Linear Regression Analysis of Life Expectancy

5/13/2022

SYDNEY ALEXANDER MUCK

Introduction

Life expectancy is the term we use to describe how many years a person can expect to live. You can find life expectancy calculator all over the internet. Many books have been published on longevity; a term used to describe living a long, healthy life. It is no wonder why people would be so interested in life expectancy and how they can increase their longevity, but depending on where you get your information, how life expectancy is calculated can vary based on many different factors. In this linear regression analysis, I try to find if the number of factors can reliably be reduced and still give us a linear correlation.

In this study I use a dataset found on Kaggle. This dataset was collected by Deeksha Russell and Duan Wang from the World Health Organizations public reports. This dataset collects several variables related to health from 193 countries and collects economic data from the countries. This data set reports data from the years 2000 to 2015. An explanation of each variable can be found in the directory below. In this study, I want to examine the variables, and see if the list can be trimmed down based on correlation and give us a linear relationship to predict life expectancy, creating a simpler basis for calculation.

Variable name	Data Type	Data Format	Description	Example
Country	Character	tttt	country	Afghanistan
Year	Year	YYYY	year	2015
Status	Character	tttt	developed or developing status	Developing
Life expectancy	Number	123	life expectancy in age	65
Adult Mortality	Number	123	adult mortality rates	263
infant deaths	Number	123	number of infant deaths per 1000 population	62
Alcohol	Number	1.23	alcohol consumption per capita, in litres of pure alcohol	0.01
percentage expenditure	Number	1.23	expenditure on health as a percentage of GDP per capita	71.27962362
Hepatitis B	Number	123	HepB immunization coverage among 1-year-olds (%)	65
Measles	Number	123	number of reported cases of measles per 1000 population	1154
BMI	Number	1.23	Average Body Mass Index of entire population	19.1
under-five deaths	Number	123	Number of under-five deaths per 1000 population	83
Polio	Number	123	Pol3 immunization coverage amont 1-year-olds (%)	6
Total expenditure	Number	1.23	General government expenditure on health as a percentage of total government expenditure (%)	8.16
Diphtheria	Number	123	DTP3 immunization coverage among 1-year-olds (%)	65
HIV/AIDS	Number	1.01	Deaths per 1000 live births HIV/AIDs	0.1
GDP	Number	123	Gross Domestic Product per capita (in USD)	584.25921
Population	Number	123	Population of the country	33736494
thinness 1-19 years	Number	1.23	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)	17.2
thinness 5-9 years	Number	1.23	Pervalence of thinness among children for Age 5 to 9 (%)	17.3
Income composition of resources	Number	1.23	Human Development Index in terms of income composition of rescources (Index range from 0 to 1)	0.479
Schooling	Number	1.23	Number of years of schooling	10.1

Methods and Results

First, I uploaded dataset and packages.

```
library(readr)
```

data1<- read_csv("C:/Users/Sydney/Desktop/Life Expectancy Data project.csv")

View(data1)

```
View(data1)
library(ggpubr)
library(car)
library(lmtest)
library(leaps)
library(asbio)
library(faraway)
library(data.table)
library(DT)
library(kableExtra)
library(knitr)
library(scales)
library(caret)
library(psych)
library(stats)
library(GGally)
library(MASS)
library(MLmetrics)
```

Next, I checked for missing values in the dataset.

```
NAvalues <-sapply(data1, function(x) sum(length(which(is.na(x)))))

kable(as.data.frame(NAvalues))
```

	NAvalues
Country	0
Year	0
Status	0
expectancy	10
Adult.Mortality	10
infant_deaths	0
Alcohol	194
percentage_expenditure	0
Hepatitis_B	553
Measles	0
BMI	34
U5deaths	0
Polio	19
Total_expenditure	226
Diphtheria	19
HIV_AIDS	0
GDP	448
Population	652
thinness1_19	34
thinness5_9	34
Income_composition	167
Schooling	163

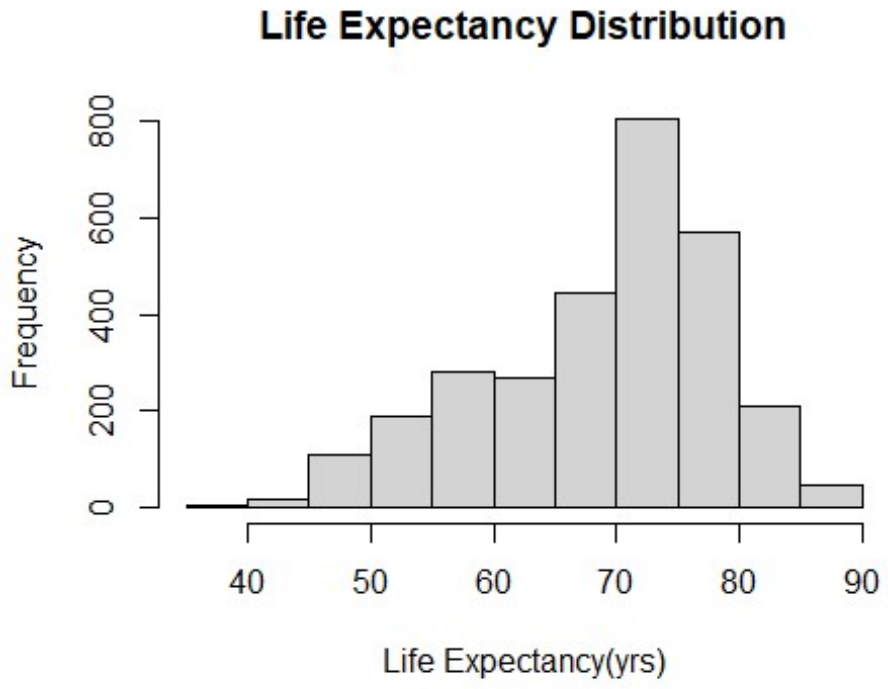
In order to get ride of missing values and still retrain the correlations, I replaced missing information with the median values. Using median values will keep from data skewing because of outliers, giving us a truer value as a representative.

```
life_med <- median(data1$expectancy,  na.rm = TRUE)
mortality_med <- median(data1$Adult.Mortality,  na.rm = TRUE)
alcohol_med<-median(data1$Alcohol,na.rm=TRUE)
BMI_med<-median(data1$BMI,na.rm=TRUE)
hepatitis_med <- median(data1$Hepatitis_B,  na.rm = TRUE)
polio_med <- median(data1$Polio,  na.rm = TRUE)
diph_med <- median(data1$Diphtheria,  na.rm = TRUE)
exp_med <- median(data1$Total_expenditure,  na.rm = TRUE)
gdp_med <- median(data1$GDP,  na.rm = TRUE)
pop_med <- median(data1$Population,  na.rm = TRUE)
thin19_med <- median(data1$thinness1_19,  na.rm = TRUE)
thin9_med <- median(data1$thinness5_9,  na.rm = TRUE)
income_med<- median(data1$Income_composition, na.rm=TRUE)
school_med <- median(data1$Schooling,  na.rm = TRUE)
```

```
#replace missing values with median
data1$expectancy[is.na(data1$expectancy)] <- life_med
data1$Adult.Mortality[is.na(data1$Adult.Mortality)] <- mortality_med
data1$Alcohol[is.na(data1$Alcohol)] <- alcohol_med
data1$BMI[is.na(data1$BMI)] <- BMI_med
data1$Hepatitis_B[is.na(data1$Hepatitis_B)] <- hepatitis_med
data1$Polio[is.na(data1$Polio)] <- polio_med
data1$Diphtheria[is.na(data1$Diphtheria)] <- diph_med
data1$Total_expenditure[is.na(data1$Total_expenditure)] <- exp_med
data1$GDP[is.na(data1$GDP)] <- gdp_med
data1$Population[is.na(data1$Population)] <- pop_med
data1$thinness1_19[is.na(data1$thinness1_19)] <- thin19_med
data1$thinness5_9[is.na(data1$thinness5_9)] <- thin9_med
data1$Income_composition[is.na(data1$Income_composition)] <- income_med
data1$Schooling[is.na(data1$Schooling)] <- school_med
```

Before making my own adjustments, I wanted to visualize the data, and see how well the data fits a linear regression model. Histogram of Life expectancy shows us the spread. Most people can expect to live between 70 and 80 years old.

```
hist(data1$expectancy,
      main = "Life Expectancy Distribution",
     xlab = "Life Expectancy(yrs)")
```



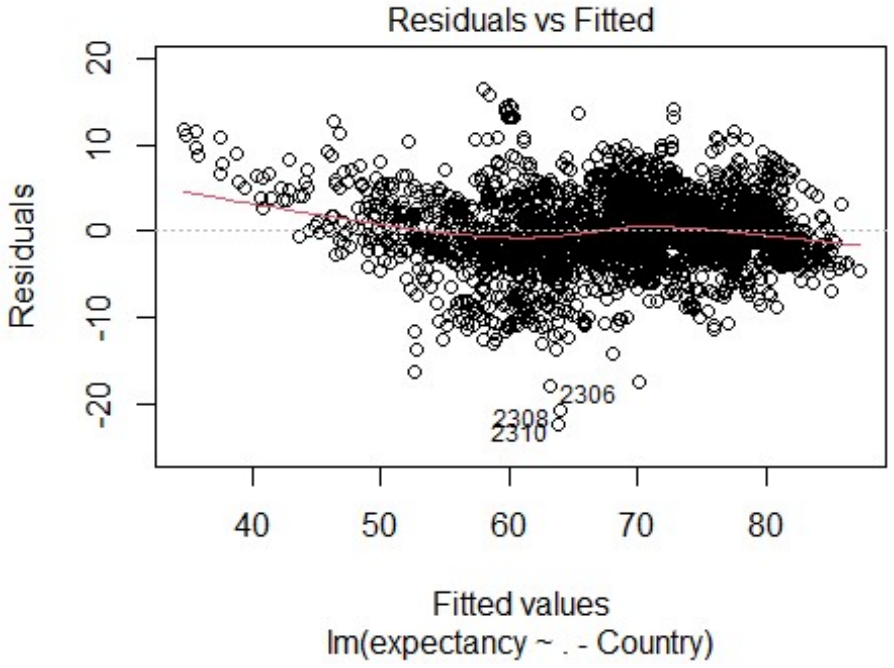
Country is a character input, and for this study it is not adding information. We will do the full model correlation with everything but Country.

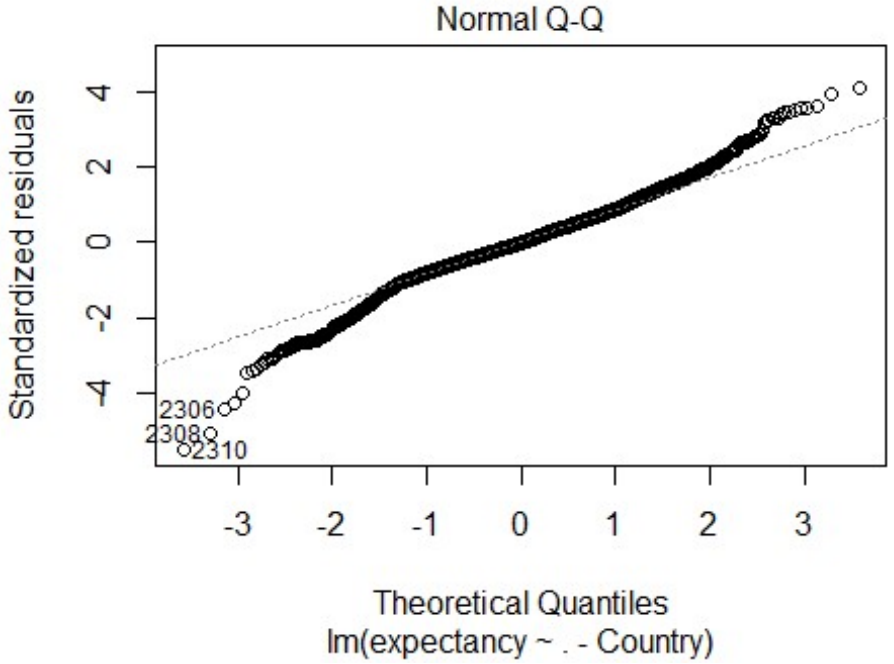
```
m1<-lm(expectancy ~. -Country,
                data = data1)
summary(m1)
```

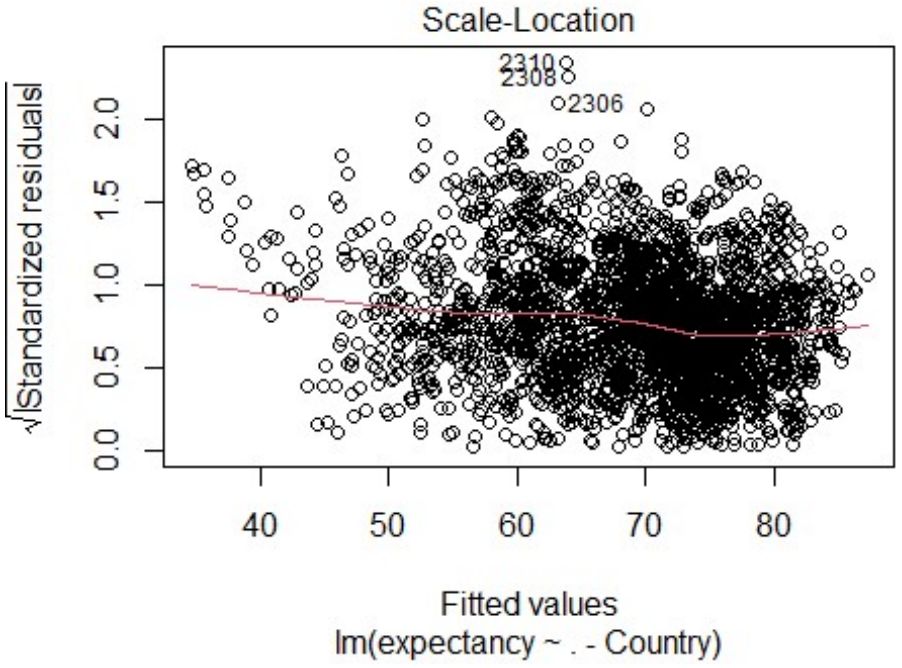
```
##
## Call:
## lm(formula = expectancy ~ . - Country, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.2661  -2.2267  -0.0901   2.3827  16.4465
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.923e+01  3.485e+01   2.274  0.02307 *
## Year                    -1.125e-02  1.742e-02  -0.646  0.51844
## StatusDeveloping        -1.601e+00  2.704e-01  -5.920 3.60e-09 ***
## Adult.Mortality         -1.990e-02  7.955e-04 -25.013  < 2e-16 ***
## infant_deaths            9.951e-02  8.441e-03  11.788  < 2e-16 ***
## Alcohol                  5.877e-02  2.622e-02   2.241  0.02508 *
## percentage_expenditure   3.409e-05  9.067e-05   0.376  0.70700
## Hepatitis_B             -1.679e-02  3.723e-03  -4.509 6.77e-06 ***
## Measles                 -1.959e-05  7.662e-06  -2.557  0.01061 *
## BMI                      4.487e-02  4.918e-03   9.123  < 2e-16 ***
## U5deaths                -7.453e-02  6.186e-03 -12.048  < 2e-16 ***
## Polio                    2.860e-02  4.453e-03   6.421 1.57e-10 ***
## Total_expenditure        7.056e-02  3.438e-02   2.052  0.04025 *
## Diphtheria               4.117e-02  4.649e-03   8.856  < 2e-16 ***
## HIV_AIDS                -4.711e-01  1.767e-02 -26.666  < 2e-16 ***
## GDP                      4.330e-05  1.381e-05   3.136  0.00173 **
## Population               8.584e-11  1.689e-09   0.051  0.95947
## thinness1_19            -8.276e-02  5.034e-02  -1.644  0.10028
## thinness5_9              1.028e-02  4.963e-02   0.207  0.83586
## Income_composition       5.569e+00  6.373e-01   8.739  < 2e-16 ***
## Schooling                6.607e-01  4.178e-02  15.814  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.05 on 2917 degrees of freedom
## Multiple R-squared:  0.8198, Adjusted R-squared:  0.8186
## F-statistic: 663.7 on 20 and 2917 DF,  p-value: < 2.2e-16
```

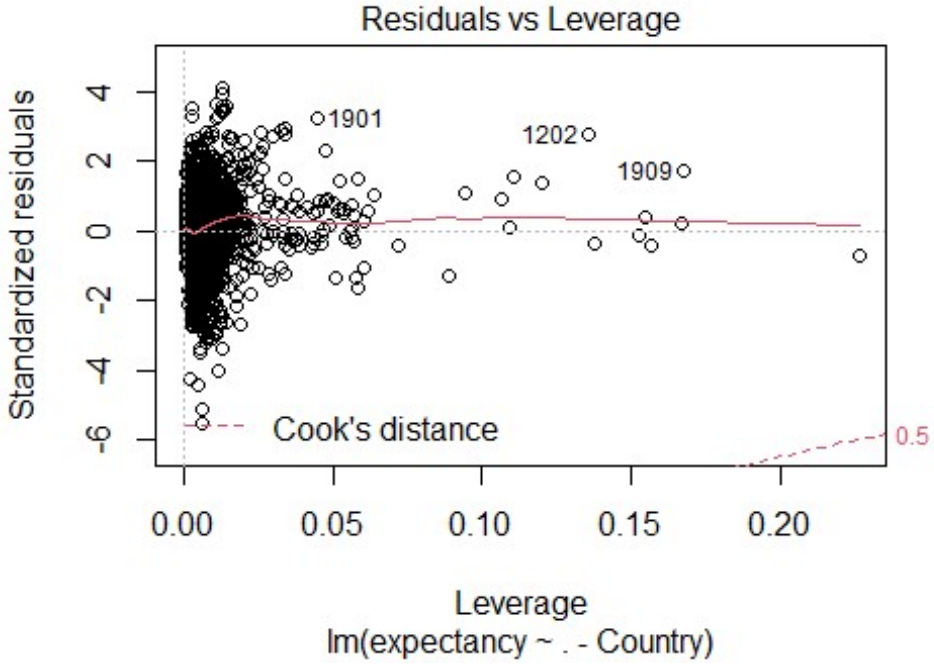
With every factor but country taken into consideration, we have an RSE of 4.05 and p-value that is less than than 2.2e-16

```
plot(m1)
```


Residuals vs Fitted
Residuals
20
10
0
-10
-20
2306
2308
2310
40
50
60
70
80
Fitted values
lm(expectancy ~ . - Country)


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
2306
2308
2310
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(expectancy ~ . - Country)



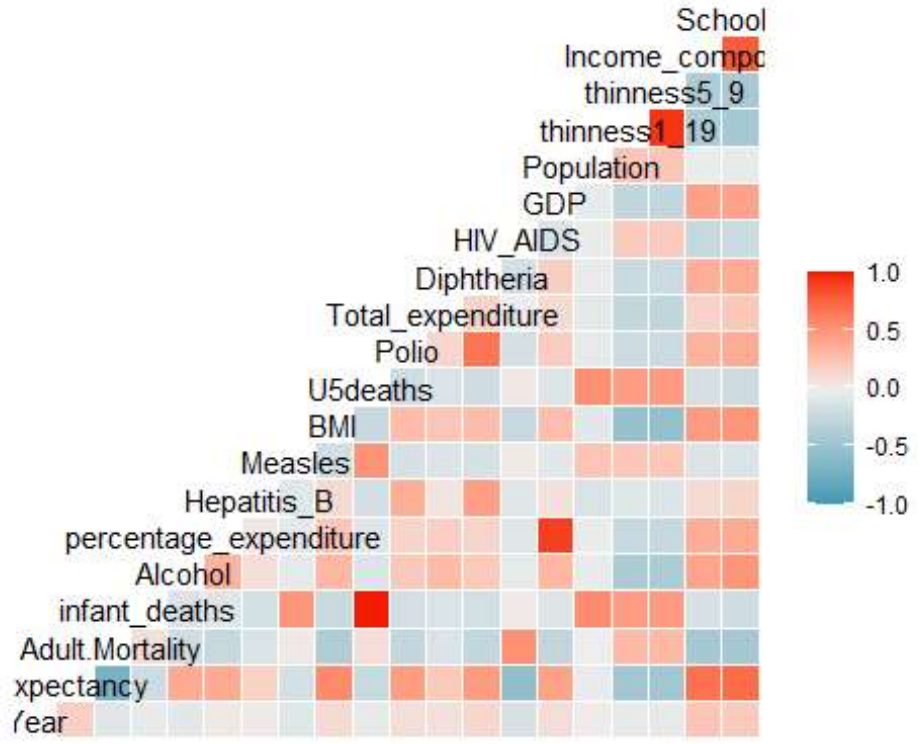


While there is noise due to outliers and the sheer amount of raw data. We can see the residuals trend around the fitted. Our data is rather normal with some fall off at the ends.

In order to determine which variables to closer into as candidates, I used a correlation to find with variables most strongly associated with life expectancy.

```
ggcorr(data1,method=c("pairwise"))
```

```
## Warning in ggcorr(data1, method = c("pairwise")): data in column(s)
'Country',
## 'Status' are not numeric and were ignored
```



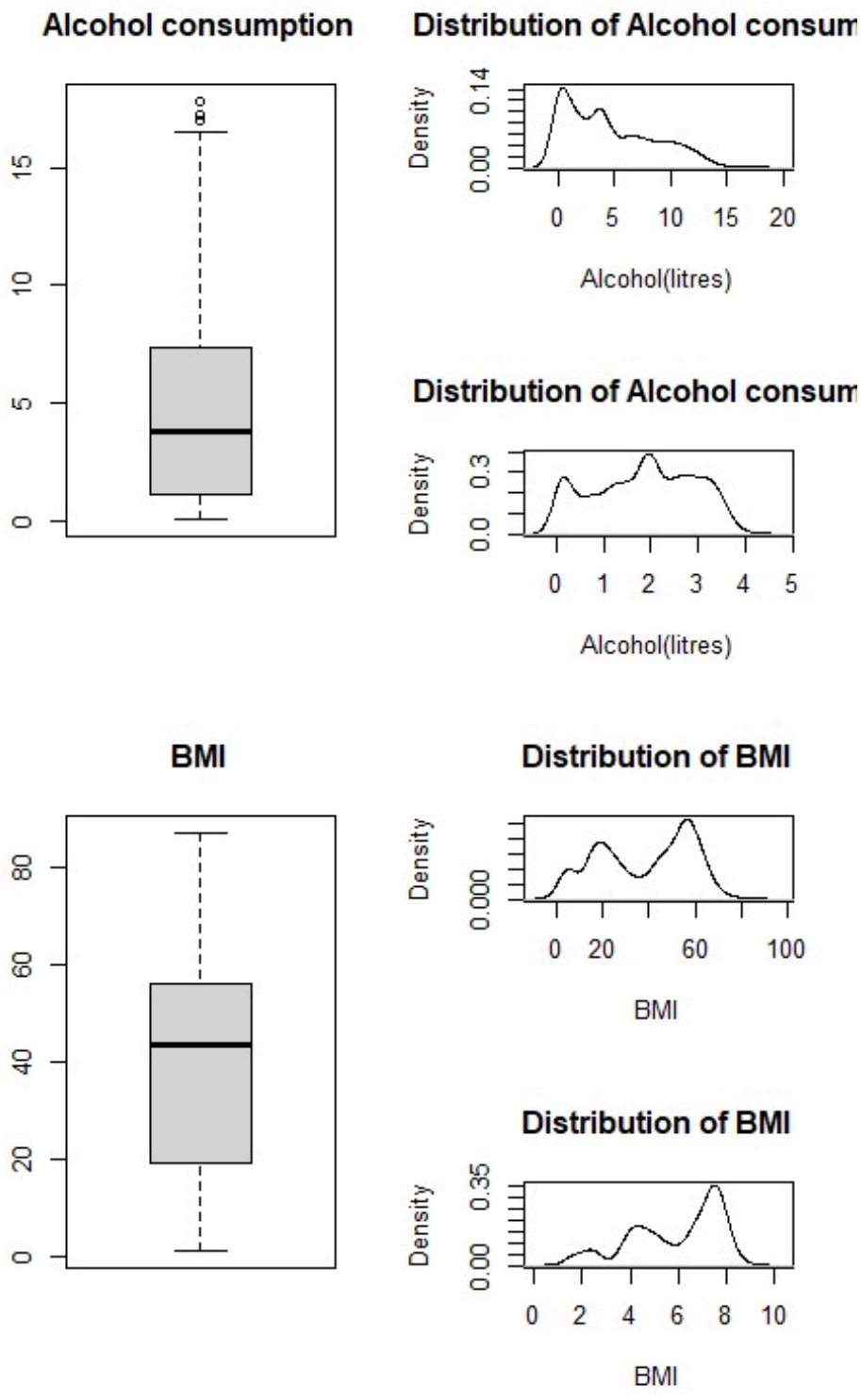
From this data, I chose to look into alcohol consumption per capita, percentage expenditure (expenditure on health as a percentage of GDP per capita), Gross Domestic Product (GDP) per capita in USD, the total expenditure (government expenditure on health as a percentage of total government expenditure), average body mass index (BMI) of entire population, human development index in terms of income composition of resources (Income composition), and number of years of schooling.

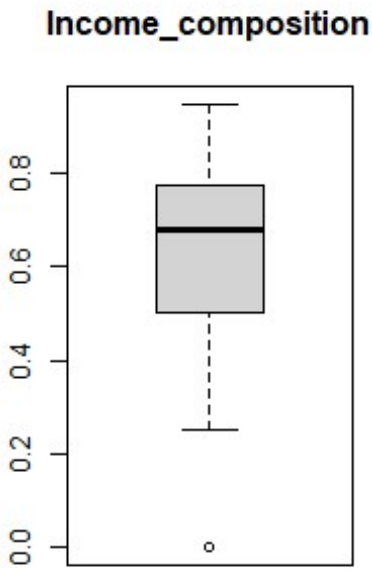
First I looked at the boxplots and distribution of each of these variables.

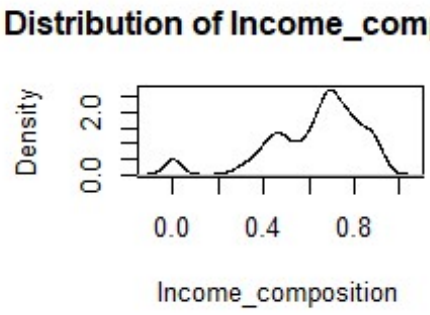
```
Example code for first factor.
```

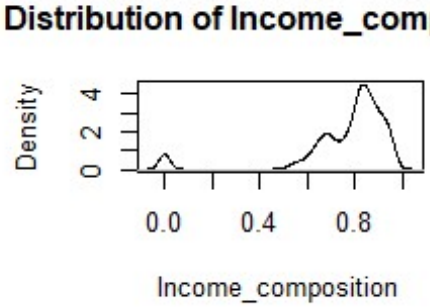
```
par(mfrow=c(2,2))
layout(matrix(c(1,1,2,3), 2, 2, byrow = F),
   widths=c(1,1), heights=c(1,1))
boxplot(data1$Alcohol,
        main = "Alcohol consumption")
plot(density(data1$Alcohol),
     main = "Distribution of Alcohol consumed",
```

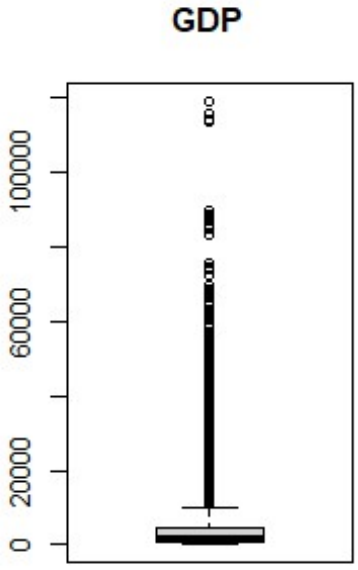
```
     xlab = "Alcohol(litres)")
plot(density(data1$Alcohol^0.5),
     main = "Distribution of Alcohol consumed",
     xlab = "Alcohol(litres)")
```

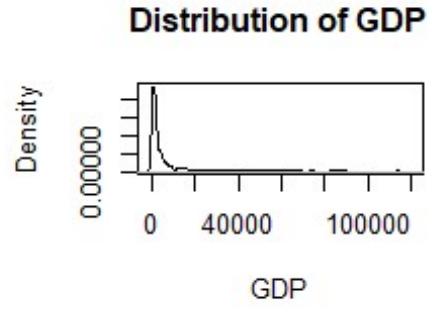


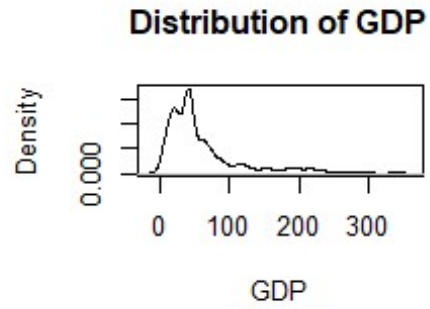

Income_composition
0.0
0.2
0.4
0.6
0.8

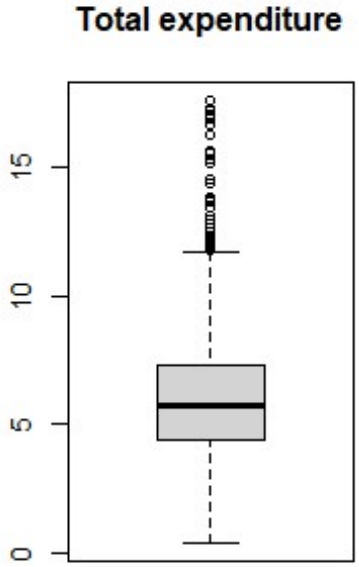

Distribution of Income_composi
Density
0.0
2.0
0.0
0.4
0.8
Income_composition

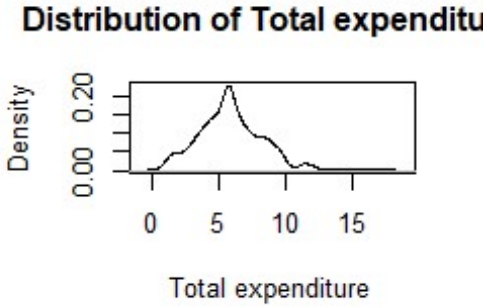

Distribution of Income_composi
Density
0
2
4
0.0
0.4
0.8
Income_composition

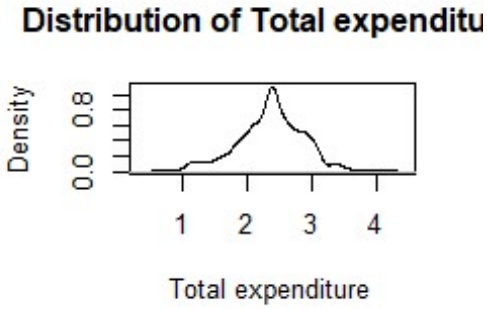

GDP
0
20000
60000
100000

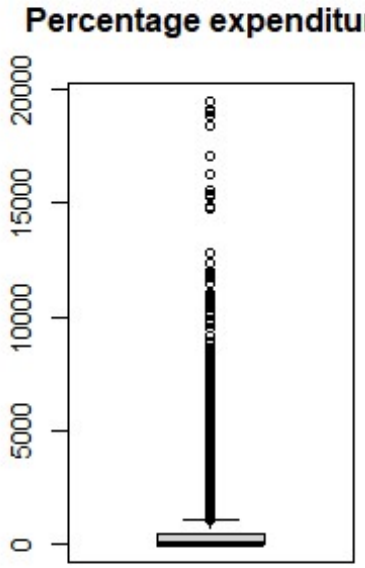

Distribution of GDP
Density
0.00000
0
40000
100000
GDP

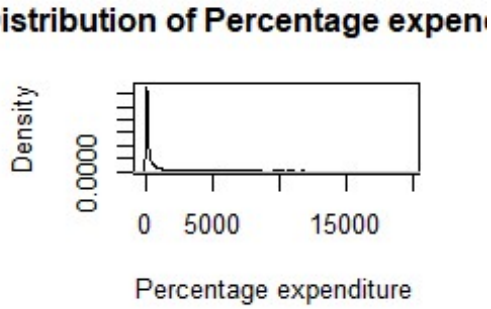

Distribution of GDP
Density
0.000
0
100
200
300
GDP

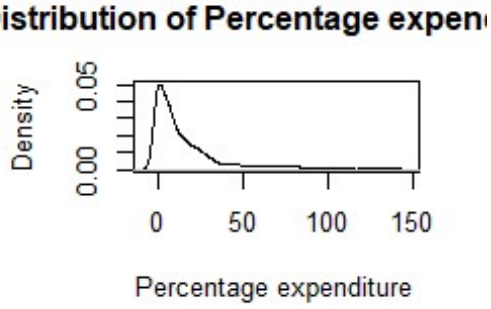

Total expenditure
0
5
10
15


Distribution of Total expenditu
Density
0.00
0.20
0
5
10
15
Total expenditure

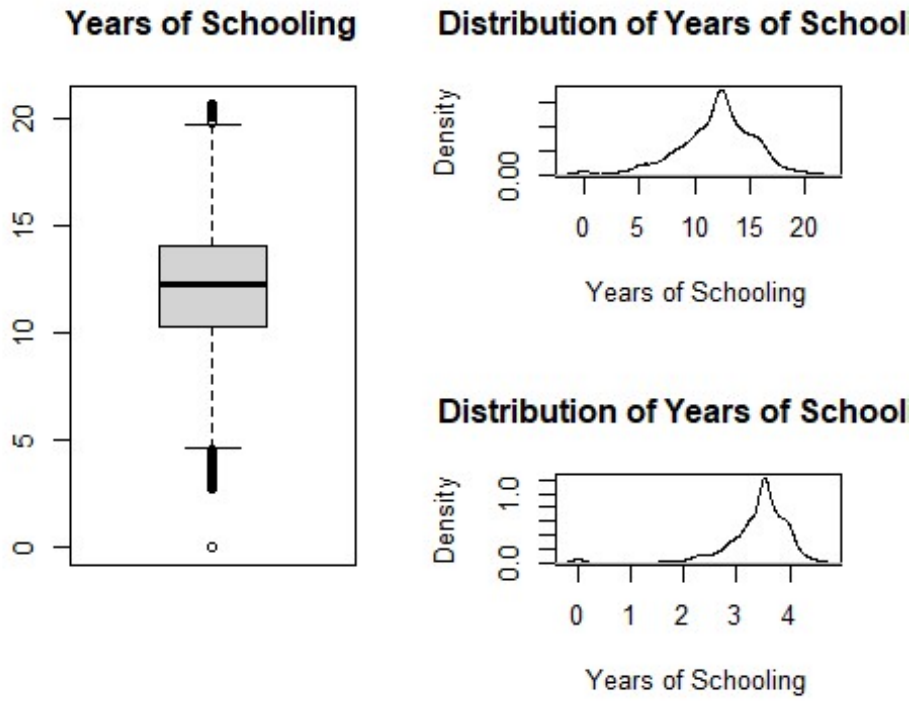

Distribution of Total expenditu
Density
0.0
0.8
1
2
3
4
Total expenditure


Percentage expenditure
0
5000
10000
15000
20000


istribution of Percentage expen
Density
0.0000
0
5000
15000
Percentage expenditure


istribution of Percentage expen
Density
0.00
0.05
0
50
100
150
Percentage expenditure

School

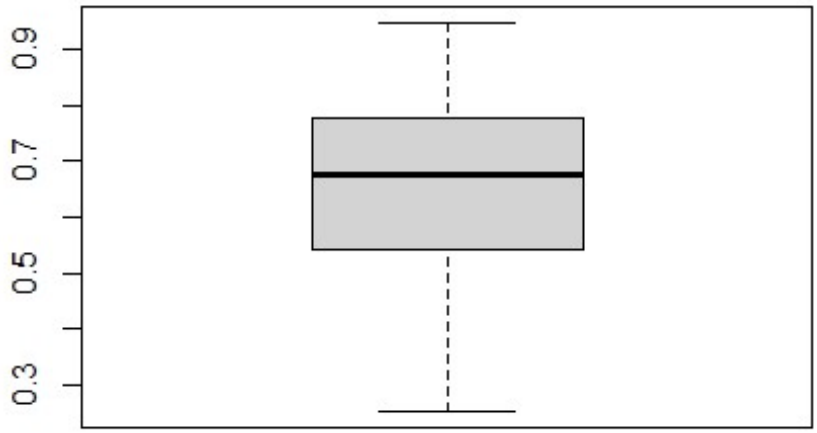


GDP and Percentage expenditure gave data with too many outliers and very skewed data and therefore I chose to drop them from the study.

Next I wanted to remove the outliers from the data points I chose to continue to look into.

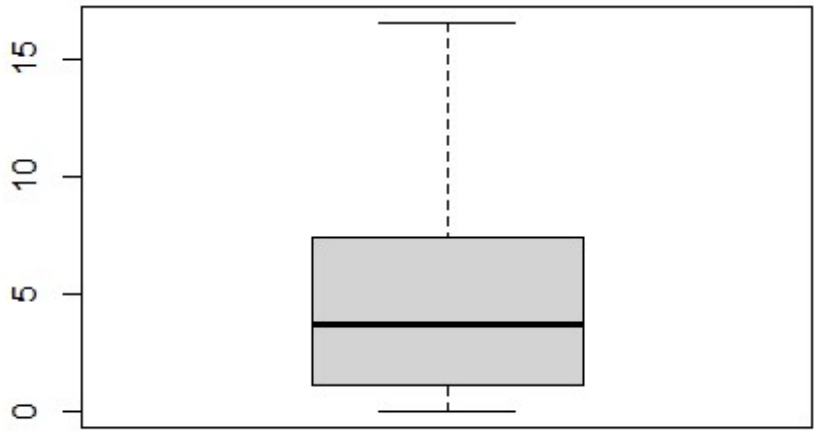
```
income_outliers <- boxplot(data1$Income_composition, plot=FALSE)$out
data2<- data1[-which(data1$Income_composition %in% income_outliers),]
```

```
boxplot(data2$Income_composition)
```



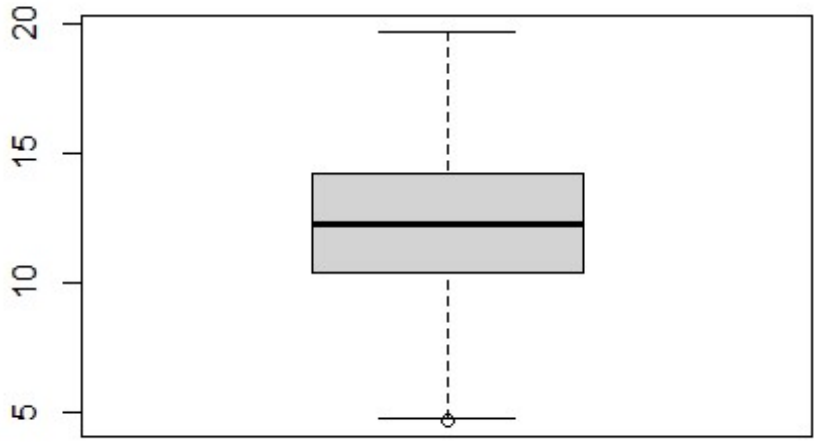

```
alcohol_outliers <- boxplot(data1$Alcohol, plot=FALSE)$out
data2<- data1[-which(data1$Alcohol %in% alcohol_outliers),]
```

```
boxplot(data2$Alcohol)
```



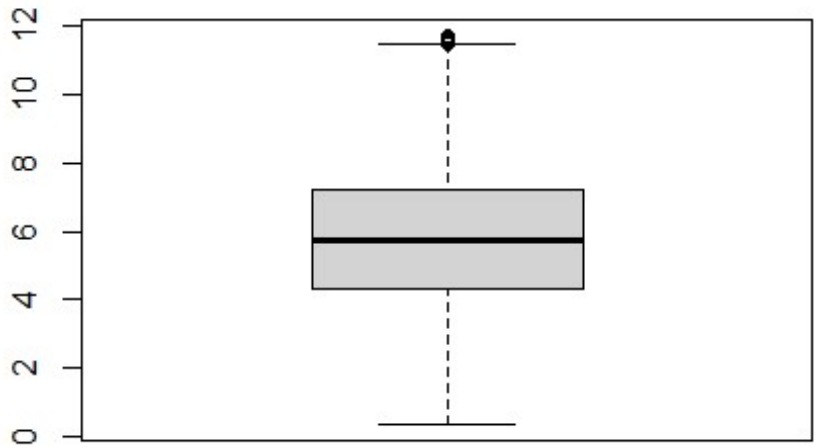

```
school_outliers <- boxplot(data1$Schooling, plot=FALSE)$out
data2<- data1[-which(data1$Schooling %in% school_outliers),]
```

```
boxplot(data2$Schooling)
```




```
expenditure_outliers <- boxplot(data1$Total_expenditure, plot=FALSE)$out
data2<- data1[-which(data1$Total_expenditure %in% expenditure_outliers),]
```

```
boxplot(data2$Total_expenditure)
```

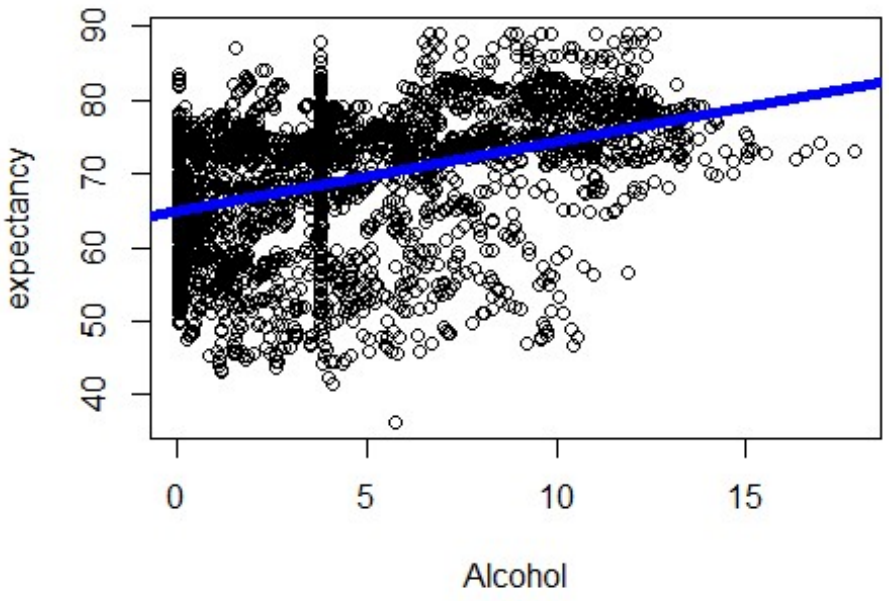


After removing outliers, I checked for each variable's linearity with life expectancy data.

```
alcohol_fit<-lm(expectancy~Alcohol, data=data2)
summary(alcohol_fit)
```

```
##
## Call:
## lm(formula = expectancy ~ Alcohol, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.084  -4.823   1.564   6.472  20.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.01562    0.24824  261.91   <2e-16 ***
## Alcohol      0.93206    0.04134   22.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.729 on 2885 degrees of freedom
## Multiple R-squared:  0.1498, Adjusted R-squared:  0.1495
## F-statistic: 508.3 on 1 and 2885 DF,  p-value: < 2.2e-16
```

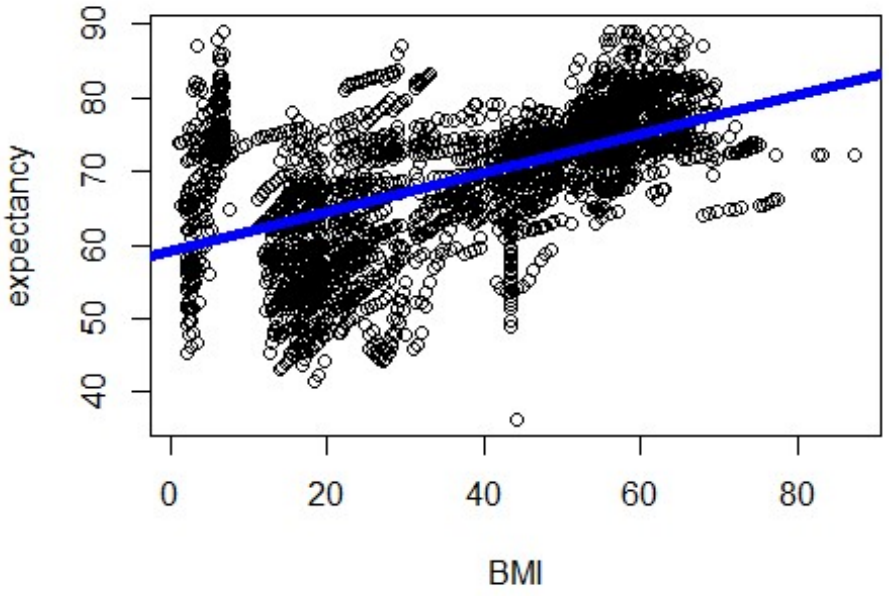
```r
plot(expectancy~Alcohol, data=data2)
abline(alcohol_fit, col='blue', lwd=5)
```




```r
BMI_fit<-lm(expectancy~BMI, data=data2)
summary(BMI_fit)
```

```
##
## Call:
## lm(formula = expectancy ~ BMI, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.556  -4.697   0.435   4.530  28.097
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.061939   0.317376  186.09   <2e-16 ***
## BMI          0.266841   0.007381   36.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.853 on 2885 degrees of freedom
## Multiple R-squared:  0.3118, Adjusted R-squared:  0.3116
## F-statistic:  1307 on 1 and 2885 DF,  p-value: < 2.2e-16
```

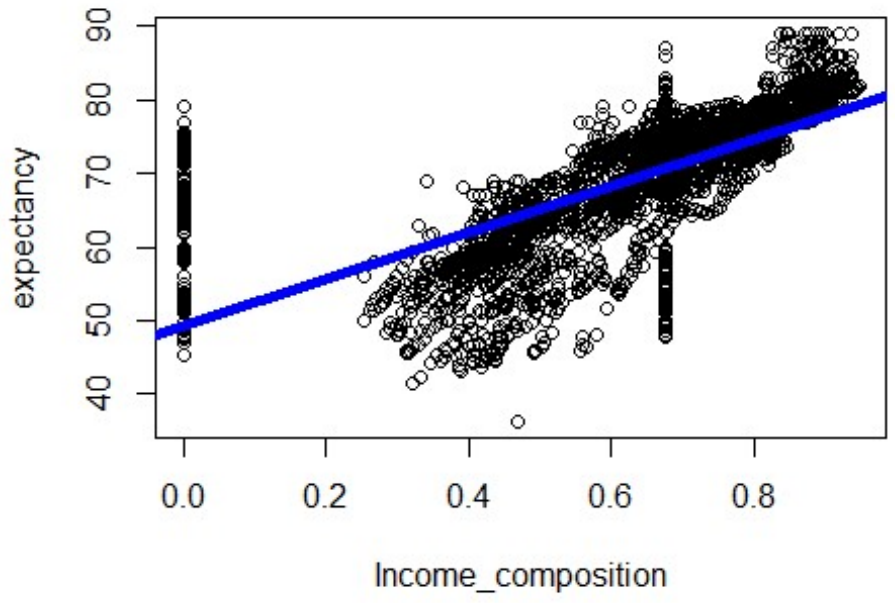
```r
plot(expectancy~BMI, data=data2)
abline(BMI_fit, col='blue', lwd=5)
```




```
income_fix<-lm(expectancy~Income_composition, data=data2)
summary(income_fix)
```

```
##
## Call:
## lm(formula = expectancy ~ Income_composition, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.879  -2.745   0.786   3.298  29.648
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         49.3518     0.4128  119.57   <2e-16 ***
## Income_composition  31.5462     0.6221   50.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.884 on 2885 degrees of freedom
## Multiple R-squared:  0.4712, Adjusted R-squared:  0.471
## F-statistic:  2571 on 1 and 2885 DF,  p-value: < 2.2e-16
```

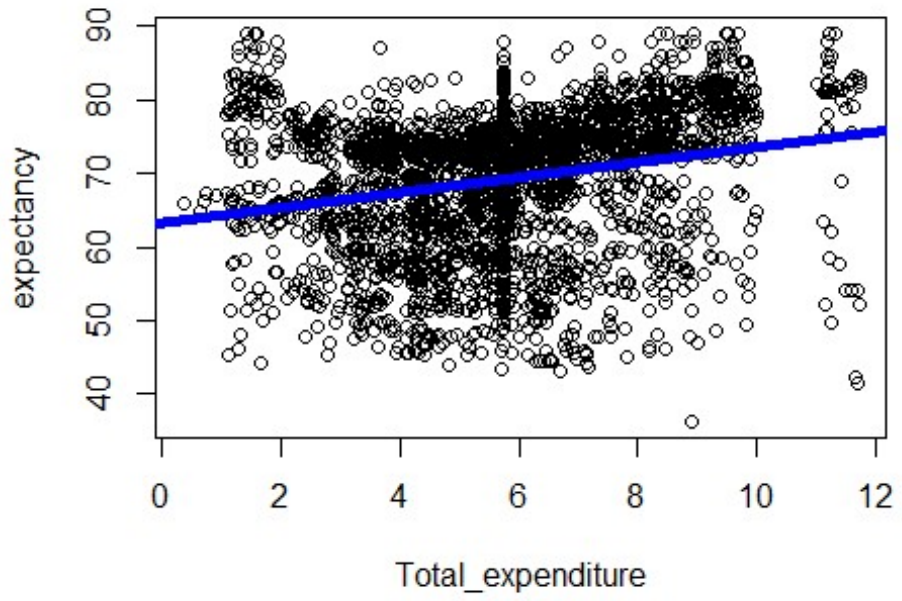
```
plot(expectancy~Income_composition, data=data2)
abline(income_fix, col='blue', lwd=5)
```




```
expenditure_fit<-lm(expectancy~Total_expenditure, data=data2)
summary(expenditure_fit)
```

```
##
## Call:
## lm(formula = expectancy ~ Total_expenditure, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.103  -5.313   2.386   6.295  24.148
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       63.38276    0.48786  129.92   <2e-16 ***
## Total_expenditure  1.01350    0.07894   12.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.207 on 2885 degrees of freedom
## Multiple R-squared:  0.05405,    Adjusted R-squared:  0.05372
## F-statistic: 164.9 on 1 and 2885 DF,  p-value: < 2.2e-16
```

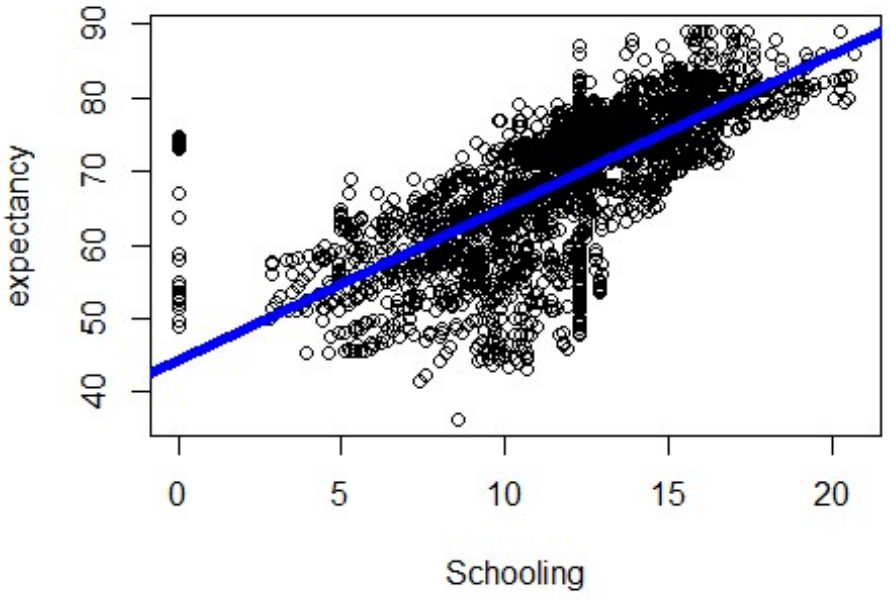
```
plot(expectancy~Total_expenditure, data=data2)
abline(expenditure_fit, col='blue', lwd=5)
```




```
school_fit<-lm(expectancy~Schooling, data=data2)
summary(school_fit)
```

```
##
## Call:
## lm(formula = expectancy ~ Schooling, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.8288  -2.8341   0.7553   4.0711  30.3503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  44.2497     0.4685   94.46   <2e-16 ***
## Schooling     2.0790     0.0376   55.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.596 on 2885 degrees of freedom
## Multiple R-squared:  0.5145, Adjusted R-squared:  0.5143
## F-statistic:  3057 on 1 and 2885 DF,  p-value: < 2.2e-16
```

```
plot(expectancy~Schooling, data=data2)
abline(school_fit, col='blue', lwd=5)
```



Before removing data that had less than desire linear correlation, I wanted to see how it correlated with the other factors.

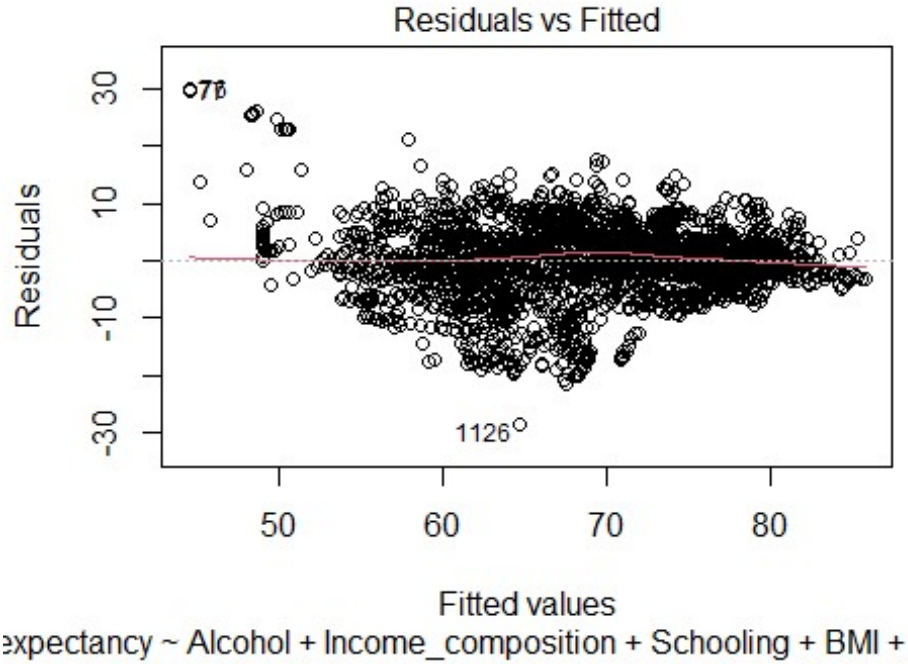
Therefore, I did a multiple linear regression with all chosen factors.

```
multi_fit<-lm(formula= expectancy~Alcohol + Income_composition + Schooling
+BMI +Total_expenditure, data=data2)
summary(multi_fit)
```

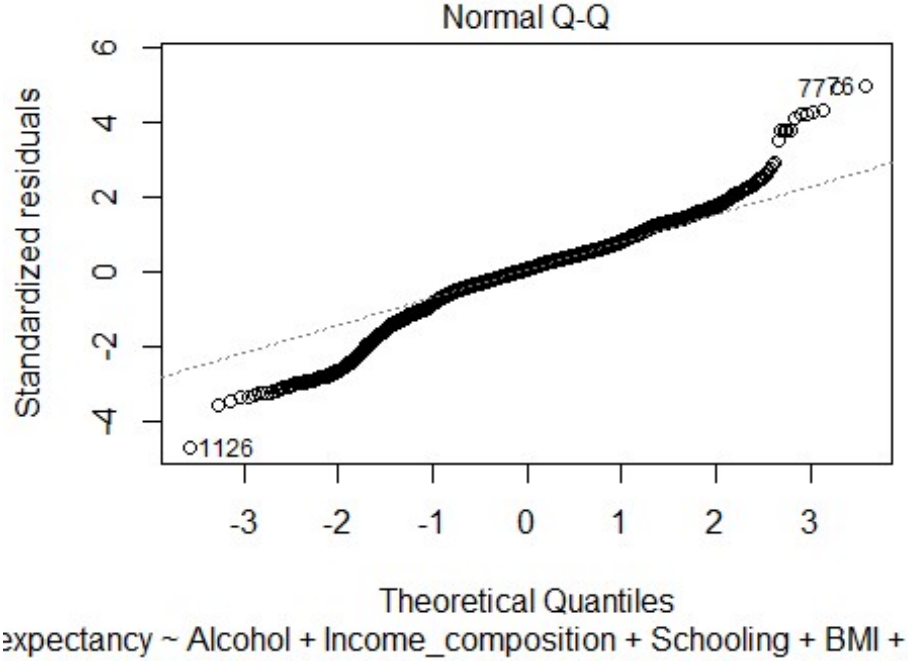
```
##
## Call:
## lm(formula = expectancy ~ Alcohol + Income_composition + Schooling +
##     BMI + Total_expenditure, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.4115  -2.6333   0.3875   3.3803  29.8837
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        43.290324   0.494412  87.559  < 2e-16 ***
## Alcohol             0.006244   0.034095   0.183  0.85470
## Income_composition 12.680151   0.919650  13.788  < 2e-16 ***
## Schooling           1.066739   0.062092  17.180  < 2e-16 ***
## BMI                 0.110297   0.006724  16.403  < 2e-16 ***
```

```
## Total_expenditure   0.154655   0.055646   2.779  0.00548 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.054 on 2881 degrees of freedom
## Multiple R-squared:  0.5915, Adjusted R-squared:  0.5908
## F-statistic: 834.5 on 5 and 2881 DF,  p-value: < 2.2e-16
```

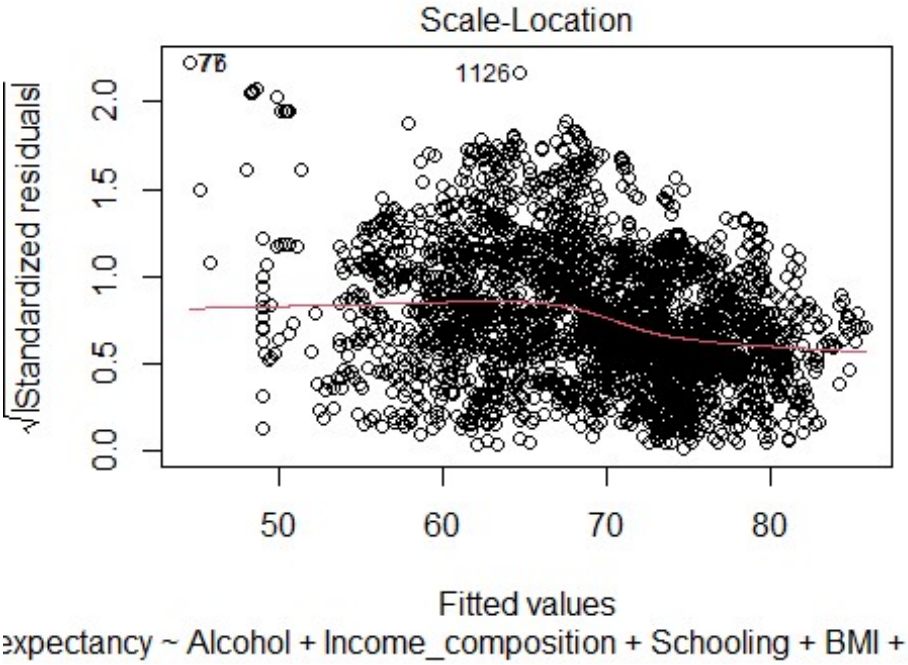
```
plot(multi_fit)
```

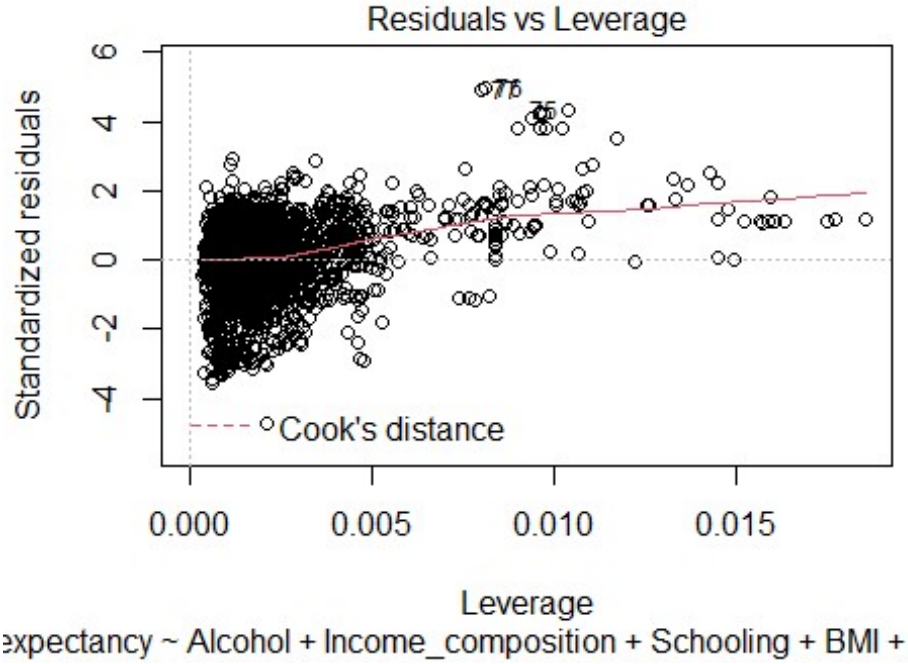


expectancy ~ Alcohol + Income_composition + Schooling + BMI + Tota



expectancy ~ Alcohol + Income_composition + Schooling + BMI + Tota





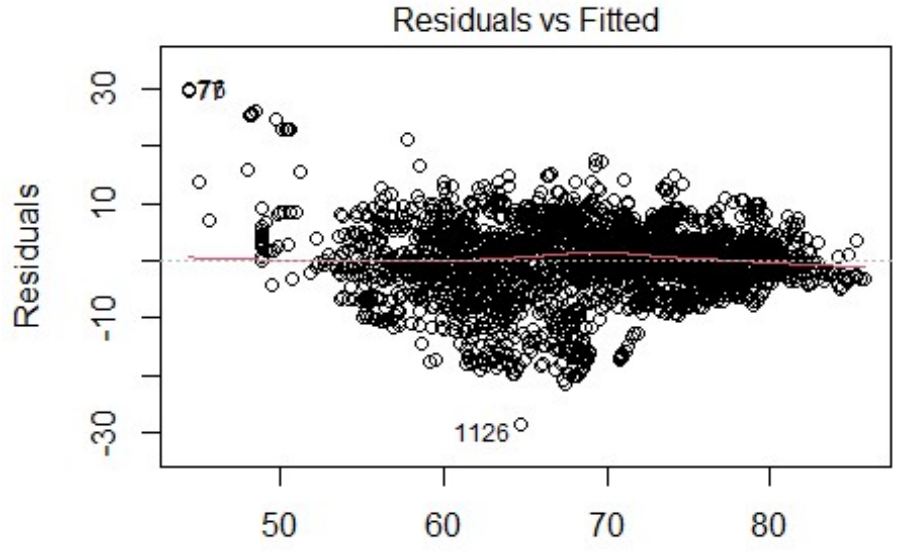
This data gained leverage, but the ends of the Q-Q became more distance from the normal.

Now I try a multiple linear regression without GDP and percent expenditure that had previously shown to have less than pleasing linear correlation.

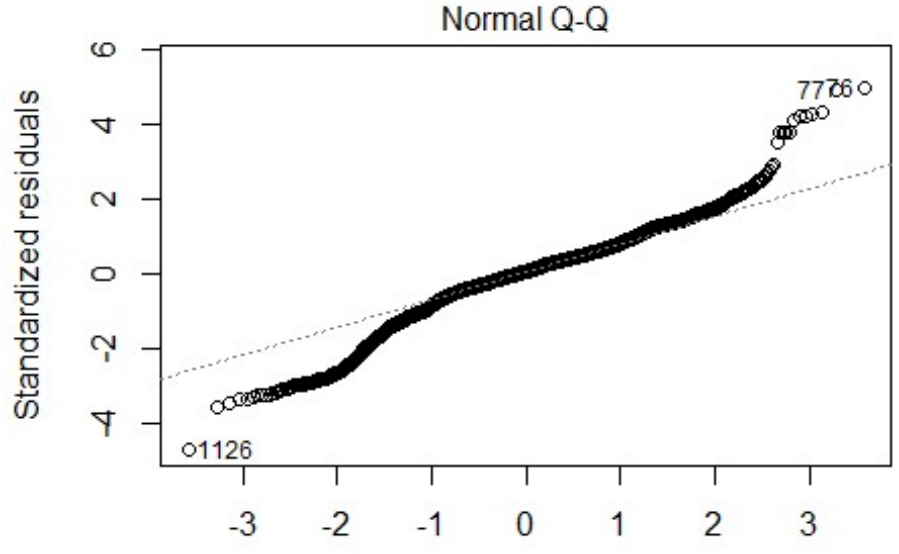
```
multi_fit2<-lm(formula= expectancy~Income_composition +
Schooling+BMI+Total_expenditure, data=data2)
summary(multi_fit2)
```

```
##
## Call:
## lm(formula = expectancy ~ Income_composition + Schooling + BMI +
##     Total_expenditure, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.4007  -2.6459   0.3928   3.3840  29.9446
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        43.264949   0.474520  91.176  < 2e-16 ***
## Income_composition 12.687681   0.918577  13.812  < 2e-16 ***
## Schooling           1.069493   0.060233  17.756  < 2e-16 ***
## BMI                 0.110368   0.006712  16.444  < 2e-16 ***
## Total_expenditure   0.156924   0.054242   2.893  0.00384 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.053 on 2882 degrees of freedom
## Multiple R-squared:  0.5915, Adjusted R-squared:  0.591
## F-statistic:  1043 on 4 and 2882 DF,  p-value: < 2.2e-16
```

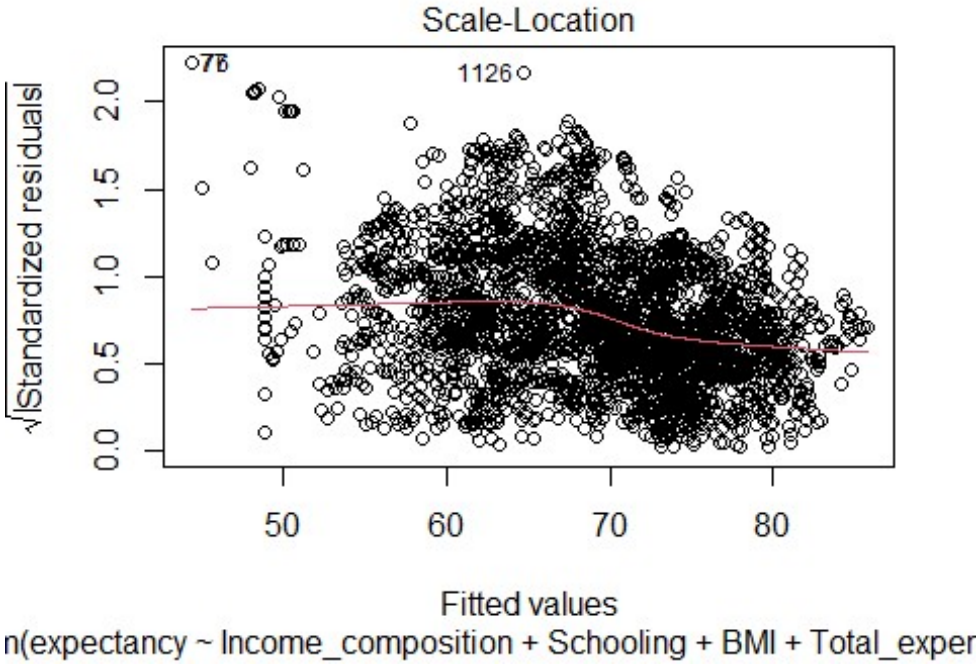
```
plot(multi_fit2)
```

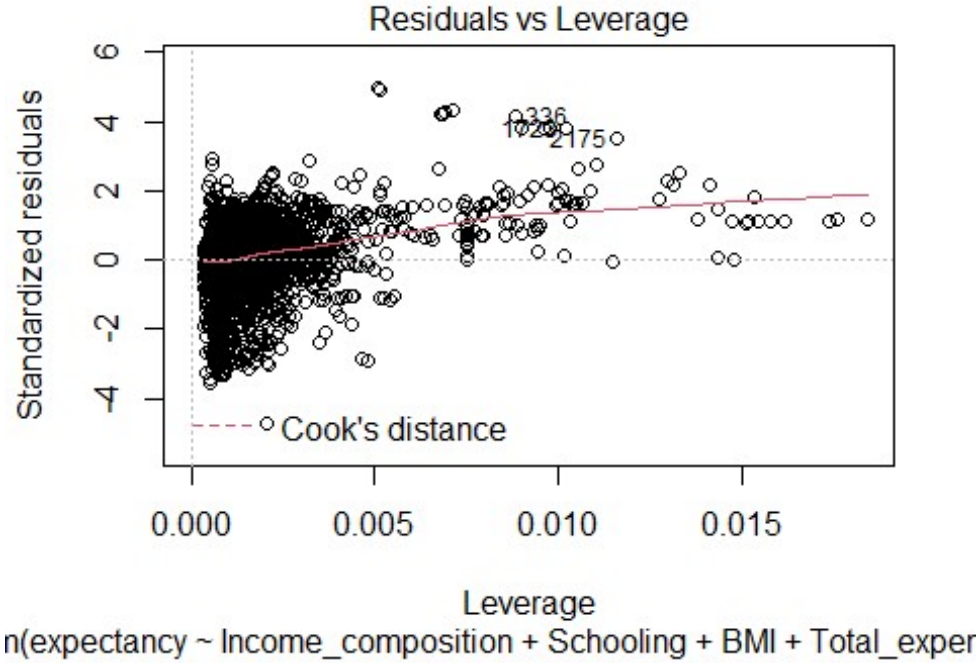


n(expectancy ~ Income_composition + Schooling + BMI + Total_exper



n(expectancy ~ Income_composition + Schooling + BMI + Total_exper





We still get similar results with those factors negated.

Conclusion and Discussion

Our original model using all the data collected by Deeksha Russell and Duan Wang from the World Health Organizations public reports gave us a Residual standard error of 4.05. The model fit with the first set of selected variables gave us an RSE of 6.054. Finally, the model with even fewer variable selected gave us an RSE of 6.053. Both of my manipulation yielded a higher RSE, and therefore, the original dataset that considers all the variables was the best fit.

In hindsight, I could have hunted down a better dataset to find a way to make a life expectancy calculator. This dataset already had a value for life expectancy which probably was a variable that already considered other factors in its own. This dataset would be interesting to visualize life expectancy across different countries and economic populations and is probably what it is intended for. I would love to investigate the implications different countries and classes and levels of education have on life expectancy using this dataset. That is what I would suggest as next steps, either a different study with this data or finding a new dataset that better correlates with the initial question.

References

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Appendix

Code will be attached separately.