MATH 6357

Multilinear regression model to predict Life Expectancy

Authors:

Phuong Delrosario

Mariah Ochoa

Patricia Sieng

Kishore Tumarada

Professor: Wenshuang Wang

Date:12/07/2020

Contents

1. Introduction

In this project, we are interested in studying the various socioeconomic and health factors that could potentially affect life expectancy (at birth) which is defined by the World Health Organization (WHO) as "the average number of years that a newborn could expect to live, if he or she were to pass through life exposed to the sex- and age-specific death rates prevailing at the time of his or her birth, for a specific year, in a given country, territory, or geographic area."

Our dataset was obtained from Kaggle which combines data related to life expectancy from the Global Health Organization (GHO) dataset repository under WHO for 193 countries over a 15-year period (2000-2015). It originally contains 2,938 observations and 21 predictor variables (Country, Year, Status, Adult Mortality, Infant deaths, Alcohol, Percentage expenditures, Hepatitis B, Measles, BMI, Under-five-deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, Thinness 1-19 years, Thinness 5-9 years, Income composition of resources, Schooling), and life expectancy as the one response variable. More details about the variables are shown in Figure 1 in appendix. The research questions for this project are as follows:

1. Is there a linear regression relationship between the response variable, life expectancy, and the set of predictor variables?
2. What are the predicting variables affecting the life expectancy?
3. Do socioeconomic or health predictors have a more significant impact on predicting life expectancy?
4. What is the parsimonious linear regression model for prediction of life expectancy?
5. Apply the linear regression model in research question 4 is to predict the life expectancy.

2. Methodology

In section 3.1, we have done exploratory data analysis to examine the missing values, outliers and correlations among all the predictors and response variable. After data cleaning, we have reduced the dataset to 19 predictors and 1649 cases. We have scaled and standardized all predictors based on their corresponding means and standard deviations. Then we have fit a first order multilinear regression model and performed hypothesis testing using an F-test to examine the statistical significance of the model in section 3.2. Here we have identified a subset of predictors that contribute to maximum variance in the response variable, based on the adjusted R-squared value. We have also appraised residuals' plots to check the basic assumptions of regression model – Linearity, Normality and constant error variance in section 3.3.

Furthermore, in section 3.4, we have probed the research question about comparative contribution of socioeconomic and health factors to predict life expectancy. Here, we have used a parsimonious model with predictors of significant contribution to variance and reduced multicollinearity by dropping predictors with high correlation and variance inflation factor (VIF). We have done hypothesis testing using a general linear test and ANOVA table for comparative analysis of both groups.

Finally, in section 3.5, we have divided the dataset into train and test datasets in 80:20 ratio. We have used stepwise regression method to fit a parsimonious model on the train data based on the "best" model suggested. Afterwards, we calculated RMSE on predictions from the test data. We have also performed ANOVA analysis, hypothesis testing for slopes, and conducted diagnostic analysis of the final parsimonious model.

3. Data Analysis

3.1.Exploratory data analysis

The structure of the dataset was first checked, and the result is shown in Figure 1.1: Data Structure in the appendix. There are 3 categorical variables and 18 quantitative variables where “life expectancy” is the response variable. The three categorical variables are “Country” (193 countries), “Year” (2000-2015) and “Status” (2 level: developed and developing).

A new dummy variable “Developed” is created based on the “status” variable, where their values were either “Developing” or “Developed,” in order to include it in our regression model. If a country is developed, “Developed” is set as one. On the other hand, if a country is developing, “Developed” is set as zero.

The two categorical variables, Country and Year, were dropped from the data. They simply are labels for the name of the county where the data for a case was obtained and the year when it was collected. Thus, along with all missing values, they were dropped from the data. After cleaning up the data, it now has 1649 observations, 19 predictors, and 1 response variable (“life expectancy”). The new data set was used for further data analysis for the remaining parts of our report.

To move on, we calculated the correlation of our response variable (life expectancy) to all the other 19 predictors. From Table 1.1 in the appendix, we see that “Adult.Mortality”, “thinness 1-19 years”, “thinness 5-9 years”, “HIV.AIDS”, “under.five.deaths”, “infant.deaths”, “Measles” and “Population” are negatively correlated with life expectancy. The remaining 11 variables have a positive correlation with life expectancy. The top 3 predictors with the highest correlations with life expectancy are: “Schooling”, “Income.composition.of.resources” and “Adult.Mortality.” From the appendix, Figure 1.3 and Figure 1.4 show the histogram and boxplot of these three variables.

We pick out the top 10 variables that are highly correlated with life expectancy. Subsequently, the correlations between the top 10 variables were checked in order to avoid multicollinearity. Figure 1.2 shows that “GDP” and “percentage.expenditure” have a higher correlation. There is a high correlation between “thinness.1.19.years” and “thinness.5.9.years”. In addition, “Schooling” and “Income.composition.of.resources” are also highly correlated.

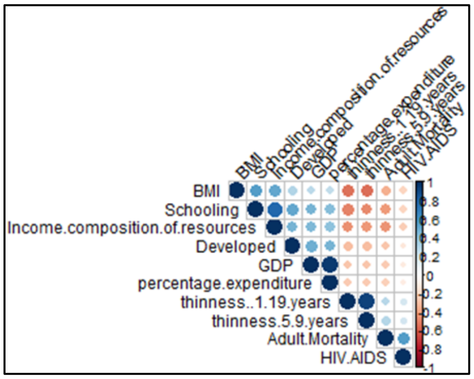


Figure 1.2: The correlations between the top 10 variables

"Schooling" is a normal distribution. "Income.composition.of.resources" is a left-skewed distribution and "Adult.Mortality" is a right-skewed distribution. There is no obvious outlier in "Schooling." There are some values of zero in "Income.composition.of.resources." From the boxplot, we speculate that there are some outliers in "Adult.Mortality." The scatterplot of "Adult.Mortality" and "Life expectancy" was constructed afterwards.

We proceeded to examine outliers and removed them if these points have a strong influence on the slope of the fitted line in the next stage of our analysis.

3.2.Test for a regression relationship between the response variables and the set of 19 predictor variables.

To test whether there is a linear regression relationship between life expectancy and the set of 19 variables, a first-order regression model based on all predictor variables was fitted to serve as a starting point. The general multilinear regression model of 19 predictors of the data is as follows below:

$$\begin{aligned} Y_i = \ & \beta_1 X_{i\ Adu.M} + \beta_2 X_{i\ inf.D} + \beta_3 X_{i\ Alco} + \beta_4 X_{i\ per.E} + \beta_5 X_{i\ Hepa.B} + \beta_6 X_{i\ Meas} + \beta_7 X_{i\ BMI} \\ & + \beta_8 X_{i\ und.five.D} + \beta_9 X_{i\ polio} + \beta_{10} X_{i\ total.E} + \beta_{11} X_{i\ diph} + \beta_{12} X_{i\ HIV} \\ & + \beta_{13} X_{i\ GDP} + \beta_{14} X_{i\ popul} + \beta_{15} X_{i\ thin.1.19} + \beta_{16} X_{i\ thin.5.9} + \beta_{17} X_{i\ income} \\ & + \beta_{18} X_{i\ school} + \beta_{19} X_{i\ devel} + \varepsilon_i \end{aligned}$$

- $i = 1, 2, \ldots\ldots, 1649$
- $Y_i = life\ expectant\ value\ in\ the\ ith\ observation$
- $X_{i\ Adu.M}, \ldots\ldots\ldots, X_{i\ devel} = values\ of\ the\ predictors\ in\ the\ ith\ observation$
- $\beta_1, \ldots\ldots, \beta_{19} = regression\ coefficient\ parameters\ of\ 19\ predictors$
- $\varepsilon_i = error\ term$

For the general multilinear regression model of 19 predictors, we would like to test if all the 19 parameters of the predictors are equal zero. The F-test for regression relation is as follows:
$H_0\colon \beta_1 = \beta_2 = \cdots \ldots . = \beta_{19} = 0$,
$H_a : not\ all\ \beta_i\ (i = 1, \ldots, 19)\ equal\ zero$
$F - statistics\ f = MSR/MSE$.

The multilinear regression (MLR) model function, lm(), was fitted using R-studio, and the result is shown in appendix - Figure 3.1. The F-statistic of 19 variables and 1629 DF is 435.7, and p-value is less than 2.2e-16. The p-value is less than the significant level α = 0.05, so the null hypothesis is rejected. There is evidence to conclude that at least one regression coefficient is not zero, and the model is statistically useful.

3.3.Build a linear regression model to test how and which variables related to the life expectancy.

The first MLR is built to model the linear relationship between life expectancy and the 19 predictor variables. Figure 3.1 in appendix shows the result of the MLR model for all 19 independent variables. With the significant level α = 0.05, there are 11 attributes that have a statistically significant correlation with life expectancy, which are Adult. Mortality, infant.deaths, Alcohol, percentage.expdenditure, BMI, under.five.deaths, Diphtheria, HIV.AIDS, Income.composition.of.resources, schooling, and Developed. Among the 11 variables, Adult.Mortality, Alcohol, under.five.deaths, and HIV.AIDS are negatively correlated with the life expectantcy. The adjusted R-squared of the model is 0.8336. It means that approximately 83.36% of variation in life expectancy can be explained by our model, and the linear regression model for estimating the life expectancy based on the 11 statistically significant variables is as follows:

$$\begin{aligned} Life\ expetancy \\ &= 53.48 - 0.0167X_{Adu.M} + 0.0935X_{inf.D} - 0.0914X_{i\ Alco} + 0.0003X_{per.E} \\ &+ 0.0338X_{BMI} - 0.0704X_{und.five.D} + 0.0149X_{diph} - 0.4370X_{HIV} \\ &+ 9.8170X_{income} + 0.8665X_{school} + 0.9684X_{\ develp} \end{aligned}$$

Furthermore, we look at the residuals' plots in Figure 3.2 in the appendix. It appears that there are constant error variances but some departure from normality at the tails of the plot. Thus, there might be potentials outliers in this dataset which we will investigate further later in this project.

3.4. Do socioeconomic or health predictors have a more significant impact on predicting life expectancy?

In this section, we will place the 11 significant variables from part 3 into two different groups to see which group of variables contribute more reduction in the variation of life expectancy. Group 1 will explore variables that relate to socioeconomic characteristics of a country: percentage.expenditure, income.composition.of.resources, schooling, and Developed. Group 2 will explore variables that relate to the health characteristics of a country: adult.mortality, infant.deaths, alcohol, BMI, under.five.deaths, diptheria, and HIV.AIDS. Before investigating each groups contribution, let's take a closer look at the correlation and Variation Inflation Factor (VIF) of the standardized data to ensure we do not need to make any further adjustments.

It is clear from Figure 4.1 that the following pair has a correlation value nearly equal to one: infant.deaths & under.five.deaths. Correlations between variables is not always an issue, but it stands to be further investigated as high multicollinearity can cause high p-values for individual predictors and low p-values for the overall model.

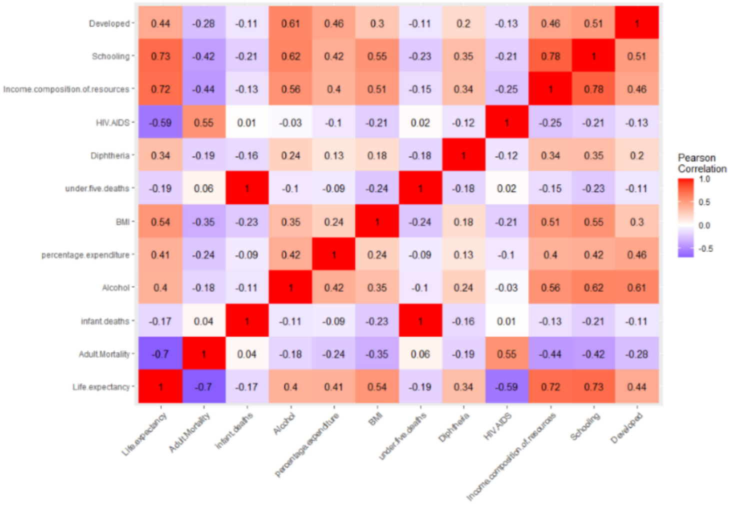


Figure 4.1: Correlation heatmap of the 11 most significant predictors

Predictor	VIF
Adult.Mortality	1.79
infant.deaths	183.79
Alcohol	2.21
percentage.expenditure	1.38
BMI	1.55
under.five.deaths	185.24
Diphtheria	1.21
HIV.AIDS	1.46
Income.composition.of.resources	2.93
Schooling	3.43
Developed	1.81

Table 4.1: VIF values of each predictor

The Table 4.1 from the appendix reveals that the two predictors (infant.death and under.five.deaths) have VIF values of 183.79 and 185.24, respectively. These VIF values are both greater than ten, which is a cause for concern. This high VIF values indicate that this multicollinearity may cause issues with the p-values of predictors in the model. Since infants are under the age of five, this indicates the infant deaths are a subset of the deaths of children under five years old, we can safely assume that dropping infant deaths will not take away the information given from this data from our model. Figure 4.3 from the appendix shows the summary of the full linear model without infant.deaths.

3.4.1 Group 1 Analysis

To tell how impactful the predictors are in group 1 and if the model could do without them, we will test the following hypotheses:

$$H_0: \beta_{per.E} = \beta_{income} = \beta_{school} = \beta_{develp} = 0$$

$$H_a: at\ least\ one\ \beta_i(i = per.E, income, school, develp)\ does\ not\ equal\ 0$$

We first created an estimated regression model with all predictors except for those related to socioeconomic factors to be part of group 1. This will be the reduced model for this hypothesis test. Running the ANOVA on the reduced model reveals an R^2 value of 0.7188 which means that 71.88-percent of the variation in life expectancy can be explained by all other predictors in the model. This reduced model also has an R^2_{adj} value of 0.7178. Comparing this to the ANOVA of the full model that excludes infant.death, which has an R^2 value of 0.826 and R^2_{adj} value of 0.825, shows that at one or more of the predictors in group 1 increases the reduction of variation in life expectancy by 0.1072 and increases the R^2_{adj} value by 0.1072, indicating that at least one of these predictors improves the fit of the model. The ANOVA also supplies a p-value of $2.2e-16$ which is less than $\alpha = 0.05$, indicating that we should reject the null hypothesis that all β_i in group 1 are equal to zero. The ANOVA table for this model is shown in appendix - Figure 4.3 reveals a coefficient of partial determination of 0.3814, which means that when group 1 is added to the model containing predictors that are not part of group 1, the sum of square error is reduced by 38.14-percent.

3.4.2 Group 2 Analysis

To tell how impactful the predictors are in group 2 and if the model could do without them, we will test the following hypotheses:

$$H_0: \beta_{Adu.M} = \beta_{Alco} = \beta_{BMI} = \beta_{und.five.D} = \beta_{diph} = \beta_{HIV} = 0$$

$$H_a: at\ least\ one\ \beta_i(i = Adu.M, Alco, BMI, und.five.D, diph, HIV\ )\ does\ not\ equal\ 0$$

We created an estimated model that contains all variables except for those in group 2. This will be the reduced model for this hypothesis test. Obtaining the ANOVA results on the reduced model also reveals an R^2 value of 0.5959 meaning that 59.59-percent of the variation in life expectancy can be explained by all other predictors in the model. This reduced model also has an R^2_{adj} value of 0.5949. Comparing this to the ANOVA of the full model, which has an R^2 value of 0.826 and R^2_{adj} value of 0.825, shows that one or more of the predictors in group 2 increases the reduction of variation in life expectancy by 19.46 and increases the R^2_{adj} value by 19.56, indicating that at least one of these predictors improves the model fit. The ANOVA table for this model is shown in appendix - Figure 4.4 gives coefficient of partial determination of 0.5695 indicating that when group 2 is added to a model that only contains the other predictors outside of group 2, the sum of square errors is reduced by 56.95-percent.

3.4.3 Comparing Outcomes

To see which group contributes more reduction in the variation of life expectancy, we will focus on the R^2_{adj} values since the R^2 values tend to increase as more predictors are added to the model. Unlike the R^2 value, R^2_{adj} penalizes the creator when non-significant predictors are included in the model. As stated earlier, the R^2_{adj} value for the group 1 (socioeconomic) reduced model is 0.7178 and the R^2_{adj} value for the group 2 (health) reduced model is 0.5949. Considering that the R^2_{adj} for the full model is 0.825, this means that at least one predictor in group 1 and at least one predictor in group 2 increased the reduction of the variation in life expectancy by 0.1072 and 0.2301, respectively. The reduction of the variation in life expectancy is higher for the predictors in group 2 than in group 1. In other words, the addition of one or more health predictors greatly improves the fit of the overall model and significantly contributes more reduction in the variation of life expectancy than if we were to add one or more socioeconomic predictors. To add on, the SSE was reduced by 38.14% when predictors in group 1 were added to a model that only contains the other predictors outside of group 1. Likewise, the SSE was reduced 56.95% when predictors in group 2 were added to a model that only contains the other predictors outside of group 2. This is higher than the reduction in SSE for the group 1 model. Thus, we can say that health predictors have a more significant impact on predicting life expectancy than the socioeconomic predictors.

3.5.Build a parsimonious linear regression model for prediction of life expectancy

Fitting a parsimonious model that explains variation in life expectancy with a small set of predictors, then predicting the life expectancy of some given cases.

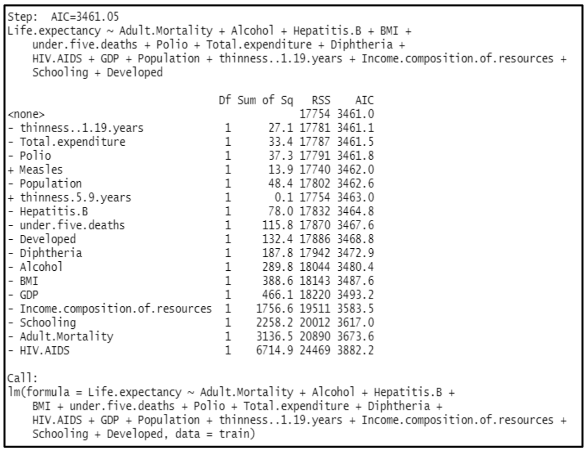
- Dividing the dataset into train (80-percent) and test data (20-percent)
- Applying Stepwise backward/forward/both to fit a parsimonious model in train data
- Predicting the dependent variable of the test data
- Calculating RMSE for the errors
- Conducting ANOVA analysis & hypothesis testing for slopes
- Conducting diagnostic analysis of the model

In order to find the best model, we start from scratch again and build our initial model with all 19 predictors as outlined in question 2. Refer again to Figure 3.2 in appendix to view the model diagnostic for the initial model. From the initial model diagnostic, we see that both the normality and constant variance assumption is valid.

In order to avoid highly correlated features and multicollinearity in our model, we investigated again the correlation between all predictors as well their corresponding VIF values. When viewing the correlation matrix (partial view shown in appendix – Figure 5.1 in appendix, we see two pairs of features that are highly correlated (nearly close to 1) with each other. The correlation between percentage.expenditure and GDP are 0.9593 while the correlation between under.five.deaths and infant.deaths is 0.9969. Looking at the VIFs values for all features (shown in appendix – Figure 5.2), we also see that the four features we listed above also have VIF values greater than 10, suggesting signs of serious multicollinearity. GDP, percentage.expenditure, infant.deaths, and under.five.deaths have the following VIF values: 13.57, 12.85, 212.19, and 202.01. We decided to solve this problem by dropping infant.deaths and percentage.expenditure. As previously explained in section 4, by definition, under.five.deaths (the morality rate of children under the age of 5) already considers the number of infant deaths (the mortality rate of children before their first birthday). Similarly, percentage.expenditure is the amount spent on health expressed as a percentage of GDP. In other words, percentage.expenditure is only a fraction of the GDP or the total amount spent on goods and services. The general multilinear regression model without infant.deaths and percentage.expenditure variables is as follows:

$$\begin{aligned} Life\ expetancy &= \beta_1 X_{i\ Adu.M} + \beta_2 X_{i\ Alco} + \beta_3 X_{i\ Hepa.B} + \beta_4 X_{i\ Meas} + \beta_5 X_{i\ BMI} + \beta_6 X_{i\ und.five.D} \\ &+ \beta_7 X_{i\ polio} + \beta_8 X_{i\ total.E} + \beta_9 X_{i\ diph} + \beta_{10} X_{i\ HIV} + \beta_{11} X_{i\ GDP} + \beta_{12} X_{i\ popul} \\ &+ \beta_{13} X_{i\ thin.1.19} + \beta_{14} X_{i\ thin.5.9} + \beta_{15} X_{i\ income} + \beta_{16} X_{i\ school} + \beta_{17} X_{i\ devel} + \varepsilon_i \end{aligned}$$

Now, our current model includes all features except infant.deaths and percentage.expenditure. We then divide the dataset into a train and test set using an 80:20 ratio. For the train set, we applied stepwise regression (backwards/forwards/both) to find the best model. For stepwise regression in all three directions, the model with the lowest AIC values includes the following features: Adult.Mortality, Alcohol, Hepatitis.B, BMI, under.five.deaths, Polio, Total.expenditure, Diptheria, HIV.AIDS, GDP, Population, Thinness.1.19, Income.composition.of resources, Schooling and Developed. Refer to Figure 5.3 to see the results from stepwise regression (direction = "both").



```
Step:  AIC=3461.05
Life.expectancy ~ Adult.Mortality + Alcohol + Hepatitis.B + BMI + 
    under.five.deaths + Polio + Total.expenditure + Diphtheria + 
    HIV.AIDS + GDP + Population + thinness..1.19.years + Income.composition.of.resources + 
    Schooling + Developed

                                  Df Sum of Sq   RSS    AIC
<none>                                         17754 3461.0
- thinness..1.19.years             1      27.1 17781 3461.1
- Total.expenditure                1      33.4 17787 3461.5
- Polio                            1      37.3 17791 3461.8
+ Measles                          1      13.9 17740 3462.0
- Population                       1      48.4 17802 3462.6
+ thinness.5.9.years               1       0.1 17754 3463.0
- Hepatitis.B                      1      78.0 17832 3464.8
- under.five.deaths                1     115.8 17870 3467.6
- Developed                        1     132.4 17886 3468.8
- Diphtheria                       1     187.8 17942 3472.9
- Alcohol                          1     289.8 18044 3480.4
- BMI                              1     388.6 18143 3487.6
- GDP                              1     466.1 18220 3493.2
- Income.composition.of.resources  1    1756.6 19511 3583.5
- Schooling                        1    2258.2 20012 3617.0
- Adult.Mortality                  1    3136.5 20890 3673.6
- HIV.AIDS                         1    6714.9 24469 3882.2

Call:
lm(formula = Life.expectancy ~ Adult.Mortality + Alcohol + Hepatitis.B + 
    BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria + 
    HIV.AIDS + GDP + Population + thinness..1.19.years + Income.composition.of.resources + 
    Schooling + Developed, data = train)
```

Figure 5.3: "Best" model suggested by stepwise regression (direction = "both") fitted on the train set

We fit the train set based on the "best" model suggested by stepwise regression and used it to predict on the test set. The residual plot for the "best" model is shown in appendix - Figure 5.4, and the summary of the model is shown in appendix – Figure 5.5. We obtained an RMSE of 3.8582 for the test set. A plot of the fitted vs expected values for life expectancy for the test set is also diagnostic and shown in appendix – Figure 5.6. The "best" MLR model is as follows:

$$\begin{aligned} &Life\ expetancy \\ &\quad = 52.72\ - 0.0163X_{Adu.M} - 0.1724X_{i\ Alco} - 0.01227X_{hepa.B} + 0.03596X_{BMI} \\ &\quad - 0.002791X_{und.five.D} + 0.01001X_{polio} + 0.07293X_{total.Exp} +\ 0.02624X_{diph} \\ &\quad - 0.445X_{HIV} + 0.7189X_{GDP} + 3.688e - 09\ X_{popul} - 0.04416X_{thin.1.19} \\ &\quad + 10.65X_{income} + 0.8675X_{school} + 1.219X_{\ develp} \end{aligned}$$

For the “best” model, we also performed an ANOVA test to test if the 15 slopes of the model equal 0 against the alternative hypothesis which states that at least one slope is not equal to 0. The ANOVA table is shown in appendix – Figure 5.7. The F-test statistic is 432.5897 (MSR/MSE = 5623.667/13) which is greater than the $F_{(0.05,\ 15,\ 1303)}$ = 1.6741. Thus, we reject the null hypothesis and conclude that the model is statistically useful.

$$\begin{aligned} H_0\text{: } &\beta_{Adu.M} =\ \beta_{Alco} = \beta_{hepa.B} =\ \ \beta_{BMI} = \beta_{und.five.D} = \beta_{polio} = \beta_{total.Exp} = \beta_{diph} \\ &= \beta_{HIV}\ = \beta_{GDP} = = \beta_{popul}\ = \beta_{thin.1.19} =\ \beta_{income} =\ \beta_{school} = \beta_{develp} = 0 \end{aligned}$$

H_a: $at\ least\ one\ \beta_i\ does\ not\ equal\ 0$

$$\begin{pmatrix} i = Adu.M, Alco, BMI, und.five.D, Polio, Total.Exp, diph, HIV\ , \\ GDP, Popul, thin.1.19, Income, School, Develop \end{pmatrix}$$

3.6.Discussing the disadvantages and future improvements that can be made

We have designed our best model as a first order multiple linear regression model based on stepwise regression mode. However, we did not closely look at distribution of the dataset, especially outliers and influential points. In future, we would like to perform Influence diagnostics using measures such as Cook’s distance, DFFITS and DFBETAS and examine the impact of outliers and influential observations on the inferences of the model. Further, we would also like to analyze the interaction effects between the 15 chosen predictors using general linear test. For example, we can test the influence of interaction effects between socio-economic and health factors on Life expectancy rather than which one of these two groups influence the life expectancy (Research question 4).

4. Conclusion

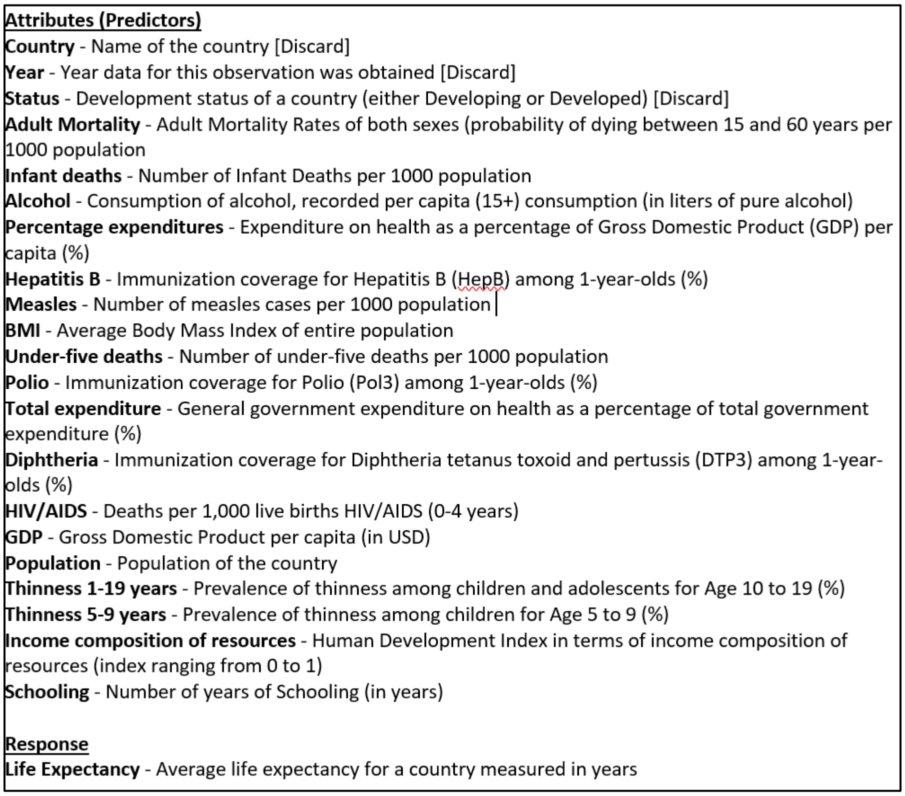
In this project, we have performed regression analysis on Life expectancy dataset with 22 predictors- socio-economic and health factors and average life expectancy of a country. In our analysis, we have done hypothesis testing and created a multilinear regression model with these predictors. We found that only 11 predictors have statistically significant relationship in the model. We have also probed various research questions based on this model. We have also examined multicollinearity among the predictors using VIF and fitted a parsimonious model with 15 predictors using stepwise regression model. In future, we would like to fit a quadratic multilinear regression model by taking interaction effects into account.

REFERENCES

1. Kutner, M. H., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). Boston, MA: McGraw-Hill.

2. KumarRajarshi .(2017). Life Expectancy (WHO): Statistical Analysis on factors influencing Life Expectancy. Retrieved November 1st , 2020 from https://www.kaggle.com/kumarajarshi/life-expectancy-who?select=Life+Expectancy+Data.csv.

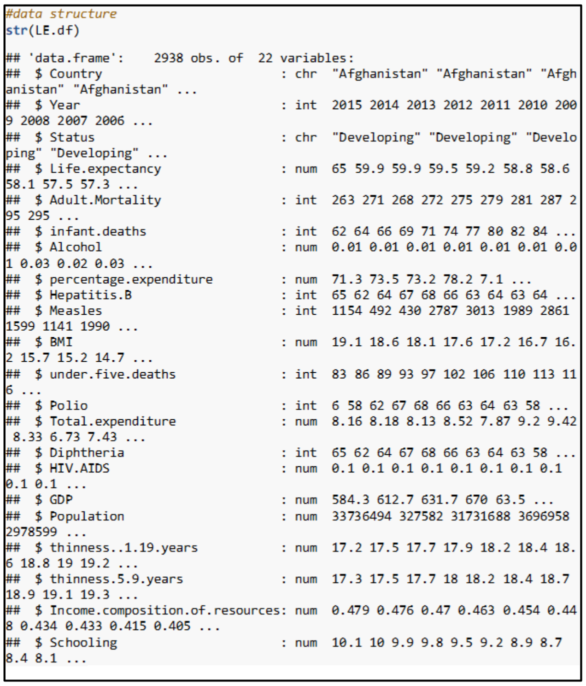
APPENDIX



Attributes (Predictors)
Country - Name of the country [Discard]
Year - Year data for this observation was obtained [Discard]
Status - Development status of a country (either Developing or Developed) [Discard]
Adult Mortality - Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population
Infant deaths - Number of Infant Deaths per 1000 population
Alcohol - Consumption of alcohol, recorded per capita (15+) consumption (in liters of pure alcohol)
Percentage expenditures - Expenditure on health as a percentage of Gross Domestic Product (GDP) per capita (%)
Hepatitis B - Immunization coverage for Hepatitis B (HepB) among 1-year-olds (%)
Measles - Number of 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 - Immunization coverage for Polio (Pol3) among 1-year-olds (%)
Total expenditure - General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria - Immunization coverage for Diphtheria tetanus toxoid and pertussis (DTP3) among 1-year-olds (%)
HIV/AIDS - Deaths per 1,000 live births HIV/AIDS (0-4 years)
GDP - Gross Domestic Product per capita (in USD)
Population - Population of the country
Thinness 1-19 years - Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
Thinness 5-9 years - Prevalence of thinness among children for Age 5 to 9 (%)
Income composition of resources - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling - Number of years of Schooling (in years)

Response
Life Expectancy - Average life expectancy for a country measured in years

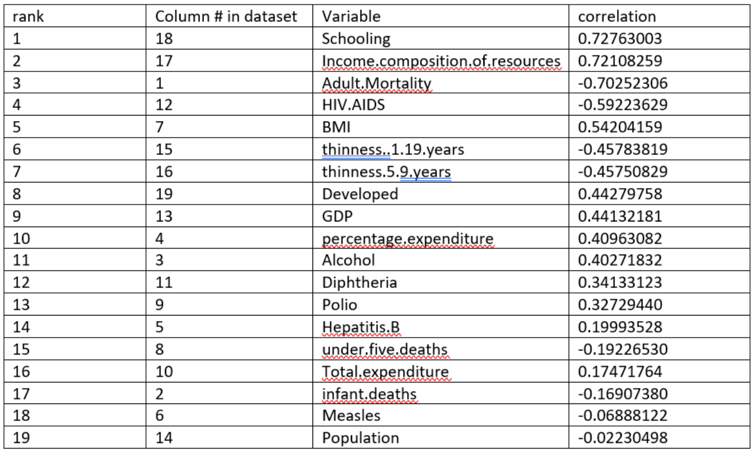
Figure 1: Variables explanation



```
#data structure
str(LE.df)

## 'data.frame':    2938 obs. of  22 variables:
##  $ Country                        : chr  "Afghanistan" "Afghanistan" "Afgh
anistan" "Afghanistan" ...
##  $ Year                           : int  2015 2014 2013 2012 2011 2010 200
9 2008 2007 2006 ...
##  $ Status                         : chr  "Developing" "Developing" "Develo
ping" "Developing" ...
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6
58.1 57.5 57.3 ...
##  $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 2
95 295 ...
##  $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.0
1 0.03 0.02 0.03 ...
##  $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : int  1154 492 430 2787 3013 1989 2861
1599 1141 1990 ...
##  $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.
2 15.7 15.2 14.7 ...
##  $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 11
6 ...
##  $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
##  $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42
 8.33 6.73 7.43 ...
##  $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
##  $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
0.1 0.1 ...
##  $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
##  $ Population                     : num  33736494 327582 31731688 3696958
2978599 ...
##  $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.
6 18.8 19 19.2 ...
##  $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7
18.9 19.1 19.3 ...
##  $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.44
8 0.434 0.433 0.415 0.405 ...
##  $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7
8.4 8.1 ...
```

Figure 1.1: Structure of the dataset



rank	Column # in dataset	Variable	correlation
1	18	Schooling	0.72763003
2	17	Income.composition.of.resources	0.72108259
3	1	Adult.Mortality	-0.70252306
4	12	HIV.AIDS	-0.59223629
5	7	BMI	0.54204159
6	15	thinness..1.19.years	-0.45783819
7	16	thinness.5.9.years	-0.45750829
8	19	Developed	0.44279758
9	13	GDP	0.44132181
10	4	percentage.expenditure	0.40963082
11	3	Alcohol	0.40271832
12	11	Diphtheria	0.34133123
13	9	Polio	0.32729440
14	5	Hepatitis.B	0.19993528
15	8	under.five.deaths	-0.19226530
16	10	Total.expenditure	0.17471764
17	2	infant.deaths	-0.16907380
18	6	Measles	-0.06888122
19	14	Population	-0.02230498

Table 1.1: The correlations between "life expectancy" and all the other variables.

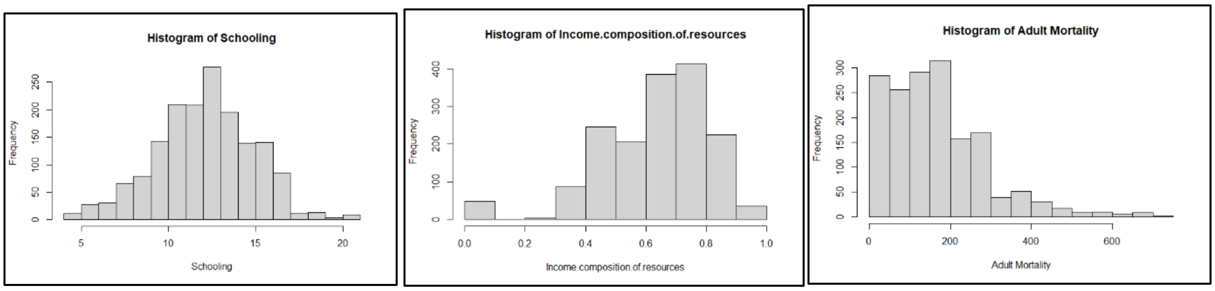


Figure 1.3: Histogram of "Schooling" (left), "Income.composition.of.resources" (middle) and "Adult.Mortality" (right)

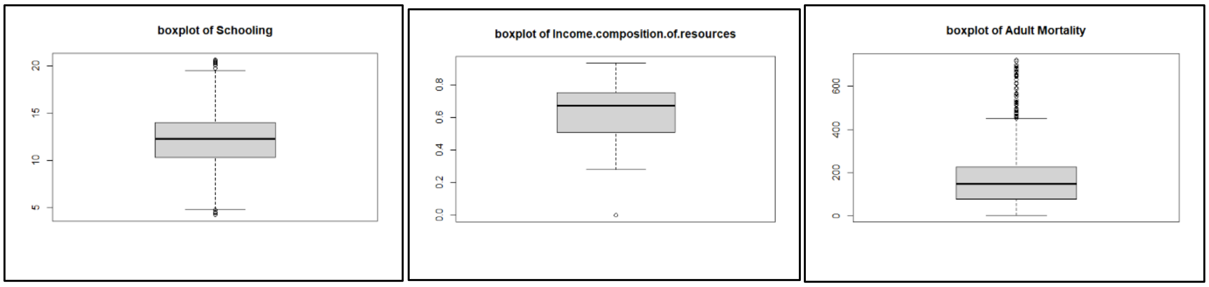


Figure 1.4: Boxplot of "Schooling" (left), "Income.composition.of.resources" (middle) and "Adult.Mortality" (right)

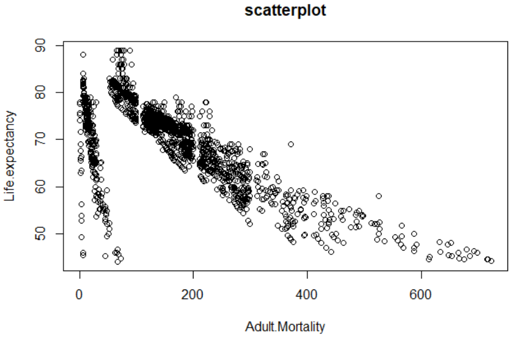
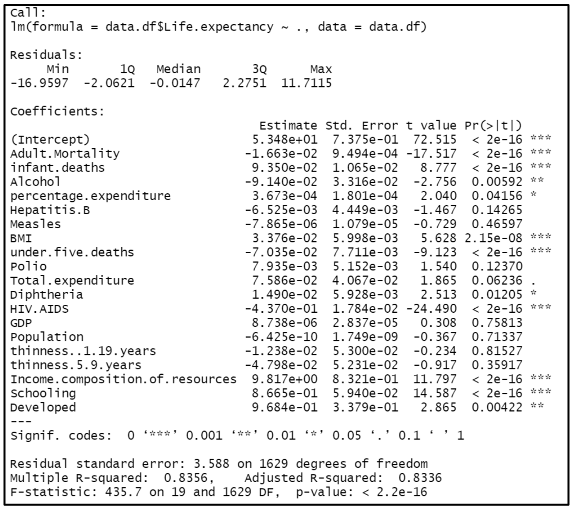


Figure 1.5: Scatterplot of “Adult.Mortality” and “Life expectancy”



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

Residuals:
     Min       1Q   Median       3Q      Max 
-16.9597  -2.0621  -0.0147   2.2751  11.7115 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     5.348e+01  7.375e-01  72.515  < 2e-16 ***
Adult.Mortality                -1.663e-02  9.494e-04 -17.517  < 2e-16 ***
infant.deaths                   9.350e-02  1.065e-02   8.777  < 2e-16 ***
Alcohol                        -9.140e-02  3.316e-02  -2.756  0.00592 ** 
percentage.expenditure          3.673e-04  1.801e-04   2.040  0.04156 *  
Hepatitis.B                    -6.525e-03  4.449e-03  -1.467  0.14265    
Measles                        -7.865e-06  1.079e-05  -0.729  0.46597    
BMI                             3.376e-02  5.998e-03   5.628 2.15e-08 ***
under.five.deaths              -7.035e-02  7.711e-03  -9.123  < 2e-16 ***
Polio                           7.935e-03  5.152e-03   1.540  0.12370    
Total.expenditure               7.586e-02  4.067e-02   1.865  0.06236 .  
Diphtheria                      1.490e-02  5.928e-03   2.513  0.01205 *  
HIV.AIDS                       -4.370e-01  1.784e-02 -24.490  < 2e-16 ***
GDP                             8.738e-06  2.837e-05   0.308  0.75813    
Population                     -6.425e-10  1.749e-09  -0.367  0.71337    
thinness..1.19.years           -1.238e-02  5.300e-02  -0.234  0.81527    
thinness.5.9.years             -4.798e-02  5.231e-02  -0.917  0.35917    
Income.composition.of.resources 9.817e+00  8.321e-01  11.797  < 2e-16 ***
Schooling                       8.665e-01  5.940e-02  14.587  < 2e-16 ***
Developed                       9.684e-01  3.379e-01   2.865  0.00422 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.588 on 1629 degrees of freedom
Multiple R-squared:  0.8356,	Adjusted R-squared:  0.8336 
F-statistic: 435.7 on 19 and 1629 DF,  p-value: < 2.2e-16
```

Figure 3.1: The result of the MLR model of life expectancy and 19 variables

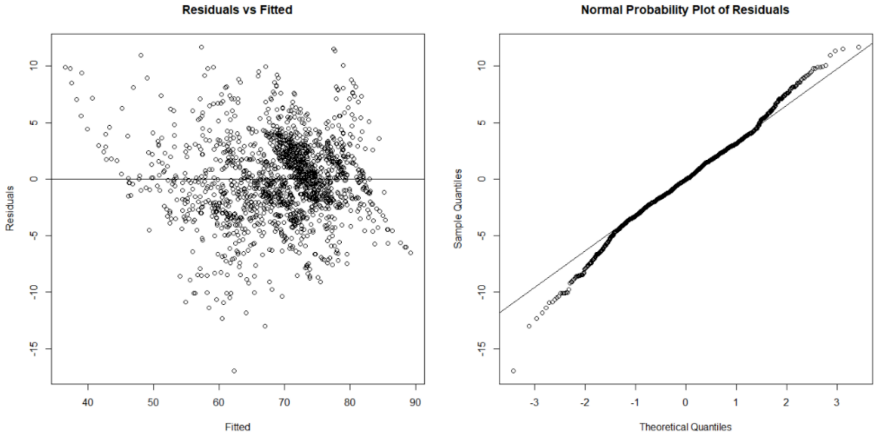
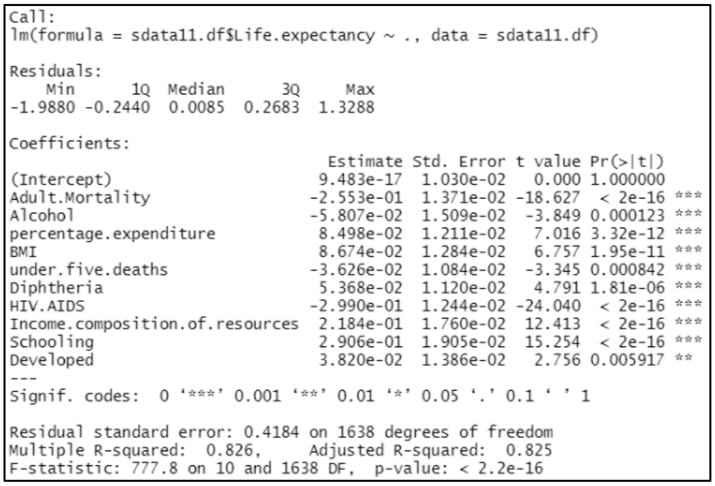


Figure 3.2: Residuals plot of the MLR model of life expectancy and 19 variables



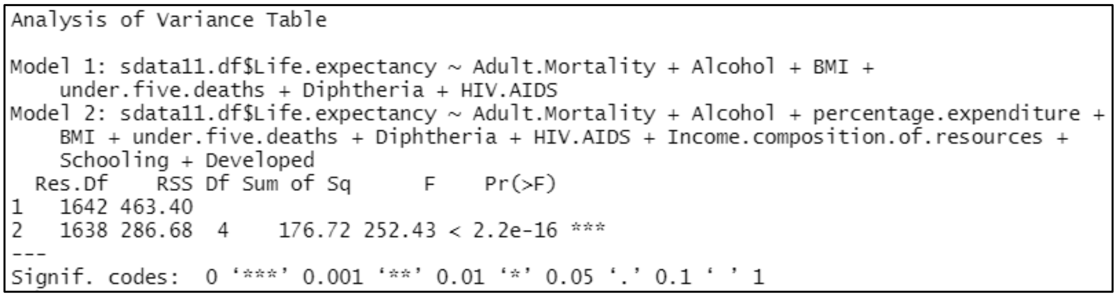
```
Call:
lm(formula = sdata11.df$Life.expectancy ~ ., data = sdata11.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9880 -0.2440  0.0085  0.2683  1.3288 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     9.483e-17  1.030e-02   0.000 1.000000    
Adult.Mortality                -2.553e-01  1.371e-02 -18.627  < 2e-16 ***
Alcohol                        -5.807e-02  1.509e-02  -3.849 0.000123 ***
percentage.expenditure          8.498e-02  1.211e-02   7.016 3.32e-12 ***
BMI                             8.674e-02  1.284e-02   6.757 1.95e-11 ***
under.five.deaths              -3.626e-02  1.084e-02  -3.345 0.000842 ***
Diphtheria                      5.368e-02  1.120e-02   4.791 1.81e-06 ***
HIV.AIDS                       -2.990e-01  1.244e-02 -24.040  < 2e-16 ***
Income.composition.of.resources  2.184e-01  1.760e-02  12.413  < 2e-16 ***
Schooling                       2.906e-01  1.905e-02  15.254  < 2e-16 ***
Developed                       3.820e-02  1.386e-02   2.756 0.005917 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4184 on 1638 degrees of freedom
Multiple R-squared:  0.826,	Adjusted R-squared:  0.825 
F-statistic: 777.8 on 10 and 1638 DF,  p-value: < 2.2e-16
```

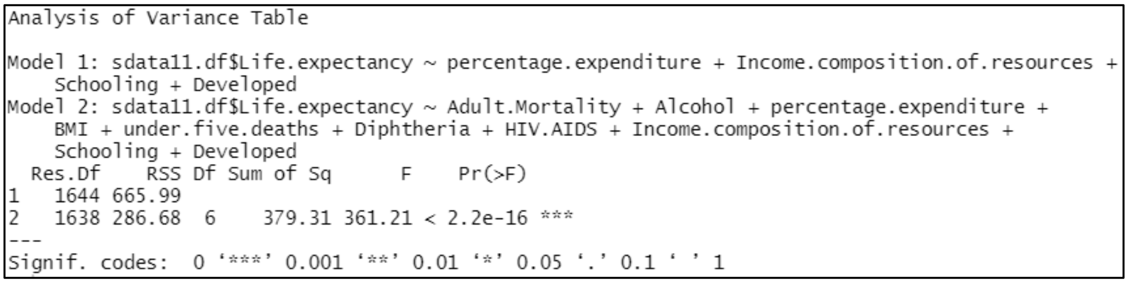
Figure 4.2: Summary of full linear model without infant.deaths



```
Analysis of Variance Table

Model 1: sdata11.df$Life.expectancy ~ Adult.Mortality + Alcohol + BMI + 
    under.five.deaths + Diphtheria + HIV.AIDS
Model 2: sdata11.df$Life.expectancy ~ Adult.Mortality + Alcohol + percentage.expenditure + 
    BMI + under.five.deaths + Diphtheria + HIV.AIDS + Income.composition.of.resources + 
    Schooling + Developed
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1   1642 463.40                                  
2   1638 286.68  4    176.72 252.43 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

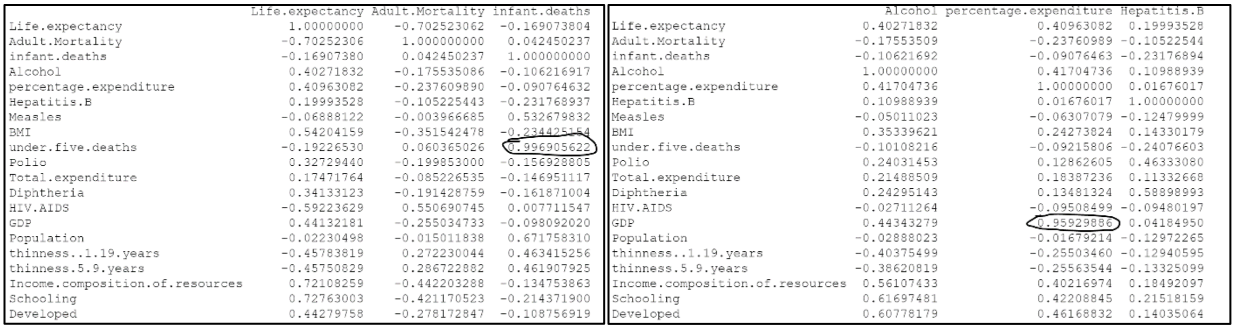
Figure 4.3: Analysis of variance with group 1 full and reduced models



```
Analysis of Variance Table

Model 1: sdata11.df$Life.expectancy ~ percentage.expenditure + Income.composition.of.resources + 
    Schooling + Developed
Model 2: sdata11.df$Life.expectancy ~ Adult.Mortality + Alcohol + percentage.expenditure + 
    BMI + under.five.deaths + Diphtheria + HIV.AIDS + Income.composition.of.resources + 
    Schooling + Developed
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1   1644 665.99                                  
2   1638 286.68  6    379.31 361.21 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

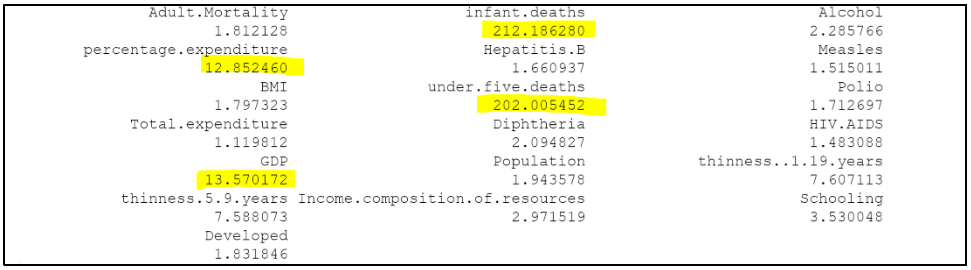
Figure 4.4: Analysis of variance with group 2 full and reduced models



```
                               Life.expectancy Adult.Mortality infant.deaths
Life.expectancy                     1.00000000    -0.702523062  -0.169073804
Adult.Mortality                    -0.70252306     1.000000000   0.042450237
infant.deaths                      -0.16907380     0.042450237   1.000000000
Alcohol                             0.40271832    -0.175535086  -0.106216917
percentage.expenditure              0.40963082    -0.237609890  -0.090764632
Hepatitis.B                         0.19993528    -0.105225443  -0.231768937
Measles                            -0.06888122    -0.003966685   0.532679832
BMI                                 0.54204159    -0.351542478  -0.234425154
under.five.deaths                  -0.19226530     0.060365026   0.996905622
Polio                               0.32729440    -0.199853000  -0.156928805
Total.expenditure                   0.17471764    -0.085226535  -0.146951117
Diphtheria                          0.34133123    -0.191428759  -0.161871004
HIV.AIDS                           -0.59223629     0.550690745   0.007711547
GDP                                 0.44132181    -0.255034733  -0.098092020
Population                         -0.02230498    -0.015011838   0.671758310
thinness..1.19.years               -0.45783819     0.272230044   0.463415256
thinness.5.9.years                 -0.45750829     0.286722882   0.461907925
Income.composition.of.resources     0.72108259    -0.442203288  -0.134753863
Schooling                           0.72763003    -0.421170523  -0.214371900
Developed                           0.44279758    -0.278172847  -0.108756919
```

```
                                    Alcohol percentage.expenditure Hepatitis.B
Life.expectancy                  0.40271832             0.40963082  0.19993528
Adult.Mortality                 -0.17553509            -0.23760989 -0.10522544
infant.deaths                   -0.10621692            -0.09076463 -0.23176894
Alcohol                          1.00000000             0.41704736  0.10988939
percentage.expenditure           0.41704736             1.00000000  0.01676017
Hepatitis.B                      0.10988939             0.01676017  1.00000000
Measles                         -0.05011023            -0.06307079 -0.12479999
BMI                              0.35339621             0.24273824  0.14330179
under.five.deaths               -0.10108216            -0.09215806 -0.24076603
Polio                            0.24031453             0.12862605  0.46333080
Total.expenditure                0.21488509             0.18387236  0.11332668
Diphtheria                       0.24295143             0.13481324  0.58898993
HIV.AIDS                        -0.02711264            -0.09508499 -0.09480197
GDP                              0.44343279             0.95929886  0.04184950
Population                      -0.02888023            -0.01679214 -0.12972265
thinness..1.19.years            -0.40375499            -0.25503460 -0.12940595
thinness.5.9.years              -0.38620819            -0.25563544 -0.13325099
Income.composition.of.resources  0.56107433             0.40216974  0.18492097
Schooling                        0.61697481             0.42208845  0.21518159
Developed                        0.60778179             0.46168832  0.14035064
```

Figure 5.1: Partial view of correlation matrix



```
                Adult.Mortality                   infant.deaths                         Alcohol 
                       1.812128                      212.186280                        2.285766 
         percentage.expenditure                     Hepatitis.B                         Measles 
                      12.852460                        1.660937                        1.515011 
                            BMI               under.five.deaths                           Polio 
                       1.797323                      202.005452                        1.712697 
              Total.expenditure                      Diphtheria                        HIV.AIDS 
                       1.119812                        2.094827                        1.483088 
                            GDP                      Population            thinness..1.19.years 
                      13.570172                        1.943578                        7.607113 
             thinness.5.9.years Income.composition.of.resources                       Schooling 
                       7.588073                        2.971519                        3.530048 
                      Developed 
                       1.831846 
```

Figure 5.2: VIF values for all features

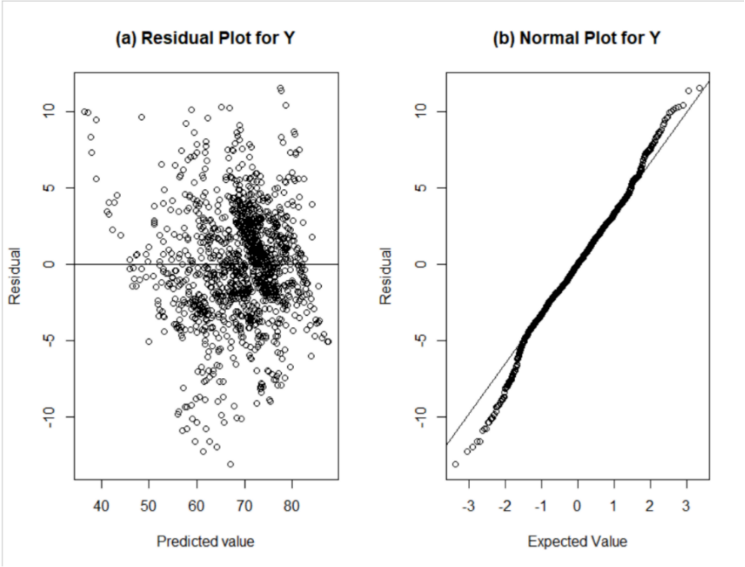
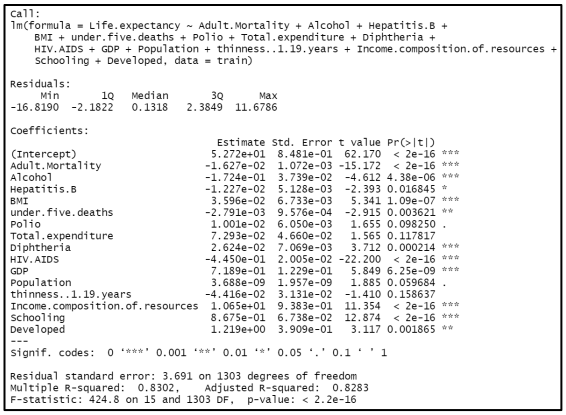


Figure 5.4: Residuals plot of the "best" model of life expectancy and 15 variables



```
Call:
lm(formula = Life.expectancy ~ Adult.Mortality + Alcohol + Hepatitis.B +
    BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria +
    HIV.AIDS + GDP + Population + thinness..1.19.years + Income.composition.of.resources +
    Schooling + Developed, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-16.8190  -2.1822   0.1318   2.3849  11.6786

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.272e+01  8.481e-01  62.170  < 2e-16 ***
Adult.Mortality                -1.627e-02  1.072e-03 -15.172  < 2e-16 ***
Alcohol                        -1.724e-01  3.739e-02  -4.612 4.38e-06 ***
Hepatitis.B                    -1.227e-02  5.128e-03  -2.393 0.016845 *
BMI                             3.596e-02  6.733e-03   5.341 1.09e-07 ***
under.five.deaths              -2.791e-03  9.576e-04  -2.915 0.003621 **
Polio                           1.001e-02  6.050e-03   1.655 0.098250 .
Total.expenditure               7.293e-02  4.660e-02   1.565 0.117817
Diphtheria                      2.624e-02  7.069e-03   3.712 0.000214 ***
HIV.AIDS                       -4.450e-01  2.005e-02 -22.200  < 2e-16 ***
GDP                             7.189e-01  1.229e-01   5.849 6.25e-09 ***
Population                      3.688e-09  1.957e-09   1.885 0.059684 .
thinness..1.19.years           -4.416e-02  3.131e-02  -1.410 0.158637
Income.composition.of.resources  1.065e+01  9.383e-01  11.354  < 2e-16 ***
Schooling                       8.675e-01  6.738e-02  12.874  < 2e-16 ***
Developed                       1.219e+00  3.909e-01   3.117 0.001865 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.691 on 1303 degrees of freedom
Multiple R-squared:  0.8302,    Adjusted R-squared:  0.8283
F-statistic: 424.8 on 15 and 1303 DF,  p-value: < 2.2e-16
```

Figure 5.5: Summary of the “best” linear model

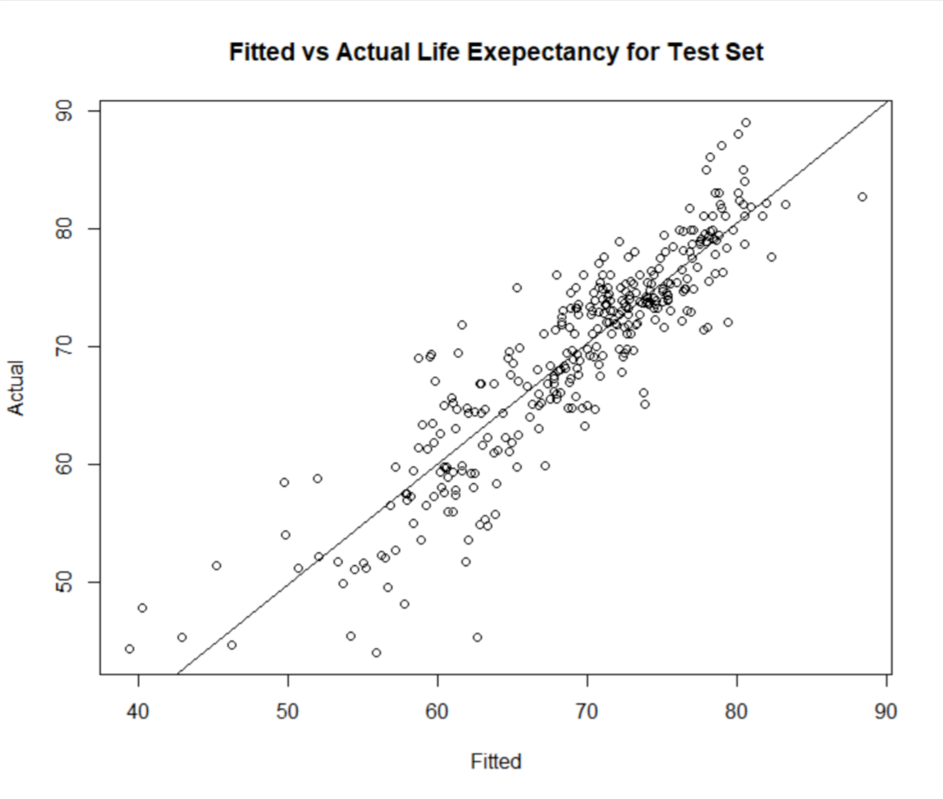
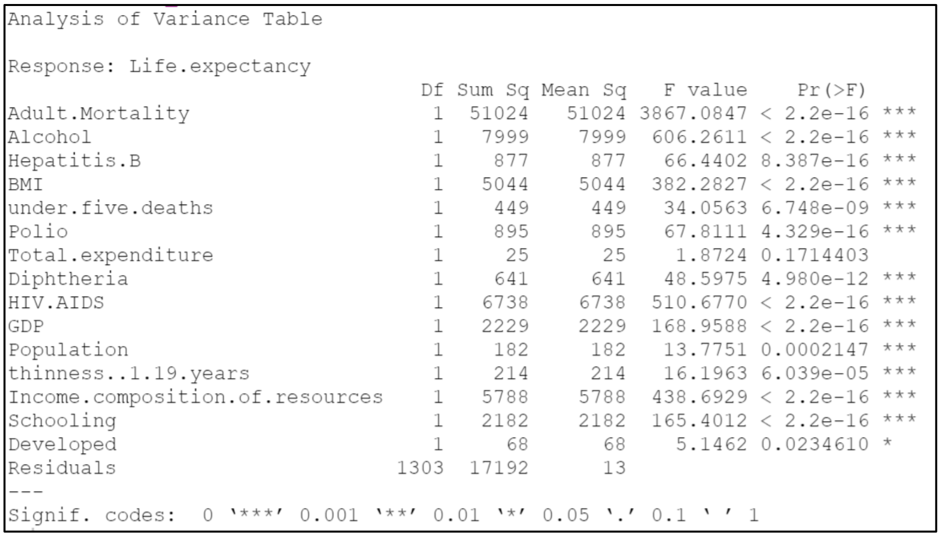


Figure 5.6: Plot of Fitted vs Actual Life Expectancy for the Test



```
Analysis of Variance Table

Response: Life.expectancy
                                  Df Sum Sq Mean Sq   F value    Pr(>F)    
Adult.Mortality                    1  51024   51024 3867.0847 < 2.2e-16 ***
Alcohol                            1   7999    7999  606.2611 < 2.2e-16 ***
Hepatitis.B                        1    877     877   66.4402 8.387e-16 ***
BMI                                1   5044    5044  382.2827 < 2.2e-16 ***
under.five.deaths                  1    449     449   34.0563 6.748e-09 ***
Polio                              1    895     895   67.8111 4.329e-16 ***
Total.expenditure                  1     25      25    1.8724 0.1714403    
Diphtheria                         1    641     641   48.5975 4.980e-12 ***
HIV.AIDS                           1   6738    6738  510.6770 < 2.2e-16 ***
GDP                                1   2229    2229  168.9588 < 2.2e-16 ***
Population                         1    182     182   13.7751 0.0002147 ***
thinness..1.19.years               1    214     214   16.1963 6.039e-05 ***
Income.composition.of.resources    1   5788    5788  438.6929 < 2.2e-16 ***
Schooling                          1   2182    2182  165.4012 < 2.2e-16 ***
Developed                          1     68      68    5.1462 0.0234610 *  
Residuals                       1303  17192      13                        
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5.7: ANOVA Table for “best” model