Predicting Life-Expectancy Based on Socio-Economic and Health Factors

Christen Ye, David Chung, Jason Xu, Mars Zan

STAT 306 Group Project

Introduction

In this project, we are interested in exploring factors affecting human life expectancy. The dataset we used in the current project is composed of health-related data from the World Health Organization and socio-economic data from the United Nations for 193 countries around the world between the years 2000-2015. We have fitted multiple regression models to test our research questions:

1. How does the baseline life expectancy in developing countries differ significantly compared to that of developed countries?
2. Which socio-economic and health-related factors exert a significant effect on life expectancy in developing and developed countries?

We expect that the expected longevity in a given country will depend linearly on socioeconomic factors and health-related factors. After applying the backward selection algorithm, we come to a final model with 9 predictors. The final selected model includes the following variables of interest:

- ***Life expectancy***: life expectancy measured in years of age
- ***Status***: a country is labeled by the UN as either a developing or developed country (the life expectancy in the developed countries is chosen as the baseline in our analysis)
- ***Adult Mortality***: number of adult death per 1000 population between 15 and 60 years of age
- ***Alcohol***: average alcohol consumption per year per person measured in liters of pure alcohol
- ***Under Five Deaths***: the number of children death under the age of 5 per 1000 population
- ***Diphtheria***: the percentage of population immunized against diphtheria tetanus toxoid and pertussis (DTP3) among 1-year-olds
- ***log(HIV/AIDS)***: the log of the number of death due to HIV/AIDS per 1000 population
- ***GDP***: Gross Domestic Product per capita measured in USD
- ***Income Composition of Resources***: a measure of the efficiency of the country's utilization of its natural resources in percentage
- ***Schooling***: number of years of schooling

Analysis

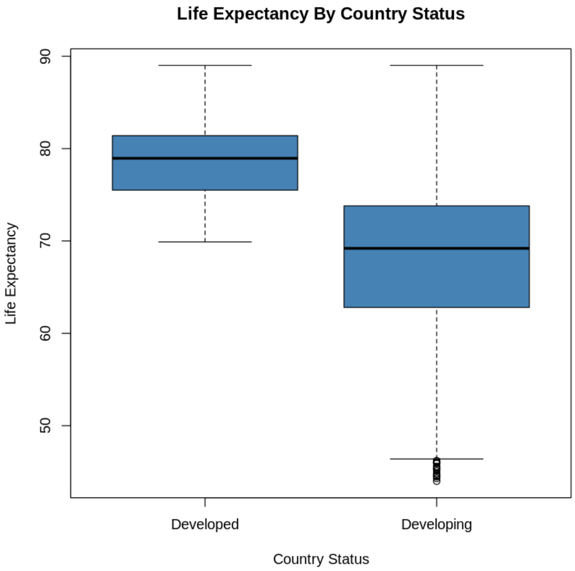


Figure 1. Life Expectancy for Developed and Developing Countries

From the above box plots, we immediately notice a lower average life expectancy for developing countries. This hinted at the possibility that there might exist a significant difference in baseline life expectancy between developing and developed countries. We also notice that the distribution of life expectancy for developing countries is wider (have a larger variance) compared to the developed countries.

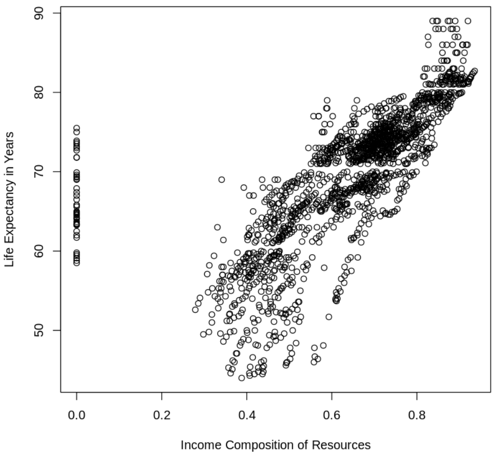


Figure 2. Scatter Plot of Life Expectancy Against Income Composition of Resources

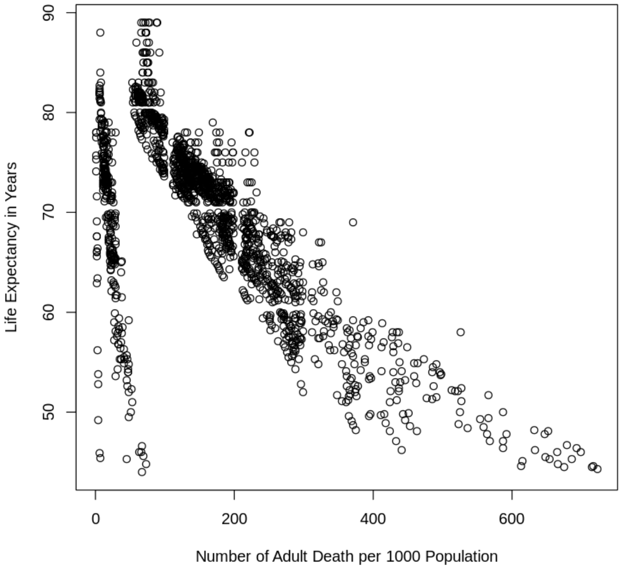


Figure 3. Scatter Plot of Life Expectancy vs. Number of Adult Death Per 1000 Population

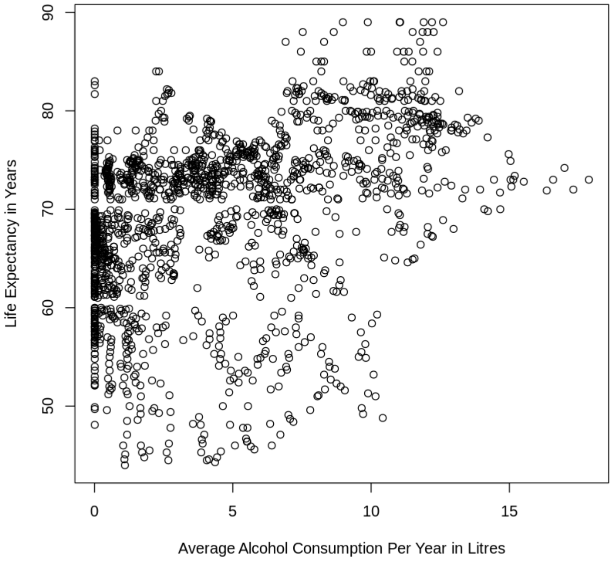


Figure 4. Scatter Plot of Life Expectancy Against Yearly Alcohol Consumption in Liters

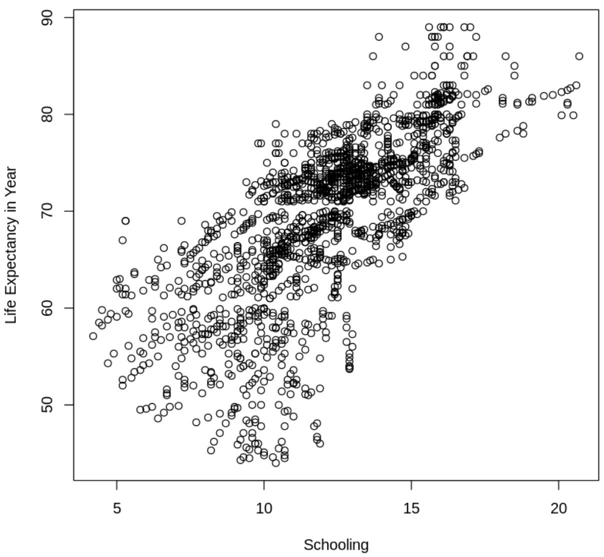


Figure 5. Scatter Plot of Life Expectancy Against Schooling in Years

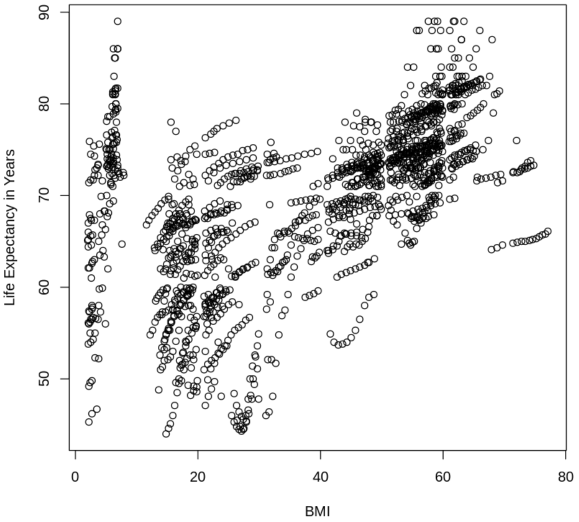


Figure 6. Scatter Plot of Life Expectancy Against Schooling in Years

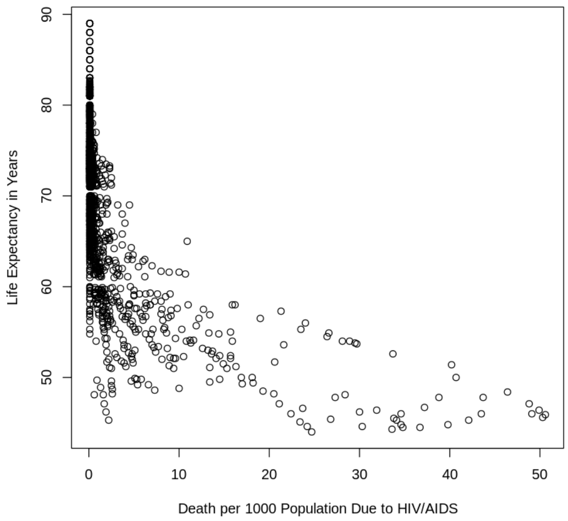
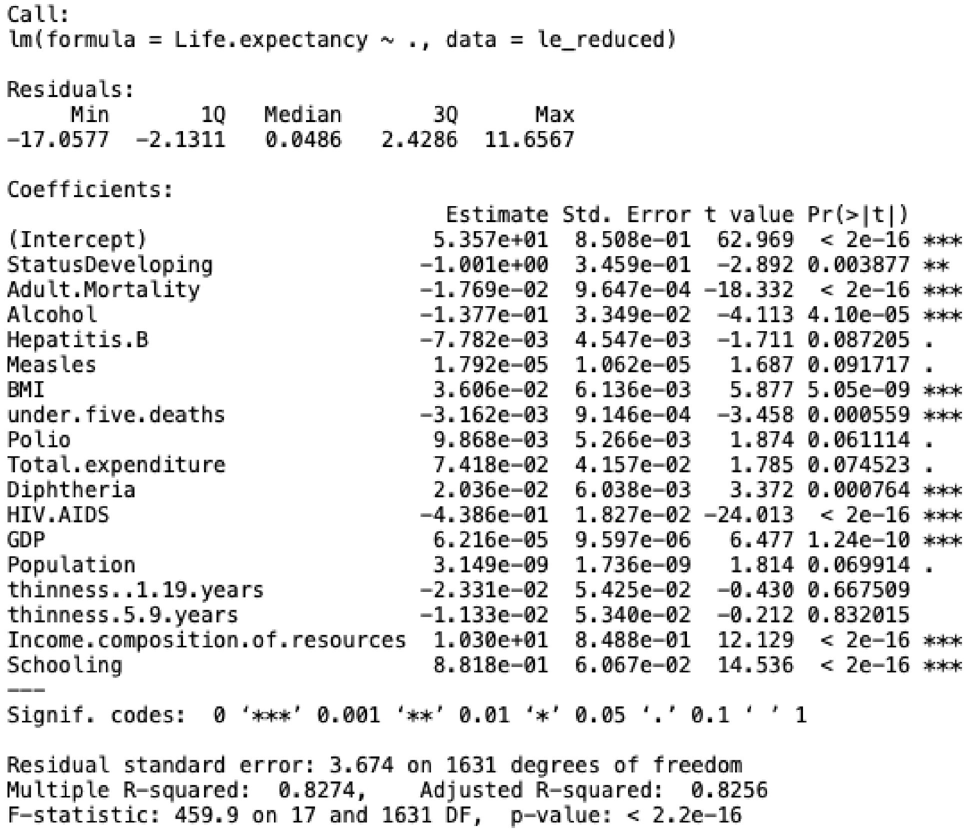


Figure 7. Scatter Plot of Life Expectancy Against HIV/AIDS Death Per 1000 Population

From the above scatterplots, we can tell that there exists a linear relationship between the response variable and the selected variables, namely *Income Composition of Resources*, *Number of Adult Death per 1000 Population*, *Alcohol Consumption*, and *Schooling*. In addition, we can tell that there are influential cases within our dataset. This is evident in Figure 2,3, and 6 as there are pockets of data points that greatly deviate from the rest of the data in their x values. We will be conducting an influence analysis later on to address this issue. We also notice that the *HIV/AIDS* variable has a wide range compared to other variables. We think a log transformation on *HIV/AIDS* might be helpful in condensing its range.

Full Model

To answer our research questions, we first fitted a full model with a selection of 17 explanatory variables that are free from data coding errors from our dataset. The summary output of the full model is shown in Figure 7 and the corresponding residual vs fitted value plot is shown in Figure 8.



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

Residuals:
     Min       1Q   Median       3Q      Max 
-17.0577  -2.1311   0.0486   2.4286  11.6567 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      5.357e+01  8.508e-01  62.969  < 2e-16 ***
StatusDeveloping                -1.001e+00  3.459e-01  -2.892 0.003877 ** 
Adult.Mortality                 -1.769e-02  9.647e-04 -18.332  < 2e-16 ***
Alcohol                         -1.377e-01  3.349e-02  -4.113 4.10e-05 ***
Hepatitis.B                     -7.782e-03  4.547e-03  -1.711 0.087205 .  
Measles                          1.792e-05  1.062e-05   1.687 0.091717 .  
BMI                              3.606e-02  6.136e-03   5.877 5.05e-09 ***
under.five.deaths               -3.162e-03  9.146e-04  -3.458 0.000559 ***
Polio                            9.868e-03  5.266e-03   1.874 0.061114 .  
Total.expenditure                7.418e-02  4.157e-02   1.785 0.074523 .  
Diphtheria                       2.036e-02  6.038e-03   3.372 0.000764 ***
HIV.AIDS                        -4.386e-01  1.827e-02 -24.013  < 2e-16 ***
GDP                              6.216e-05  9.597e-06   6.477 1.24e-10 ***
Population                       3.149e-09  1.736e-09   1.814 0.069914 .  
thinness..1.19.years            -2.331e-02  5.425e-02  -0.430 0.667509    
thinness.5.9.years              -1.133e-02  5.340e-02  -0.212 0.832015    
Income.composition.of.resources  1.030e+01  8.488e-01  12.129  < 2e-16 ***
Schooling                        8.818e-01  6.067e-02  14.536  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.674 on 1631 degrees of freedom
Multiple R-squared:  0.8274,	Adjusted R-squared:  0.8256 
F-statistic: 459.9 on 17 and 1631 DF,  p-value: < 2.2e-16
```

Figure 8. Summary Statistics of the Full Model

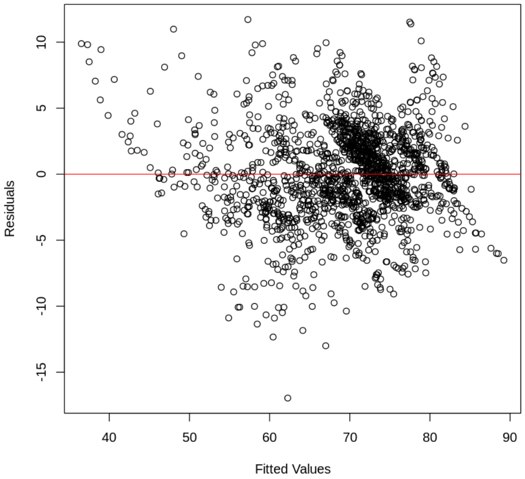


Figure 9. Residual vs. Fitted Values Plot for the Full Model

The above summary statistics of our full model suggest that 11 out of the 17 variables have p-values lower than the 0.05 alpha level. The residual plots somewhat resemble a funnel shape which suggests to us that the residuals might not be randomly distributed. The results from the full model suggested the following actions to take:

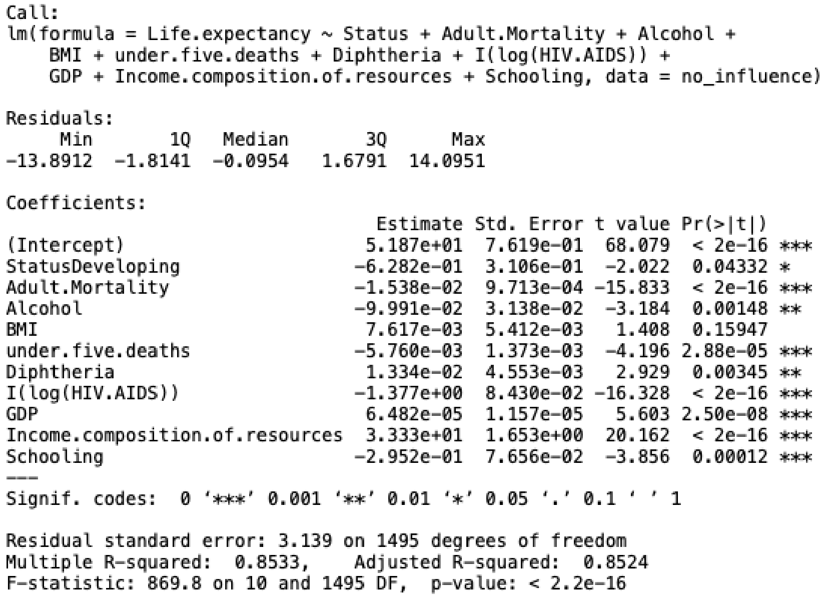
- Conduct an influence analysis to identify the influential cases because we have seen many outliers in our predictor variables.
- Filter out the high influence cases from the dataset
- Apply a log transformation to the HIV.AIDS variable
- Fit a reduced model with the remaining 11 explanatory variables to the filtered data from step 2.

The influence analysis was performed in the following steps in R (see Appendix for R codes):

- Construct a design matrix X with the 11 significant explanatory variables
- Computed the “hat” matrix and extracted the diagonal entries from the “hat” matrix
- Filtered the data set to include only data points with a moderate level of leverage

Reduced Model

After removing high influential cases, dropping non-significant predictors from the full model, and applying log transformation to the HIV.AIDS variable, the resulting reduced model with 10 variables is shown below.



```
Call:
lm(formula = Life.expectancy ~ Status + Adult.Mortality + Alcohol +
    BMI + under.five.deaths + Diphtheria + I(log(HIV.AIDS)) +
    GDP + Income.composition.of.resources + Schooling, data = no_influence)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8912  -1.8141  -0.0954   1.6791  14.0951

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.187e+01  7.619e-01  68.079  < 2e-16 ***
StatusDeveloping                -6.282e-01  3.106e-01  -2.022  0.04332 *
Adult.Mortality                 -1.538e-02  9.713e-04 -15.833  < 2e-16 ***
Alcohol                         -9.991e-02  3.138e-02  -3.184  0.00148 **
BMI                              7.617e-03  5.412e-03   1.408  0.15947
under.five.deaths               -5.760e-03  1.373e-03  -4.196 2.88e-05 ***
Diphtheria                       1.334e-02  4.553e-03   2.929  0.00345 **
I(log(HIV.AIDS))                -1.377e+00  8.430e-02 -16.328  < 2e-16 ***
GDP                              6.482e-05  1.157e-05   5.603 2.50e-08 ***
Income.composition.of.resources  3.333e+01  1.653e+00  20.162  < 2e-16 ***
Schooling                       -2.952e-01  7.656e-02  -3.856  0.00012 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.139 on 1495 degrees of freedom
Multiple R-squared:  0.8533,    Adjusted R-squared:  0.8524
F-statistic: 869.8 on 10 and 1495 DF,  p-value: < 2.2e-16
```

Figure 10. Summary Statistics of the Reduced Model

In this reduced model, the predictor BMI is no longer significant. The adjusted R-squared increased to 0.8524. The reduced model performs exceptionally well because it accounts for a higher percentage of variation in the response variable than the full model (R^2 = 0.8256) while using fewer variables and fewer data points.

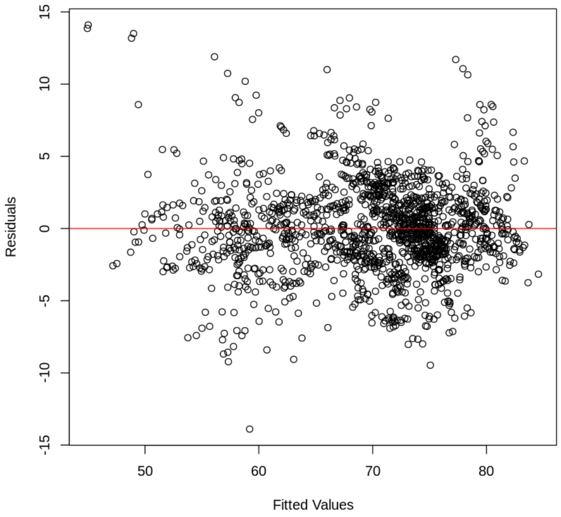
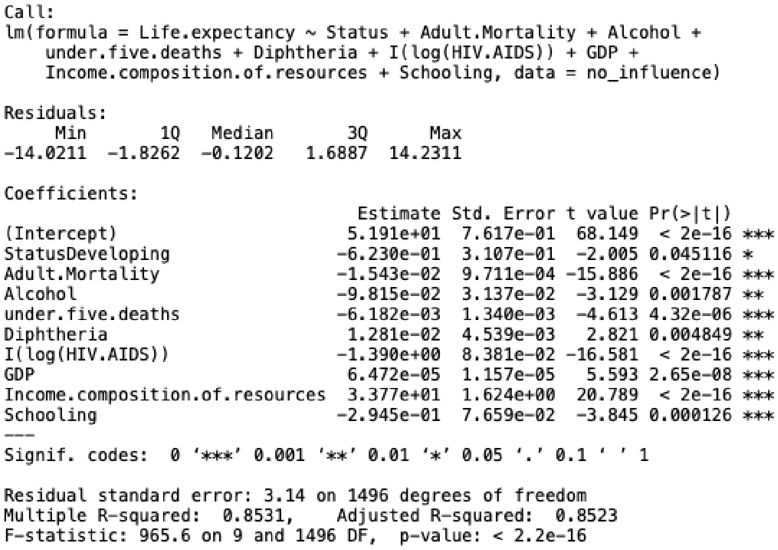


Figure 11. Residual vs. Fitted Values Plot for the Full Model

The residual plot of the reduced model shows an appreciable level of improvement compared to the full model. The residuals are more randomly distributed and devoid of pattern.

Final Model

In the final model, we further simplified our model by dropping the non-significant predictor variable, BMI. The resulting model is composed of 9 predictor variables. The summary result from R is displayed below:



```
Call:
lm(formula = Life.expectancy ~ Status + Adult.Mortality + Alcohol +
    under.five.deaths + Diphtheria + I(log(HIV.AIDS)) + GDP +
    Income.composition.of.resources + Schooling, data = no_influence)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0211  -1.8262  -0.1202   1.6887  14.2311

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.191e+01  7.617e-01  68.149  < 2e-16 ***
StatusDeveloping                -6.230e-01  3.107e-01  -2.005 0.045116 *
Adult.Mortality                 -1.543e-02  9.711e-04 -15.886  < 2e-16 ***
Alcohol                         -9.815e-02  3.137e-02  -3.129 0.001787 **
under.five.deaths               -6.182e-03  1.340e-03  -4.613 4.32e-06 ***
Diphtheria                       1.281e-02  4.539e-03   2.821 0.004849 **
I(log(HIV.AIDS))                -1.390e+00  8.381e-02 -16.581  < 2e-16 ***
GDP                              6.472e-05  1.157e-05   5.593 2.65e-08 ***
Income.composition.of.resources  3.377e+01  1.624e+00  20.789  < 2e-16 ***
Schooling                       -2.945e-01  7.659e-02  -3.845 0.000126 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.14 on 1496 degrees of freedom
Multiple R-squared:  0.8531,	Adjusted R-squared:  0.8523
F-statistic: 965.6 on 9 and 1496 DF,  p-value: < 2.2e-16
```

Figure 12. Summary Statistics of the Final Model

The final model is an improvement over the reduced model due to the reduction in the number of predictors used. The final model is able to account for 85.23% of the variation in the response variable by using 9 predictors compared to the reduced model which accounts for 85.24% of the variation in response using 10 predictors. We think the reduction in the number of predictors outweighs the marginal decrease in adjusted R^2.

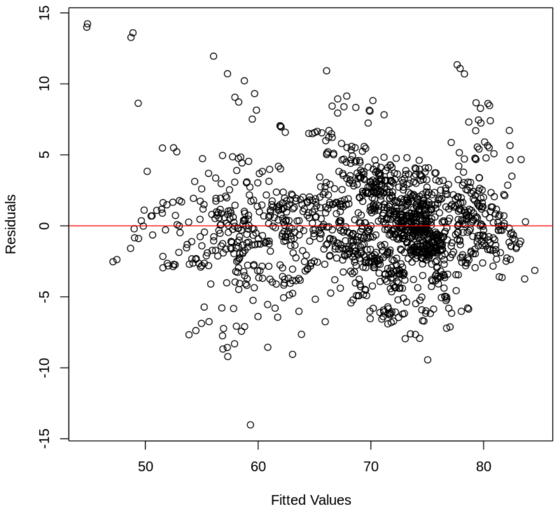


Figure 13. Summary Statistics of the Final Model

The residual plot for the final model is much similar to the one from the reduced model. The residuals of the final model are reasonably scattered with no observable patterns with only a few outliers.

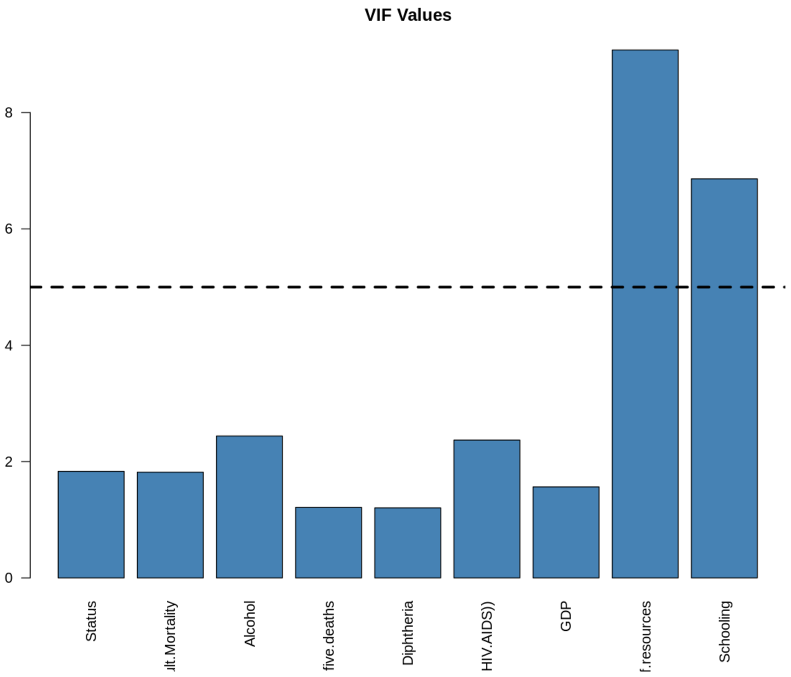


Figure 14. VIF Values for Each Predictors in Final Model

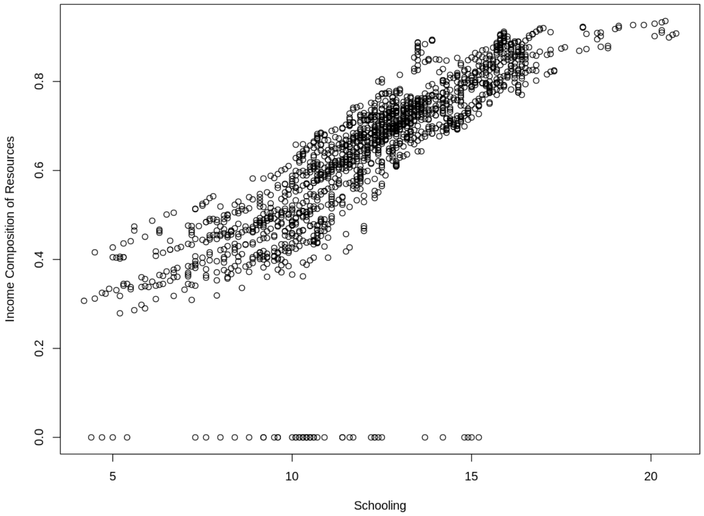


Figure 15. Scatter Plot of Income Composition of Resources Against Schooling

In our final model, we observed that Schooling has a negative coefficient even though it is positively correlated with Life Expectancy in Figure 5. Therefore we computed VIF values for each predictor in our final model as we suspect Schooling might be correlated with another variable. We use VIF > 5 (the dotted line in Figure 13) as a cutoff for collinearity. We noted that both Percentage of Resources Composition and Schooling has a VIF value greatly exceeding this threshold. Indeed, we found a strong positive linear association between the two variables in Figure 14.

Discussion

Our final least squares regression model is :

$$Y = 51.91 - 0.6230z_1 - 0.01543x_1 - 0.09815x_2 - 0.006182x_3 + 0.01281x_4 - 1.390\, x_5 - 0.00006472x_6 + 33.77x_7 - 0.02945x_8 + \epsilon$$

Where

- Y is the life expectancy in years,
- z_1 is the development status of a country (1 for developing, 0 for developed),
- x_1 is number of adult death per 1000 population,
- x_2 is the number of liters of alcohol consumed per year,
- x_3 is the number of children death under the age of 5 per 1000 population,
- x_4 is the percentage of population immunized against diphtheria tetanus toxoid and pertussis (DTP3) among 1-year-olds,
- x_5 is the log of the number of death due to HIV/AIDS per 1000 population,
- x_6 is the gross domestic production in USD,
- x_7 is the income composition of resources in percentage which measures the efficiency of the utilization of natural resources ,
- x_8 is the number of years of schooling ,

Our final model suggest two results:

1. There is a baseline difference in life expectancy between developed and developing countries.
2. Life expectancy depends linearly on the aforementioned socio-economic and health-related factors.

One surprising result from our final model is that number of years of schooling (x_8) has a negative coefficient which suggests that as the average number of years of schooling increases, the predicted life expectancy will decrease. This is likely caused by multicollinearity with the predictor variable income composition of resources. We first check our predictor variables for multicollinearity by computing their variance inflation factor (VIF). We find that both schooling and utilization of resources have VIF exceeds our chosen cutoff and the scatter plot between the two suggests a strong positive correlation. Collinearity can cause overfitting and make certain predictor variables have opposite relationships of what is expected.

We see that the other predictor variables have expected relationships with life expectancy. The predicted life expectancy for developing countries when all other variables are held constant is about 0.623 years lower than the developed countries (baseline). The largest positive effect on life expectancy is caused by income composition of resources (x_7) that for every one percent increase in a country's efficiency in utilizing its natural resources, the life expectancy increases by 33.77 years. Whereas the largest negative effect is caused by deaths due to HIV/AIDS per 1000 people(x_5). We found that for every one unit increase in log(HIV/AIDS) the life expectancy is expected to decrease by 1.39 years.

There are several limitations to our project. In the process of removing influential data points to better fit the model, we reduced our data set from n= 1649 to n = 1506 (minus 143 data points or 8.67% of the original data). We believe this is a noticeable reduction in sample size and a trade-off for removing influential cases. Another limitation is that many of the predictor variables removed during the backward selection were marginally significant. We think this hinted at the probability of making type 2 errors during our model selection process. We might be missing out on some predictors that in fact have a significant relationship with the response variable. The final limitation of our project is that the variable Schooling and Income Composition of Resources variables are positively correlated. This resulted in the variable Schooling having a negative coefficient which is opposite to our expectation. However, we think both variables are unique measures of facets of a country's socio-economic status. For this reason, we are keeping both variables in our model. In the future, we could be applying corrections to address the correlation between the two variables.

Conclusion

In this project, we investigated the relationship between human life expectancy and a selection of socio-economic factors. In our final model, we have found that there exists a baseline difference in life expectancy between developing and developed countries. Furthermore, the best-fitted final model indicates a significant linear relationship between life expectancy and the other eight continuous predictors (Adult.Mortality, Alcohol, under.five.deaths, Diphtheria, log(HIV.AIDS), GDP, Income.composition.of.resources, Schooling). The final 9-predictor regression model was able to account for 85.23 percent of the variation in Life expectancy. For future research, it is suggested to improve the data quality and data size to study the marginal significant predictors.

Appendix

August 13, 2022

```
[1]: library(tidyverse)
     library(GGally)
     library(car)
```

```
Warning message in system("timedatectl", intern = TRUE):
"running command 'timedatectl' had status 1"
  Attaching packages                       tidyverse
1.3.1

  ggplot2 3.3.5      purrr   0.3.4
  tibble  3.1.6      dplyr   1.0.7
  tidyr   1.1.3      stringr 1.4.0
  readr   2.1.1      forcats 0.5.1

  Conflicts
tidyverse_conflicts()
  dplyr::filter() masks stats::filter()
  dplyr::lag()    masks stats::lag()

Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2

Loading required package: carData


Attaching package: ‘car’


The following object is masked from ‘package:dplyr’:

    recode


The following object is masked from ‘package:purrr’:

    some
```

[2]:

```
le_data <- read.csv("~/STAT 306/Life Expectancy Data.csv")
le_data$Status <- as.factor(le_data$Status)
```

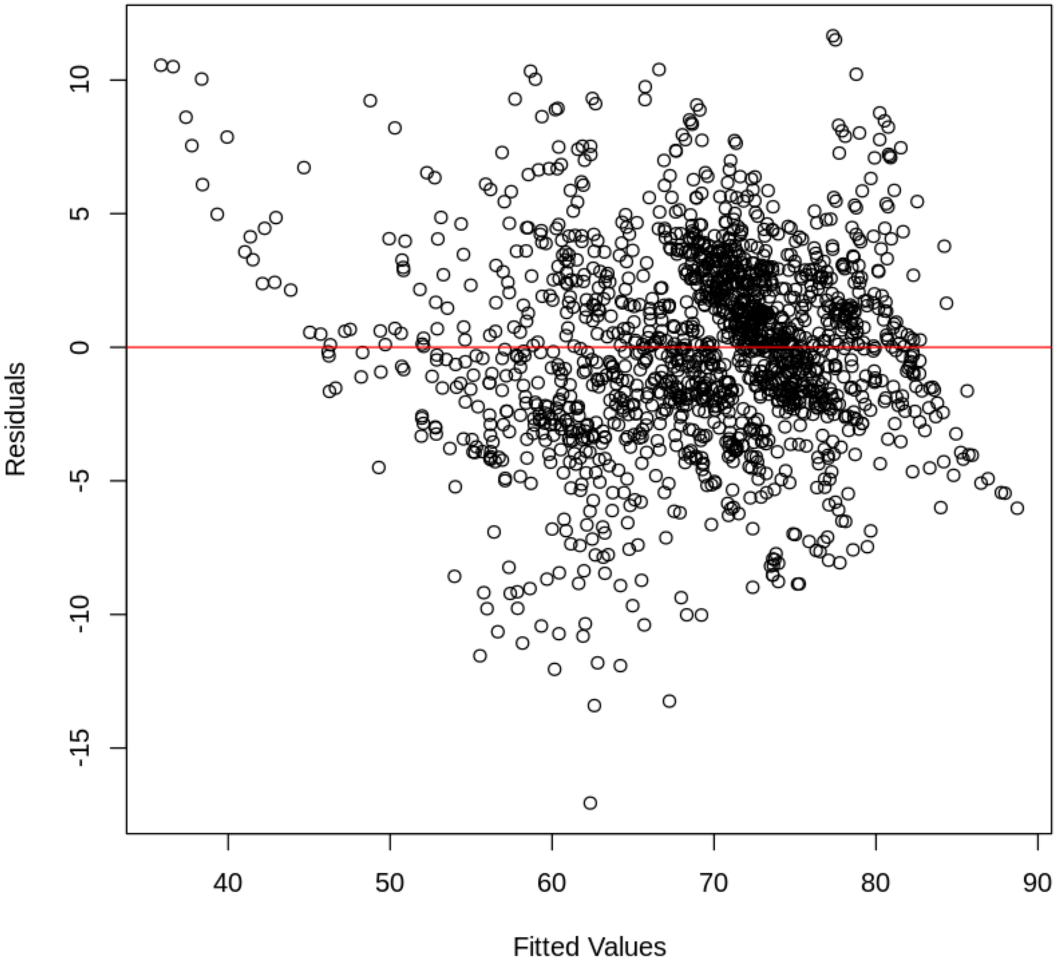
[3]:

```
le_reduced <- le_data %>%
    select(c(-Country, -Year, -percentage.expenditure, -infant.deaths)) %>%
    na.omit()
full_model <- lm(Life.expectancy ~. , data = le_reduced)
summary(full_model)
dim(le_reduced)
```

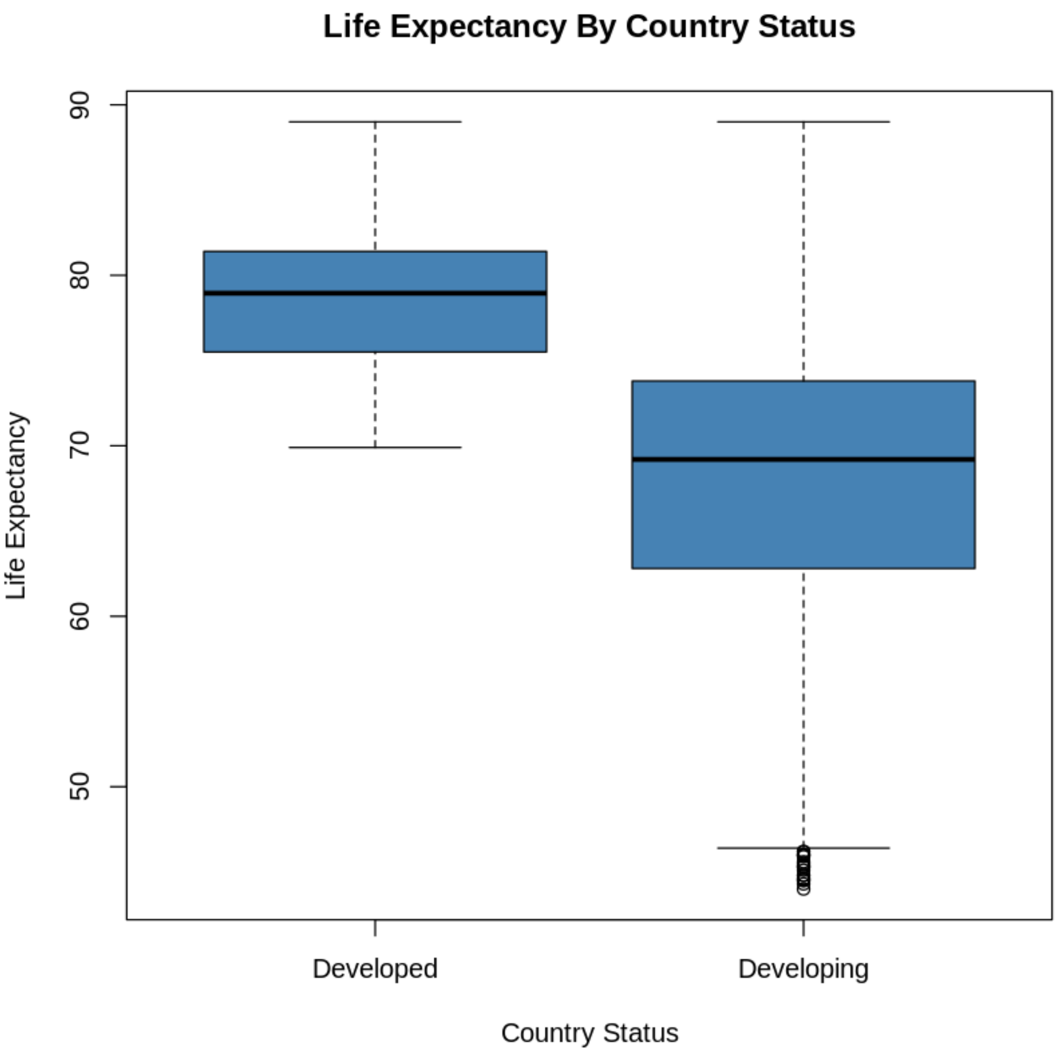
```
Call:
lm(formula = Life.expectancy ~ ., data = le_reduced)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.0577  -2.1311   0.0486   2.4286  11.6567 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      5.357e+01  8.508e-01  62.969  < 2e-16 ***
StatusDeveloping                -1.001e+00  3.459e-01  -2.892 0.003877 ** 
Adult.Mortality                 -1.769e-02  9.647e-04 -18.332  < 2e-16 ***
Alcohol                         -1.377e-01  3.349e-02  -4.113 4.10e-05 ***
Hepatitis.B                     -7.782e-03  4.547e-03  -1.711 0.087205 .  
Measles                          1.792e-05  1.062e-05   1.687 0.091717 .  
BMI                              3.606e-02  6.136e-03   5.877 5.05e-09 ***
under.five.deaths               -3.162e-03  9.146e-04  -3.458 0.000559 ***
Polio                            9.868e-03  5.266e-03   1.874 0.061114 .  
Total.expenditure                7.418e-02  4.157e-02   1.785 0.074523 .  
Diphtheria                       2.036e-02  6.038e-03   3.372 0.000764 ***
HIV.AIDS                        -4.386e-01  1.827e-02 -24.013  < 2e-16 ***
GDP                              6.216e-05  9.597e-06   6.477 1.24e-10 ***
Population                       3.149e-09  1.736e-09   1.814 0.069914 .  
thinness..1.19.years            -2.331e-02  5.425e-02  -0.430 0.667509    
thinness.5.9.years              -1.133e-02  5.340e-02  -0.212 0.832015    
Income.composition.of.resources  1.030e+01  8.488e-01  12.129  < 2e-16 ***
Schooling                        8.818e-01  6.067e-02  14.536  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.674 on 1631 degrees of freedom
Multiple R-squared:  0.8274,	Adjusted R-squared:  0.8256 
F-statistic: 459.9 on 17 and 1631 DF,  p-value: < 2.2e-16
```

1. 1649 2. 18

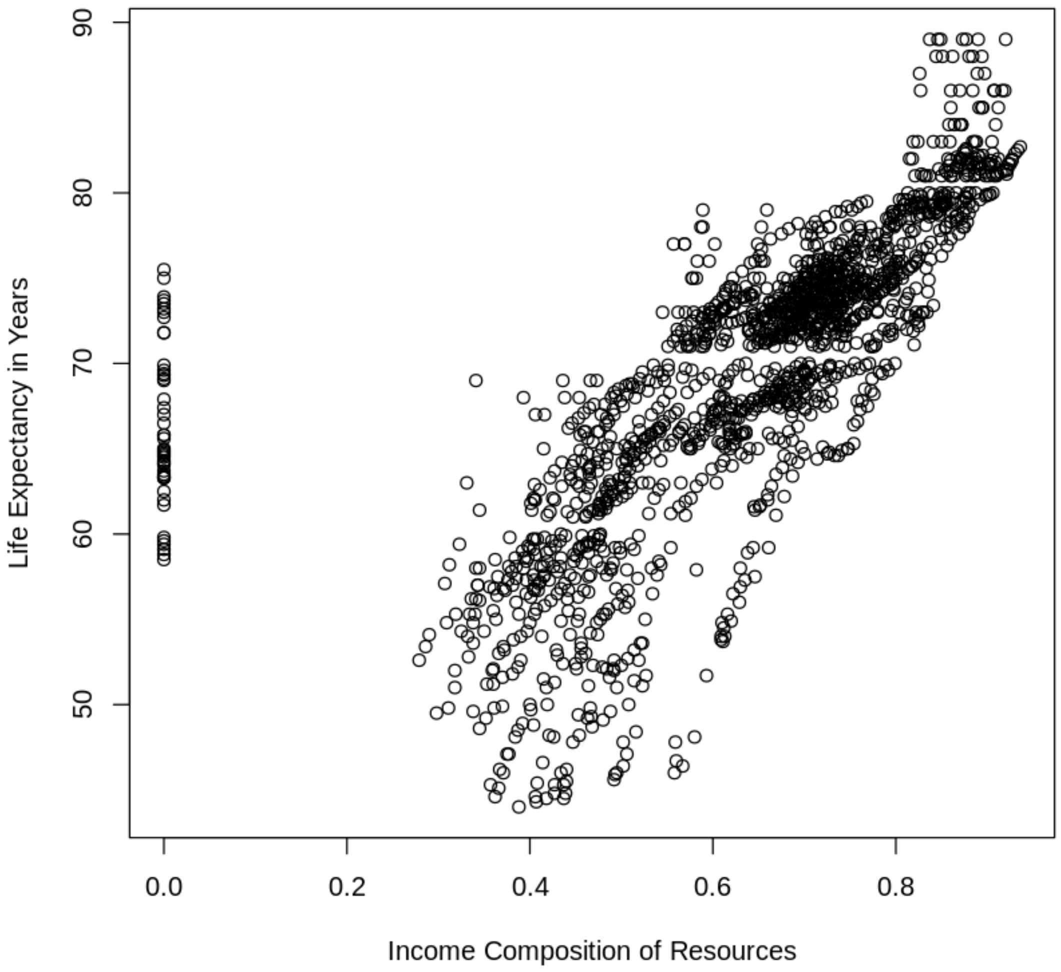
```
[4]: plot(x = full_model$fitted.values, y = full_model$residuals, xlab = "Fitted␣
     ↪Values", ylab = "Residuals")
     abline(h=0, col="red")
```



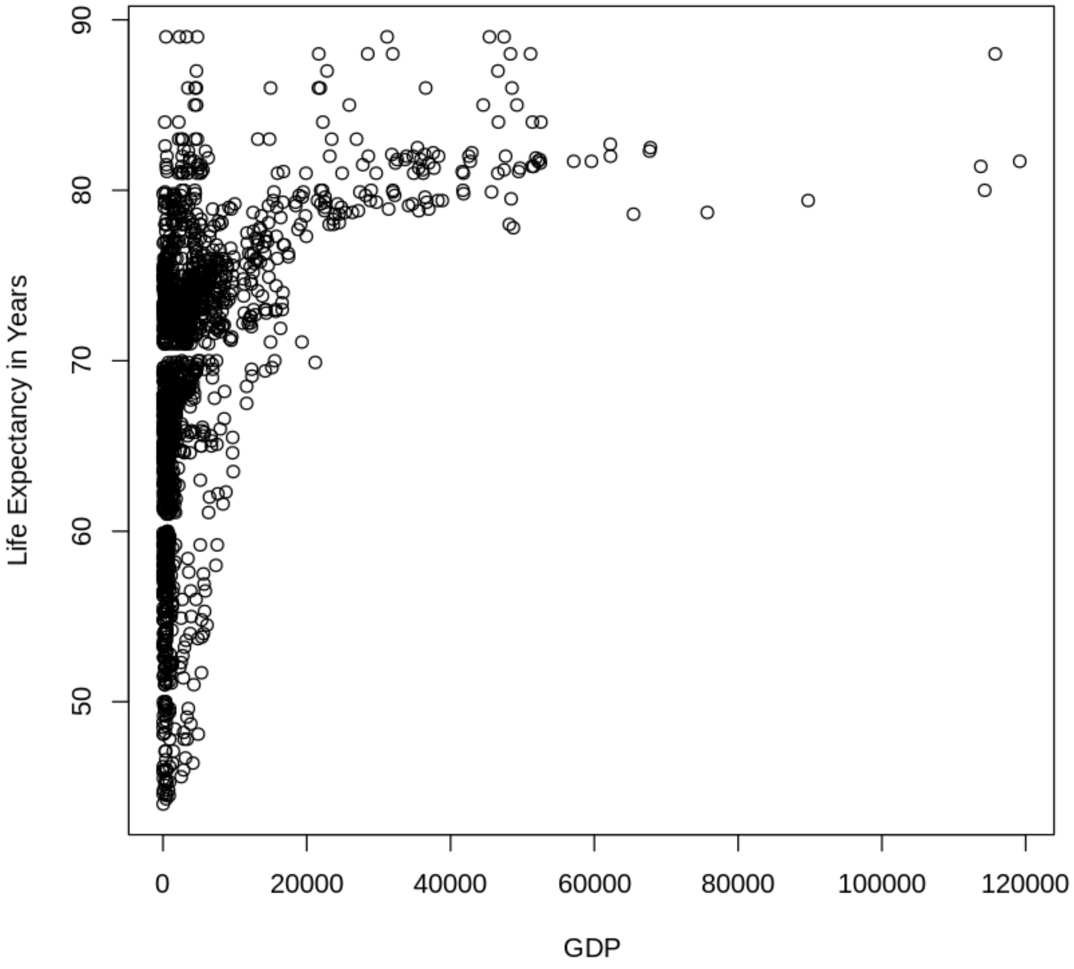
```
[5]: boxplot(le_reduced$Life.expectancy ~ le_reduced$Status,
             col='steelblue',
             main='Life Expectancy By Country Status',
             xlab='Country Status',
             ylab='Life Expectancy')
```



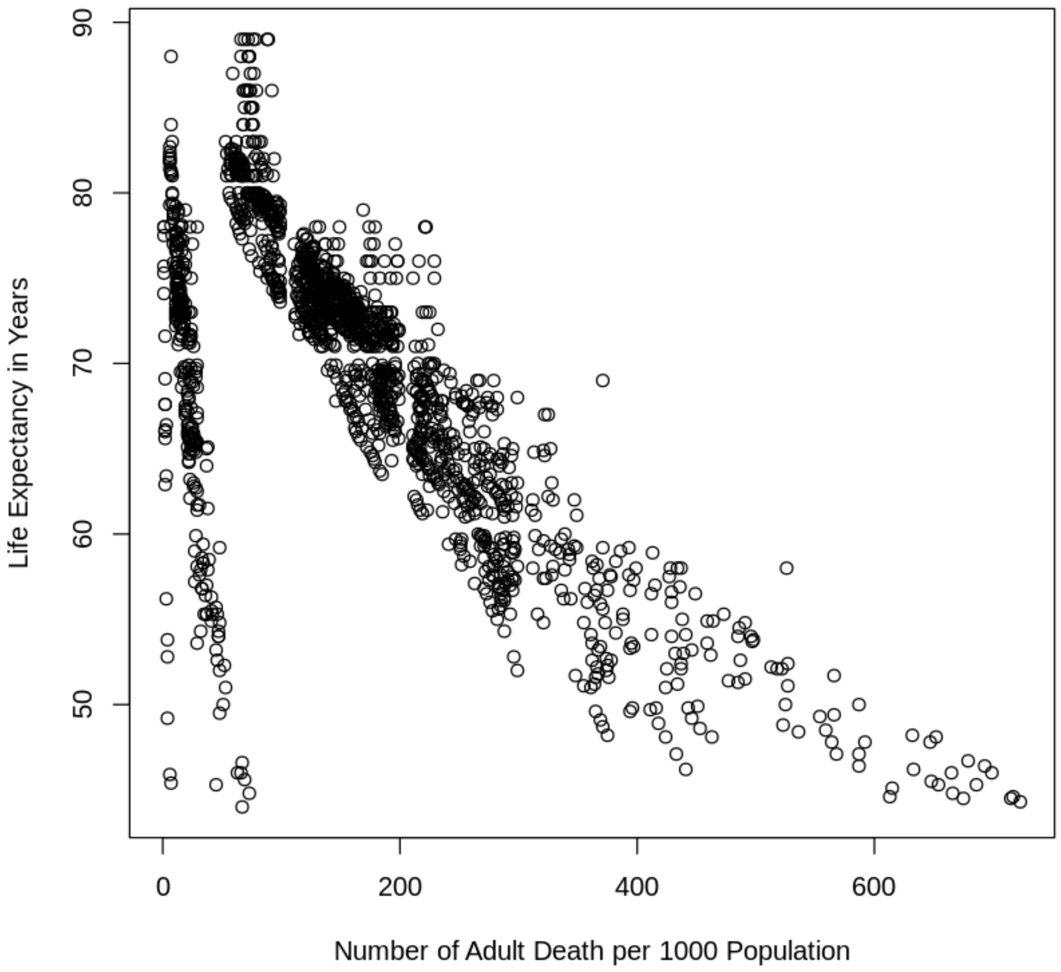
```
[6]: plot(x = le_reduced$Income.composition.of.resources, y = le_reduced$Life.
     ↪expectancy,
         xlab = "Income Composition of Resources",
         ylab = "Life Expectancy in Years")
```



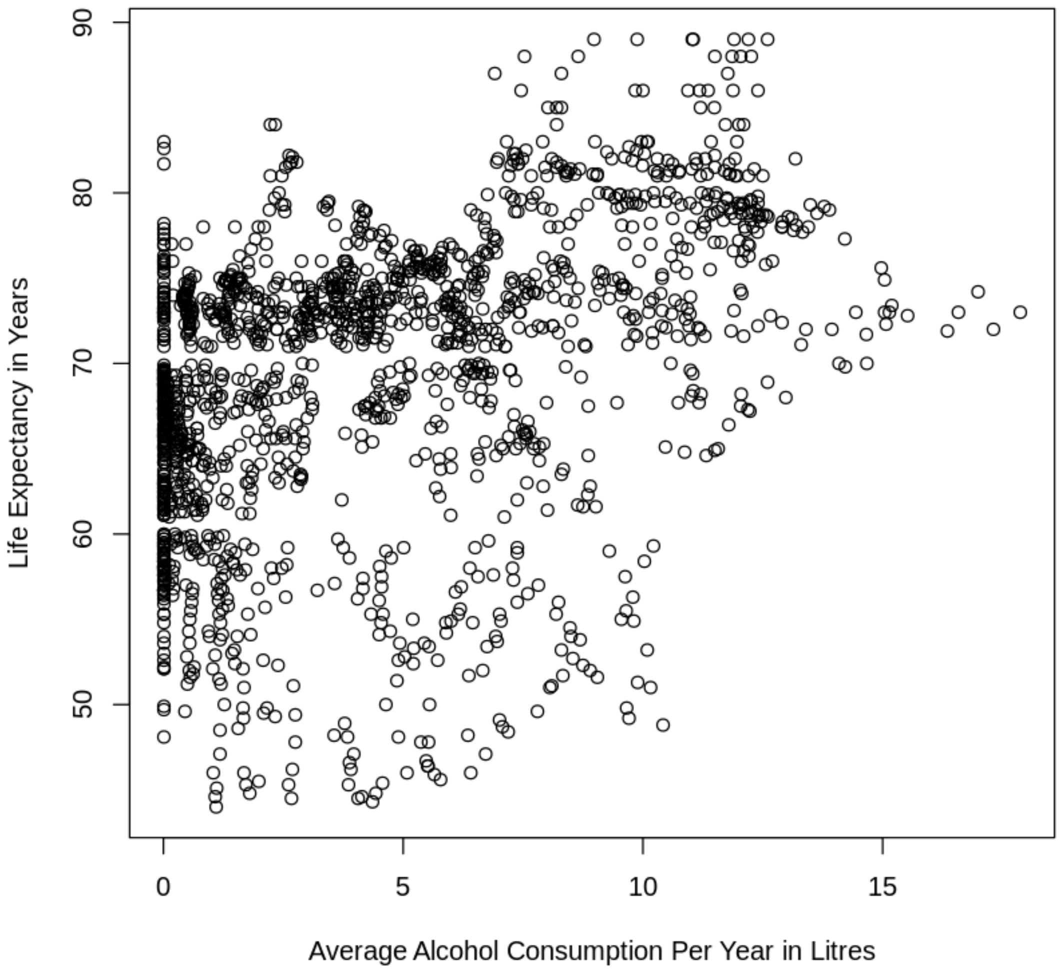
```
[7]: plot(x = le_reduced$GDP, y = le_reduced$Life.expectancy,
         xlab = "GDP",
         ylab = "Life Expectancy in Years")
```



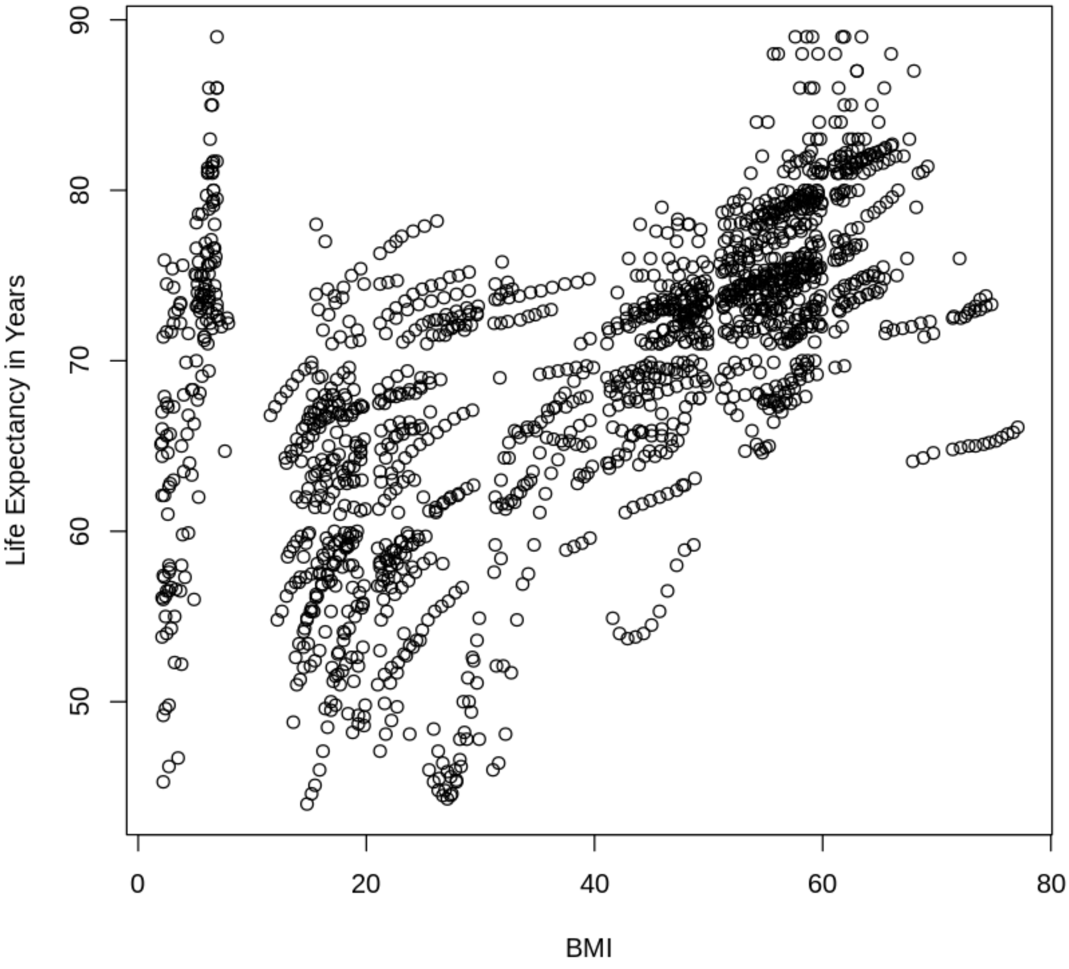
[8]:
```
plot(x = le_reduced$Adult.Mortality, y = le_reduced$Life.expectancy,
    xlab = "Number of Adult Death per 1000 Population",
    ylab = "Life Expectancy in Years")
```



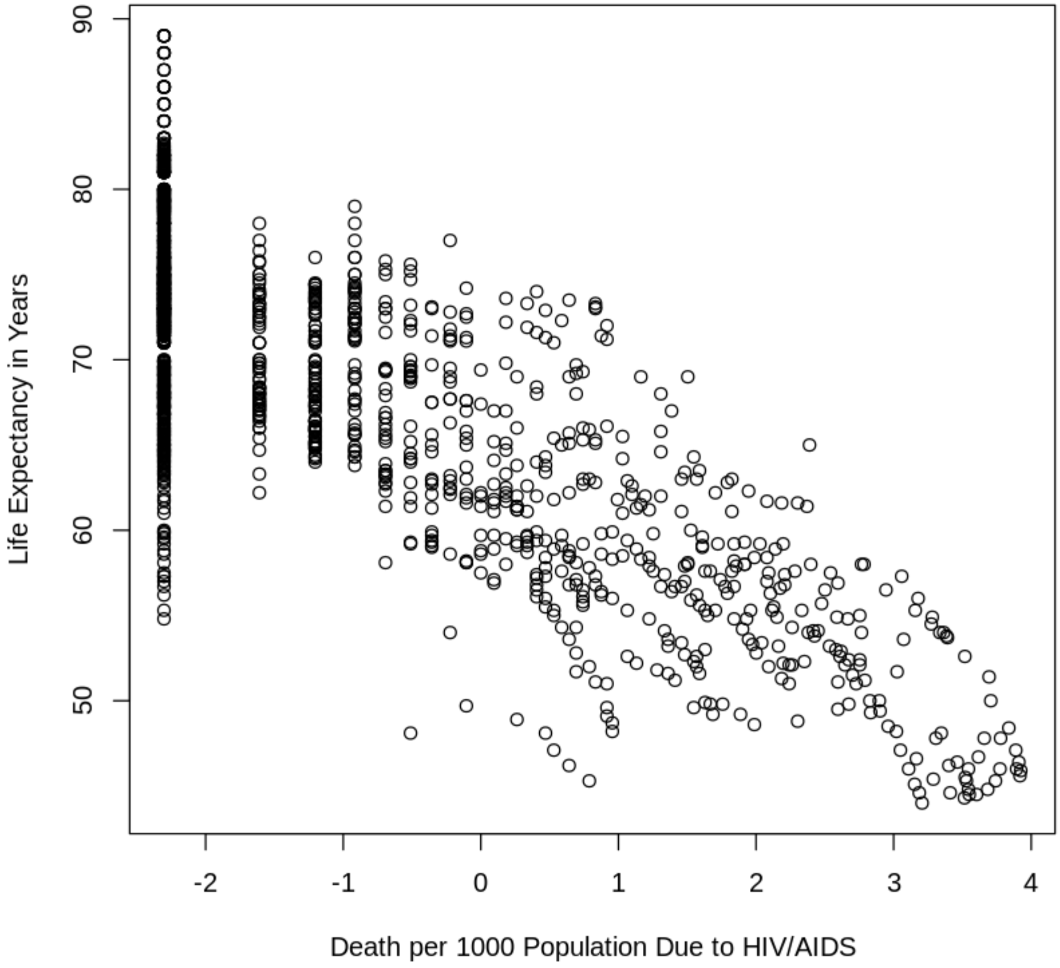
```
[9]: plot(x = le_reduced$Alcohol, y = le_reduced$Life.expectancy,
         xlab = "Average Alcohol Consumption Per Year in Litres",
         ylab = "Life Expectancy in Years")
```



[10]:
```
plot(x = le_reduced$BMI, y = le_reduced$Life.expectancy,
    xlab = "BMI",
    ylab = "Life Expectancy in Years")
```

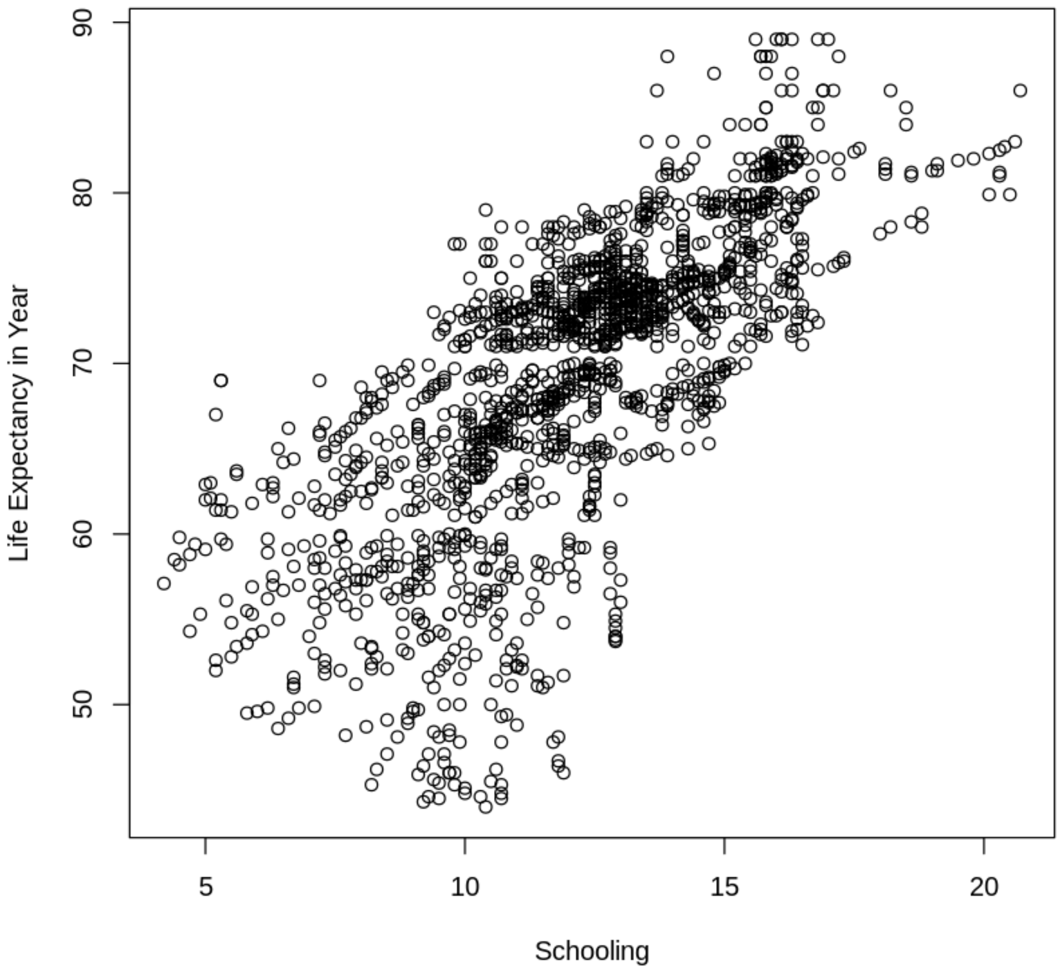


```
[11]: plot(x = log(le_reduced$HIV.AIDS), y = le_reduced$Life.expectancy,
          xlab = "Death per 1000 Population Due to HIV/AIDS",
          ylab = "Life Expectancy in Years")
```



[12]:

```
plot(x = le_reduced$Schooling, y = le_reduced$Life.expectancy,
    xlab = "Schooling",
    ylab = "Life Expectancy in Year")
```



[13]:

```
le_reduced$intercept <- c(rep(1,1649))
X <- le_reduced %>%
    select(intercept, Status, Adult.Mortality, Alcohol,
           BMI, under.five.deaths, Diphtheria, HIV.AIDS, GDP
           ,Income.composition.of.resources , Schooling) %>%
    mutate(Status = as.numeric(Status))

X = data.matrix(X)
dim(X)
```

1. 1649 2. 11

[14]:

```
Xt = t(X)
inv_XtX = solve(Xt %*% X)
P = X %*% inv_XtX %*% Xt
leverage = diag(P)
length(leverage)
```

1649

[15]:

```
le_reduced$leverage <- leverage
cut_off <- 2*11/length(leverage)
dim(le_reduced)
```

1. 1649 2. 20

[16]:

```
no_influence <- le_reduced %>%
    filter(leverage < cut_off)
dim(no_influence)
```

1. 1506 2. 20

[17]:

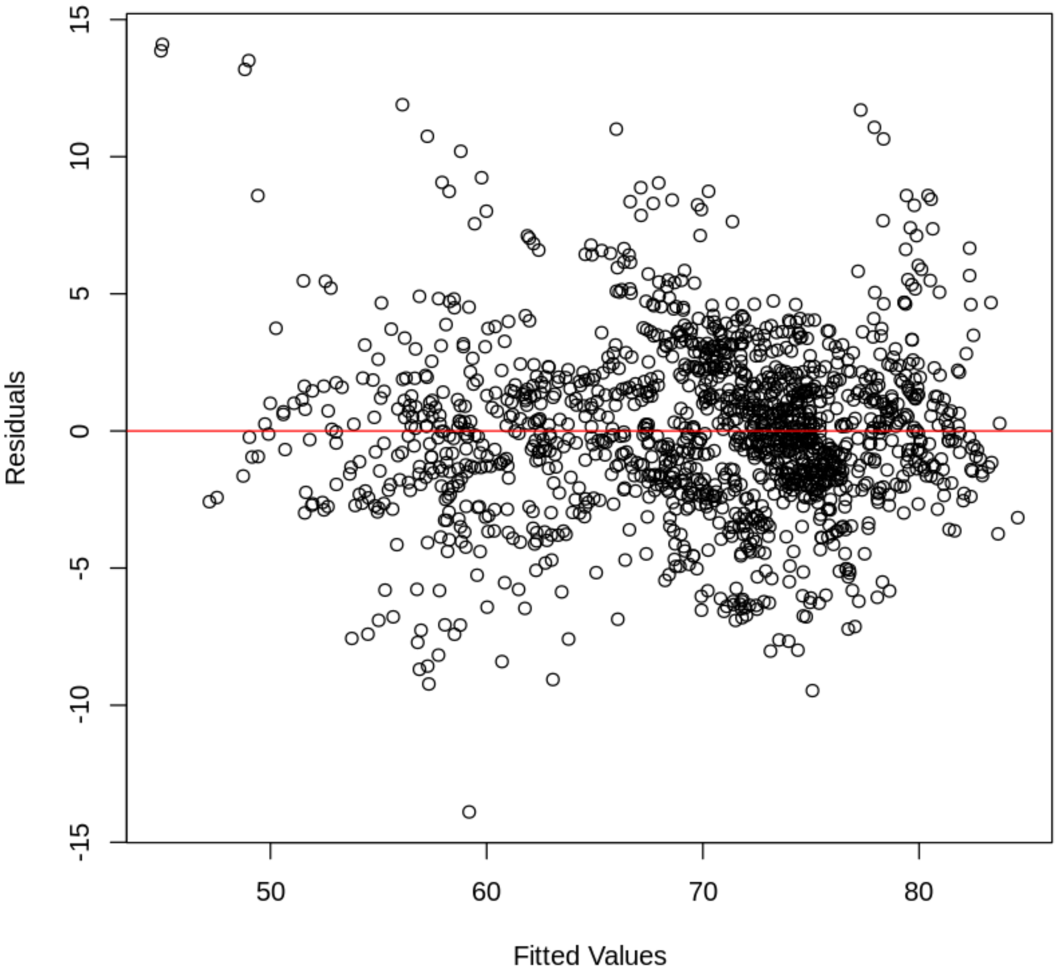
```
reduced_model_1 <- lm(Life.expectancy ~ Status + Adult.Mortality + Alcohol +
           BMI + under.five.deaths + Diphtheria + I(log(HIV.AIDS)) + GDP
           + Income.composition.of.resources + Schooling,
                      data = no_influence)
summary(reduced_model_1)
```

```
Call:
lm(formula = Life.expectancy ~ Status + Adult.Mortality + Alcohol +
    BMI + under.five.deaths + Diphtheria + I(log(HIV.AIDS)) +
    GDP + Income.composition.of.resources + Schooling, data = no_influence)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8912  -1.8141  -0.0954   1.6791  14.0951

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.187e+01  7.619e-01  68.079  < 2e-16 ***
StatusDeveloping                -6.282e-01  3.106e-01  -2.022  0.04332 *
Adult.Mortality                 -1.538e-02  9.713e-04 -15.833  < 2e-16 ***
Alcohol                         -9.991e-02  3.138e-02  -3.184  0.00148 **
BMI                              7.617e-03  5.412e-03   1.408  0.15947
under.five.deaths               -5.760e-03  1.373e-03  -4.196 2.88e-05 ***
Diphtheria                       1.334e-02  4.553e-03   2.929  0.00345 **
I(log(HIV.AIDS))                -1.377e+00  8.430e-02 -16.328  < 2e-16 ***
GDP                              6.482e-05  1.157e-05   5.603 2.50e-08 ***
Income.composition.of.resources  3.333e+01  1.653e+00  20.162  < 2e-16 ***
Schooling                       -2.952e-01  7.656e-02  -3.856  0.00012 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.139 on 1495 degrees of freedom
Multiple R-squared:  0.8533,	Adjusted R-squared:  0.8524
F-statistic: 869.8 on 10 and 1495 DF,  p-value: < 2.2e-16
```

[18]:

```
plot(x = reduced_model_1$fitted.values, y = reduced_model_1$residuals, xlab =␣
 ↪"Fitted Values", ylab = "Residuals")
abline(h=0, col="red")
```



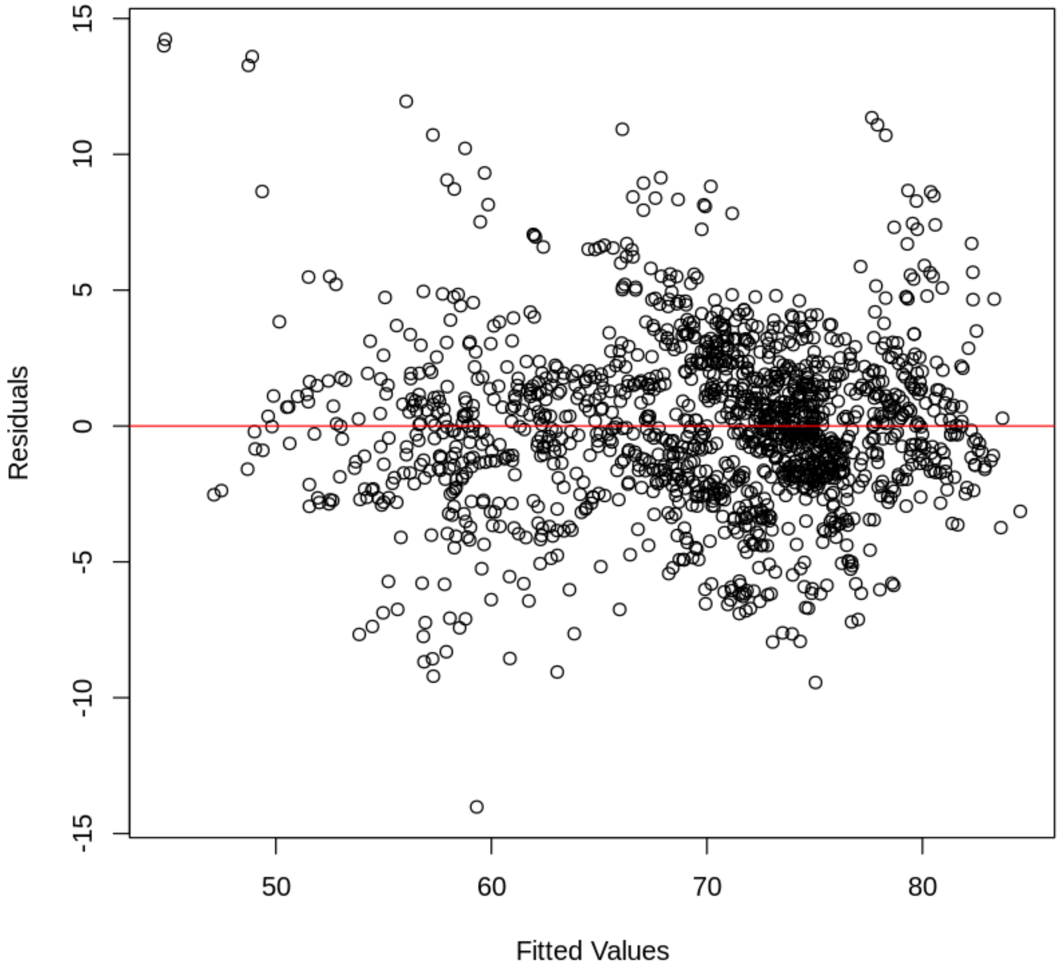
[19]:

```
reduced_model_2 <- lm(Life.expectancy ~ Status + Adult.Mortality + Alcohol +
           under.five.deaths + Diphtheria + I(log(HIV.AIDS)) + GDP
           + Income.composition.of.resources + Schooling,
                      data = no_influence)
summary(reduced_model_2)
```

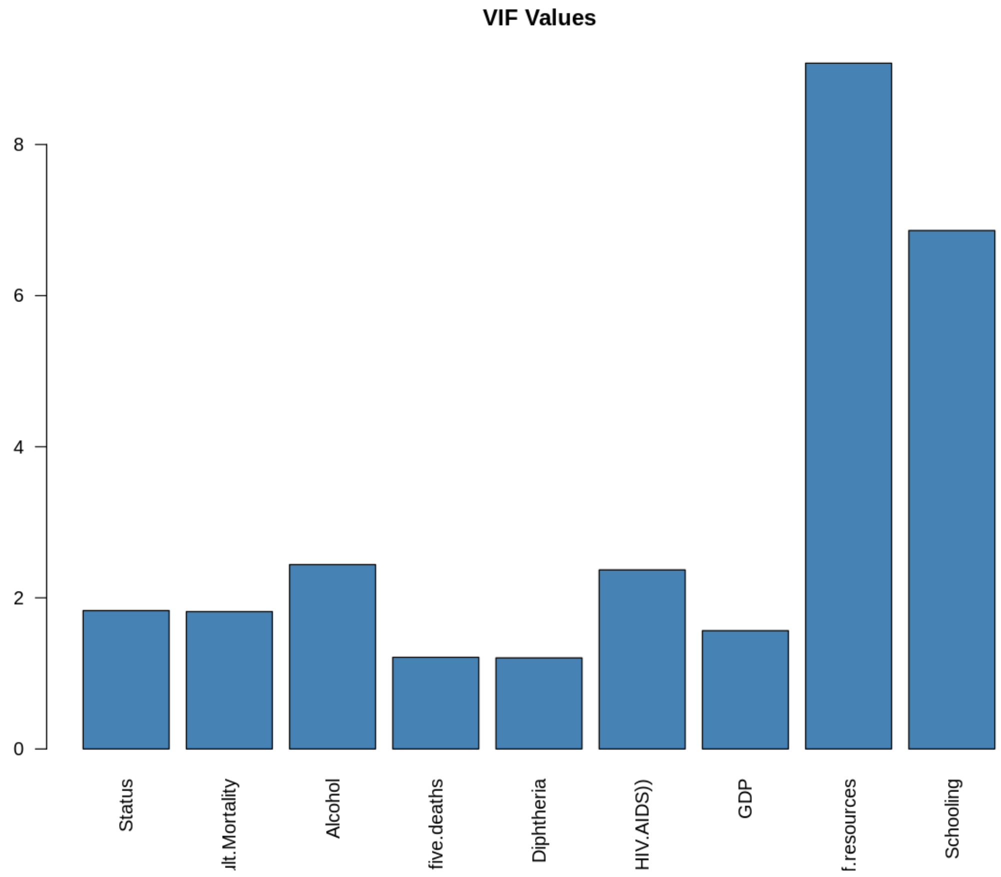
```
Call:
lm(formula = Life.expectancy ~ Status + Adult.Mortality + Alcohol +
    under.five.deaths + Diphtheria + I(log(HIV.AIDS)) + GDP +
    Income.composition.of.resources + Schooling, data = no_influence)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0211  -1.8262  -0.1202   1.6887  14.2311

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.191e+01  7.617e-01  68.149  < 2e-16 ***
StatusDeveloping                -6.230e-01  3.107e-01  -2.005 0.045116 *
Adult.Mortality                 -1.543e-02  9.711e-04 -15.886  < 2e-16 ***
Alcohol                         -9.815e-02  3.137e-02  -3.129 0.001787 **
under.five.deaths               -6.182e-03  1.340e-03  -4.613 4.32e-06 ***
Diphtheria                       1.281e-02  4.539e-03   2.821 0.004849 **
I(log(HIV.AIDS))                -1.390e+00  8.381e-02 -16.581  < 2e-16 ***
GDP                              6.472e-05  1.157e-05   5.593 2.65e-08 ***
Income.composition.of.resources  3.377e+01  1.624e+00  20.789  < 2e-16 ***
Schooling                       -2.945e-01  7.659e-02  -3.845 0.000126 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.14 on 1496 degrees of freedom
Multiple R-squared:  0.8531,	Adjusted R-squared:  0.8523
F-statistic: 965.6 on 9 and 1496 DF,  p-value: < 2.2e-16
```

[20]:

```
plot(x = reduced_model_2$fitted.values, y = reduced_model_2$residuals, xlab =␣
 ↪"Fitted Values", ylab = "Residuals")
abline(h=0, col="red")
```

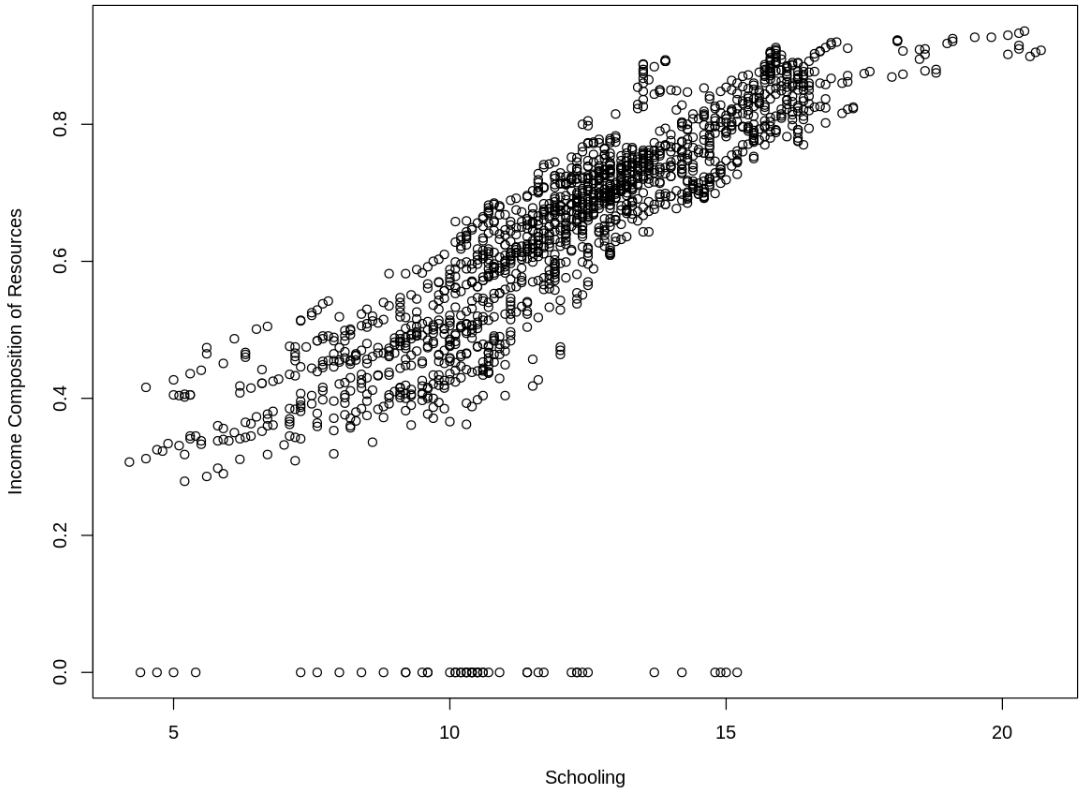


[21]:
```
options(repr.plot.width = 10, repr.plot.height = 8)
vif_values <- vif(reduced_model_2)
barplot(vif_values, main = "VIF Values", horiz = FALSE, col = "steelblue", las␣
↪= 2, cex.names = 1)
abline(h = 10, lwd = 3, lty = 2)
```



```
[22]: plot(x = le_reduced$Schooling, y = le_reduced$Income.composition.of.resources,
          xlab = "Schooling",
          ylab = "Income Composition of Resources")
      cor(x = le_reduced$Schooling, y = le_reduced$Income.composition.of.resources)
```

```
0.784740581168297
```


Income Composition of Resources
0.0
0.2
0.4
0.6
0.8
5
10
15
20
Schooling