MATH 3340

**Life Expectancy Multiple Regression**

**Group 43 dataset 10**

**Yichuan Hu**

**Tony Sun**

**Xunze Zheng**

**2012/12/11**

**Life Expectancy Multiple Regression**

**Abstract**

The life expectancy is very indicator of diversity in the world population, which represents the overall health of the country and its people in terms other than death. Thus, it is significant to understand the different factors or policies that directly reflect the health indicator of the expected remaining years of functional well-being. This paper aims to assess each factor that directly affects the mortality conditions of a particular year, demonstrating their possible impact with a linear regression model drawn on the World Health Organization (WHO) dataset. Then, this information can be used for actuarial purposes in planning and policymaking. It is also useful in identifying high-risk populations for which preventive health care and medical care can compress morbidity during the last years of life. Therefore, the research wants to promote awareness regarding the sensitivity of life expectancy to different factors.

**Introduction**

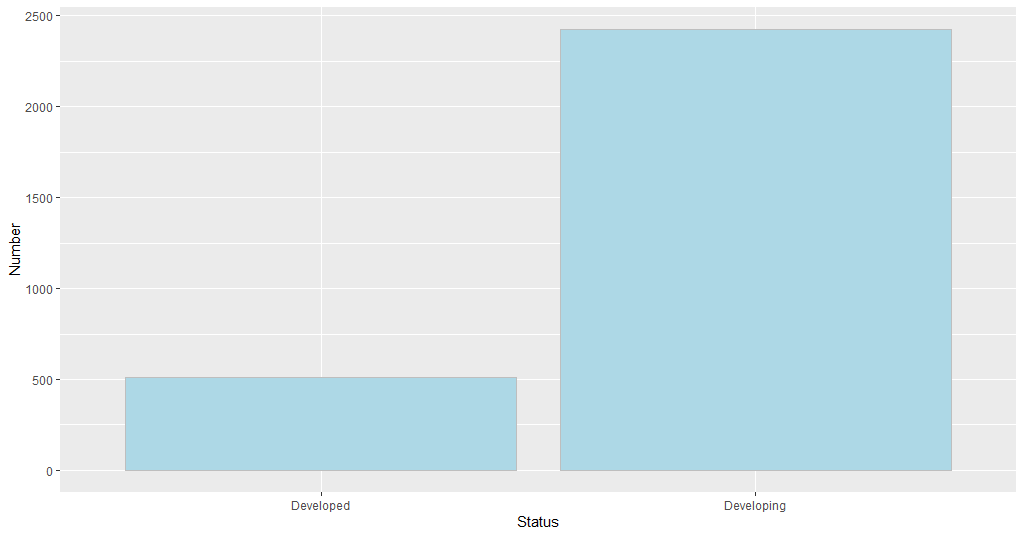
Life expectancy is one of the most commonly used summary statistics and determinants for the overall health of the population. In the simplest term, life expectancy is the arithmetic average of the number of years lived by a certain group of people. For example, five people lived until 50 years, and the other five people lived until 80 years. Thus, the life expectancy of these ten people will be 65 years. Demographers rarely use such an indicator because it is inconvenient. For example, to find out the life expectancy of a group of people born in the same year in a particular country or region (demographers call such groups birth cohorts), you have to wait until all its representatives die. In this case, the demographers are more interested in knowing what is happening and will happen to the people living now. These limitations force researchers when answering a seemingly simple question - how long do people live - to resort to mathematical modelling, i.e. regression modelling. However, the apparent simplicity of the indicator leads to numerous errors in interpretation. Speaking about people of the past over the age of 50, it is traditionally usual to note that "at that time" these people would be considered to be very old because, on general, people lived for about 30 years. However, that is where life expectancy misleads its viewers. Indeed, the average life expectancy of all live births in traditional societies was in the region of 30 years. But the reason for this is the very high mortality rate in infancy and early childhood. For example, the average age of five people who lived one year, four years, ten years, sixty years, and eighty years, will be thirty-one years. And a similar distribution of the age of death was observed until very recently. The smaller part of the people survived to their adulthood. And among this segment of the population, the perception of age was hardly radically different from the current one. There was just a phenomenal selection, selection of the healthiest and luckiest. Although the life expectancy is one of the commonly used indicators of the overall health of the population, it should not be misinterpreted by the age at which people die when it is actually the average measure of age which is affected by extreme values such as a death at the early age or death at the very old age. Therefore, these assumptions will be taken into the research when analysing the factors that affect the life expectancy rate.

**Data Description**

The given dataset contains data collected by the World Health Organization (WHO) and the United Nations to monitor factors that impact life expectancy. The data has 2938 data entries and 22 variables (or columns), which include country as “Country”, year as “Year”, developing status as “Status”, adult mortality as “Adult.Mortality”, life expectancy as “Life.expectancy”, infant deaths as “infant.deaths”, alcohol consumption per capita as “Alcohol”, country’s expenditure on health as “percentage.expenditure” and “Total.expenditure”, BMI as “BMI”, deaths under 5-years-old as “under.five.deaths”, deaths due to HIV/AIDS as “HIV.AIDS”, deaths due to Hepatitis B as “Hepatitis.B”, deaths due to Measles as “Measles”, deaths due to Polio as “Polio”, deaths due to Diphtheria as “Diphtheria”, deaths due to GDP as “GDP”, the population as “Population”, income information as “Income.composition.of.resources”, education as “Schooling”, and thinness as “thinness.5.9.years” and “thinness..1.19.years”. Three first variables i.e. Country, Year, and Status are categorical variable where the order is not important. There are presented 193 countries. All countries have sixteen observations per each country, except the Cook Islands, Dominica, Marshall Islands, Monaco, Nauru, Niue, Palau, Saint Kitts and Nevis, San Marino, Tuvalu who has only one observation per each country. The majority of these data entries (2426) belong to the countries with developing status while only 512 data entries belong to the countries with developed as Status which can be seen from Figure 1 or Figure 2.

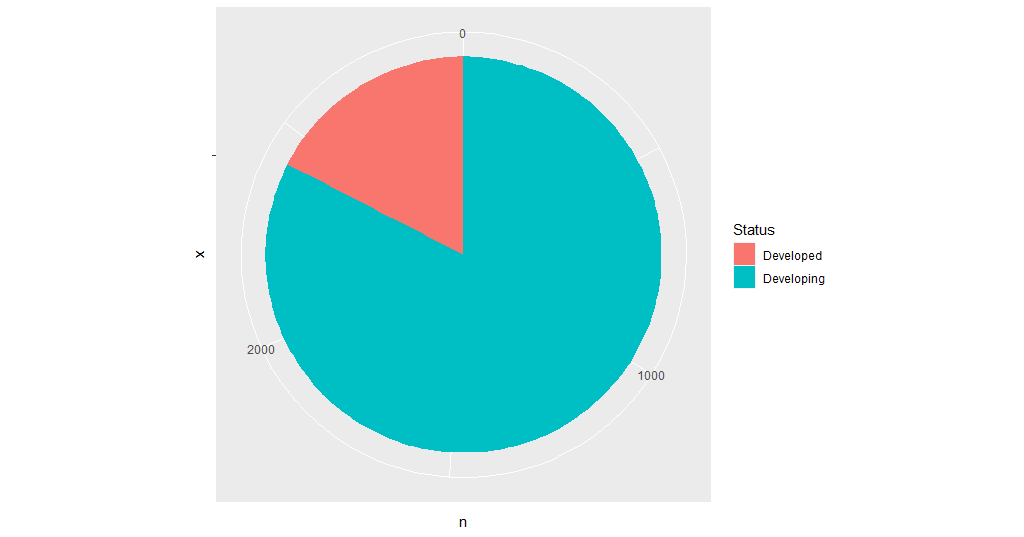
**Figure 1**

*Bar Chart for the Number of Data Entries by Status*



**Figure 2**

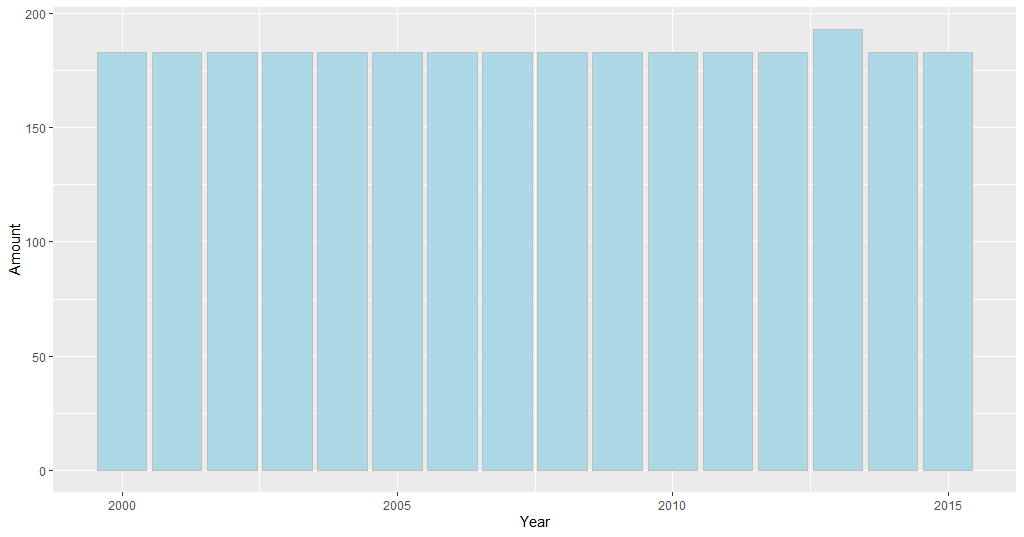
*Pie Chart for the Number of Data Entries by Status*



In this case, the data cover the period from 2000 to 2015, where each year has an equal number of data entries (183), except the year 2013 which has 193 data entries, which can be seen in Figure 3.

**Figure 3**

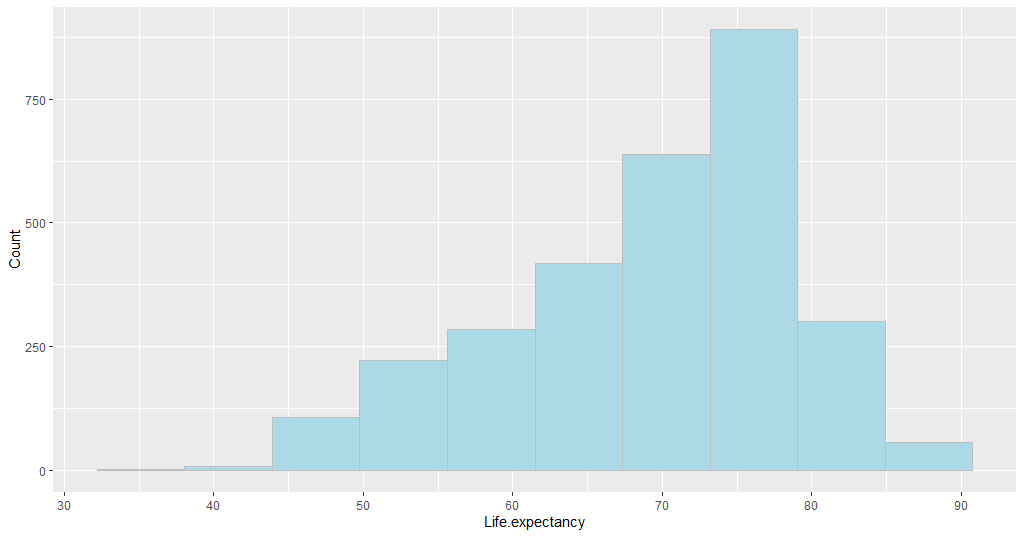
*Bar Chart for the Number of Data Entries by Year*



After the analysis of the categorical data, there can be analyzed the rest data, which are numerical ones. The first variable is the life expectancy with mean and median 69.22 and 72.10, respectively, meaning that the left tail of the distribution is going to be longer and thus the distribution is skewed to the left. It is evident in Figure 4.

**Figure 4**

*Histogram for the Life Expectancy*

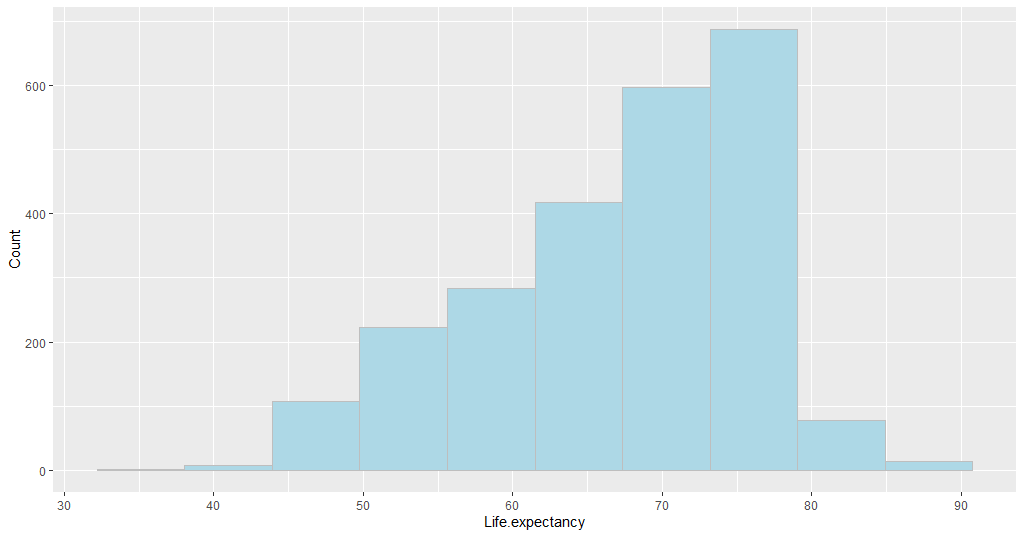


The minimum of the distribution is 36.30 years old, and the maximum is 89 years old.

Let check if the life expectancy differs by Status (see fig. 5, fig. 6, and fig. 7). Thus, there can be said that life expectancy in developed countries is much higher for people above 65 years old. This means that more people die at the age of 65 years old or older. According to the data analysis, the average life expectancy index, as well as median in developed countries, is 79.2 years old. At the same, in developing countries, few people survive to celebrate their sixty-fifth, seventieth, or eightieth birthday.

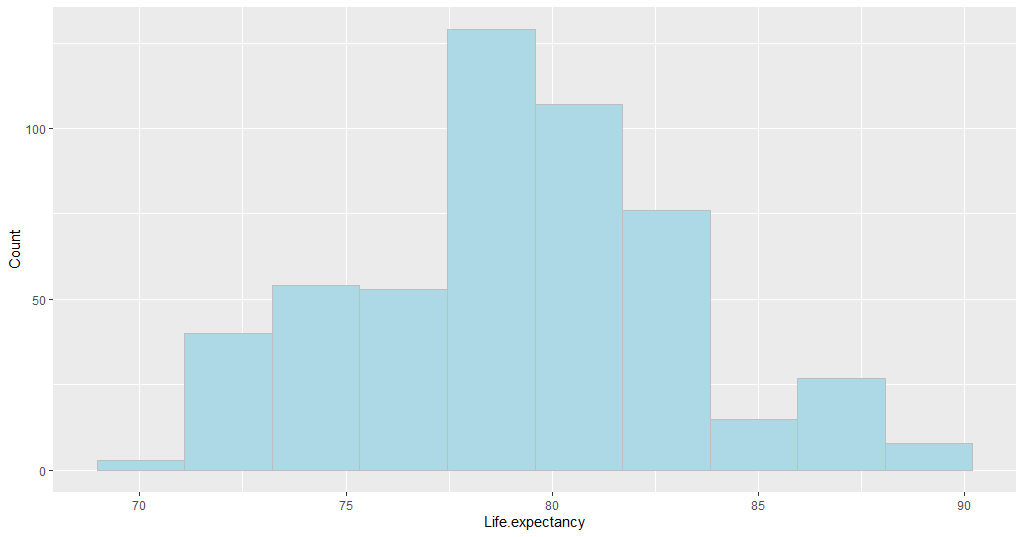
**Figure 5**

*Histogram for the Life Expectancy by Status Developing*



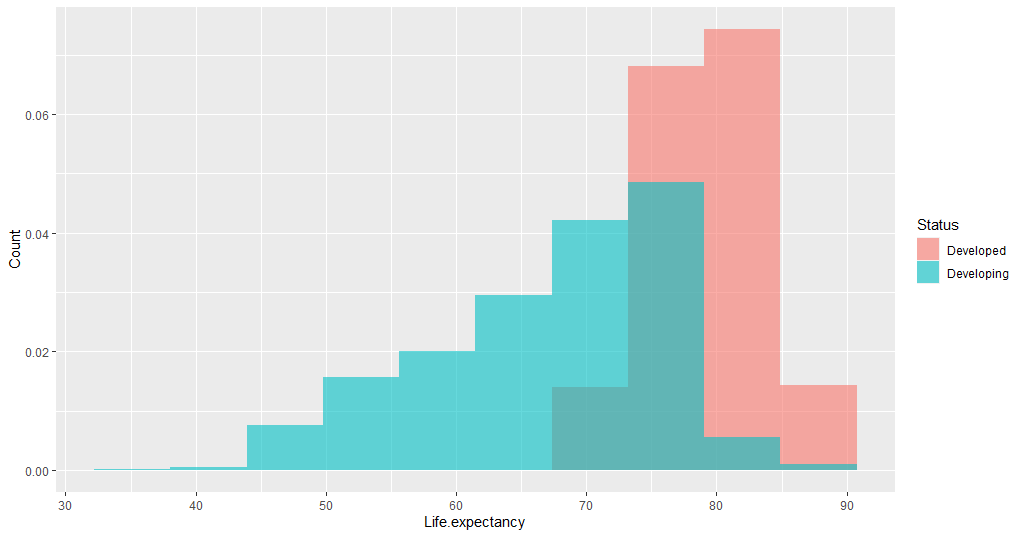
**Figure 6**

*Histogram for the Life Expectancy by Status Developed*



**Figure 7**

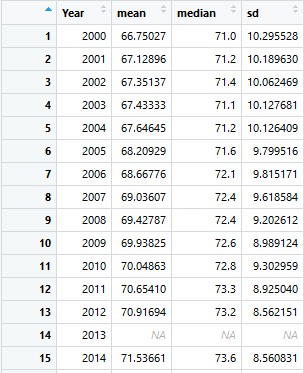
*Histogram for the Life Expectancy by Status*



Let also check if the life expectancy differs by Year (see fig. 8). From the table, there can be seen that both life expectancy mean and median have been increased from 2000 to 2014 while the standard deviation has been decreased.

**Figure 8**

*Table of Life Expectancy Mean, Median, and Standard Deviation by Year*



Such numbers mean that the average life expectancy has been increased by 4.786 years, on average. If there is taken the median, life expectancy has been increased by 2.6 years. The standard deviation demonstrates that the variability in life expectancy has been reduced by 1.73 years, which also is a positive signal. All these data are in Figure 9.

**Figure 9**

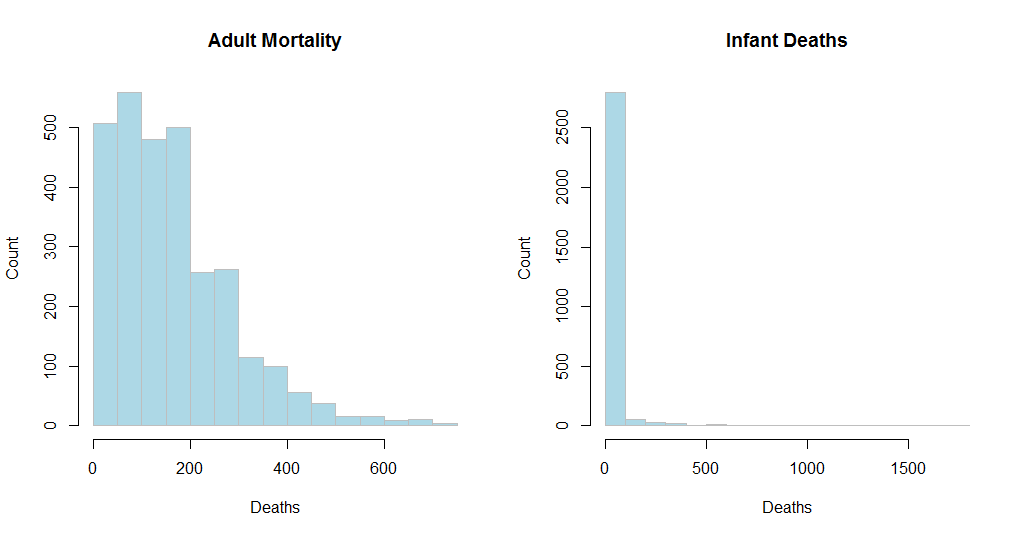
*The Difference between the Last Year (2014) and the First Year (2000) in the Dataset*



The next two variables are adult mortality and infant deaths. From Figure 10, both these variables are left-skewed, which means that their means are less than their medians. This could mean that the average does not display the actual picture and that the average number of deaths might be higher than the mean. The averages, as well as medians and standard deviations, cannot be found due to the presence of the missing values in the dataset.

**Figure 10**

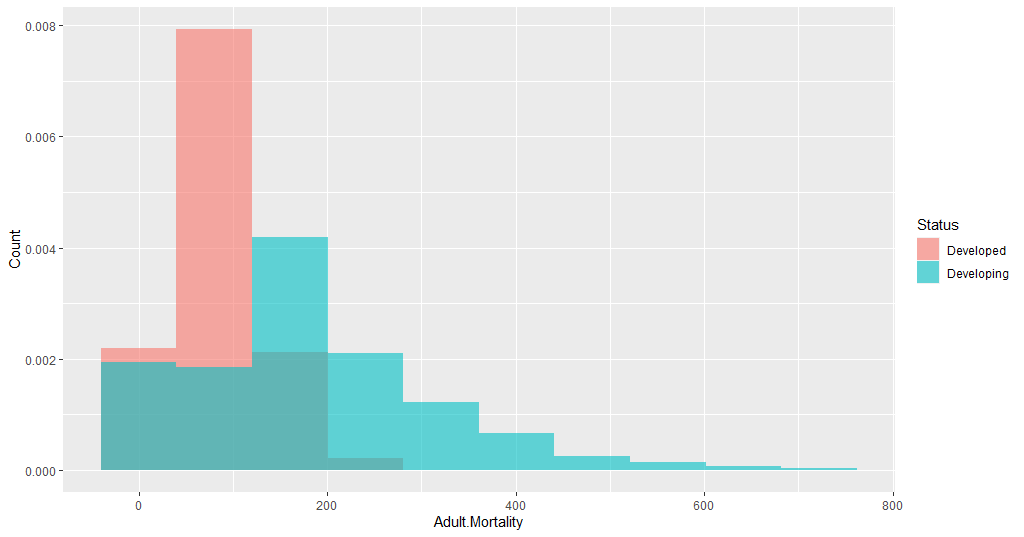
*Histogram for the Adult Mortality and Infant Deaths*



Let check how these values differ by Status. First, adult mortality in developed countries is much lower than in developing ones and does not exceed 300 deaths (see fig. 11). Second, infant deaths in developed countries also are much lower than in developed countries, it is very close to zero while in developing this number can be different from zero (see fig, 12). It seems like developed countries have better indicators than developing countries have. Thus, there will not be conducted any further analysis by Status to reduce some time.

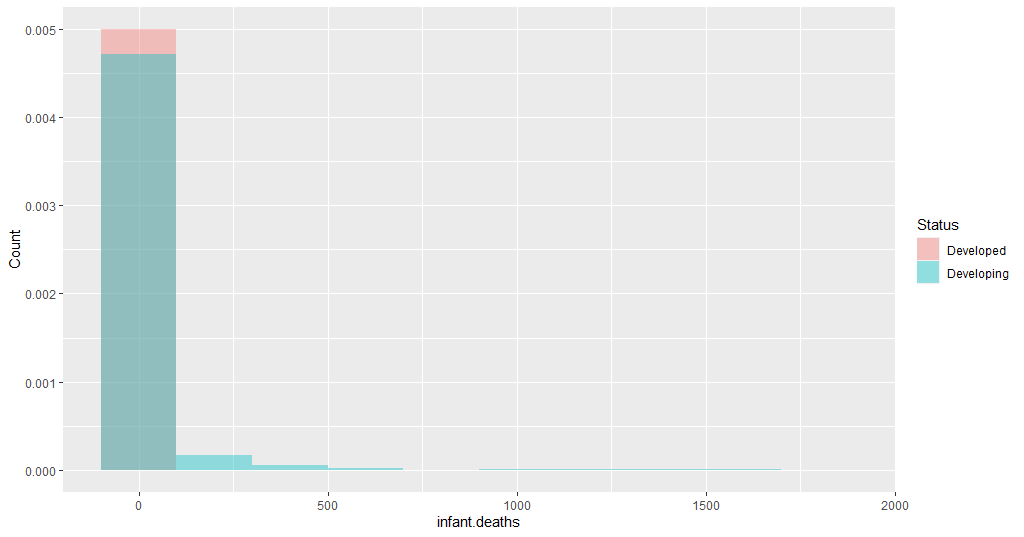
**Figure 11**

*Histogram of the Adult Mortality by Status*



**Figure 12**

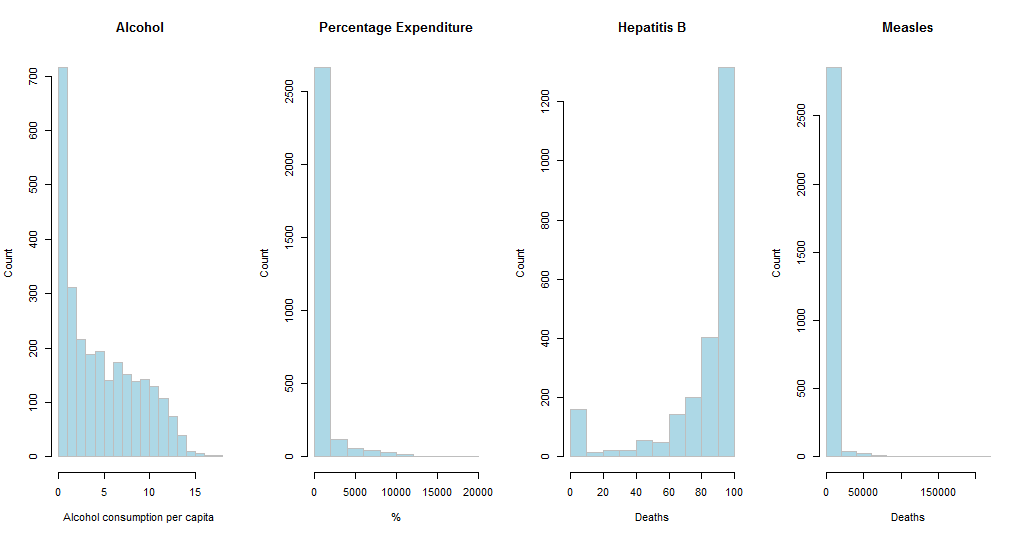
*Histogram of the Infant Deaths by Status*



The next four variables are Alcohol, Percentage.expenditure, Hepatitis.B, and Measles. From Figure 13, all the variables are right-skewed, except Hepatitis.B, which is left-skewed as the life expectancy variable. Thus, all the variables except Hepatitis B tend to be around zero or so. For example, Alcohol, Percentage.expenditure and Measles have the strongest values in zero, while Hepatitis.B – in 100.

**Figure 13**

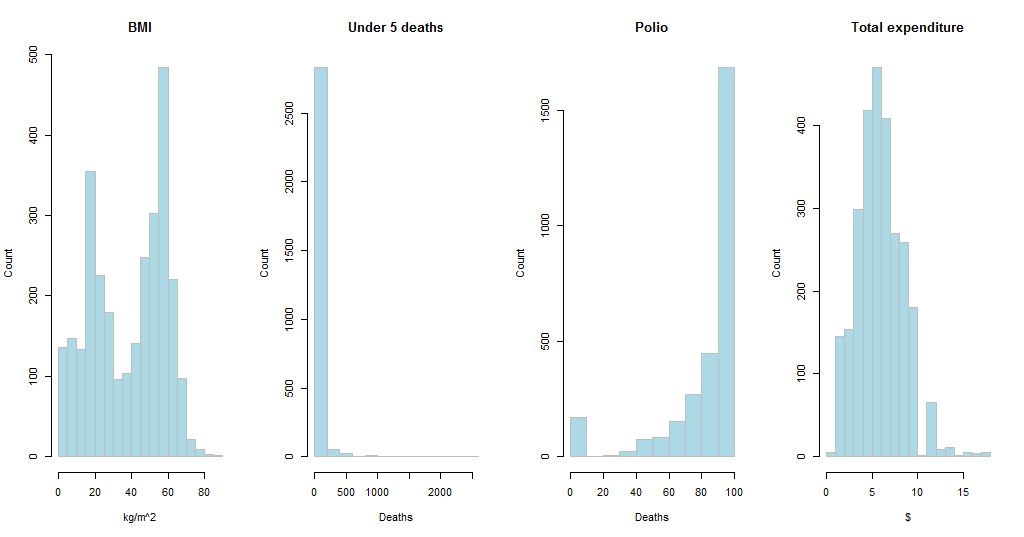
*Histogram for the Alcohol, Percentage Expenditure, Hepatitis B, and Measles*



The next four variables are BMI, under.five.deaths (or Under 5 deaths), Polio, and Total.expenditure. From Figure 13, the variable under.five.deaths is right-slewed which means that the majority of distribution is placed around zero or so. On the other hand, Polio as Hepatitis.B is left-skewed which means that these data can be very similar and redundant the linear regression model. Other variables such as BMI and Total expenditure demonstrate some values which are close to normality, though BMI may be considered to have two modes, which reduces their chances to effectively present the life expectancy.

**Figure 13**

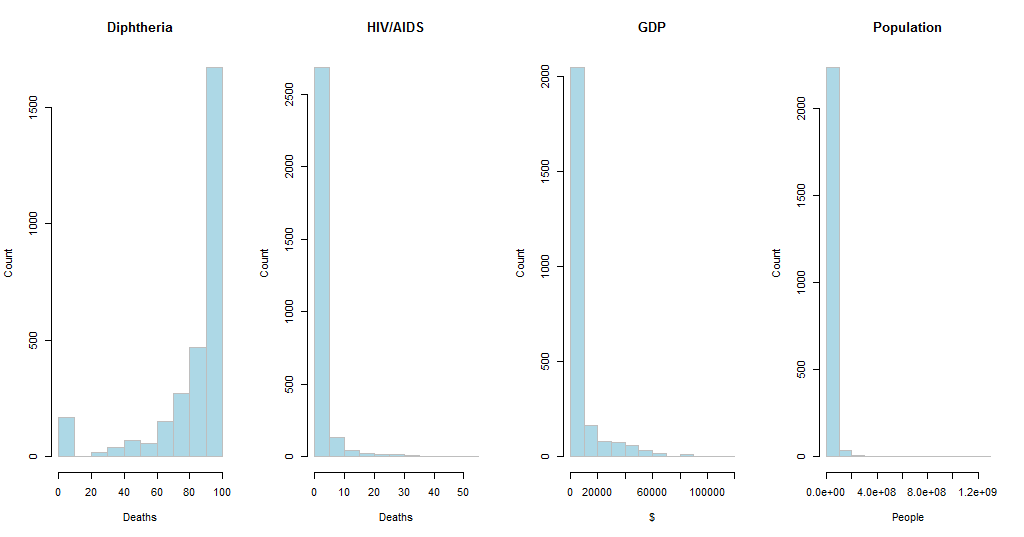
*Histogram for the BMI, Under 5 deaths, Polio, and Total Expenditure*



The next four variables are Diphtheria, HIV/AIDS, GDP, and Population. From Figure 14, all the variables except Diphtheria, are skewed to the right which means that the majority of distribution is placed around zero or so. On the other hand, Diphtheria as Polio and Hepatitis.B is left-skewed which means that confirms that some illnesses tend to be very similar in their distribution.

**Figure 14**

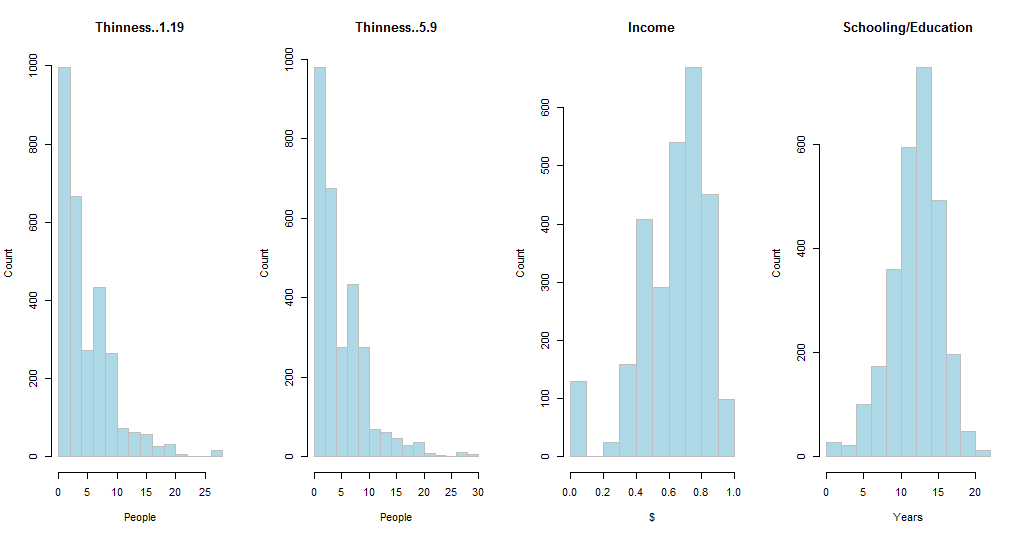
*Histogram for the Diphtheria, HIV/AIDS, GDP, and Population*



The last four variables are Thinness..1..19, Thinness..5..9, Income (Income.composition.of.resources), and Schooling. From Figure 15, two last variables demonstrate some normality in their distribution while the first two variables are right-skewed, meaning that the majority of their distribution is placed around zero or so.

**Figure 15**

*Histogram for the Diphtheria, HIV/AIDS, GDP, and Population*



In this case, there is an easier way to confirm the skewness of the data distribution: if the median is greater than mean, then the distribution is skewed to the left; if the median is less than mean, then the distribution is skewed to the right; and if the mean is equal to the median, then the distribution is normal. For example, from the summary statistics, there can be found that Adult.Mortality, infant.deaths, Alcohol, percentage.expenditure, Measles, under.five.deaths, Total.expenditure, HIV.AIDS, GDP are right-skewed, and thus, are less coherent with the distribution of the life expectancy which is left-skewed.

**Methods**

First, there will be removed all data entries with missing values by using command na.omit(data). Without this step, it is impossible to build the model. Next, all categorical variables, except Status, will be removed from the dataset. In this case, the Status variable will be transformed into dummy variable (0 for Developing countries, and 1 for Developed countries) since this variable proved to be a major factor that defines the level of life expectancy. Then, there will be conducted correlation analysis using command cor(data) to find the variables which are similar to each other and, thus, will be redundant to the linear regression. As a result, one of the variables that have a relationship with a correlation coefficient of 0.95 or higher will be removed from the dataset while building a regression model. It is also expected that variables with similar distributions might be highly correlated with each other. There also will be built a correlation matrix for more clarity. Then, there will be built a regression model. First, there will be built the first and full multiple linear regression using the command lm(variables, data). Second, using the summary statistics, there will be found which variables are not significant to the model. Third, in this case, the variable with the largest p-value will be ignored in the second multiple regression. Next, the first and second model will be compared using the command anova. If the p-value is insignificant (greater than 0.05), then there cannot be rejected the Null hypothesis that the full model is better. Fifth, this means that we go with the reduced (or second) model. The steps 1-4 will be repeated until there is rejected the Null hypothesis.

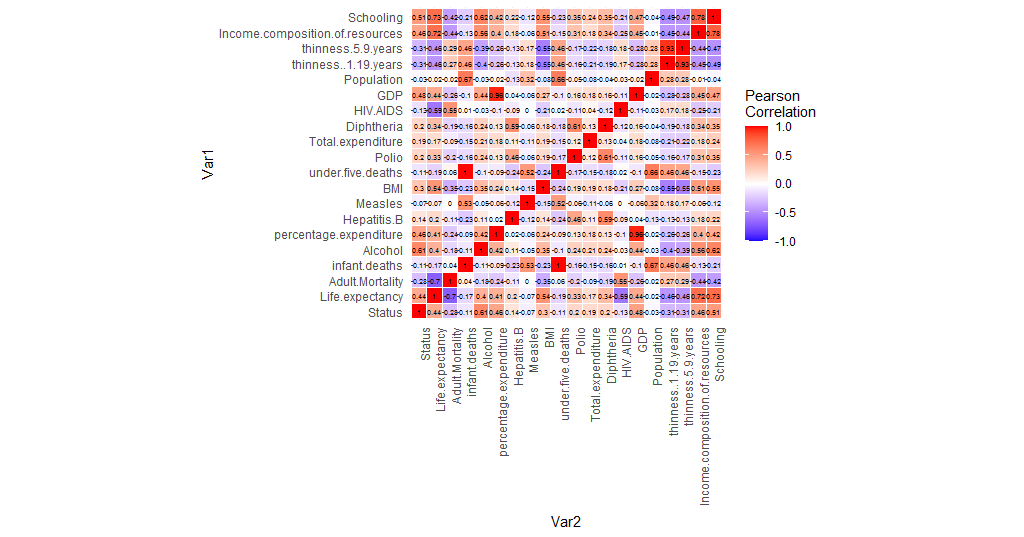
**Results**

After deleting all data entries, there will be only left 1649 observations, compared to 2938 observations previously. After removing the first categorical variables and changing Status on the dummy variable, there will be only twenty variables in the dataset, all of which will be of numerical type. The phase of cleaning the data has ended with these activities.

Next, there will be conducted correlation analysis of the data, which is displayed in Figure 16. From the correlations matrix, GDP and percentage.expenditure have a correlation coefficient that is equal to 0.96, and thus, one of these data (percentage.expenditure) should be removed from the dataset. In this case, infant.deaths and under.five.deaths have a greater coefficient of correlation which is 1, showing a perfect positive correlation, and thus, under.five.deaths will be removed from the current dataset. As a result, there will be left only 18 variables in the dataset.

**Figure 16**

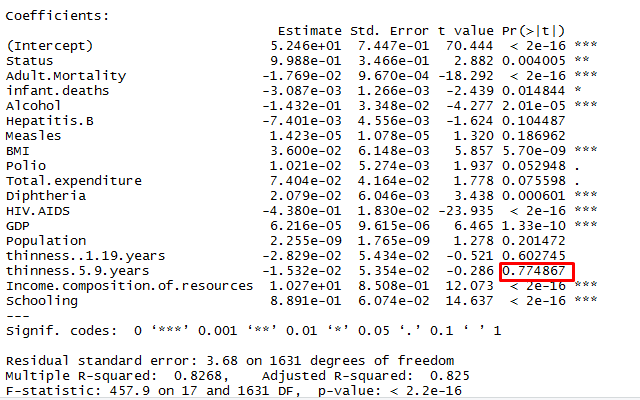
*Correlation Matrix*



The next step is building the full model. The summary statistic of the full regression model is in Figure 17.

**Figure 17**

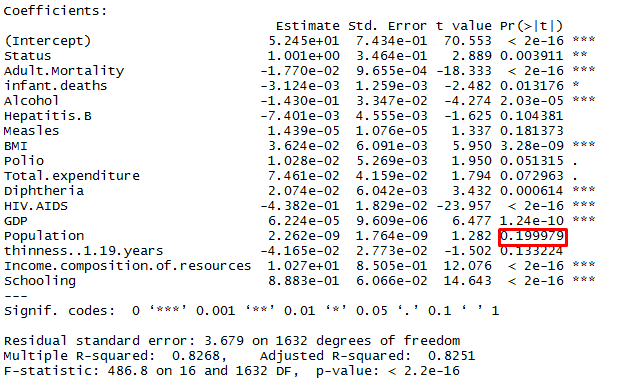
*Full Regression Summary Statistics*



In this case, the variable with the least insignificant (or the largest) p-value is thinness.5.9.years that has a p-value of 0.774867. Thus, this variable will be excluded in the next regression which summary statistic is in Figure 18.

**Figure 18**

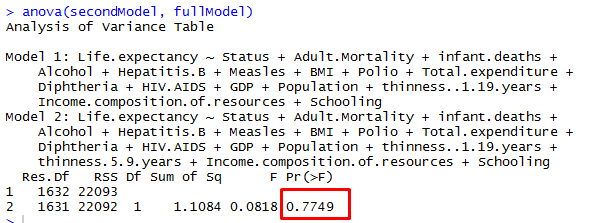
*Second Regression Summary Statistics*



In this case, the ANOVA statistics of these two variables (see fig. 19) show that the Null hypothesis cannot be rejected, and thus, it is necessary to continue to reduce the model.

**Figure 19**

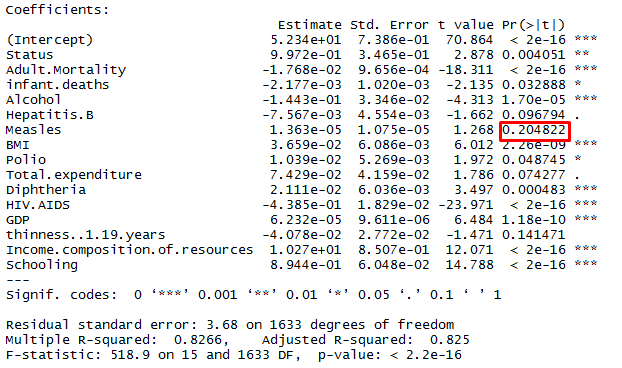
*ANOVA Statistics*



From Figure 18, Population has the largest p-value, and thus, will be removed from the next regression. The summary statistics of the third regression is in Figure 20.

**Figure 20**

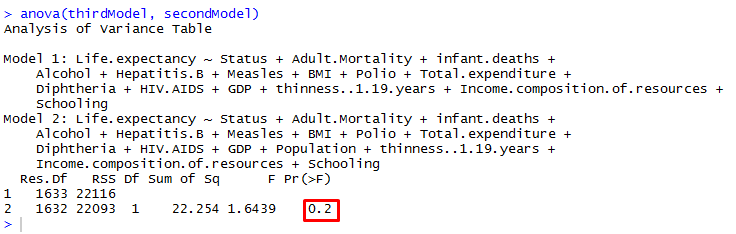
*Third Regression Summary Statistics*



In this case, the ANOVA statistics of these two variables (see fig. 21) show that the Null hypothesis cannot be rejected, and thus, it is necessary to continue to reduce the model.

**Figure 21**

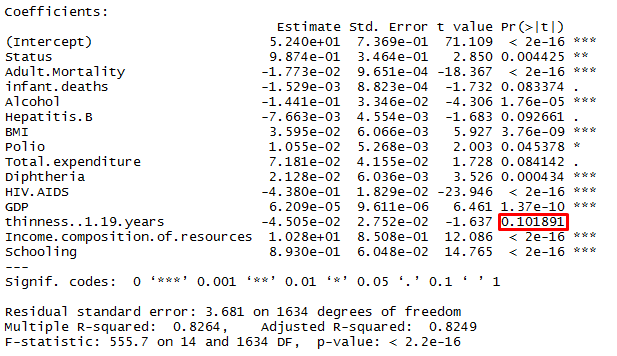
*ANOVA Statistics*



From Figure 20, Measles has the largest p-value, and thus, will be removed from the next regression. The summary statistics of the third regression is in Figure 22.

**Figure 22**

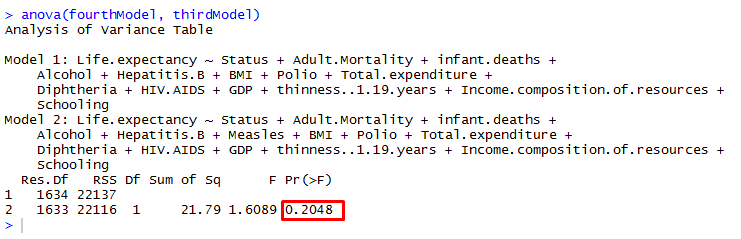
*Fourth Regression Summary Statistics*



In this case, the ANOVA statistics of these two variables (see fig. 23) show that the Null hypothesis cannot be rejected, and thus, it is necessary to continue to reduce the model.

**Figure 23**

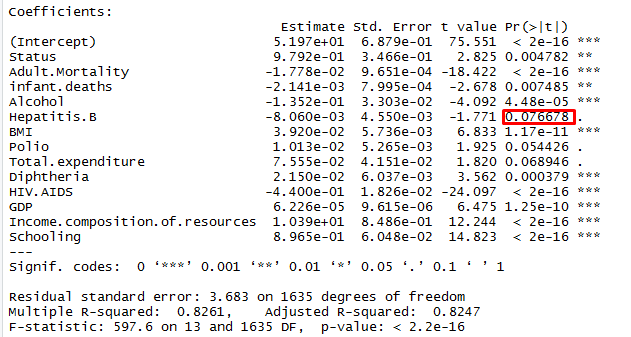
*ANOVA Statistics*



From Figure 22, thinness..1.19.years has the largest p-value, and thus, will be removed from the next regression. The summary statistics of the third regression is in Figure 24.

**Figure 24**

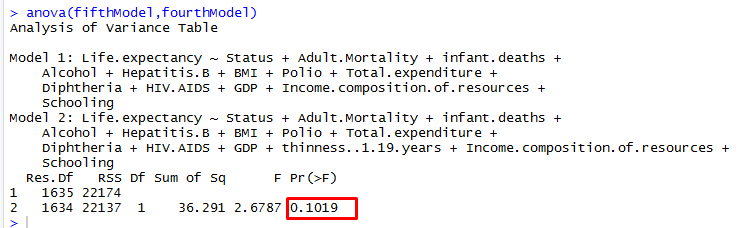
*Fifth Regression Summary Statistics*



In this case, the ANOVA statistics of these two variables (see fig. 25) show that the Null hypothesis cannot be rejected, and thus, it is necessary to continue to reduce the model.

**Figure 25**

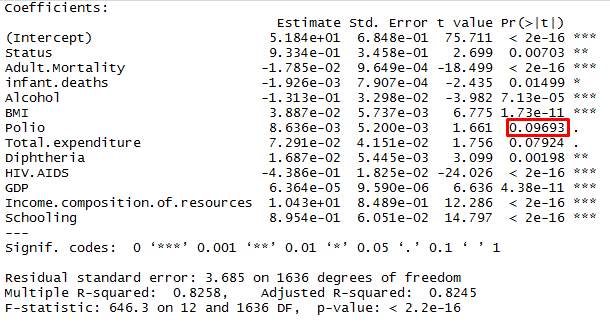
*ANOVA Statistics*



From Figure 24, Hepatitis.B has the largest p-value, and thus, will be removed from the next regression. The summary statistics of the third regression is in Figure 26.

**Figure 26**

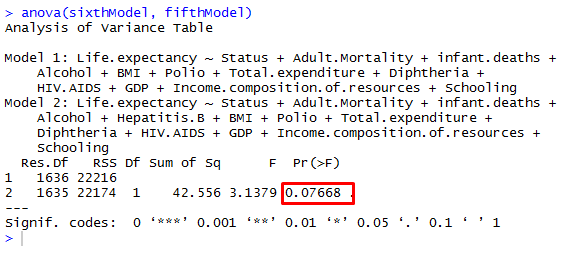
*Sixth Regression Summary Statistics*



In this case, the ANOVA statistics of these two variables (see fig. 27) show that the Null hypothesis cannot be rejected, and thus, it is necessary to continue to reduce the model.

**Figure 27**

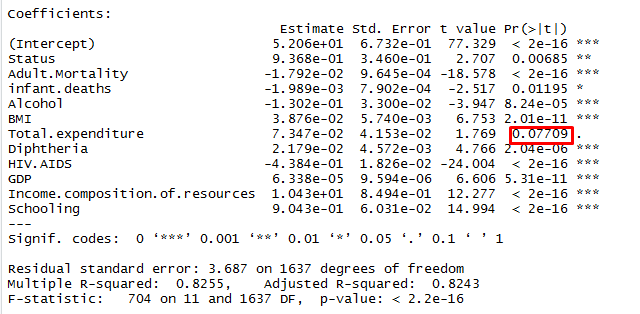
*ANOVA Statistics*



From Figure 26, Polio has the largest p-value, and thus, will be removed from the next regression. The summary statistics of the third regression is in Figure 28.

**Figure 28**

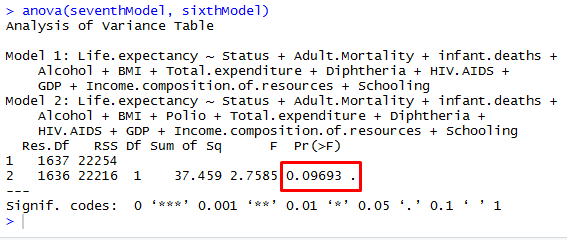
*Seventh Regression Summary Statistics*



In this case, the ANOVA statistics of these two variables (see fig. 29) show that the Null hypothesis cannot be rejected, and thus, it is necessary to continue to reduce the model.

**Figure 29**

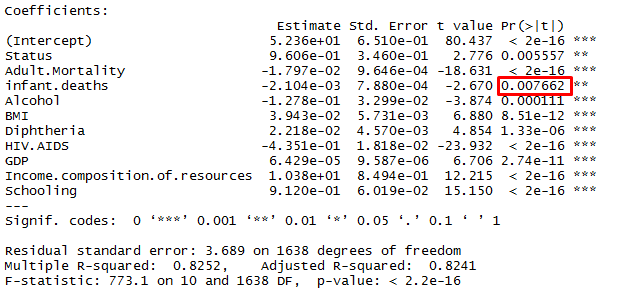
*ANOVA Statistics*



From Figure 28, Total.expenditure has the largest p-value, and thus, will be removed from the next regression. The summary statistics of the third regression is in Figure 30.

**Figure 30**

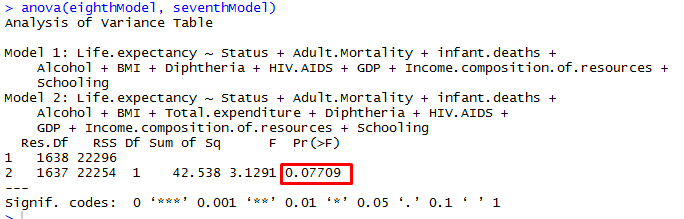
*Eighth Regression Summary Statistics*



In this case, the ANOVA statistics of these two variables (see fig. 31) show that the Null hypothesis cannot be rejected, and thus, it is necessary to continue to reduce the model.

**Figure 31**

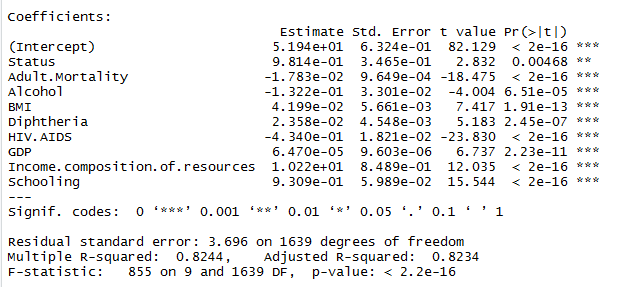
*ANOVA Statistics*



From Figure 30, infant.deaths has the largest p-value, and thus, will be removed from the next regression though it is significant. The summary statistics of the third regression is in Figure 32.

**Figure 32**

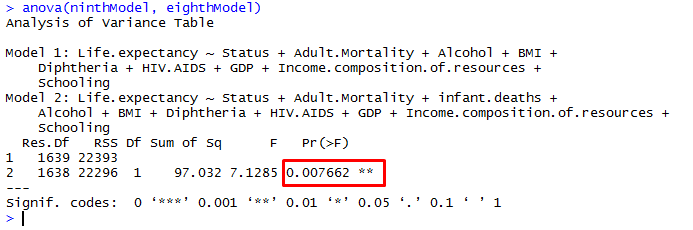
*Ninth Regression Summary Statistics*



In this case, the ANOVA statistics of these two variables (see fig. 33) show that the Null hypothesis can be rejected (p-value is less than 0.05), and thus, the final prediction model is eighth regression model, which includes such variables as Status, Adult.Mortality, infant.deaths, Alcohol, BMI, Diphtheria, HIV.AIDS, GDP, Income.composition.of.resources, Schooling. It also explains 82.52 of observations in the dataset. From Figure 30, life expectancy will be 5.236e+01 if all other variables are zero. Increasing one Status results in a 9.606e-01 increase in life expectancy if all other predictors held constant. BMI, Diphtheria, GDP, Income, Schooling also positively affect life expectancy, while all other variables such as Adult.Mortality, Infant.deaths, Alcohol, and HIV.AIDS negatively affect life expectancy.

**Figure 33**

*ANOVA Statistics*



**Conclusion**

Therefore, during this analysis, there have been learned how to draw bar charts, histograms, pie charts, correlation matrix, and how to apply them in the analysis. Also, there have been learned how to build multiple linear regressions and compare them using ANOVA summary statistics, defining the best regression model for predicting the dataset.

**Appendix A**

**Data**



**Appendix B**

**R Markdown**

---

title: "Life Expectancy Multiple Regression"

output: html\_document

---

## Installing necessary packages

```{r}

install.packages("dplyr")

install.packages("ggplot2")

```

## Loading these packages into our project

```{r}

library(dplyr)

library(ggplot2)

```

## Setting up working directory

```{r}

setwd("Documents")

```

## Importing data

```{r}

LifeExpectancyData = read.csv("Life Expectancy Data.csv", stringsAsFactors = FALSE)

```

## Determining class object of LifeExpectancyData

```{r}

class(LifeExpectancyData)

```

## Finding out the number of data entries

```{r}

count(LifeExpectancyData)

```

## Finding out the number of dimensions (variables)

```{r}

dim(LifeExpectancyData)

```

## Displaying the names of variables

```{r}

names(LifeExpectancyData)

```

## Determining the internal structure of LifeExpectancyData

```{r}

str(LifeExpectancyData)

```

## Getting a summary of all the variables

```{r}

summary <- summary(LifeExpectancyData)

```

## Finding out the number of countries and the countries itself

```{r}

unique(LifeExpectancyData$Country)

```

## Getting a glimpse of LifeExpectancyData

```{r}

glimpse(LifeExpectancyData)

```

## Finding out the number of data entries by Status

```{r}

LifeExpectancyData %>%

count(Status) %>%

arrange(n)

```

## Displaying the number of data entries by Status (Bar chart)

```{r}

ggplot(LifeExpectancyData, aes(x = Status)) +

geom\_bar(fill = "lightblue", col="grey") +

ylab("Number")

```

## Displaying the number of data entries by Status (Pie chart)

```{r}

var <- LifeExpectancyData %>%

count(Status) %>%

arrange(n)

ggplot(var, aes(x= "", y = n, fill = Status)) +

geom\_bar(width = 1, stat = "identity") +

coord\_polar("y", start=0)

```

## Finding out the number of data entries by Year

```{r}

LifeExpectancyData %>%

count(Year) %>%

arrange(Year)

```

##---- Displaying the number of data entries by Year

```{r}

ggplot(LifeExpectancyData, aes(x = Year))+

geom\_bar(fill = "lightblue",

col="grey")+

ylab("Amount")

```

## Plotting the life expectancy

```{r}

ggplot(LifeExpectancyData, aes(x=Life.expectancy)) +

geom\_histogram(bins=10, color="grey", fill="lightblue") +

ylab('Count')

```

## Plotting the life expectancy by Status Developing

```{r}

ggplot(LifeExpectancyData[LifeExpectancyData$Status == "Developing",], aes(x=Life.expectancy)) +

geom\_histogram(bins=10, color="grey", fill="lightblue") +

ylab('Count')

```

## Plotting the life expectancy by Status Developed

```{r}

ggplot(LifeExpectancyData[LifeExpectancyData$Status == "Developed",], aes(x=Life.expectancy)) +

geom\_histogram(bins=10, color="grey", fill="lightblue") +

ylab('Count')

```

## Plotting the life expectancy by Status

```{r}

ggplot(LifeExpectancyData, aes(Life.expectancy, fill = Status)) +

geom\_histogram(bins=10, alpha = 0.6, aes(y = ..density..), position = 'identity') +

ylab('Count')

```

## Finding out the mean, median, and standard deviation by Status

```{r}

LifeExpectancyData %>%

group\_by(Status) %>%

summarise(mean = mean(Life.expectancy), median = median(Life.expectancy), sd = sd(Life.expectancy)) %>%

arrange(desc(mean))

```

## Finding out the mean, median, and standard deviation by Year

```{r}

years <- LifeExpectancyData %>%

group\_by(Year) %>%

summarise(mean = mean(Life.expectancy), median = median(Life.expectancy), sd = sd(Life.expectancy))

```

## Finding out the difference between last and first mean, median, and standard deviation

```{r}

dif <- years[years$Year == 2014,] - years[years$Year == 2000,]

```

## Plotting the adult mortality

```{r}

par(mfrow=c(1,2))

hist(LifeExpectancyData$Adult.Mortality,main = "Adult Mortality", ylab = "Count", xlab = "Deaths",col = "lightblue", border = "grey")

```

## Plotting the adult mortality

```{r}

hist(LifeExpectancyData$infant.deaths,main = "Infant Deaths", ylab = "Count", xlab = "Deaths",col = "lightblue", border = "grey")

```

## Plotting the adult mortality by Status

```{r}

ggplot(LifeExpectancyData, aes(x=Adult.Mortality, fill = Status)) +

geom\_histogram(bins=10, alpha = 0.4, aes(y = ..density..), position = 'identity') +

ylab('Count')

```

## Plotting the infant deaths by Status

```{r}

ggplot(LifeExpectancyData, aes(x=infant.deaths, fill = Status)) +

geom\_histogram(bins=10, alpha = 0.4, aes(y = ..density..), position = 'identity') +

ylab('Count')

```

## Plotting other 4 variables

```{r}

par(mfrow=c(1,4))

hist(LifeExpectancyData$Alcohol,main = "Alcohol", ylab = "Count", xlab = "Alcohol consumption per capita ",col = "lightblue", border = "grey")

hist(LifeExpectancyData$percentage.expenditure,main = "Percentage Expenditure", ylab = "Count", xlab = "%",col = "lightblue", border = "grey")

hist(LifeExpectancyData$Hepatitis.B,main = "Hepatitis B", ylab = "Count", xlab = "Deaths",col = "lightblue", border = "grey")

hist(LifeExpectancyData$Measles,main = "Measles", ylab = "Count", xlab = "Deaths",col = "lightblue", border = "grey")

```

## Plotting other 4 variables

```{r}

par(mfrow=c(1,4))

hist(LifeExpectancyData$BMI,main = "BMI", xlab = "kg/m^2", ylab = "Count", col = "lightblue", border = "grey")

hist(LifeExpectancyData$under.five.deaths,main = "Under 5 deaths",xlab = "Deaths", ylab = "Count",col = "lightblue", border = "grey")

hist(LifeExpectancyData$Polio,main = "Polio", xlab = "Deaths", ylab = "Count",col = "lightblue", border = "grey")

hist(LifeExpectancyData$Total.expenditure,main = "Total expenditure", xlab = "$", ylab = "Count", col = "lightblue", border = "grey")

```

## Plotting other 4 variables

```{r}

par(mfrow=c(1,4))

hist(LifeExpectancyData$Diphtheria,main = "Diphtheria", xlab = "Deaths", ylab = "Count", col = "lightblue", border = "grey")

hist(LifeExpectancyData$HIV.AIDS,main = "HIV/AIDS",xlab = "Deaths", ylab = "Count",col = "lightblue", border = "grey")

hist(LifeExpectancyData$GDP,main = "GDP", xlab = "$", ylab = "Count",col = "lightblue", border = "grey")

hist(LifeExpectancyData$Population,main = "Population", xlab = "People", ylab = "Count", col = "lightblue", border = "grey")

```

## Plotting the last 4 variables

```{r}

par(mfrow=c(1,4))

hist(LifeExpectancyData$thinness..1.19.years,main = "Thinness..1.19", xlab = "People", ylab = "Count", col = "lightblue", border = "grey")

hist(LifeExpectancyData$thinness.5.9.years,main = "Thinness..5.9",xlab = "People", ylab = "Count",col = "lightblue", border = "grey")

hist(LifeExpectancyData$Income.composition.of.resources,main = "Income", xlab = "$", ylab = "Count",col = "lightblue", border = "grey")

hist(LifeExpectancyData$Schooling,main = "Schooling/Education", xlab = "Years", ylab = "Count", col = "lightblue", border = "grey")

```

## Removing all data entries with missing value

```{r}

LifeExpectancyData\_new <- na.omit(LifeExpectancyData)

```

## Removing two first categorical variables

```{r}

install.packages("tidyverse")

library(tidyverse)

LifeExpectancyData\_new <- select(LifeExpectancyData\_new, -Country)

LifeExpectancyData\_new <- select(LifeExpectancyData\_new, -Year)

```

## Making dummy variable out of Status

```{r}

LifeExpectancyData\_new$Status <- ifelse(LifeExpectancyData\_new$Status == "Developed", 1, 0)

```

## Checking

```{r}

str(LifeExpectancyData\_new)

```

## Correlation matrix

```{r}

corMatrix = cor(LifeExpectancyData\_new, method = "pearson")

corMatrix

install.packages("reshape2")

library(reshape2)

corMatrix <- melt(corMatrix)

ggheatmap <- ggplot(corMatrix, aes(Var2, Var1, fill = value))+

geom\_tile(color = "white")+

scale\_fill\_gradient2(low = "blue", high = "red", mid = "white",

midpoint = 0, limit = c(-1,1), space = "Lab",

name="Pearson\nCorrelation") +

theme\_minimal()+ # minimal theme

theme(axis.text.x = element\_text(angle = 90, vjust = 1,

size = 9, hjust = 1))+

coord\_fixed()

ggheatmap +

geom\_text(aes(Var2, Var1, label = round(value, digits = 2)), color = "black", size = 1.8)

```

## Removing the variables after correlation analysis

```{r}

LifeExpectancyData\_new <- select(LifeExpectancyData\_new, -percentage.expenditure)

LifeExpectancyData\_new <- select(LifeExpectancyData\_new, -under.five.deaths)

```

## Full model

```{r}

fullModel <- lm(Life.expectancy~Status+Adult.Mortality+infant.deaths+

Alcohol+Hepatitis.B+Measles+BMI+Polio+Total.expenditure+

Diphtheria+HIV.AIDS+GDP+Population+thinness..1.19.years+

thinness.5.9.years+Income.composition.of.resources+

Schooling, data = LifeExpectancyData\_new)

summary(fullModel)

```

## Model without thinness.5.9.year

```{r}

secondModel <- lm(Life.expectancy~Status+Adult.Mortality+infant.deaths+

Alcohol+Hepatitis.B+Measles+BMI+Polio+Total.expenditure+

Diphtheria+HIV.AIDS+GDP+Population+thinness..1.19.years+

Income.composition.of.resources+

Schooling, data = LifeExpectancyData\_new)

summary(secondModel)

anova(secondModel, fullModel)

```

## Model without thinness.5.9.year and Population

```{r}

thirdModel <- lm(Life.expectancy~Status+Adult.Mortality+infant.deaths+

Alcohol+Hepatitis.B+Measles+BMI+Polio+Total.expenditure+

Diphtheria+HIV.AIDS+GDP+thinness..1.19.years+

Income.composition.of.resources+

Schooling, data = LifeExpectancyData\_new)

summary(thirdModel)

anova(thirdModel, secondModel)

```

## Model without thinness.5.9.year, Population, and Measles

```{r}

fourthModel <- lm(Life.expectancy~Status+Adult.Mortality+infant.deaths+

Alcohol+Hepatitis.B+BMI+Polio+Total.expenditure+

Diphtheria+HIV.AIDS+GDP+thinness..1.19.years+

Income.composition.of.resources+

Schooling, data = LifeExpectancyData\_new)

summary(fourthModel)

anova(fourthModel, thirdModel)

```

## Model without thinness.5.9.year, Population, Measles, and thinness..1.19.years

```{r}

fifthModel <- lm(Life.expectancy~Status+Adult.Mortality+infant.deaths+

Alcohol+Hepatitis.B+BMI+Polio+Total.expenditure+

Diphtheria+HIV.AIDS+GDP+

Income.composition.of.resources+

Schooling, data = LifeExpectancyData\_new)

summary(fifthModel)

anova(fifthModel,fourthModel)

```

## Model without thinness.5.9.year, Population, Measles,thinness..1.19.years, and Hepatitis.B

```{r}

sixthModel <- lm(Life.expectancy~Status+Adult.Mortality+infant.deaths+

Alcohol+BMI+Polio+Total.expenditure+

Diphtheria+HIV.AIDS+GDP+

Income.composition.of.resources+

Schooling, data = LifeExpectancyData\_new)

summary(sixthModel)

anova(sixthModel, fifthModel)

```

## Model without thinness.5.9.year, Population, Measles,thinness..1.19.years, Hepatitis.B, and Polio

```{r}

seventhModel <- lm(Life.expectancy~Status+Adult.Mortality+infant.deaths+

Alcohol+BMI+Total.expenditure+

Diphtheria+HIV.AIDS+GDP+

Income.composition.of.resources+

Schooling, data = LifeExpectancyData\_new)

summary(seventhModel)

anova(seventhModel, sixthModel)

```

## Model without thinness.5.9.year, Population, Measles,thinness..1.19.years, Hepatitis.B, Polio, and Total.expenditure

```{r}

eighthModel <- lm(Life.expectancy~Status+Adult.Mortality+infant.deaths+

Alcohol+BMI+

Diphtheria+HIV.AIDS+GDP+

Income.composition.of.resources+

Schooling, data = LifeExpectancyData\_new)

summary(eighthModel)

anova(eighthModel, seventhModel)

```

## Model without thinness.5.9.year, Population, Measles,thinness..1.19.years, Hepatitis.B, Polio, Total.expenditure, infant.deaths

```{r}

ninthModel <- lm(Life.expectancy~Status+Adult.Mortality+

Alcohol+BMI+

Diphtheria+HIV.AIDS+GDP+

Income.composition.of.resources+

Schooling, data = LifeExpectancyData\_new)

summary(ninthModel)

anova(ninthModel, eighthModel)

```