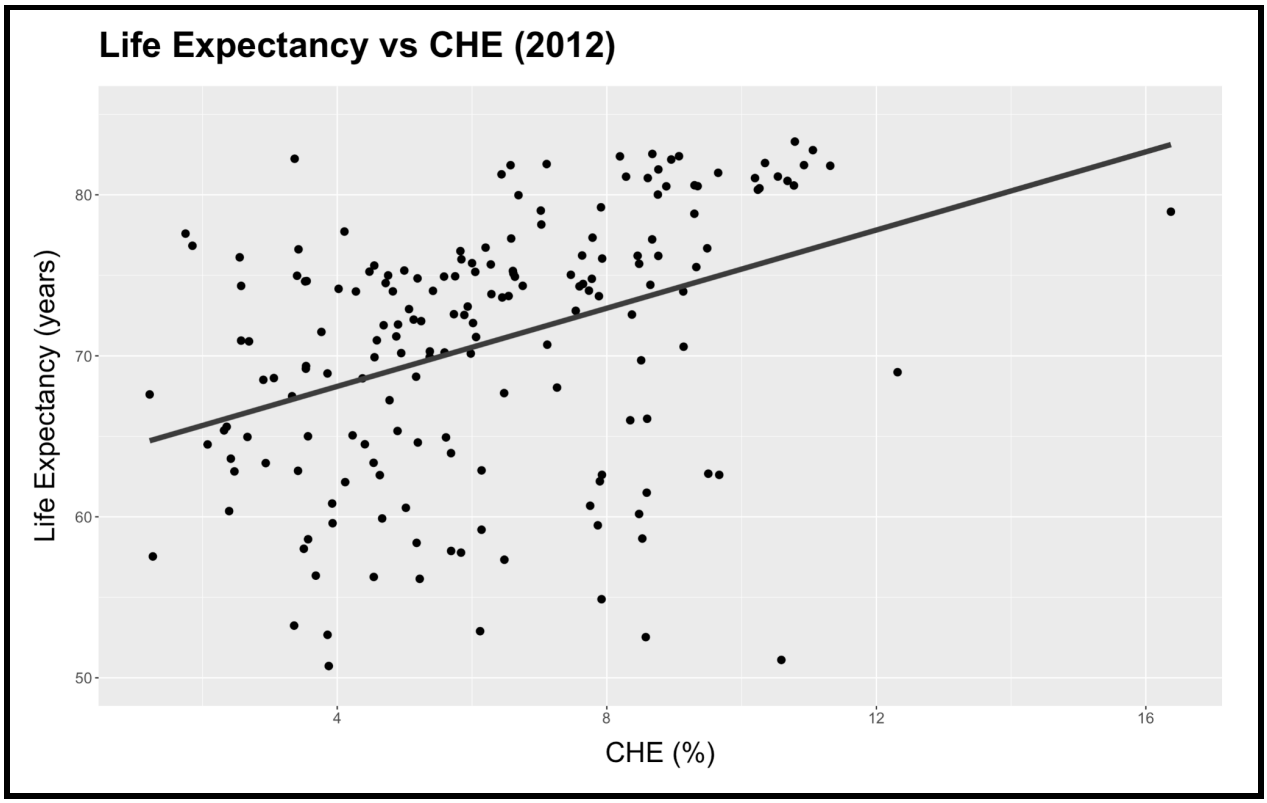

Life Expectancy vs Age 5-19 Thinness (2012)
Life Expectancy (years)
80
70
60
50
0
10
20
Age 5-19 Thinness (%)

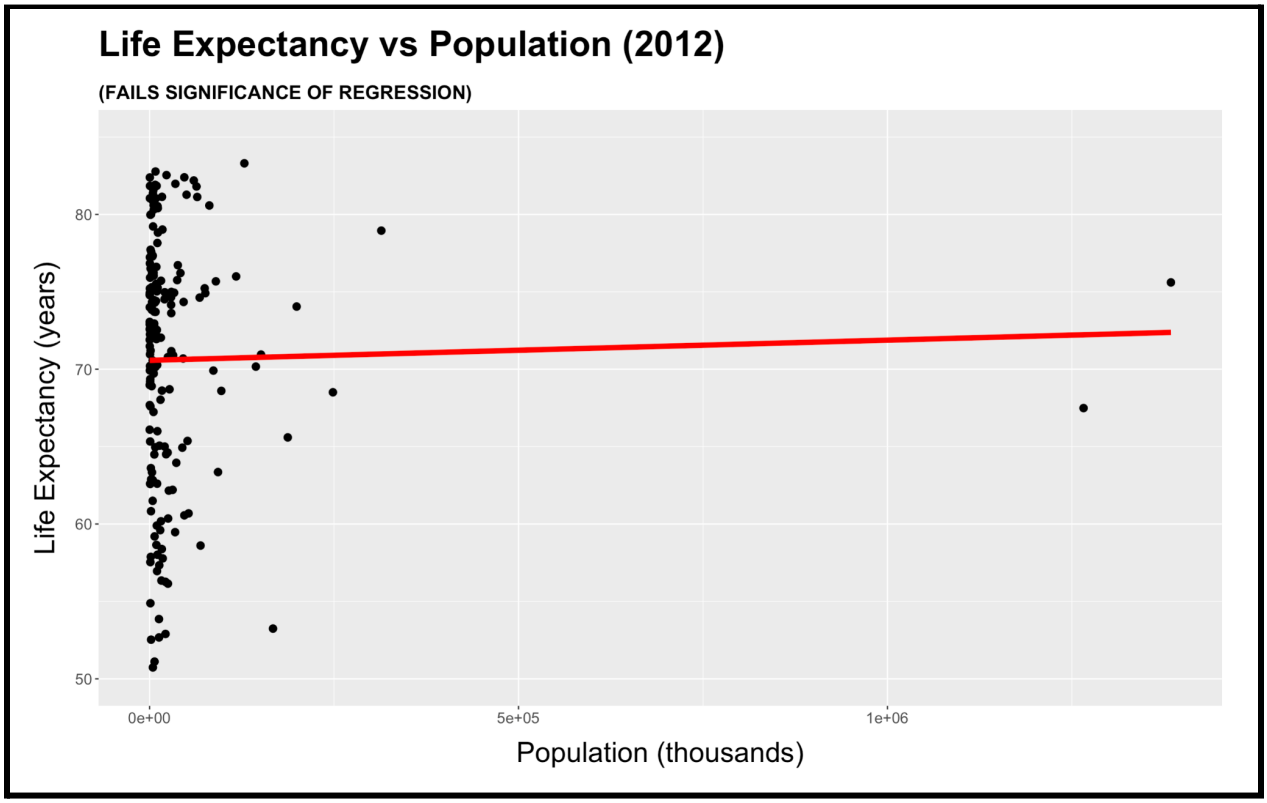

Life Expectancy vs Age 5-19 Obesity (2012)
Life Expectancy (years)
80
70
60
50
0
5
10
15
20
Age 5-19 Obesity (%)


Life Expectancy vs Hepatitis B Vaccination (2012)
Life Expectancy (years)
80
70
60
50
20
40
60
80
100
Hepatitis B Vaccination (%)


Life Expectancy vs Polio Vaccination (2012)
Life Expectancy (years)
80
70
60
50
40
60
80
100
Polio Vaccination (%)


Life Expectancy vs Diphtheria Vaccination (2012)
Life Expectancy (years)
80
70
60
50
40
60
80
100
Diphtheria Vaccination (%)


Life Expectancy vs CHE (2012)
Life Expectancy (years)
80
70
60
50
4
8
12
16
CHE (%)


Life Expectancy vs Population (2012)
(FAILS SIGNIFICANCE OF REGRESSION)
Life Expectancy (years)
80
70
60
50
0e+00
5e+05
1e+06
Population (thousands)

Removal of Similar Variables

This was a non-statistical method that required knowledge about the variables. Among the variables that were tested to be significant, we did not include variables into the multiple regression model if they had definitions similar to or highly related to other variables. For example, we did not include *bmi* because the average BMI is influenced by the proportion of people who are thin or obese. This is already being represented by the *age5_19thinness* and *age5-19obesity* variables. Thus, we removed BMI rather than those two variables since *bmi* is less informative than they are. *che_gdp* was also removed because it is very closely related to *gghe_d* in definition and calculation, so only one of them would suffice in the model. We chose to maintain gghe_d instead of che_gdp based on better results from the univariate testing.

Elimination of Multicollinearity

At this point, we fit the remaining variables into a multiple regression model. Now whether a variable passes the significance of regression test is dependent on whether it is being influenced by multicollinearity. We sought out the variables that induced high amounts of multicollinearity and dealt with them. Those variables were *polio* and *diphtheria*. They both had VIF values larger than 24. Removal of one or the other reduced their values to around 8 or so. However, *measles* would still remain at around the value of 7 unless both were removed. Despite these values being less than 10, they are still quite near it. Also, if a country vaccinates for one disease, they likely vaccinate for others as well, explaining why the vaccination variables might be so highly collinear. Thus, it would suffice to only maintain one or two of them to prevent high VIF values. Once both *polio* and *diphtheria* were removed, *measles* and *hepatitis* would have VIF values of around 3 and 2 respectively. Now all the variables have relatively low VIF values and we can move on to the next step.

Before removal of *polio* and *diphtheria*

alcohol	age5_19thinness	age5_19obesity	hepatitis	measles
1.832833	2.139592	2.544653	3.103809	7.097183
polio	diphtheria	basic_water	gni_capita	gghe_d
24.732528	24.954528	2.730303	1.729370	2.358398

After removal

```
      alcohol age5_19thinness  age5_19obesity       hepatitis         measles
     1.820923        2.134455        2.464529        2.324233        3.239812
  basic_water      gni_capita           gghe_d
     2.605736        1.700044        2.263410
```

Testing Significance of Regression

Among the remaining variables, we removed (one by one) the variables that had p-values greater than ~0.05 (we say around 0.05 because we will accept if the variable has a p-value of no more than 0.06 or so as long as it improves the model in other aspects such as residual standard error and R^2), starting with the largest ones. We removed *hepatitis* followed by *age5_19thinness*, *alcohol*, then finally *age5_19obesity*. The following linear model summary shows what the model looked like before removal of insignificant variables.

Before removal of insignificant variables:

```
Call:
lm(formula = life_expectancy ~ alcohol + age5_19thinness + age5_19obesity +
    hepatitis + measles + basic_water + gni_capita + gghe_d,
    data = LED12)

Residuals:
     Min       1Q   Median       3Q      Max
-15.0265  -1.5877   0.5165   2.2034   9.5794

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     40.217922   2.884500  13.943  < 2e-16 ***
alcohol         -0.084842   0.103367  -0.821 0.413074
age5_19thinness -0.046606   0.096376  -0.484 0.629389
age5_19obesity   0.109325   0.098678   1.108 0.269682
hepatitis        0.000754   0.031785   0.024 0.981107
measles          0.079530   0.047452   1.676 0.095817 .
basic_water      0.224348   0.027357   8.201 9.91e-14 ***
gni_capita       0.087535   0.023572   3.713 0.000288 ***
gghe_d           0.899901   0.220305   4.085 7.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.743 on 150 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared:  0.7861,    Adjusted R-squared:  0.7747
F-statistic: 68.89 on 8 and 150 DF,  p-value: < 2.2e-16
```

Model Fitting

The variables that remained significant after the variable selection process were *measles, basic_water, gni_capita,* and *gghe_d.* The following linear model summary shows the linear model after the removal of insignificant variables.

After removal of insignificant variables:

```
Call:
lm(formula = life_expectancy ~ measles + basic_water + gni_capita +
    gghe_d, data = LED12)

Residuals:
     Min       1Q   Median       3Q      Max
-15.1511  -1.7401   0.3308   2.2315   9.6551

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.94293    2.31691  16.808  < 2e-16 ***
measles      0.08779    0.03114   2.819  0.00541 **
basic_water  0.23119    0.02357   9.807  < 2e-16 ***
gni_capita   0.09579    0.02052   4.669 6.21e-06 ***
gghe_d       0.91540    0.16558   5.528 1.23e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.662 on 166 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.7991,    Adjusted R-squared:  0.7943
F-statistic: 165.1 on 4 and 166 DF,  p-value: < 2.2e-16
```

We can see that in the model, each variable has a positive relationship with life expectancy. Interpreting these coefficients shows:

- As the Measles-containing-vaccine first-dose (MCV1) immunization coverage among 1-year-olds increases by 1%, life expectancy increases by 0.08779 years on average while all other regressors are held constant.
- As the percentage of the population using basic drinking water services increases by 1%, life expectancy increases by 0.23119 years on average while all other regressors are held constant.
- As the GNI per capita of a country increases by $1000, life expectancy increases by 0.09579 years on average while all other regressors are held constant.

- As the GGHE-D of a country increases by 1%, life expectancy increases by 0.91540 years on average while all other regressors are held constant.

The R^2value for this model shows that 79.91% of the variance in the data is explained by the fitted model. The model also shows that the *measles* variable is less significant than the other variables included. *basic_water, gni_capita,* and *gghe_d* are all extremely significant to the model.

We also need to note that 12 countries were not used in the formation of the linear model due to missing data points. Those countries are the following:

```
 [1] "Cuba"                                  "Brunei Darussalam"
 [3] "Argentina"                             "Montenegro"
 [5] "Kuwait"                                "Libya"
 [7] "Democratic People's Republic of Korea" "Myanmar"
 [9] "Syrian Arab Republic"                  "Djibouti"
[11] "South Sudan"                           "Somalia"
```

Residual Analysis

Variance Inflation Factors

```
> vif(model)
    measles basic_water  gni_capita      gghe_d
   1.634502    2.183281    1.574373    1.651320
```

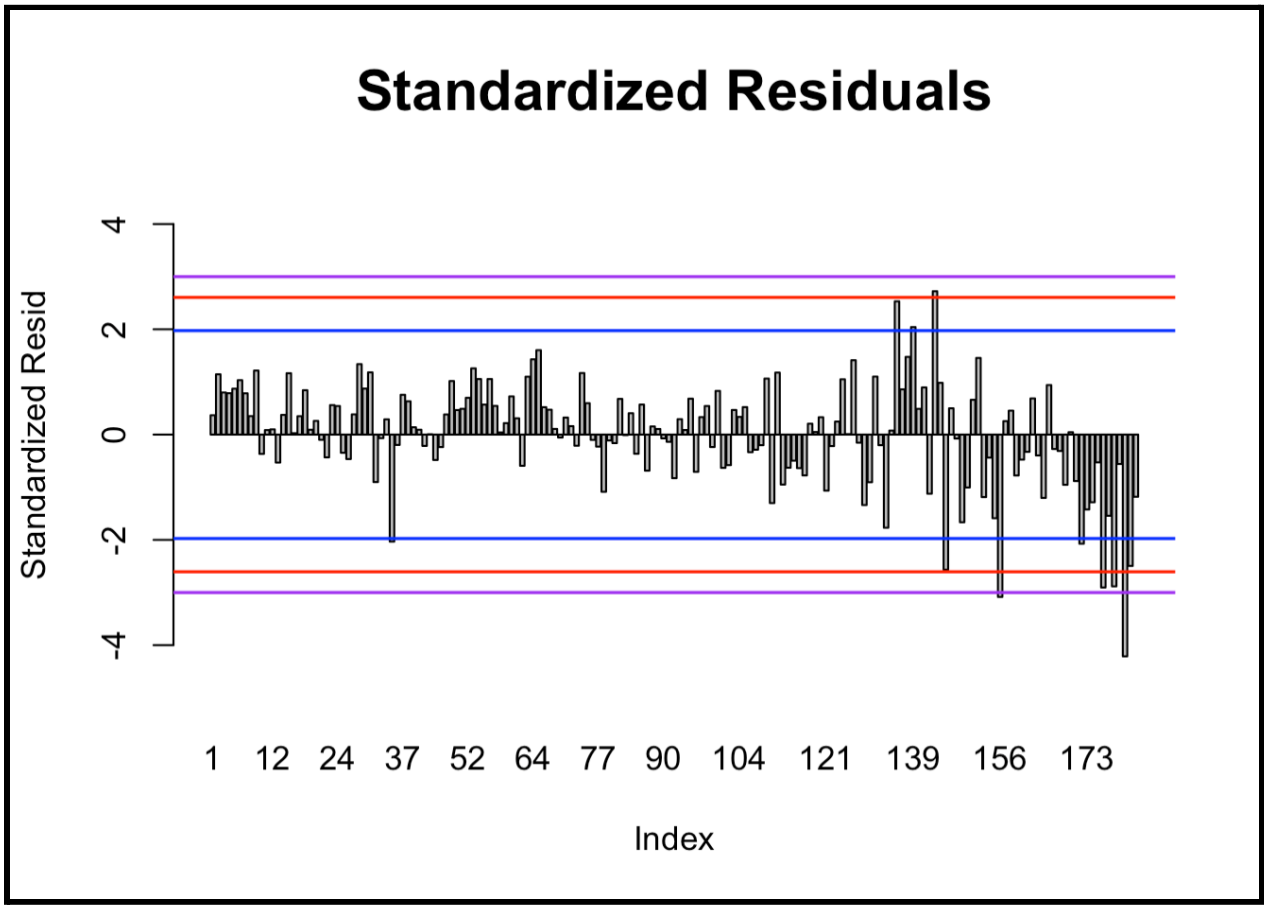
The VIFs are acceptably low for our variables. The highest VIF value is for *basic_water* and is only a little over 2. We can safely assume there is little to no multicollinearity in the model.

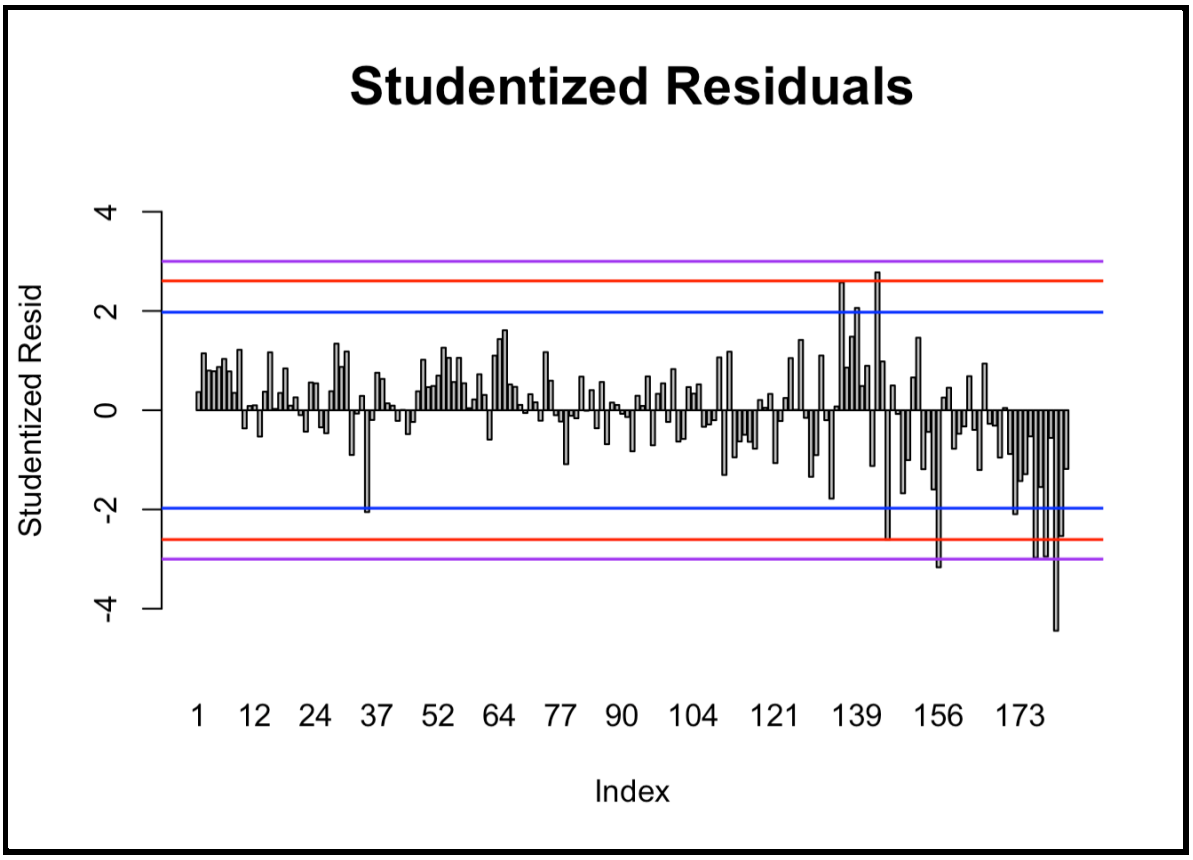
Residual Bar-plots & Summaries

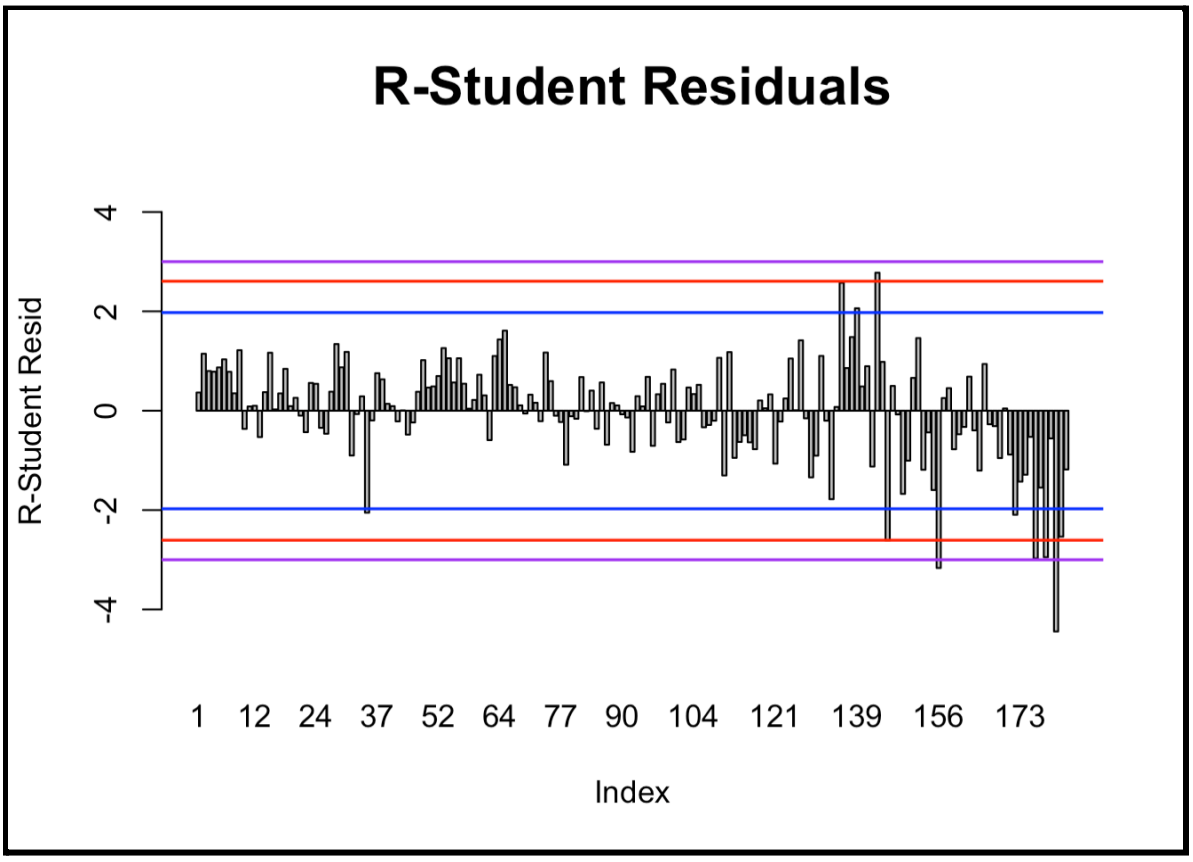
The following is a summary of the unscaled residuals:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-15.1511	-1.7401	0.3308	0.0000	2.2315	9.6551

We can see that the most deviant sample is one that has a life expectancy about 15 years less than what was predicted by the model. However, residuals are more informative when scaled. The following plots show the standardized, studentized, and R-student residuals and their summaries. The three bars on the plots indicate the three cutoff t-values. The blue bar represents the 95% cutoff value, which is around a t-value of ~2. The red bar is a 99% cutoff value, around ~2.6, and purple is a flat t-value of 3 (used to indicate major outliers).




Studentized Residuals
Studentized Resid
4
2
0
-2
-4
1
12
24
37
52
64
77
90
104
121
139
156
173
Index


R-Student Residuals
R-Student Resid
4
2
0
-2
-4
1
12
24
37
52
64
77
90
104
121
139
156
173
Index

```
> summary(stdres(model))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-4.212620 -0.478951  0.092105 -0.002409  0.614263  2.722682
>
> summary(studres(model))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-4.444176 -0.477836  0.091829 -0.005227  0.613109  2.777187
>
> summary(rstudent(model))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-4.444176 -0.477836  0.091829 -0.005227  0.613109  2.777187
```

The summaries from the standardized, studentized, and rstudent models show maximum values all around 2.7, and min values around -4.4.

Since the countries were arranged in descending order based on life expectancy, we can see that the countries with a higher life expectancy (samples from 1-70 or so) tend to have positive residuals, indicating that the country's life expectancy is larger than expected. Meanwhile, countries with lower life expectancies (samples 120-183) tend to have either large positive residuals or large negative residuals, indicating a mix between lower or higher than expected life expectancies.

We also notice that the majority of samples that pass the 95% and 99% cutoff values (our possible outliers) are countries with large negative residuals. However some outliers do have positive residuals as well.

Influence Analysis

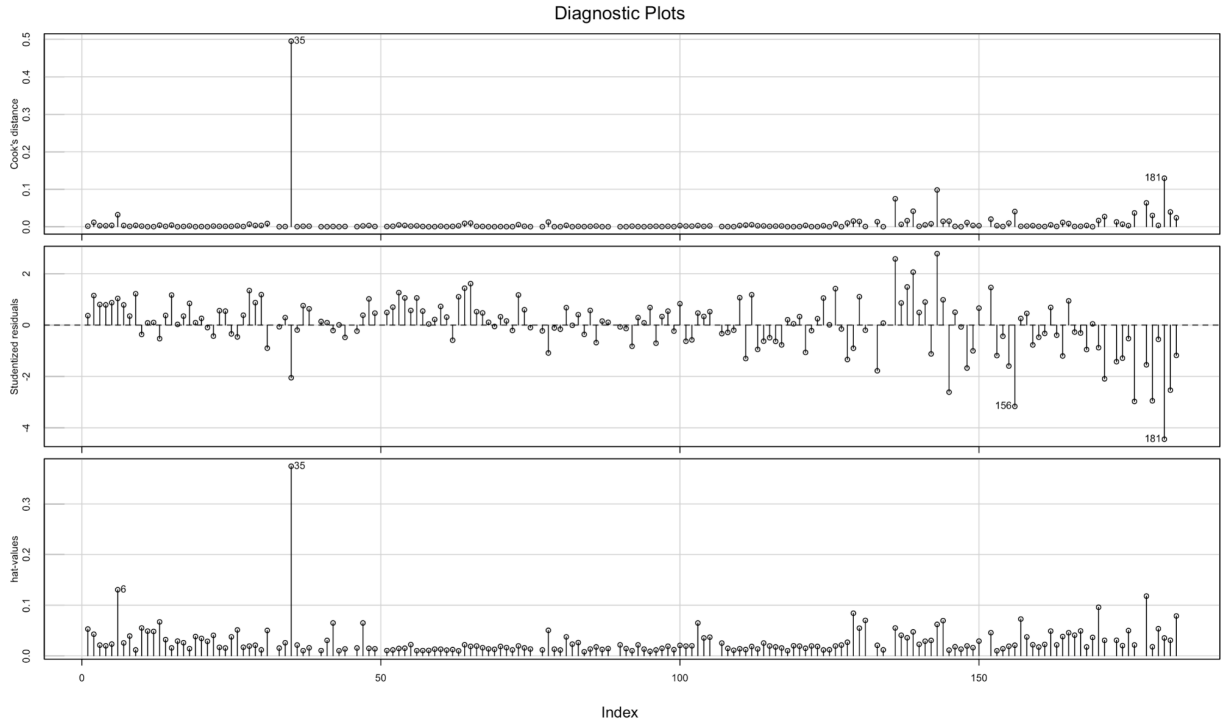
The following report shows the summary of potentially influential samples on the data set.

```
Potentially influential observations of
         lm(formula = life_expectancy ~ measles + basic_water + gni_capita +      gghe_d, data = LED12) :

    dfb.1_ dfb.msls dfb.bsc_ dfb.gn_c dfb.ggh_ dffit   cov.r   cook.d hat
6   -0.02   0.05    -0.02     0.36    -0.24     0.40    1.15_*  0.03   0.13_*
13  -0.03   0.01     0.04    -0.11    -0.04    -0.14    1.10_*  0.00   0.07
35  -0.04  -0.19     0.32    -1.53_*   0.77    -1.59_*  1.45_*  0.50   0.37_*
42   0.00  -0.01     0.01    -0.05     0.03    -0.06    1.10_*  0.00   0.06
47  -0.01   0.04    -0.03     0.09    -0.05     0.10    1.10_*  0.00   0.06
103  0.07  -0.11     0.08    -0.01    -0.02     0.12    1.09_*  0.00   0.06
129 -0.05  -0.04     0.16     0.08    -0.24    -0.27    1.10_*  0.02   0.08
131 -0.03   0.05    -0.04     0.01     0.01    -0.06    1.11_*  0.00   0.07
136  0.40   0.00    -0.49     0.11     0.17     0.62_*  0.90_*  0.07   0.05
143  0.59  -0.18    -0.45     0.13     0.14     0.71_*  0.87_*  0.10   0.06
145  0.10  -0.17     0.10     0.03    -0.06    -0.28    0.85_*  0.01   0.01
156 -0.23   0.34    -0.22     0.17    -0.18    -0.46    0.78_*  0.04   0.02
157 -0.01   0.05    -0.06     0.01     0.00     0.07    1.11_*  0.00   0.07
170 -0.25   0.23    -0.03    -0.11     0.04    -0.29    1.11_*  0.02   0.10_*
176 -0.10  -0.15     0.33     0.04    -0.19    -0.44    0.81_*  0.04   0.02
178 -0.49   0.52    -0.18     0.00     0.01    -0.57_*  1.09    0.06   0.12_*
179 -0.25   0.22    -0.09     0.06     0.11    -0.40    0.81_*  0.03   0.02
181 -0.11  -0.24     0.53     0.26    -0.60    -0.85_*  0.61_*  0.13   0.04
182 -0.07  -0.21     0.34    -0.05     0.06    -0.45    0.88_*  0.04   0.03
```

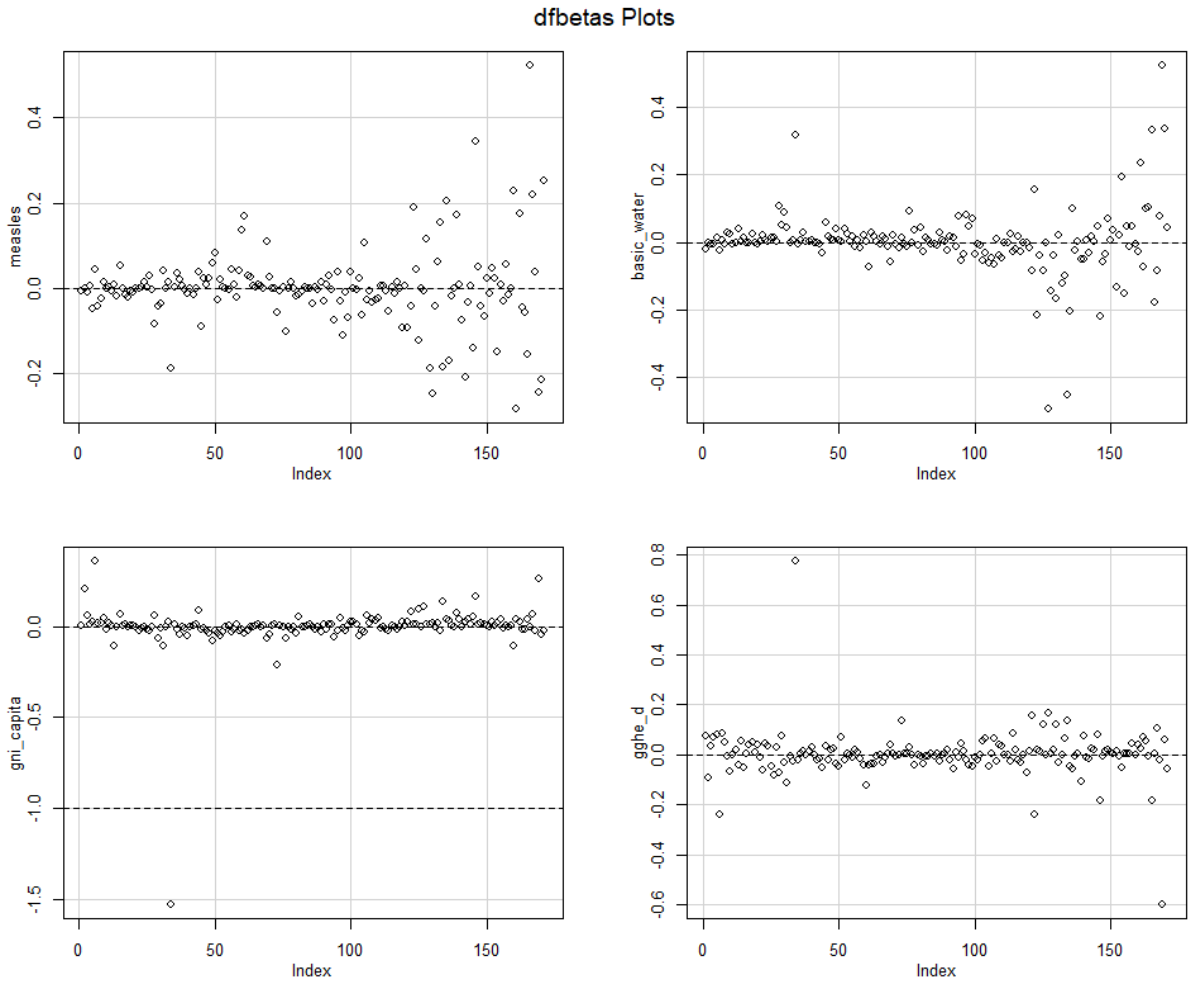
From the list of observations, we can see that sample 35 (Qatar) affects *gni_capita* significantly. We expect this observation to pull down the slope by a large amount since Qatar has the largest GNI per capita but is only ranked 35th in life expectancy. That is an indication that there is a limit

to how much money can improve life expectancy. The columns for DFFITS and the hat values show only 4-5 influential measures each. DFFITS influential values generally correspond to the same data points with large residuals, meanwhile that is not the case for hat values. However, the COVRATIO column shows this category is responsible for classifying the majority of the values as influential. The majority of points with COVRATIO values significantly less than 1 generally have negative DFFITS values (large negative residuals), although 2 samples show positive DFFITS.

We then visualize these relationships:



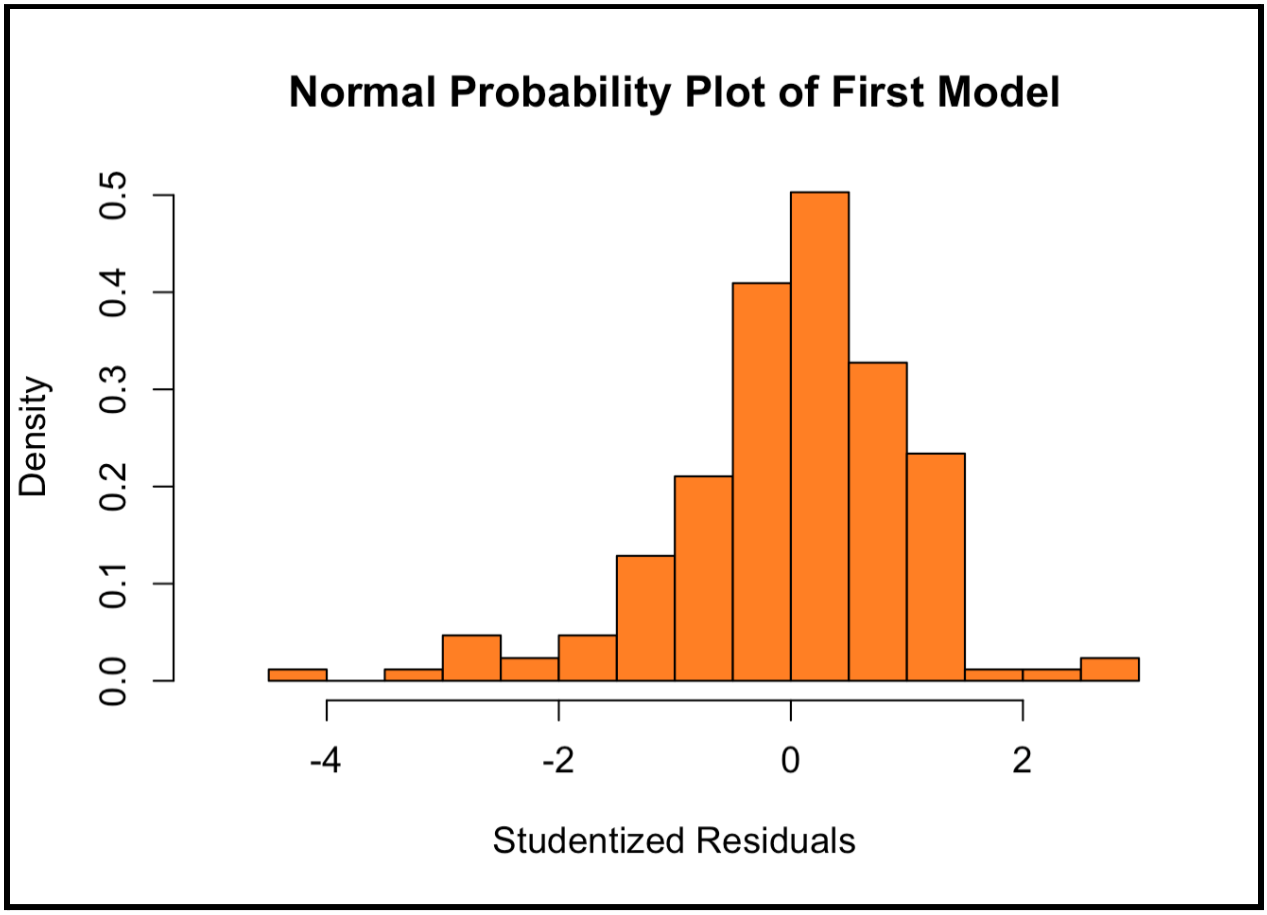
In the index plot we can see that Cook's distance plot shows sample 35 (Qatar) and 181 (Lesotho) are the largest values (with Qatar being overwhelmingly larger). For Studentized residuals, the largest values belong to samples 156 (South Africa) and 181 (Lesotho). For hat values, the largest values belong to samples 6 (Singapore) and 35 (Qatar). So, a general trend has shown that Qatar is an influential point necessary for consideration.



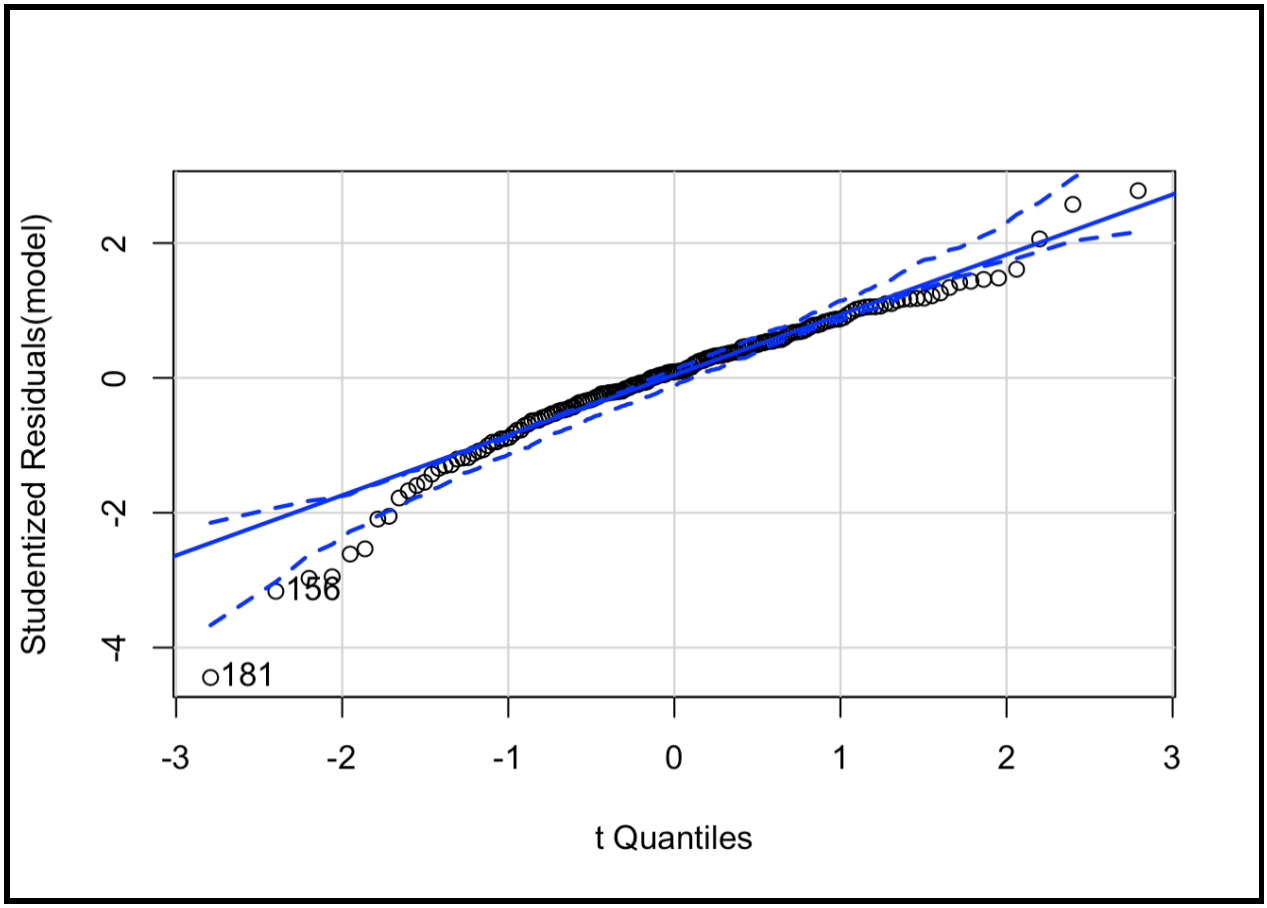
Looking at the DFBETAS plots, we can see the relationships in each variable. For measles, the values are consistently around 0 until around sample 120 where the values diverge from 0 in both directions. For basic water access, the values are consistently around 0 until around sample 130 where the values diverge in the negative direction and then diverge back in the positive direction. For GNI per capita, almost all values are around 0 except for sample 35 (Qatar) which has a large negative value compared to the rest (around -1.5). As explained before, this is due to an extremely large GNI per capita value while the life expectancy is not so large. For GGHE-D, almost all values are around zero except two samples where one has a large positive value around 0.8 and the other with a large negative value around -0.6, however neither are significant enough to be influential. The large positive value could be due to a sample with a small *gghe_d* value but a large life expectancy, and the negative value is vice versa.

Testing the Normality Assumption

Next, we constructed a normal probability plot to test if our residuals are normally distributed. We will use studentized residuals for the plot. The number of breaks in the histogram is calculated according to the square-root choice.

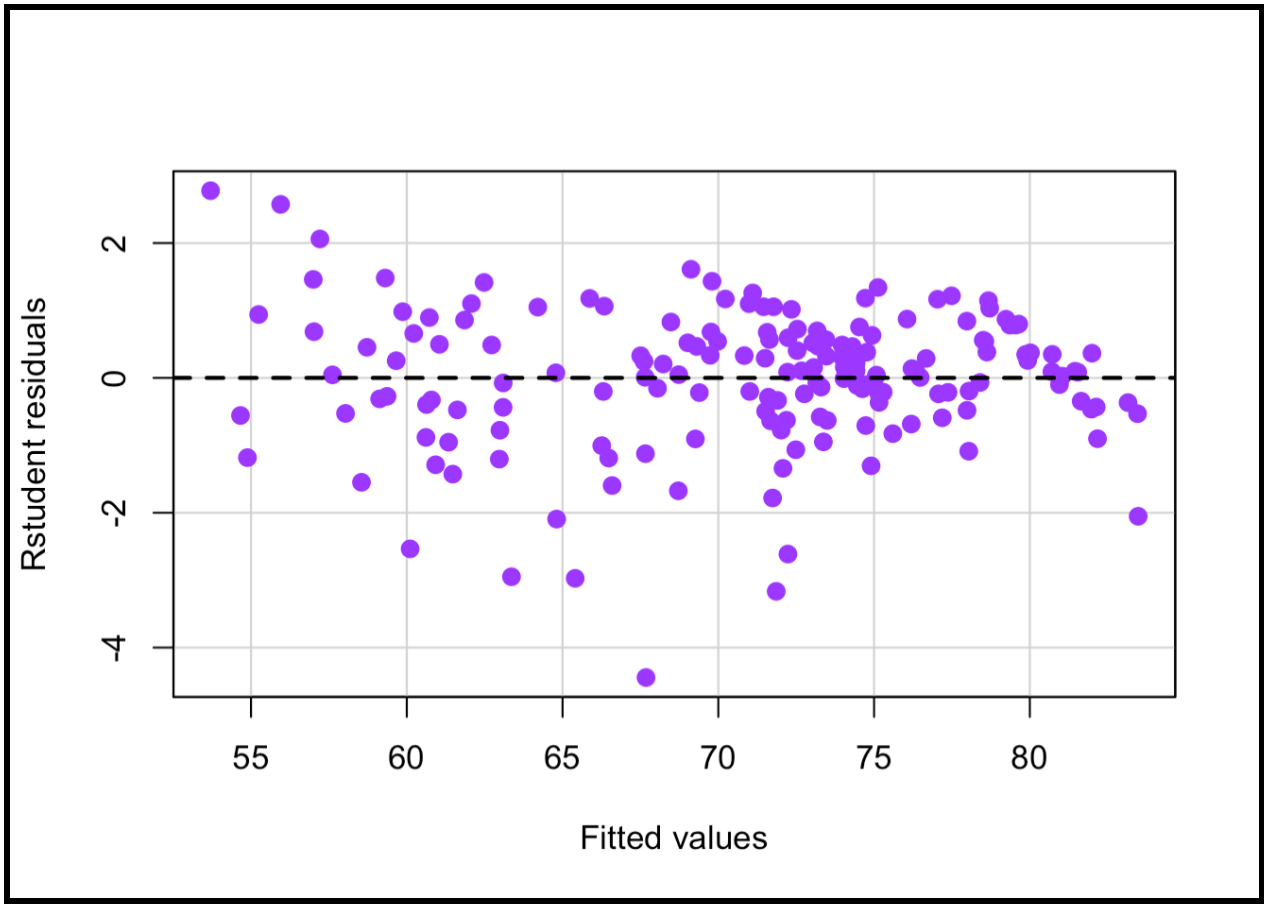


Normally, the tails in a normal distribution plot would become too small to detect with the histogram after the values of 2 and -2. However, we notice that while this histogram does maintain a normal distribution around 0, there is some heavy tailing in both the positive and, more significantly, the negative direction where we are still able to see values such as -4. We can use a QQ plot to confirm our thoughts.

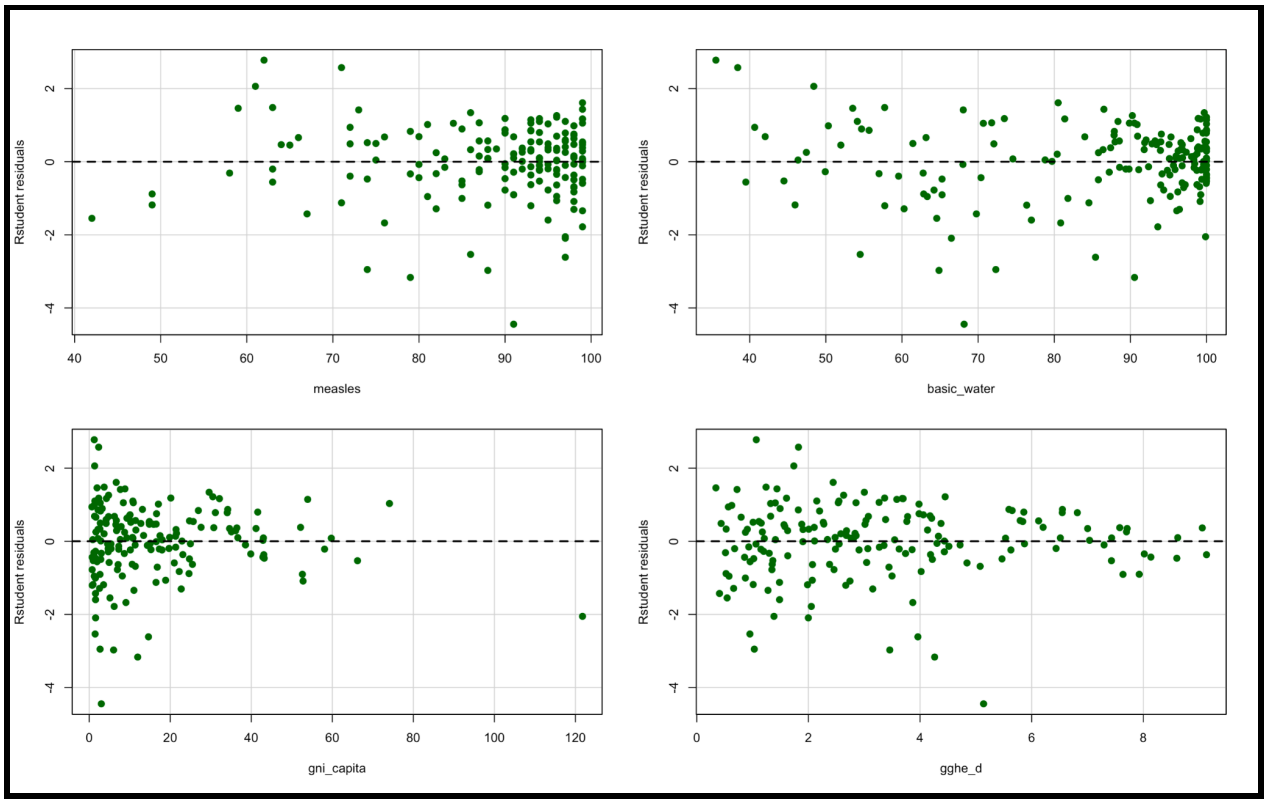


We notice two main details from the QQ-plot. The first one is the slight vertical asymptote at the -3 quantile. This is a feature that heavy tailed distributions normally have. However, we cannot call this plot heavy tailed since there is no vertical asymptote in the positive direction. Therefore this is a one-sided heavy tail graph. The second main detail we notice is the slight concave down shape of the residual distribution. This is an indication of a negatively-skewed distribution. This would indicate that there is a substantial amount of data where the samples have significantly low life expectancy compared to the determined fit.

Our next step is to plot the distribution of the residuals against the fitted values from our model. This will point out any model inadequacies.



Besides the few outliers that have extremely low life expectancies, the majority of the residuals lie randomly between values of 2 and -2. There does not seem to be any signs of funneling, double bows, or nonlinear errors. We do notice a larger density of residuals in the 70-75 year range and a lighter density in the <65 range. We should also consider the residual distribution against each regressor.

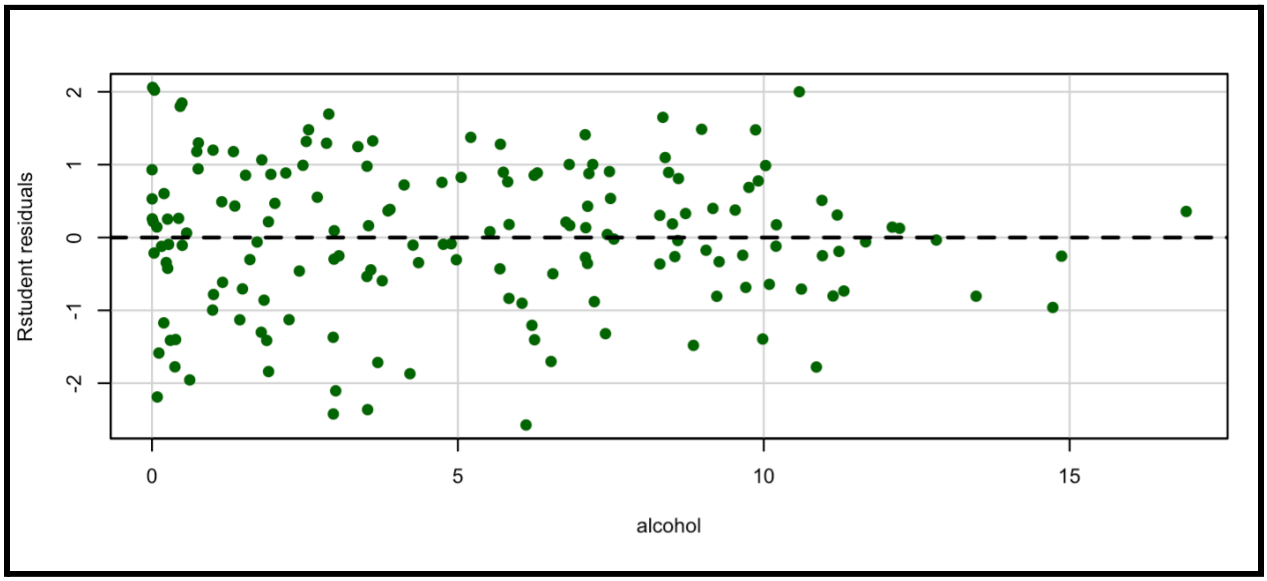


Although the plot of residuals against fitted values did not display anything unusual, we see more patterns with the plot against regressors. We are seeing two main details that need to be addressed. First, the majority of the data points are skewed to either extreme of their range, depending on the regressor. This causes most residuals to be compact. This may not be necessarily harmful, but may indicate the need for better scaling of the variable. The residuals for measles vaccination (top left) and basic water access (top right) are highly dense around the 90-100% range, while the residuals are highly dense in the lower range of GNI per capita. This indicates that the average vaccination rates and basic water access are not at 50% As for GNI per capita, this indicates a wealth gap where few countries have large GNI per capita values and the rest of the world have little variance in wealth among themselves. The second main detail to point out is that, besides basic water access and measles vaccination, we notice that the residuals follow a non-constant variance. For GNI per capita and GGHE-D, the variance is large at first and decreases as the residual density decreases (funnel shape). This suggests that we need to consider a transformation.

Our Attempt to Remove Outliers without a Transformation

We would like to mention that, while analyzing data, we attempted to improve the model by removing outliers and not performing any transformations. This was due to the ongoing learning process where we were still learning about transformations and indicator variables. To summarize our results, the model improved slightly, but problems like non-constant variance in GNI per capita residuals were not fixed. Also, the R^2 value was not largely changed, so not much more of the variance was becoming explained by our model. Our RMarkdown work will include the code for proof of our work, but we will talk much about it here. We advise any future work that it considers transformations before attempting to remove outliers. This is because certain data points that seem like outliers, such as Qatar which seemed influential, ends up becoming consistent with the model.

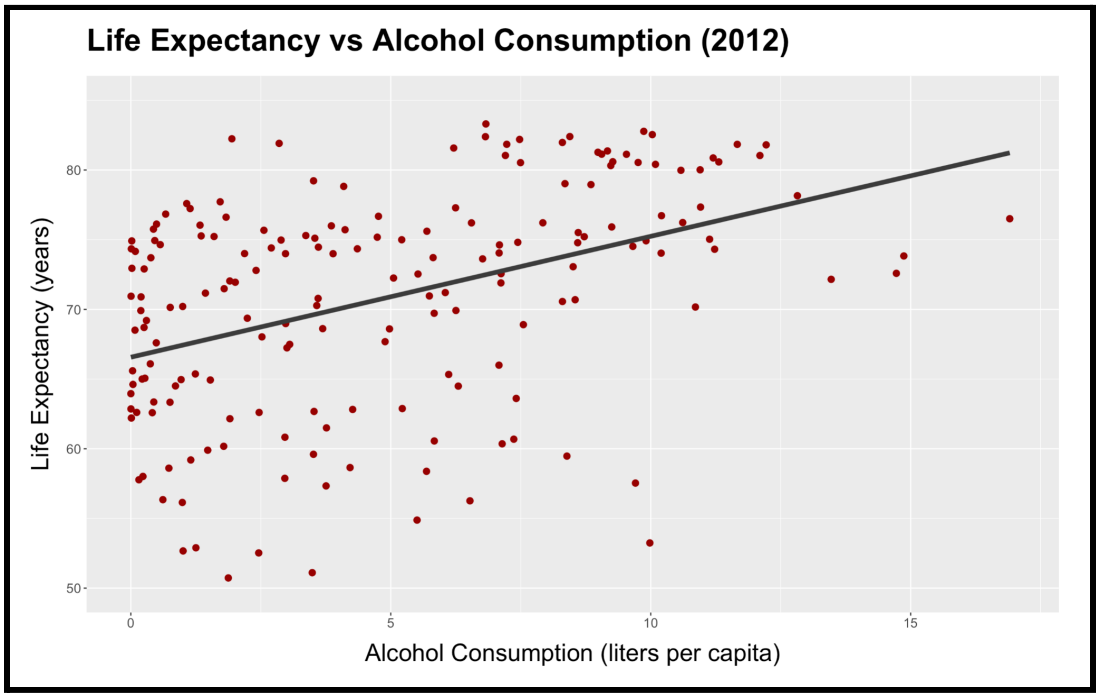
One extremely important note to make is that while performing outlier removal, we noticed that *alcohol* became a significant variable. Therefore when we consider if our regressors require transformations, we will also consider if the variable *alcohol* requires a transformation as well as check if it is still significant when no outliers have been removed but a transformed variable is introduced. A relevant graph to consider is the plot of residuals against the *alcohol* as a regressor. When outliers are removed (without a transformation), we notice a normal distribution. This indicates that *alcohol* possibly doesn't require a transformation but rather a better model in order to contribute properly.



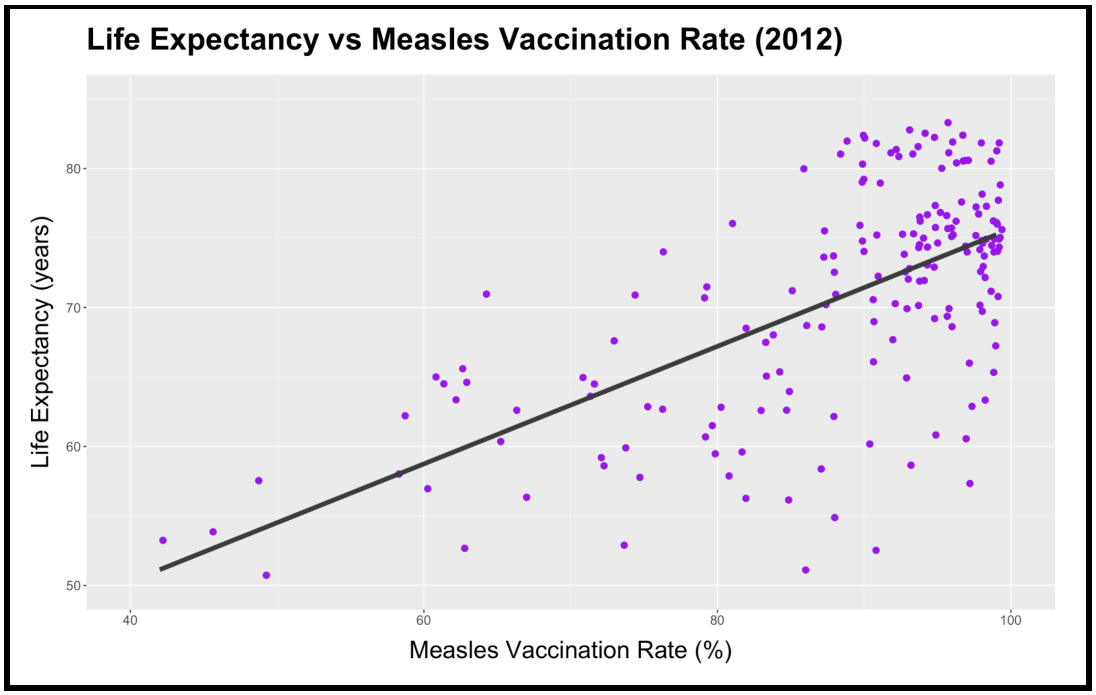
Transformations

Given our data from the plot of residuals against the regressors, we would like to further investigate if transformations on them will improve the model. So, before removing any outliers, we graph the univariate forms of each of our variables to inspect how the values are distributed.

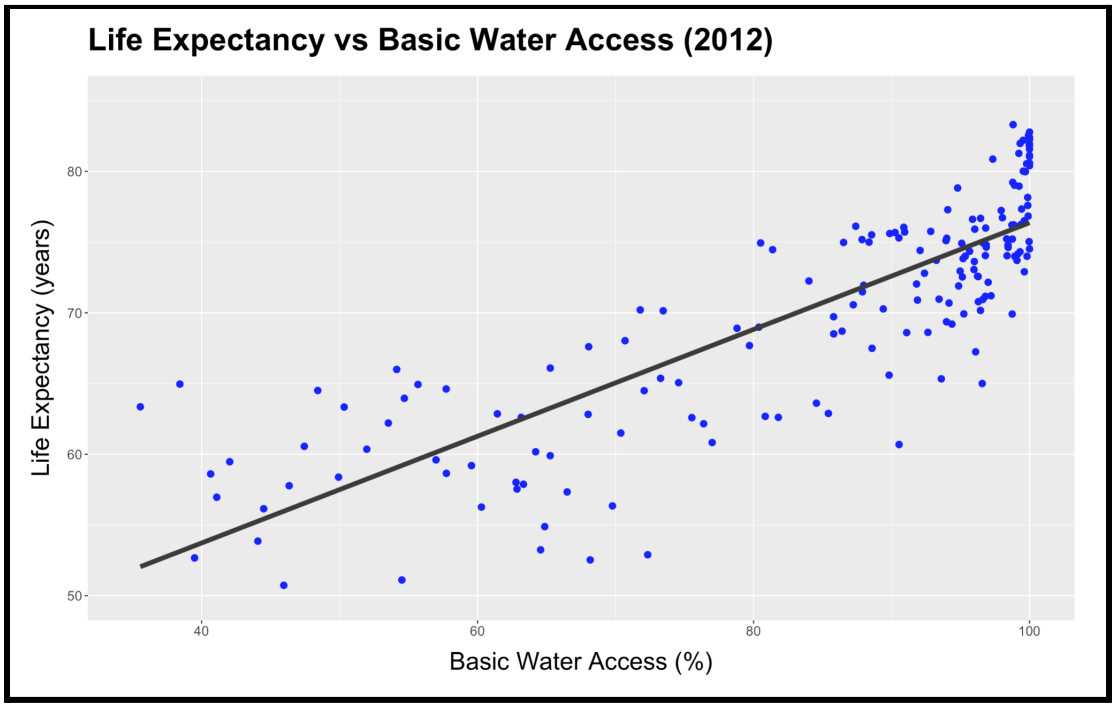
The following plots show the univariate, simple regression models of each of our regressors, including an alcohol (which was previously mentioned to be a potential regressor when outliers are removed or the model is improved).



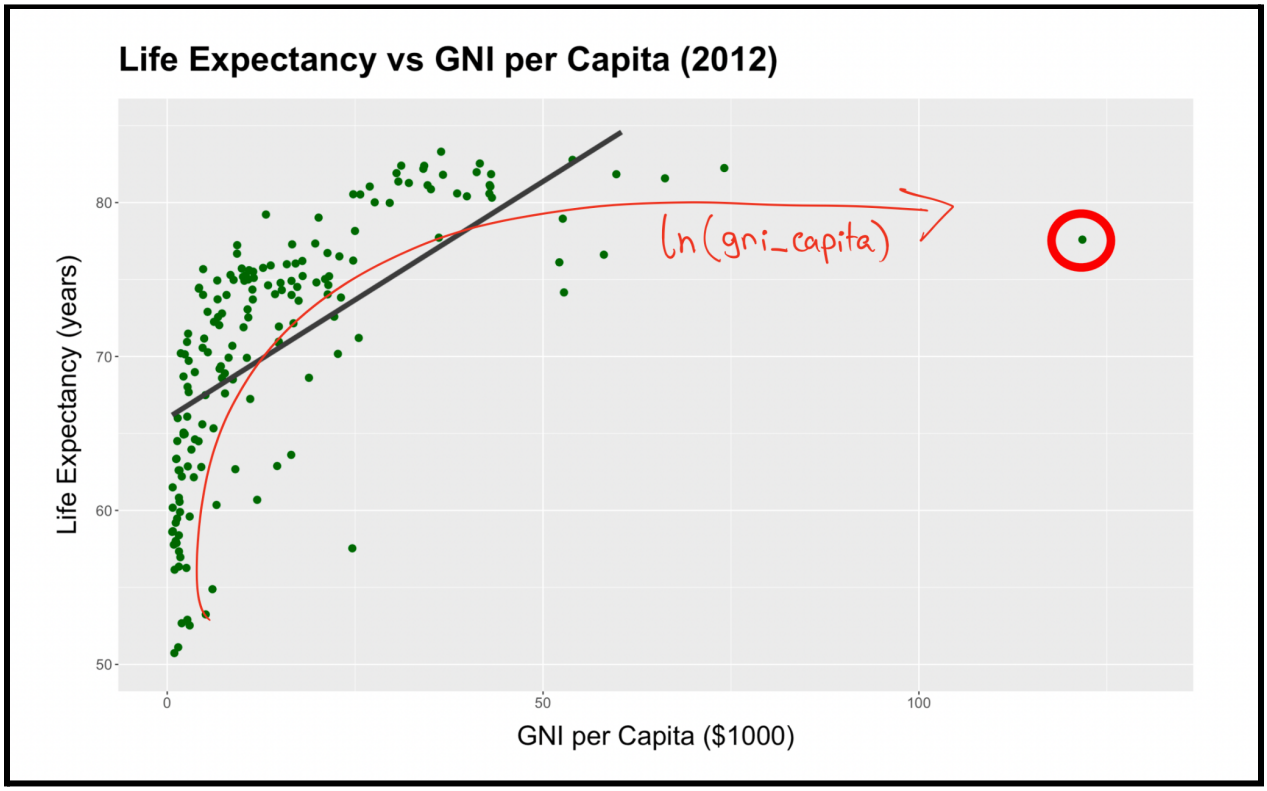
For alcohol consumption, there doesn't seem to be a clear distribution of the points where we can perform a transformation confidently. Given how the residuals are normally distributed in the residual plot against the regressor (in a model where outliers are removed), we don't see a need to perform a transformation.



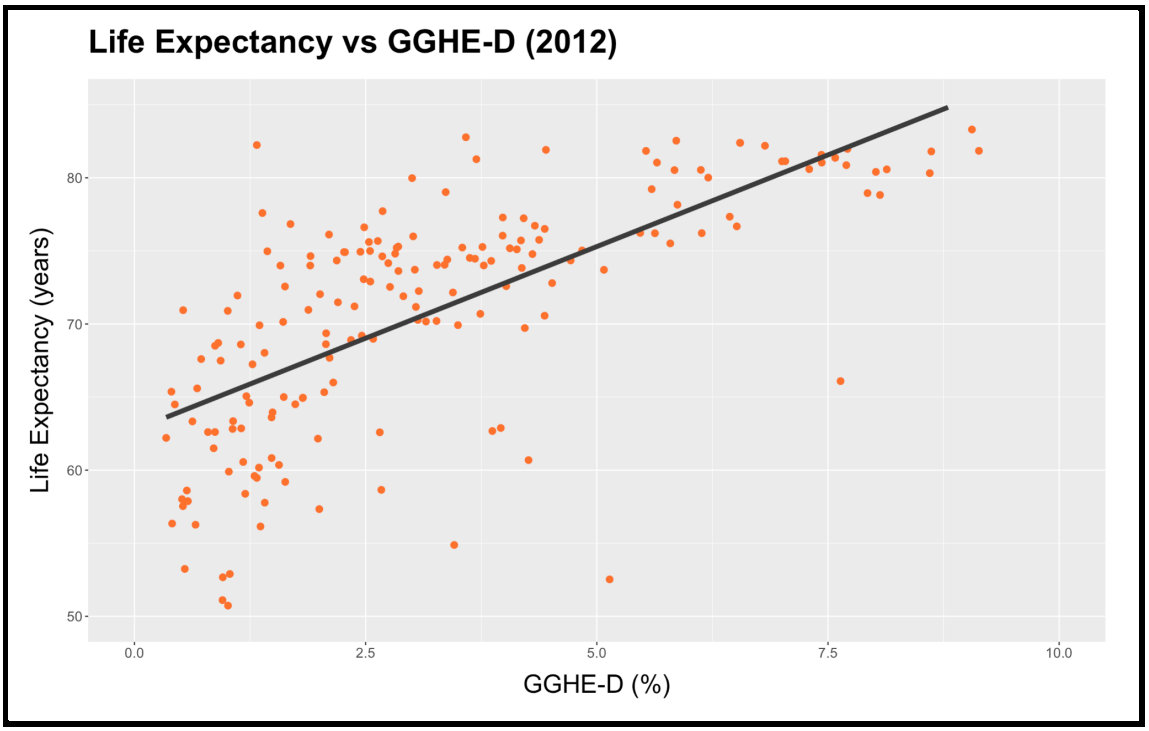
The relationship seems to be linear with *measles*, but the variance of the points seems to also increase slightly as the value of *measles* increases. However, the change in variance is minuscule and the plot of residuals against *measles* shows that the variance hardly changed. There is no priority or enough information to use a transformation.



The distribution with basic water access has a very strong linear relationship with a clear constant variance; no transformation needed.

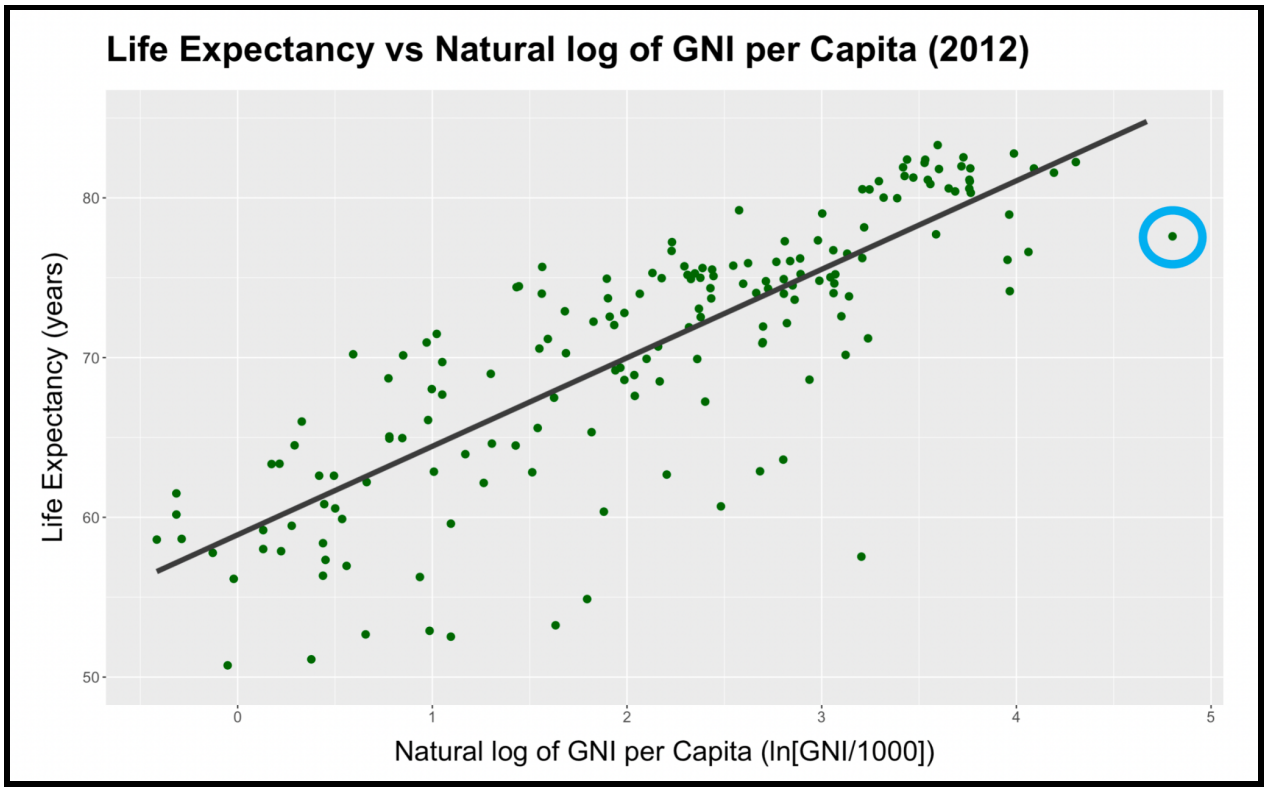


The distribution of GNI per capita looks strongly logarithmic. In fact, we notice that Qatar, marked with the red circle, would be a perfectly acceptable sample if the model followed a logarithmic trend, as shown with the red curve. Given how the plot of the residuals against the GNI regressor was not optimal, we can perform a transformation where we take the natural logarithm of the values and use that as the regressor instead. Whether we use a natural logarithm or a base 10 logarithm, the results will roughly be the same.



The distribution seems to have a slight curve to it, but given how the errors were normally distributed with a constant variance for the majority of points, the transformation isn't quite necessary. The limited range of GGHE-D values makes it less necessary for a logarithmic-type transformation. With this in mind, we decide not to transform the data, but we would investigate if a square root transformation would help.

Overall, we will transform the GNI per capita regressor by applying the natural log function to it. This will change the logarithmic distribution of the points into a linear one. After creating a new data set where the outliers that were removed are reintroduced, a new variable, called *ln_gni*, is introduced where its values are the natural logarithm of each corresponding *gni_capita* value. Once the model is created we can re-perform the linear fitting. The same procedure is applied where variables with too many missing values, high multicollinearity, and large p-values (above ~0.5 or so) are excluded. The following plot is the distribution of the points in *ln_gni* with a simple regression line added. We notice that our prediction of Qatar, labeled with the blue circle, being a consistent point after transformation is true.



We also performed a square root transformation of *gghe_d* to check if it would help with the multiple regression model during the variable selection process. Before performing the variable selection process, we wanted to check three different scenarios: only *gni_capita* is transformed, only *gghe_d* is transformed, or both variables transformed. We would then perform the variable selection process on each of these scenarios and compare the final models with each other. In all three scenarios, the same variables (*alcohol, measles, basic_water,* GNI per capita transformed or not, and GGHE-D transformed or not) came out as the significant regressors for our model. Using the values given to us by the summary such as R^2, adjusted R^2, t-values, and residual standard error, we determined that the best model only transformed *gni_capita* (i.e. included *ln_gni* as the regressor instead of *gni_capita*). However, there is one important thing to note. We found that another model, with very similar statistics (p-values, R^2 values, etc.), can be created by replacing *measles* with *polio* or *diphtheria*. It seemed that their high multicollinearity does not allow more than one of them to exist in the model at once as their p-values became large when two or more were included. So, we chose the model that only included *measles* for the sake of consistency in our model. Our new model is as follows:

```
Call:
lm(formula = life_expectancy ~ alcohol + measles + basic_water +
    ln_gni + gghe_d, data = LEDtrans)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6883  -1.3090   0.2055   2.1059   8.9497

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.94507    2.34689  17.447  < 2e-16 ***
alcohol     -0.20157    0.08865  -2.274  0.02427 *
measles      0.10145    0.03092   3.281  0.00126 **
basic_water  0.16078    0.03034   5.299 3.69e-07 ***
ln_gni       2.38290    0.44160   5.396 2.33e-07 ***
gghe_d       0.96659    0.17791   5.433 1.96e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.593 on 165 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.8078,    Adjusted R-squared:  0.802
F-statistic: 138.7 on 5 and 165 DF,  p-value: < 2.2e-16
```

We notice some details such as a smaller residual standard error, a larger adjusted R^2value, and a larger t-value compared to the original model. The median residual is closer to 0, and the maximum residual is smaller. Meanwhile, the absolute value of the minimum residual is actually larger.

Residual Analysis: Part 2

Variance Inflation Factors

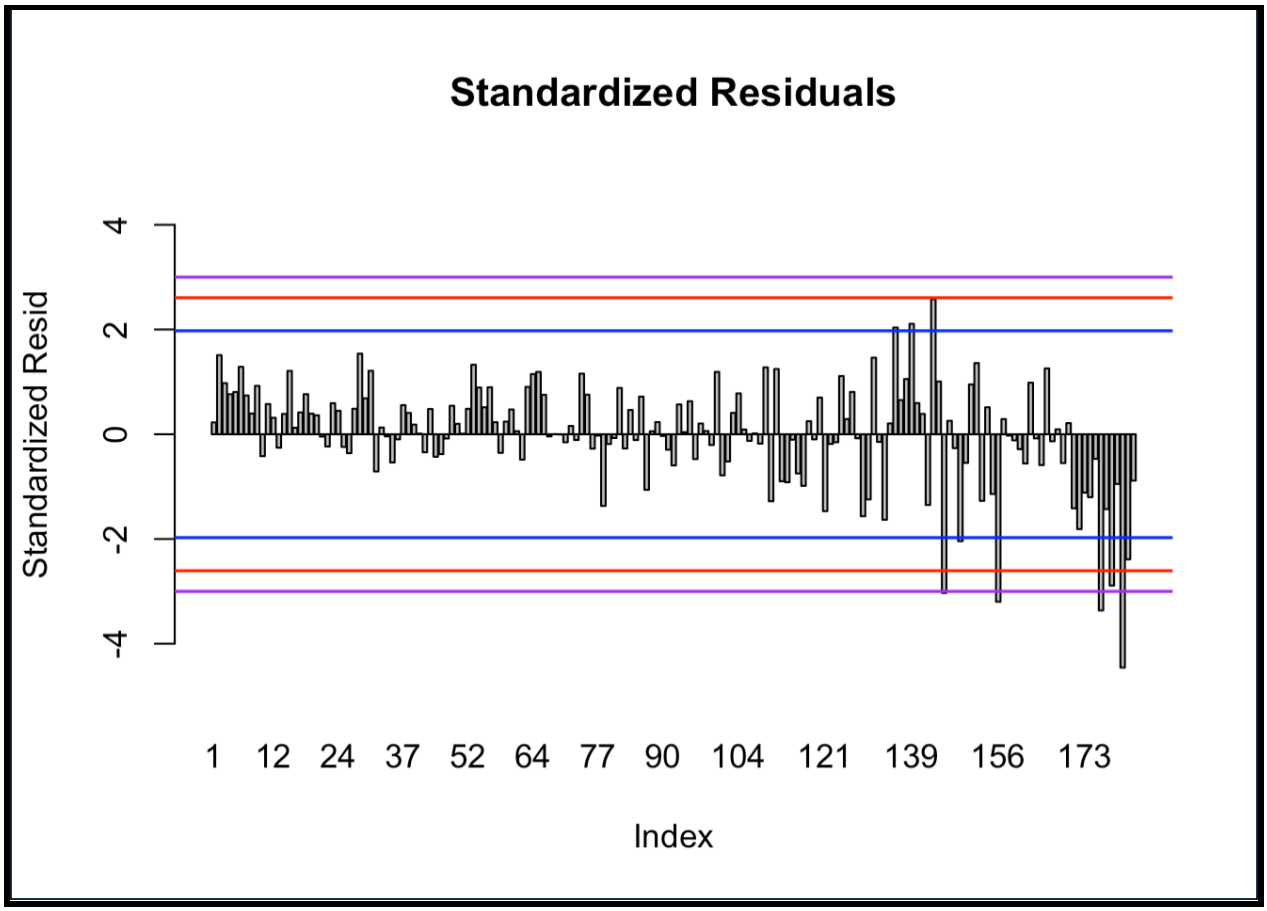
The following values are the VIFs of each of our new and old regressors:

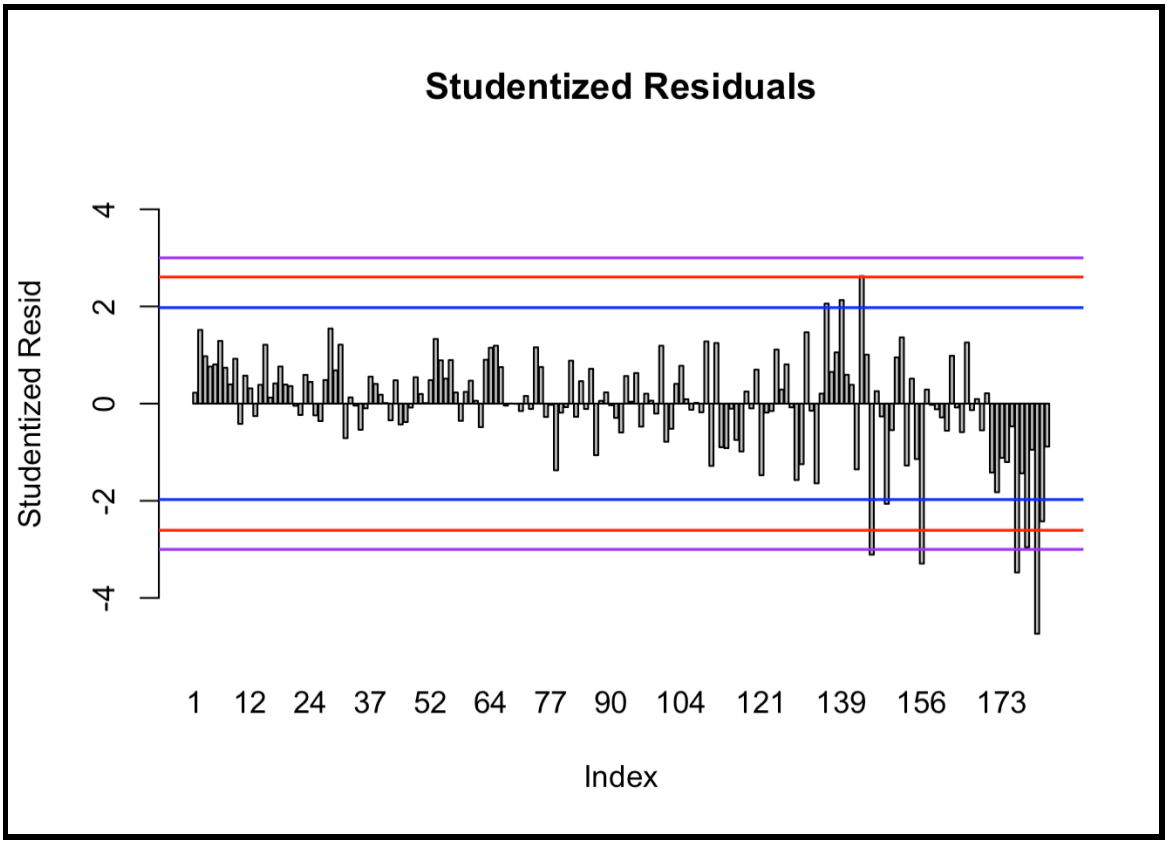
```
> vif(trModel)
    alcohol     measles basic_water      ln_gni       gghe_d
   1.587896    1.673845    3.757163    3.684309    1.980343
```

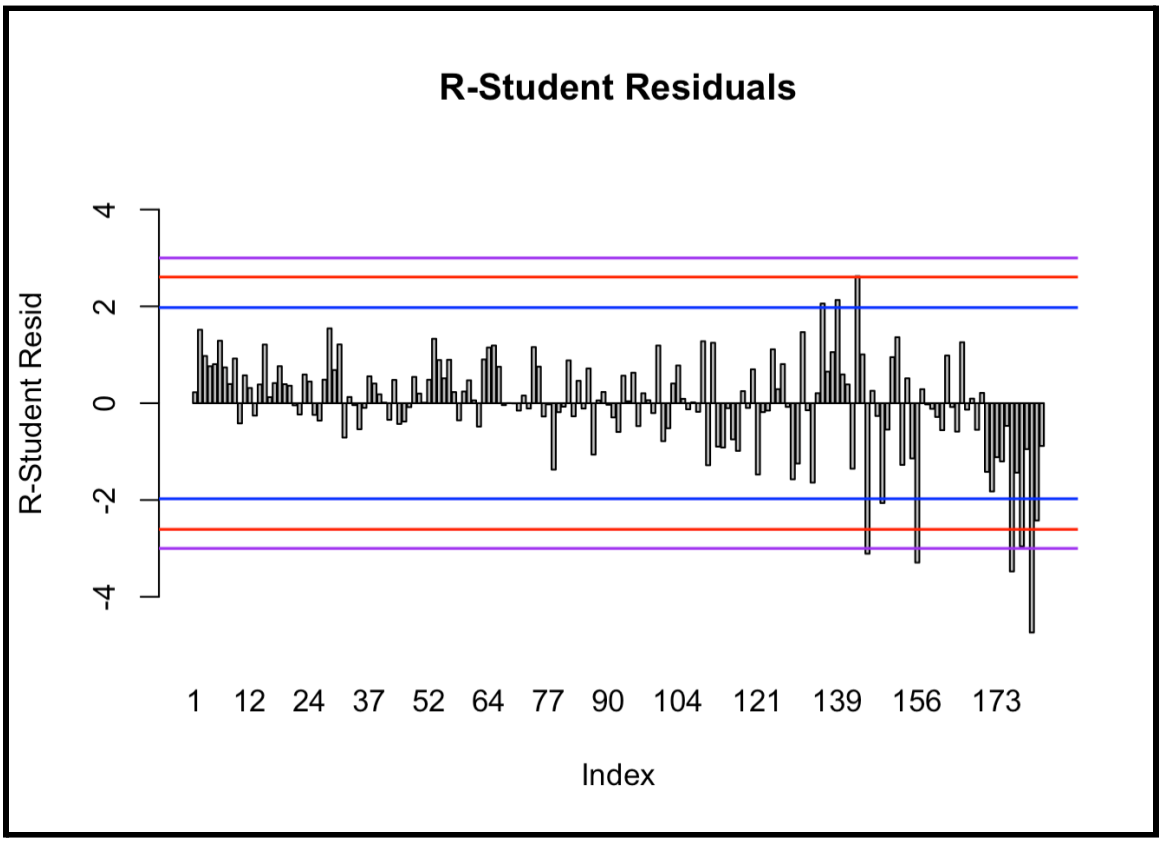
We notice that the multicollinearity between the natural logarithm of GNI per capita and basic water access is larger than that of the untransformed GNI regressor. This relationship is, however, normal. Countries with higher income will normally have better water access for its citizens. These variables are expected to be slightly related in their values. However, since these two variables, although slightly related, still have VIFs much less than 10, we will continue to incorporate both. Removing either one of them reduced the adjusted R^2 value and thus the model became weaker. We can tolerate slight amounts of collinearity as long as the variance isn't majorly inflated.

Residual Bar-plots & Summaries

The following plots show the standardized, studentized, and R-student residuals of our new transformed model. As mentioned before, the blue and red bars represent the 95% and 99% cutoff values. However, due to an additional regressor (*alcohol*), the degrees of freedom for the cutoff changed and thus the values are slightly different (although the difference is minuscule).


Standardized Residuals
Standardized Resid
4
2
0
-2
-4
1
12
24
37
52
64
77
90
104
121
139
156
173
Index


Studentized Residuals
Studentized Resid
4
2
0
-2
-4
1
12
24
37
52
64
77
90
104
121
139
156
173
Index


R-Student Residuals
R-Student Resid
4
2
0
-2
-4
1
12
24
37
52
64
77
90
104
121
139
156
173
Index

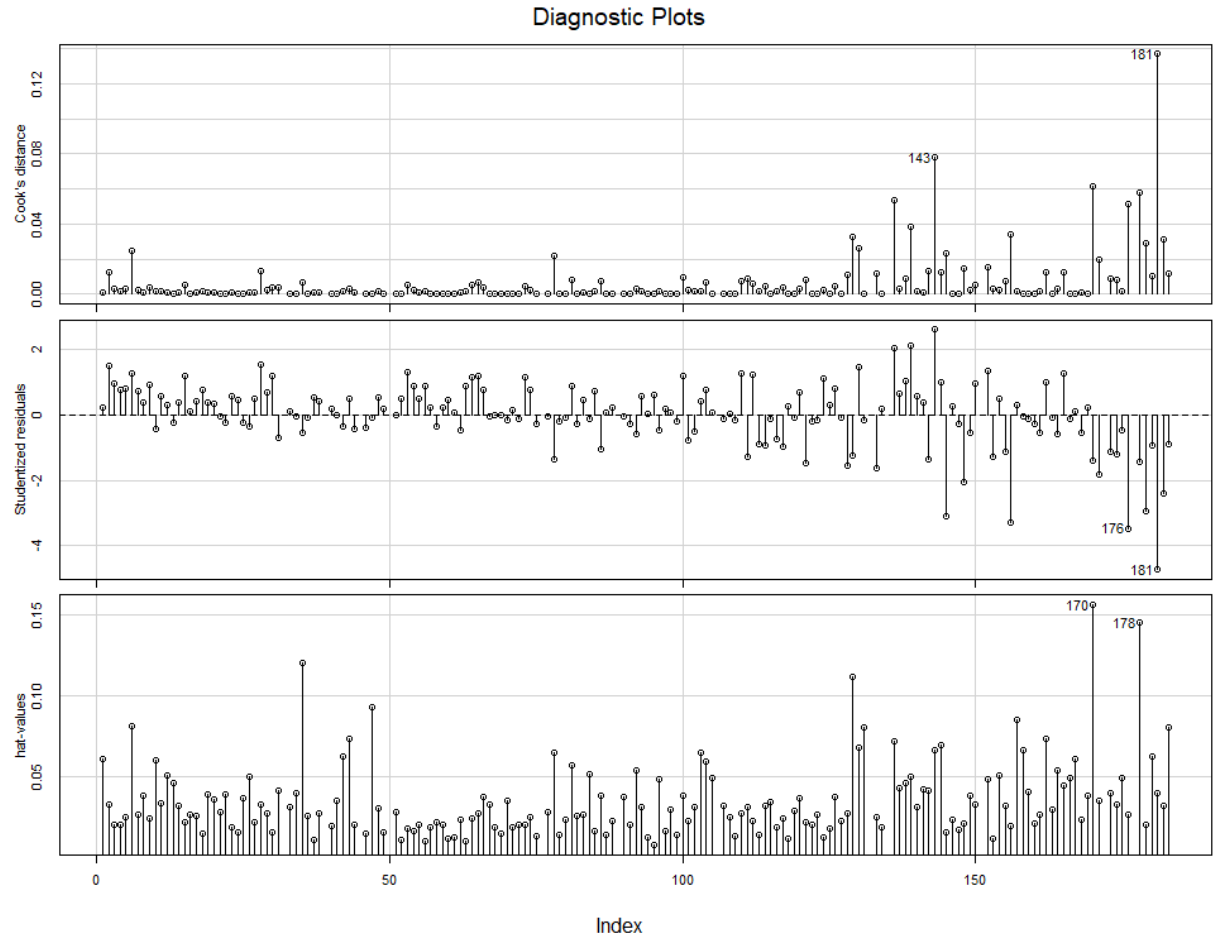
A consistent feature is that there are about 5-6 samples that pass the 99% cutoff. One slight feature that we noticed is that sample 35 (Qatar) which used to have a large negative residual now has a much smaller residual value. Meanwhile, some samples such as 181 (Lesotho) are still an outlier just as in the non-transformed model.

Influence Analysis

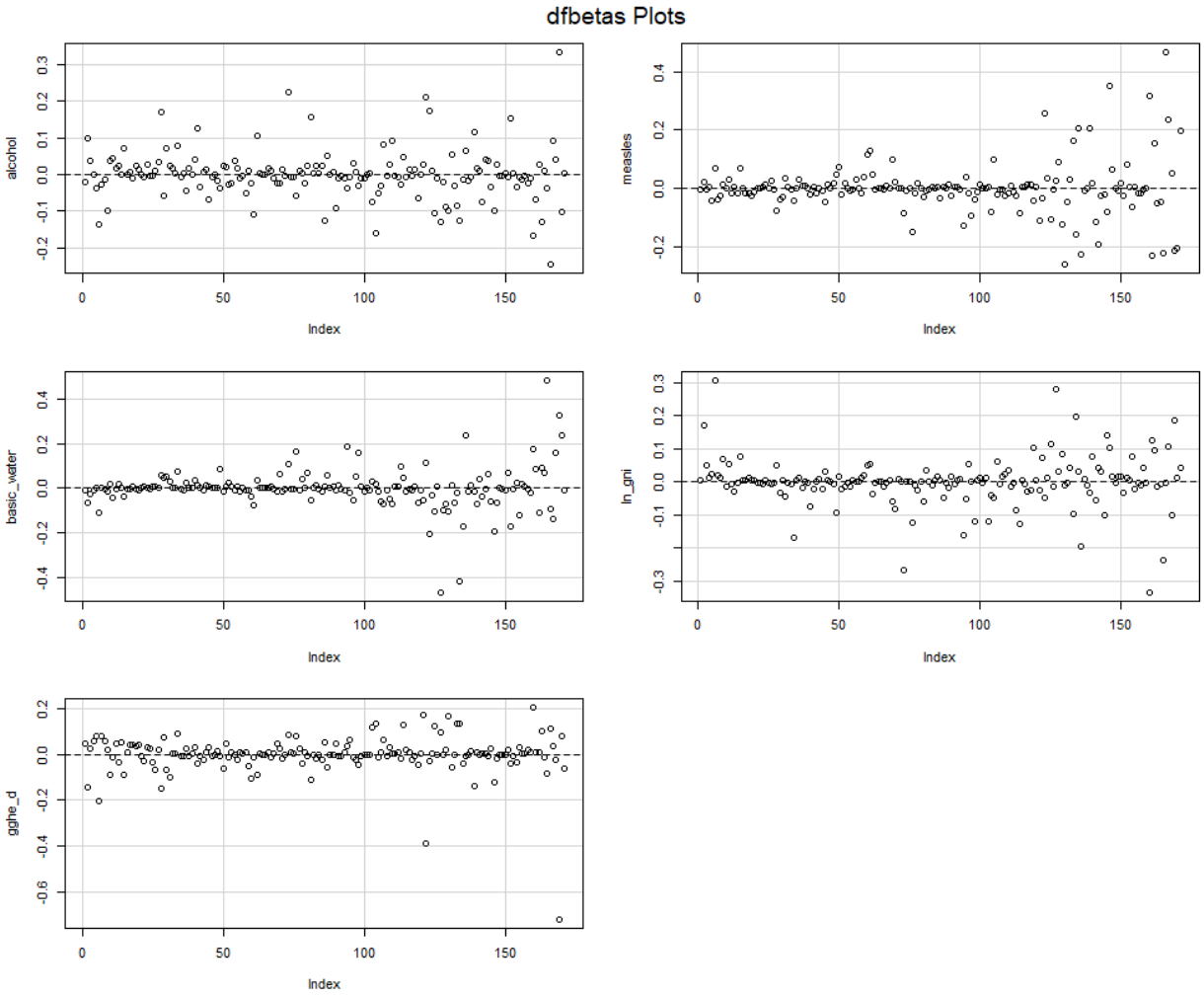
The following report shows the influential points in the transformed model:

```
Potentially influential observations of
         lm(formula = life_expectancy ~ alcohol + measles + basic_water +      ln_gni + gghe_d, data = LEDtrans) :

    dfb.1_ dfb.alch dfb.msls dfb.bsc_ dfb.ln_g dfb.ggh_ dffit   cov.r   cook.d hat
35  -0.12   0.08    -0.04     0.08    -0.17     0.09    -0.20    1.17_*  0.01   0.12_*
43  -0.03   0.13     0.00     0.02    -0.02    -0.04     0.13    1.11_*  0.00   0.07
47  -0.02   0.01    -0.01     0.02    -0.02     0.01    -0.03    1.14_*  0.00   0.09
129  0.01   0.21    -0.03     0.12     0.07    -0.39    -0.44    1.10    0.03   0.11_*
131  0.00   0.01     0.03    -0.03     0.01     0.00    -0.04    1.13_*  0.00   0.08
143  0.45  -0.13    -0.16    -0.42     0.20     0.13     0.70_*  0.87_*  0.08   0.07
145 -0.11   0.06    -0.23     0.23    -0.20    -0.01    -0.38    0.75_*  0.02   0.01
156 -0.05  -0.10     0.35    -0.19     0.10    -0.12    -0.47    0.72_*  0.03   0.02
157  0.01   0.03     0.06    -0.06     0.01    -0.02     0.09    1.13_*  0.00   0.09
158 -0.01   0.00     0.00     0.00     0.00     0.00    -0.01    1.11_*  0.00   0.07
170 -0.48  -0.17     0.32     0.18    -0.33     0.21    -0.61_*  1.14_*  0.06   0.16_*
176 -0.26  -0.04    -0.22     0.48    -0.24    -0.08    -0.57_*  0.70_*  0.05   0.03
178 -0.24  -0.25     0.47    -0.09     0.00     0.11    -0.59_*  1.13_*  0.06   0.15_*
179 -0.04   0.09     0.24    -0.14     0.11     0.04    -0.43    0.78_*  0.03   0.02
181  0.07   0.33    -0.21     0.33     0.18    -0.72    -0.96_*  0.50_*  0.14   0.04
182 -0.02  -0.10    -0.21     0.24     0.01     0.08    -0.44    0.87_*  0.03   0.03
```

We notice that the influence only lies in the areas of DFFITS, COVRATIO, and the hat values. We see that Qatar, sample 35, no longer has any major influence in the DFBETA column of GNI per capita (now called *ln_gni*). We also maintain the consistency where samples with large negative residuals (indicated by negative DFFITS values) generally have COVRATIO values less than 1. The following plot helps visualize the Cook's D values, Standardized residuals, and hat values.



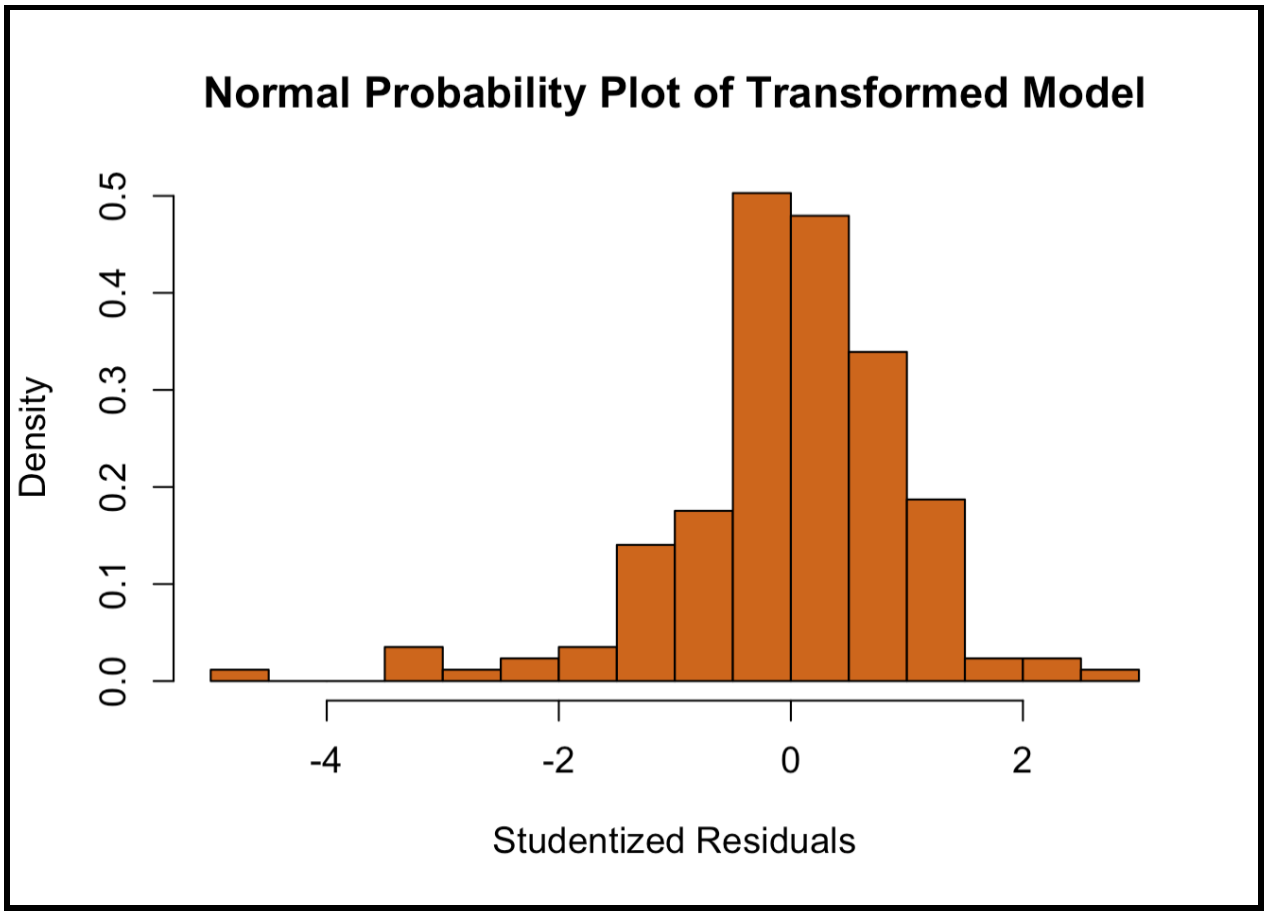
In contrast to the original model, this model shows that sample 35 (Qatar) isn't having an overwhelming amount of influence, and the other samples start showing their true influence. Although samples with smaller life expectancies have larger Cook's D values compared to high life expectancy samples, their overall Cook's D values are actually smaller now (the graph may be misleading due to the change in scaling). Samples 171 (Zimbabwe) and 178 (Nigeria) also now show larger hat values compared to before. Next we check the DFBETAS plots for each regressor:



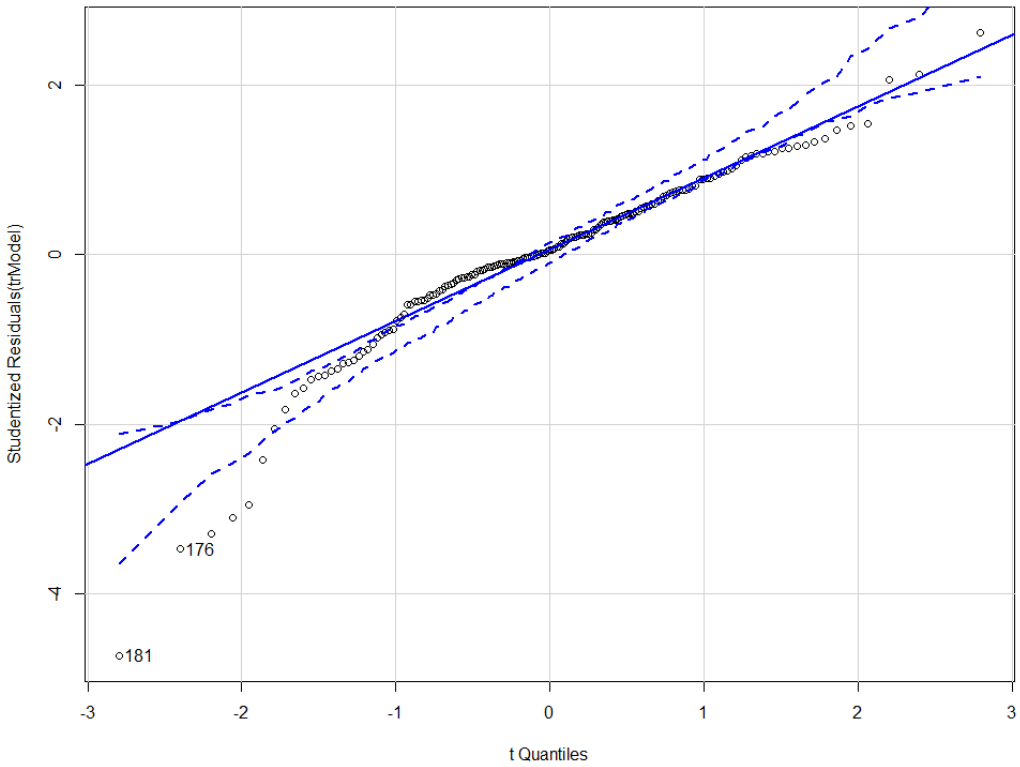
A consistent theme among all of the DFBETAS plots is that the variance in the influence is always slightly different as the index gets larger, including some points that seem to be outliers. Since there are no influential values in this category, there isn't much to say about it.

Testing the Normality Assumption

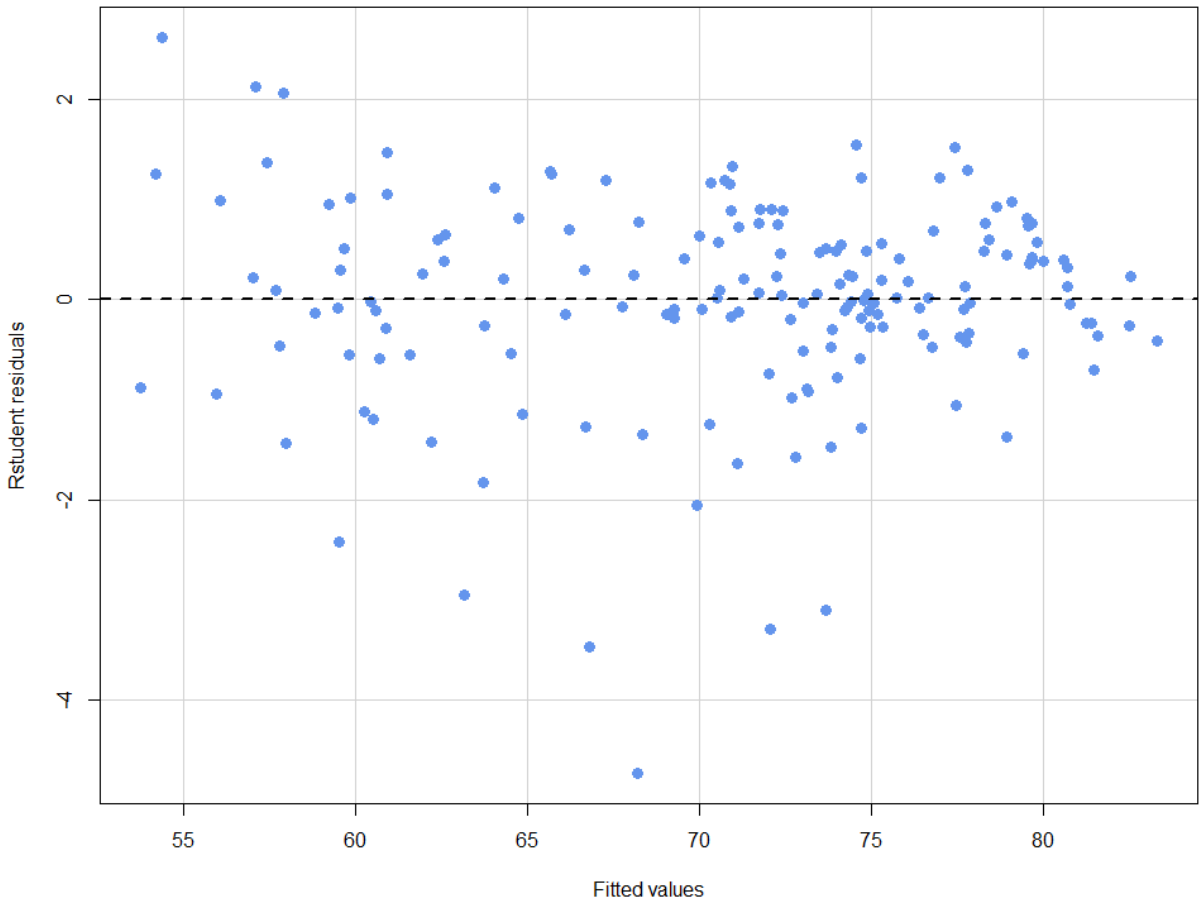
Our original reason for performing the transformation was to improve the plot of residuals, especially against the regressors such as GNI per capita (previously *gni_capita*). The following plot shows the histogram of the Studentized residuals:



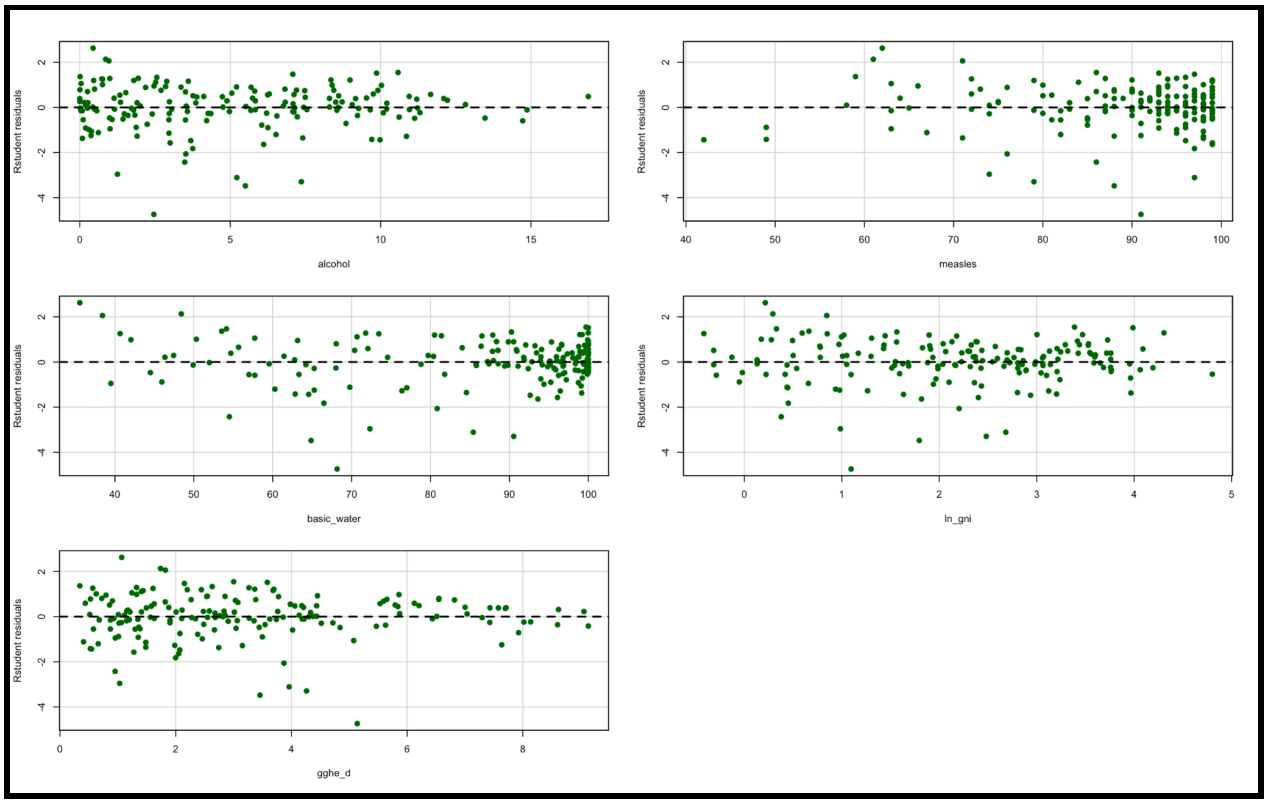
We still see that same trend where, due to the large negative residuals found in the data set, the probability plot seems to be left skewed. However, ignoring the data points outside the range of (-3,3), the plot looks normally distributed. This might be an indication that the removal of outliers is the only step required to perfect this distribution.



The plot isn't perfect with many of the points near the larger t-quantiles exceeding the barriers of a normal probability distribution. This is probably due to the existence of outliers. The concave down shape presented by the original model's QQ-plot isn't as nearly as apparent in this model, but there are still hints that it exists. The main concern this time is the deviation of the points rather than the slight concave shape of the distribution. We can support this by looking at the residual plots for R-student residuals:



We see a constant variance shape in the distribution (ignoring the outliers), so this is an indication for removal of outliers. We also notice (as seen with the residual bar plots) that the majority of the outliers have large negative residuals.



The distribution of points is similar to the non-transformed model. As for the logarithmic GNI per capita regressor, the distribution is more appropriate and less condensed in one area. GGHE-D has slightly tighter variance towards the larger values, but our previous testing showed that transforming it wouldn't have created a better model anyways. This slight difference is acceptable. This brings us to an overall conclusion: remove samples that match the outlier criteria. If a sample matches any of the following criteria, it is classified as an outlier and removed:

- Studentized or R-Student residual passes the 99% cutoff
- COVRATIO value is influential and less than 1

The samples that got removed were Ethiopia, Botswana, South Africa, Eswatini, Côte d'Ivoire, Lesotho, and Sierra Leone. It is interesting to see that the majority of the outliers are countries in South Africa, and all of them belong to the continent of Africa. This indicates that the region of Africa as a whole is an important piece to consider. Analysis of the region of Africa alone can provide some insightful relationships. However, since our analysis deals with the world as a whole, we will proceed to not include whichever countries end up being outliers, even if they are all African. We will simply note that the advice we give countries based on our analysis do not apply to many African countries which may require their own separate analysis.

New and Improved Model

After removing the outliers from the previous model and going through the variable selection process again, we get the following model:

```
Call:
lm(formula = life_expectancy ~ alcohol + measles + basic_water +
    ln_gni + gghe_d, data = LEDtrans.2)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6363 -1.4643 -0.0285  1.8740  6.7173

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.52777    1.86007  22.326  < 2e-16 ***
alcohol     -0.21360    0.06846  -3.120  0.00215 **
measles      0.11468    0.02423   4.734 4.87e-06 ***
basic_water  0.14205    0.02416   5.879 2.37e-08 ***
ln_gni       2.32812    0.34218   6.804 1.99e-10 ***
gghe_d       1.08733    0.13862   7.844 6.13e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.759 on 158 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.8734,    Adjusted R-squared:  0.8694
F-statistic:   218 on 5 and 158 DF,  p-value: < 2.2e-16
```

First, we note that no new variables became significant with the variable selection process. The same variables that were insignificant, such as *hepatitis* and *age5_19thinness,* are still insignificant. We also still have the same trend that *polio* or *diphtheria* could replace *measles* and maintain a similar model.

The major details of this model are the decreased RSE, increased Adj. R^2, and increase in the F-statistic and absolute value of the t-values. A minor detail is the decreased degrees of freedom due to less data points which will affect the 95% and 99% cutoff values (although extremely slightly). We also

notice that the absolute value of the minimum and maximum residuals are now smaller.

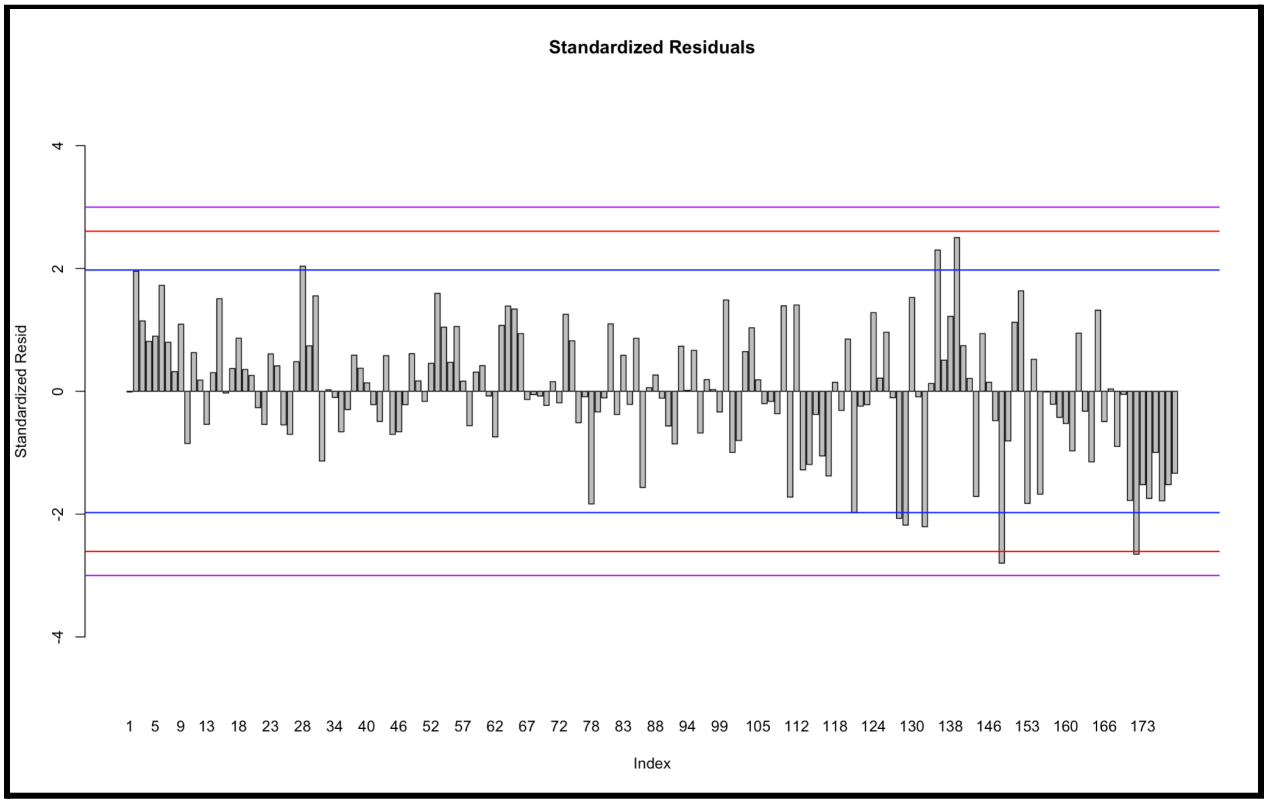
Variance Inflation Factors

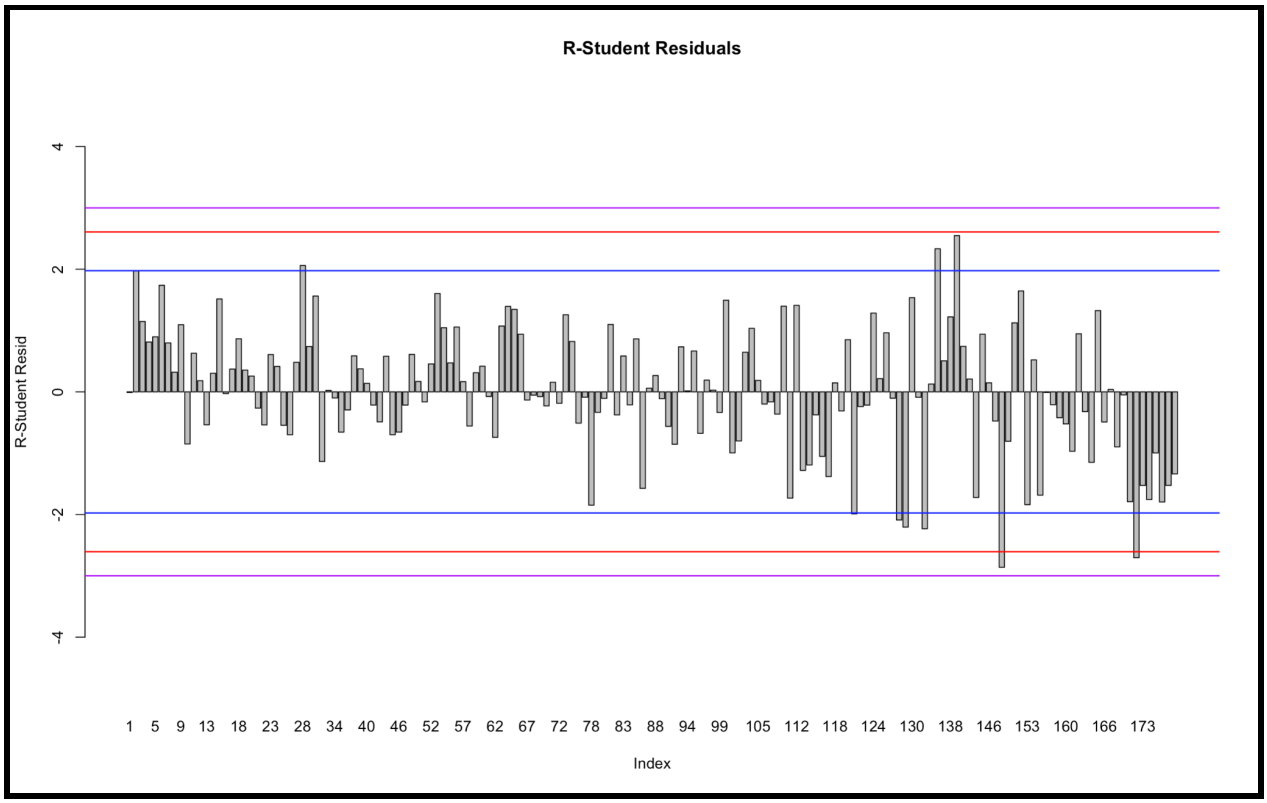
The following values show the VIF values for each regressor. In general, there is an overall decrease in multicollinearity. However, the difference is very slight and does not require attention.

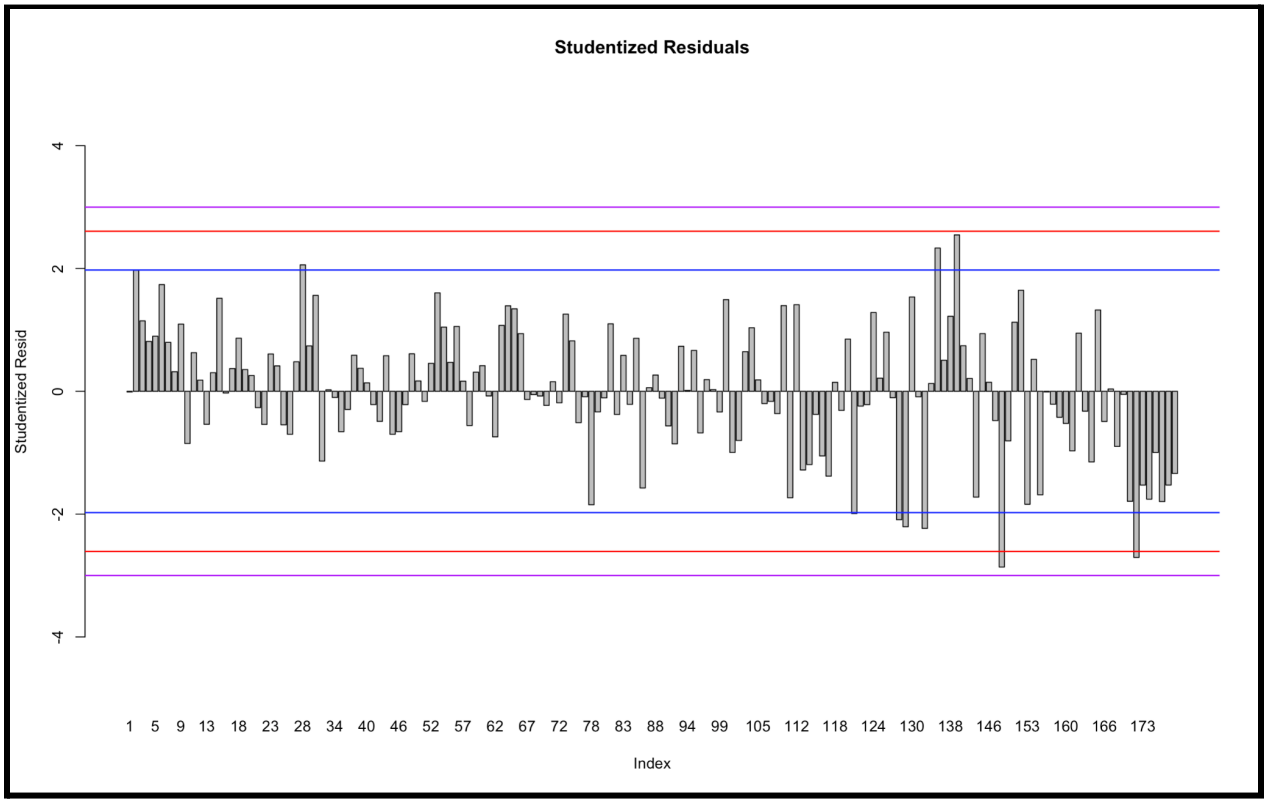
```
> vif(trModel.2)
    alcohol     measles basic_water      ln_gni      gghe_d
   1.574827    1.656872    3.705979    3.596684    1.988085
```

Residual Bar-plots & Summaries

The following plots show the scaled residuals after removing outliers:




R-Student Residuals
R-Student Resid
4
2
0
-2
-4
1 5 9 13 18 23 28 34 40 46 52 57 62 67 72 78 83 88 94 99 105 112 118 124 130 138 146 153 160 166 173
Index


Studentized Residuals
Studentized Resid
4
2
0
-2
-4
1 5 9 13 18 23 28 34 40 46 52 57 62 67 72 78 83 88 94 99 105 112 118 124 130 138 146 153 160 166 173
Index

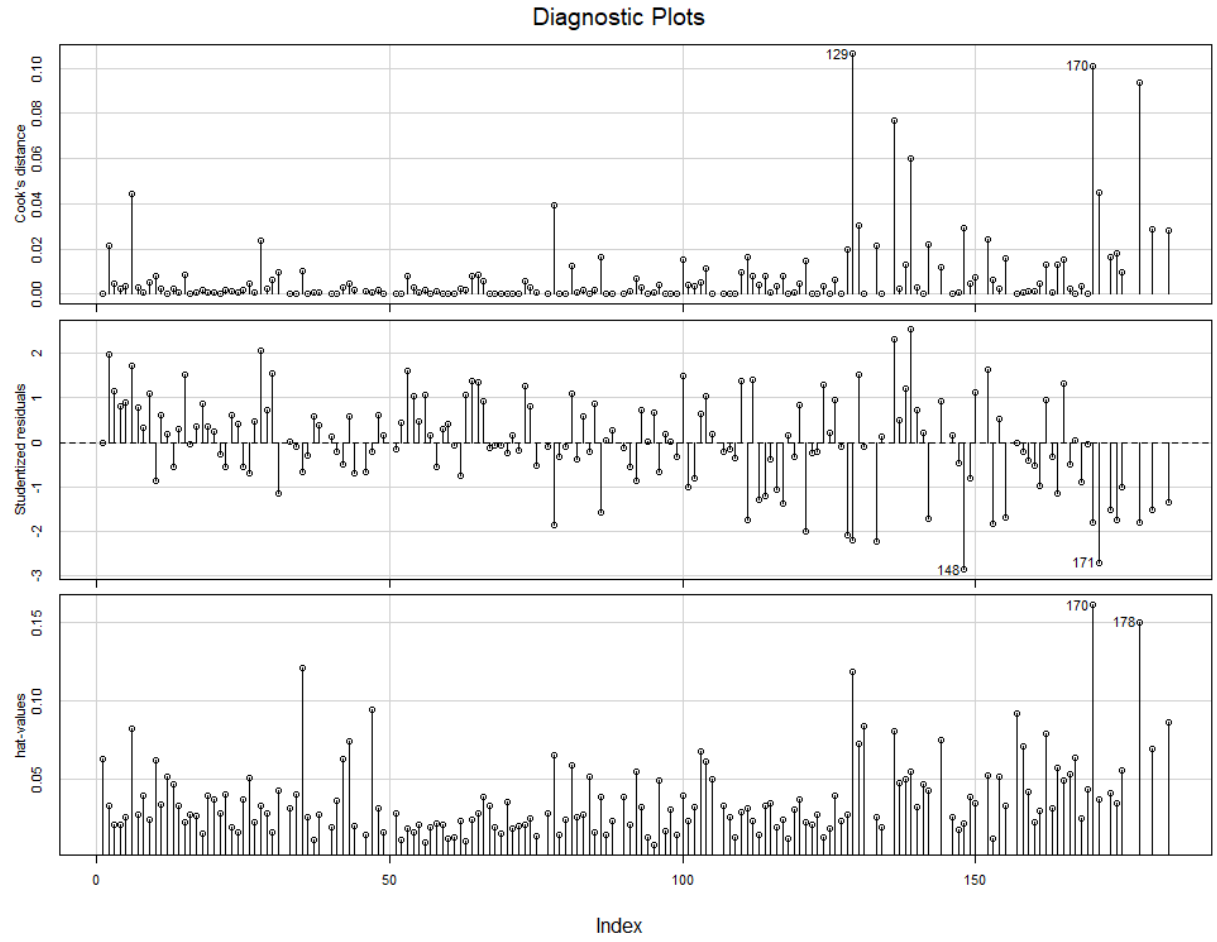
The main consistent detail is that there are two new outliers that pass the 99% range, samples 148 (Namibia) and 171 (Zimbabwe). Consistently, these countries are also African countries. We speculate that these samples were nearly coincident with the previous outlying cases, therefore their values seemed to be consistent with a model that was heavily skewed in their direction. However, since outliers were already removed once, we can be more conservative and instead label any scaled residuals larger than 3 or -3 as outliers. If we can show that the model can be improved by using the 99% cutoff instead, we will remove those two samples.

Influence Analysis

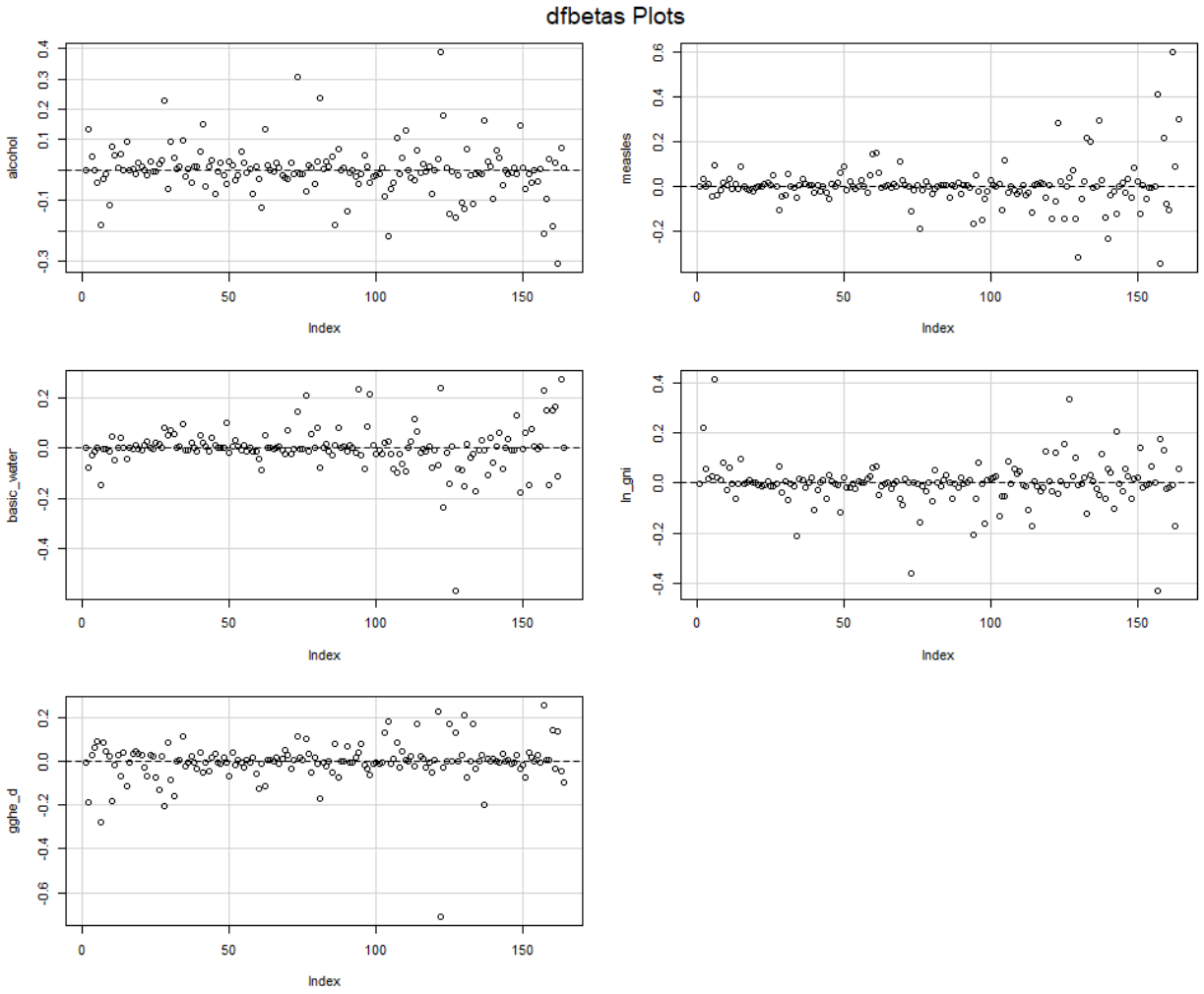
Now we measure influence:

```
Potentially influential observations of
        lm(formula = life_expectancy ~ alcohol + measles + basic_water +      ln_gni + gghe_d, data = LEDtrans.2) :

    dfb.1_ dfb.alch dfb.msls dfb.bsc_ dfb.ln_g dfb.ggh_ dffit   cov.r   cook.d hat
35  -0.15   0.10    -0.05     0.09    -0.21     0.11    -0.24    1.16_*  0.01   0.12_*
47  -0.04   0.03    -0.03     0.04    -0.06     0.02    -0.07    1.14_*  0.00   0.09
129  0.00   0.39    -0.07     0.24     0.12    -0.71    -0.81_*  0.98    0.11   0.12_*
131  0.00   0.01     0.02    -0.02     0.01     0.00    -0.03    1.13_*  0.00   0.08
133  0.23  -0.14    -0.14    -0.14     0.16     0.17    -0.36    0.88_*  0.02   0.03
136  0.49  -0.16     0.04    -0.57     0.33     0.13     0.69_*  0.92    0.08   0.08
139  0.29  -0.13    -0.32    -0.15    -0.01     0.21     0.61_*  0.86_*  0.06   0.05
148 -0.21   0.16     0.29    -0.01    -0.05    -0.20    -0.43    0.78_*  0.03   0.02
157  0.00   0.00     0.00     0.00     0.00     0.00     0.00    1.14_*  0.00   0.09
158 -0.05  -0.01     0.01     0.04    -0.03     0.01    -0.06    1.12_*  0.00   0.07
170 -0.62  -0.21     0.41     0.23    -0.43     0.26    -0.78_*  1.10    0.10   0.16_*
171  0.21  -0.10    -0.35     0.15     0.18     0.01    -0.53    0.82_*  0.05   0.04
178 -0.31  -0.31     0.60    -0.11    -0.01     0.14    -0.76_*  1.08    0.09   0.15_*
```

Many of the original influential values (mainly the ones with COVRATIO > 1) remain. There are new influential values that arose with COVRATIO < 1, samples 148 (Namibia) and 171 (Zimbabwe). Now we confirm that not only do these samples have large residuals, but they also decrease the precision of the model. This implies that their removal could perhaps improve the model by using the 99% t-value as the cutoff. Visualizing the influence measures may also help.



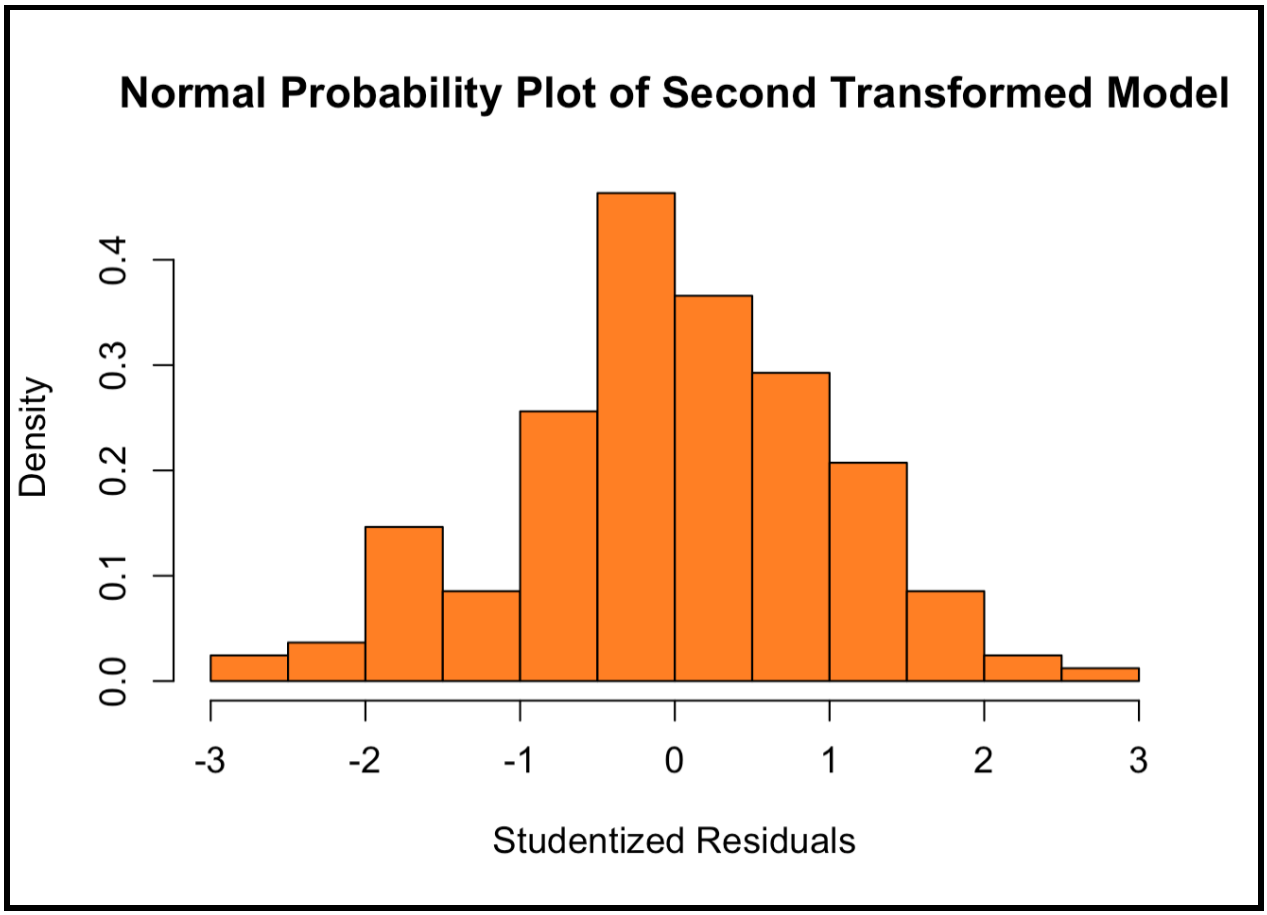
It is clear that sample 129 (Kiribati), 170 (Equatorial Guinea), and 178 (Nigeria) have large Cook's D and hat values. This is probably due to the high values of GNI per capita and alcohol consumption for Equatorial Guinea and the large alcohol consumption of Nigeria, meanwhile their life expectancies are one of the lowest. As for Kiribati, it is probably due to its large GGHE-D value but relatively low life expectancy. Again, we also see samples 148 (Namibia) and 171 (Zimbabwe) showing the two largest negative residuals in the Studentized residuals section.



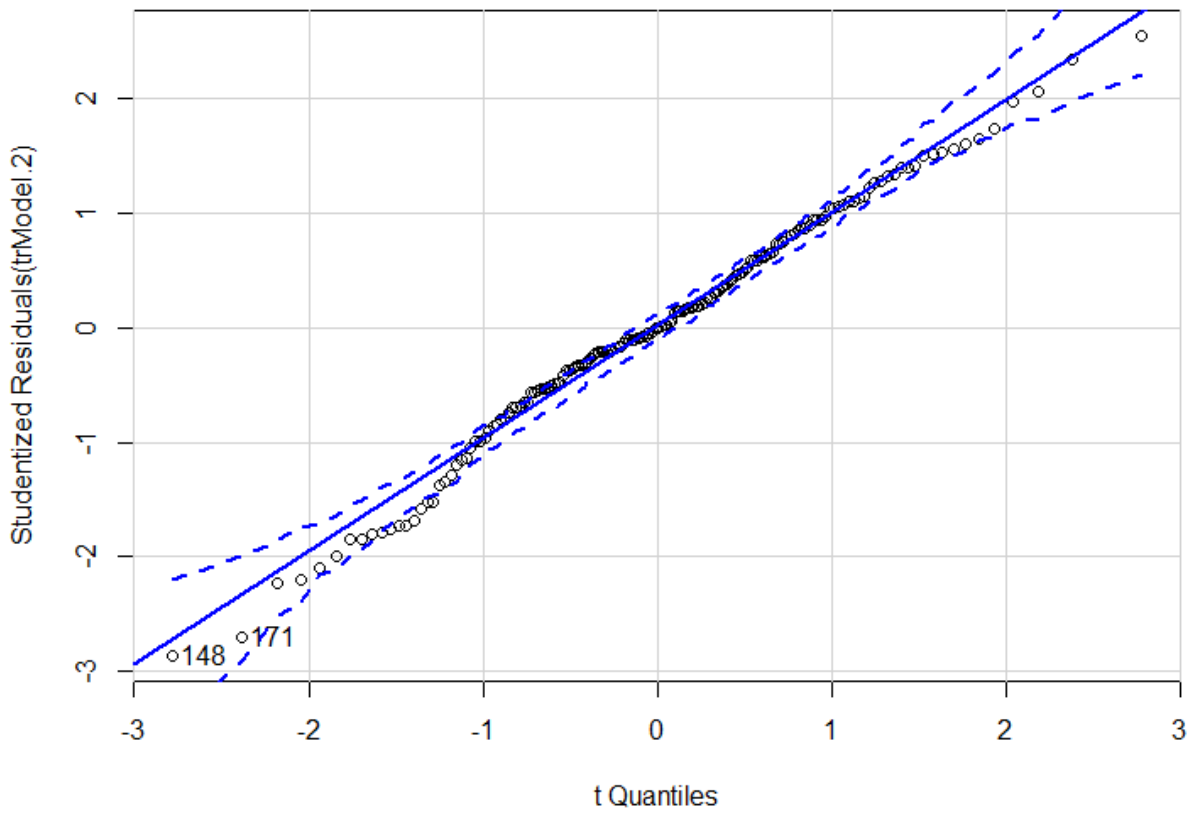
There seem to be some values, such as sample 129 (Kiribati) with -0.71 for *gghe_d*, that stand out from the rest of the distributed points. Overall, most points seem to lie around 0 in terms of their influence value. None of the samples with large deviations are considered influential, although they might be for DFFITS such as Kiribati. The explanation for it is that the country has a really large GGHE-D value with a relatively small life expectancy. This is probably because the country is in transition to improve their life expectancy in the future by providing better medical services through increasing the GGHE-D (*gghe_d*).

Testing the Normality Assumption

The following plot shows the histogram of the Studentized residuals after outliers were removed:

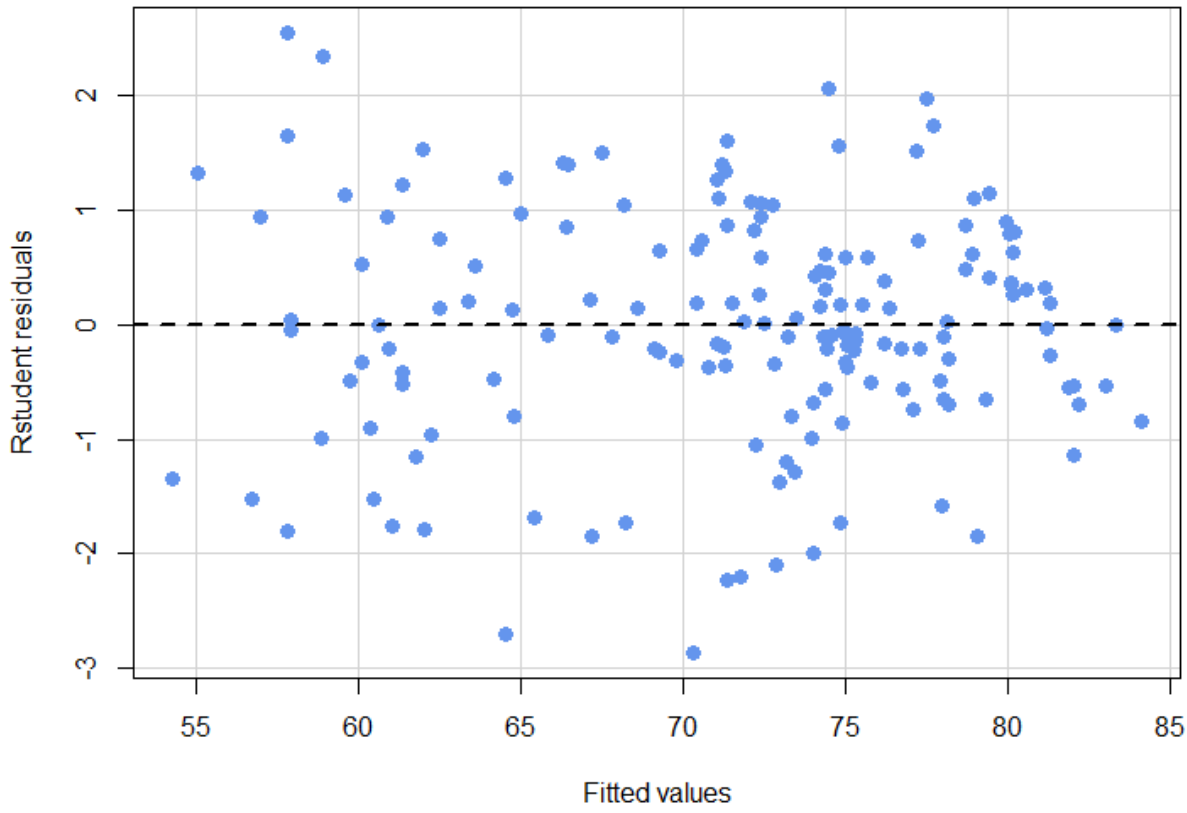


The plot is much better than it was before. All values are between 3 and -3, although it would be better if there was a little less density with the large negative residuals. Although there is larger density in the small positive residuals region compared to the small negative residuals region, the model is quite normally distributed. The QQ plot also confirms our thoughts on the normality distribution:

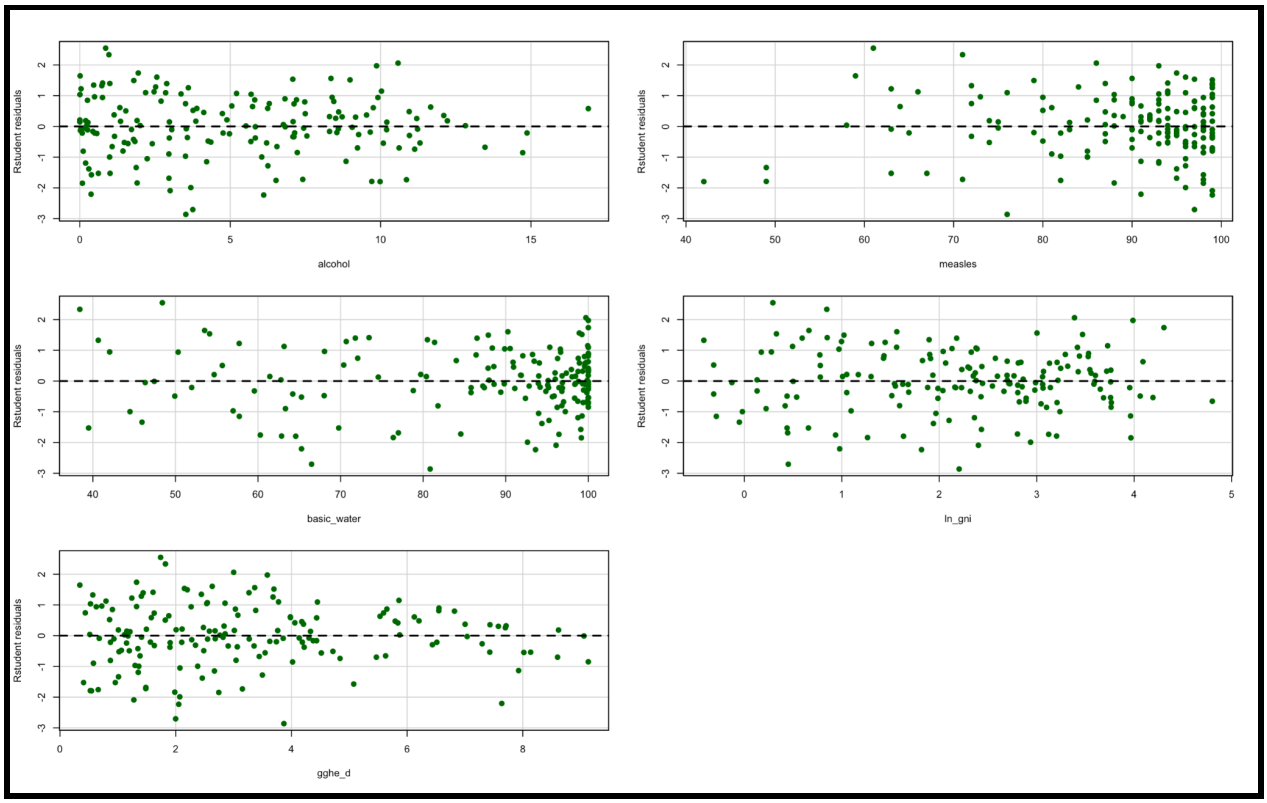


Unlike the previous QQ plot, this QQ plot fits much better. There is no evidence of any concavity, up or down, and the majority of the points lie in the acceptability range. However, we still slightly see the trend where the values towards the negative t-quantiles tend to vary a bit more. Samples 148 (Namibia) and 171 (Zimbabwe) seem to be marked for us, indicating that the model may be better off without them.

We can also inspect the constant variance assumption using the plot of residuals against the fitted values:



We see a constant variance between -3 and ~2.5 (which is near -3 to 3). There are less outliers now than before. We can now concern ourselves with the plot of residuals against the regressors:



Distribution is relatively the same, although variance seems to be more stabilized in all of the variables. No further action is needed, although investigating the removal of Namibia and Zimbabwe would help reveal if the model can be finalized.

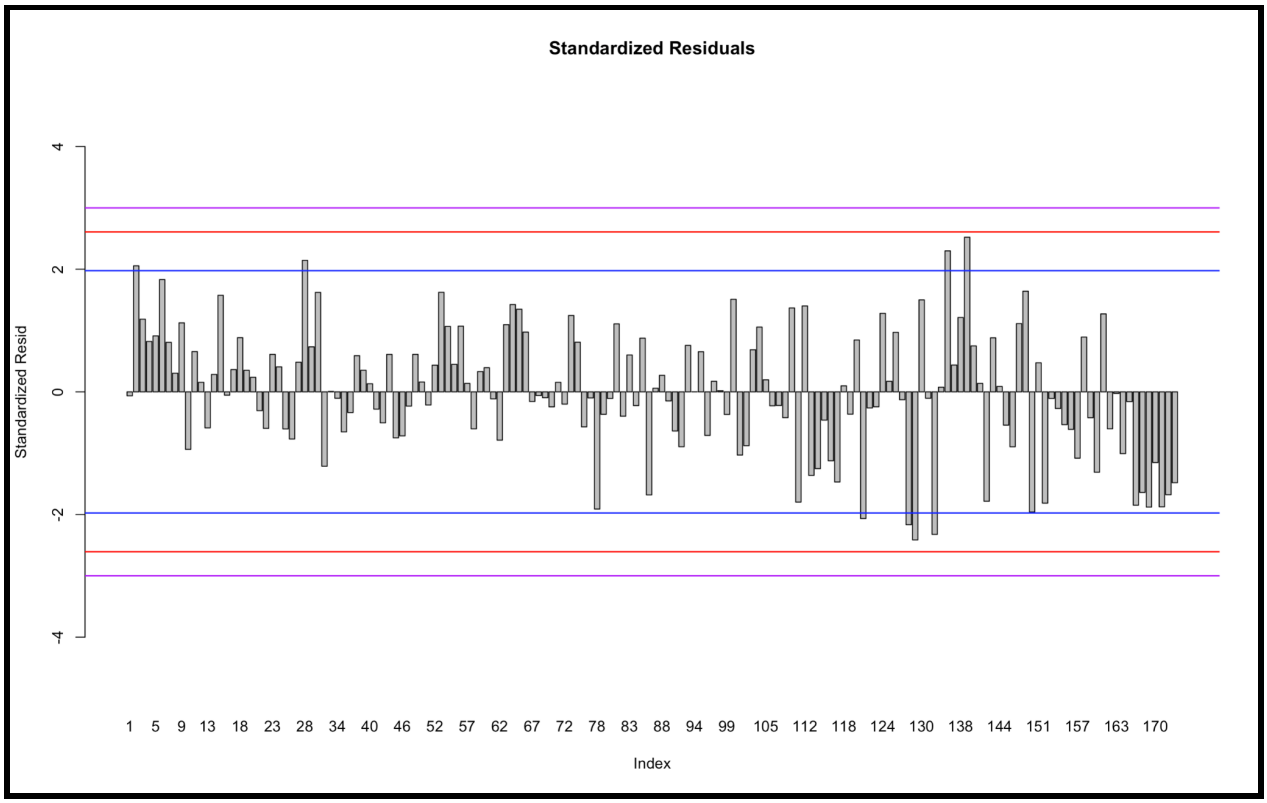
Finalizing the Model

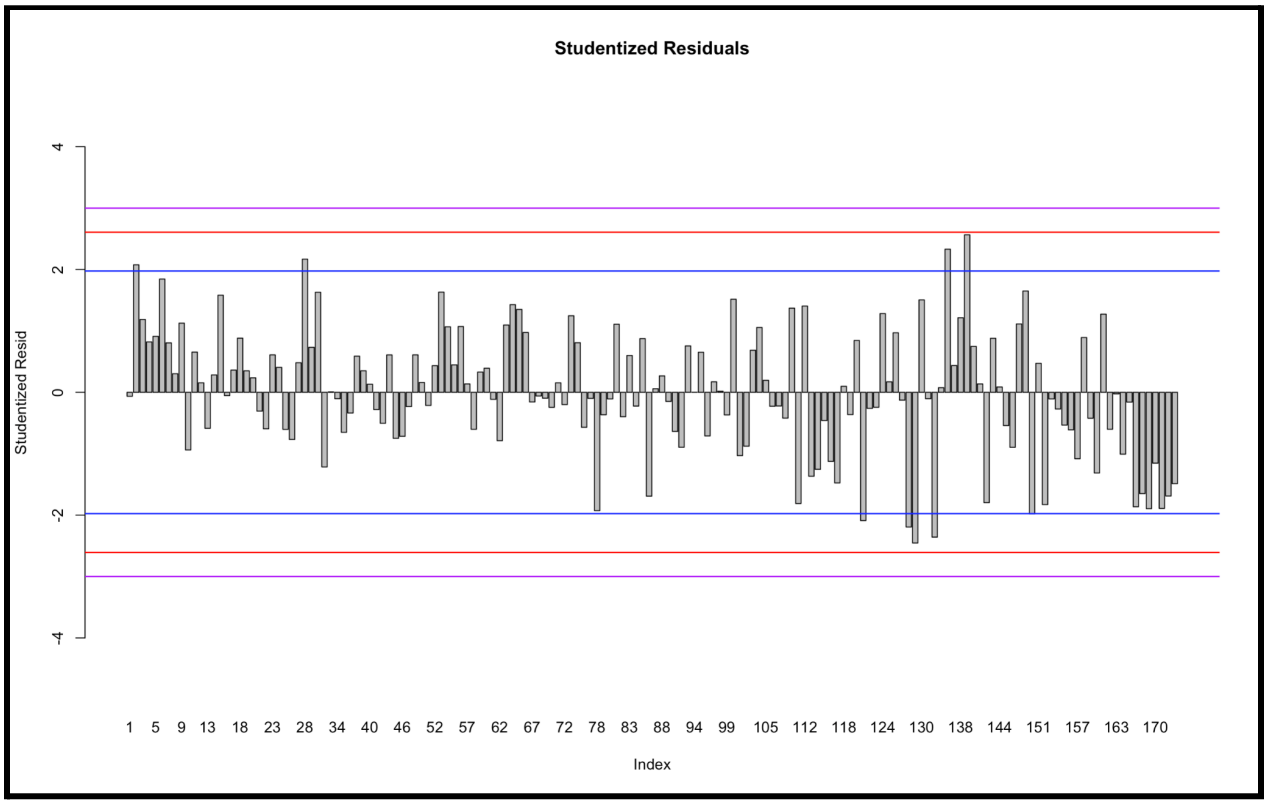
We noticed that Zimbabwe (sample 171) and Namibia (sample 148) had some large negative residuals and low COVRATIO values. They also had relatively large, although not significant, negative DFFITS value. We were curious to see if the model would improve without their existence, so we removed them and redid the variable selection process, arriving at the same set of regressors. What we generally found was that the model did improve slightly. The improved model is shown below with only the relevant residual analysis:

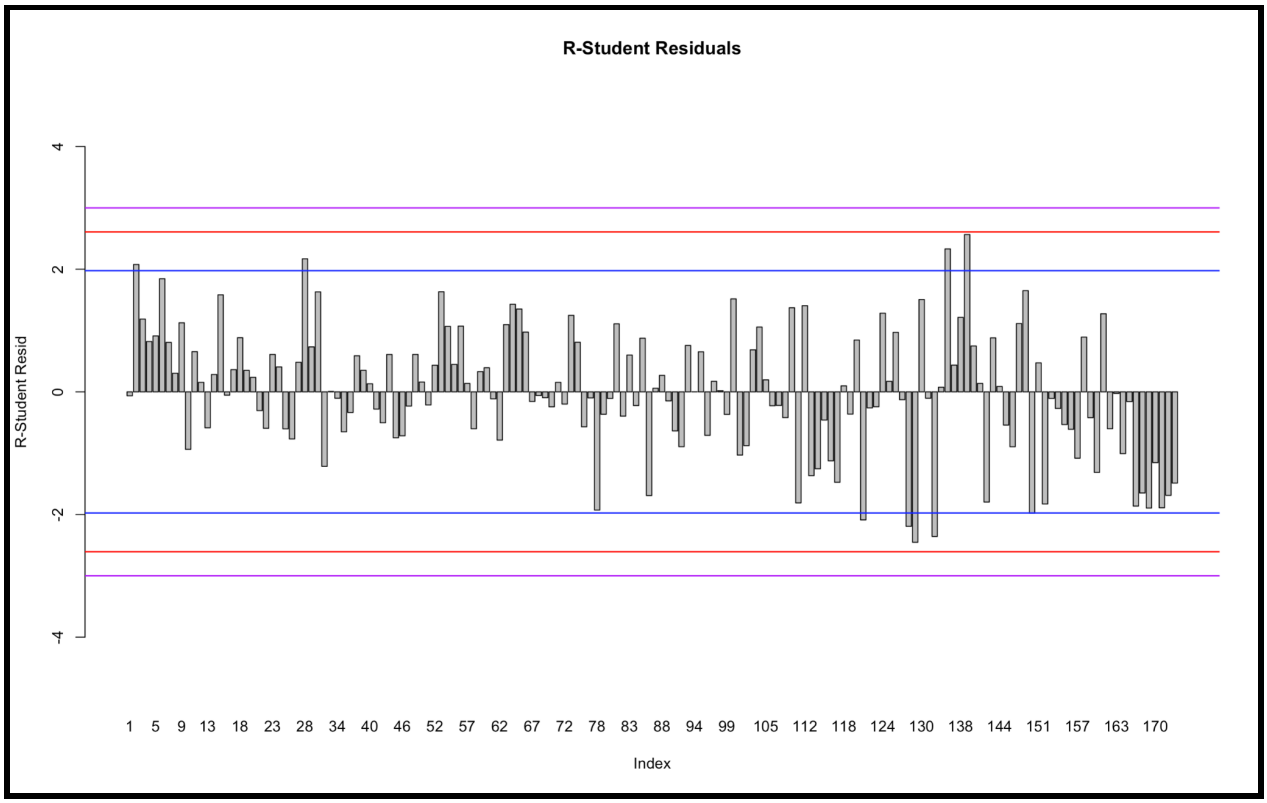
```
Call:
lm(formula = life_expectancy ~ alcohol + measles + basic_water +
    ln_gni + gghe_d, data = LEDtrans.3)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0658 -1.5477 -0.1026  1.8726  6.4780

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.81453    1.79659  23.274  < 2e-16 ***
alcohol     -0.21817    0.06576  -3.318  0.00113 **
measles      0.11604    0.02353   4.931 2.07e-06 ***
basic_water  0.13880    0.02319   5.986 1.42e-08 ***
ln_gni       2.28475    0.32866   6.952 9.28e-11 ***
gghe_d       1.11265    0.13314   8.357 3.32e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.644 on 156 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.8818,    Adjusted R-squared:  0.878
F-statistic: 232.7 on 5 and 156 DF,  p-value: < 2.2e-16
```

We can see that the RSE decreased again, our R^2increased, and the f-statistic and t-values both increased as well.

The residual plots also indicate that there were no more outliers being masked by Namibia and Zimbabwe, leaving all residuals below the 99% confidence range (around a t-value of ~2.6).


Standardized Residuals
Standardized Resid
4
2
0
-2
-4
1 5 9 13 18 23 28 34 40 46 52 57 62 67 72 78 83 88 94 99 105 112 118 124 130 138 144 151 157 163 170
Index


Studentized Residuals
Studentized Resid
4
2
0
-2
-4
1 5 9 13 18 23 28 34 40 46 52 57 62 67 72 78 83 88 94 99 105 112 118 124 130 138 144 151 157 163 170
Index

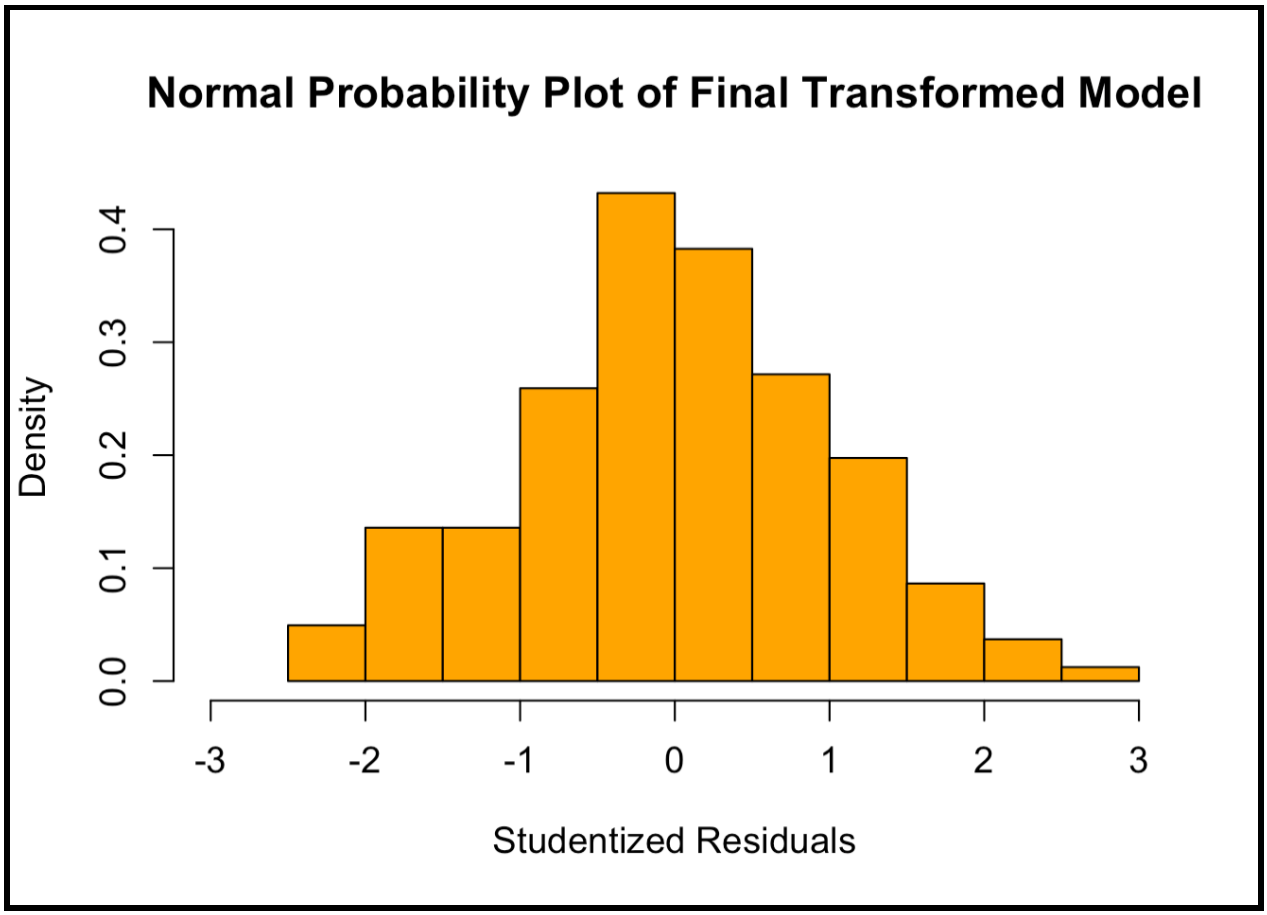


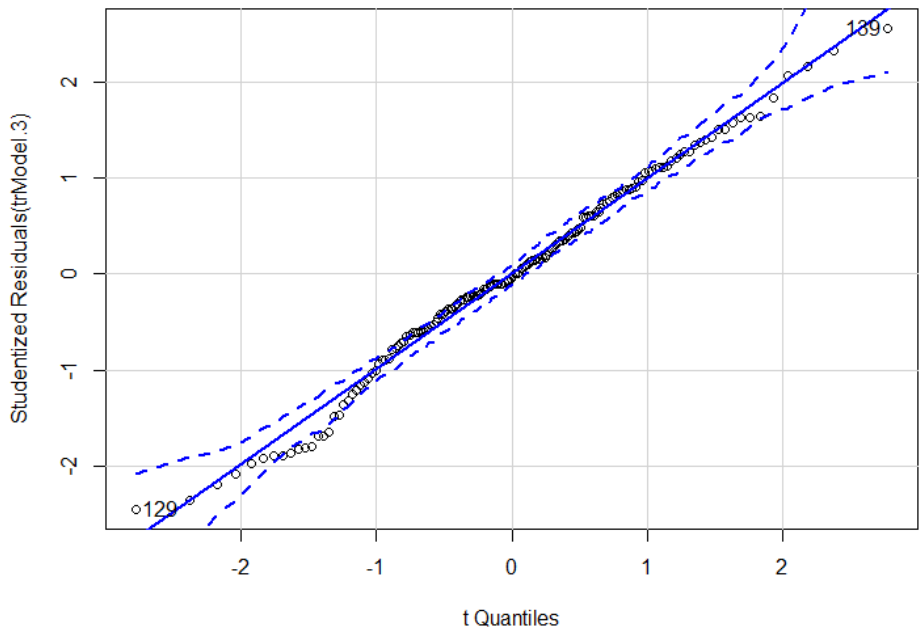
Our influence analysis also helped show that there were no samples coincident to our outliers as all influential points maintain relatively the same values for their influence measures:

```
Potentially influential observations of
        lm(formula = life_expectancy ~ alcohol + measles + basic_water +      ln_gni + gghe_d, data = LEDtrans.3) :

    dfb.1_ dfb.alch dfb.msls dfb.bsc_ dfb.ln_g dfb.ggh_ dffit   cov.r   cook.d hat
35  -0.15   0.10    -0.05     0.09    -0.21     0.11    -0.24    1.16_*  0.01   0.12_*
47  -0.05   0.03    -0.03     0.04    -0.06     0.02    -0.07    1.15_*  0.00   0.09
129  0.00   0.43    -0.08     0.27     0.14    -0.79    -0.90_*  0.94    0.13   0.12_*
131  0.00   0.01     0.03    -0.02     0.01     0.00    -0.03    1.14_*  0.00   0.09
133  0.25  -0.16    -0.16    -0.15     0.17     0.19    -0.38    0.86_*  0.02   0.03
136  0.49  -0.16     0.04    -0.57     0.33     0.13     0.69_*  0.92    0.08   0.08
139  0.29  -0.13    -0.32    -0.15    -0.01     0.22     0.62_*  0.86_*  0.06   0.06
153  0.00  -0.01    -0.02     0.03    -0.01     0.01    -0.03    1.15_*  0.00   0.09
154 -0.06  -0.02     0.02     0.05    -0.04     0.01    -0.07    1.12_*  0.00   0.07
166 -0.66  -0.21     0.44     0.24    -0.45     0.26    -0.82_*  1.09    0.11   0.16_*
172 -0.33  -0.32     0.63    -0.12    -0.01     0.14    -0.80_*  1.07    0.11   0.15_*
```

There are still relatively the same number of influential values, but overall none have large residuals along with a small COVRATIO value, so we can stop here with the removal of outliers.

Our normal probability plot also now (arguably) has a better shape to it where all values are between the 99% cutoff limits (-2.6 and +2.6), although our conclusion on the normality of the residuals hasn't changed.


Normal Probability Plot of Final Transformed Model
Density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
Studentized Residuals


Studentized Residuals(trModel.3)
-2
-1
0
1
2
t Quantiles
129
139

We now notice that the QQ plot does not call out Namibia and Zimbabwe for being outliers with new points replacing them. However, the new outliers are not nearly as influential as were our previous outliers, so this is not an issue. As for the distribution of the points, it is still relatively the same.

Any residual analysis not shown such as the DFBETAS plots, residual vs fitted value plot, etc. had no identifiable change, so it was not worth mentioning.

Interpretation of Final Model

Based on the final results of our analysis, we can say the following about our model:

- All samples deviate no more than 6.5 years from the fitted model
- 12 samples were not included in the creation of the model due to missing data in either *gni_capita* or *gghe_d*
- All outliers removed were African countries, mainly South African countries and countries with low life expectancies
- GNI per capita was better modeled when a logarithm of its values were taken
- Polio and Diphtheria vaccinations can be represented by the regression coefficient of Measles vaccinations due to extremely high multicollinearity.
- Hepatitis vaccinations and youth obesity or thinness did not have significant effects on life expectancy
- No information can be pulled out of the other variables due to unviability (missing data, definition problems, etc.)
- 88.18% of the variance in the data is explained by the linear model

Interpretation of the Regression Coefficients

- The intercept indicates that the base life expectancy of a country is 41.81453 years
- *alcohol* indicates that as recorded per capita (15+) consumption (in litres of pure alcohol) increases by 1 litre, life expectancy decreases by 0.21817 years on average while all other regressors are held constant.
- *measles* indicates that As the Measles-containing-vaccine first-dose (MCV1) immunization coverage among 1-year-olds increases by 1%, life expectancy increases by 0.11604 years on average while all other regressors are held constant.

- *basic_water* indicates that as the percent of the population using at least basic drinking-water services increases by 1%, life expectancy increases by 0.13880 years on average while all other regressors are held constant.
- *ln_gni* indicates that every time the GNI per capita is multiplied by Euler's constant (2.71828), life expectancy increases by 2.28475 years on average while all other regressors are held constant.
- *gghe_d* indicates that as the GGHE-D of a country increases by 1%, life expectancy increases by 1.11265 years on average while all other regressors are held constant.

Conclusion

Our goal was to answer the big question of how a country can improve their life expectancy. First and foremost, we had to narrow down the selection of which predictors would be best. We chose to exclude any variables with excessive amounts of missing data or variables that were similar to each other, such as BMI and thinness/obesity factors. This decision may be done oppositely in future studies where BMI is kept instead. Other variables such as region were too specific as we wanted to give advice to every country. However, as we saw that a lot of African countries were considered outliers compared to the world, it may be helpful to include the region as an indicator variable to help maintain as many countries as possible with more specific advice for each region. In addition, the mortality variables *adult_mort*, *infant_mort*, and *age1_4mort* directly predict life expectancy and would skew the model making it harder to find the effect of other predictors. However, one can create a strong model for predicting how the life expectancy of a country might change given new data regarding the mortality of that country. For example, if a country goes to war and mortality rates increase, one might use the model with the mortality variables to predict how the country's life expectancy might change.

From our final model we found that alcohol consumption, Measles vaccination (including Polio and Diphtheria), basic access to drinking water services, GNI per capita, and GGHE-D were all important factors to the multiple linear regression model against life expectancy. Seeing the importance of these variables in the model help us to formulate general advice for these countries.

We believe that by lowering alcohol consumption, a country can improve its life expectancy. This may be due to the fact that excessive

alcohol consumption can lead to fatal incidents such as drunk driving and alcohol poisoning. Furthermore, by increasing accessibility to Measles, Polio, and Hepatitis vaccinations in 1 year olds, these fatal diseases can be avoided in adulthood and prevent any outbreaks. Also, increasing the population's access to basic drinking water services may increase the life expectancy by helping the population avoid water borne diseases from polluted waters. .

Meanwhile, it is much more difficult to give advice regarding the factors of GNI per capita and GGHE-D. The income of a country is not fully under the control of its government and can be largely influenced by the political decisions of other countries as well. GGHE-D is also difficult to advise about largely because each country's situation might be different and may require funding to be allocated to other factors. For example, if a country can improve its GNI per capita by reducing the GGHE-D and spending more on other factors, then it should do so because an increased GNI per capita opens doors to improving all of the significant factors mentioned today. Advising countries regarding these matters would require a lot more knowledge in the field of political science.

We also want to mention that our advice does not apply directly to the African countries that were marked as outliers. It is highly likely that these countries, due to their large differences from the rest of the world, require analysis dedicated to them and countries with similar statistics. Our advice is best catered to other regions of the world, especially countries with higher life expectancies as they normally had little to no outliers in the model.

Reflective Process

Within our project, we set specific guidelines on choosing data which in turn created limitations. By excluding variables with substantial missing data we are limiting the scope of our advice and model regardless of importance to life expectancy. One such variable was *une_hiv*, the prevalence of HIV/AIDS, this factor seemed to have a correlation to life expectancy. In fact, intermediate models in our ongoing analysis showed that a linear model including HIV included HIV as a significant regressor. However, the lack of data prevented us from fully utilizing its information. Other examples include education and literacy variables which were not used in our model. Notably, the data visualization showed that a linear relationship might have existed between life expectancy and those variables. In addition, outliers within the model that were removed typically represented developing countries. Using

indicator variables may improve the model as mentioned before. The final limitation, as mentioned early in the report, was our use of only one year. Intermediate analysis showed us that utilizing all samples from all the years results in some variables, such as youth obesity and thinness, being significant, despite a lot of missing data.

We would like to conclude our report by saying that no model is perfect. Our model has its advantages and disadvantages. Much more analysis would be required to fully transform our correlations into causations. We wish the best of luck to those who will use our work for their future studies.

APPENDIX

Group Roles

Hamza - Programming, Preliminary Data Researcher, Analyzer, ggplot2
Luke - Programming, Data Cleaning, Visualizer, Writer
Rachel - Programming, Analyzer, Background Researcher

References

"Data Finder - Health, United States - Products." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 2 Mar. 2021,www.cdc.gov/nchs/hus/contents2019.htm?search=Life_expectancy%2C.

MMattson. "WHO National Life Expectancy." *Kaggle*, 6 Oct. 2020, www.kaggle.com/mmattson/who-national-life-expectancy.

"World Life Expectancy 1950-2021." *MacroTrends*, www.macrotrends.net/countries/WLD/world/life-expectancy.