Final Project- Life Expectancy Estimation Analysis

Leo Hong

May 09, 2019

1.Introduction

Life expectancy is a measure of the average time an organism is expected to live based on the year of its birth, its current age and other many factors. Life expectancy is often used to gauge the overall health of one nation. Shifts in life expectency are often used to describe the trends in mortality. Being able to predict how populations will age has enormous implications for the planning and provision of services and support. Small increases in life expectancy translate into large increases in the population. The purpose of this study is as follow. First, I explored the importance of the antecedents that contribute to life expectancy and find the highest predicting factors based on prior literature review. Second, I explored whether there are any ways to improve my model that justifies a future research. Based on extensive literature, I initially identified 11 predicting variables that are worth considering in the study. I started by building an initial model using these 11 predicting variables followed by finding the best predictors for the model. I also included possible interaction variables and compared with best regression model. For data collection, I used World Bank website, launched in 2001, as it is a reliable resource available. My dataset consists of three seperate files; GDP file, population file, and trade file. Each file contains 8 countries categorized by 4 developed countries and 4 developing countries.

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.2.1 --
```

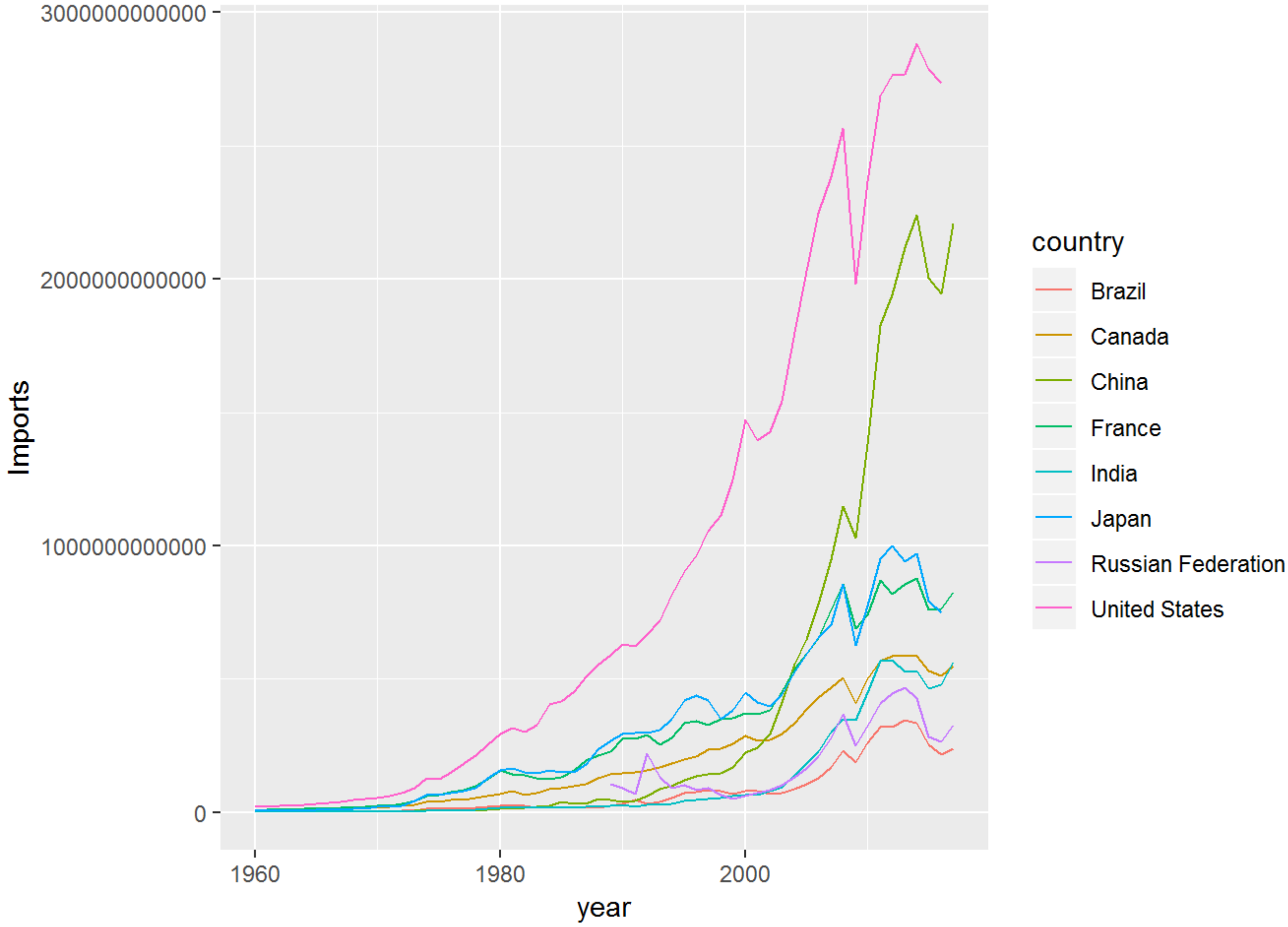
```
## v ggplot2 3.1.0      v purrr   0.3.0
## v tibble  2.0.1      v dplyr   0.8.0.1
## v tidyr   0.8.3      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.4.0
```

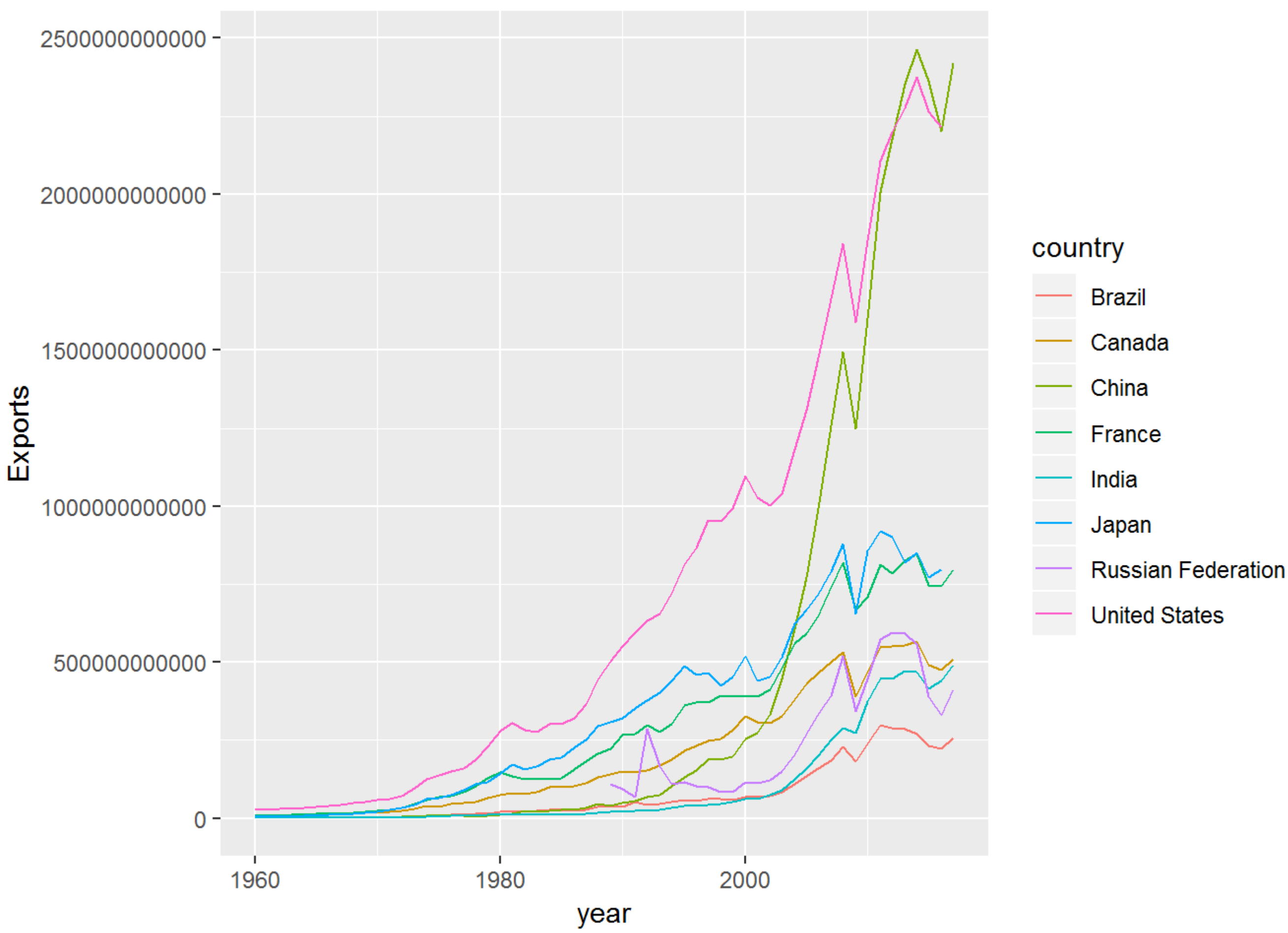
```
## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

2.Data analysis on world trade file

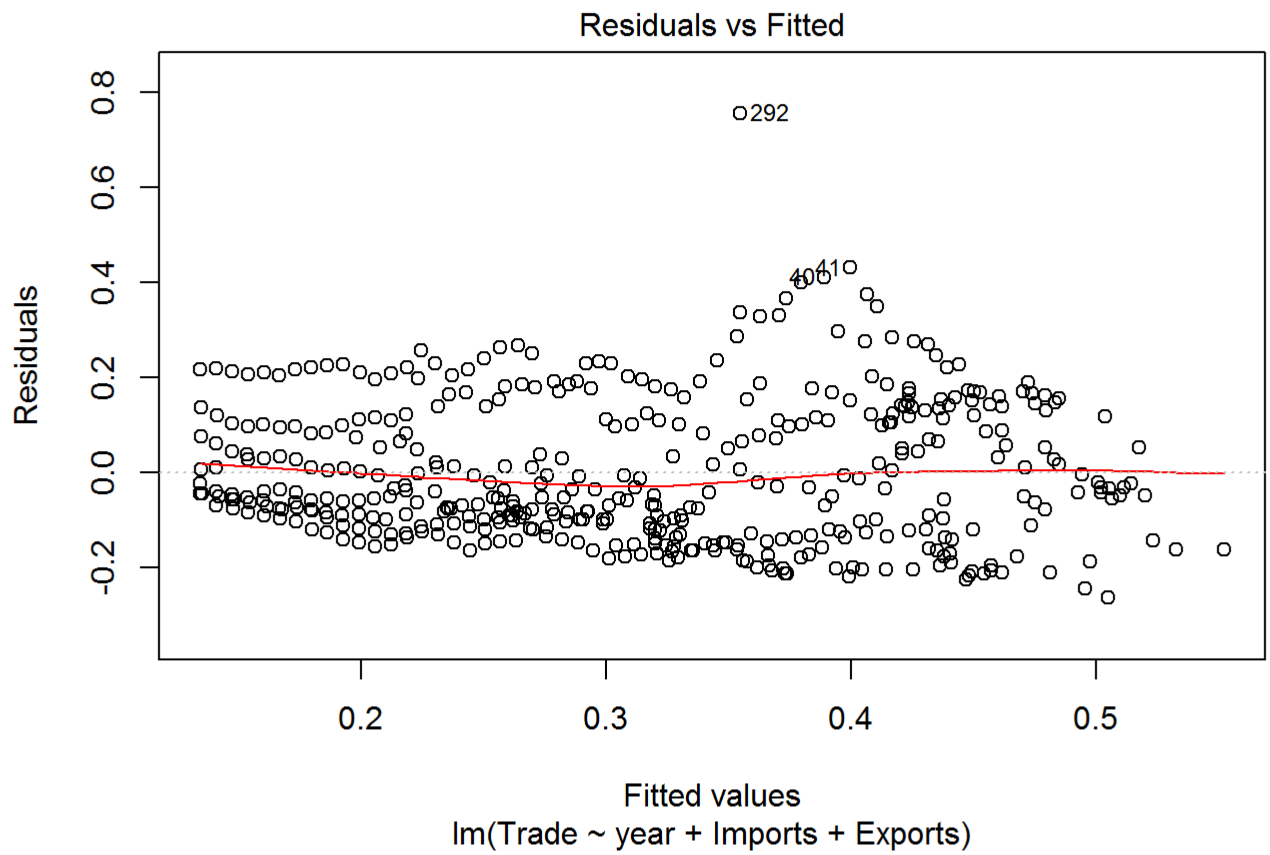
In this analysis, I analyzed world trade file. World trade file includes country, year, exports, imports and trade. Initial data cleaning was required. First, I cleaned rows containing 0 values using world_trade_new <-filter (world_trade, Imports>0, Exports>0) followed by round larger numbers down using option(scipen=999). I plotted a linear relationship to find out whether the dataset I chose is significant. I compared two models which include (y=imports, x=year) and (y=exports, x=year). After plotting, I ran a multiple regression (y=trade, x=year+imports+exports) to check the significance of dataset. Data results showed that all three variables are significant and did show a liner relationship. Imports and exports are higher in both United States and China which showed that both countries trade more goods/services than other countries.

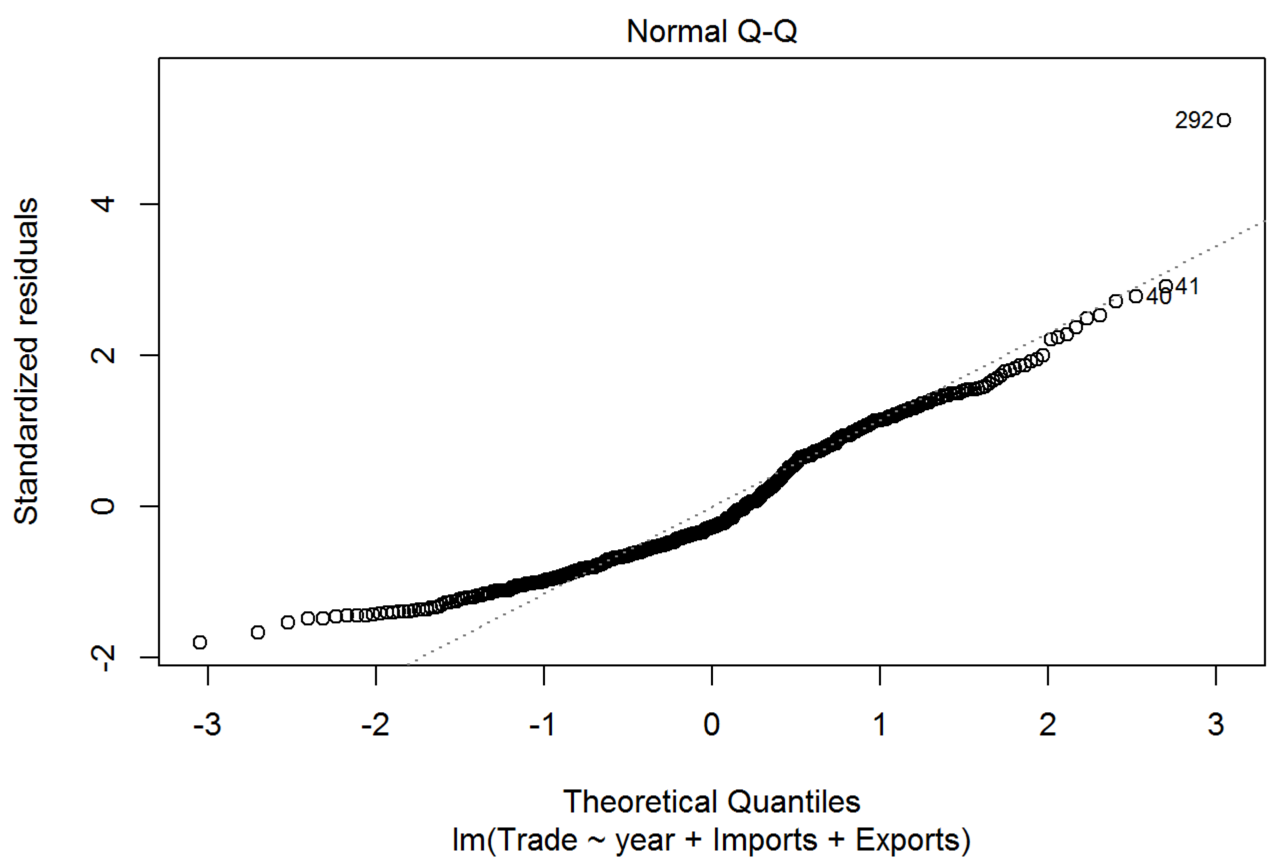
```
## Parsed with column specification:
## cols(
##   country = col_character(),
##   year = col_double(),
##   Imports = col_number(),
##   Exports = col_number(),
##   Trade = col_double()
## )
```

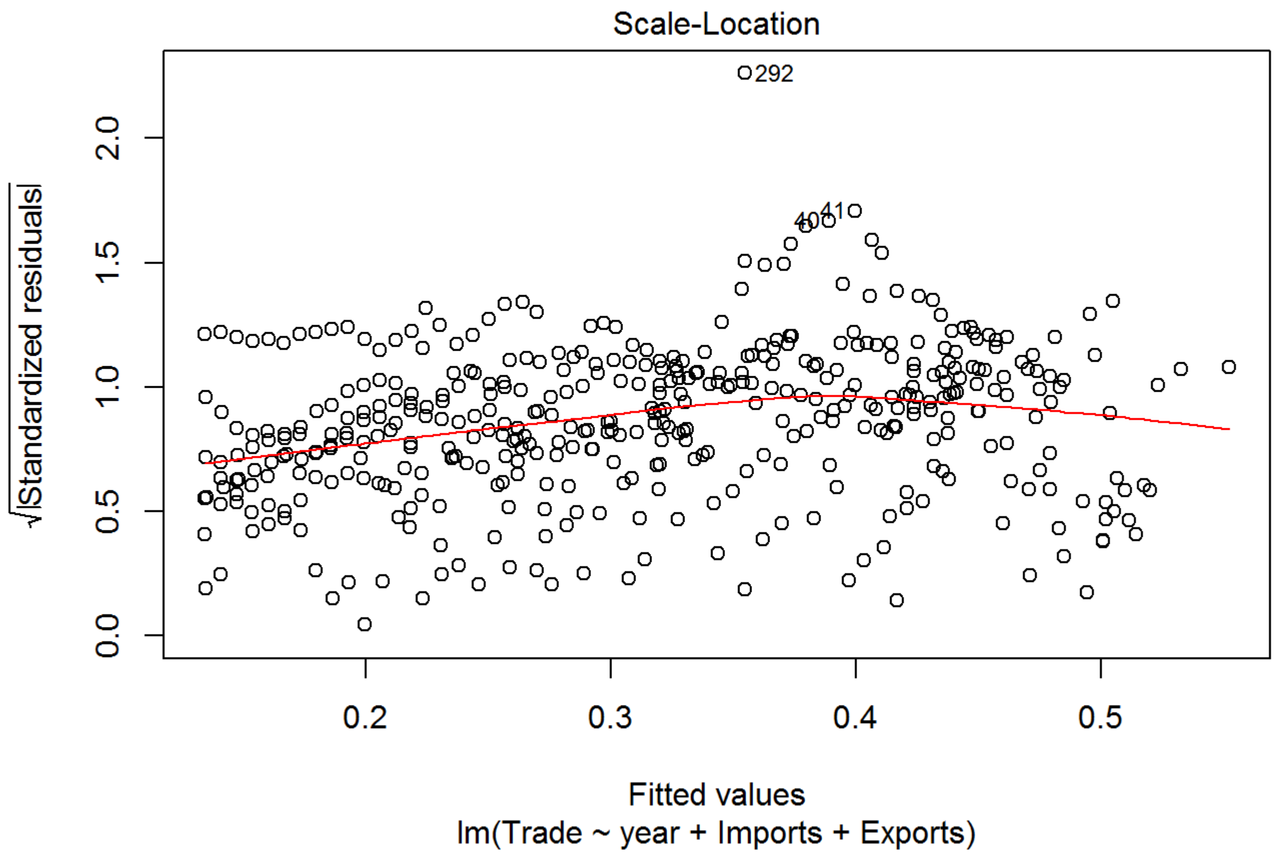

3000000000000
2000000000000
1000000000000
0
Imports
1960
1980
2000
year
country
Brazil
Canada
China
France
India
Japan
Russian Federation
United States

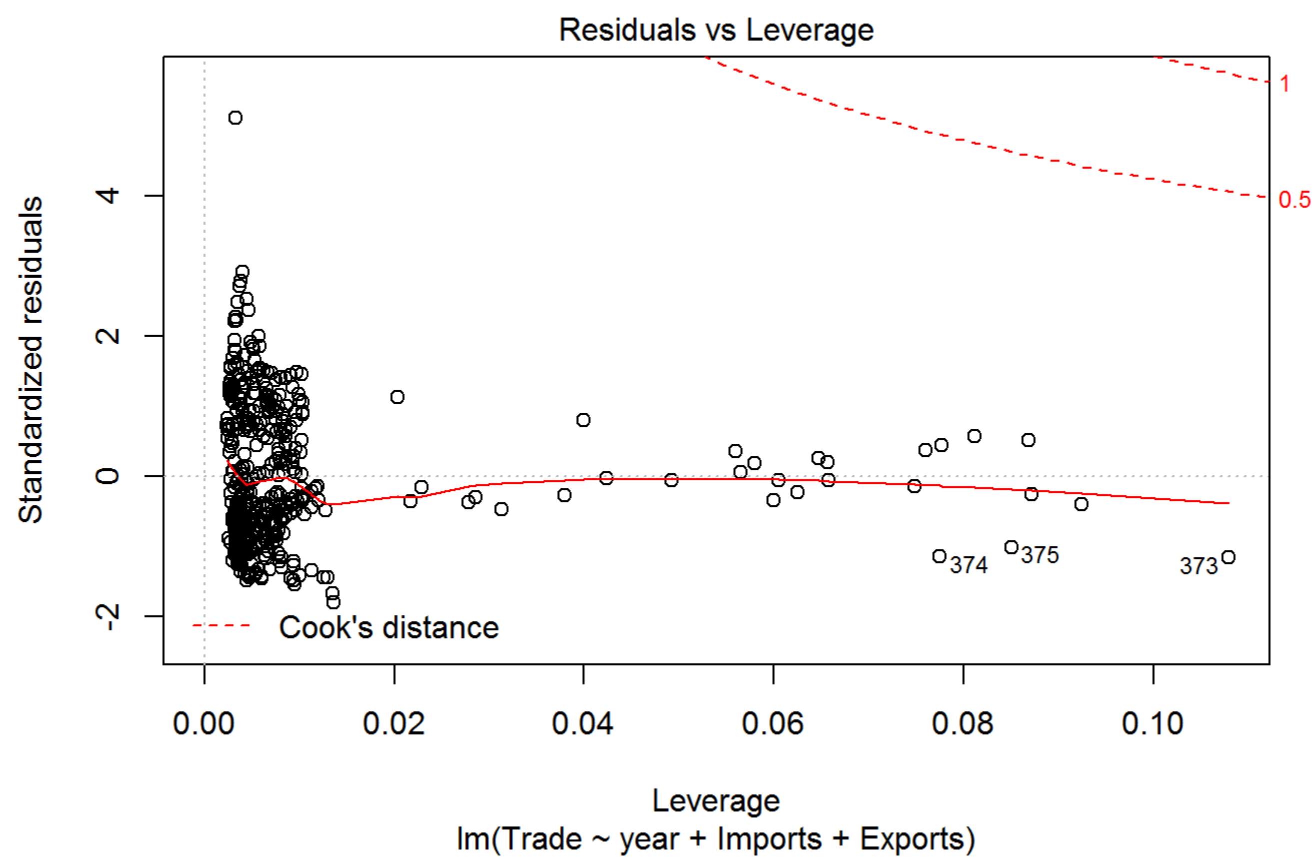

2500000000000
2000000000000
1500000000000
1000000000000
500000000000
0
Exports
1960
1980
2000
year
country
Brazil
Canada
China
France
India
Japan
Russian Federation
United States

```
##
## Call:
## lm(formula = Trade ~ year + Imports + Exports, data = world_trade_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26512 -0.11454 -0.04055  0.11418  0.75502
##
## Coefficients:
##                          Estimate           Std. Error t value
## (Intercept) -12.54451799032254300   1.09000263620837612 -11.509
## year          0.00646896470629074   0.00054985246794619  11.765
## Imports      -0.00000000000027413   0.00000000000006068  -4.518
## Exports       0.00000000000025869   0.00000000000007056   3.666
##                         Pr(>|t|)
## (Intercept) < 0.0000000000000002 ***
## year        < 0.0000000000000002 ***
## Imports                0.0000081 ***
## Exports                 0.000277 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1479 on 429 degrees of freedom
## Multiple R-squared:  0.3421, Adjusted R-squared:  0.3375
## F-statistic: 74.36 on 3 and 429 DF,  p-value: < 0.00000000000000022
```


Residuals vs Fitted
Residuals
Fitted values
lm(Trade ~ year + Imports + Exports)
292
40
41
0.8
0.6
0.4
0.2
0.0
-0.2
0.2
0.3
0.4
0.5


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Trade ~ year + Imports + Exports)
292
40
41
4
2
0
-2
-3
-2
-1
0
1
2
3

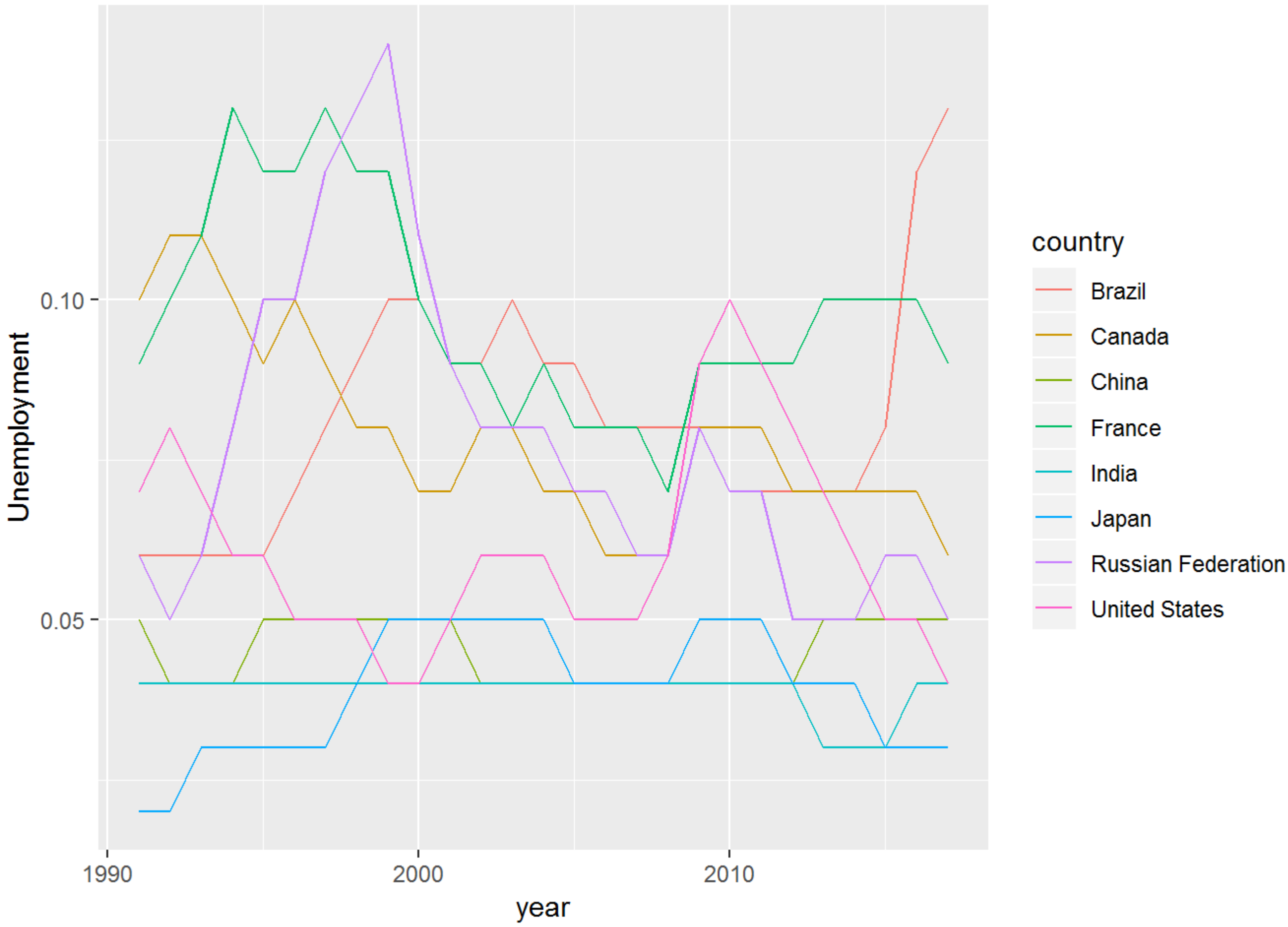

Scale-Location
292
40
41
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0.2
0.3
0.4
0.5
Fitted values
lm(Trade ~ year + Imports + Exports)


Residuals vs Leverage
1
0.5
374
375
373
Cook's distance
Standardized residuals
-2
0
2
4
0.00
0.02
0.04
0.06
0.08
0.10
Leverage
lm(Trade ~ year + Imports + Exports)

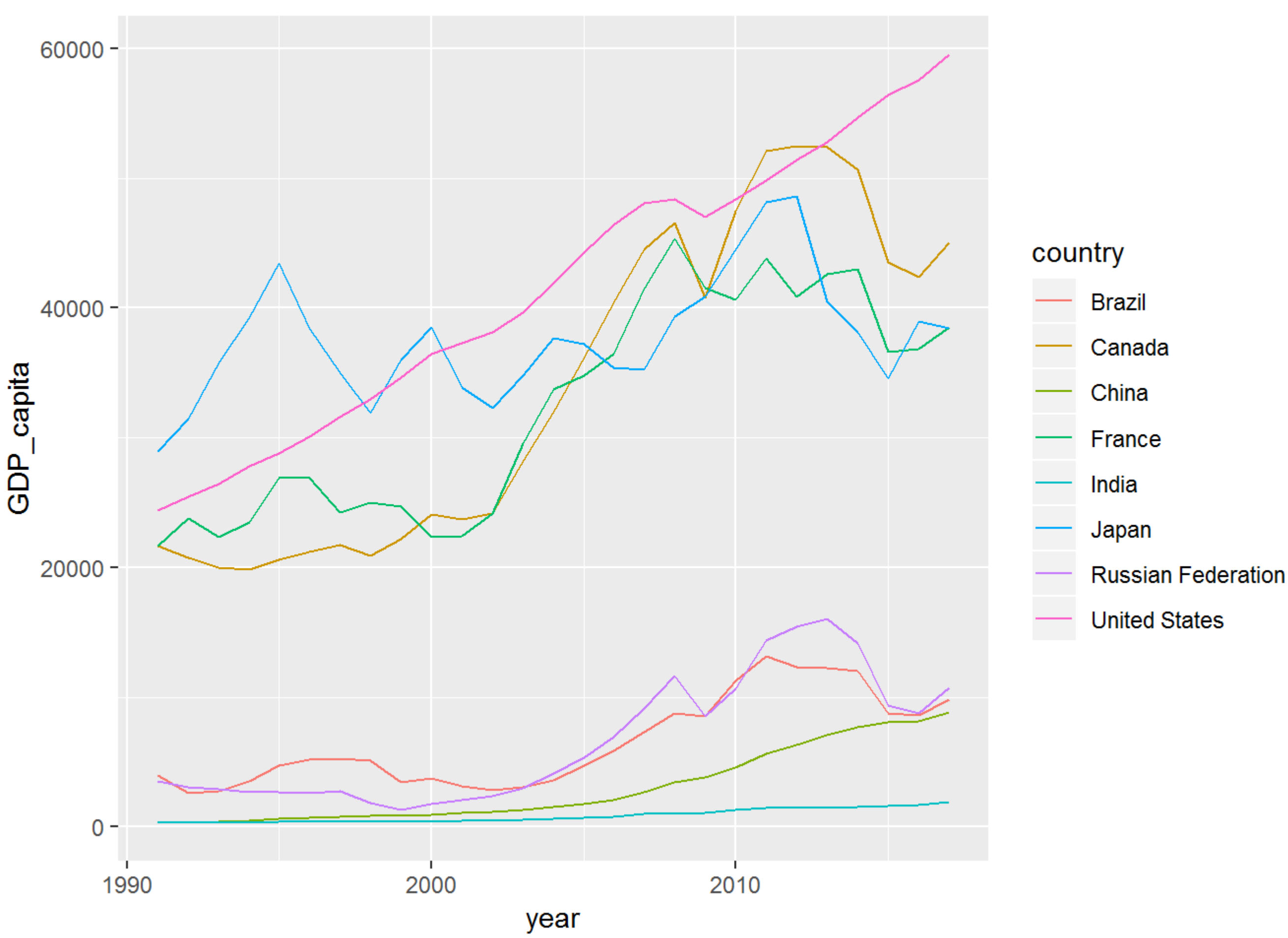
3.Data analysis on world GDP file

In this analysis, I analyzed world GDP file. World GDP file includes country, year, unemployment, GDP, and GDP_capita. Initial data cleaning was required. First, I cleaned rows containing 0 values using world_gdp_new <- filter(world_gdp, unemployment>0) followed by round larger numbers down using option(scipen=999).I plotted a linear relationship to find out whether the dataset I chose is significant. I compared two models which include (y=unemployment, x=year) and (y=GDP_capita, x=year). After plotting, I ran a multiple regression (y=unemployment, x=all X variables) to check the significance of dataset. Data results showed that two variables are significant (GDP, GDP_capita) except year. Our plot shows there is a linear relationship in the model. Fluctuation in unemployment rates are common in developing countries than developed countries. GDP_capita which measures the country's living standard (e.g. affordable power) is relatively higher in developed countries. This shows that buying power in developed countries far exceeds the buying power in developing countries.

```
## Parsed with column specification:
## cols(
##   country = col_character(),
##   year = col_double(),
##   Unemployment = col_double(),
##   GDP = col_number(),
##   GDP_capita = col_number()
## )
```

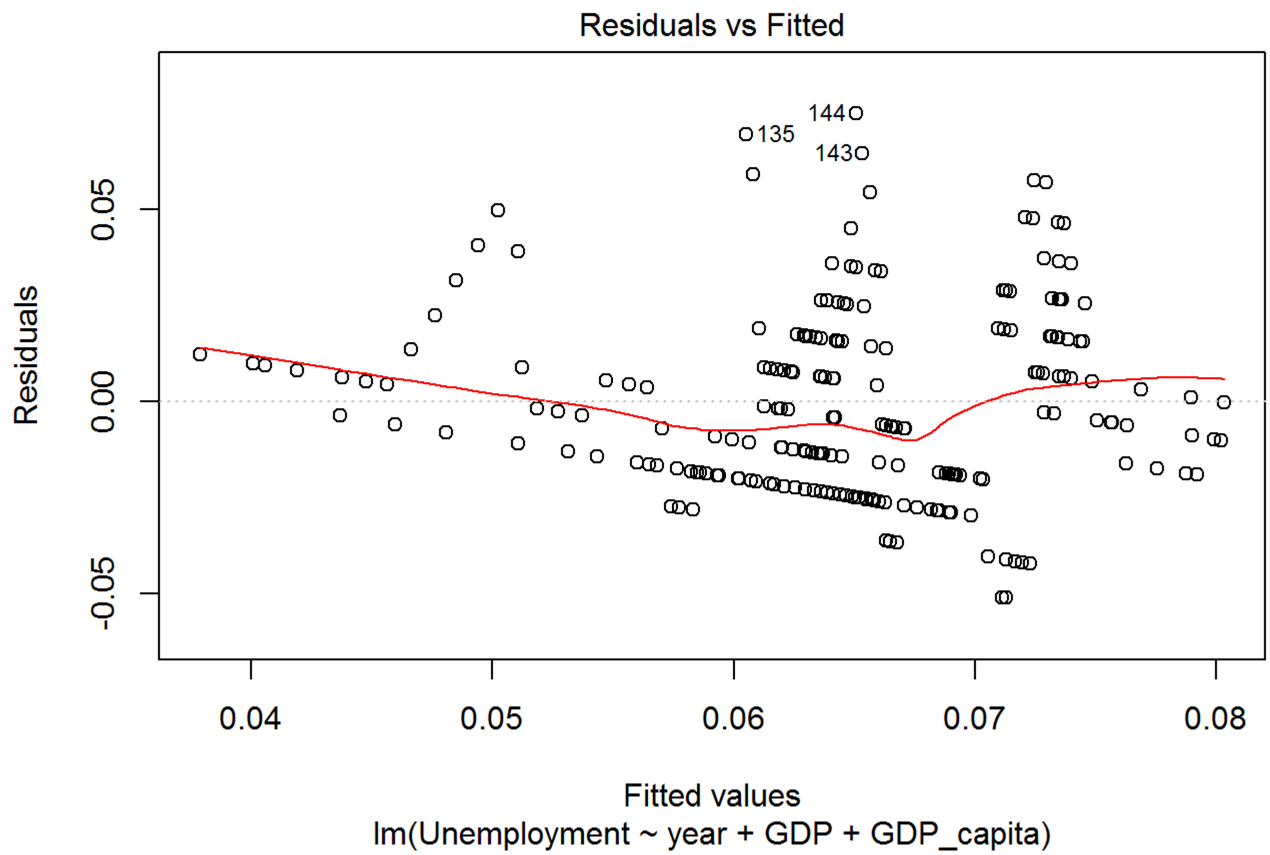


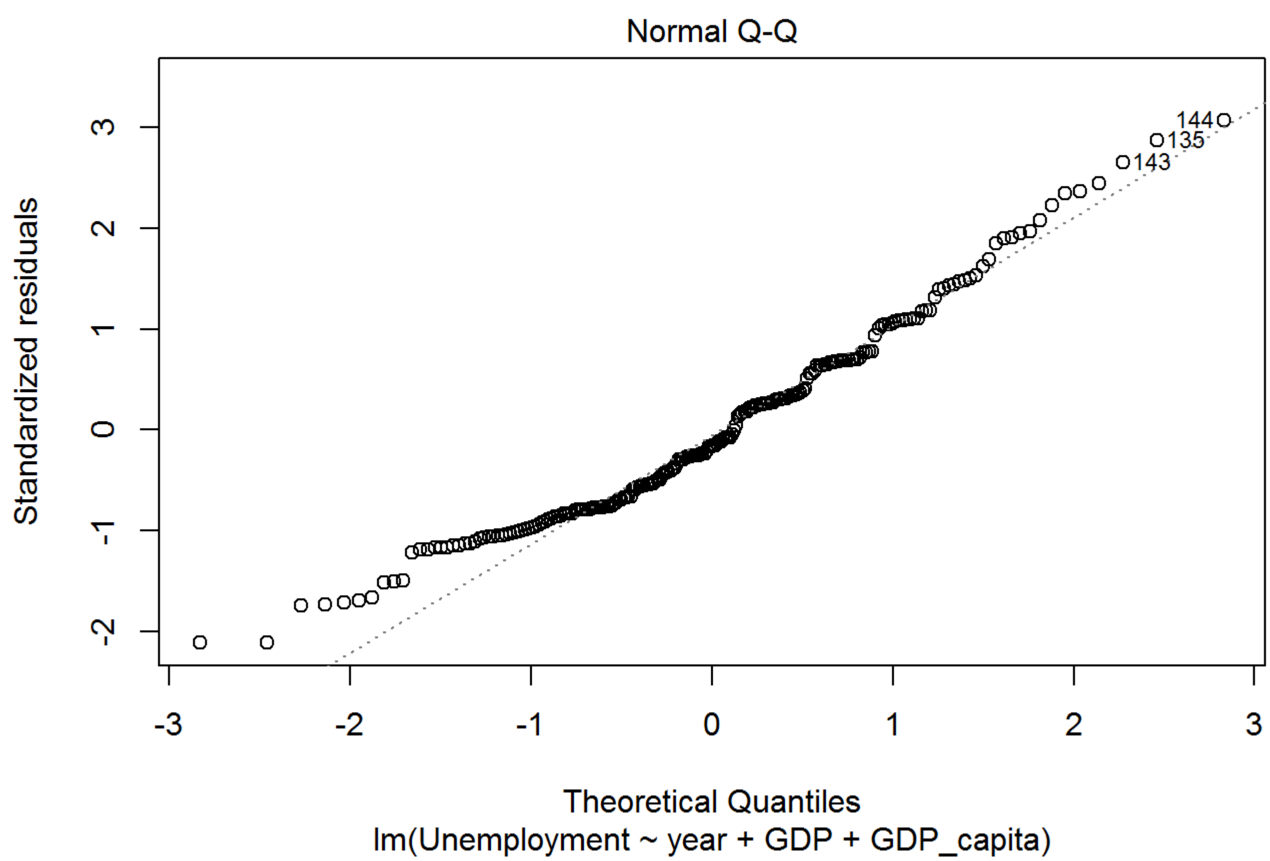
Unemployment
0.10
0.05
1990
2000
2010
year
country
Brazil
Canada
China
France
India
Japan
Russian Federation
United States

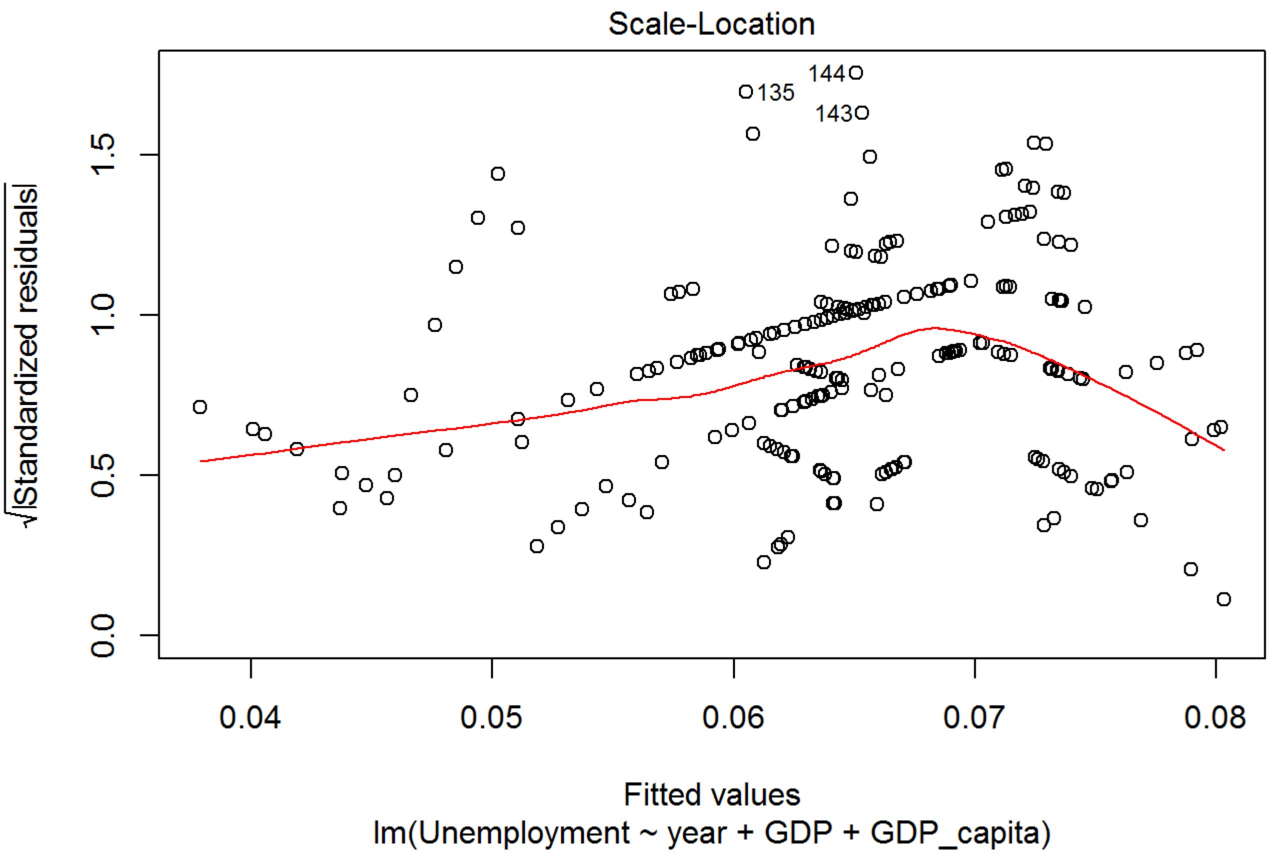


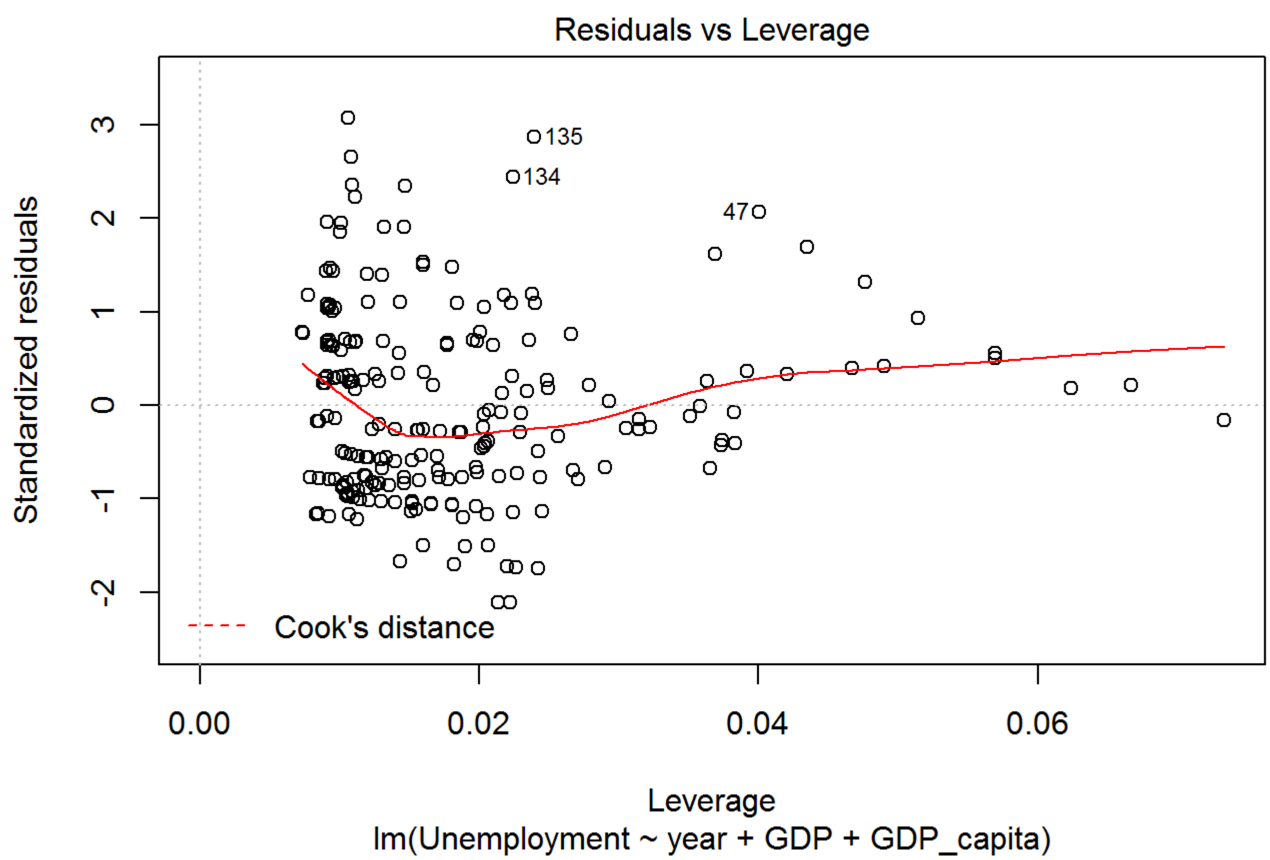
GDP_capita
60000
40000
20000
0
1990
2000
2010
year
country
Brazil
Canada
China
France
India
Japan
Russian Federation
United States

```
##
## Call:
## lm(formula = Unemployment ~ year + GDP + GDP_capita, data = world_gdp_new)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.051272 -0.019052 -0.003928  0.016430  0.074939
##
## Coefficients:
##                          Estimate             Std. Error t value
## (Intercept)  0.5167571232409954307  0.4562084115615206437   1.133
## year        -0.0002260286505821016  0.0002279703929513176  -0.991
## GDP         -0.0000000000000021805  0.0000000000000004961  -4.395
## GDP_capita   0.0000004222955354609  0.0000001150199379179   3.671
##              Pr(>|t|)
## (Intercept)  0.258611
## year         0.322580
## GDP         0.0000175 ***
## GDP_capita   0.000305 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0245 on 212 degrees of freedom
## Multiple R-squared:  0.1025, Adjusted R-squared:  0.08981
## F-statistic: 8.071 on 3 and 212 DF,  p-value: 0.00004074
```


Residuals vs Fitted
Residuals
Fitted values
lm(Unemployment ~ year + GDP + GDP_capita)
0.04
0.05
0.06
0.07
0.08
0.05
0.00
-0.05
144
135
143

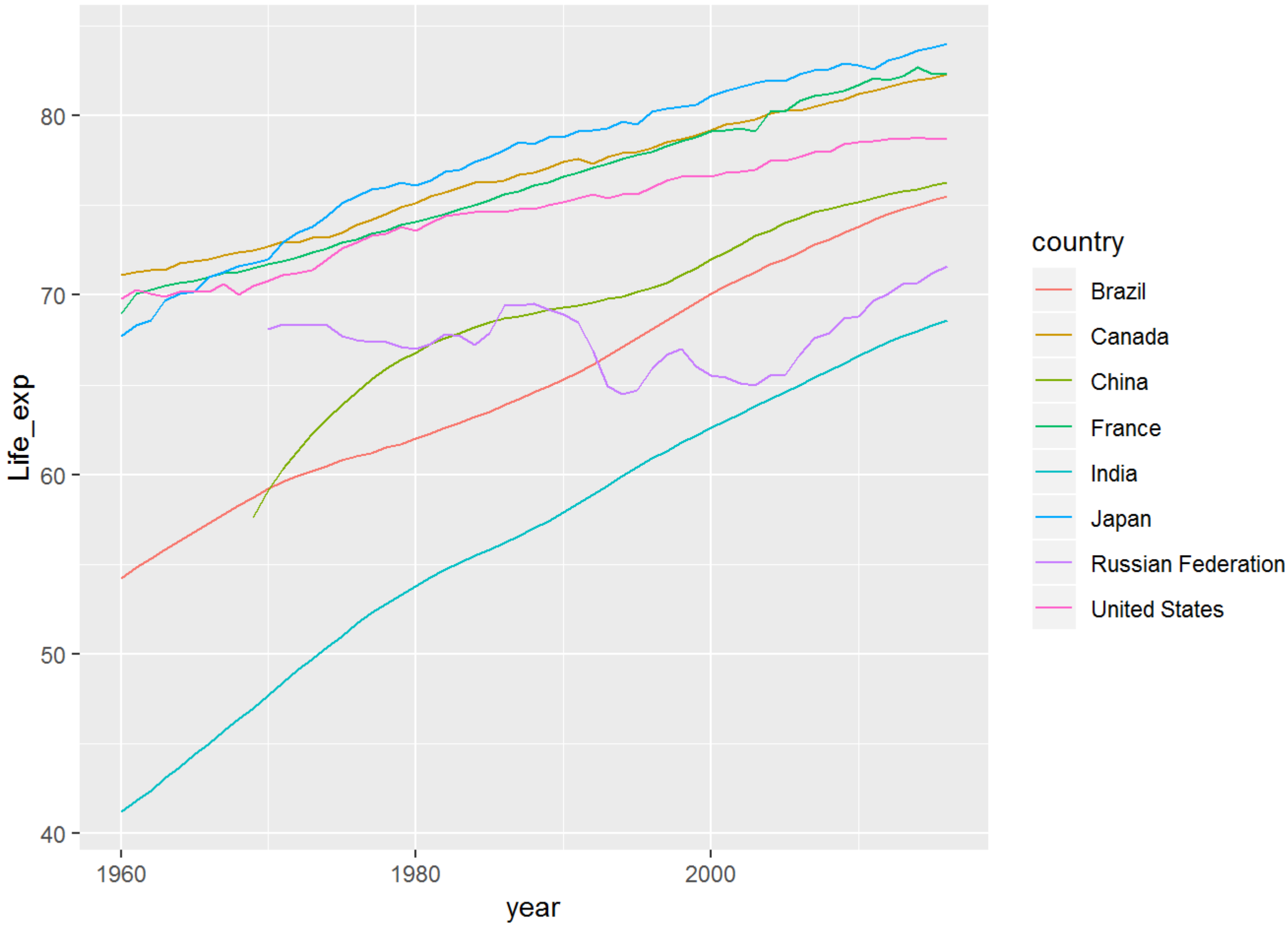

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Unemployment ~ year + GDP + GDP_capita)
-3
-2
-1
0
1
2
3
144
135
143

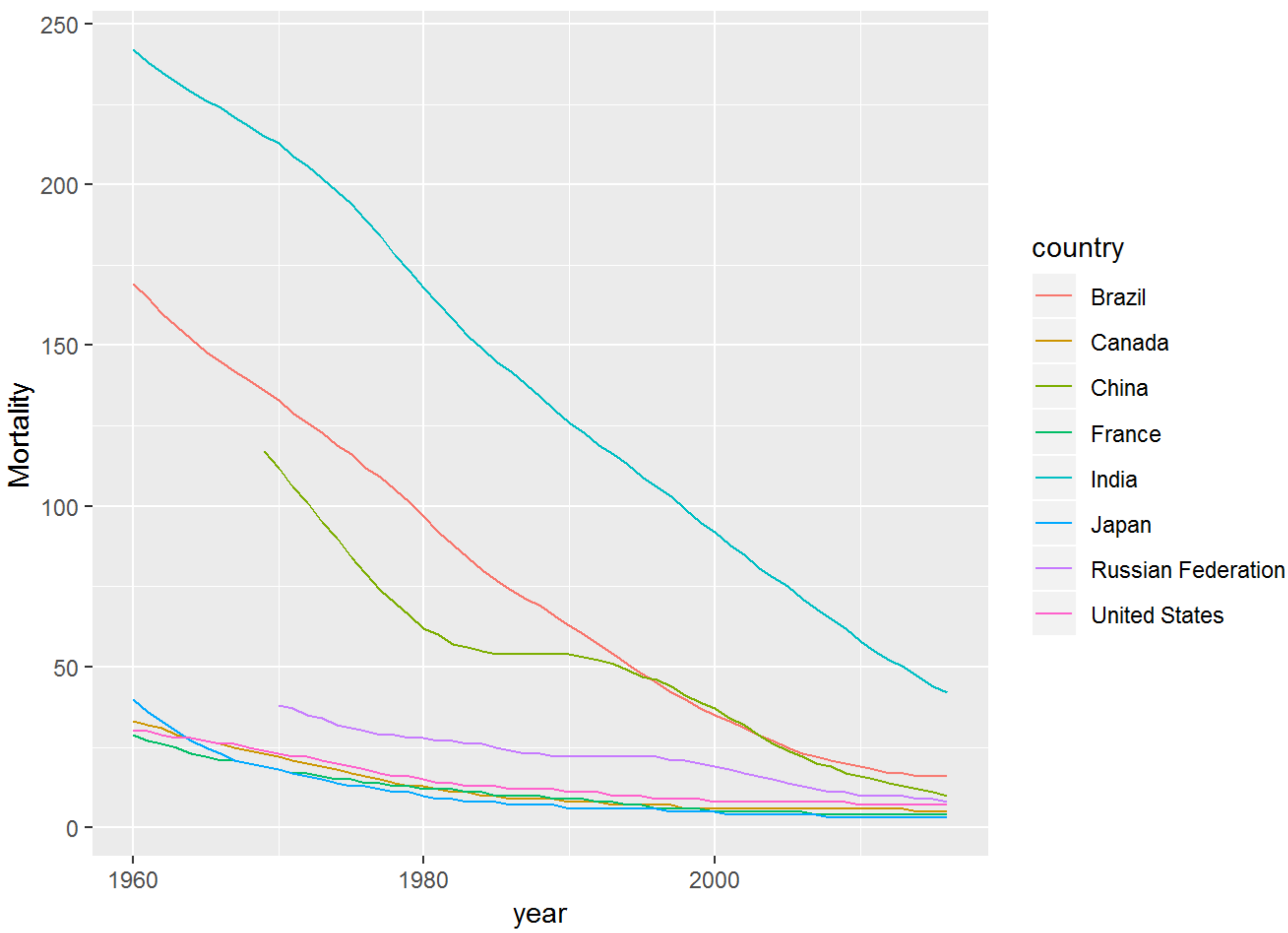

Scale-Location
√|Standardized residuals|
Fitted values
lm(Unemployment ~ year + GDP + GDP_capita)
135
144
143
0.0
0.5
1.0
1.5
0.04
0.05
0.06
0.07
0.08


Residuals vs Leverage
Standardized residuals
Leverage
lm(Unemployment ~ year + GDP + GDP_capita)
Cook's distance
135
134
47
-2
-1
0
1
2
3
0.00
0.02
0.04
0.06

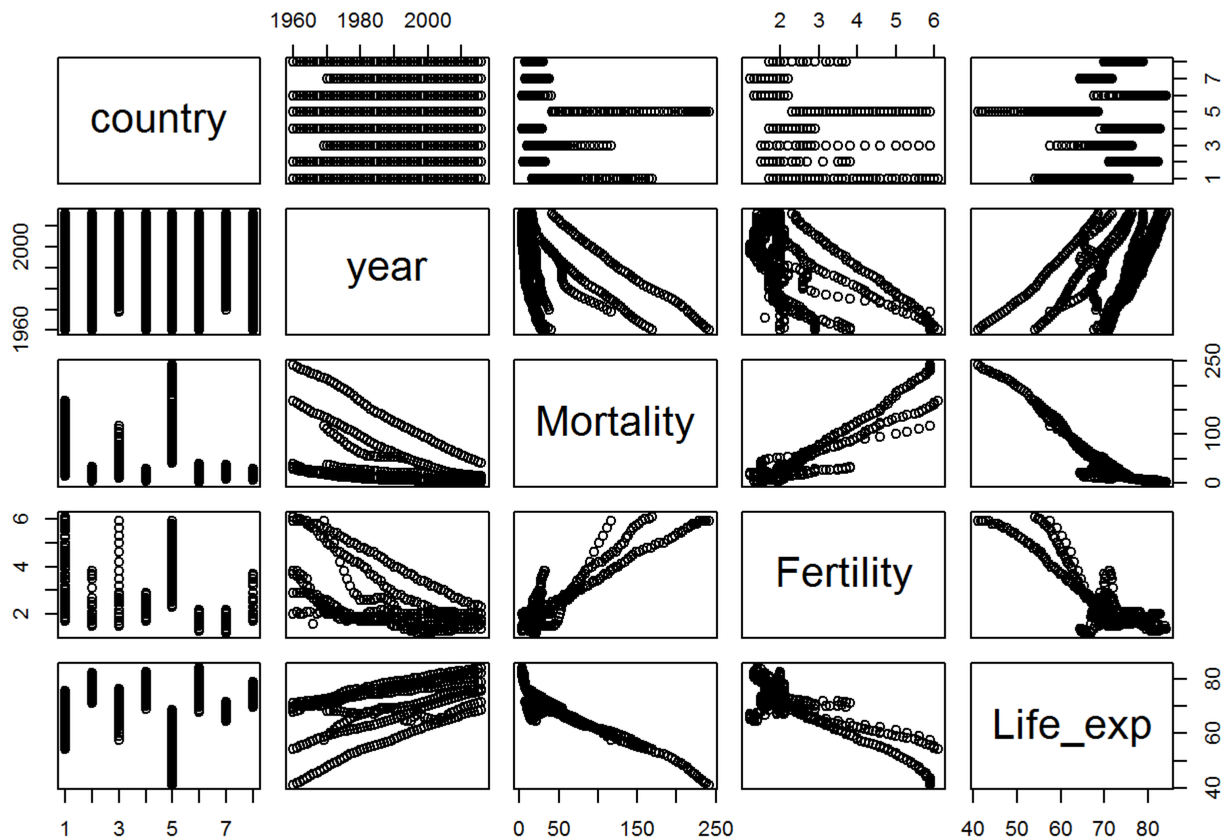
4.Data analysis on world population file

In this analysis, I analyzed world population file. World population file includes country, year, mortality, fertility, Life_exp, and population. Initial data cleaning was required. First, I cleaned rows containing 0 values using world_population_new <-filter(world_population, Mortality>0, Fertility>0) followed by round larger numbers down using option(scipen=999).I also believed that the size of population does not contribute to life expectancy. Therefore, I deleted population column using world_population_new1 <-select(world_population_new, -Population) %>% distinct. I plotted a linear relationship to find out whether the dataset I chose is significant. I compared two models which include (y=Life_exp, x=year) and (y=mortality, x=year). After plotting, I ran a multiple regression (y=Life_exp, x=all X variables) to check the significance of dataset. Data results showed that all three variables are significant (year, mortility, fertility). Our plot shows there is a linear relationship in the model. Throughout the years, the rates of life expectency increased in both developed countries and developing countries. Moreover, the rates of mortality decreased significantly in both developed and developing countries.


Life_exp
40
50
60
70
80
1960
1980
2000
year
country
Brazil
Canada
China
France
India
Japan
Russian Federation
United States


Mortality
0
50
100
150
200
250
1960
1980
2000
year
country
Brazil
Canada
China
France
India
Japan
Russian Federation
United States

```
##
## Call:
## lm(formula = Life_exp ~ year + Mortality + Fertility, data = world_population_new1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3926 -1.4884  0.2681  1.8346  6.0111
##
## Coefficients:
##                Estimate  Std. Error t value            Pr(>|t|)
## (Intercept) -112.682662  21.797139  -5.170          0.000000359 ***
## year           0.093622   0.010800   8.669 < 0.0000000000000002 ***
## Mortality     -0.171972   0.007327 -23.470 < 0.0000000000000002 ***
## Fertility      1.747014   0.350226   4.988          0.000000884 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.927 on 433 degrees of freedom
## Multiple R-squared:  0.881,  Adjusted R-squared:  0.8802
## F-statistic:  1068 on 3 and 433 DF,  p-value: < 0.00000000000000022
```



5. Merging all three tables

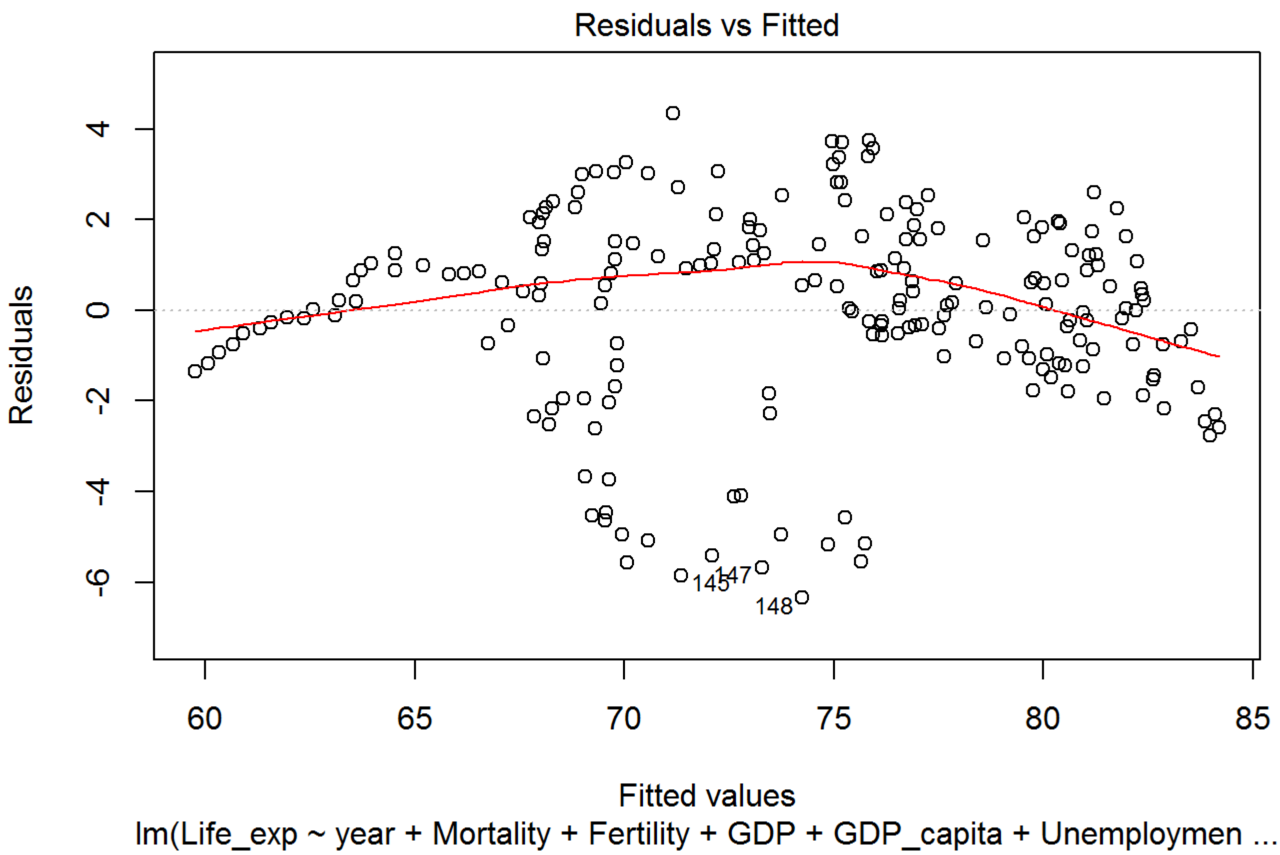
In this section, I merged all three tables together so that I can find which variables are significant on life expectancy. First, I deleted country and year from world_gdp file followed by left join with world_population. At the end, I also used left join to merge the final file world_trade. After successful merge, I deleted all rows containing 0 using world_final <- filter(world_final,Unemployment>0, Fertility>0, Mortality>0, Life_exp>0). Moreover, population does not contribute to our dinal model. I, therefore, deleted population column using world_final1 <- select(world_final, -Population)

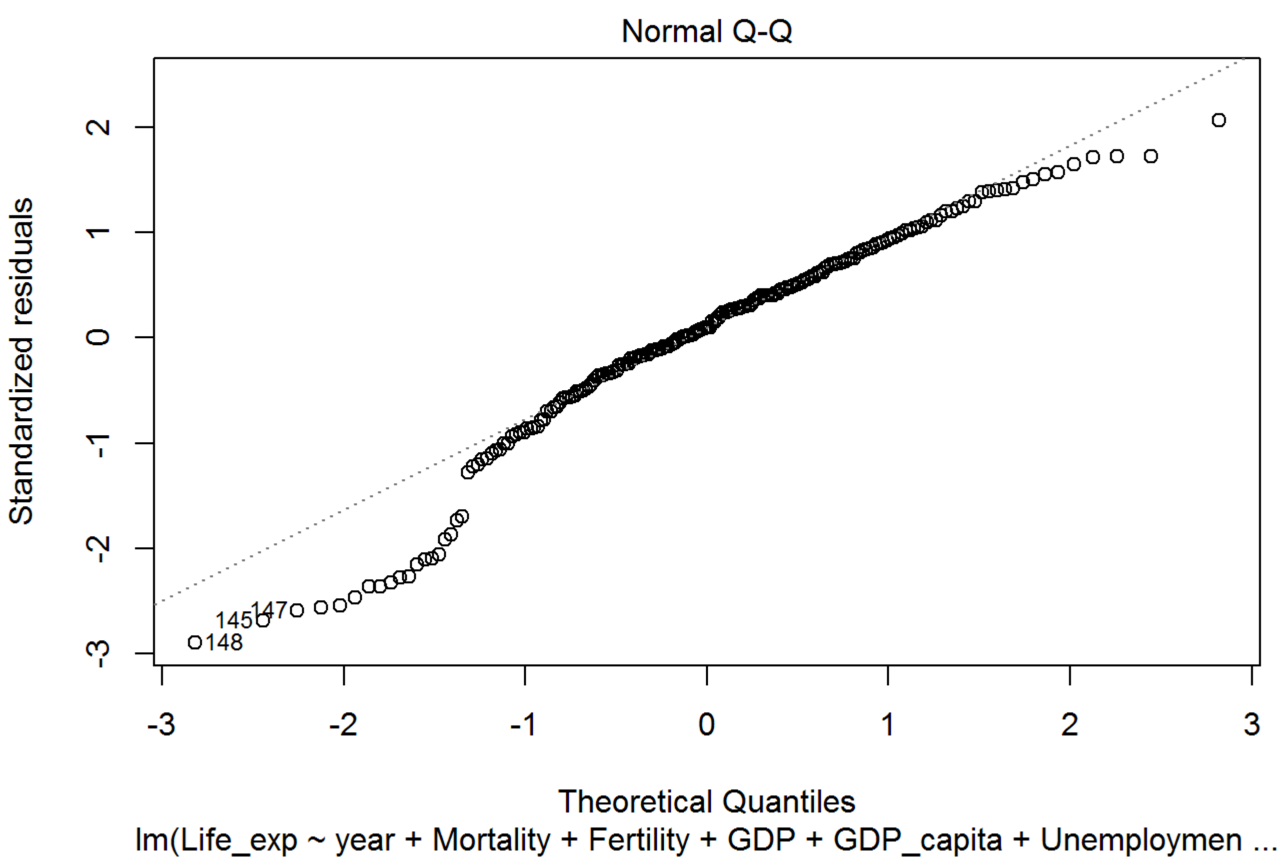
```
## Joining, by = c("country", "year")
## Joining, by = c("country", "year")
```

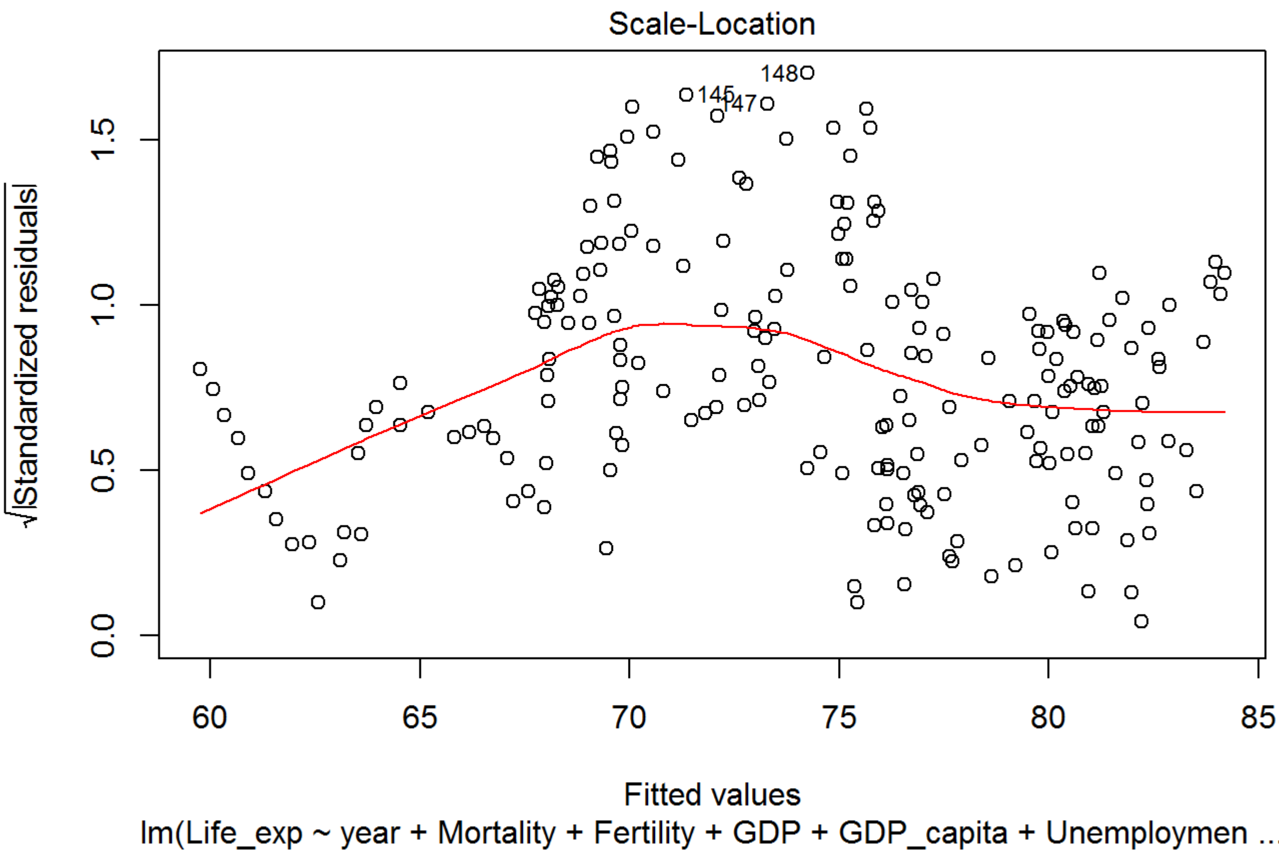
6. Run regression on merged file

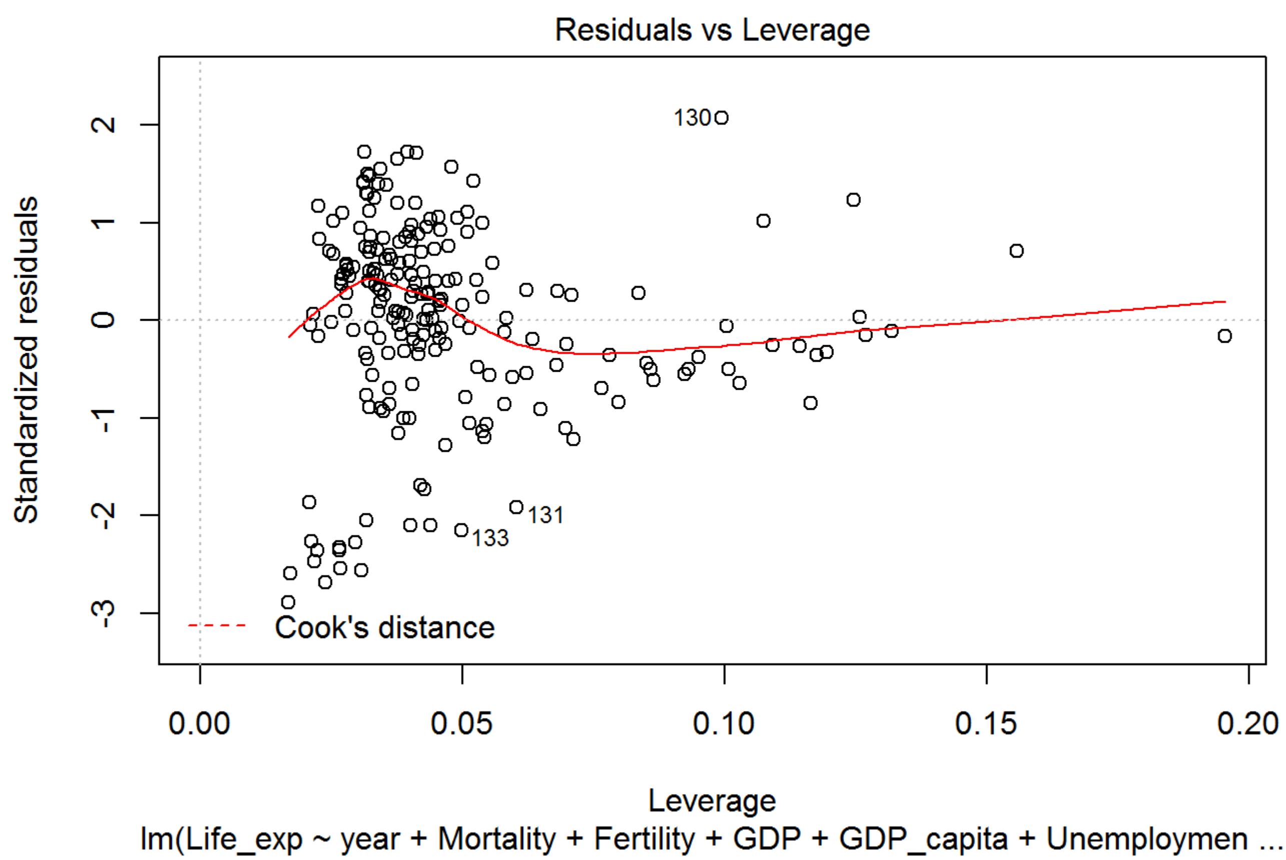
In this section, I ran multi regression using all variables from the merged file. Our final results showed that 7 predicting variables are significant (year, mortality, GDP, GDP_capita, unemployment, export and trade) with R squared value of 88%.

```
##
## Call:
## lm(formula = Life_exp ~ year + Mortality + Fertility + GDP +
##     GDP_capita + Unemployment + Imports + Exports + Trade, data = world_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3481 -1.0345  0.2104  1.4554  4.3436
##
## Coefficients:
##                          Estimate          Std. Error t value
## (Intercept)  190.5560858655312018  56.0717262777693151   3.398
## year          -0.0562934498222768   0.0278649824551405  -2.020
## Mortality     -0.1344590330423817   0.0191194056067424  -7.033
## Fertility      0.0988381101883995   0.6775933747615330   0.146
## GDP           -0.0000000000014903   0.0000000000002477  -6.017
## GDP_capita     0.0002404214464975   0.0000158487754040  15.170
## Unemployment -25.2711300875059024   8.2741169304071551  -3.054
## Imports        0.0000000000034447   0.0000000000022495   1.531
## Exports        0.0000000000046678   0.0000000000014520   3.215
## Trade         -8.3697531883892680   1.5563258272946050  -5.378
##                          Pr(>|t|)
## (Intercept)              0.000819 ***
## year                     0.044708 *
## Mortality        0.0000000000322 ***
## Fertility                0.884175
## GDP              0.0000000084382 ***
## GDP_capita   < 0.0000000000000002 ***
## Unemployment             0.002567 **
## Imports                  0.127289
## Exports                  0.001524 **
## Trade            0.0000002112166 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.211 on 198 degrees of freedom
## Multiple R-squared:  0.8874, Adjusted R-squared:  0.8822
## F-statistic: 173.3 on 9 and 198 DF,  p-value: < 0.00000000000000022
```


Residuals vs Fitted
Residuals
Fitted values
lm(Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita + Unemploymen ...
145
147
148
4
2
0
-2
-4
-6
60
65
70
75
80
85


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita + Unemploymen ...
145
147
148
2
1
0
-1
-2
-3
3


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
60
65
70
75
80
85
148
145
147
Fitted values
lm(Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita + Unemploymen ...


Residuals vs Leverage
Standardized residuals
-3
-2
-1
0
1
2
0.00
0.05
0.10
0.15
0.20
130
131
133
Cook's distance
Leverage
lm(Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita + Unemploymen ...

7. Remove outliers

However, there are a few outliers (row: 145, 147, 148) that need to be deleted before further analysis. I, therefore, deleted all outliers using world_final1 <- world_final1[-145,, drop=FALSE], world_final1 <- world_final1[-147,, drop=FALSE], world_final1 <- world_final1[-148,, drop=FALSE].

```
##                country year Mortality Fertility Life_exp Unemployment
## 145 Russian Federation 2005        14       1.3     65.5         0.07
##              GDP GDP_capita      Imports      Exports Trade
## 145 764017107992       5323 164341474452 268957446508  0.57
```

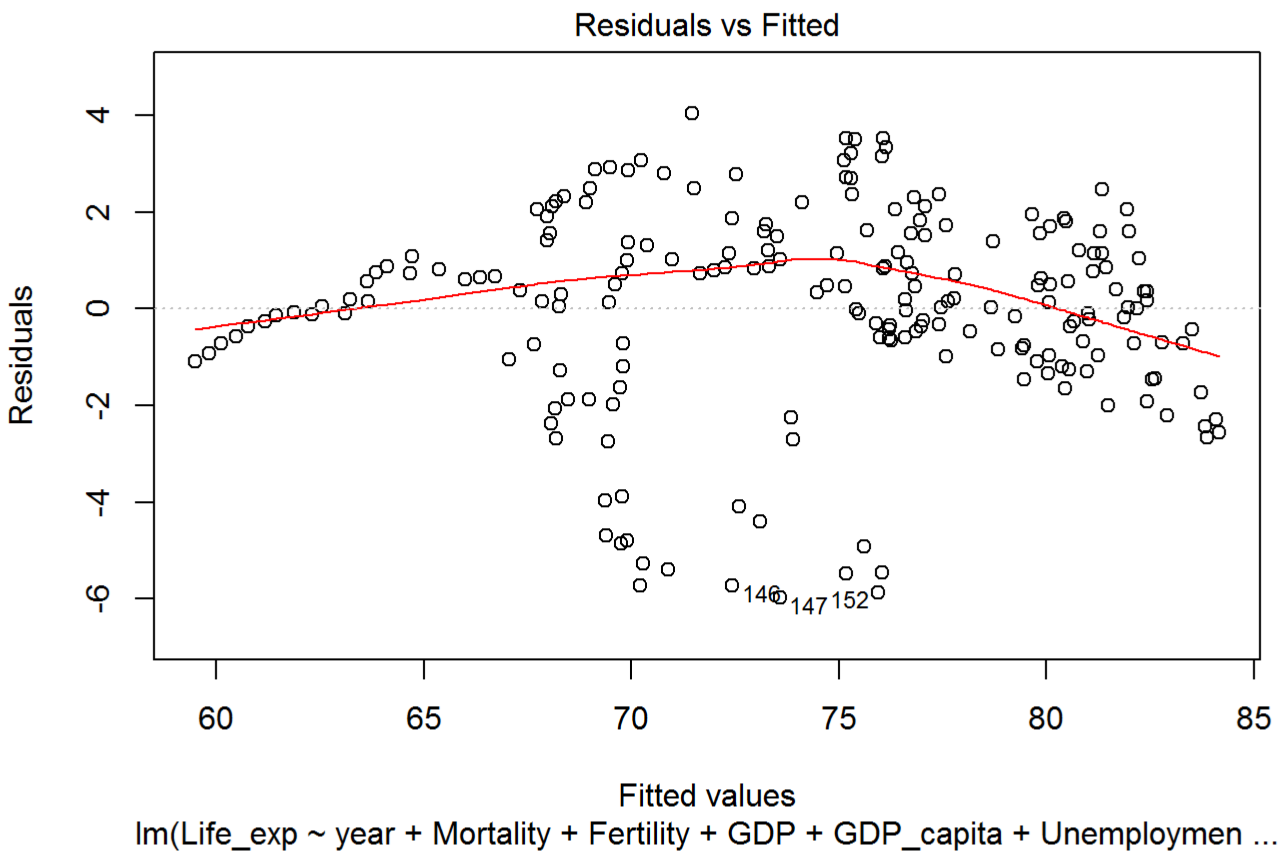
```
##                country year Mortality Fertility Life_exp Unemployment
## 147 Russian Federation 2007        12       1.4     67.6         0.06
##               GDP GDP_capita      Imports      Exports Trade
## 147 1299705247686       9101 279983425069 392044033025  0.52
```

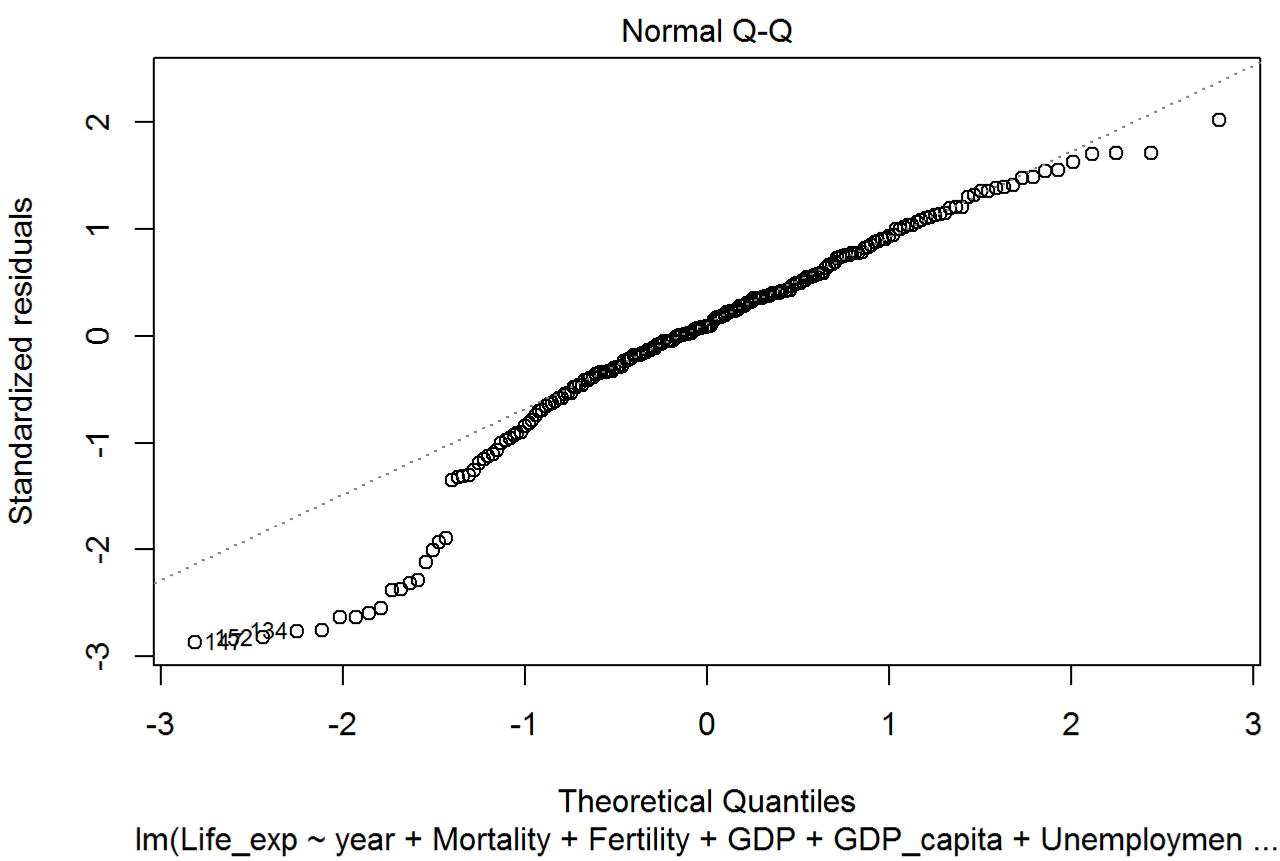
```
##                country year Mortality Fertility Life_exp Unemployment
## 148 Russian Federation 2008        11       1.5     67.9         0.06
##               GDP GDP_capita      Imports      Exports Trade
## 148 1660844408500      11635 366597057084 520003701781  0.53
```

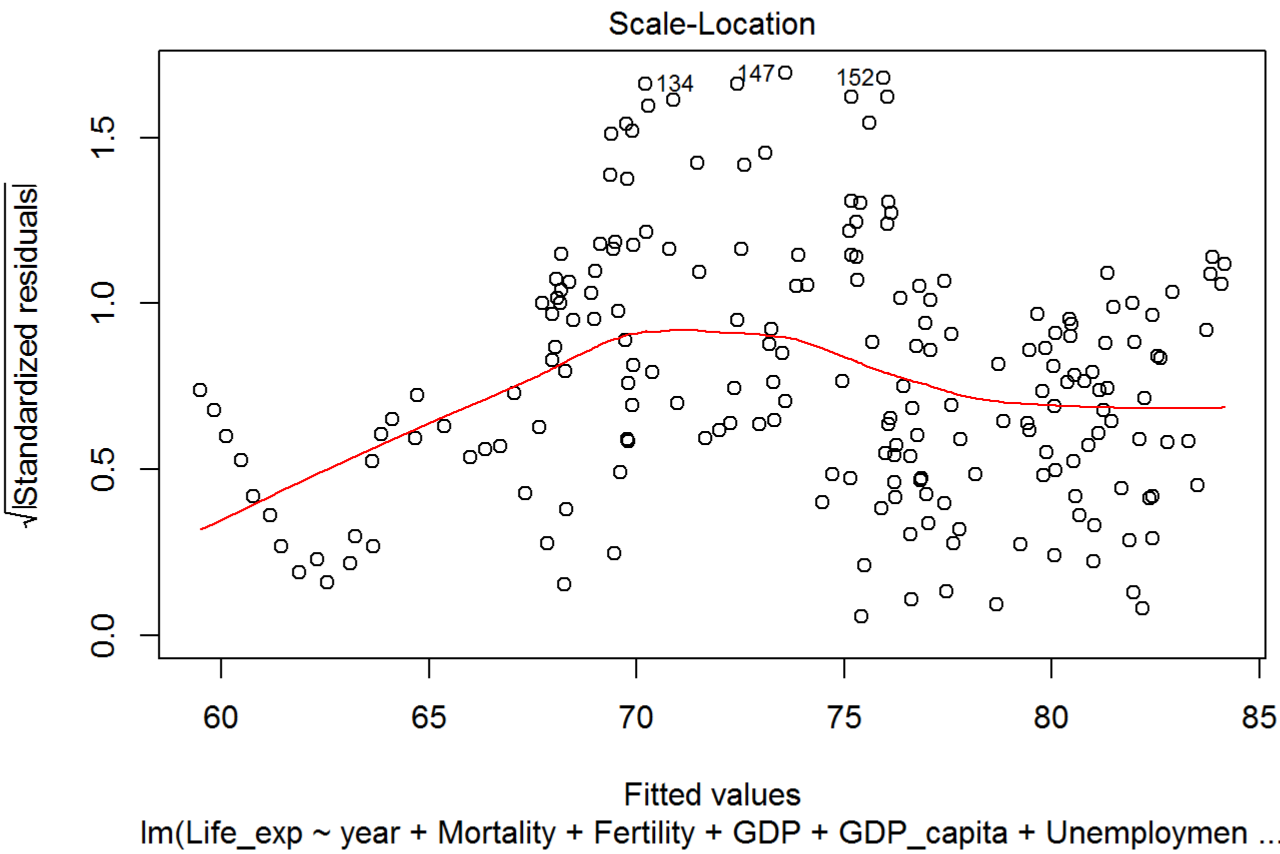
8. Run regression after removing outliers

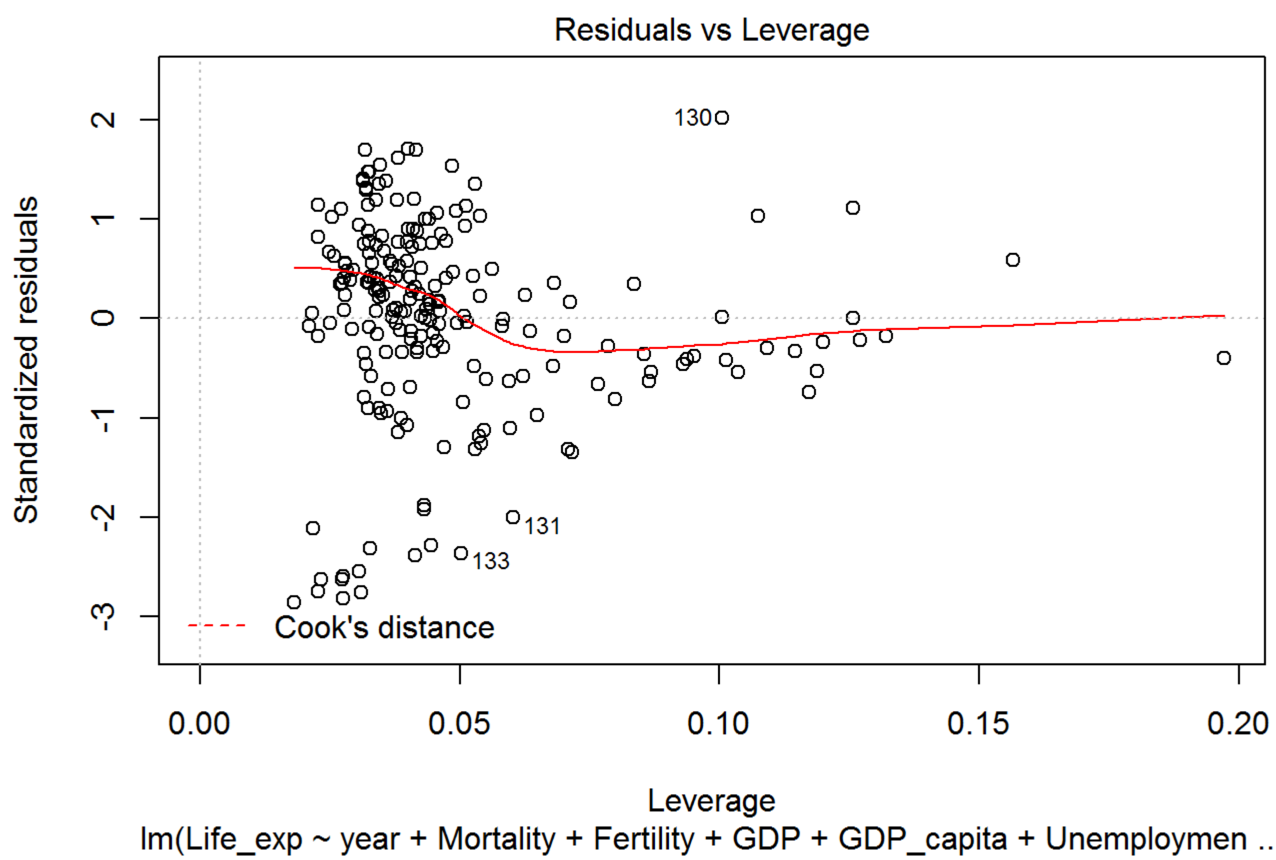
We ran multiple regression after removing initial outliers to see whether the model fits better. The results showed that the r squared value was slightly improved by 1%.

```
##
## Call:
## lm(formula = Life_exp ~ year + Mortality + Fertility + GDP +
##     GDP_capita + Unemployment + Imports + Exports + Trade, data = world_final1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9798 -0.8347  0.1843  1.3763  4.0373
##
## Coefficients:
##                         Estimate          Std. Error t value
## (Intercept)  164.1000829194836399  53.6777054428390556   3.057
## year          -0.0430606164613111   0.0266767678324724  -1.614
## Mortality     -0.1373821681891240   0.0182299877591978  -7.536
## Fertility      0.1233863368643090   0.6461745724629087   0.191
## GDP           -0.0000000000013921   0.0000000000002366  -5.883
## GDP_capita     0.0002334251974594   0.0000151564488882  15.401
## Unemployment -25.6323259421471974   7.8765483346237541  -3.254
## Imports        0.0000000000026430   0.0000000000021475   1.231
## Exports        0.0000000000048618   0.0000000000013830   3.515
## Trade         -7.7789035929090673   1.4863711981100036  -5.233
##                         Pr(>|t|)
## (Intercept)             0.002548 **
## year                    0.108109
## Mortality       0.00000000000178 ***
## Fertility               0.848764
## GDP             0.00000001724119 ***
## GDP_capita   < 0.0000000000000002 ***
## Unemployment            0.001340 **
## Imports                 0.219911
## Exports                 0.000546 ***
## Trade           0.00000042795218 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.104 on 195 degrees of freedom
## Multiple R-squared:  0.8978, Adjusted R-squared:  0.893
## F-statistic: 190.3 on 9 and 195 DF,  p-value: < 0.00000000000000022
```


Residuals vs Fitted
Residuals
4
2
0
-2
-4
-6
146
147
152
60
65
70
75
80
85
Fitted values
lm(Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita + Unemploymen ...


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-3
147
152
134
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita + Unemploymen ...


Scale-Location
134
147
152
√|Standardized residuals|
0.0
0.5
1.0
1.5
60
65
70
75
80
85
Fitted values
lm(Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita + Unemploymen ...


Residuals vs Leverage
130
131
133
Cook's distance
Standardized residuals
-3
-2
-1
0
1
2
0.00
0.05
0.10
0.15
0.20
Leverage
lm(Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita + Unemploymen ...

9. Best subset regression

In this section, I conducted best subset regression. Best Subsets compares all possible models using a specified set of predictors, and displays the best-fitting models that contain one predictor, two predictors, and so on. In best regression model, BIC tells us the best predicting model. Therfore, I used the following function to draw BIC value. The results showed that #6 is the best regression model which includes mortality, GDP, GDP_capita, unemployment, explorts, and trade.

res.sum <- summary(best) data.frame(Adj.R2 = which.max(res.sum$adjr2), CP = which.min(res.sum$cp), BIC = which.min(res.sum$bic))

```
##    country              year        Mortality        Fertility    
##  Length:205         Min.   :1991   Min.   :  3.00   Min.   :1.200  
##  Class :character   1st Qu.:1997   1st Qu.:  6.00   1st Qu.:1.500  
##  Mode  :character   Median :2003   Median : 10.00   Median :1.800  
##                     Mean   :2003   Mean   : 22.86   Mean   :1.903  
##                     3rd Qu.:2010   3rd Qu.: 29.00   3rd Qu.:2.000  
##                     Max.   :2016   Max.   :123.00   Max.   :4.000  
##     Life_exp      Unemployment          GDP            
##  Min.   :58.40   Min.   :0.02000   Min.   :  195905767669  
##  1st Qu.:69.40   1st Qu.:0.04000   1st Qu.:  734547898221  
##  Median :75.80   Median :0.06000   Median : 1660287965660  
##  Mean   :74.35   Mean   :0.06498   Mean   : 3379975208960  
##  3rd Qu.:79.50   3rd Qu.:0.08000   3rd Qu.: 4515264514430  
##  Max.   :84.00   Max.   :0.14000   Max.   :18624475000000  
##    GDP_capita       Imports                 Exports           
##  Min.   :  298   Min.   :  22887476747   Min.   :  22875165149  
##  1st Qu.: 2695   1st Qu.: 151757004451   1st Qu.: 168142004496  
##  Median :20017   Median : 351430953969   Median : 391450612675  
##  Mean   :20220   Mean   : 582045744234   Mean   : 556174712331  
##  3rd Qu.:36450   3rd Qu.: 719974000000   3rd Qu.: 720939000000  
##  Max.   :57589   Max.   :2883157000000   Max.   :2462839435100  
##      Trade       
##  Min.   :0.1600  
##  1st Qu.:0.2500  
##  Median :0.3800  
##  Mean   :0.4067  
##  3rd Qu.:0.5500  
##  Max.   :1.1100  
```

```
## Subset selection object
## Call: regsubsets.formula(Life_exp ~ year + Mortality + Fertility +
##     GDP + GDP_capita + Unemployment + Imports + Exports + Trade,
##     data = world_final1)
## 9 Variables  (and intercept)
##              Forced in Forced out
## year             FALSE      FALSE
## Mortality        FALSE      FALSE
## Fertility        FALSE      FALSE
## GDP              FALSE      FALSE
## GDP_capita       FALSE      FALSE
## Unemployment     FALSE      FALSE
## Imports          FALSE      FALSE
## Exports          FALSE      FALSE
## Trade            FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          year Mortality Fertility GDP GDP_capita Unemployment Imports
## 1  ( 1 ) " "  " "       " "       " " "*"        " "          " "
## 2  ( 1 ) " "  "*"       " "       " " "*"        " "          " "
## 3  ( 1 ) " "  "*"       " "       "*" "*"        " "          " "
## 4  ( 1 ) " "  "*"       " "       " " "*"        " "          "*"
## 5  ( 1 ) " "  "*"       " "       "*" "*"        " "          " "
## 6  ( 1 ) " "  "*"       " "       "*" "*"        "*"          " "
## 7  ( 1 ) "*"  "*"       " "       "*" "*"        "*"          " "
## 8  ( 1 ) "*"  "*"       " "       "*" "*"        "*"          "*"
##          Exports Trade
## 1  ( 1 ) " "     " "
## 2  ( 1 ) " "     " "
## 3  ( 1 ) " "     " "
## 4  ( 1 ) "*"     " "
## 5  ( 1 ) "*"     "*"
## 6  ( 1 ) "*"     "*"
## 7  ( 1 ) "*"     "*"
## 8  ( 1 ) "*"     "*"
```

```
##   Adj.R2 CP BIC
## 1      8  6   6
```

10. Best subset regression analysis

I included all the variables based on BIC and ran regression to check whether all variables from BIC are significant. The results showed that all 6 variables are significant with r-squared value of 0.8961. This tells us that best subset regression model predicts the best.

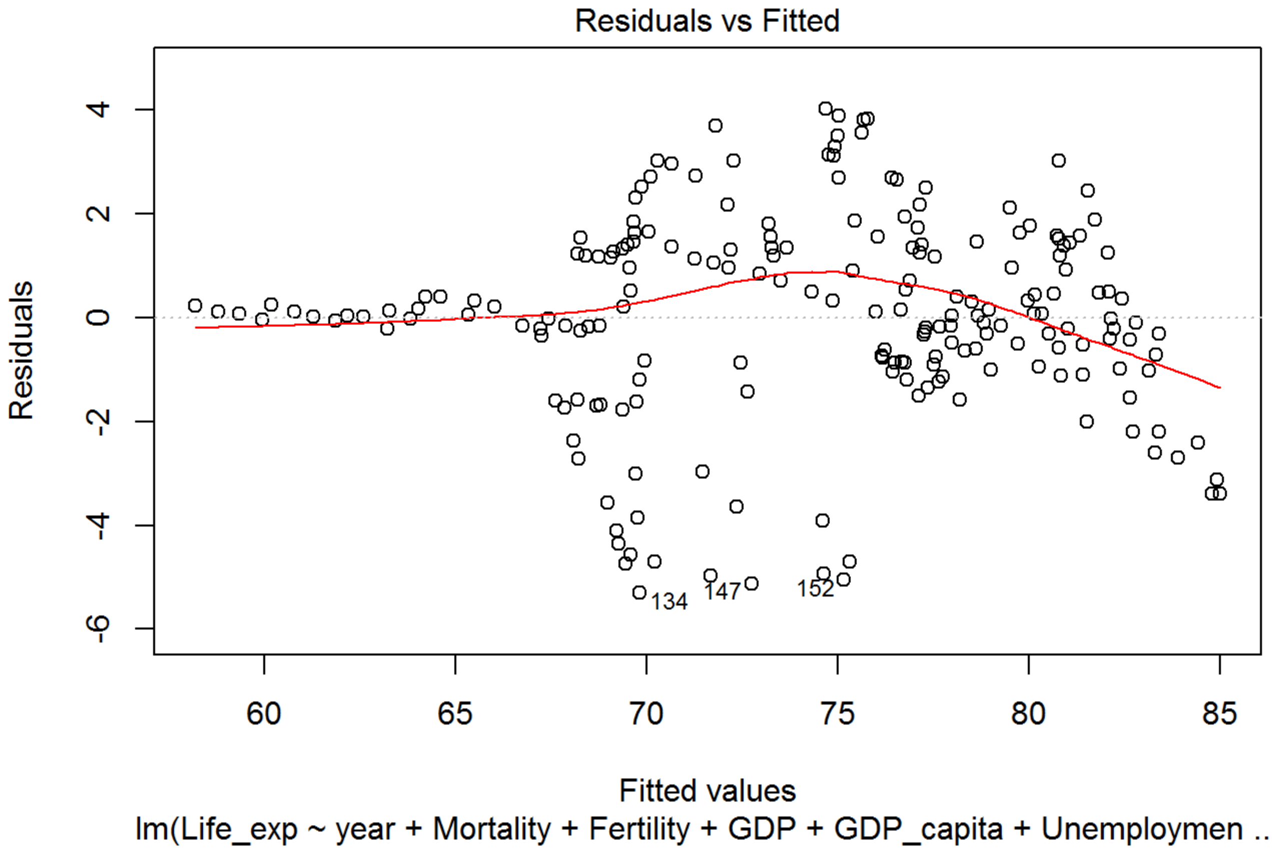
```
##
## Call:
## lm(formula = Life_exp ~ Mortality + GDP + GDP_capita + Unemployment +
##     Exports + Trade, data = world_final1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1471 -0.7207  0.2438  1.2675  3.6566
##
## Coefficients:
##                          Estimate          Std. Error t value
## (Intercept)   77.0048866684559812   0.8287658115865543  92.915
## Mortality     -0.1262318703637217   0.0084126505594981 -15.005
## GDP           -0.0000000000011012   0.0000000000001165  -9.455
## GDP_capita     0.0002369463845649   0.0000130141161464  18.207
## Unemployment -20.0243579229795294   6.7562695933817745  -2.964
## Exports        0.0000000000058363   0.0000000000007460   7.824
## Trade         -6.8363684668103151   1.2006568893771445  -5.694
##                          Pr(>|t|)
## (Intercept)  < 0.0000000000000002 ***
## Mortality    < 0.0000000000000002 ***
## GDP          < 0.0000000000000002 ***
## GDP_capita   < 0.0000000000000002 ***
## Unemployment              0.00341 **
## Exports        0.000000000000301 ***
## Trade          0.000000044349033 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.105 on 198 degrees of freedom
## Multiple R-squared:  0.8961, Adjusted R-squared:  0.8929
## F-statistic: 284.6 on 6 and 198 DF,  p-value: < 0.00000000000000022
```

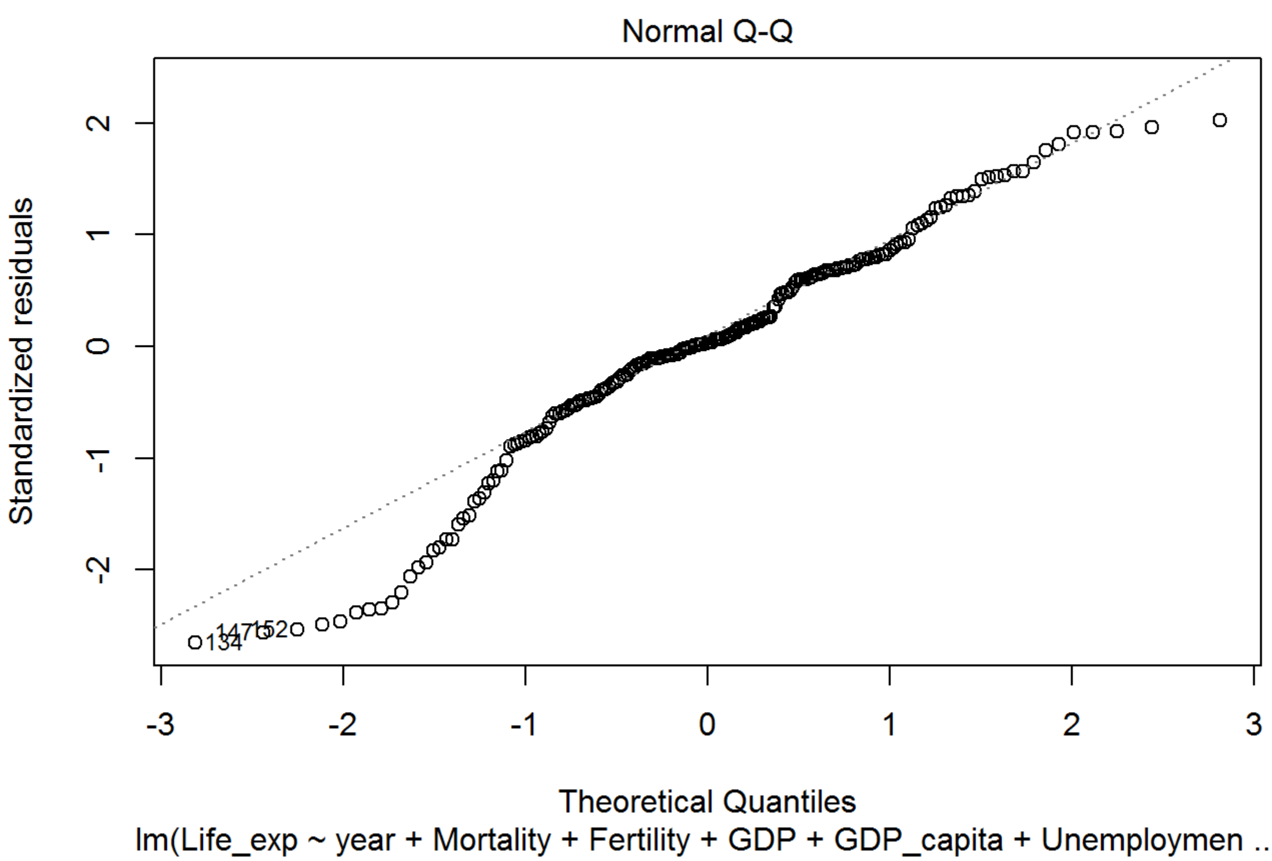
11. Interaction Effects

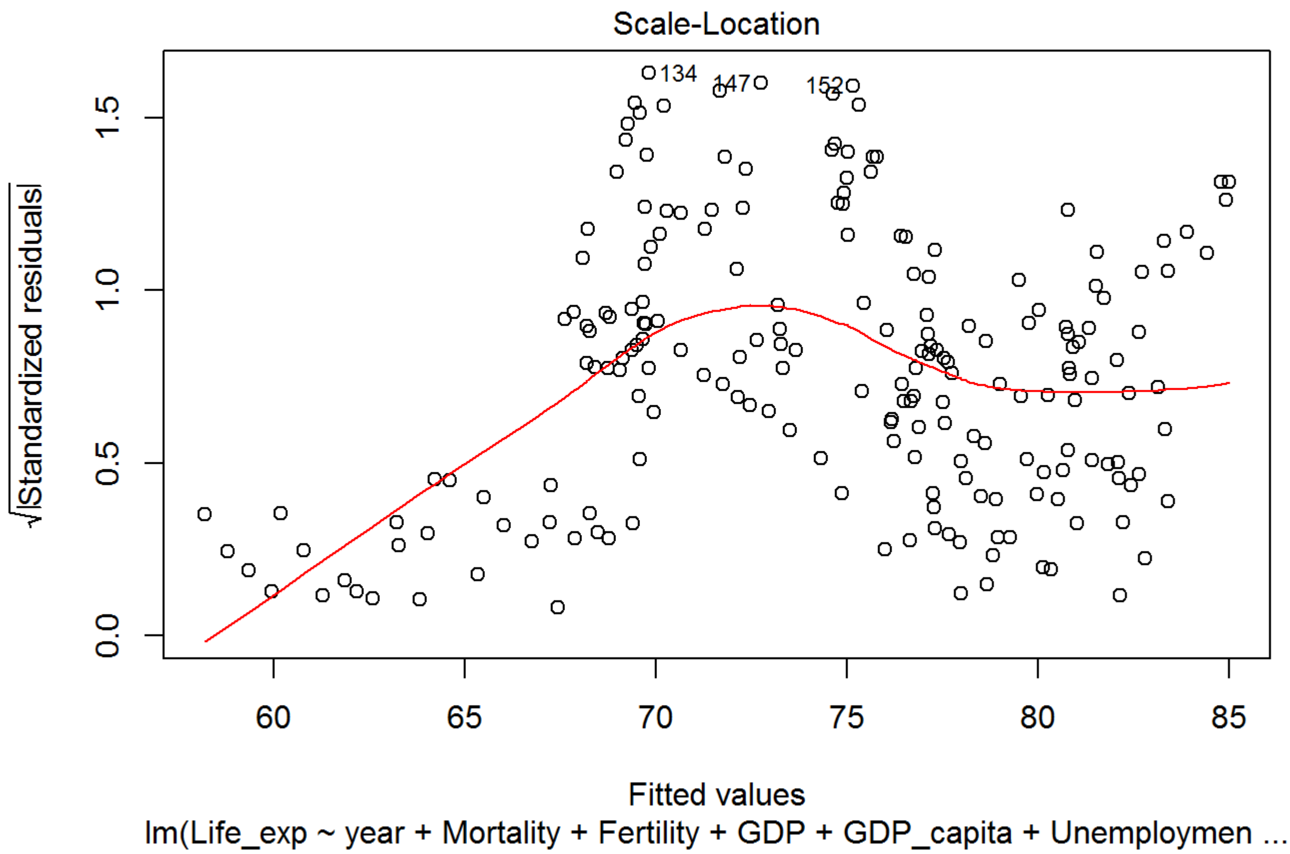
In this analysis, I analyzed interaction effects.In more complex study areas, the independent variables might interact with each other. Interaction effects indicate that a third variable influences the relationship between an independent and dependent variable. This type of effect makes the model more complex, but if the real world behaves this way, it is critical to incorporate it in your model. I, therefore, included the following variables to the original model. 1. (Mortality:Fertility) 2. (Imports:Exports) 3. (GDP: GDP_capita) The results showed that (mortality:fertility) and (GDP:GDP-capita) are significant factors with r squared value 0.9058.

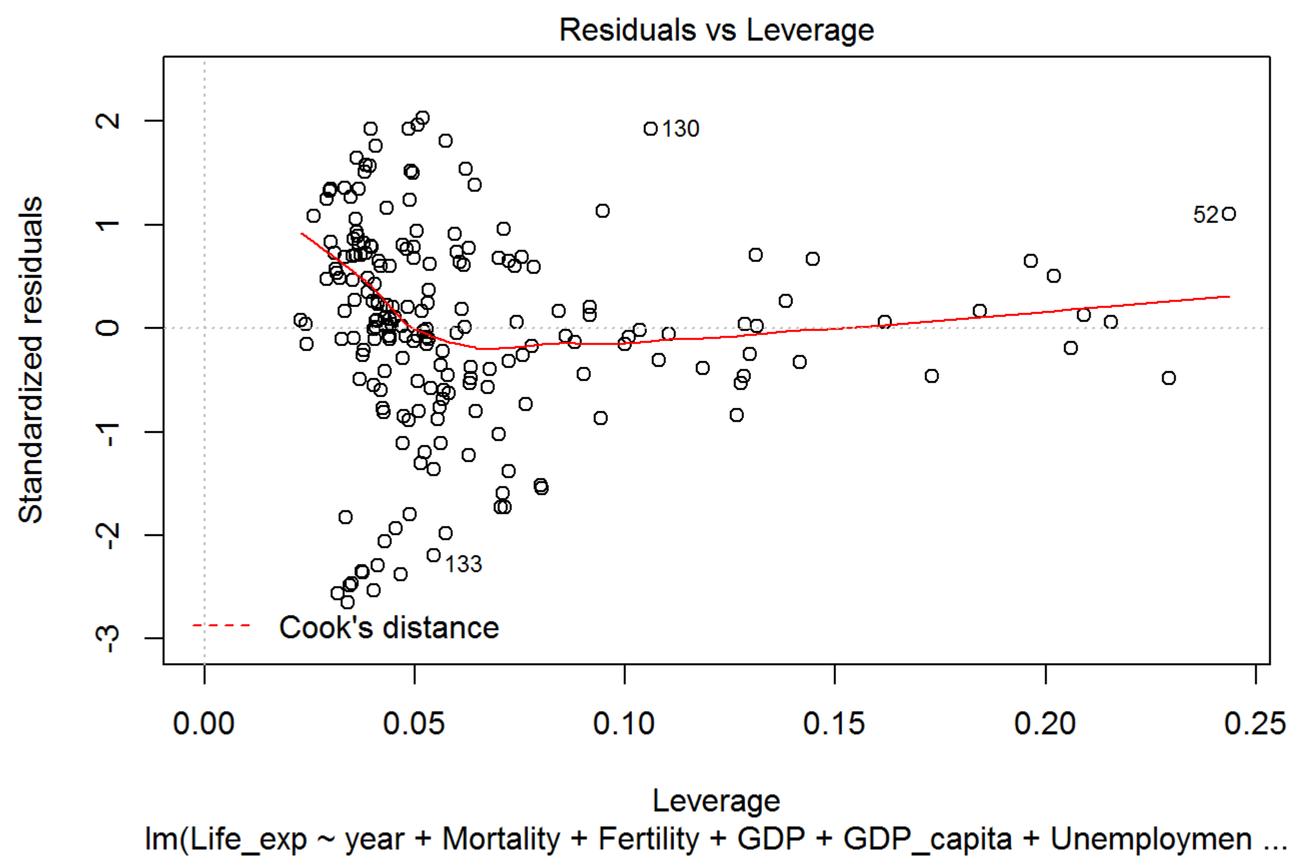
```
##
## Call:
## lm(formula = Life_exp ~ year + Mortality + Fertility + GDP +
##     GDP_capita + Unemployment + Imports + Exports + Trade + (Mortality:Fertility) +
##     (Imports:Exports) + (GDP:GDP_capita), data = world_final1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3062 -0.9031  0.0677  1.3378  4.0214
##
## Coefficients:
##                                              Estimate
## (Intercept)         146.8725661804992341785691678524
## year                 -0.0365046781543425027938276628
## Mortality            -0.0130972877014524976407860990
## Fertility            -0.1589333462850445188863091062
## GDP                  -0.0000000000007691814269256995
## GDP_capita            0.0002939737291619568329559264
## Unemployment        -10.1526682609102714138771261787
## Imports               0.0000000000029143770346812612
## Exports               0.0000000000041933674949460033
## Trade                -6.6183743318951844258890560013
## Mortality:Fertility  -0.0249602193569374367076996180
## Imports:Exports      -0.0000000000000000000000004474
## GDP:GDP_capita       -0.00000000000000000099677020242
##                                            Std. Error t value
## (Intercept)          53.5438525894610322097832977306   2.743
## year                  0.0266005126244921660805253794  -1.372
## Mortality             0.0432149607933211346577628831  -0.303
## Fertility             0.7767547332983560925967481126  -0.205
## GDP                   0.000000000003124670767226541  -2.462
## GDP_capita            0.0000247386571932429841398637  11.883
## Unemployment          8.9421049735574538175342240720  -1.135
## Imports               0.0000000000024997670800029518   1.166
## Exports               0.0000000000022243478888512911   1.885
## Trade                 1.5265572535844182944231306465  -4.335
## Mortality:Fertility   0.0108960955423224627180989188  -2.291
## Imports:Exports       0.0000000000000000000000005694  -0.786
## GDP:GDP_capita        0.00000000000000000041455866371  -2.404
##                               Pr(>|t|)
## (Intercept)                    0.00666 **
## year                           0.17156
## Mortality                      0.76216
## Fertility                      0.83809
## GDP                            0.01471 *
## GDP_capita          < 0.0000000000000002 ***
## Unemployment                   0.25763
## Imports                        0.24512
## Exports                        0.06091 .
## Trade                        0.0000235 ***
## Mortality:Fertility            0.02306 *
## Imports:Exports                0.43295
## GDP:GDP_capita                 0.01715 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.035 on 192 degrees of freedom
## Multiple R-squared:  0.9058, Adjusted R-squared:    0.9
## F-statistic: 153.9 on 12 and 192 DF,  p-value: < 0.00000000000000022
```


Residuals vs Fitted
Residuals
Fitted values
lm(Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita + Unemploymen ...
134
147
152


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita + Unemploymen ...
134
147
152


Scale-Location
√|Standardized residuals|
Fitted values
lm(Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita + Unemploymen ...
134
147
152


Residuals vs Leverage
Standardized residuals
Leverage
lm(Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita + Unemploymen ...
130
52
133
Cook's distance

12. Which model is better? (ANOVA test: best subset vs interaction)

A good model not only needs to fit data well, it also needs to be parsimonious. That is, a good model should be only be as complex as necessary to describe a dataset. If you are choosing between a very simple model with 1 IV, and a very complex model with, say, 10 IVs, the very complex model needs to provide a much better fit to the data in order to justify its increased complexity. If it can't, then the more simpler model should be preferred.If the resulting p-value is sufficiently low (usually less than 0.05), we conclude that the more complex model is significantly better than the simpler model, and thus favor the more complex model. If the p-value is not sufficiently low (usually greater than 0.05), we should favor the simpler model. I compared two models. First model is best subset including 6 predicting variables and second model includes 12 predicting variables. As you can see, the result indicates that the more complex model has six additional degree of freedom, and a very small p-value (< .001). This means that adding interaction effects to the model did lead to a significantly improved fit over the model.

```
## Analysis of Variance Table
##
## Model 1: Life_exp ~ Mortality + GDP + GDP_capita + Unemployment + Exports +
##     Trade
## Model 2: Life_exp ~ year + Mortality + Fertility + GDP + GDP_capita +
##     Unemployment + Imports + Exports + Trade + (Mortality:Fertility) +
##     (Imports:Exports) + (GDP:GDP_capita)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    198 877.66
## 2    192 795.38  6    82.282 3.3104 0.004005 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

13. Conclusion

In this research, we have defined which variables predict life expectancy the most. To achieve our goal, we have analyzed several analysis in order to find the best fit model. Future researches can take our model into account and this can be used as a tool to measure life expectancy of a certain country. Our researech was done based on a reliable source from World Bank and this is a useful website to refer for future research on Life expectancey. For a future research, researchers should look into more predicting variables not included in this study.