Analyzing Life Expectancy Using Linear Models

Thanh Ho, Akshay Sakanaveeti, Malavika Mampally

2023-12-04

This dataset called Life Expectancy (WHO), sourced from Kaggle.com and originally obtained from the World Health Organization's Global Health Observatory (GHO), is a comprehensive study examining the impact of immunization and the Human Development Index (HDI) on life expectancy. This suggests that the data is likely to be authoritative and reliable for studying health-related factors. It takes into account critical immunization factors such as Hepatitis B, Polio, and Diphtheria, alongside economic indicators, social variables, education metrics, and other health-related factors. Through the analysis of this dataset, we aim to predict life expectancy based on these variables and formulate evidence-based policies to enhance public health outcomes. Analyzing such a dataset can lead to valuable insights that may help governments, healthcare organizations, and policymakers make informed decisions to improve public health and life expectancy.

Data

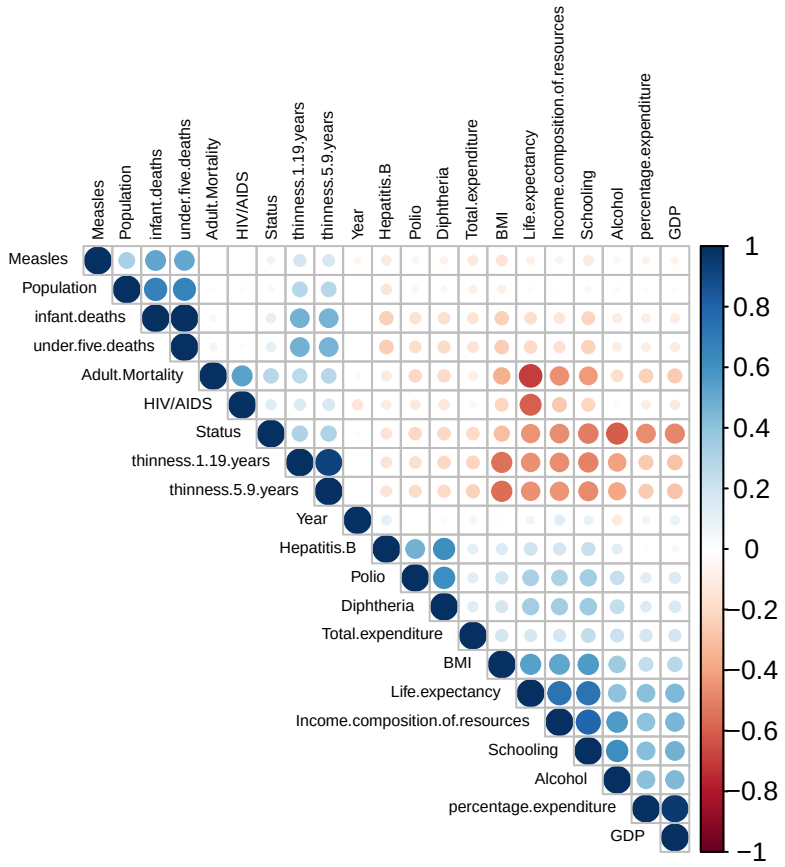
The dataset has 2938 observations. Each observation has information about life expectancy of a country in a year along with 22 economical and immunization related factors. The predicting variables can be broadly classified into different categories. The following is a list of some of the features.

- Alcohol-Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
- percentage expenditure-Expenditure on health as a percene of Gross Domestic Product per capita(%)
- Adult Mortality-Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- infant deaths-Number of Infant Deaths per 1000 population
- Life expectancy-Life Expectancy in age
- Status-Developed or Developing status
- Year-Year
- Country-Country
- Hepatitis B-Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- Measles-Measles - number of reported cases per 1000 population
- BMI-Average Body Mass Index of entire population
- under-five deaths-Number of under-five deaths per 1000 population
- Polio-Polio (Pol3) immunization coverage among 1-year-olds (%)
- Total expenditure-General government expenditure on health as a percene of total government expenditure (%)

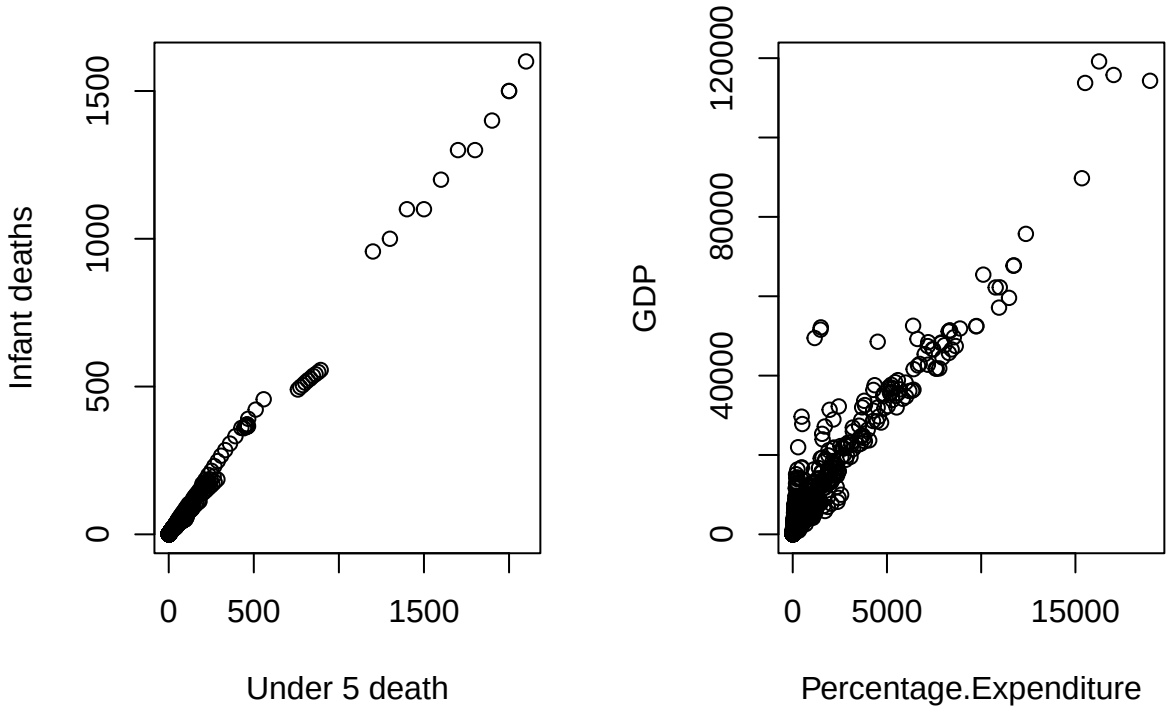
- Diphtheria-Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
- HIV/AIDS-Deaths per 1 000 live births HIV/AIDS (0-4 years)
- GDP-Gross Domestic Product per capita (in USD)
- Population-Population of the country-
- thinness 1-19 years-Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
- thinness 5-9 years-Prevalence of thinness among children for Age 5 to 9(%)
- Income composition of resources-Income composition of resources
- Schooling - Number of years of Schooling(years)

To begin analyzing the dataset, it is essential to address missing data by removing entries that pertain to countries without complete population data spanning the entire 15-year period (2000-2015). This step ensures the integrity and completeness of the dataset for further examination.

We initiate our exploration of the dataset by conducting an overall analysis, including the visualization of correlation relationships among variables. The examination reveals positive correlations between life expectancy and income, as well as life expectancy and schooling. Conversely, a negative correlation is observed between life expectancy and adult mortality. Additionally, positive correlations are identified between variables such as thinness 1.19 years and thinness 5.9 years, percentage expenditure and GDP, and infant death and under-five death. Notably, some variables exhibit reasonable correlations, such as percentage expenditure and GDP, where expenditure utilizes the same variables as GDP. Given the dataset's numerous variables, a subsequent step involves checking for multicollinearity to ensure the accuracy of the model.



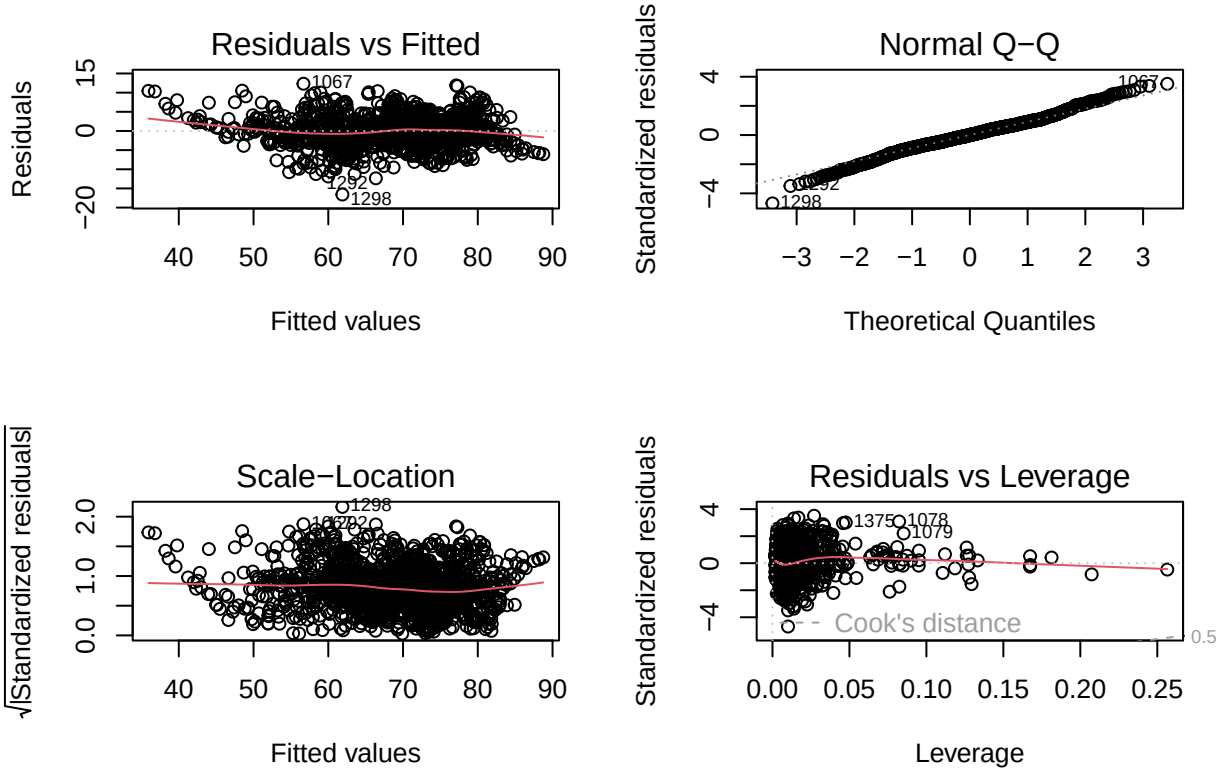
Plots of some highly correlated variables



We fit the model to data to identify any potential issues with data.

```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5893  -2.1296  -0.0041   2.1758  12.3207
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    2.920e+02  4.972e+01   5.874 5.19e-09 ***
## Year                          -1.183e-01  2.483e-02  -4.763 2.08e-06 ***
## Status                        -9.818e-01  3.437e-01  -2.857  0.00434 **
## Adult.Mortality               -1.595e-02  9.561e-04 -16.684  < 2e-16 ***
## infant.deaths                  8.644e-02  1.069e-02   8.083 1.25e-15 ***
## Alcohol                       -1.383e-01  3.421e-02  -4.043 5.54e-05 ***
## percentage.expenditure         3.440e-04  1.801e-04   1.911  0.05622 .
## Hepatitis.B                   -5.081e-03  4.658e-03  -1.091  0.27555
## Measles                       -1.067e-05  1.080e-05  -0.989  0.32303
## BMI                            3.020e-02  6.080e-03   4.966 7.57e-07 ***
## under.five.deaths             -6.487e-02  7.739e-03  -8.382  < 2e-16 ***
## Polio                          6.264e-03  5.207e-03   1.203  0.22916
## Total.expenditure              9.866e-02  4.108e-02   2.402  0.01644 *
```

```
## Diphtheria                          1.496e-02  6.171e-03   2.424  0.01546 *
## `HIV/AIDS`                         -4.636e-01  1.853e-02 -25.018  < 2e-16 ***
## GDP                                 1.545e-05  2.843e-05   0.543  0.58688
## Population                         -5.831e-10  1.748e-09  -0.334  0.73878
## thinness.1.19.years                -1.030e-02  5.301e-02  -0.194  0.84595
## thinness.5.9.years                 -5.172e-02  5.237e-02  -0.988  0.32349
## Income.composition.of.resources     1.134e+01  8.856e-01  12.810  < 2e-16 ***
## Schooling                           8.661e-01  6.093e-02  14.215  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.558 on 1567 degrees of freedom
## Multiple R-squared:  0.8402, Adjusted R-squared:  0.8381
## F-statistic: 411.9 on 20 and 1567 DF,  p-value: < 2.2e-16
```



- There are some variables which seem to be insignificant.
- The correlation plot suggests the presence of multicollinearity.
- The qq plot suggests that the data is nearly normal. We will later check to see if a transformation can make the data better.
- The residual vs fitted plot indicates some of the residuals are extremely high. This raises the suspicion for outliers in the data.

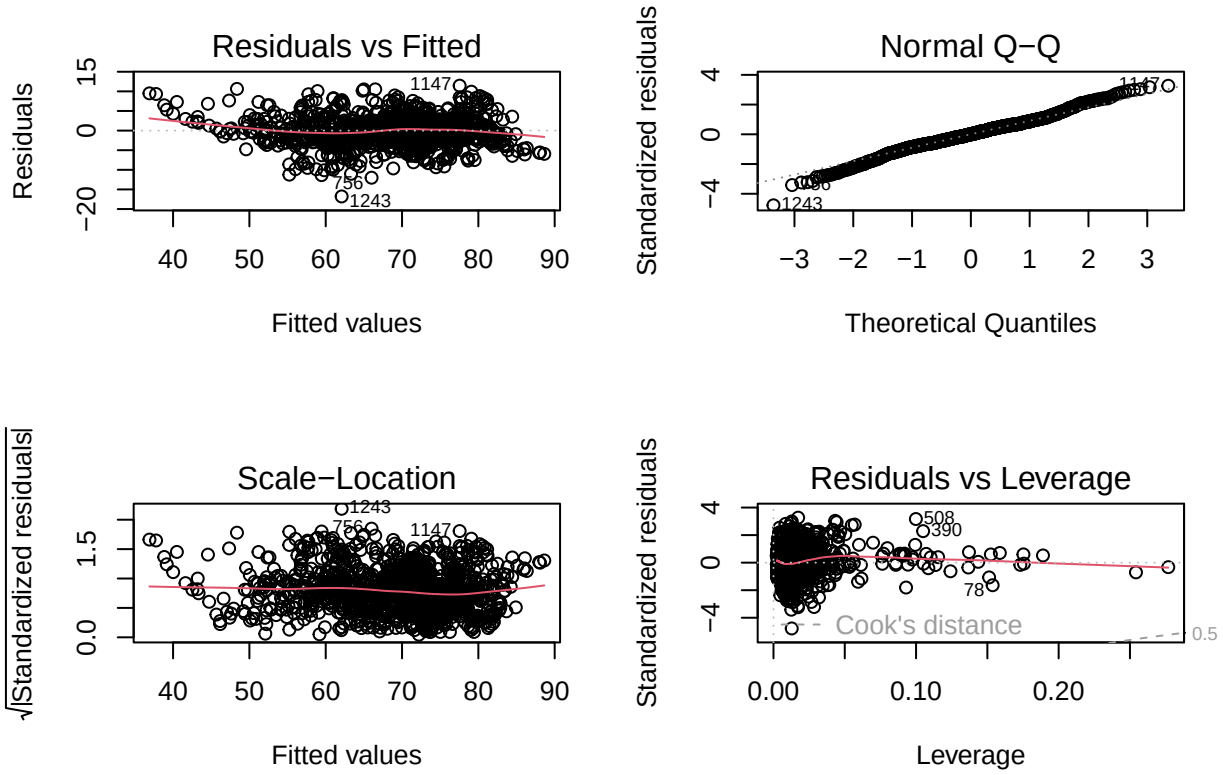
  We will explore these issues next.

Splitting the data

Before proceeding with further data analysis, we will split the data into training and testing sets with an 80:20 ratio to evaluate the model's performance on new, unseen data and prevent overfitting, ensuring its ability to generalize beyond the training set.

Following is the summary of model fit on the training data.

```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.7897  -2.1294   0.0213   2.1490  11.4495
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.047e+02  5.547e+01   5.492 4.80e-08 ***
## Year                           -1.242e-01  2.770e-02  -4.483 8.05e-06 ***
## Status                         -1.067e+00  3.824e-01  -2.790  0.00535 **
## Adult.Mortality                -1.639e-02  1.069e-03 -15.329  < 2e-16 ***
## infant.deaths                   9.066e-02  1.152e-02   7.872 7.54e-15 ***
## Alcohol                        -1.217e-01  3.826e-02  -3.182  0.00150 **
## percentage.expenditure          3.655e-04  1.919e-04   1.905  0.05698 .
## Hepatitis.B                    -4.531e-03  5.151e-03  -0.880  0.37925
## Measles                        -1.545e-05  1.215e-05  -1.272  0.20370
## BMI                             3.337e-02  6.886e-03   4.845 1.42e-06 ***
## under.five.deaths              -6.787e-02  8.349e-03  -8.129 1.03e-15 ***
## Polio                           5.491e-03  5.976e-03   0.919  0.35838
## Total.expenditure               6.038e-02  4.497e-02   1.343  0.17963
## Diphtheria                      1.208e-02  6.948e-03   1.739  0.08235 .
## `HIV/AIDS`                     -4.415e-01  2.052e-02 -21.519  < 2e-16 ***
## GDP                             1.085e-05  3.015e-05   0.360  0.71895
## Population                     -7.992e-10  1.827e-09  -0.438  0.66177
## thinness.1.19.years             3.360e-02  5.995e-02   0.560  0.57529
## thinness.5.9.years             -9.987e-02  5.950e-02  -1.678  0.09350 .
## Income.composition.of.resources  1.183e+01  1.021e+00  11.590  < 2e-16 ***
## Schooling                       8.246e-01  6.875e-02  11.994  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.537 on 1249 degrees of freedom
## Multiple R-squared:  0.8459, Adjusted R-squared:  0.8434
## F-statistic: 342.7 on 20 and 1249 DF,  p-value: < 2.2e-16
```



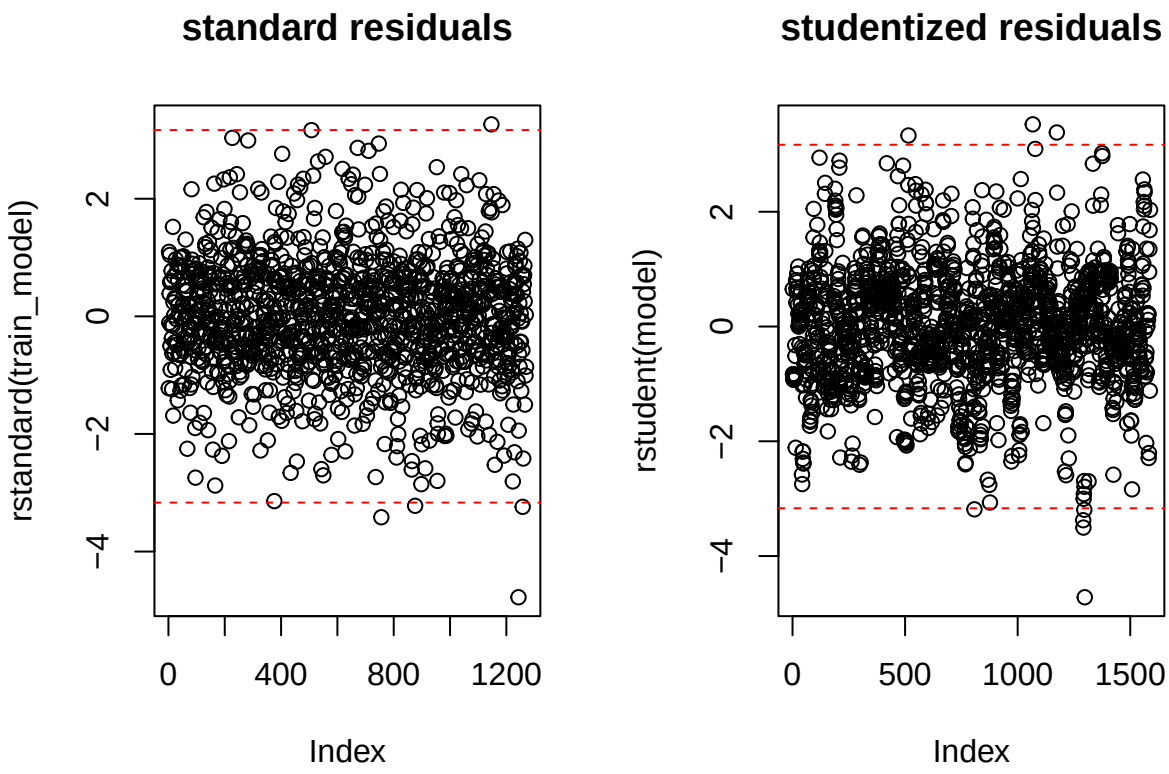
- **Residual Analysis**

  The data has n= 1270 and p =21 predictors right now.

  We investigate the presence of outliers first using *standard residuals and studentised residuals.*

The cutoff we use is (1-1/n)% quantile for t_{n-p-1} distribution which turns out to be 2.498. (red lines in the plot)

By utilizing the "rstandard" and "rstudent" functions, we examine potential outliers within the dataset and identify 6 out of 1270 observations as outliers. Opting to address these outliers separately, our scrutiny begins with an analysis of their behaviors.



Upon scrutinizing these outliers, it becomes apparent that Sierra Leone accounts for more than half of the identified instances with life expectancy values between 39 and 54, with the average being 46. Initiating a closer examination, Sierra Leone stands out primarily due to an exceptionally low life expectancy value recorded between 2000 and 2015. Delving deeper, we uncover that Sierra Leone faced formidable challenges during this period, including civil conflict, a substantial disease burden, and pervasive poverty. Specifically, during this period, Sierra Leone underwent post-civil war recovery (2002-2007), followed by an Ebola outbreak (2014-2016). Sierra Leone serves as a case study to understand the negative impacts on life expectancy, allowing us to formulate policies tailored to address such challenges.

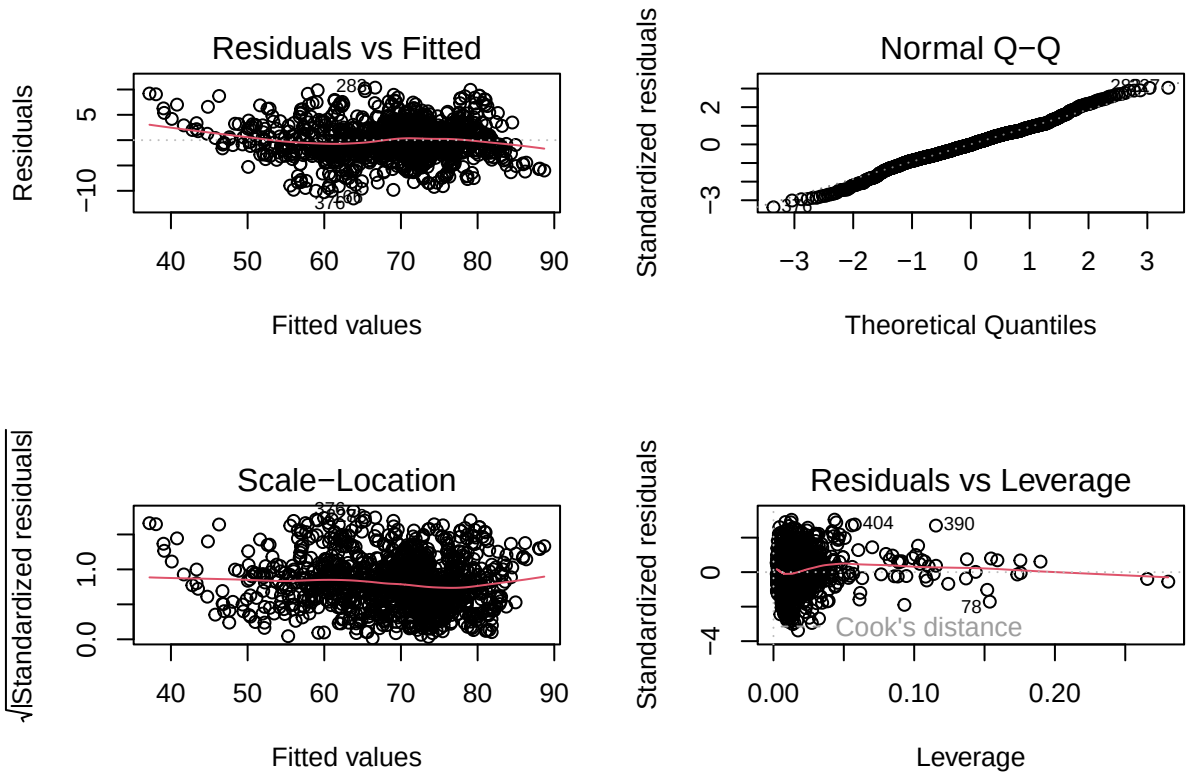
```
## # A tibble: 6 x 23
##   Country      Year Status Life.expectancy Adult.Mortality infant.deaths Alcohol
##   <fct>       <dbl>  <int>           <dbl>           <dbl>         <dbl>   <dbl>
## 1 Nigeria      2008      2            59               386           536    9.3
## 2 Sierra Leo~  2013      2            54                47            23    0.01
## 3 Malawi       2002      2            44                67            46    1.1
## 4 France       2007      2            89                89             3   12.2
## 5 Sierra Leo~  2007      2            45.3              45            29    3.86
## 6 Sierra Leo~  2014      2            48.1             463            23    0.01
## # i 16 more variables: percentage.expenditure <dbl>, Hepatitis.B <dbl>,
## #   Measles <dbl>, BMI <dbl>, under.five.deaths <dbl>, Polio <dbl>,
## #   Total.expenditure <dbl>, Diphtheria <dbl>, 'HIV/AIDS' <dbl>, GDP <dbl>,
## #   Population <dbl>, thinness.1.19.years <dbl>, thinness.5.9.years <dbl>,
## #   Income.composition.of.resources <dbl>, Schooling <dbl>, le1 <dbl>
```

On the other hand, France exhibits an exceptionally high life expectancy value of 89, serving as an example of positive impacts on life expectancy. Upon closer inspection of the data for France, it is observed that

indicators such as death under five and HIV/AIDS have the lowest values compared to those of other countries. Meanwhile, variables such as Income composition of resources, schooling, population, diphtheria, BMI, Polio, and alcohol are among the highest. These indications suggest the potential for regression models based on these variables, highlighting avenues for further exploration and analysis.

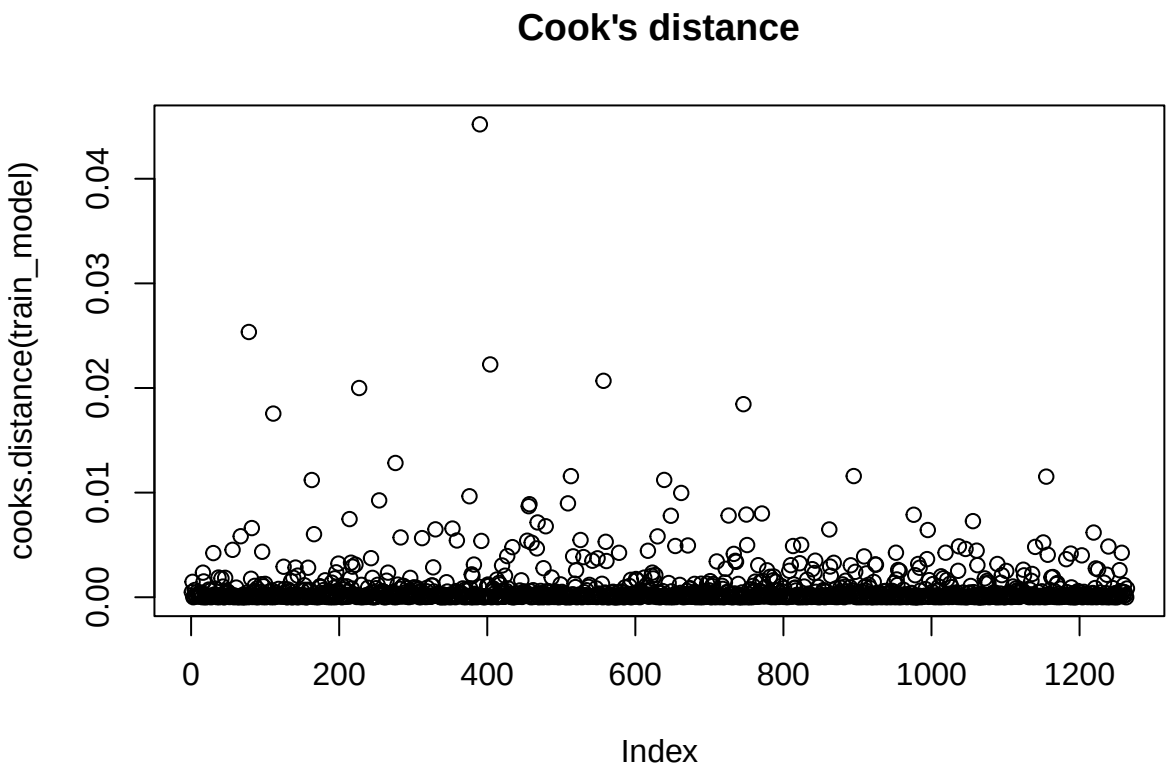
The summary of the model after removing the outliers.

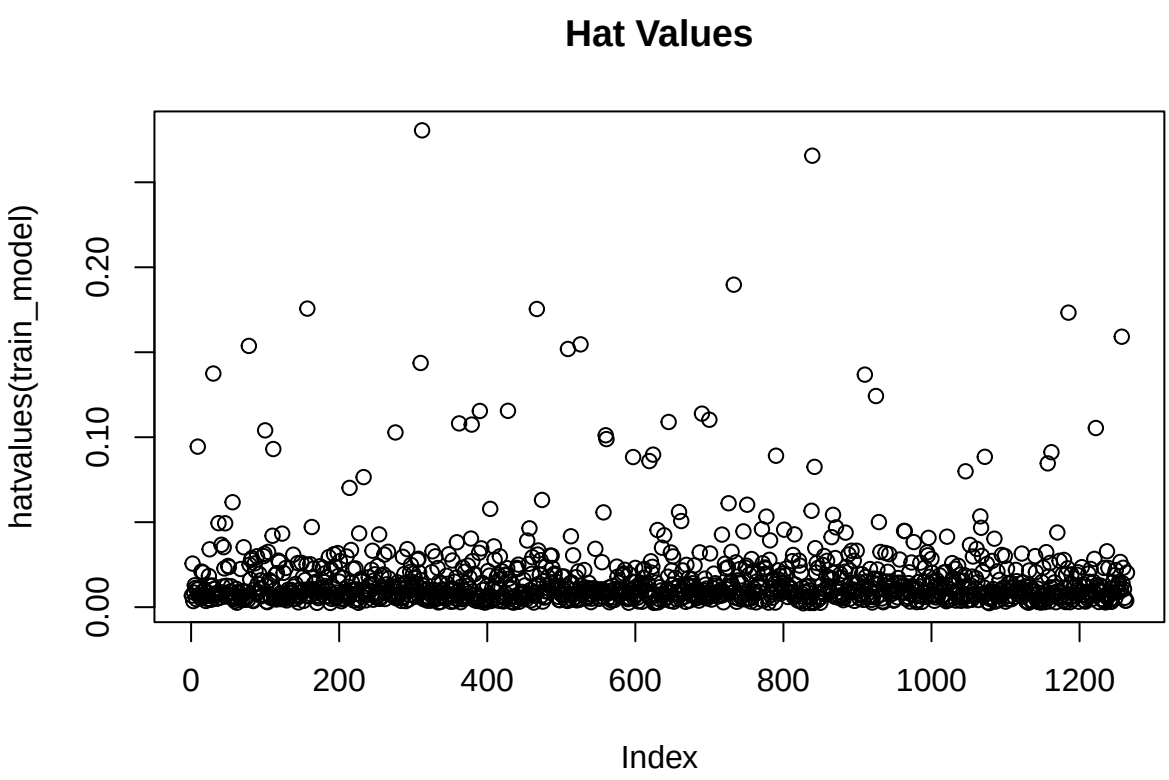
```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5074  -2.1227  -0.0371   2.1613  10.3592
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.939e+02  5.398e+01   5.444 6.28e-08 ***
## Year                           -1.185e-01  2.696e-02  -4.395 1.20e-05 ***
## Status                         -1.167e+00  3.720e-01  -3.137  0.00175 **
## Adult.Mortality                -1.700e-02  1.052e-03 -16.160  < 2e-16 ***
## infant.deaths                   9.811e-02  1.149e-02   8.538  < 2e-16 ***
## Alcohol                        -1.295e-01  3.728e-02  -3.474  0.00053 ***
## percentage.expenditure          3.626e-04  1.864e-04   1.946  0.05193 .
## Hepatitis.B                    -3.495e-03  5.015e-03  -0.697  0.48605
## Measles                        -1.607e-05  1.180e-05  -1.362  0.17346
## BMI                             3.083e-02  6.697e-03   4.603 4.58e-06 ***
## under.five.deaths              -7.382e-02  8.369e-03  -8.820  < 2e-16 ***
## Polio                           6.968e-03  5.820e-03   1.197  0.23148
## Total.expenditure               7.974e-02  4.411e-02   1.808  0.07092 .
## Diphtheria                      9.956e-03  6.761e-03   1.473  0.14111
## `HIV/AIDS`                     -4.318e-01  2.017e-02 -21.407  < 2e-16 ***
## GDP                             1.227e-05  2.929e-05   0.419  0.67543
## Population                     -3.311e-10  1.779e-09  -0.186  0.85241
## thinness.1.19.years             3.215e-02  5.823e-02   0.552  0.58104
## thinness.5.9.years             -9.884e-02  5.780e-02  -1.710  0.08747 .
## Income.composition.of.resources  1.142e+01  9.929e-01  11.497  < 2e-16 ***
## Schooling                       8.157e-01  6.683e-02  12.206  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.436 on 1243 degrees of freedom
## Multiple R-squared:  0.8518, Adjusted R-squared:  0.8494
## F-statistic: 357.2 on 20 and 1243 DF,  p-value: < 2.2e-16
```



Given the relatively low number of outliers in comparison to the dataset, the choice was made to eliminate these outliers. Subsequent to this outlier removal, a meticulous examination of influential observations was conducted using Cook's distance and hat value. Since the overall model appears to be well-fitted, we utilize Cook's distance as a metric to assess the influence of individual observations on the regression model. 50% quantile for F_p,n-p is distribution is 0.9688527

10% quantile for F_p,n-p is distribution is 0.629304


Cook's distance
cooks.distance(train_model)
0.00
0.01
0.02
0.03
0.04
0
200
400
600
800
1000
1200
Index



The results from both metrics indicate the absence of influential observations. With this confirmation, we proceed to the next phase of analysis.

Multicollinearity.

```
##                                 sort.VIF.lm_train...decreasing...TRUE.
## infant.deaths                                               232.811664
## under.five.deaths                                           222.157207
## GDP                                                          13.858565
## percentage.expenditure                                       12.950975
## thinness.5.9.years                                            7.764920
## thinness.1.19.years                                           7.641826
## Schooling                                                     3.753012
## Income.composition.of.resources                               3.387432
## Alcohol                                                       2.430942
## Diphtheria                                                    2.203107
## Population                                                    2.114794
## Adult.Mortality                                               1.923149
## Status                                                        1.891969
## BMI                                                           1.875111
## Polio                                                         1.774716
## Hepatitis.B                                                   1.752430
## Measles                                                       1.609341
## `HIV/AIDS`                                                    1.564262
## Year                                                          1.157609
## Total.expenditure                                             1.118153
```

It is not surprising to note that the variable pairs: infant.deaths and under.five.deaths, GDP and percentage.expenditure have high VIF indicating the presence of multicollinearity.

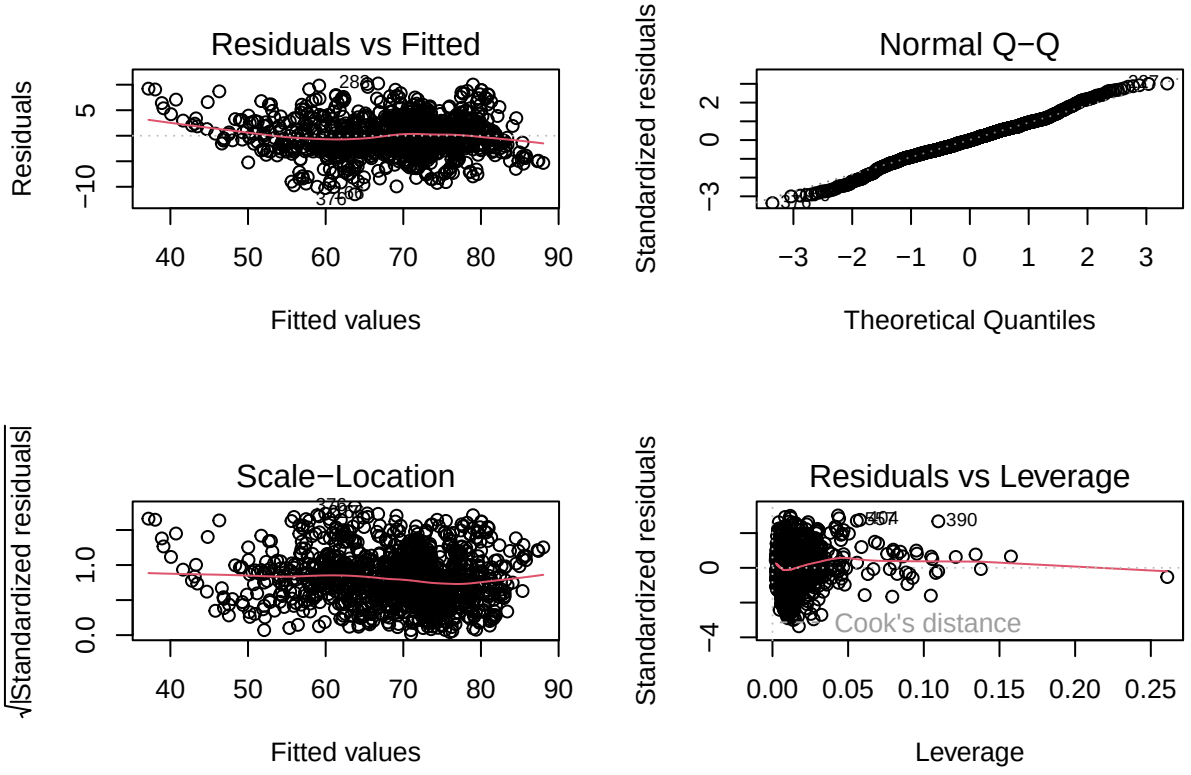
According to WHO, about 60% of deaths of infants are part of under.five.deaths. That explains the high correlation between these two variables.We choose to eliminate the redundant variable (infant deaths).

Percentage of expenditure is defined as the percentage of GDP being spent on the healthcare system of the country. Because of this overlap, the variables depict high correlation and hence we eliminate one of them. (percentage expenditure)

Another pair that seems to be obviously highly related is thinness in age group 5-9 and thinness in the age group 10-19.This variable, as defined in the data source, captured the 'thinness' of children in that age group. We would ideally believe that this has some sort of effect on life expectancy, but there was no further information on the units they were measured in. If it were to involve any bodily measurements, we already have it recorded in BMI. In that case both the variables might be redundant. On the other hand, if it compares waist, hip and other relevant body parts width measurements, then it definitely stands as a separate entity.

Despite the fact that both are relevant in the 2nd case, we believe it is fair to exclude the age group 10-19 for two reasons:First, at this stage, the thinness is dependent on the lifestyle of these kids' lives. The country as a whole will have no part affecting the life expectancy of these kids.Second, the thinness at an older age can be also caused due to genetic reasons, or faster metabolism. It is hard to pin down the fact that this variable can provide a logical explanation about life expectancy.

Model after fitting removing the correlated variables.




```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = train)
```

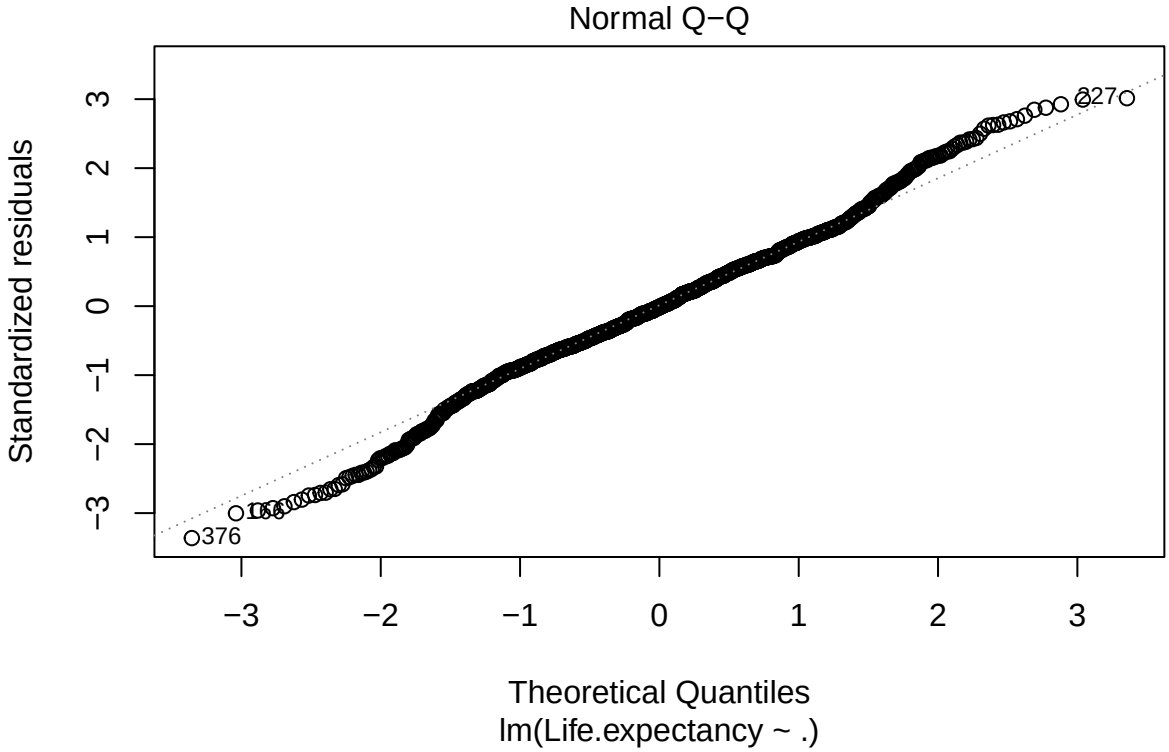
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.467  -2.079  -0.049   2.162  10.241
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.990e+02  5.389e+01   5.548 3.53e-08 ***
## Year                           -1.210e-01  2.692e-02  -4.495 7.60e-06 ***
## Status                         -1.191e+00  3.724e-01  -3.198 0.001420 **
## Adult.Mortality                -1.718e-02  1.050e-03 -16.355  < 2e-16 ***
## infant.deaths                   9.527e-02  1.118e-02   8.521  < 2e-16 ***
## Alcohol                        -1.282e-01  3.733e-02  -3.433 0.000616 ***
## Hepatitis.B                    -4.084e-03  5.016e-03  -0.814 0.415615
## Measles                        -1.403e-05  1.170e-05  -1.200 0.230557
## BMI                             3.266e-02  6.625e-03   4.930 9.33e-07 ***
## under.five.deaths              -7.205e-02  8.259e-03  -8.724  < 2e-16 ***
## Polio                           7.371e-03  5.818e-03   1.267 0.205361
## Total.expenditure               8.620e-02  4.410e-02   1.955 0.050857 .
## Diphtheria                      1.014e-02  6.770e-03   1.498 0.134443
## `HIV/AIDS`                     -4.322e-01  2.019e-02 -21.405  < 2e-16 ***
## GDP                             6.643e-05  9.759e-06   6.806 1.55e-11 ***
## thinness.1.19.years            -5.272e-02  3.004e-02  -1.755 0.079500 .
## Income.composition.of.resources  1.133e+01  9.936e-01  11.405  < 2e-16 ***
## Schooling                       8.023e-01  6.656e-02  12.055  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.441 on 1246 degrees of freedom
## Multiple R-squared:  0.851,  Adjusted R-squared:  0.849
## F-statistic: 418.6 on 17 and 1246 DF,  p-value: < 2.2e-16
```

Normality

The following is the Q-Q plot for the model on training data.

```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.467  -2.079  -0.049   2.162  10.241
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.990e+02  5.389e+01   5.548 3.53e-08 ***
## Year                           -1.210e-01  2.692e-02  -4.495 7.60e-06 ***
## Status                         -1.191e+00  3.724e-01  -3.198 0.001420 **
## Adult.Mortality                -1.718e-02  1.050e-03 -16.355  < 2e-16 ***
## infant.deaths                   9.527e-02  1.118e-02   8.521  < 2e-16 ***
## Alcohol                        -1.282e-01  3.733e-02  -3.433 0.000616 ***
## Hepatitis.B                    -4.084e-03  5.016e-03  -0.814 0.415615
## Measles                        -1.403e-05  1.170e-05  -1.200 0.230557
```

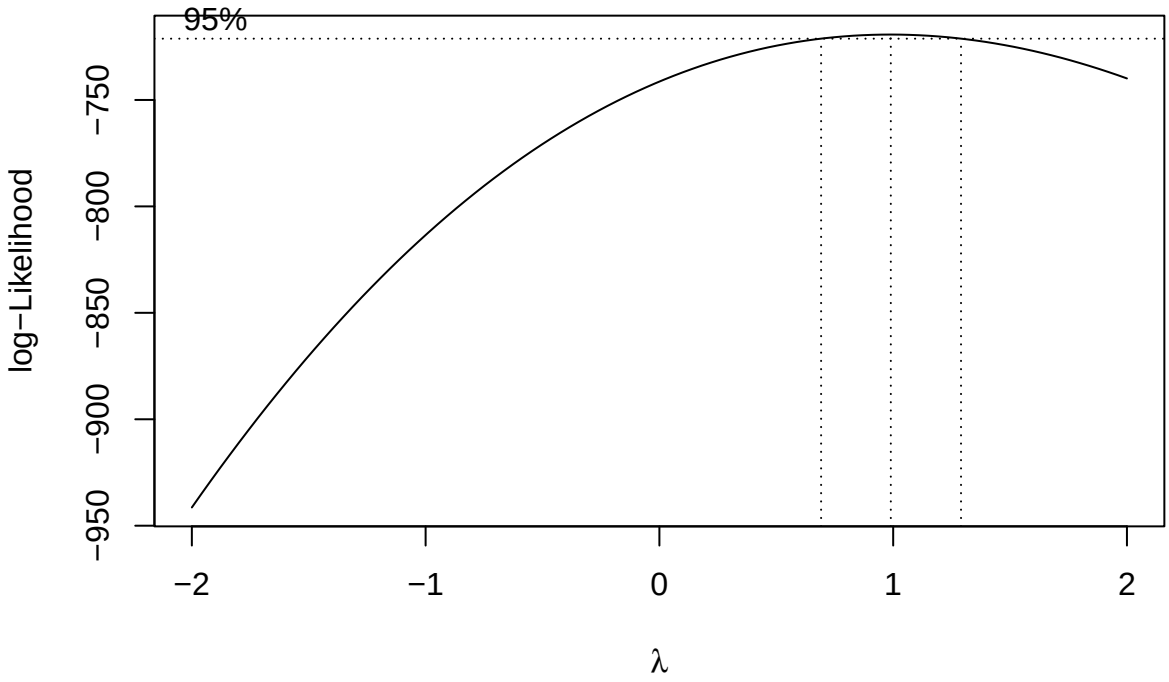
```
## BMI                               3.266e-02  6.625e-03   4.930 9.33e-07 ***
## under.five.deaths                -7.205e-02  8.259e-03  -8.724  < 2e-16 ***
## Polio                             7.371e-03  5.818e-03   1.267 0.205361
## Total.expenditure                 8.620e-02  4.410e-02   1.955 0.050857 .
## Diphtheria                        1.014e-02  6.770e-03   1.498 0.134443
## `HIV/AIDS`                       -4.322e-01  2.019e-02 -21.405  < 2e-16 ***
## GDP                               6.643e-05  9.759e-06   6.806 1.55e-11 ***
## thinness.1.19.years              -5.272e-02  3.004e-02  -1.755 0.079500 .
## Income.composition.of.resources   1.133e+01  9.936e-01  11.405  < 2e-16 ***
## Schooling                         8.023e-01  6.656e-02  12.055  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.441 on 1246 degrees of freedom
## Multiple R-squared:  0.851,  Adjusted R-squared:  0.849
## F-statistic: 418.6 on 17 and 1246 DF,  p-value: < 2.2e-16
```



p-value of Kolmogorov-Smirnov test is 0.06748. This is just barely above the 0.05 threshold.

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  sres
## D = 0.036614, p-value = 0.06748
## alternative hypothesis: two-sided
```

We now look at the BoxCox plot to see if a transformation can improve the normality. The best value of lambda is very close to 1, which suggests us to not transform the variables.



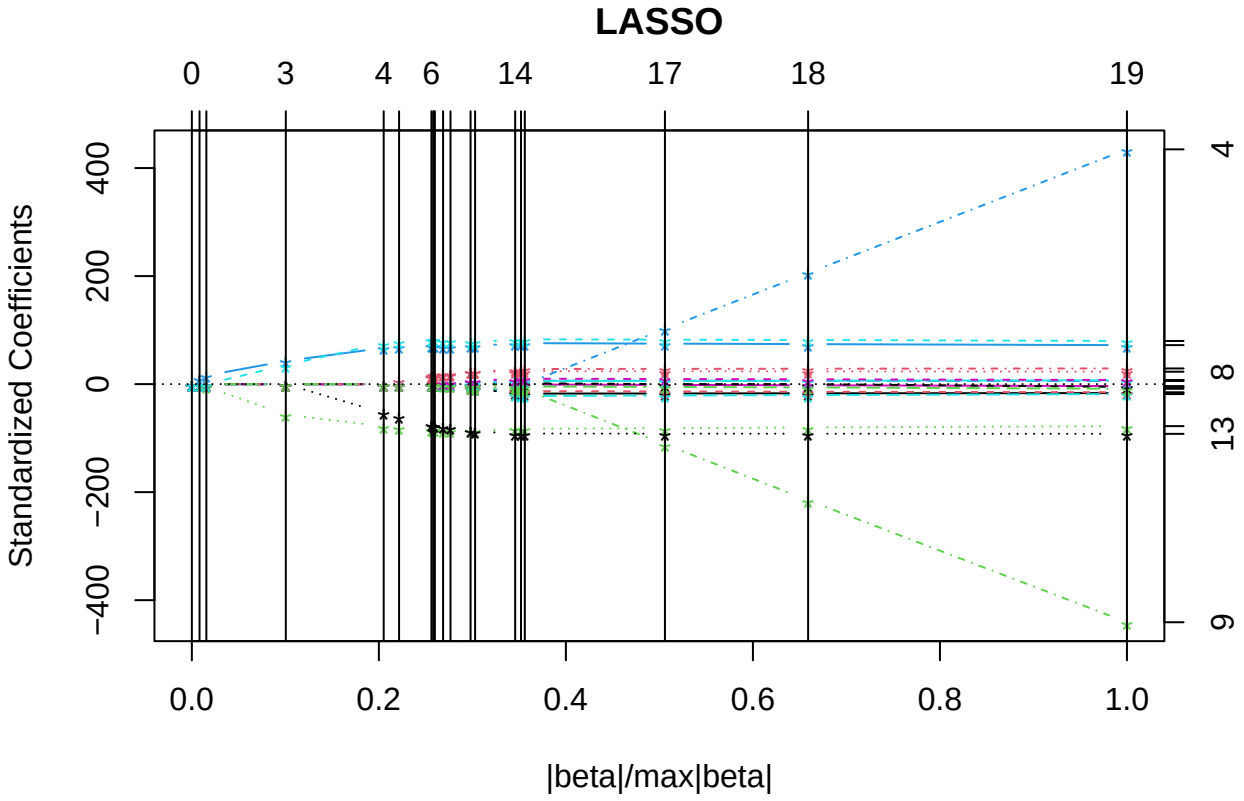
Variable Selection

Given Lasso's capability to induce sparsity in the model by driving certain coefficients to zero, it serves as a potent tool for both variable selection and the identification of multicollinearity. Consequently, we will utilize this method to perform variable selection and address multicollinearity in our modeling process

```
## Loaded lars 1.3
```

```
##
## Attaching package: 'lars'
```

```
## The following object is masked from 'package:psych':
##
##     error.bars
```



Root Mean Square for the Lasso model.

Ridge Regression

```
## [1] 12.41944
```

```
##                                  sorted_coef
## (Intercept)                     2.755773e+02
## Income.composition.of.resources  1.131396e+01
## Schooling                        7.556512e-01
## Total.expenditure                7.287476e-02
## BMI                              3.790075e-02
## Diphtheria                       1.615334e-02
## Polio                            1.142253e-02
## infant.deaths                    2.120245e-03
## GDP                              6.438820e-05
## Measles                          6.903660e-06
## Hepatitis.B                     -2.840249e-03
## under.five.deaths               -3.115502e-03
## Adult.Mortality                 -1.783977e-02
## thinness.1.19.years             -4.766145e-02
## Year                            -1.095834e-01
## Alcohol                         -1.291136e-01
## HIV/AIDS                        -4.106652e-01
## Status                          -1.189118e+00
```

Principal Least Square Regression

```
##
## Attaching package: 'pls'

## The following object is masked from 'package:corrplot':
##
##     corrplot

## The following object is masked from 'package:stats':
##
##     loadings

## [1] 15.53798
```

Ordinary least squares

```
## [1] 13.33926

##
## Call:
## lm(formula = Life.expectancy ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.467  -2.079  -0.049   2.162  10.241
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.990e+02  5.389e+01   5.548 3.53e-08 ***
## Year                           -1.210e-01  2.692e-02  -4.495 7.60e-06 ***
## Status                         -1.191e+00  3.724e-01  -3.198 0.001420 **
## Adult.Mortality                -1.718e-02  1.050e-03 -16.355  < 2e-16 ***
## infant.deaths                   9.527e-02  1.118e-02   8.521  < 2e-16 ***
## Alcohol                        -1.282e-01  3.733e-02  -3.433 0.000616 ***
## Hepatitis.B                    -4.084e-03  5.016e-03  -0.814 0.415615
## Measles                        -1.403e-05  1.170e-05  -1.200 0.230557
## BMI                             3.266e-02  6.625e-03   4.930 9.33e-07 ***
## under.five.deaths              -7.205e-02  8.259e-03  -8.724  < 2e-16 ***
## Polio                           7.371e-03  5.818e-03   1.267 0.205361
## Total.expenditure               8.620e-02  4.410e-02   1.955 0.050857 .
## Diphtheria                      1.014e-02  6.770e-03   1.498 0.134443
## `HIV/AIDS`                     -4.322e-01  2.019e-02 -21.405  < 2e-16 ***
## GDP                             6.643e-05  9.759e-06   6.806 1.55e-11 ***
## thinness.1.19.years            -5.272e-02  3.004e-02  -1.755 0.079500 .
## Income.composition.of.resources  1.133e+01  9.936e-01  11.405  < 2e-16 ***
## Schooling                       8.023e-01  6.656e-02  12.055  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.441 on 1246 degrees of freedom
## Multiple R-squared:  0.851,  Adjusted R-squared:  0.849
## F-statistic: 418.6 on 17 and 1246 DF,  p-value: < 2.2e-16
```

Upon examining the root mean square error (RMSE), it was evident that ridge regression yielded the optimal RMSE, utilizing a model encompassing all 18 variables. Notably, variables such as income composition of resources, schooling, and total expenditure exhibited the most significant positive influence in contrast to factors like development status (developing/developed), HIV/AIDS, and alcohol use. Interestingly, unlike the observed outliers, the highest levels of alcohol use were positively correlated with higher life expectancy values.

Conclusion In summary, our exploration of the Life Expectancy dataset involved a thorough analysis of correlation relationships, detection of outliers, and addressing multicollinearity. Splitting the data into training and testing sets allowed us to assess the model's performance and address potential overfitting. Notably, Sierra Leone emerged as a significant outlier with unique challenges impacting life expectancy, while France stood out as a positive example.

The removal of outliers and examination of influential observations confirmed the overall model's fitness. Multicollinearity was addressed by eliminating redundant variables, enhancing the accuracy of subsequent analyses. Ridge regression, determined through RMSE evaluation, yielded an optimal model encompassing all 18 variables. Key influencers on life expectancy included income composition of resources, schooling, and total expenditure.

Surprisingly, the analysis revealed that the highest levels of alcohol use were positively correlated with higher life expectancy values, challenging conventional outlier patterns. This comprehensive study provides valuable insights for policymakers and healthcare organizations, offering evidence-based recommendations to enhance public health outcomes and life expectancy.

In our attempt to incorporate one year's life expectancy as a predictor for the next, akin to an AR(1) time series model, we discovered that including this variable suggests the need for a nonlinear regression model. However, this falls beyond our current knowledge scope, and we express the intention to delve into this complex model in future research endeavors. Exploring the intricacies of a nonlinear regression model could provide valuable insights into the dynamic relationships affecting life expectancy over time, opening avenues for more nuanced and accurate predictions.