Project Life Expectancy

Natallia Casey

2/11/2020

# Description of steps taken for analysis of the variables possibly influencing life expectancy.

## Checking the class of the object I opened to confirm it is a data frame.

data.frame

## Checking the dimensions of the imported data to make sure the number of rows and columns coincides with the original version.

2938 22

The dimensions are the same as those of the original data set.

## Looking at the names of the columns to make sure they are readable and look good.

[1] "Country" "Year"   
 [3] "Status" "Life.expectancy"   
 [5] "Adult.Mortality" "infant.deaths"   
 [7] "Alcohol" "percentage.expenditure"   
 [9] "Hepatitis.B" "Measles"   
[11] "BMI" "under.five.deaths"   
[13] "Polio" "Total.expenditure"   
[15] "Diphtheria" "HIV.AIDS"   
[17] "GDP" "Population"   
[19] "thinness..1.19.years" "thinness.5.9.years"   
[21] "Income.composition.of.resources" "Schooling"

I would like all of the names to start with **capital letter**, and also remove the **extra period** for the name “thinness..1.19.years”.

**Updated names of the columns**

[1] "Country" "Year"   
 [3] "Status" "Life.expectancy"   
 [5] "Adult.Mortality" "Infant.deaths"   
 [7] "Alcohol" "Percentage.expenditure"   
 [9] "Hepatitis.B" "Measles"   
[11] "BMI" "Under.five.deaths"   
[13] "Polio" "Total.expenditure"   
[15] "Diphtheria" "HIV.AIDS"   
[17] "GDP" "Population"   
[19] "Thinness.1.19.years" "Thinness.5.9.years"   
[21] "Income.composition.of.resources" "Schooling"

## Looking at the data type information

'data.frame': 2938 obs. of 22 variables:  
 $ Country : Factor w/ 193 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...  
 $ Year : int 2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...  
 $ Status : Factor w/ 2 levels "Developed","Developing": 2 2 2 2 2 2 2 2 2 2 ...  
 $ Life.expectancy : num 65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...  
 $ Adult.Mortality : int 263 271 268 272 275 279 281 287 295 295 ...  
 $ infant.deaths : int 62 64 66 69 71 74 77 80 82 84 ...  
 $ Alcohol : num 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...  
 $ percentage.expenditure : num 71.3 73.5 73.2 78.2 7.1 ...  
 $ Hepatitis.B : int 65 62 64 67 68 66 63 64 63 64 ...  
 $ Measles : int 1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...  
 $ BMI : num 19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...  
 $ under.five.deaths : int 83 86 89 93 97 102 106 110 113 116 ...  
 $ Polio : int 6 58 62 67 68 66 63 64 63 58 ...  
 $ Total.expenditure : num 8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...  
 $ Diphtheria : int 65 62 64 67 68 66 63 64 63 58 ...  
 $ HIV.AIDS : num 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...  
 $ GDP : num 584.3 612.7 631.7 670 63.5 ...  
 $ Population : num 33736494 327582 31731688 3696958 2978599 ...  
 $ thinness..1.19.years : num 17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...  
 $ thinness.5.9.years : num 17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...  
 $ Income.composition.of.resources: num 0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...  
 $ Schooling : num 10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...

The variables of the data set are mostly of **numeric**, **integer**, and **factor** types.

## Summery statistics and missing values for each of the variables

Country Year Status Life.expectancy  
 Afghanistan : 16 Min. :2000 Developed : 512 Min. :36.30   
 Albania : 16 1st Qu.:2004 Developing:2426 1st Qu.:63.10   
 Algeria : 16 Median :2008 Median :72.10   
 Angola : 16 Mean :2008 Mean :69.22   
 Antigua and Barbuda: 16 3rd Qu.:2012 3rd Qu.:75.70   
 Argentina : 16 Max. :2015 Max. :89.00   
 (Other) :2842 NA's :10   
 Adult.Mortality infant.deaths Alcohol percentage.expenditure  
 Min. : 1.0 Min. : 0.0 Min. : 0.0100 Min. : 0.000   
 1st Qu.: 74.0 1st Qu.: 0.0 1st Qu.: 0.8775 1st Qu.: 4.685   
 Median :144.0 Median : 3.0 Median : 3.7550 Median : 64.913   
 Mean :164.8 Mean : 30.3 Mean : 4.6029 Mean : 738.251   
 3rd Qu.:228.0 3rd Qu.: 22.0 3rd Qu.: 7.7025 3rd Qu.: 441.534   
 Max. :723.0 Max. :1800.0 Max. :17.8700 Max. :19479.912   
 NA's :10 NA's :194   
 Hepatitis.B Measles BMI under.five.deaths  
 Min. : 1.00 Min. : 0.0 Min. : 1.00 Min. : 0.00   
 1st Qu.:77.00 1st Qu.: 0.0 1st Qu.:19.30 1st Qu.: 0.00   
 Median :92.00 Median : 17.0 Median :43.50 Median : 4.00   
 Mean :80.94 Mean : 2419.6 Mean :38.32 Mean : 42.04   
 3rd Qu.:97.00 3rd Qu.: 360.2 3rd Qu.:56.20 3rd Qu.: 28.00   
 Max. :99.00 Max. :212183.0 Max. :87.30 Max. :2500.00   
 NA's :553 NA's :34   
 Polio Total.expenditure Diphtheria HIV.AIDS   
 Min. : 3.00 Min. : 0.370 Min. : 2.00 Min. : 0.100   
 1st Qu.:78.00 1st Qu.: 4.260 1st Qu.:78.00 1st Qu.: 0.100   
 Median :93.00 Median : 5.755 Median :93.00 Median : 0.100   
 Mean :82.55 Mean : 5.938 Mean :82.32 Mean : 1.742   
 3rd Qu.:97.00 3rd Qu.: 7.492 3rd Qu.:97.00 3rd Qu.: 0.800   
 Max. :99.00 Max. :17.600 Max. :99.00 Max. :50.600   
 NA's :19 NA's :226 NA's :19   
 GDP Population thinness..1.19.years  
 Min. : 1.68 Min. :3.400e+01 Min. : 0.10   
 1st Qu.: 463.94 1st Qu.:1.958e+05 1st Qu.: 1.60   
 Median : 1766.95 Median :1.387e+06 Median : 3.30   
 Mean : 7483.16 Mean :1.275e+07 Mean : 4.84   
 3rd Qu.: 5910.81 3rd Qu.:7.420e+06 3rd Qu.: 7.20   
 Max. :119172.74 Max. :1.294e+09 Max. :27.70   
 NA's :448 NA's :652 NA's :34   
 thinness.5.9.years Income.composition.of.resources Schooling   
 Min. : 0.10 Min. :0.0000 Min. : 0.00   
 1st Qu.: 1.50 1st Qu.:0.4930 1st Qu.:10.10   
 Median : 3.30 Median :0.6770 Median :12.30   
 Mean : 4.87 Mean :0.6276 Mean :11.99   
 3rd Qu.: 7.20 3rd Qu.:0.7790 3rd Qu.:14.30   
 Max. :28.60 Max. :0.9480 Max. :20.70   
 NA's :34 NA's :167 NA's :163

Since the **questions I originally formulated** include only a few variables, I will be focusing my analysis on these 6 variables.

**Number of missing values for Life.expectancy:**

NA's :10

**Number of missing values for Alcohol:**

NA's :194

**Number of missing values for BMI:**

NA's :34

**Number of missing values for Total.expenditure:**

NA's :226

**Number of missing values for Schooling:**

NA's :163

**Number of missing values for Population:**

NA's :652

## Looking at the variable Population closer

The number of NA is 652, which is high. Also, if we print out a few observations we can see that the values of population for different years for the same country differ greatly. At times an 8-digit number for one year and a 6-digit number for another. There are even more extreme differences present for the population for the same country throughout the data set which happens for almost every country. This means a lot of data about the population is incorrect.

Country Year Population  
2933 Zimbabwe 2005 129432  
2934 Zimbabwe 2004 12777511  
2935 Zimbabwe 2003 12633897  
2936 Zimbabwe 2002 125525  
2937 Zimbabwe 2001 12366165  
2938 Zimbabwe 2000 12222251

I decided not to include Population in my project due to the fact that there are a lot of observations missing and quite a bit of data is invalid. I compared the data on the website [kaggle](https://www.kaggle.com/kumarajarshi/life-expectancy-who) that describes the variables of this data set to what I see in my data, and the errors I see are true for the dataset on [kaggle](https://www.kaggle.com/kumarajarshi/life-expectancy-who), which means these errors couldn’t have happened when I was importing the data, they were already in the data set.

## Choosing the variables that I need to answer the questions I have and creating a new data frame that has no missing values. Presenting it in a condensed way.

Life.expectancy Alcohol BMI Total.expenditure Schooling  
1 65.0 0.01 19.1 8.16 10.1  
2 59.9 0.01 18.6 8.18 10.0  
3 59.9 0.01 18.1 8.13 9.9  
4 59.5 0.01 17.6 8.52 9.8  
5 59.2 0.01 17.2 7.87 9.5  
6 58.8 0.01 16.7 9.20 9.2

## Summary statistics and proof of absence of missing values

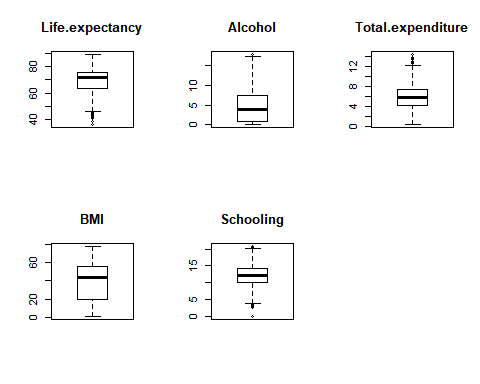
Life.expectancy Alcohol BMI Total.expenditure  
 Min. :36.30 Min. : 0.010 Min. : 1.00 Min. : 0.370   
 1st Qu.:63.65 1st Qu.: 0.850 1st Qu.:19.10 1st Qu.: 4.260   
 Median :72.20 Median : 3.760 Median :43.80 Median : 5.710   
 Mean :69.32 Mean : 4.542 Mean :38.24 Mean : 5.871   
 3rd Qu.:75.40 3rd Qu.: 7.530 3rd Qu.:55.80 3rd Qu.: 7.455   
 Max. :89.00 Max. :17.870 Max. :77.10 Max. :14.390   
 Schooling   
 Min. : 0.00   
 1st Qu.:10.10   
 Median :12.30   
 Mean :12.04   
 3rd Qu.:14.20   
 Max. :20.70

**Dimensions of the new data set: the first number represents the number of rows and the second number - the number of columns**

2563 5

# Looking for outliers

## Creating boxplots for the 5 variables we need to see if we have some possible outliers



## Looking at each of the variables highest and lowest values to find potential errors in the data

**Highest 20 values for Life.expectancy:**

89 89 89 89 89 89 89 89 89 89 89 88 88 88 88 88 88 88 88 88

**Lowest 10 values for Life.expectancy:**

36.3 39 41 41.5 42.3 43.1 43.3 43.5 43.8 44

The value of 36.3 is very different from what I could find when researching life expectancy for Haiti. Looks like it is a mistake and I will not include it in the data.

**Highest 10 values for Alcohol (recorded per capita (15+) consumption (in litres of pure alcohol)):**

17.87 17.31 16.99 16.58 16.35 15.52 15.19 15.14 15.07 15.04

The highest values seem to be correct after looking up some information about alcohol consumption in Estonia, Belarus and Lithuania.

**Lowest 10 values for Alcohol:**

0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01

The values seem to be correct after checking information on countries of Mauritania (the dry country), Malawi, and Maldives and etc. that have the lowest values for alcohol consumption according to the data set.

**Highest 10 values for BMI:**

77.1 76.7 76.2 75.7 75.2 74.8 74.6 74.3 74.3 74.1

These values are way too high. According to [WHO](https://en.wikipedia.org/wiki/Body_mass_index) BMI over 40 is class III obese. I will remove values over 50 as they are highly unlikely to be correct.

**Lowest 10 values for BMI:**

1 1.4 1.4 1.8 1.9 2 2.1 2.1 2.1 2.1

These numbers are way too low. BMI of less than 18.5 is considered underweight. I will remove values below 10 as they are highly unlikely to be correct.

**Highest 10 values for Total.expenditure:**

14.39 13.83 13.76 13.73 13.71 13.66 13.63 13.44 13.38 13.13

**Lowest 10 values for Total.expenditure:**

0.37 0.65 0.74 0.76 0.92 1.1 1.1 1.12 1.12 1.12

These values seem to be within possible boundaries.

**Highest 10 values for Schooling:**

20.7 20.6 20.5 20.4 20.4 20.3 20.3 20.3 20.3 20.1

**Lowest 20 values for Schooling:**

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2.8 2.9 2.9 2.9 2.9

0 years of schooling is used for a few countries that in practice definitely have greater number of years of school attended on average. For example, Montenegro has 0, which is not the case for this country according to this [article](https://en.wikipedia.org/wiki/Education_in_Montenegro). I am removing values that equal to 0.

## Creating a new data set without the errors that were identified earlier. Presenting it in a condensed way.

Life.expectancy Alcohol BMI Total.expenditure Schooling  
1 65.0 0.01 19.1 8.16 10.1  
2 59.9 0.01 18.6 8.18 10.0  
3 59.9 0.01 18.1 8.13 9.9  
4 59.5 0.01 17.6 8.52 9.8  
5 59.2 0.01 17.2 7.87 9.5  
6 58.8 0.01 16.7 9.20 9.2

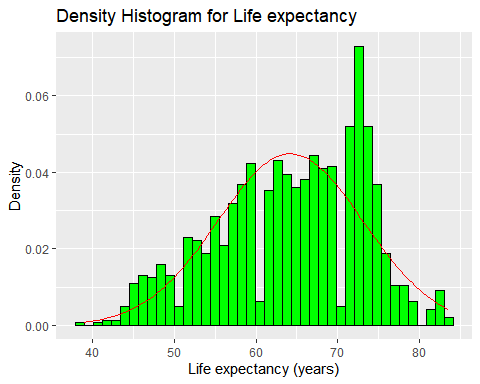
### Summary statistics for the new data set

Life.expectancy Alcohol BMI Total.expenditure  
 Min. :39.00 Min. : 0.010 Min. :11.20 Min. : 0.370   
 1st Qu.:57.90 1st Qu.: 0.420 1st Qu.:18.10 1st Qu.: 3.962   
 Median :65.20 Median : 2.250 Median :26.00 Median : 5.190   
 Mean :64.20 Mean : 3.036 Mean :29.23 Mean : 5.328   
 3rd Qu.:71.97 3rd Qu.: 4.997 3rd Qu.:42.20 3rd Qu.: 6.447   
 Max. :83.50 Max. :12.680 Max. :49.90 Max. :14.390   
 Schooling   
 Min. : 2.80   
 1st Qu.: 8.60   
 Median :10.60   
 Mean :10.27   
 3rd Qu.:12.20   
 Max. :15.80

## Visualizing normality of the variables as part of the exploratory data analysis.

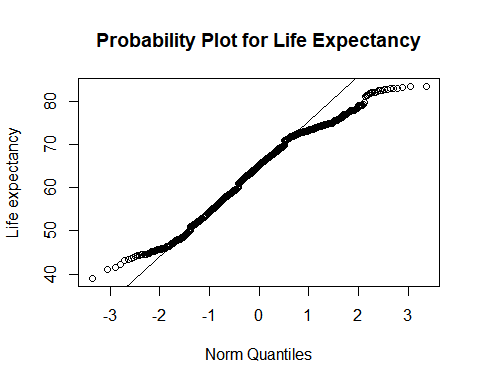
### Histogram for the variable Life.expectancy

*This sample is large which makes sense to choose histograms as the best way to look at the distribution*



The distribution is not symmetrical. It is slightly negatively skewed with a slight negative kurtosis.

### Probability plot for Life.expectancy



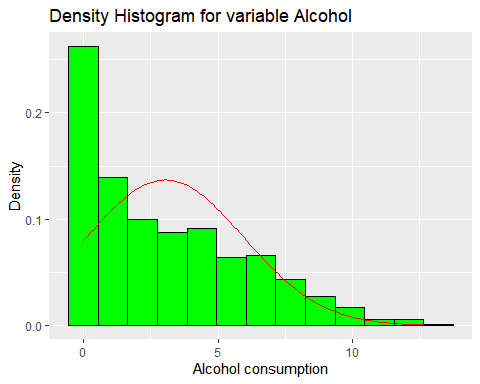
The plot shows deviations from the normal distribution. The shape of the curvature is indicative of the slight left(negatively) skewed distribution.

### Quantifying normality for the variable Life.expectancy

median mean SE.mean CI.mean.0.95 var   
 6.520000e+01 6.420305e+01 2.457653e-01 4.821369e-01 7.912476e+01   
 std.dev coef.var skewness skew.2SE kurtosis   
 8.895210e+00 1.385481e-01 -3.435883e-01 -2.541352e+00 -6.241604e-01   
 kurt.2SE normtest.W normtest.p   
-2.310055e+00 9.757802e-01 5.010022e-14

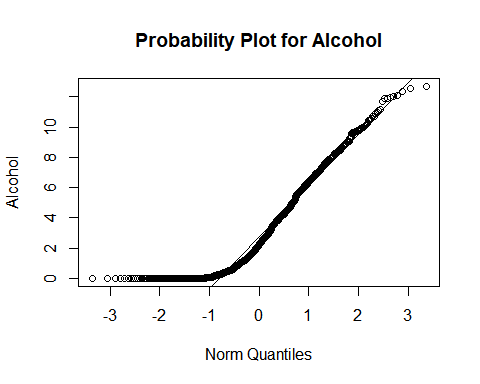
z-scores for kurtosis - 2.3 and skewedness of -2.5 (both deviating from the values required for normal distribution confirm that the distribution is slightly negatively skewed with a slight negative kurtosis).

### Histogram for the variable Alcohol



The distribution is not symmetrical. It is positively skewed.

### Probability plot for the variable Alcohol



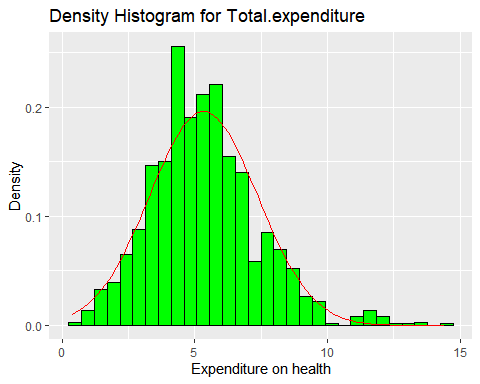
The plot shows deviations from the normal distribution. The shape of the curvature is indicative of the positively skewed distribution.

### Quantifying normality for the variable Alcohol

median mean SE.mean CI.mean.0.95 var   
 2.250000e+00 3.036366e+00 8.052077e-02 1.579639e-01 8.493508e+00   
 std.dev coef.var skewness skew.2SE kurtosis   
 2.914362e+00 9.598191e-01 8.425273e-01 6.231756e+00 -1.762404e-01   
 kurt.2SE normtest.W normtest.p   
-6.522761e-01 8.913898e-01 2.106830e-29

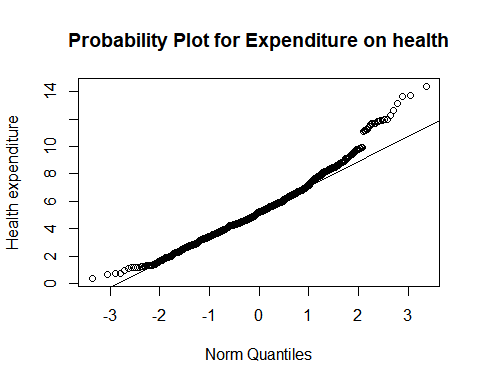
z-score for skewedness of 6.2 is too high and indicates positive skewedness. The z-score for kurtosis is -0.7 and is indicative of a slight negative kurtosis.

### Histogram for the variable Total.expenditure



The distribution is not fully symmetrical. It is positively skewed with a positive kurtosis.

### Probability plot for the variable Total.expenditure



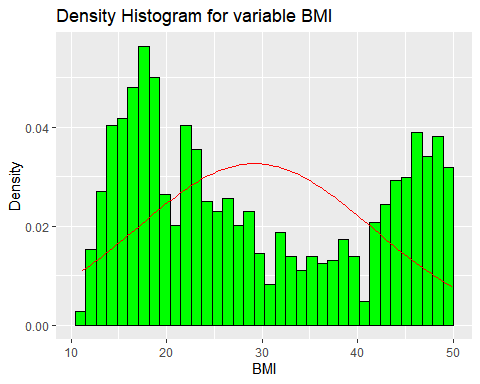
The plot shows deviations from the normal distribution. The shape of the curvature is indicative of the positively skewed distribution.

### Quantifying normality for the variable Total.expenditure

median mean SE.mean CI.mean.0.95 var std.dev   
5.190000e+00 5.327931e+00 5.618180e-02 1.102162e-01 4.134877e+00 2.033440e+00   
 coef.var skewness skew.2SE kurtosis kurt.2SE normtest.W   
3.816565e-01 6.703167e-01 4.958000e+00 1.135437e+00 4.202318e+00 9.750390e-01   
 normtest.p   
2.821728e-14

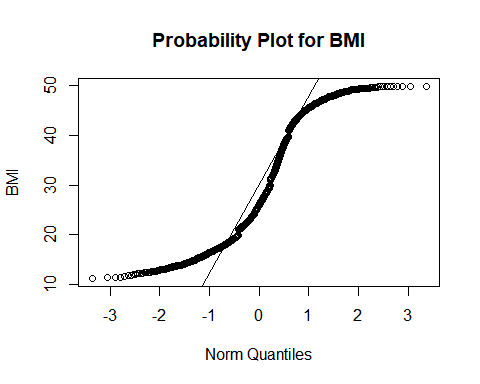
z-scores for skewedness of 4.9 indicates positive skewedness. The z-score for kurtosis is 4.2 and is indigative of a positive kurtosis.

### Histogram for the variable BMI



The distribution is not symmetrical, and looks bimodal.

### Probability plot for the variable BMI

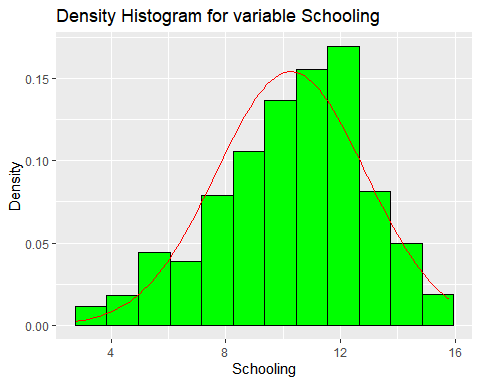


### Quantifying normality for the variable BMI

median mean SE.mean CI.mean.0.95 var   
 2.600000e+01 2.922656e+01 3.379616e-01 6.630055e-01 1.496256e+02   
 std.dev coef.var skewness skew.2SE kurtosis   
 1.223215e+01 4.185286e-01 3.256202e-01 2.408451e+00 -1.389314e+00   
 kurt.2SE normtest.W normtest.p   
-5.141932e+00 8.996072e-01 1.907747e-28

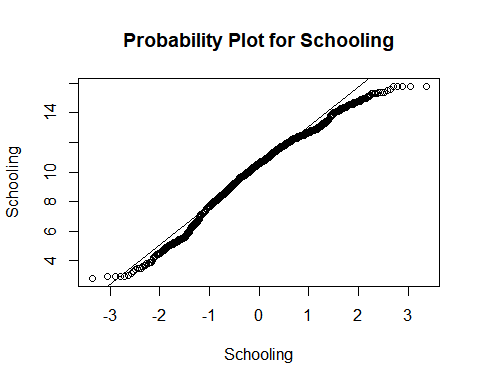
z-score for skewedness is positive and equals to 2.4 . The z-score for kurtosis is negative and equals to -5.1.

### Histogram for the variable Schooling



The distribution looks slightly negatively skewed.

### Probability plot for the variable Schooling

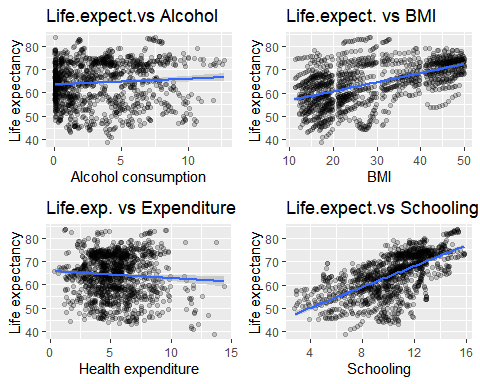


### Quantifying normality for the variable Schooling

median mean SE.mean CI.mean.0.95 var   
 1.060000e+01 1.027122e+01 7.157724e-02 1.404186e-01 6.711524e+00   
 std.dev coef.var skewness skew.2SE kurtosis   
 2.590661e+00 2.522252e-01 -4.709767e-01 -3.483581e+00 -1.261683e-01   
 kurt.2SE normtest.W normtest.p   
-4.669563e-01 9.800245e-01 1.721581e-12

z-score for skewedness of -3.5 which indicates negative skewedness. The z-score for kurtosis is -0.5 and is indicative of a very slight negative kurtosis.

## To visualize relationships between all of the variables with Life expectancy I created these scatterplots.



The linearity is best reflected in Life.expectancy vs Schooling scatterplot, and Life.expectancy vs BMI.

## I would like to build a model to predict life expectancy based on the level of consumption of alcohol, BMI, health expenditure, and schooling:

*With the help of this multiple regression I will test the strength of the relationship between the dependent and each independent variable while the rest of the independent variables are held* ***constant****.*

Call:  
lm(formula = Life.expectancy ~ Schooling + Alcohol + BMI + Total.expenditure,   
 data = newdataset2)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-20.0512 -3.3638 0.7012 4.1051 15.1290   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 42.05352 0.83954 50.091 < 2e-16 \*\*\*  
Schooling 1.94263 0.08334 23.311 < 2e-16 \*\*\*  
Alcohol -0.60737 0.06472 -9.385 < 2e-16 \*\*\*  
BMI 0.21254 0.01697 12.523 < 2e-16 \*\*\*  
Total.expenditure -0.40750 0.08639 -4.717 2.65e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 6.216 on 1305 degrees of freedom  
Multiple R-squared: 0.5131, Adjusted R-squared: 0.5116   
F-statistic: 343.8 on 4 and 1305 DF, p-value: < 2.2e-16

The F-statistic p-value : < 2.2e-16 is highly significant, which implies at least one of the independent variables is significantly related to the dependent variable. R squared statistic explains how much variability in the dependent variable is accounted for by the independent variables. In my case it is **51.3%** The *adjusted* R squared at **51.2%** is pretty close to the value of *R squared*. The difference is small, about **0.2%**. The shrinkage means that the model derived from a population and not a sample would account for 0.2% less variance in the outcome. Using **Stein’s formula** I get the version of R squared that equals to 0.5101. The difference between this value and the value of unadjusted R squared is 0.5131-0.5101= 0.003, or about **0.3%** percent. This value helps me with cross-validation of the model. The value is very similar to the observed value of R squared, which is a good indicator of a cross validity of the model.

## Examining the estimated coefficients and the t-statistic p-values:

Estimate Std. Error t value Pr(>|t|)  
(Intercept) 42.0535179 0.83954286 50.090972 3.250412e-306  
Schooling 1.9426289 0.08333630 23.310718 9.166319e-101  
Alcohol -0.6073670 0.06471757 -9.384886 2.698343e-20  
BMI 0.2125363 0.01697173 12.522961 4.653475e-34  
Total.expenditure -0.4075041 0.08638609 -4.717242 2.647637e-06

The column “Estimate” provides estimates for the b-values that give us information about the relationship between the predictors and the dependent variable. Predictors Schooling and BMI have positive b-values which is an indicator of a positive relationship Predictors Alcohol and Total.expenditure have negative b-values which indicate a negative relationship. If we look at the p-values we can see that they all are significantly less than .05, thus I can say that all the predictors are significant predictors of this model. By looking at the t-statistic (the third column) I can tell that **Schooling** and **BMI** have the highest impact and Total. expenditure has the lowest impact. Since all of these variables use different units of measurement, I am going to look at standardized beta estimates. Since standardized betas are directly comparable due to being measured in standard deviation units, they give better information about the “importance” of the independent variables.

Schooling Alcohol BMI Total.expenditure   
 0.56577562 -0.19899333 0.29226709 -0.09315519

**Schooling**. Standardized beta is 0.566. This value shows that when Schooling increases by one standard deviation (2.59), life expectancy increases by 0.566 standard deviations. Standard deviation for Life.expectancy is 8.9, so the change that will be observed is about 5 years (5.0374)

**BMI**. Standardized beta is 0.292.

This value shows that when BMI increases by one standard deviation (12.23), life expectancy increases by 0.292 standard deviations. Standard deviation for Life.expectancy is 8.9, so the change that will be observed is 2.5988.

**Alcohol**. Standardized beta is -0.199.

This value shows that when Alcohol increases by one standard deviation (2.91), life expectancy decreases by -0.199 standard deviations. Standard deviation for Life.expectancy is 8.9, so the change that will be observed is -1.7711.

**Total.expenditure**. Standardized beta is -0.093.

This value shows that when Health expenditure increases by one standard deviation (2.03), life expectancy decreases by -0.093 standard deviations. Standard deviation for Life.expectancy is 8.9, so the change that will be observed is -0.8277.

## Checking the confidence intervals

2.5 % 97.5 %  
(Intercept) 40.4065166 43.7005192  
Schooling 1.7791411 2.1061167  
Alcohol -0.7343288 -0.4804051  
BMI 0.1792415 0.2458312  
Total.expenditure -0.5769749 -0.2380333

*Printing the estimated coefficients so it is easier to compare the results*

Estimate Std. Error t value Pr(>|t|)  
(Intercept) 42.0535179 0.83954286 50.090972 3.250412e-306  
Schooling 1.9426289 0.08333630 23.310718 9.166319e-101  
Alcohol -0.6073670 0.06471757 -9.384886 2.698343e-20  
BMI 0.2125363 0.01697173 12.522961 4.653475e-34  
Total.expenditure -0.4075041 0.08638609 -4.717242 2.647637e-06

None of the confidence intervals cross 0 which means that the predictors are related to the outcome. The confidence intervals are small, which is an indicator of a good model.

## Casewise diagnostics to identify outliers or influential cases

Showing some of it:

Life.expectancy Alcohol BMI Total.expenditure Schooling residuals  
1 65.0 0.01 19.1 8.16 10.1 2.597794  
2 59.9 0.01 18.6 8.18 10.0 -2.193525  
3 59.9 0.01 18.1 8.13 9.9 -1.913369  
4 59.5 0.01 17.6 8.52 9.8 -1.853912  
5 59.2 0.01 17.2 7.87 9.5 -1.750986  
6 58.8 0.01 16.7 9.20 9.2 -0.919949  
 standardized.residuals studentized.residuals cooks.distance  
1 0.4188494 0.4187170 1.596094e-04  
2 -0.3536758 -0.3535572 1.148977e-04  
3 -0.3084974 -0.3083905 8.654087e-05  
4 -0.2989960 -0.2988916 9.151289e-05  
5 -0.2822584 -0.2821588 6.591288e-05  
6 -0.1484357 -0.1483800 2.659939e-05  
 dfbeta.(Intercept) dfbeta.Schooling dfbeta.Alcohol dfbeta.BMI  
1 -7.040590e-03 8.820671e-04 -9.608053e-04 -2.155813e-04  
2 5.713462e-03 -7.264624e-04 8.013493e-04 1.863923e-04  
3 4.646636e-03 -6.151706e-04 6.873575e-04 1.656883e-04  
4 4.999366e-03 -5.934464e-04 6.728815e-04 1.678286e-04  
5 2.904637e-03 -4.573298e-04 5.866434e-04 1.479875e-04  
6 2.410523e-03 -2.218825e-04 3.229674e-04 8.247965e-05  
 dfbeta.Total.expenditure dffit leverage covariance.ratios  
1 1.725024e-03 0.02824080 0.004528367 1.007729  
2 -1.466750e-03 -0.02396045 0.004571730 1.007967  
3 -1.262193e-03 -0.02079433 0.004526036 1.008037  
4 -1.365363e-03 -0.02138329 0.005092188 1.008633  
5 -1.061555e-03 -0.01814751 0.004119588 1.007685  
6 -7.937891e-04 -0.01152811 0.006000003 1.009814

### Looking at the cases with large residuals and identifying the quantity:

The number of cases with large residuals is 76.

The calculated number represents 5.8% of cases which means that the sample is within 1% of what I would expect.

## Leverage, cooks distance, covariance ratios

**cooks.distance**

Cases with cooks.distance greater than 1:

[1] Life.expectancy Alcohol BMI Total.expenditure  
[5] Schooling cooks.distance   
<0 rows> (or 0-length row.names)

There are no cases with cooks.distance greater than 1.

**leverage**

The average leverage would be calculated as the number of predictors (4) +1 , divided by the size of the sample(1310), which equals to 0.0038168. We are looking for the value that is at least 3 times greater than the average leverage (a value that is greater than 0.0115)

[1] Life.expectancy Alcohol BMI Total.expenditure  
[5] Schooling leverage   
<0 rows> (or 0-length row.names)

It looks like all the cases are within the boundary of three times the average.

**Looking at covariance ratios**

*Top 6 cases with abnormal covariance.ratios with large residuals sorted in ascending order:*

Life.expectancy Alcohol BMI Total.expenditure Schooling covariance.ratios  
672 44.5 2.67 27.4 6.30 10.7 0.9657809  
164 46.4 5.51 31.6 4.65 11.8 0.9658981  
673 44.8 1.80 26.9 6.96 10.7 0.9669365  
165 46.0 6.41 31.1 6.47 11.9 0.9693882  
671 45.3 2.61 27.9 7.12 10.7 0.9699516  
718 43.1 1.18 14.1 6.70 10.7 0.9709145

*Top 6 cases with covariance.ratios not within normal range and with large residuals sorted in the descending order*:

Life.expectancy Alcohol BMI Total.expenditure Schooling covariance.ratios  
1123 47.8 5.53 28.2 6.81 9.9 0.9879872  
1094 54.0 8.50 44.3 7.57 12.9 0.9878869  
713 46.0 1.04 15.9 8.20 9.7 0.9878010  
76 77.0 0.01 16.4 3.80 9.9 0.9876887  
1293 47.9 2.46 18.0 7.33 10.5 0.9875844  
1054 44.3 3.80 19.7 1.68 8.0 0.9874061

Some of the values barely exceed the boundaries (0.9885 - 1.0115). No cases of real concern.

## Testing for multicollinearity

**VIF**

Schooling Alcohol BMI Total.expenditure   
 1.578936 1.205054 1.459934 1.045262

None of the VIF are greater than 10.

**Average of VIF**

[1] 1.322297

The average of VIF values is not substantially greater than 1, which helps us make a conclusion that there is no collinearity within the data.

**Tolerance**

Schooling Alcohol BMI Total.expenditure   
 0.6333378 0.8298384 0.6849625 0.9566977

None of the values have tolerance that is lower than 0.2, which indicates absence of collinearity.

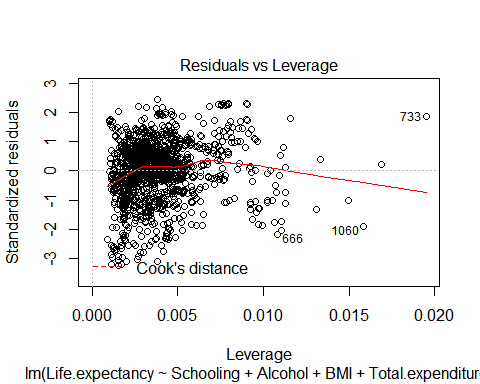
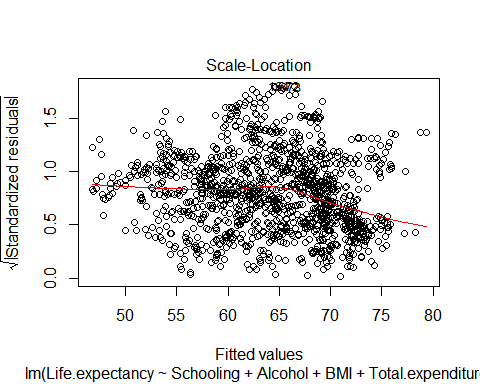
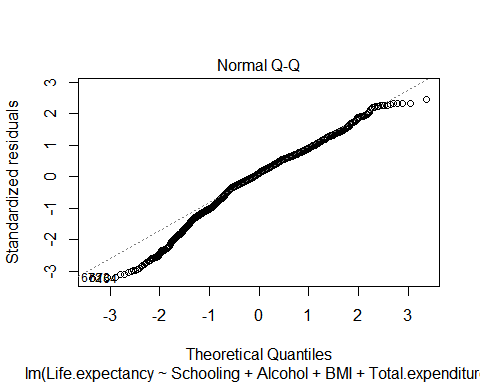
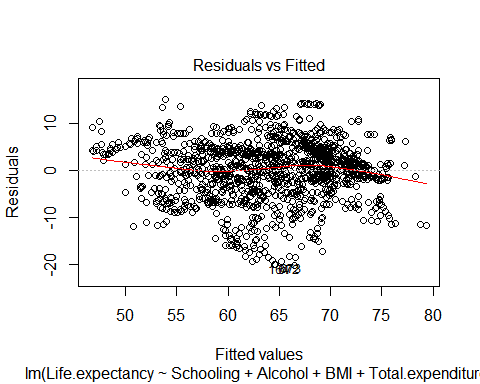
Assumption of no multicollinearity has been met.

## Assessing the assumption of independence

lag Autocorrelation D-W Statistic p-value  
 1 0.8607623 0.272879 0  
 Alternative hypothesis: rho != 0

The D-W Statistic is less than 1, which indicates a positive correlation between adjacent residuals, and also a case of concern. The value of p-value is 0, and is less than 0.05, which indicates evidence to reject the null-hypothesis (there is no correlation among the residuals). The assumption of independence hasn’t been met.

## Checking assumptions about the residuals



It looks like the degree of residuals scattering around zero is not the same for all fitted values (first graph) but close to the normal pattern. When we look at the Q-Q plot we can see that not all of the points on the Q-Q plot lie on the straight line, which shows abnormal distribution of the residuals.

### Conclusions about the first model:

Even though the model overall looks good, some of the assumptions have been violated. I could draw conclusions about this particular data but not generalize my findings beyond this sample since the assumptions were violated.

# Next Step:

The biggest concern about the previous model was the violation of the assumption of independence which could deteriorate the results of multiple regression. In order to get valid results, I will attempt to see why the assumption of Independence was violated. Usually the way data was collected is to blame. After looking at my data closely I realized that the same country with different values for the rest of the variables is analyzed for years 2000-2015, which means it is analyzed 16 times. Even though the values for each year for all of the variables are usually different they are still very close to the previously analyzed year. I believe this caused the possible lack of independence. I found that [Handbook of Biological Statistics](http://www.biostathandbook.com/independence.html) was a good reading for this topic.

## Solution:

I am going to create a new data frame with observations for the most recent year. Since year 2015 is the most recent one but it has a lot of observations that are missing, I will use year 2014.

## New data frame created with the help of filtering by year (2014) only for 5 variables I chose before and without missing values

*10 highest values for Life.expectancy*

[1] 89.0 89.0 89.0 89.0 88.0 87.0 84.0 83.5 83.2 83.0

*10 lowest values for Life.expectancy*

[1] 48.1 51.7 52.1 52.6 53.6 56.7 56.7 57.6 57.8 57.9

*Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol).* The highest values seem to be correct after looking up some information about alcohol consumption in Estonia, Belarus and Lithuania.

[1] 15.19 13.94 12.60 12.32 12.14 12.03 11.50 11.12 11.03 10.75

*10 lowest values for Alcohol consumption.* The values seem to be correct after checking information on countries like Mauritania (the dry country), Malawi, and Maldives and etc.

[1] 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01

*10 highest values for BMI.* These values are way too high. According to [WHO](https://en.wikipedia.org/wiki/Body_mass_index) BMI over 40 is class III obese. I will be removing extreme values.

[1] 77.1 74.8 74.3 69.2 68.7 68.4 67.3 66.9 66.4 66.1

*10 lowest values for BMI*

[1] 2.0 2.8 2.9 3.1 3.1 3.5 4.2 4.4 5.1 5.2

These numbers are way too low. BMI of less than 18.5 is considered underweight. I will be removing values below 10 as they are highly unlikely to be correct.

*10 highest values for Health expenditure*

[1] 13.73 13.71 11.93 11.90 11.66 11.60 11.54 11.38 11.30 11.30

*10 lowest values for Health expenditure*

[1] 1.21 1.23 1.37 1.40 1.45 1.48 1.57 1.59 1.62 1.80

*10 highest values for Schooling*

[1] 20.4 19.2 19.2 19.0 18.6 18.1 17.7 17.6 17.6 17.3

These values seem to be within possible boundaries.

*10 lowest values for Schooling*

[1] 5.0 5.3 6.3 7.1 7.3 7.7 7.8 8.2 8.4 8.5

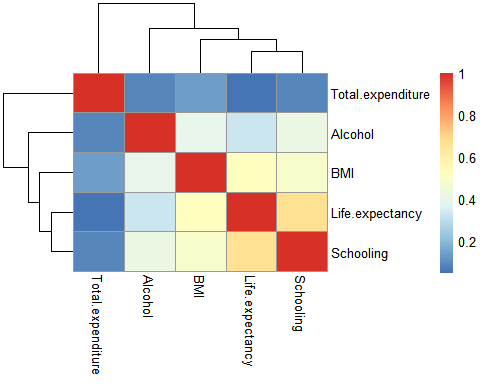
These values seem to be within possible boundaries.

## Data frame after removing the extreme values of BMI

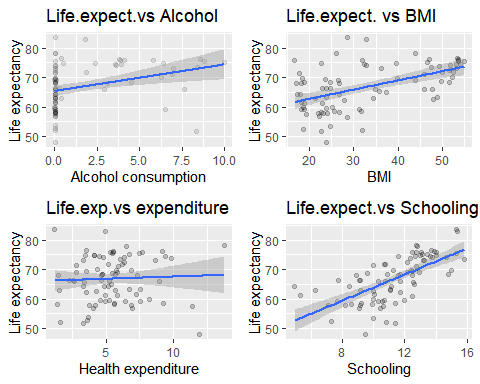
Life.expectancy Alcohol BMI Total.expenditure Schooling  
1 59.9 0.01 18.6 8.18 10.0  
2 51.7 8.33 22.7 3.31 11.4  
3 76.2 8.56 47.0 5.54 13.9  
4 74.6 3.91 54.1 4.48 12.7  
5 72.5 0.01 51.5 6.40 12.2  
6 71.4 0.01 17.7 2.82 10.0

## Creating a heatmap for correlation coefficients.

I will have a better idea of the correlations after using summary function of the model, because the other variables are held constant when we get correlation coefficients seen in the output from the summary function.



## To visualize relationships between all of the variables with Life expectancy I created these scatterplots.



## Creating a new model and looking at the summary results

Call:  
lm(formula = Life.expectancy ~ BMI + Schooling + Alcohol + Total.expenditure,   
 data = newdataset\_LED2)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-14.1084 -3.8951 0.3572 3.2929 11.1133   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 41.76533 3.48571 11.982 < 2e-16 \*\*\*  
BMI 0.15755 0.05704 2.762 0.00716 \*\*   
Schooling 1.81364 0.31882 5.689 2.14e-07 \*\*\*  
Alcohol -0.02183 0.25298 -0.086 0.93147   
Total.expenditure -0.11451 0.26433 -0.433 0.66607   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 5.568 on 78 degrees of freedom  
Multiple R-squared: 0.4966, Adjusted R-squared: 0.4708   
F-statistic: 19.24 on 4 and 78 DF, p-value: 4.809e-11

Looks like the *p*-values for the variables Alcohol and Total.expenditure are greater than .05 which means that there is no sufficient evidence to determine that effect exists for these variables at the population level.

The variables BMI and Schooling are statistically significant.

Since it is pretty common to use *p*-values of the coefficients to decide which values should be in the final model and which shouldn’t, I am keeping BMI and Schooling for my final model.

## Final model

Call:  
lm(formula = Life.expectancy ~ BMI + Schooling, data = newdataset\_LED2)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-14.0387 -3.8862 0.4495 3.2670 11.2657   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 41.35360 3.05001 13.559 < 2e-16 \*\*\*  
BMI 0.15368 0.05432 2.829 0.0059 \*\*   
Schooling 1.80290 0.30227 5.964 6.35e-08 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 5.505 on 80 degrees of freedom  
Multiple R-squared: 0.4954, Adjusted R-squared: 0.4828   
F-statistic: 39.27 on 2 and 80 DF, p-value: 1.315e-12

The F-statistic *p*-value of 1.315e-12 is highly significant, which implies at least one of the independent variables is significantly related to the dependent variable. R squared statistic explains how much variability in the dependent variable is accounted for by the independent variables. In my case it is **49.5%** The *adjusted* R squared at **48.3%** is pretty close to the value of *R squared*. The difference is small, about **1.2%**. The shrinkage means that the model derived from a population and not a sample would account for 1.2% less variance in the outcome. Using **Stein’s formula**, I get the version of R squared that equals to 0.463. The difference between this value and the value of unadjusted R squared is 0.495-0.463= 0.032, or about **3.2%** percent. This value helps me with cross-validation of the model. The value is similar to the observed value of R squared, which is a good indicator of a cross validity of the model.

**Standardized betas**

BMI Schooling   
0.2571045 0.5420347

**Schooling**. Standardized beta is 0.542. This value shows that when Schooling increases by one standard deviation (2.3), life expectancy increases by 0.542 standard deviations. Standard deviation for Life.expectancy is 7.65, so the change that will be observed is about (4.1463)

**BMI**. Standardized beta is 0.257. This value shows that when BMI increases by one standard deviation (12.81), life expectancy increases by 0.257 standard deviations. Standard deviation for Life.expectancy is 7.65, so the change that will be observed is 1.96605.

## Checking the confidence intervals

2.5 % 97.5 %  
(Intercept) 35.28388475 47.4233148  
BMI 0.04557858 0.2617836  
Schooling 1.20135112 2.4044434

*Printing the estimated coefficients so it is easier to compare the results*

Estimate Std. Error t value Pr(>|t|)  
(Intercept) 41.3535998 3.05001085 13.558509 1.975326e-22  
BMI 0.1536811 0.05432115 2.829121 5.897901e-03  
Schooling 1.8028972 0.30227485 5.964430 6.350759e-08

None of the confidence intervals cross 0 which means that the predictors are related to the outcome. The confidence intervals are small, which is an indicator of a good model.

## Testing for multicollinearity

**VIF**

BMI Schooling   
 1.309279 1.309279

None of the VIF are greater than 10.

**Average of VIF**

[1] 1.309279

The average of VIF values is not substantially greater than 1, which helps us make a conclusion that there is no collinearity within the data.

**Tolerance**

BMI Schooling   
0.7637793 0.7637793

None of the values have tolerance that is lower than 0.2, which indicates absence of collinearity.

Assumption of no multicollinearity has been met.

## Assessing the assumption of independence

lag Autocorrelation D-W Statistic p-value  
 1 -0.1799472 2.345003 0.13  
 Alternative hypothesis: rho != 0

DW statistic is close to 2 and *p*-value is greater than .05, which indicate that assumption of independence has been met.

## Casewise diagnostics to identify outliers or influential cases

Showing some of it:

Life.expectancy BMI Schooling residuals standardized.residuals  
1 59.9 18.6 10.0 -2.341041 -0.4312766  
2 51.7 22.7 11.4 -13.695189 -2.5170909  
3 76.2 47.0 13.9 2.563117 0.4732809  
4 74.6 54.1 12.7 2.035457 0.3784500  
5 72.5 51.5 12.2 1.236477 0.2290973  
6 71.4 17.7 10.0 9.297272 1.7145493  
 studentized.residuals cooks.distance dfbeta.(Intercept) dfbeta.BMI  
1 -0.4290718 0.0017654467 -0.117289043 0.002571830  
2 -2.6066580 0.0499968701 -0.170894225 0.014546904  
3 0.4709734 0.0024818439 -0.151498931 0.001774806  
4 0.3764143 0.0022721759 -0.060789637 0.003577050  
5 0.2277357 0.0007054436 -0.019532937 0.002001580  
6 1.7359936 0.0299813687 0.468885598 -0.011074605  
 dfbeta.Schooling dffit leverage covariance.ratios  
1 0.000279183 -0.07240390 0.02768664 1.0605953  
2 -0.042415473 -0.40106724 0.02312622 0.8299476  
3 0.011027365 0.08586680 0.03217041 1.0639879  
4 -0.002823190 0.08211817 0.04543113 1.0820486  
5 -0.002761561 0.04573017 0.03875928 1.0782053  
6 0.001159956 0.30365784 0.02968821 0.9565328

### Looking at the cases with large residuals and identifying the quantity:

The number of cases with large residuals is 4.

The calculated number represents 4.82% of cases which means that the sample is within 1% of what I would expect.

## Leverage, cooks distance, covariance ratios

**cooks.distance**

Cases with cooks.distance greater than 1:

[1] Life.expectancy BMI Schooling cooks.distance   
<0 rows> (or 0-length row.names)

There are no cases with cooks.distance greater than 1.

**leverage**

The average leverage would be calculated as the number of predictors(2) +1 , divided by the size of the sample(83), which equals to 0.0361446. We are looking for the value that is at least 3 times greater than the average leverage (a value that is greater than 0.11)

[1] Life.expectancy BMI Schooling leverage   
<0 rows> (or 0-length row.names)

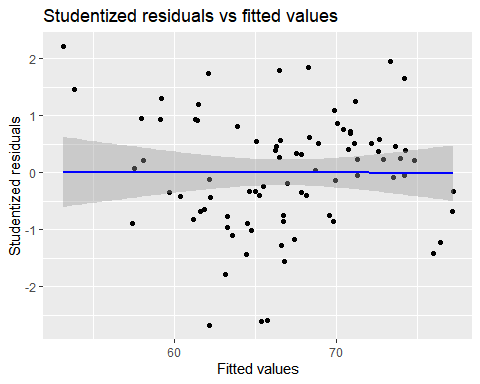
It looks like all the cases are within the boundary of three times the average.

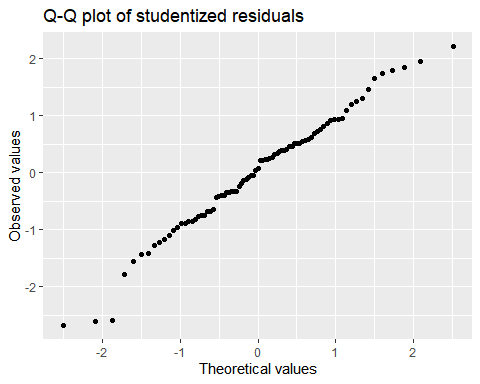
**Looking at covariance ratios**

Life.expectancy BMI Schooling covariance.ratios  
2 51.7 22.7 11.4 0.8299476  
42 52.1 32.0 10.8 0.8249948  
67 48.1 23.8 9.5 0.8180947

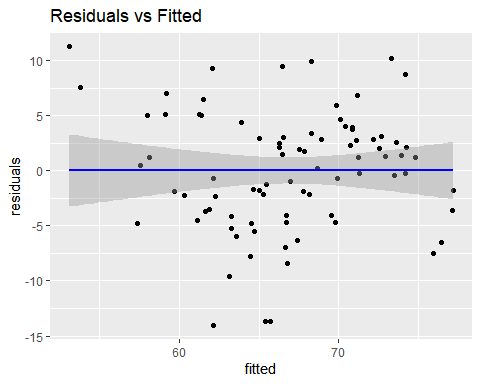
Some of the values barely exceed the boundaries (0.89 - 1.11). No cases of real concern especially with no cases with cooks.distance greater than 1.

## Assesing assumptions about residuals





It appears that the Studentized residuals are normally distributed, the scatters for the most part are on or very close to the normal distribution line.



Looking at the scatterplot of Residuals vs Fitted values I can tell that there is no evidence of nonlinear pattern to the residuals.

# Conclusion about the model

After cross-validating the final model with the help of Stein’s formula above, and assessing if all of the necessary assumptions have been met, I can tell that the model looks good without violations of assumptions, which makes me believe the results would be more correct than for the first model I built.

# Implications

**1. What are the possible factors and what areas (social, economic, health and etc.) could be influencing variations in life expectancy?** The two factors that seem to have a significant relationship with the length of life expectancy turned out to be **Schooling** which belongs to the social factor, and **BMI** which is associated with the health factor.

**2. Is there a negative relationship between life expectancy and alcohol consumption?** No significant effect of the variable Alcohol was discovered. I can’t conclude if Alcohol positively or negatively affected life expectancy from the analysis I performed since the results were not statistically significant.

**3. Is there a positive relationship between life expectancy and schooling?** This analysis makes me believe that there is a significant positive relationship between **Schooling** and the length of life expectancy.

**4. Is there a negative relationship between life expectancy and population density?** The data weren’t letting me use this variable where I wouldn’t be concerned about incorrect results. There were a lot of missing values for this variable, and a lot of the values were invalid. The analysis for this variable wasn’t performed in the end.

**5. What is the relationship between BMI and life expectancy?** The results of the estimated coefficients that I got with the help of my model indicate that there is a positive relationship between BMI and life expectancy. The discrepancy between high values of BMI and gains in life expectancy at first seemed confusing because my initial belief about BMI is that the greater, the worse the effects might be for the life expectancy. After reading a few studies( [Association of All-Cause Mortality With Overweight and Obesity Using Standard Body Mass Index Categories](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4855514/), [How Much Should We Weigh for a Long and Healthy Life Span? The Need to Reconcile Caloric Restriction versus Longevity with Body Mass Index versus Mortality Data](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4115619/), [Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2662372/)). I realized that the variable BMI is more complicated than it seems. There are a few main ranges for BMI (underweight, normal, overweight and obese), and for each of these ranges the effects on the life expectancy could diifer. For the underweight (especially for malnourished, or those whose weight loss happened due to effect of deseases) the increase in BMI might mean longer life expectancy. For normal BMI range, according to the studies the BMI increase that stays within the normal range is known not to affect life expectancy significantly (a small increase has been mentioned). It is the overweight and obese categories that might benefit from the decrease of BMI since increase of weight affects the likelihood of getting diseases associated with increased mortality rates, at the same time the increased effective treatments of the diseases known to be caused by increased weight in the obesity range also changes the overall picture of the effect of BMI on life expectancy. The quality of life might be signficantly worse for people with the BMI in the overweight range but their life expectancy is known to have benefitted from the advanced treatments. After researching these findings with the help of the study [Obesity and Trends in Life Expectancy](https://www.hindawi.com/journals/jobe/2012/107989/), I had a better understanding why I got the results that I didn’t expect. It also makes sense to suggest to look at different ranges of BMI separately to get a clear picture of the findings and possible reasons for the findings.

**6. Is there a positive relationship between expenditure on health and life expectancy?** No significant effect of the variable Total.expenditure was discovered. I can’t conclude if expenditure on health positively or negatively affected life expectancy from the analysis I performed because the effect was not statistically significant.

**7. What could be done to prolong life expectancy?** I will answer this by trying to predict the variable of life expectancy by comparing a few predictions:

With the value of Schooling = 15, and BMI = 20,

*Life expectancy* would be

71

With the value of Schooling = 19, and BMI = 20,

*Life expectancy* would be

79

With the value of Schooling = 15, and BMI = 24,

*Life expectancy* would be

72

It looks like incraesing the level of education would possibly increase life expectancy. The same could be told about BMI, but it is important to remember that final conclusion yet to be made about BMI, beacuse it is a complicated variable, and analyzing seperate ranges of BMI might provide more insights.

## Limitations

The project has some limitations. Even though it provides results for effects of BMI and Schooling, it doesn’t explain other factors that could be influencing life expectancy. Considering performing the analyzes of other variables from the data set can allow for discovery of additional factors influencing life expectancy.

Incorrect data points present in the data set and missing values significantly reduced the amount of observations available thus reducing the possibility of accounting for all of the information for the variables. Finding the original data set and verifying the values would help with substituting the incorrect values of the secondary data set analyzed in the project.

The project only covered the values for the year of 2014. There are 15 more years which the project hasn’t analyzed. Further analysis of the rest of the years and comparison of the results could help with confirming or questioning the validity of the results obtained so far, or identifying the trends of effects on the life expectancy over the given time frame.

Looking at different ranges of BMI separately could help with getting a clearer picture of the findings and more possible reasons for explanation of the results of this project.

## Conclusion

This project examined the hypotheses about the relationships between factors of Schooling, BMI, expenditure on health, alcohol and life expectancy with the help of multiple regression analysis. Since the effect of only two of the variables I analyzed turned out to be statistically significant, I could only suggest that if we increase the number of years spent on education, our life expectancy might increase. The same is true for BMI, but the increase seen is a lot smaller than that for Schooling. Also, as mentioned above BMI is a complicated variable and its effect might possibly depend on which range of BMI we are looking at, which makes Schooling the variable whose results seem to be the most influential. It is important to remember that the relationships discovered do not imply causation.