

# Life Expectancy Prediction

The goal of this project is to build a Linear Regression model that can predict the life expectancy of the human population based on several factors such as the amount of alcohol consumption, average Body Mass Index (BMI), immunization of various vaccines among 1-year-olds such as Hepatitis B, Polio, and Diphtheria vaccines, and more, and also derive insights into what factors are significant in determining a higher or lower life expectancy of the human population.

The dataset contains data of each country recorded during the years between 2000 and 2015.  
The dataset is available at <https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>.

## Load Libraries

```
library(corrplot)

## corrplot 0.92 loaded

library(car)

## Loading required package: carData

library(olsrr)

##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

## Load Dataset

```
life = read.csv('Life Expectancy Data.csv')
```

The Life Expectancy Dataset contains the following fields:

- **Country** - Country Observed.
- **Year** - Year Observed.
- **Status** - Developed or Developing status.
- **Life expectancy** - Life Expectancy in age.
- **Adult.Mortality** - Adult Mortality Rates on both sexes (probability of dying between 15-60 years/1000 population).
- **infant.deaths** - Number of Infant Deaths per 1000 population.
- **Alcohol** - Alcohol recorded per capita (15+) consumption (in litres of pure alcohol).
- **percentage.expenditure** - Expenditure on health as a percentage of Gross Domestic Product per capita (%).
- **Hepatitis.B** - Hepatitis B (HepB) immunization coverage among 1-year-olds (%).
- **Measles** - Number of reported Measles 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 percentage of total government expenditure (%).
- **Diphtheria** - Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%).
- **HIV.AIDS** - Deaths per 1000 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** - Human Development Index in terms of income composition of resources (index ranging from 0 to 1).
- **Schooling** - Number of years of Schooling (years).

In total, there are 2938 observations of 22 variables with 20 of them being numerical and 2 categorical (Country and Status).

We will be using **Life.expectancy** to predict the life expectancy of the human population with the given dependent variables in the dataset.

## Clean Data

We will drop any observation that does not contain any value in any of its columns.

```
life = na.omit(life)
```

This shrinks our dataset to 1649 observations.

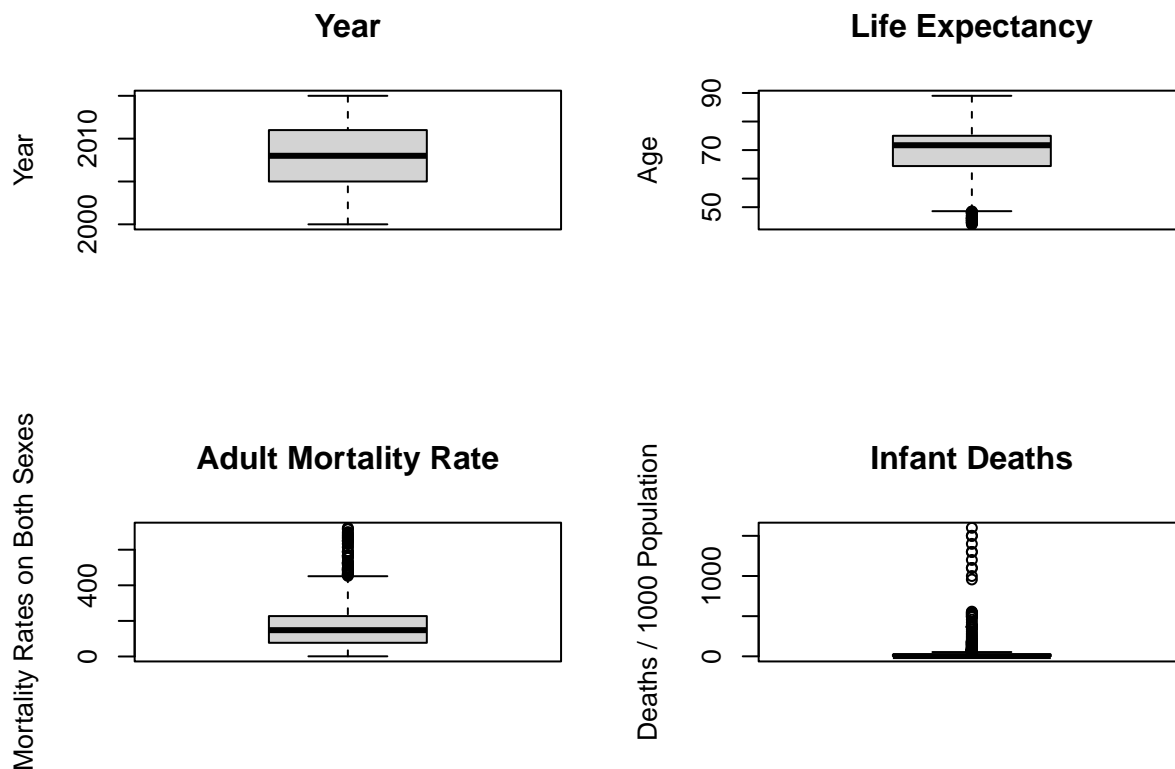
## Data Exploration

```
summary(life)
```

```
##      Country              Year      Status      Life.expectancy
## Length:1649      Min.      :2000      Length:1649      Min.      :44.0
## Class :character  1st Qu.:2005      Class :character  1st Qu.:64.4
## Mode  :character  Median :2008      Mode  :character  Median :71.7
##                               Mean  :2008      Mean  :69.3
##                               3rd Qu.:2011      3rd Qu.:75.0
##                               Max.   :2015      Max.   :89.0
## Adult.Mortality infant.deaths      Alcohol      percentage.expenditure
## Min.      : 1.0      Min.      : 0.00      Min.      : 0.010      Min.      : 0.00
## 1st Qu.: 77.0      1st Qu.: 1.00      1st Qu.: 0.810      1st Qu.: 37.44
## Median :148.0      Median : 3.00      Median : 3.790      Median : 145.10
## Mean   :168.2      Mean   : 32.55      Mean   : 4.533      Mean   : 698.97
## 3rd Qu.:227.0      3rd Qu.: 22.00      3rd Qu.: 7.340      3rd Qu.: 509.39
## Max.   :723.0      Max.   :1600.00      Max.   :17.870      Max.   :18961.35
## Hepatitis.B      Measles      BMI      under.five.deaths
## Min.      : 2.00      Min.      : 0      Min.      : 2.00      Min.      : 0.00
## 1st Qu.:74.00      1st Qu.: 0      1st Qu.:19.50      1st Qu.: 1.00
## Median :89.00      Median : 15      Median :43.70      Median : 4.00
## Mean   :79.22      Mean   : 2224      Mean   :38.13      Mean   : 44.22
## 3rd Qu.:96.00      3rd Qu.: 373      3rd Qu.:55.80      3rd Qu.: 29.00
## Max.   :99.00      Max.   :131441      Max.   :77.10      Max.   :2100.00
## Polio      Total.expenditure      Diphtheria      HIV.AIDS
## Min.      : 3.00      Min.      : 0.740      Min.      : 2.00      Min.      : 0.100
## 1st Qu.:81.00      1st Qu.: 4.410      1st Qu.:82.00      1st Qu.: 0.100
## Median :93.00      Median : 5.840      Median :92.00      Median : 0.100
```

```
## Mean :83.56 Mean : 5.956 Mean :84.16 Mean : 1.984
## 3rd Qu.:97.00 3rd Qu.: 7.470 3rd Qu.:97.00 3rd Qu.: 0.700
## Max. :99.00 Max. :14.390 Max. :99.00 Max. :50.600
## GDP Population thinness..1.19.years
## Min. : 1.68 Min. :3.400e+01 Min. : 0.100
## 1st Qu.: 462.15 1st Qu.:1.919e+05 1st Qu.: 1.600
## Median : 1592.57 Median :1.420e+06 Median : 3.000
## Mean : 5566.03 Mean :1.465e+07 Mean : 4.851
## 3rd Qu.: 4718.51 3rd Qu.:7.659e+06 3rd Qu.: 7.100
## Max. :119172.74 Max. :1.294e+09 Max. :27.200
## thinness.5.9.years Income.composition.of.resources Schooling
## Min. : 0.100 Min. :0.0000 Min. : 4.20
## 1st Qu.: 1.700 1st Qu.:0.5090 1st Qu.:10.30
## Median : 3.200 Median :0.6730 Median :12.30
## Mean : 4.908 Mean :0.6316 Mean :12.12
## 3rd Qu.: 7.100 3rd Qu.:0.7510 3rd Qu.:14.00
## Max. :28.200 Max. :0.9360 Max. :20.70
```

```
par(mfrow = c(2, 2))
boxplot(life$Year, main = 'Year', ylab = 'Year')
boxplot(life$Life.expectancy, main = 'Life Expectancy', ylab = 'Age')
boxplot(life$Adult.Mortality, main = 'Adult Mortality Rate', ylab = 'Mortality Rates on Both Sexes')
boxplot(life$infant.deaths, main = 'Infant Deaths', ylab = 'Deaths / 1000 Population')
```

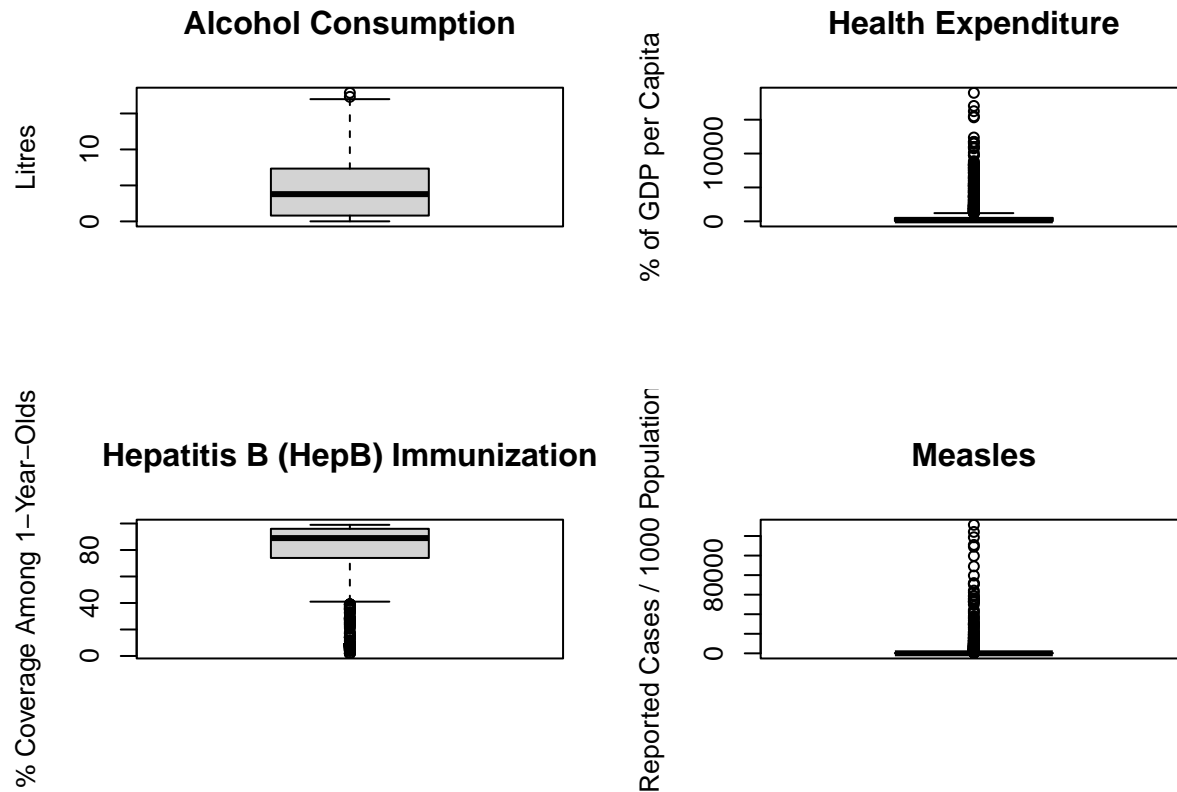


```
par(mfrow = c(2, 2))
boxplot(life$Alcohol, main = 'Alcohol Consumption', ylab = 'Litres')
boxplot(life$percentage.expenditure,
```

```

    main = 'Health Expenditure',
    ylab = '% of GDP per Capita')
boxplot(life$Hepatitis.B, main = 'Hepatitis B (HepB) Immunization', ylab =
    '% Coverage Among 1-Year-Olds')
boxplot(life$Measles, main = 'Measles', ylab = 'Reported Cases / 1000 Population')

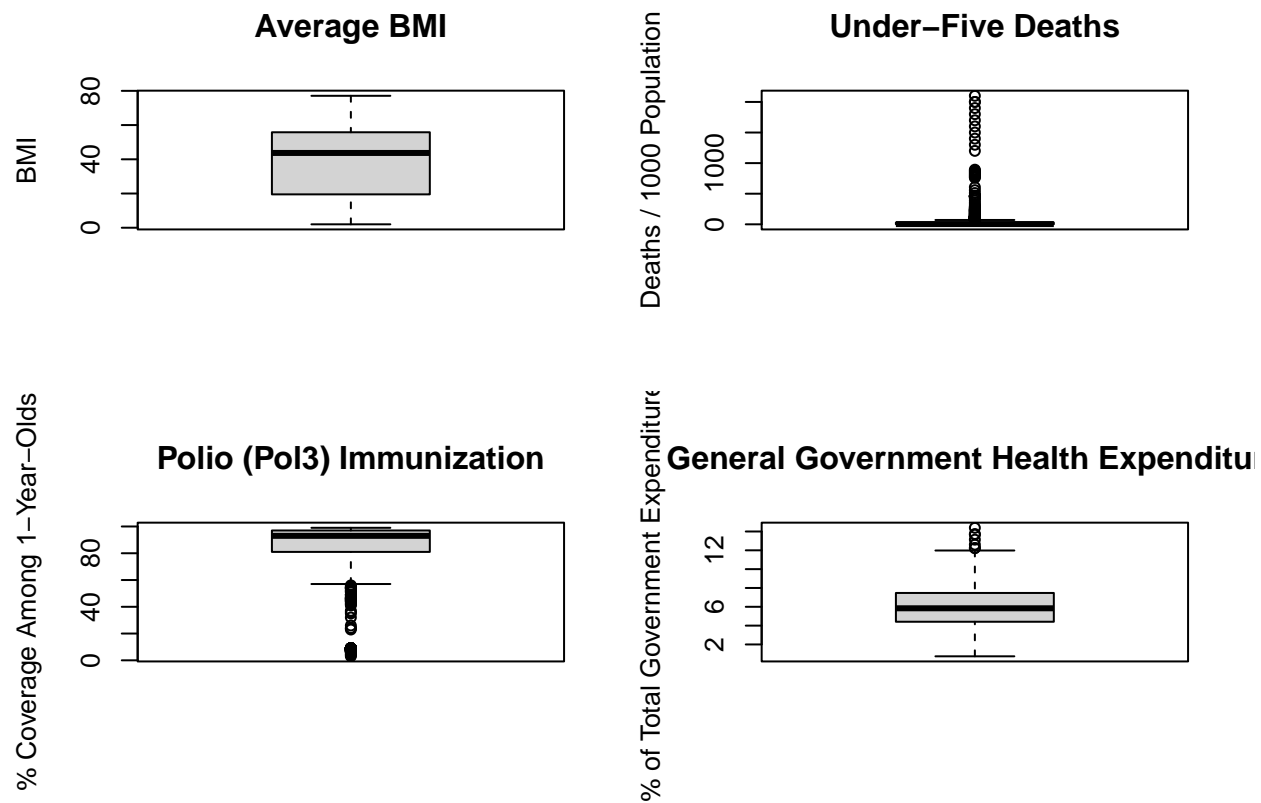
```



```

par(mfrow = c(2, 2))
boxplot(life$BMI, main = 'Average BMI', ylab = 'BMI')
boxplot(life$under.five.deaths, main = 'Under-Five Deaths', ylab = 'Deaths / 1000 Population')
boxplot(life$Polio, main = 'Polio (Pol3) Immunization', ylab = '% Coverage Among 1-Year-Olds')
boxplot(life$Total.expenditure, main = 'General Government Health Expenditure', ylab =
    '% of Total Government Expenditure')

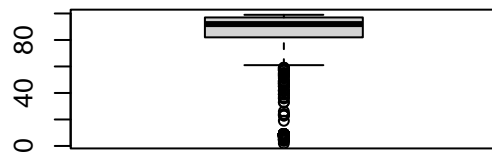
```



```
par(mfrow = c(2, 2))
boxplot(life$Diphtheria, main = 'DTP3 Immunization', ylab = '% Coverage Among 1-Year-Olds')
boxplot(life$HIV.AIDS, main = 'HIV/AIDS (0-4 Years)', ylab = 'Deaths / 1000 Live Births')
boxplot(life$GDP, main = 'GDP per Capita (in USD)')
boxplot(life$Population, main = 'Country Population')
```

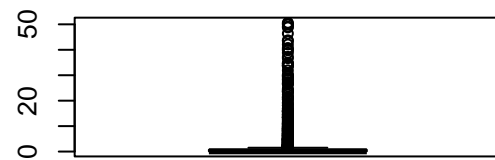
% Coverage Among 1–Year–Olds

**DTP3 Immunization**

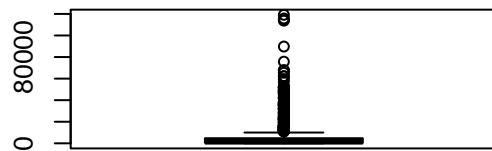


Deaths / 1000 Live Births

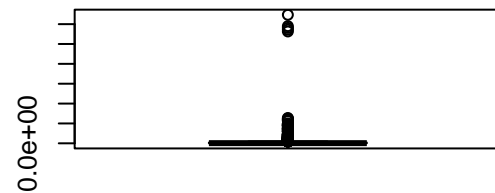
**HIV/AIDS (0–4 Years)**



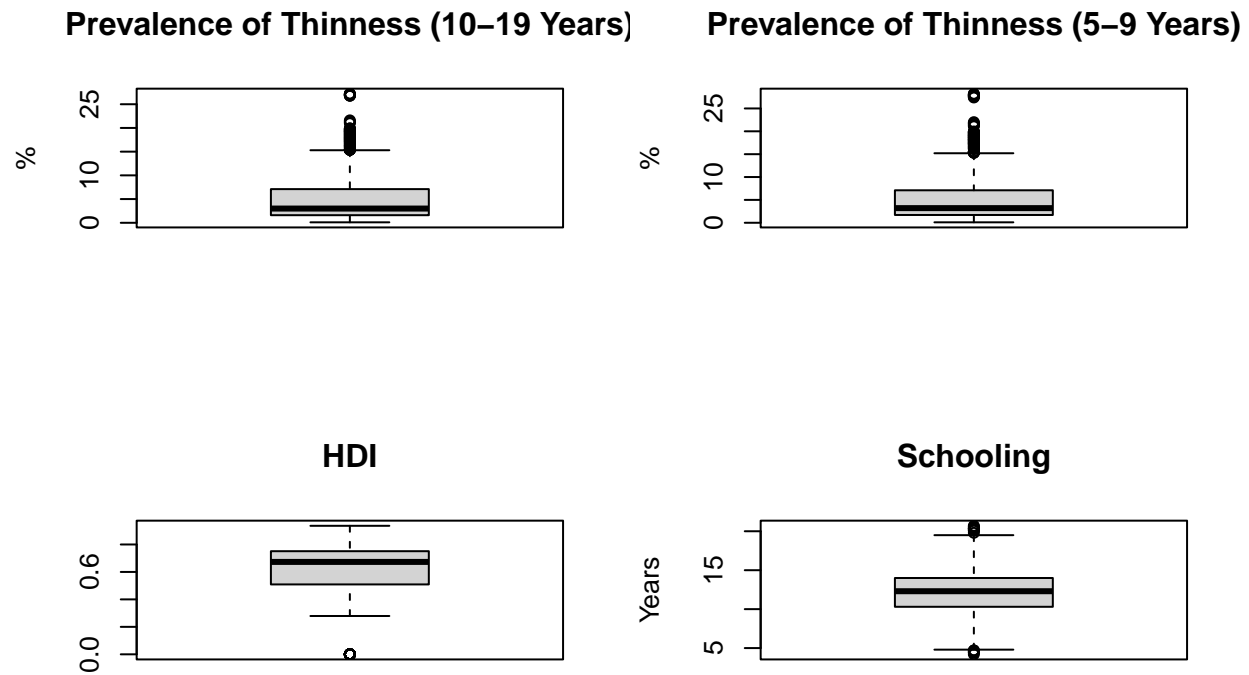
**GDP per Capita (in USD)**



**Country Population**



```
par(mfrow = c(2, 2))
boxplot(life$thinness..1.19.years, main = 'Prevalence of Thinness (10-19 Years)', ylab =
'%')
boxplot(life$thinness.5.9.years, main = 'Prevalence of Thinness (5-9 Years)', ylab =
'%')
boxplot(life$Income.composition.of.resources, main = 'HDI')
boxplot(life$Schooling, main = 'Schooling', ylab = 'Years')
```



## Feature Selection

We will be removing some of the variables for building the model due to the reasons mentioned below:

**Country** - Contains too many levels with no additional information to predict `Life expectancy`.

**Year** - Contains time series data with no additional information to predict `Life expectancy`.

```
life = life[, !(names(life) %in% c('Country', 'Year'))]
```

We will be mutating `Hepatitis.B`, `Polio` and `Diphtheria` for building the model since their range between the minimum value and the 1st Quartile is too wide. We will be mutating their values into 2 categorical values: '<90% Covered' and '>=90% Covered'.

```
life$Hepatitis.B = ifelse(life$Hepatitis.B < 90, '<90% Covered', '>=90% Covered')
life$Polio = ifelse(life$Polio < 90, '<90% Covered', '>=90% Covered')
life$Diphtheria = ifelse(life$Diphtheria < 90, '<90% Covered', '>=90% Covered')
```

This leaves us with 1649 observations of 20 variables with 16 of them being numerical and 4 categorical (`Status`, `Hepatitis.B`, `Polio` and `Diphtheria`).

```
summary(life)
```

##	Status	Life expectancy	Adult Mortality	infant.deaths
##	Length:1649	Min. :44.0	Min. : 1.0	Min. : 0.00
##	Class :character	1st Qu.:64.4	1st Qu.: 77.0	1st Qu.: 1.00
##	Mode :character	Median :71.7	Median :148.0	Median : 3.00
##		Mean :69.3	Mean :168.2	Mean : 32.55
##		3rd Qu.:75.0	3rd Qu.:227.0	3rd Qu.: 22.00

```

##           Max.      :89.0    Max.      :723.0    Max.      :1600.00
##   Alcohol      percentage.expenditure Hepatitis.B      Measles
##   Min.      : 0.010    Min.      : 0.00      Length:1649      Min.      : 0
##   1st Qu.: 0.810    1st Qu.: 37.44      Class :character  1st Qu.: 0
##   Median : 3.790    Median : 145.10      Mode  :character  Median : 15
##   Mean      : 4.533    Mean      : 698.97      Mean      : 2224
##   3rd Qu.: 7.340    3rd Qu.: 509.39      3rd Qu.: 373
##   Max.      :17.870    Max.      :18961.35      Max.      :131441
##   BMI          under.five.deaths      Polio          Total.expenditure
##   Min.      : 2.00    Min.      : 0.00      Length:1649      Min.      : 0.740
##   1st Qu.:19.50    1st Qu.: 1.00      Class :character  1st Qu.: 4.410
##   Median :43.70    Median : 4.00      Mode  :character  Median : 5.840
##   Mean      :38.13    Mean      : 44.22      Mean      : 5.956
##   3rd Qu.:55.80    3rd Qu.: 29.00      3rd Qu.: 7.470
##   Max.      :77.10    Max.      :2100.00      Max.      :14.390
##   Diphtheria      HIV.AIDS          GDP          Population
##   Length:1649      Min.      : 0.100    Min.      : 1.68      Min.      :3.400e+01
##   Class :character  1st Qu.: 0.100    1st Qu.: 462.15      1st Qu.:1.919e+05
##   Mode  :character  Median : 0.100    Median : 1592.57      Median :1.420e+06
##   Mean      : 1.984    Mean      : 5566.03      Mean      :1.465e+07
##   3rd Qu.: 0.700    3rd Qu.: 4718.51      3rd Qu.:7.659e+06
##   Max.      :50.600    Max.      :119172.74      Max.      :1.294e+09
##   thinness..1.19.years thinness.5.9.years Income.composition.of.resources
##   Min.      : 0.100      Min.      : 0.100      Min.      :0.0000
##   1st Qu.: 1.600      1st Qu.: 1.700      1st Qu.:0.5090
##   Median : 3.000      Median : 3.200      Median :0.6730
##   Mean      : 4.851      Mean      : 4.908      Mean      :0.6316
##   3rd Qu.: 7.100      3rd Qu.: 7.100      3rd Qu.:0.7510
##   Max.      :27.200      Max.      :28.200      Max.      :0.9360
##   Schooling
##   Min.      : 4.20
##   1st Qu.:10.30
##   Median :12.30
##   Mean      :12.12
##   3rd Qu.:14.00
##   Max.      :20.70

```

## Correlations

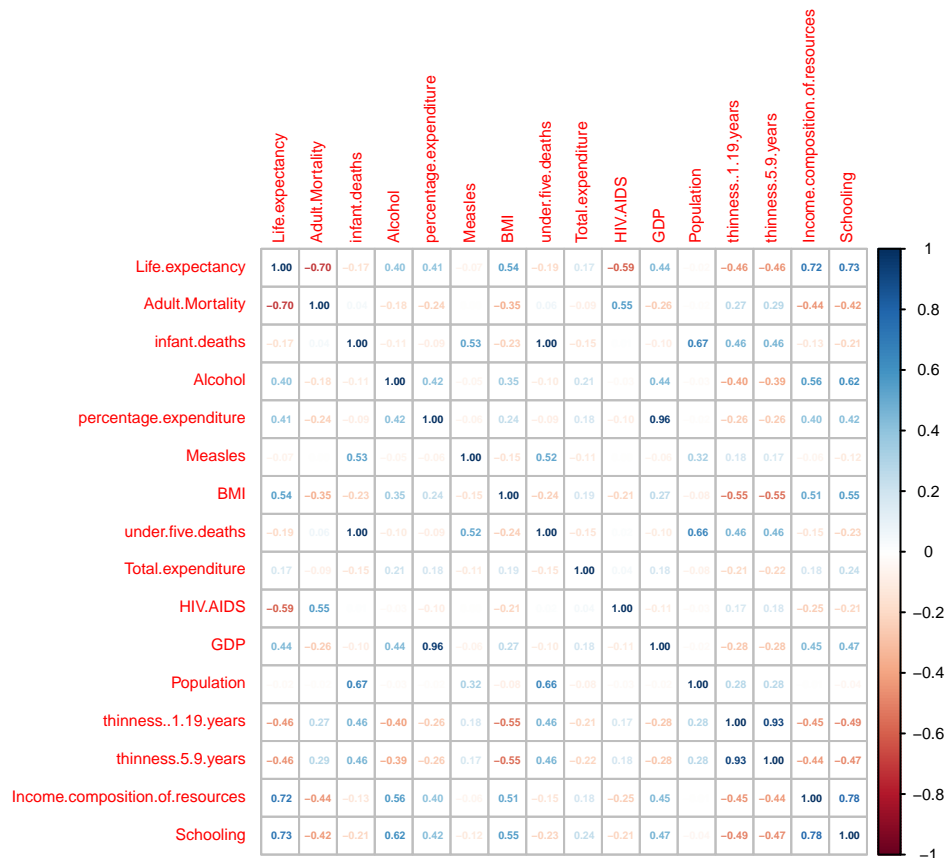
Plot the correlation matrix of the dataset.

```

life_nums = unlist(lapply(life, is.numeric), use.names = FALSE)
corrplot(
  cor(life[, life_nums]),
  method = 'number',
  tl.cex = 0.5,
  number.cex = 0.33,
  cl.cex = 0.5
)

```





There are a few takeaways from this correlation plot:

- Life expectancy has a somewhat strong positive correlation with Income.composition.of.resources and Schooling.
- Life expectancy has a negative correlation with Adult.Mortality, which makes sense since if the mortality rate of adult is high, then obviously the life expectancy will be low.
- Life expectancy has a very weak correlation with Measles and Population.
- There is a very strong correlation between infant.deaths and under.five.deaths, indicating multicollinearity between them. Therefore, we will remove under.five.deaths for building the model.

```
life = life[, !(names(life) %in% c('under.five.deaths'))]
```

## Model Building

We will now build a Linear Regression Model using all the remaining variables to predict the life expectancy of the human population.

```
lmod = lm(Life.expectancy ~ ., data = life)
summary(lmod)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.0291  -2.1529   0.0557   2.3893  11.5018
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.500e+01  8.108e-01  67.833 < 2e-16 ***
## StatusDeveloping -9.815e-01  3.464e-01  -2.834  0.00466 **
## Adult.Mortality -1.780e-02  9.674e-04 -18.399 < 2e-16 ***
## infant.deaths   -3.007e-03  1.266e-03  -2.376  0.01762 *
## Alcohol         -1.552e-01  3.380e-02  -4.590  4.77e-06 ***
## percentage.expenditure 3.491e-04  1.862e-04   1.875  0.06094 .
## Hepatitis.B>=90% Covered -6.372e-01  3.192e-01  -1.996  0.04611 *
## Measles         1.683e-05  1.079e-05   1.560  0.11906
## BMI             3.585e-02  6.161e-03   5.819  7.13e-09 ***
## Polio>=90% Covered  5.680e-01  4.439e-01   1.280  0.20087
## Total.expenditure  6.994e-02  4.179e-02   1.674  0.09439 .
## Diphtheria>=90% Covered  9.097e-01  4.899e-01   1.857  0.06352 .
## HIV.AIDS        -4.279e-01  1.849e-02 -23.142 < 2e-16 ***
## GDP             9.181e-06  2.925e-05   0.314  0.75368
## Population      2.496e-09  1.766e-09   1.414  0.15769
## thinness..1.19.years -5.018e-02  5.469e-02  -0.918  0.35899
## thinness.5.9.years  1.519e-03  5.374e-02   0.028  0.97745
## Income.composition.of.resources 1.048e+01  8.507e-01  12.316 < 2e-16 ***
## Schooling       8.843e-01  6.172e-02  14.328 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.686 on 1630 degrees of freedom
## Multiple R-squared:  0.8263, Adjusted R-squared:  0.8244
## F-statistic: 430.9 on 18 and 1630 DF, p-value: < 2.2e-16
```

There are a few takeaways from this model:

- The p-value of the model is  $2.2e-16 < 0.05$ , indicating that it is significant.
- The Adj R-squared value of the model is 0.8244, indicating that about 82.44% of the observed variation can be explained by the variables in the model, which is quite a good result and can possibly be improved even further with model selection.
- Adult.Mortality, Alcohol, BMI, HIV.AIDS, Income.composition.of.resources and Schooling are the most significant variables with p-value  $< 0.5$ .
- From the model we can interpret that StatusDeveloping, Adult.Mortality, infant.deaths, Alcohol, HIV.AIDS, and thinness..1.19.years may have a negative effect on life expectancy.
- From the model we can interpret that Income.composition.of.resources has a strong positive effect on life expectancy.
- A peculiar result we can interpret from the model is that Hepatitis.B90% Covered and Schooling also have a negative effect on life expectancy.

## Model Selection

We will now generate models by using different techniques like Forward Selection Method, Backward Elimination Method and Stepwise Selection Method.

Build Model using Forward Selection Method.

```
ols_step_forward_p(lmod)
```

```
##
##              Selection Summary
## -----
##              Variable              Adj.
```

## Step	Entered	R-Square	R-Square	C(p)	AIC	RMSE
## 1	Schooling	0.5294	0.5292	2771.7513	10612.7157	6.0362
## 2	HIV.AIDS	0.7304	0.7301	887.6286	9696.3271	4.5704
## 3	Adult.Mortality	0.7871	0.7867	357.3801	9308.9473	4.0627
## 4	Income.composition.of.resources	0.8092	0.8087	152.1307	9130.3986	3.8474
## 5	percentage.expenditure	0.8147	0.8141	102.1617	9083.8457	3.7924
## 6	BMI	0.8201	0.8194	54.0203	9037.6049	3.7384
## 7	Diphtheria	0.8218	0.8211	39.2920	9023.1915	3.7210
## 8	Alcohol	0.8231	0.8222	29.5343	9013.5567	3.7090
## 9	thinness..1.19.years	0.8240	0.8230	22.9694	9007.0292	3.7006
## 10	Status	0.8249	0.8238	16.6366	9000.6904	3.6924
## 11	Hepatitis.B	0.8252	0.8240	15.5038	8999.5443	3.6900
## 12	Total.expenditure	0.8255	0.8242	14.8813	8998.9062	3.6881
## 13	infant.deaths	0.8257	0.8243	14.8516	8998.8614	3.6870
## 14	Measles	0.8259	0.8244	14.7734	8998.7652	3.6858
## 15	Population	0.8262	0.8246	14.7661	8998.7380	3.6846
## 16	Polio	0.8263	0.8246	15.0990	8999.0524	3.6839

```

lmod_forward = lm(
  Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources + percent
  BMI + Diphtheria + Alcohol + thinness..1.19.years + Status + Hepatitis.B +
  Total.expenditure + infant.deaths + Measles + Population + Polio,
  data = life
)
summary(lmod_forward)

```

```

##
## Call:
## lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
##   Income.composition.of.resources + percentage.expenditure +
##   BMI + Diphtheria + Alcohol + thinness..1.19.years + Status +
##   Hepatitis.B + Total.expenditure + infant.deaths + Measles +
##   Population + Polio, data = life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.0291  -2.1512   0.0485   2.3846  11.4744
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.499e+01  8.094e-01  67.942  < 2e-16 ***
## Schooling       8.858e-01  6.141e-02  14.426  < 2e-16 ***
## HIV.AIDS      -4.279e-01  1.848e-02 -23.157  < 2e-16 ***
## Adult.Mortality -1.779e-02  9.656e-04 -18.428  < 2e-16 ***
## Income.composition.of.resources 1.050e+01  8.481e-01  12.378  < 2e-16 ***
## percentage.expenditure  4.043e-04  6.128e-05   6.597 5.64e-11 ***
## BMI            3.579e-02  6.096e-03   5.871 5.24e-09 ***
## Diphtheria>=90% Covered  9.024e-01  4.888e-01   1.846  0.06505 .
## Alcohol       -1.551e-01  3.378e-02 -4.591  4.75e-06 ***
## thinness..1.19.years -4.903e-02  2.788e-02 -1.758  0.07885 .
## StatusDeveloping -9.882e-01  3.454e-01 -2.861  0.00428 **
## Hepatitis.B>=90% Covered -6.299e-01  3.180e-01 -1.981  0.04780 *
## Total.expenditure  6.940e-02  4.169e-02   1.664  0.09621 .

```

```
## infant.deaths          -2.996e-03  1.259e-03  -2.379  0.01746 *
## Measles                1.682e-05  1.077e-05   1.561  0.11869
## Population             2.486e-09  1.764e-09   1.409  0.15892
## Polio>=90% Covered     5.728e-01  4.433e-01   1.292  0.19657
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## Residual standard error: 3.684 on 1632 degrees of freedom
```

```
## Multiple R-squared:  0.8263, Adjusted R-squared:  0.8246
```

```
## F-statistic: 485.3 on 16 and 1632 DF,  p-value: < 2.2e-16
```

Build Model using Backward Elimination Method.

```
ols_step_backward_p(lmod)
```

```
##
```

```
##
```

```
## Elimination Summary
```

```
## -----
##          Variable                Adj.
## Step      Removed      R-Square  R-Square    C(p)      AIC      RMSE
## -----
##    1  thinness.5.9.years    0.8263    0.8245   17.0008   9000.9530   3.6849
##    2      GDP              0.8263    0.8246   15.0990   8999.0524   3.6839
## -----
```

```
lmod_backward = lm(
  Life.expectancy ~ Status + Adult.Mortality + infant.deaths + Alcohol +
  percentage.expenditure + Hepatitis.B + Measles + BMI + Polio + Total.expenditure +
  Diphtheria + HIV.AIDS + Population + thinness..1.19.years + Income.composition.of.resources +
  Schooling,
  data = life
)
summary(lmod_backward)
```

```
##
```

```
## Call:
```

```
## lm(formula = Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
##      Alcohol + percentage.expenditure + Hepatitis.B + Measles +
##      BMI + Polio + Total.expenditure + Diphtheria + HIV.AIDS +
##      Population + thinness..1.19.years + Income.composition.of.resources +
##      Schooling, data = life)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -17.0291  -2.1512   0.0485   2.3846  11.4744
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.499e+01  8.094e-01  67.942 < 2e-16 ***
## StatusDeveloping -9.882e-01  3.454e-01  -2.861  0.00428 **
## Adult.Mortality  -1.779e-02  9.656e-04 -18.428 < 2e-16 ***
## infant.deaths    -2.996e-03  1.259e-03  -2.379  0.01746 *
## Alcohol          -1.551e-01  3.378e-02  -4.591  4.75e-06 ***
## percentage.expenditure  4.043e-04  6.128e-05   6.597  5.64e-11 ***
## Hepatitis.B>=90% Covered -6.299e-01  3.180e-01  -1.981  0.04780 *
```

```
## Measles          1.682e-05  1.077e-05  1.561  0.11869
## BMI              3.579e-02  6.096e-03  5.871  5.24e-09 ***
## Polio>=90% Covered 5.728e-01  4.433e-01  1.292  0.19657
## Total.expenditure 6.940e-02  4.169e-02  1.664  0.09621 .
## Diphtheria>=90% Covered 9.024e-01  4.888e-01  1.846  0.06505 .
## HIV.AIDS         -4.279e-01  1.848e-02 -23.157 < 2e-16 ***
## Population        2.486e-09  1.764e-09  1.409  0.15892
## thinness..1.19.years -4.903e-02  2.788e-02 -1.758  0.07885 .
## Income.composition.of.resources 1.050e+01  8.481e-01  12.378 < 2e-16 ***
## Schooling         8.858e-01  6.141e-02  14.426 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.684 on 1632 degrees of freedom
## Multiple R-squared:  0.8263, Adjusted R-squared:  0.8246
## F-statistic: 485.3 on 16 and 1632 DF,  p-value: < 2.2e-16
```

Build Model using Stepwise Selection Method.

```
ols_step_both_p(lmod)
```

```
##
##                               Stepwise Selection Summary
## -----
```

## Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC
## 1	Schooling	addition	0.529	0.529	2771.7510	10612.71
## 2	HIV.AIDS	addition	0.730	0.730	887.6290	9696.32
## 3	Adult.Mortality	addition	0.787	0.787	357.3800	9308.94
## 4	Income.composition.of.resources	addition	0.809	0.809	152.1310	9130.39
## 5	percentage.expenditure	addition	0.815	0.814	102.1620	9083.84
## 6	BMI	addition	0.820	0.819	54.0200	9037.60
## 7	Diphtheria	addition	0.822	0.821	39.2920	9023.19
## 8	Alcohol	addition	0.823	0.822	29.5340	9013.55
## 9	thinness..1.19.years	addition	0.824	0.823	22.9690	9007.02
## 10	Status	addition	0.825	0.824	16.6370	9000.69
## 11	Hepatitis.B	addition	0.825	0.824	15.5040	8999.54

```
## -----
```

```
lmod_stepwise = lm(
  Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
    percentage.expenditure + BMI + Diphtheria + Alcohol + thinness..1.19.years +
    Status + Hepatitis.B,
  data = life
)
summary(lmod_stepwise)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
##     Income.composition.of.resources + percentage.expenditure +
##     BMI + Diphtheria + Alcohol + thinness..1.19.years + Status +
##     Hepatitis.B, data = life)
##
## Residuals:
```

```
##      Min      1Q   Median      3Q      Max
## -17.2593 -2.1481  0.0745  2.4046 11.5838
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.534e+01  7.750e-01  71.401 < 2e-16 ***
## Schooling      9.061e-01  6.102e-02  14.848 < 2e-16 ***
## HIV.AIDS      -4.239e-01  1.833e-02 -23.122 < 2e-16 ***
## Adult.Mortality -1.779e-02  9.636e-04 -18.464 < 2e-16 ***
## Income.composition.of.resources 1.037e+01  8.444e-01  12.280 < 2e-16 ***
## percentage.expenditure 4.098e-04  6.119e-05  6.698 2.90e-11 ***
## BMI            3.610e-02  6.071e-03  5.946 3.36e-09 ***
## Diphtheria>=90% Covered 1.439e+00  3.443e-01  4.181 3.05e-05 ***
## Alcohol       -1.605e-01  3.353e-02 -4.788 1.84e-06 ***
## thinness..1.19.years -7.223e-02  2.491e-02 -2.900 0.00378 **
## StatusDeveloping -1.014e+00  3.454e-01 -2.934 0.00339 **
## Hepatitis.B>=90% Covered -5.567e-01  3.149e-01 -1.768 0.07723 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.69 on 1637 degrees of freedom
## Multiple R-squared:  0.8252, Adjusted R-squared:  0.824
## F-statistic: 702.7 on 11 and 1637 DF, p-value: < 2.2e-16
```

Build Model using All Possible Regressions Method.

```
# ols_step_all_possible(lmod, sbc = TRUE)
```

In summary, variables chosen by the methods:

Model Selection Method	Status	Life.expectancy	Adult.Mortality	infant.deaths
Forward Selection	x		x	x
Backward Elimination	x	x	x	x
Stepwise Selection	x		x	

Model Selection Method	Alcohol	percentage.expenditure	Hepatitis.B	Measles	BMI
Forward Selection	x	x	x	x	x
Backward Elimination	x	x	x	x	x
Stepwise Selection	x	x	x		x

Model Selection Method	Polio	Total.expenditure	Diphtheria	HIV.AIDS	GDP
Forward Selection	x	x	x	x	
Backward Elimination	x	x	x	x	
Stepwise Selection			x	x	

Model Selection Method	Population	thinness..1.19.years	thinness.5.9.years
Forward Selection	x	x	
Backward Elimination	x	x	

Model Selection Method	Population	thinness..1.19.years	thinness.5.9.years
Stepwise Selection		x	

Model Selection Method	Income.composition.of.resources	Schooling
Forward Selection	x	x
Backward Elimination	x	x
Stepwise Selection	x	x

Adj. R-squared values of the above models:

```
data.frame(
  model = c('lmod', 'lmod_forward', 'lmod_backward', 'lmod_stepwise'),
  AdjRsquare = c(
    summary(lmod)$adj.r.square,
    summary(lmod_forward)$adj.r.square,
    summary(lmod_backward)$adj.r.square,
    summary(lmod_stepwise)$adj.r.square
  )
)
```

```
##          model AdjRsquare
## 1          lmod  0.8244244
## 2 lmod_forward  0.8246289
## 3 lmod_backward 0.8246289
## 4 lmod_stepwise 0.8240486
```

We will be choosing the model chosen by Forward Selection method `lmod_forward` as it has the highest Adj. R-squared value.

```
lmod_final = lmod_forward
summary(lmod_final)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
##      Income.composition.of.resources + percentage.expenditure +
##      BMI + Diphtheria + Alcohol + thinness..1.19.years + Status +
##      Hepatitis.B + Total.expenditure + infant.deaths + Measles +
##      Population + Polio, data = life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.0291  -2.1512   0.0485   2.3846  11.4744
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.499e+01  8.094e-01  67.942 < 2e-16 ***
## Schooling       8.858e-01  6.141e-02  14.426 < 2e-16 ***
## HIV.AIDS      -4.279e-01  1.848e-02 -23.157 < 2e-16 ***
## Adult.Mortality -1.779e-02  9.656e-04 -18.428 < 2e-16 ***
## Income.composition.of.resources 1.050e+01  8.481e-01  12.378 < 2e-16 ***
## percentage.expenditure  4.043e-04  6.128e-05   6.597 5.64e-11 ***
## BMI            3.579e-02  6.096e-03   5.871 5.24e-09 ***
```

```
## Diphtheria>=90% Covered      9.024e-01  4.888e-01   1.846  0.06505 .
## Alcohol                     -1.551e-01  3.378e-02  -4.591  4.75e-06 ***
## thinness..1.19.years        -4.903e-02  2.788e-02  -1.758  0.07885 .
## StatusDeveloping            -9.882e-01  3.454e-01  -2.861  0.00428 **
## Hepatitis.B>=90% Covered    -6.299e-01  3.180e-01  -1.981  0.04780 *
## Total.expenditure           6.940e-02  4.169e-02   1.664  0.09621 .
## infant.deaths               -2.996e-03  1.259e-03  -2.379  0.01746 *
## Measles                     1.682e-05  1.077e-05   1.561  0.11869
## Population                  2.486e-09  1.764e-09   1.409  0.15892
## Polio>=90% Covered          5.728e-01  4.433e-01   1.292  0.19657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.684 on 1632 degrees of freedom
## Multiple R-squared:  0.8263, Adjusted R-squared:  0.8246
## F-statistic: 485.3 on 16 and 1632 DF,  p-value: < 2.2e-16
```

## Model Error Estimation

We will now use our final model to see how well it performs in predicting the life expectancy of the human population.

```
result = predict(lmod_final, life)
```

Mean Squared Error:

```
mse = mean((life$Life.expectancy - result) ^ 2)
mse
```

```
## [1] 13.43106
```

Root Mean Squared Error:

```
rmse = sqrt(mse)
rmse
```

```
## [1] 3.664841
```

Mean Absolute Error:

```
n = length(result)
sum = 0

for (i in 1:n) {
  sum = sum + abs(life$Life.expectancy[i] - result[i])
}

mae = sum / n
mae
```

```
##      1
## 2.817618
```

In summary,

```
data.frame(
  Method = c('MSE', 'RMSE', 'MAE'),
  Result = c(mse, rmse, mae)
)
```

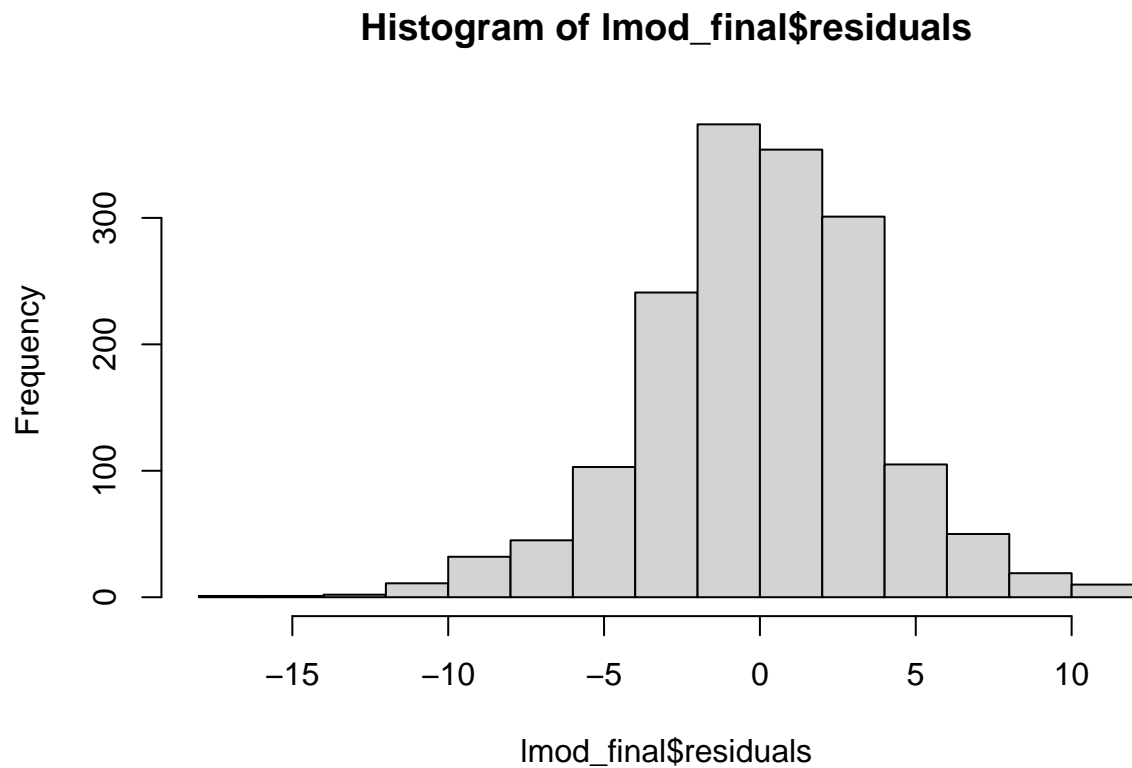


```
##    Method    Result
## 1    MSE 13.431062
## 2   RMSE  3.664841
## 3    MAE  2.817618
```

## Model Adequacy Checking

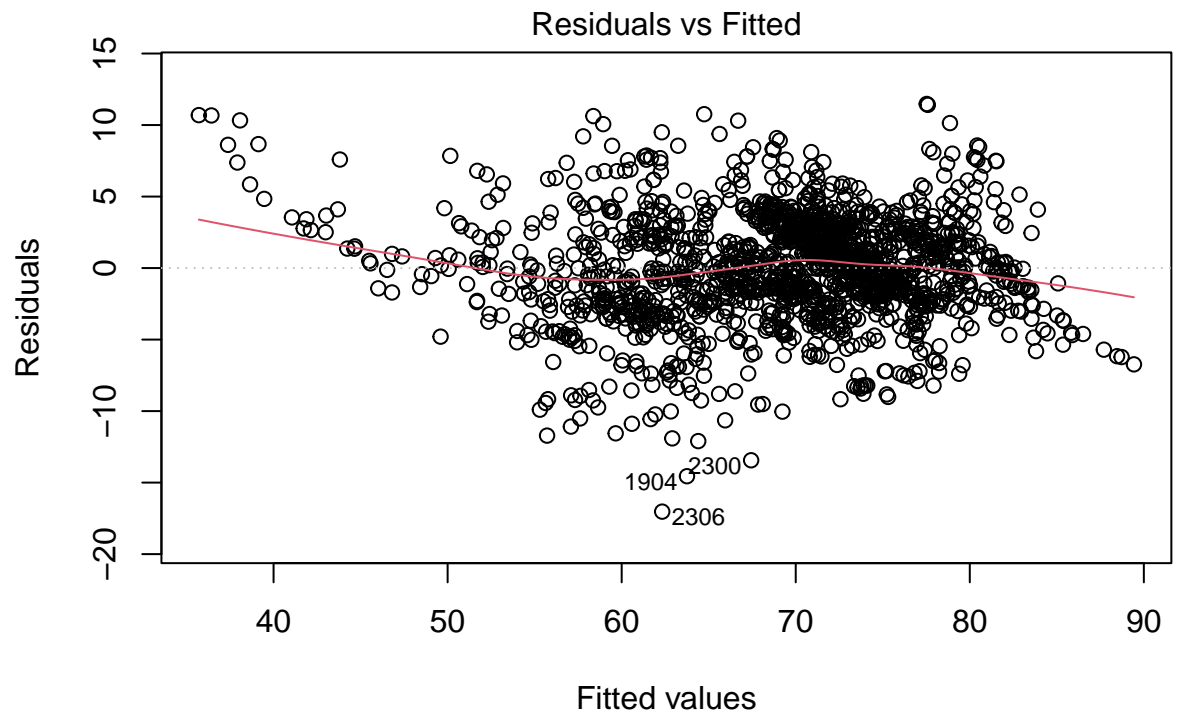
Normality Testing:

```
hist(lmod_final$residuals, breaks = 20)
```

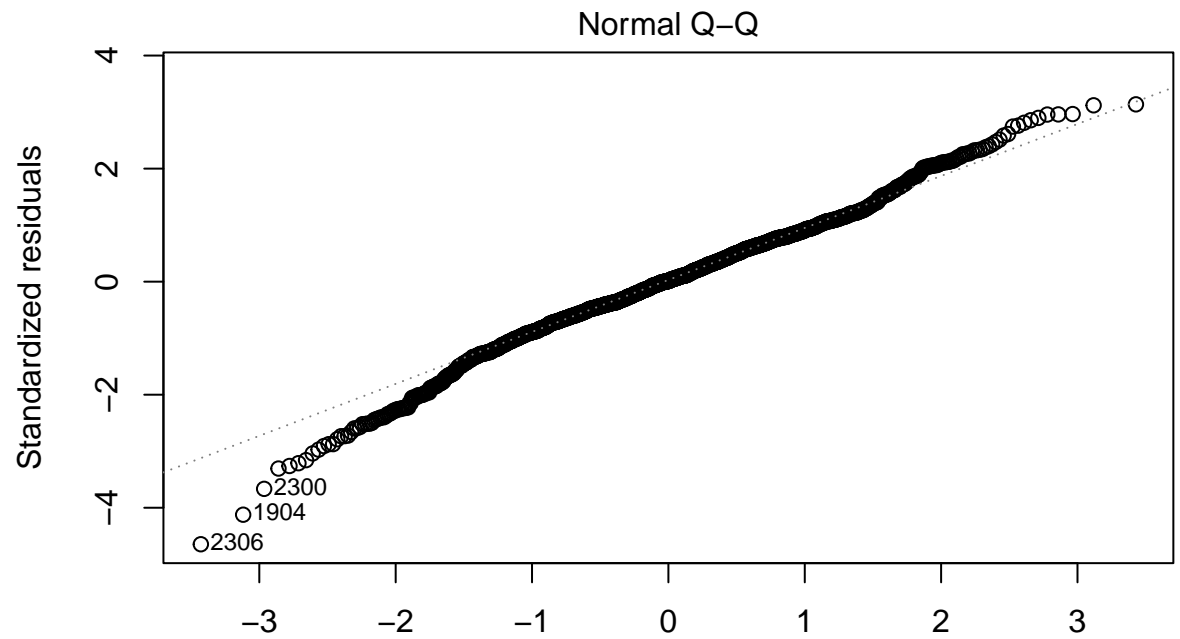


Most of the residuals seem to be distributed in the center, indicating that they are distributed normally.

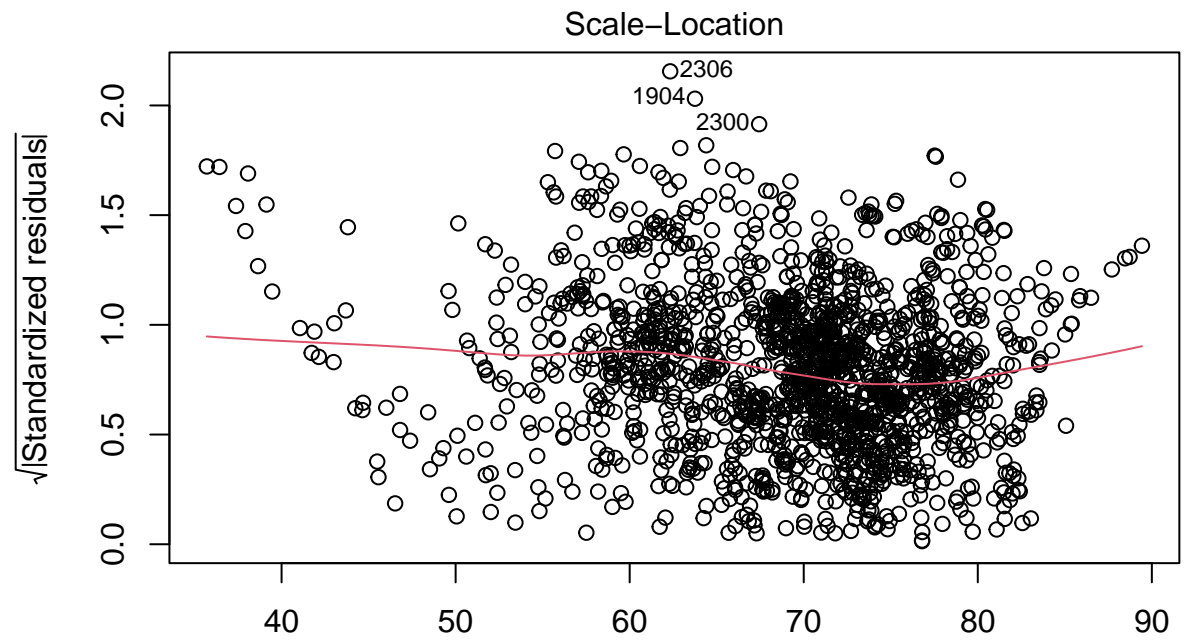
```
plot(lmod_final, which = c(1:6))
```



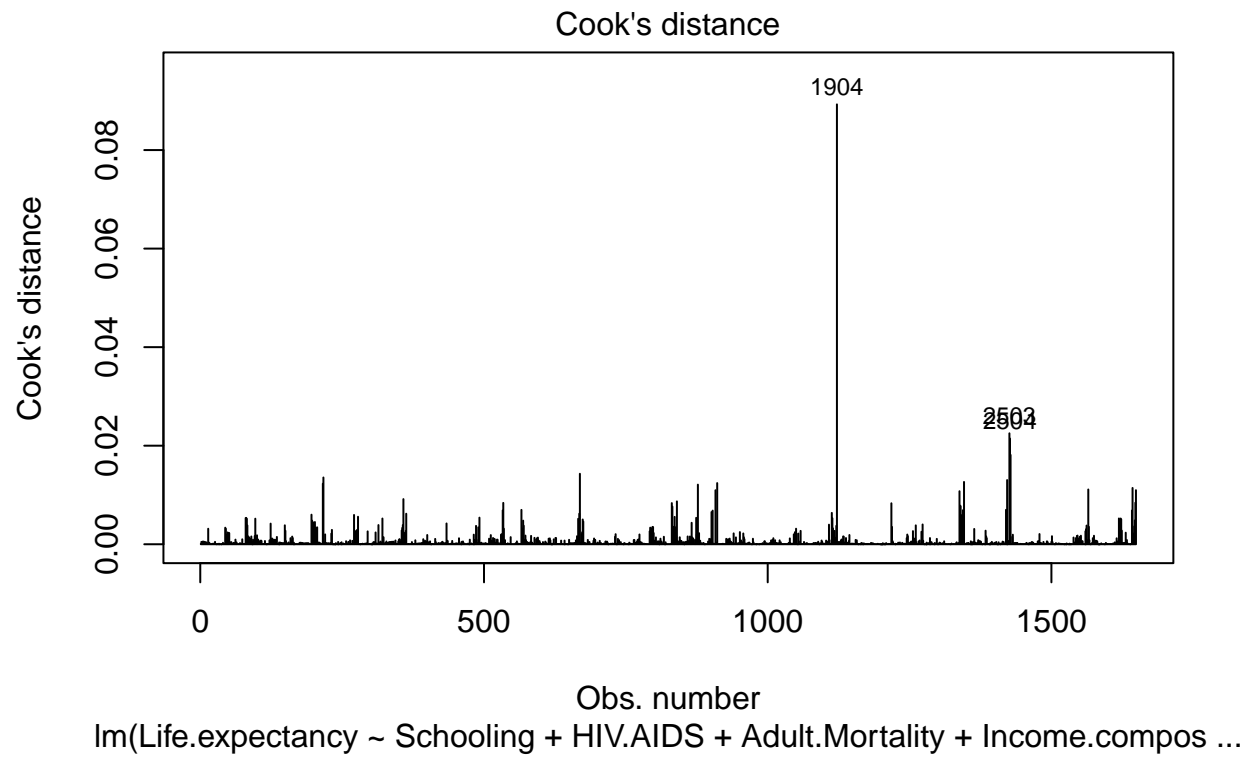
lm(Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.compos ...

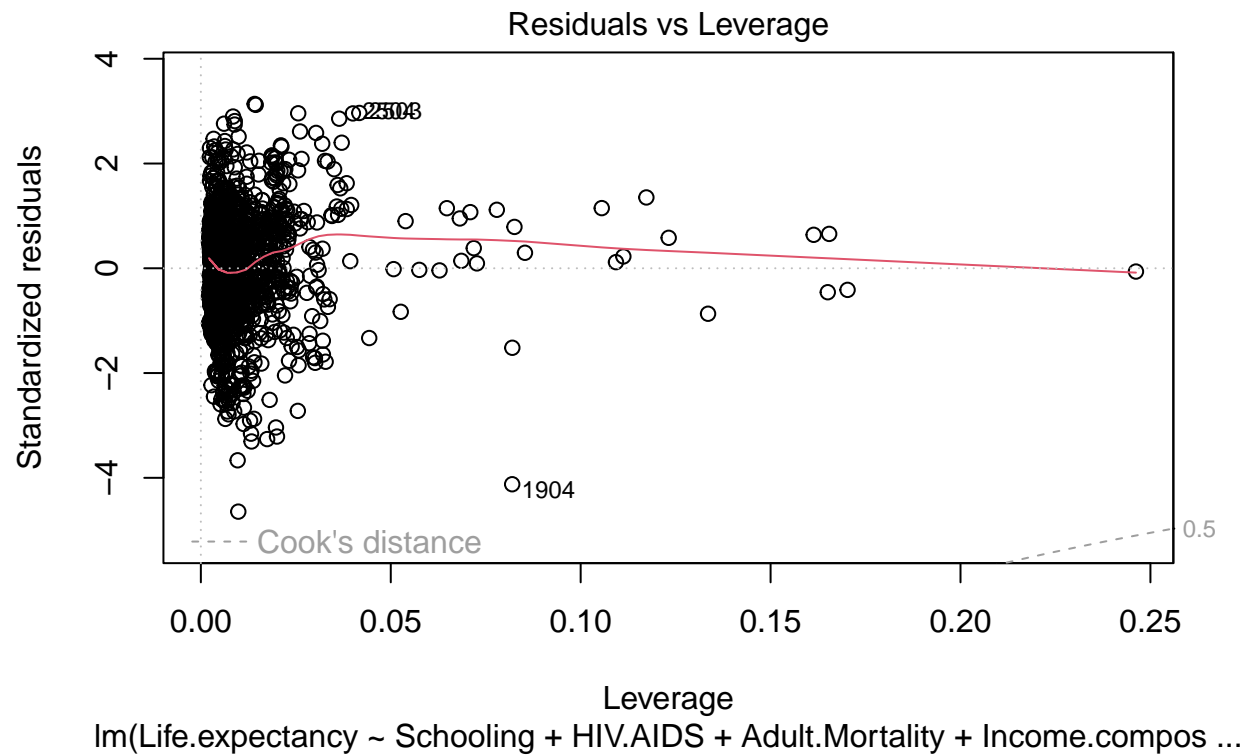


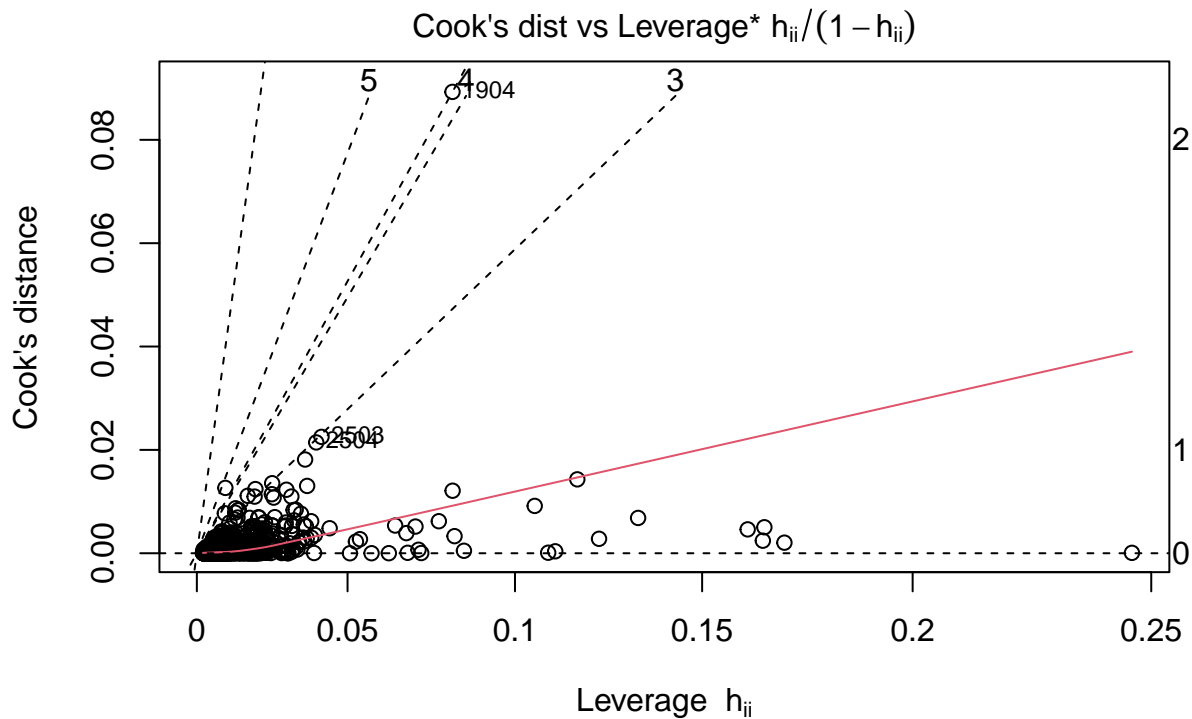
Im(Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.compos ...



Fitted values  
 $\text{lm}(\text{Life.expectancy} \sim \text{Schooling} + \text{HIV.AIDS} + \text{Adult.Mortality} + \text{Income.compos} \dots)$







lm(Life expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.compos ...

There is no obvious observable pattern in the above plots, indicating that the model is appropriate.

Multicollinearity Test:

```
vif(lmod_final)
```

##	Schooling	HIV.AIDS
##	3.578091	1.509013
##	Adult.Mortality	Income.composition.of.resources
##	1.778090	2.927679
##	percentage.expenditure	BMI
##	1.411270	1.761017
##	Diphtheria	Alcohol
##	7.102613	2.249650
##	thinness..1.19.years	Status
##	1.996547	1.815140
##	Hepatitis.B	Total.expenditure
##	3.072344	1.116175
##	infant.deaths	Measles
##	2.811727	1.433389
##	Population	Polio
##	1.876386	5.834447

A VIF > 10 implies serious problems with multicollinearity.

Since the VIF for all of the predictors is less than 10, there seems to be no issue with multicollinearity.