MSDS Project 1

# Introduction

“Life expectancy at birth reflects the overall mortality level of a population. It summarizes the mortality pattern that prevails across all age groups in a given year – children and adolescents, adults, and the elderly.” 1

What factors correlate with a high life expectancy? We chose to examine a dataset from the World Health Organization (WHO) to understand the factors that correlate life expectancy. We obtained this dataset from Kaggle.2 While this data is observational, it may help us to understand which factors may leads to higher life expectancy. The data includes both health factors, health expenditure information, and country financial information.

While the data runs from 2000 to 2015, we chose to analyze the data for a single year - 2014. We wanted to use 2015 initially, but there was quite a bit of missing data for 2015 (in particular Alcohol and Percentage expenditure, and Total expenditure data was missing), so we settled on 2014. In order to analyze these factors for that data, we will use several approaches to find an optimal regression model; forward automated feature selection, step-wise automated feature select, and LASSO. We will also do a two-way ANOVA using two categorical features of the data.

# Data Description

We obtained dataset from Kaggle containing data originally sourced from the Global Health Observatory (GHO) data repository under the WHO. The dataset contains 20 predictive variables as well as the expected life expectancy for 193 countries. The data contains repeated measures for each country, once per year, from 2000-2015. The source of the data was individual per year data sets from GHO – these were aggregated into this single dataset which we downloaded.

Looking at the full dataset, from 2000 to 2015, we saw that there was quite a bit of data missing in general, but more specifically from certain years. 2014 was the most recent data for which we had most of the data present.

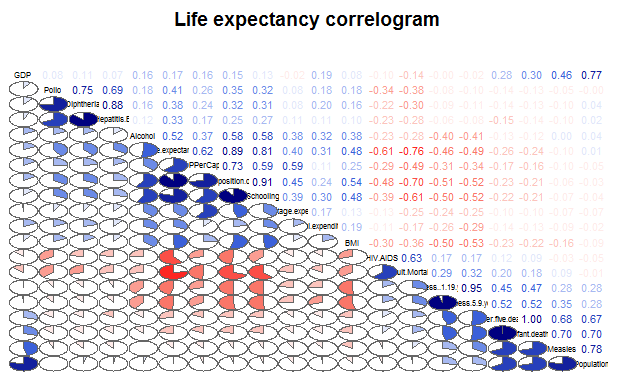
**Variable Information if for 2014 Year Only**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable Name | Description | Min | Max | Median | Mean |
| Status | Developed or Developing status | N/A | N/A | N/A | N/A |
| Life.expectancy | Life Expectancy in age | 48.10 | 89.00 | 73.60 | 71.54 |
| Adult.Mortality | Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population) | 1.0 | 522.0 | 135.0 | 148.7 |
| Infant.deaths | Number of Infant Deaths per 1000 population | 0.00 | 957.00 | 2.00 | 24.56 |
| Alcohol | Alcohol, records per capita (15+) consumption (in litres of pure alcohol) | 0.010 | 15.190 | 0.320 | 3.271 |
| Percentage.expenditure | Expenditure on health as a percentage of Gross Domestic Product per capita(%) | 0.00 | 19479.91 | 151.10 | 1001.91 |
| Hepatitis.B | Hepatitis B (HepB) immunization coverage among 1-year-olds (%) | 2.00 | 99.00 | 93.00 | 83.12 |
| Measles | Measles - number of reported cases per 1000 population | 0 | 79563 | 13 | 1831 |
| BMI | Average Body Mass Index of entire population | 2.00 | 77.10 | 47.40 | 41.03 |
| Under.five.deaths | Number of under-five deaths per 1000 population | 0.00 | 1200.00 | 3.00 | 32.89 |
| Polio | Polio (Pol3) immunization coverage among 1-year-olds (%) | 8.00 | 99.00 | 94.00 | 84.73 |
| Total.expenditure | General government expenditure on health as a percentage of total government expenditure (%) | 1.210 | 17.140 | 5.840 | 6.201 |
| Diptheria | Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%) | 2.00 | 99.00 | 94.00 | 84.08 |
| HIV.AIDS | Deaths per 1 000 live births HIV/AIDS (0-4 years) | 0.100 | 9.400 | 0.100 | 0.682 |
| GDP | Gross Domestic Product per capita (in USD) | 193 | 18228356 | 72127 | 630291 |
| Population | Population of the country | 9.136  \* | 1.364\* | 9.344\* | 4.064\* |
| thinness..1.19.years | Prevalence of thinness among children and adolescents for Age 10 to 19 (% ) | 0.100 | 26.800 | 3.300 | 4.533 |
| thinness.5.9.years | Prevalence of thinness among children for Age 5 to 9(%) | 0.100 | 27.400 | 3.400 | 4.676 |
| Income.composition.of.resources | Human Development Index in terms of income composition of resources (index ranging from 0 to 1) | 0.3450 | 0.9450 | 0.7220 | 0.6884 |
| Schooling | Number of years of Schooling(years) | 4.90 | 20.40 | 13.00 | 12.89 |
| GDPPerCapita | GDP / Population | 0.000395 | 0.122925 | 0.011721 | 0.018501 |

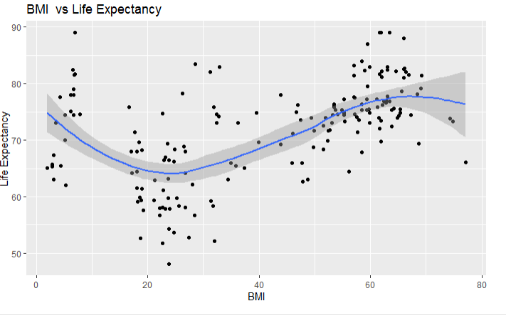
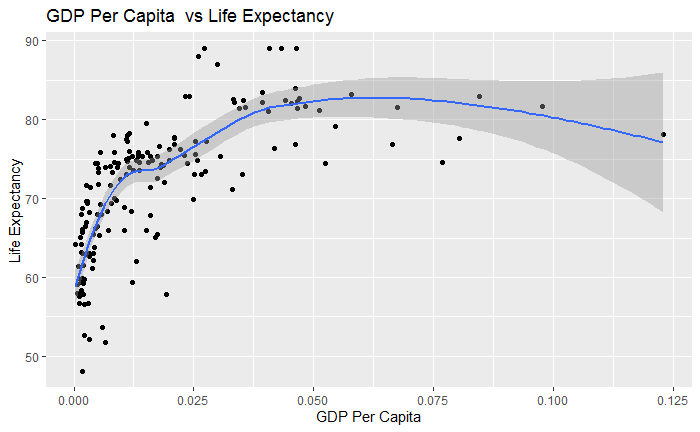
Looking at the min and max values we can see there are some values that do not make sense. The percentage.expeditur and Measles variables are all percentages, and all contain values over 100. Looking at the ranges for these same fields, we can see that they all in fact have ranges that include values over 1000 – it’s not just outliers. Under.five.deaths also is described as having a meaningful maximum of 1000 but has values above 1000. It is not clear whether this data is erroneous or if the scale is simply different than what we should expect. We chose to keep the data for analysis despite not understanding it fully. The Population also has some errors – e.g. Maldives is listed as population of 44, and it should be around 435,000. Since this data was erroneous, missing in many cases, and because it is easily available from other sources, we replaced the Population data with data from Wikipedia (IMF Estimates)3, and replaced the GDP data from data from the World Bank.4

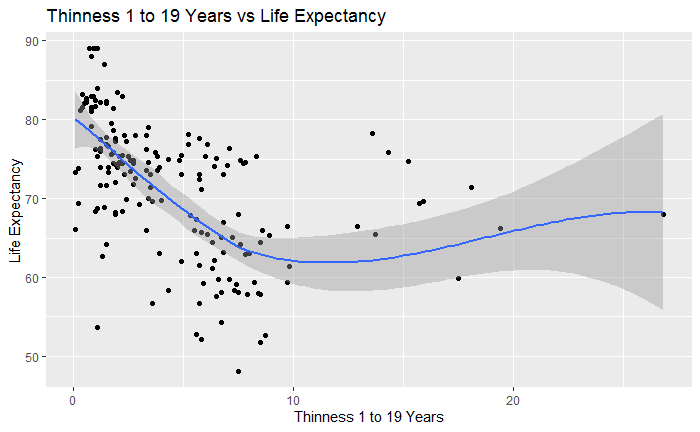
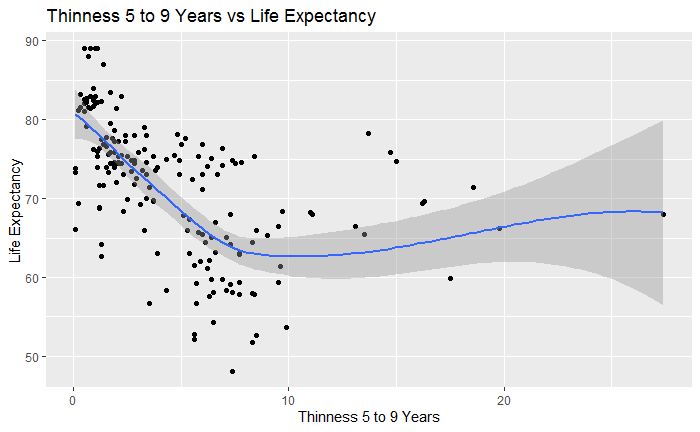
# Exploratory Analysis

Looking at correlation matrix shows strong positive correlation between life expectancy and Alcohol (0.52), Income.composition.of.resources, (0.89) Schooling (0.81), Total Expenditure (0.31), BMI (0.48). There are negative correlations between life expectancy and HIV AIDS (-0.61), Adult.Mortality (-.076), thinness 1-19 years (-0.46), thinness 5-9 years (-0.49). We also see a strong correlation between “Thinness 1-19”, “Thinness 5-9”, and “infant.deaths”. Also a strong positive correlation exists between Polio and Hepetatis vaccination levels. There is also a strong correlation between GDPPerCapita and Income.composition and Schooling.



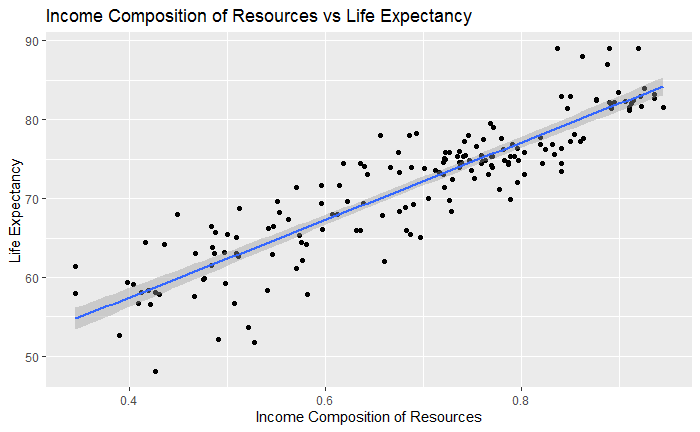
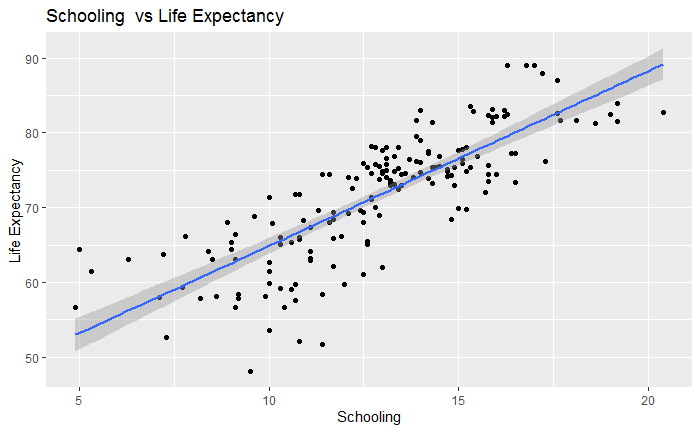
Looking at the BMI vs Life Expectancy, it looks like there may be a relationship to the square or cube of BMI. Life expectancy may also have a relationship to the square of “Thinness 5 to 9 Years” and “Thinness 1 to 19”. Adult Mortality vs. Life Expectancy appears to have a subset of the data going off on a second curve, but we could not identify a common factor among the low adult mortality/low life expectancy subset.

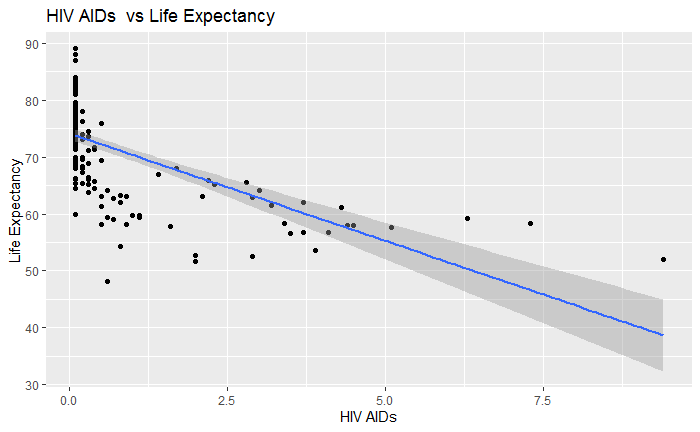
 



## 

We identified several predictive variables that have a strong correlation with Life Expectancy based on scatter plots ; Income Composition of Resources and Schooling. While Adult Mortality (noted above) and HIV Aids shows a strong negative correlation with Life Expectancy.



## Missing Data

For the year 2014, we found the following data was missing (after replacing the Population and GDP data):

|  |  |
| --- | --- |
| **Field** | **Missing Percent** |
| GDP Per Capita | 7.1 |
| GDP | 6.6 |
| Population | 3.8 |
| Hepatitis.B | 5.5 |
| Income.Composition.of.resources | 5.5 |
| Schooling | 5.5 |
| BMI | 1.1 |
| Total.Expenditure | 1.1 |
| Thinness..1.19.years | 1.1 |
| Alcohol | 0.5 |

## Missing Data Approach

We used the MICE package to impute missing values for our data.5 The MICE package works best when there is a random pattern of data missing such that the distribution of the data is not affected. Most of the fields only had < 5% missing data and we feel comfortable with the possible risk vs. being able to include the full data in our analysis.6

# Objective 1 - Regression

We would like to evaluate the relationship between the different variables and life expectancy from the chosen dataset. We also would like to accurately predict life expectancy from the current database. In order to achieve the mentioned objectives, we will evaluate and select the appropriate explanatory variables and perform a multilinear regression analysis on the life expectancy outcome.

### **Model Selection**

First, we added complexity to the BMI variable due to no linear correlation with life expectancy. We square BMI. We also log transformed thinnes.1.19, thinness.5.9.years, and GDPPerCapita to account for the presence of a particular pattern.

We run a manual selection by starting with all the explanatory variables and having life expectancy as the outcome in a regression model. After the first run of the model we found that **Adult.Mortality, Total.expenditure, HIV.AIDS, Income.composition.of.resources, GDP, and Population** was statistically significant. So, we removed the non-significant variables and kept the significant variables. We proceed to check for collinearity, and we did not have any concerning value. We re-run the model and **GDP** and **population** became non statistically significant. The final model has VIF less than 10.

Adult.Mortality Total.expenditure HIV.AIDS

2.476265 1.138418 1.669668

Income.composition.of.resources GDP Population

2.139845 2.952073 2.793680

### **Automated model selection**

We used four automated approaches to model selection (**Forward, Stepwise, backward, and LASSO**) and compared them to our manual model. More details can be found in the appendix.

For each automated model selection process, we chose the model with the lowest ASE (**in order to perform this task we used the ASE test/training plot, please see the appendix**). For the forward selection process, the model with 4 variables had lowest ASE. This model has a collinearity issue with HIV, which has a VIF of 10.479316 correlated to adult mortality. For the backward selection process, the model with 2 variables had lowest ASE, which did not have any collinearity issue. For the step-wise model selection process, a model with 3 variables has lowest ASE, which did not have any collinearity issue. The chosen models for each selection process (**Forward, Stepwise, and backward selection model)** is in the table below. See figures 5, 6, and 7 from the appendix for the Test ASE of each coefficient step from the iterative process for the Forward, Stepwise and Backward model section.

Finally, we run the lasso selection model, it gave us more variables than the manual selection, the forward selection, the backward selection, and the stepwise selection model. We chose the best lambda coefficient in order to get the variables for our lasso model. In this model we square thinness 1.19 and thinnes5.9 instead of log transform as we did for all the previous model selection. The lasso model performs better with the addition of complexity in comparison to the other models. We refit the model and we got similar results in comparison to the original model model.

|  |  |
| --- | --- |
| **Selection Algorithm** | **Model Found** |
| Manual | life.model2<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+HIV.AIDS+Income.composition.of.resources+GDP+Population, data=completeData2) |
| Forward | forward<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+HIV.AIDS+ Income.composition.of.resources, data=completeData2) |
| Backward | backward<-lm(Life.expectancy~Adult.Mortality+Income.composition.of.resources, data=completeData2) |
| Stepwise | stepwise<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+Income.composition.of.resources, data=completeData2) |
| LASSO | lasso.model<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+HIV.AIDS+Income.composition.of.resources+ BMI +Diphtheria+Population+thinness.5.9.years+I(thinness.5.9.years^2)+Schooling+log(GDPPerCapita)+Hepatitis.B+I(BMI^2)+I(BMI^3)+Polio+GDP+thinness..1.19.years+I(thinness..1.19.years^2), data=completeData2) |

The lasso model has many parameters, of which many are not statistically significant. Adult mortality, total expenditure and income composition of resources is present in all the model and are statistically significant. The manual model has GDP and population, but those variables became non statistically significant after the final run of the model.

### Checking Assumptions for Stepwise model:

A screenshot of a map

Description automatically generated

A close up of a map

Description automatically generated A close up of a map

Description automatically generated

We reviewed the plots for assumptions and found:

1. The normality assumption is met as per the qqplot
2. The variances assumption is met as per the residual fit plot
3. Independence is met as per the residual fit plot, which looks random, and the way that the values are collected makes us assume with caution that the residuals are independent.

We found the following outliers: 1-) 40 = Côte d'Ivoire. 2-)144 = Sierra Leone

3-) Equatorial Guinea = 75.

But this outlier does not seen to be a mistake, except the 144. But the leverage and cook value are less than 1. Therefore, we decided to continue without taken this value out.

### Compare Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Selection Algorithm** | **AIC** | **BIC** | **ADJ R2** | **ASE** |
| OLS | 663.4621 | 685.7621 | 0.825542 | 15.8096 |
| Forward | 991.8287 | 1011.086 | 0.8262806 | 12.16686 |
| Backward | 1014.394 | 1027.232 | 0.8013724 | 12.14428 |
| Stepwise | 1008.63 | 1024.678 | 0.8085589 | 12.14428 |
| LASSO | 983.8228 | 1048.013 | 0.8451271 | 8.922948 |

We decided to use the stepwise model to describe the relationship between the outcome which is life expectancy and the explanatory variables that are statistically significantly correlated with the outcome. The forward model has better AIC and BIC in comparison to the stepwise model, but there is some collinearity, which makes the description of the mentioned relationship more complicated. Backward model has slightly worse AIC and BIC in comparison to the stepwise model and in our view was missing an important variable that is related to life expectancy, reason why we did not choose as the final model. The manual model definitely has better AIC and BIC, but it has worse ASE in comparison to the stepwise model, and also it has non-significant variables. The stepwise model is well balanced from the bias/variance’s standpoint. The stepwise model’s bias is low due to having enough variables and avoid high variance that would occur with too many variables. The stepwise model does not have multicollinearity, therefore, allowing us to have a clear interpretation of the regression model, which led us it as our final model. Also, the variables that were present in the stepwise model were present in all the other model and were significant in all the other model. The R-square is close among all the model.

We propose to use the LASSO model for prediction, because the lasso model has the best ASE among all the model, ASE of 8.922948. This model is more complicated for describing the relationship between predictive variables and life expectancy because the variables have high multicollinearity and in some cases lack significance. However, multicollinearity is not an issue when creating a predictive model. The refit of the model did not improve the performance from the descriptive prospective, it gives us the same ASE, therefore we will the LASSO model as a predictive model, and the stepwise model as the descriptive model.

We also performed the 10K cross validation on the forward, backward, and stepwise model; it did not change the outcome in comparison to the previous model selection. Please see the code in appendix.

The graphic for the BIC, RSS, adjusted R-square, and ASE can find in the appendices.

**Parameter Interpretation – Descriptive Model**

The final descriptive model was the stepwise model as mentioned above. We check for interaction between the total expenditure and income composition of resources, but it was not statistically significant, we also check the interaction between adult mortality and income composition of resources, which was not significant as well. Adult mortality and total expenditure were slightly significant but have multicollinearity. Therefore, our final model was the following model without interaction, due to easy explicability:

**stepwise<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+Income.composition.of.resources, data=completeData2)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 48.535398 | 2.198599 | 22.076 | < 0.05 |
| Adult.Mortality | -0.023995 | 0.003651 | -6.572 | <0.05 |
| Total.expenditure | 0.293395 | 0.105343 | 2.785 | <0.05 |
| Income.composition.of.resources | 35.833474 | 2.586263 | 13.855 | <0.05 |

**Confidence intervals**

|  |  |  |
| --- | --- | --- |
| Coefficient | 2.5% | 97.5% |
| (Intercept) | 44.19689011 | 52.87390650 |
| Adult.Mortality | -0.03120021 | -0.01678983 |
| Total.expenditure | 0.08552096 | 0.50126898 |
| Income.composition.of.resources | 30.72998740 | 40.93696152 |

1-For any unit increment in adult mortality, the life expectancy decreased by 0.023995 years, p value less than 0.05, with a 95% confidence interval (0.03120021 -- 0.01678983).

2-For any unit increment in total expenditure, the life expectancy increased by 0.293395 years, p value less than 0.05, with a 95% confidence interval (0.08552096 -- 0.50126898).

3- For any unit increment in total income composition of resources, the life expectancy increased by 35.833474 years, p value less than 0.05, with a 95% confidence interval (30.72998740 -- 40.93696152).

4-When the adult mortality, the total expenditure, the HIV, and the income composition of resources are negligible the life expectancy is 44.19689011, p value less than 0.05, with a 95% confident interval (44.19689011 -- 40.93696152).

**Final conclusions from the analyses of Objective**

After evaluating all the 20 variables from the dataset, adult mortality, total expenditure, and income composition of resources are statistically significantly correlated with life expectancy. Adult mortality impacts negatively life expectancy, whereas total expenditure and income composition of resources impact positively life expectancy. Income composition of resources seen to impact the most on life expectancy after controlling for adult mortality, and total expenditure. This was purely an observational study, where different countries were not randomly assigned different country attributes, so we cannot draw causal relationships between the attributes and life expectancy. As a matter of public policy, the message to groups working on improving life expectancy would be to focus on research related to income composition of resources.

#### Objective 2: 2-Way ANOVA

# Objective

For the second objective, we ran a 2-way ANOVA on the data using Status and Total expenditure as the categorical variables. Status has two levels Developing and Developed. Total expenditure is continuous variable, which is converted to 3 levels, Total expenditure from 0 to 6 as 1st level, 7 to 12 as 2nd level and 13 to 18 as 3rd level. Below box plot shows Developed countries has higher life expectancy than developing countries. The assumption of homoscedasticity (Figure 1) is not met with original data so we log transform the life expectancy (Figure 2).

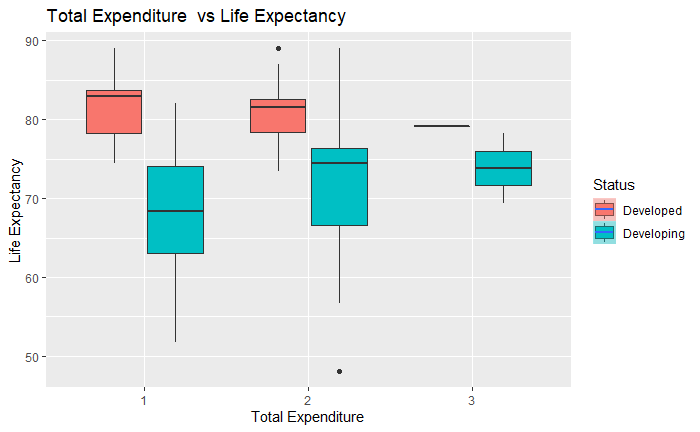


Figure 1: Total Expenditure vs Life Expectancy

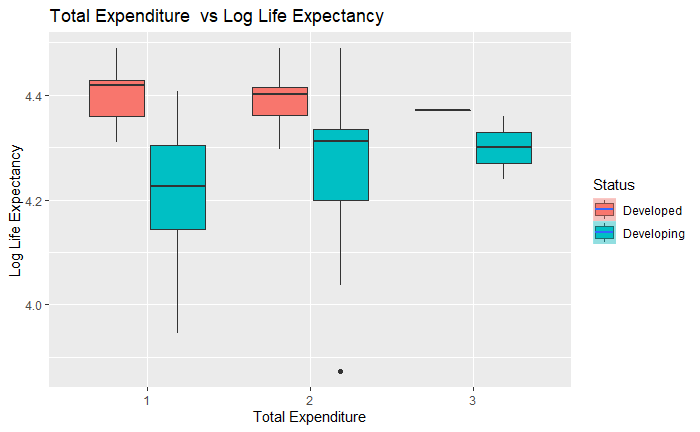


Figure 2: Total Expenditure vs Log Life Expectancy

With the two levels of Status and three levels of Total Expenditure Level, below Table 1 provides the sample size, mean, standard deviation, minimum and maximum for each of the 6 combinations of the two factors combined. Developed countries with Total expenditure level 3 has only one observation.

| **Total.expenditurelevel** <fctr> | **Status**  <fctr> | **N**  <dbl> | **Mean**  <dbl> | **SD**  <dbl> | **Minimum**  <dbl> | **Maximum**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Developed | 7 | 81.47143 | *5.224849* | 74.4 | 89.0 |
| 2 | Developed | 24 | 81.12500 | 4.001223 | 73.4 | 89.0 |
| 3 | Developed | 1 | 79.10000 | NA | 79.1 | 79.1 |
| 1 | Developing | 90 | 67.76000 | 7.194611 | 51.7 | 82.0 |
| 2 | Developing | 59 | 72.01356 | 8.198895 | 48.1 | 89.0 |
| 3 | Developing | 2 | 73.80000 | 6.222540 | 69.4 | 78.2 |

Table 1: Means for Status and total expenditure

*General Model:*

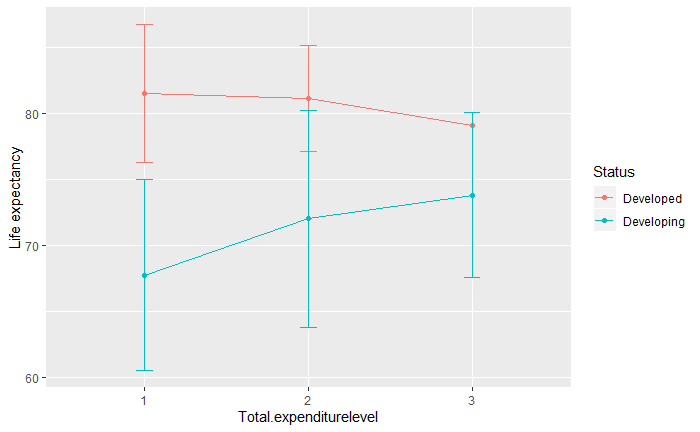


Figure 3: Interaction Plot

Looking at the interaction plot above (Figure 3), it’s not entirely clear if the data conforms to an additive model. Total expenditure Levels that are higher tending to stay lower for developed countries and Total expenditure Levels that are lower tending to stay higher for developing countries, but there doesn’t seem to be some cross-over. The overall model has a p-value of less than significant level (0.05), meaning that the model as a whole is significant that suggest there is at least one difference in the means for the given factor of interest that the F test is being reported for.

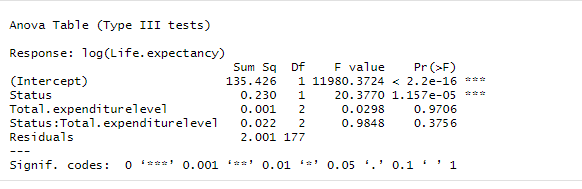


Table 2: Type III SS table

Looking at the Type III SS table, we can see that the interaction term between status and Total expenditure level is not significant. This means we fail to reject the null hypothesis that the interaction term is not equal to zero, and also means and that we have an additive model. The main effects S*tatus is significant and Total expenditure Level is not* significant. Next, we verified the assumptions for a two-way ANOVA and ran multiple comparisons between the different levels of Total expenditure *and Status* to see which were significant.

*Assumptions:*  
The assumptions for a two-way ANOVA are independence, constant variance, and normality.  The residual plots for the ANOVA are shown in below figure 4. The qq-plot does not indicate that there are any normality violations. There may be different variations between the groups, as seen in Figure 2. Also, the group for Total.expenditurelevel = 3 and Status = Developed may not have enough data for analysis. For this analysis, we will continue with the assumption that there are constant variances and that the data is independent.

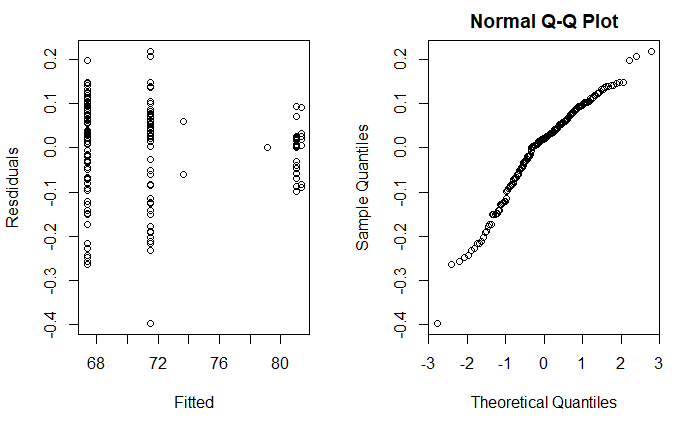
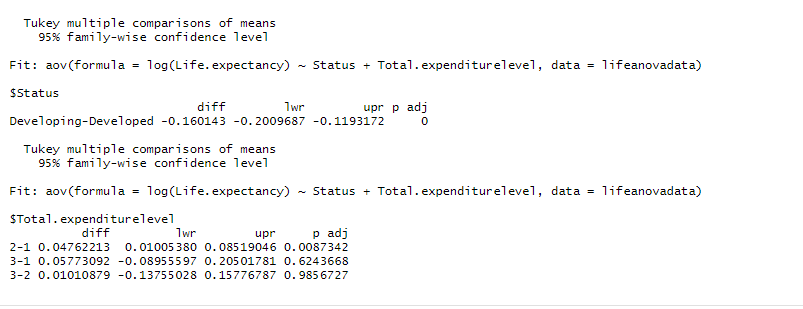


Figure 4: Fit diagnostics

*Pairwise Analysis:*

Since the model is additive, we did not need to analyze the interaction term between Status and Total expenditure Level. We focused our pairwise analysis on the main effect of Total expenditure *Level and Status*. We ran pairwise analysis of all of the combinations in status and Total expenditure Level with a Bonferroni adjustment. The results are shown below.



Life expectancy in developing countries is significantly different from Developed countries. Mean Life expectancy for Total expenditure Level 2 is significant from Mean Life expectancy for Total expenditure Level 1 with 95% confidence interval (0.0100538, 0.08519046), whereas all other combination of total expenditure levels is not significant.

*Conclusion:*

From two-way ANOVA, at a significance level of 0.05, we see that both Status and Total expenditure Levels are significant p-values. Developed countries have higher life expectancy than developing countries and there is significant difference in mean life expectancy between total expenditure level 2 and total expenditure level1. Access to health services and also proper nutrition is significantly higher in developed countries which correlates with higher life expectancy. This was purely an observational study, where different countries were not randomly assigned different country attributes, so we cannot draw causal relationships between the attributes and life expectancy.

# Appendix

## Reference

1. <https://www.who.int/gho/mortality_burden_disease/life_tables/situation_trends_text/en/>  
   Life expectancy definition
2. Kaggle.com   
   <https://www.kaggle.com/kumarajarshi/life-expectancy-who>
3. Wikipedia list of countries by past and projected GDP (IMF Estimates between 2010 and 2015)  
    <https://en.wikipedia.org/wiki/List_of_countries_by_past_and_projected_GDP_(nominal)#IMF_estimates_between_2010_and_2019>
4. Country Population – Work Bank  
   <https://data.worldbank.org/indicator/SP.POP.TOTL>
5. Mice Package for Imputing Missing Data  
   <https://cran.r-project.org/web/packages/mice/mice.pdf>
6. Mice Package Discussions  
   <https://datascienceplus.com/imputing-missing-data-with-r-mice-package/>

Figure 5 - Forward Automated Model Selection Iterations

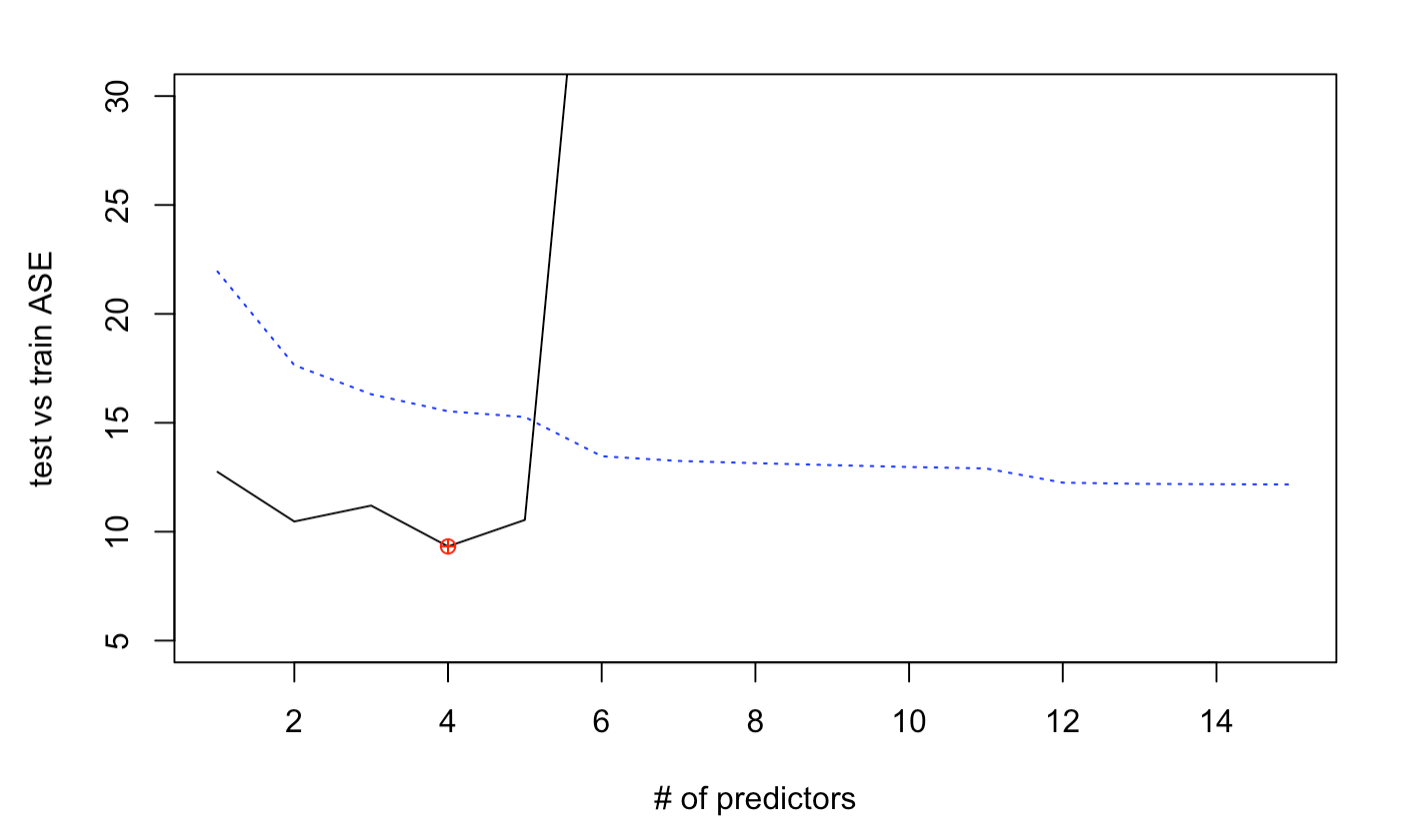


Figure 6 - Backward Automated Model Selection Steps

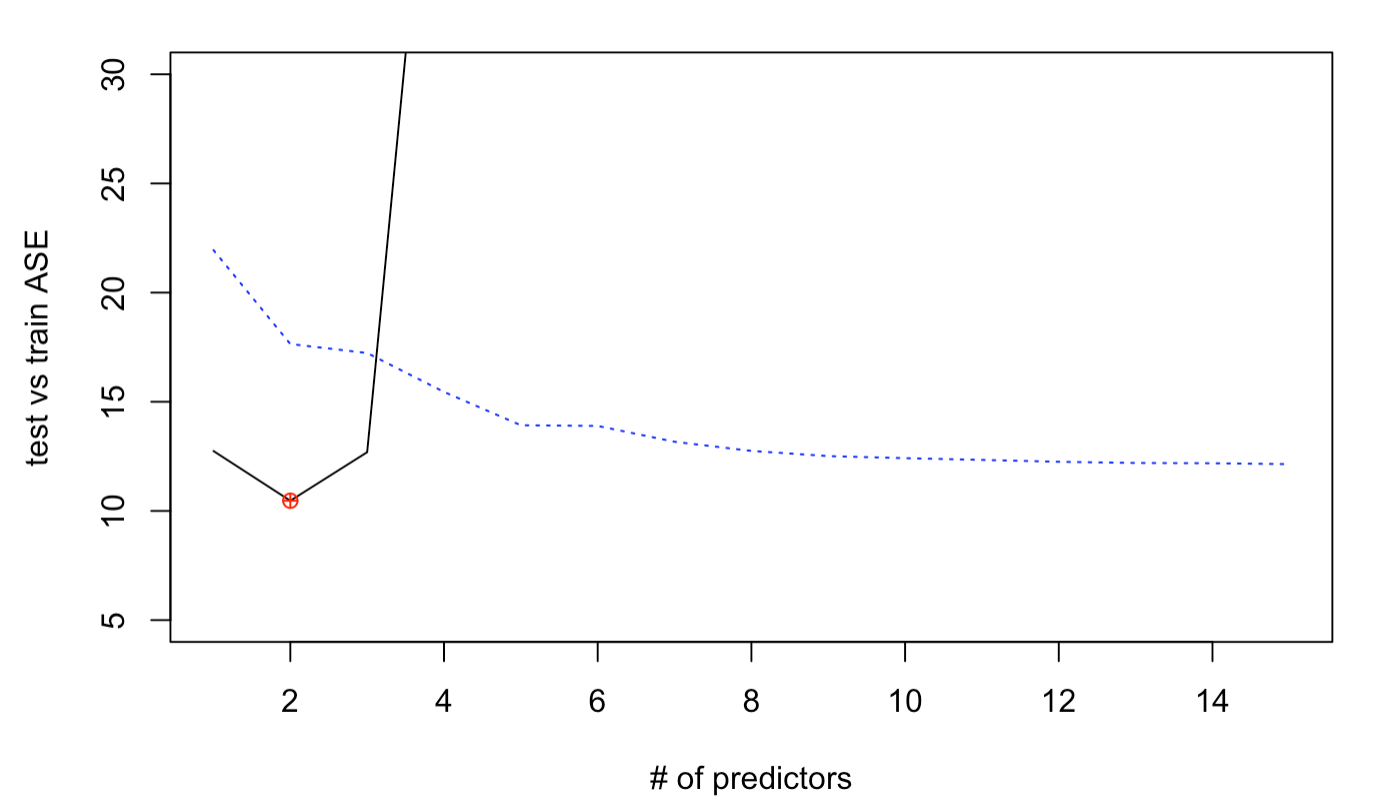
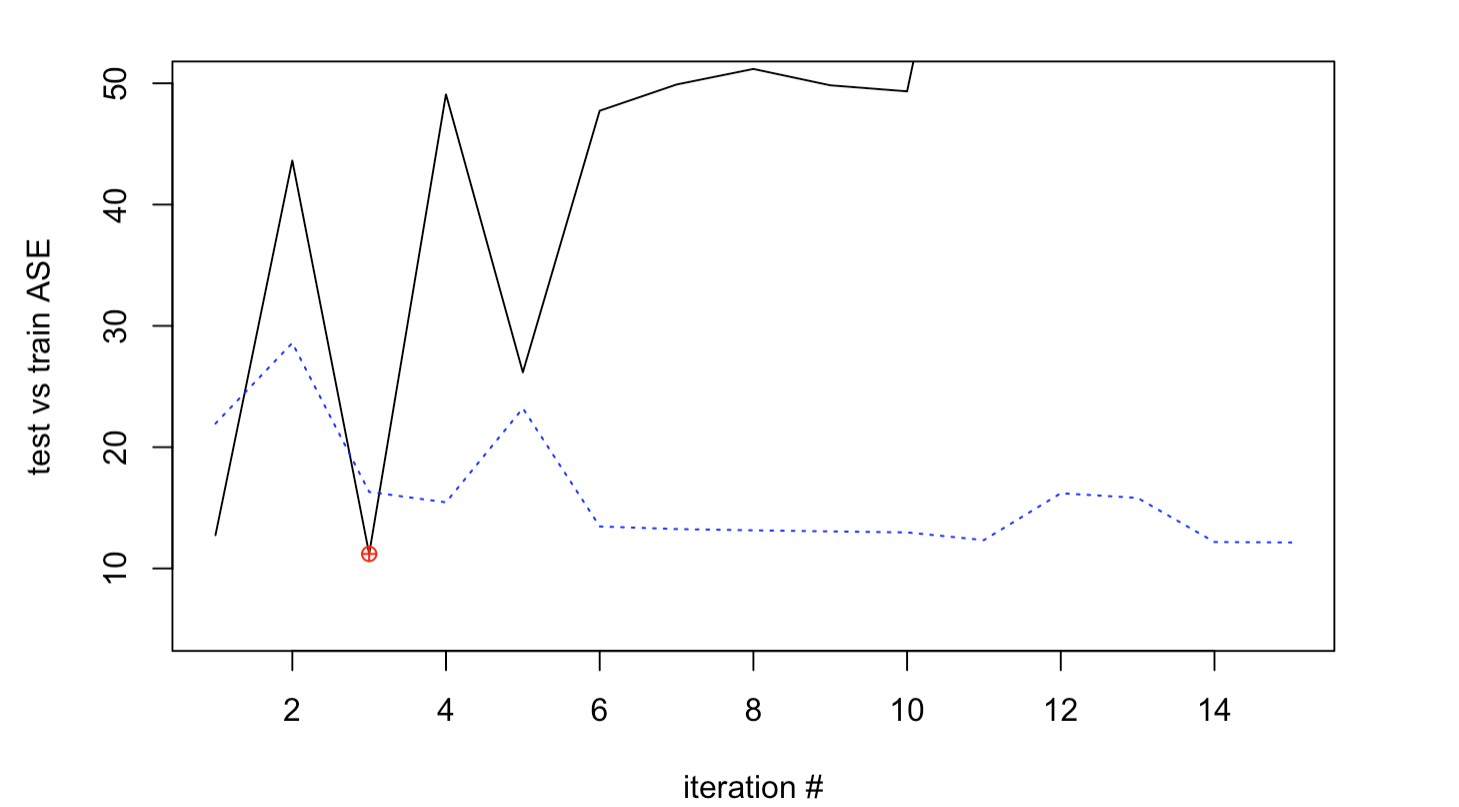


Figure 7 - Stepwise Automated Model Selection Steps



# R Language Source Code:

---

title: "Project\_1"

author: "Fabio\_Savorgnan"

date: "1/30/2020"

output: word\_document

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

\_\_Library

```{r }

library(dplyr)

library(ggplot2)

library(tidyr)

library(SDMTools)

library(readr)

library(digest)

library(ISLR)

library(car)

library(leaps)

library( Matrix)

library(foreach)

library(glmnet)

library(VIM)

library(mice)

library(corrgram)

library(car)

library(tidyverse)

library(plyr)

library(limma)

library(gridExtra)

```

## Load the data

```{r }

#df<- read.csv("Life Expectancy Data 2.csv")

df<- read.csv(file.choose())

head(df)

```

```{r }

summary(df)

dim(df)

```

# Plot the data to perform DEA

```{r visual1, fig.height=2}

lifedata = subset(df,Year==2014)

# remove GDP and pop from the dataset since we are going to pull it in from external file below

lifedata = lifedata %>% select(-c(GDP, Population))

GDPPopulation = read.csv(file.choose())

lifedata = merge(lifedata, GDPPopulation, by.x = "Country", by.y = "Country", all.x = TRUE)

lifedata$GDPPerCapita<-lifedata$GDP/lifedata$Population

```

# check for missing . So we see below for 2015, we are missing data for several columns for several countries. This will add a challenge to our analysis

```{r}

nrow(lifedata)

sapply(lifedata, function(x) sum(is.na(x)))

```

# repeat analysis for 2014 data

```{r }

summary(lifedata)

dim(lifedata)

```

# let's add another cateogorical varible for ANOVA analysis

```{r}

#Add new categorical variable "SchoolingLevel"

lifedata$SchoolingLevel<-cut(lifedata$Schooling, seq(4,24,5), right=FALSE, labels=c("1","2", "3", "4"))

```

# plot life expactancy itself

```{r}

ggplot(data=lifedata, aes(lifedata$Life.expectancy)) +

geom\_histogram(binwidth = 1) + ggtitle("Histogram of Life Expectancy") + xlab("Life Expectancy")

```

# compare categorical variable vs life expectancy

```{r}

# histogram for life expectancy by status

mu <- ddply(lifedata, "Status", summarise, grp.mean=mean(lifedata$Life.expectancy))

#head(mu)

p <- ggplot(lifedata, aes(x=lifedata$Life.expectancy, fill=lifedata$Status, color=lifedata$Status)) +

geom\_histogram(position="identity")

# Use semi-transparent fill

p <- ggplot(lifedata, aes(x=lifedata$Life.expectancy, fill=lifedata$Status, color=lifedata$Status)) +

geom\_histogram(position="identity", alpha=0.5)

# Add mean lines

p <- p+geom\_vline(data=mu, aes(xintercept=grp.mean, color=Status),

linetype="dashed")

p

t(aggregate(lifedata$Life.expectancy~lifedata$Status,data=lifedata,summary))

```

#historgram vs school Level

```{r}

# histogram for life expectancy by status

mu <- ddply(lifedata, "SchoolingLevel", summarise, grp.mean=mean(lifedata$Life.expectancy))

#head(mu)

p <- ggplot(lifedata, aes(x=lifedata$Life.expectancy, fill=lifedata$SchoolingLevel, color=lifedata$SchoolingLevel)) +

geom\_histogram(position="identity")

# Use semi-transparent fill

p <- ggplot(lifedata, aes(x=lifedata$Life.expectancy, fill=lifedata$SchoolingLevel, color=lifedata$SchoolingLevel)) +

geom\_histogram(position="identity", alpha=0.5)

# Add mean lines

p <- p+geom\_vline(data=mu, aes(xintercept=grp.mean, color=SchoolingLevel),

linetype="dashed")

p

t(aggregate(lifedata$Life.expectancy~lifedata$SchoolingLevel,data=lifedata,summary))

```

# scatter plots of life expectancy vs. various data

```{r}

ggplot(lifedata, aes(x=lifedata$Adult.Mortality, y=lifedata$Life.expectancy)) + geom\_point(data=lifedata) + geom\_smooth(method=lm) +

xlab("Adult Mortality") + ylab("Life Expectancy") + ggtitle("Adult Mortality vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$infant.deaths, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Infact Deaths") + ylab("Life Expectancy") + ggtitle("Infant Deaths vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Alcohol, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Alcohol") + ylab("Life Expectancy") + ggtitle("Alcohol vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Hepatitis.B, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Hebapitis B") + ylab("Life Expectancy") + ggtitle("Hepatitis B vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Measles, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Measles") + ylab("Life Expectancy") + ggtitle("Measles vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$BMI, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("BMI") + ylab("Life Expectancy") + ggtitle("BMI vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$BMI, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=loess) +

xlab("BMI") + ylab("Life Expectancy") + ggtitle("BMI vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$under.five.deaths, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Under Five Deaths") + ylab("Life Expectancy") + ggtitle("Under Five Deaths vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Polio, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Polio") + ylab("Life Expectancy") + ggtitle("Polio vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Total.expenditure, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Total Expenditure") + ylab("Life Expectancy") + ggtitle("Total Expenditure vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Diphtheria, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Diptheria") + ylab("Life Expectancy") + ggtitle("Diptheria vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$HIV.AIDS, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("HIV AIDs") + ylab("Life Expectancy") + ggtitle("HIV AIDs vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$GDP, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("GDP") + ylab("Life Expectancy") + ggtitle("GDP vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Population, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Population") + ylab("Life Expectancy") + ggtitle("Population vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$thinness..1.19.years, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Thinness 1 to 19 Years") + ylab("Life Expectancy") + ggtitle("Thinness 1 to 19 Years vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$thinness..1.19.years, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=loess) +

xlab("Thinness 1 to 19 Years") + ylab("Life Expectancy") + ggtitle("Thinness 1 to 19 Years vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$thinness.5.9.years, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Thinness 5 to 9 Years") + ylab("Life Expectancy") + ggtitle("Thinness 5 to 9 Years vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$thinness.5.9.years, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=loess) +

xlab("Thinness 5 to 9 Years") + ylab("Life Expectancy") + ggtitle("Thinness 5 to 9 Years vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Income.composition.of.resources, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Income Composition of Resources") + ylab("Life Expectancy") + ggtitle("Income Composition of Resources vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Schooling, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Schooling") + ylab("Life Expectancy") + ggtitle("Schooling vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$percentage.expenditure, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Percentage Expenditure") + ylab("Life Expectancy") + ggtitle("Percentage Expenditure vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$GDPPerCapita, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=lm) +

xlab("GDP Per Capita") + ylab("Life Expectancy") + ggtitle("GDP Per Capita vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$GDPPerCapita, y=lifedata$Life.expectancy)) + geom\_point() + geom\_smooth(method=loess) +

xlab("GDP Per Capita") + ylab("Life Expectancy") + ggtitle("GDP Per Capita vs Life Expectancy")

```

# redo the plots, but this time include the SchoolingLevel as a grouping/color

```{r}

ggplot(lifedata, aes(x=lifedata$Adult.Mortality, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point(data=lifedata) + geom\_smooth(method=lm) +

xlab("Adult Mortality") + ylab("Life Expectancy") + ggtitle("Adult Mortality vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$infant.deaths, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Infact Deaths") + ylab("Life Expectancy") + ggtitle("Infant Deaths vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Alcohol, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Alcohol") + ylab("Life Expectancy") + ggtitle("Alcohol vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Hepatitis.B, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Hebapitis B") + ylab("Life Expectancy") + ggtitle("Hepatitis B vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Measles, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Measles") + ylab("Life Expectancy") + ggtitle("Measles vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$BMI, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("BMI") + ylab("Life Expectancy") + ggtitle("BMI vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$BMI, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=loess) +

xlab("BMI") + ylab("Life Expectancy") + ggtitle("BMI vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$under.five.deaths, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Under Five Deaths") + ylab("Life Expectancy") + ggtitle("Under Five Deaths vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Polio, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Polio") + ylab("Life Expectancy") + ggtitle("Polio vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Total.expenditure, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Total Expenditure") + ylab("Life Expectancy") + ggtitle("Total Expenditure vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Diphtheria, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Diptheria") + ylab("Life Expectancy") + ggtitle("Diptheria vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$HIV.AIDS, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("HIV AIDs") + ylab("Life Expectancy") + ggtitle("HIV AIDs vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$GDP, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("GDP") + ylab("Life Expectancy") + ggtitle("GDP vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Population, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Population") + ylab("Life Expectancy") + ggtitle("Population vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$thinness..1.19.years, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Thinness 1 to 19 Years") + ylab("Life Expectancy") + ggtitle("Thinness 1 to 19 Years vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$thinness..1.19.years, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=loess) +

xlab("Thinness 1 to 19 Years") + ylab("Life Expectancy") + ggtitle("Thinness 1 to 19 Years vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$thinness.5.9.years, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Thinness 5 to 9 Years") + ylab("Life Expectancy") + ggtitle("Thinness 5 to 9 Years vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$thinness.5.9.years, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=loess) +

xlab("Thinness 5 to 9 Years") + ylab("Life Expectancy") + ggtitle("Thinness 5 to 9 Years vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Income.composition.of.resources, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Income Composition of Resources") + ylab("Life Expectancy") + ggtitle("Income Composition of Resources vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Schooling, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Schooling") + ylab("Life Expectancy") + ggtitle("Schooling vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$percentage.expenditure, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Percentage Expenditure") + ylab("Life Expectancy") + ggtitle("Percentage Expenditure vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$GDPPerCapita, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=lm) +

xlab("GDP Per Capita") + ylab("Life Expectancy") + ggtitle("GDP Per Capita vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$GDPPerCapita, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point() + geom\_smooth(method=loess) +

xlab("GDP Per Capita") + ylab("Life Expectancy") + ggtitle("GDP Per Capita vs Life Expectancy")

```

# what countries are on this separate curve on life expectancy - very low adult mortality, and low life expectancy

```{r}

p <- ggplot(lifedata, aes(x=lifedata$Adult.Mortality, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point(data=lifedata) + geom\_smooth(method=lm) +

xlab("Adult Mortality") + ylab("Life Expectancy") + ggtitle("Adult Mortality vs Life Expectancy") +

geom\_label(

label=lifedata$Country,

nudge\_x = 0.25, nudge\_y = 0.25,

check\_overlap = T

)

p

ggsave("Mortality vs Life Expectancy.jpg", plot = p, width = 20, height = 20, units = "in")

p <- ggplot(lifedata, aes(x=lifedata$thinness..1.19.years, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=loess) +

xlab("Thinness 1 to 19 Years") + ylab("Life Expectancy") + ggtitle("Thinness 1 to 19 Years vs Life Expectancy")+

geom\_label(

label=lifedata$Country,

nudge\_x = 0.25, nudge\_y = 0.25,

check\_overlap = T

)

p

ggsave("Thinness1.19 vs Life Expectancy.jpg", plot = p, width = 20, height = 20, units = "in")

```

# redo the plots, but this time include the Status (developed, developing) as a grouping/color

```{r}

ggplot(lifedata, aes(x=lifedata$Adult.Mortality, y=lifedata$Life.expectancy, color=lifedata$SchoolingLevel)) + geom\_point(data=lifedata) + geom\_smooth(method=lm) +

xlab("Adult Mortality") + ylab("Life Expectancy") + ggtitle("Adult Mortality vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$infant.deaths, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Infact Deaths") + ylab("Life Expectancy") + ggtitle("Infant Deaths vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Alcohol, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Alcohol") + ylab("Life Expectancy") + ggtitle("Alcohol vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Hepatitis.B, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Hebapitis B") + ylab("Life Expectancy") + ggtitle("Hepatitis B vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Measles, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Measles") + ylab("Life Expectancy") + ggtitle("Measles vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$BMI, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("BMI") + ylab("Life Expectancy") + ggtitle("BMI vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$BMI, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=loess) +

xlab("BMI") + ylab("Life Expectancy") + ggtitle("BMI vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$under.five.deaths, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Under Five Deaths") + ylab("Life Expectancy") + ggtitle("Under Five Deaths vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Polio, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Polio") + ylab("Life Expectancy") + ggtitle("Polio vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Total.expenditure, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Total Expenditure") + ylab("Life Expectancy") + ggtitle("Total Expenditure vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Diphtheria, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Diptheria") + ylab("Life Expectancy") + ggtitle("Diptheria vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$HIV.AIDS, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("HIV AIDs") + ylab("Life Expectancy") + ggtitle("HIV AIDs vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$GDP, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("GDP") + ylab("Life Expectancy") + ggtitle("GDP vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Population, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Population") + ylab("Life Expectancy") + ggtitle("Population vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$thinness..1.19.years, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Thinness 1 to 19 Years") + ylab("Life Expectancy") + ggtitle("Thinness 1 to 19 Years vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$thinness..1.19.years, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=loess) +

xlab("Thinness 1 to 19 Years") + ylab("Life Expectancy") + ggtitle("Thinness 1 to 19 Years vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$thinness.5.9.years, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Thinness 5 to 9 Years") + ylab("Life Expectancy") + ggtitle("Thinness 5 to 9 Years vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$thinness.5.9.years, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=loess) +

xlab("Thinness 5 to 9 Years") + ylab("Life Expectancy") + ggtitle("Thinness 5 to 9 Years vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Income.composition.of.resources, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Income Composition of Resources") + ylab("Life Expectancy") + ggtitle("Income Composition of Resources vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$Schooling, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Schooling") + ylab("Life Expectancy") + ggtitle("Schooling vs Life Expectancy")

ggplot(lifedata, aes(x=lifedata$percentage.expenditure, y=lifedata$Life.expectancy, color=lifedata$Status)) + geom\_point() + geom\_smooth(method=lm) +

xlab("Percentage Expenditure") + ylab("Life Expectancy") + ggtitle("Percentage Expenditure vs Life Expectancy")

```

```{r }

#df\_filtered <- df %>% filter(Year == 2015 | Year == 2014)

#head(df\_filtered)

```

```{r}

pairs(lifedata[,c(4,5,6,7,8,9,10,11,12,13)])

pairs(lifedata[,c(4,14,15,16,17,18,19,20,21,22)])

```

```{r }

pairs(lifedata[,c(4,5,6,7,8)])

pairs(lifedata[,c(4,9,10,11,12,13)])

pairs(lifedata[,c(4,14,15,16,17,18)])

pairs(lifedata[,c(4,19,20,21,22)])

```

```{r }

pairs(lifedata[,-c(1,2,3)])

lifedata

pairs(lifedata[,c(4,5,6,7,8,9,10)], col=lifedata$Status)

pairs(lifedata[,c(4,11,12,13,14,15,16,17)], col=lifedata$Status)

pairs(lifedata[,c(4,18,19,20,21,22)], col=lifedata$Status)

```

# Clean the data

```{r }

#lifedata = subset(df,Year==2014)

# add this so we can impute the missing values

lifedata$GDPPerCapita<-lifedata$GDP/lifedata$Population

#lifedata <-df

# Verify any missing values in any variable

lifedata[!complete.cases(lifedata), ]

#Display missing-data patterns

md.pattern(lifedata, plot=TRUE, rotate.names = TRUE)

#Display missing-data in a bar-plot

mice\_plot <- aggr(lifedata, col=c('navyblue','yellow'),

numbers=TRUE, sortVars=TRUE,

labels=names(lifedata), cex.axis=.7,

gap=3, ylab=c("Missing data","Pattern"))

#Data frame with all continuous variables

#lifedata1 = lifedata %>% select(c(Country, Status, Life.expectancy, Adult.Mortality, Alcohol, Hepatitis.B, BMI, Polio,

# Total.expenditure, Diphtheria, GDP,Population, Income.composition.of.resources,

# thinness..1.19.years , thinness.5.9.years, Schooling, HIV.AIDS, under.five.deaths,

# Measles, percentage.expenditure, infant.deaths))

lifedata1 = lifedata %>% select(-c(Year))

#correlation plot for all continuous variables

corrgram(lifedata1, order=TRUE,

upper.panel=panel.cor, lower.panel=panel.pie, main="Life expectancy correlogram")

#Impute missing data with "CART"(Classification and Regression Trees) method

imputed\_Data2 <- mice(lifedata1, m=5, maxit =5, method = 'cart', seed = 500)

#Data frame after imputation

completeData1 <- mice::complete(imputed\_Data2,)

str(completeData1)

#we readd this variable since it's based on schooliing, not really it's own varible

#completeData1$SchoolingLevel<-cut(completeData1$Schooling, seq(4,24,5), right=FALSE, labels=c("1","2", "3", "4"))

completeData1$Total.expenditurelevel<-cut(completeData1$Total.expenditure, seq(0,18,6), right=FALSE, labels=c("1","2", "3"))

#Display missing-data in a bar-plot after imputation

mice\_plot <- aggr(completeData1, col=c('navyblue','red'),

numbers=TRUE, sortVars=TRUE,

labels=names(completeData1), cex.axis=.7,

gap=3, ylab=c("Missing data","Pattern"))

```

#Create a regression manual regrsion model

```{r }

#manual model

completeData2 = completeData1 %>% select(-c(Country, Status, Total.expenditurelevel))

life.model1<-lm(Life.expectancy~Adult.Mortality+infant.deaths+Alcohol+percentage.expenditure+Hepatitis.B+Measles+

BMI+I(BMI^2)+under.five.deaths+Polio+Total.expenditure+Diphtheria+HIV.AIDS+GDP+Population+log(thinness..1.19.years)+log(thinness.5.9.years)+Income.composition.of.resources+Schooling+ log(GDPPerCapita), data=completeData2)

summary(life.model1)

vif(life.model1)

```

## Below variables are statistically significant

Adult.Mortality

GDP

Population

Total.expenditure

HIV.AIDS

Income.composition.of.resources

## Re run the statistical significant variables and call VIF

```{r }

#statistical significant variables

life.model2<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+HIV.AIDS+Income.composition.of.resources+GDP+Population, data=completeData2)

summary(life.model2)

vif(life.model2)

confint(life.model2)

```

# summary and confint

```{r }

life.model2<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+HIV.AIDS+Income.composition.of.resources+GDP+Population, data=completeData2)

summary(life.model2)

confint(life.model2)

```

# Manual model plot

```{r }

#final manual model

OLS.model<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+HIV.AIDS+Income.composition.of.resources+GDP+Population, data=completeData2)

summary(OLS.model)

plot(OLS.model)

confint(OLS.model)

vif(OLS.model)

```

# ASE, AIC, BIC for manual model

```{r }

OLStrain <- completeData2 [1:120, ]

OLStest <-completeData2 [121:183, ]

life.model2<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+HIV.AIDS+Income.composition.of.resources+GDP+Population, data=OLStrain)

pred<-predict(life.model2, newdata=OLStest)

ASE<-mean((OLStest$Life.expectancy-pred)^2)

ASE

AIC(life.model2)

BIC(life.model2)

summary(life.model2)$adj.r.squared

```

#Log transformed variables to feed the forward model

```{r }

completeData2$"log(GDPPperCapita)"<-log(completeData2$GDPPerCapita)

completeData2$"thinness..1.19.years"<-log(completeData2$thinness..1.19.years)

completeData2$"thinness.5.9.years"<-log(completeData2$thinness.5.9.years)

```

# Model selection "Forward"

```{r }

completeData2 = completeData1 %>% select(-c(Country, Status, Total.expenditurelevel))

reg.fwd=regsubsets(Life.expectancy~., data=completeData2, method="forward", nvmax=20)

reg.fwd

```

# Plot ASE for forward

```{r }

set.seed(1234)

index<-sample(1:dim(completeData2)[1], 100, replace=F)

fwdtrain<-completeData2[index,]

fwdtest<-completeData2[-index,]

reg.fwd=regsubsets(Life.expectancy~., data=fwdtrain, method="forward", nvmax=15)

reg.fwd

predict.regsubsets =function (object , newdata ,id ,...){

form=as.formula (object$call [[2]])

mat=model.matrix(form ,newdata )

coefi=coef(object ,id=id)

xvars=names(coefi)

mat[,xvars]%\*%coefi

}

testASE<-c()

#note my index is to 20 since that what I set it in regsubsets

for (i in 1:15){

predictions<-predict.regsubsets(object=reg.fwd,newdata=fwdtest,id=i)

testASE[i]<-mean((fwdtest$Life.expectancy-predictions)^2)

}

par(mfrow=c(1,1))

plot(1:15,testASE,type="l",xlab="# of predictors",ylab="test vs train ASE", ylim=c(5, 30))

index<-which(testASE==min(testASE))

points(index,testASE[index],col="red",pch=10)

rss<-summary(reg.fwd)$rss

lines(1:15,rss/100,lty=3,col="blue")

```

### BIC, adj R-squared and RSS plots for forward selection

```{r }

par(mfrow=c(1,3))

bics<-summary(reg.fwd)$bic

plot(1:15,bics,type="l",ylab="BIC",xlab="# of predictors")

index<-which(bics==min(bics))

points(index,bics[index],col="red",pch=10)

adjr2<-summary(reg.fwd)$adjr2

plot(1:15,adjr2,type="l",ylab="Adjusted R-squared",xlab="# of predictors")

index<-which(adjr2==max(adjr2))

points(index,adjr2[index],col="red",pch=10)

rss<-summary(reg.fwd)$rss

plot(1:15,rss,type="l",ylab="train RSS",xlab="# of predictors")

index<-which(rss==min(rss))

points(index,rss[index],col="red",pch=10)

summary(reg.fwd)$bic[which(bics==min(bics))]

summary(reg.fwd)$adjr2[which(adjr2==max(adjr2))]

summary(reg.fwd)$rss[which(rss==min(rss))]/100

```

# Pick up the model for forward with AIC, BIC, R-square

```{r }

m1<-coef(reg.fwd,4)

m1

#test the forward model

forward<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+HIV.AIDS+ Income.composition.of.resources, data=completeData2)

summary(forward)

confint(forward)

AIC(forward)

BIC(forward)

summary(forward)$adj.r.squared

plot(forward)

```

#VIF for forward model

```{r }

vif(forward)

```

# Model selection "Backward"

```{r }

completeData2$"log(GDPPperCapita)"<-log(completeData2$GDPPerCapita)

completeData2$"thinness..1.19.years"<-log(completeData2$thinness..1.19.years)

completeData2$"thinness.5.9.years"<-log(completeData2$thinness.5.9.years)

completeData2 = completeData1 %>% select(-c(Country, Status, Total.expenditurelevel))

reg.bwd=regsubsets(Life.expectancy~., data=completeData2, method="backward", nvmax=20)

reg.bwd

```

# Plot ASE for backward

```{r }

set.seed(1234)

index<-sample(1:dim(completeData2)[1], 100, replace=F)

bwdtrain<-completeData2[index,]

bwdtest<-completeData2[-index,]

reg.bwd=regsubsets(Life.expectancy~., data=bwdtrain, method="backward", nvmax=15)

reg.bwd

predict.regsubsets =function (object , newdata ,id ,...){

form=as.formula (object$call [[2]])

mat=model.matrix(form ,newdata )

coefi=coef(object ,id=id)

xvars=names(coefi)

mat[,xvars]%\*%coefi

}

bwdtestASE<-c()

#note my index is to 20 since that what I set it in regsubsets

for (i in 1:15){

predictions<-predict.regsubsets(object=reg.bwd,newdata=bwdtest,id=i)

bwdtestASE[i]<-mean((bwdtest$Life.expectancy-predictions)^2)

}

par(mfrow=c(1,1))

plot(1:15,bwdtestASE,type="l",xlab="# of predictors",ylab="test vs train ASE", ylim=c(5, 30))

index<-which(bwdtestASE==min(bwdtestASE))

points(index,bwdtestASE[index],col="red",pch=10)

bwdrss<-summary(reg.bwd)$rss

lines(1:15,bwdrss/100,lty=3,col="blue")

```

### BIC, adj R-squared and RSS plots for backward selection

```{r }

par(mfrow=c(1,3))

bics<-summary(reg.bwd)$bic

plot(1:15,bics,type="l",ylab="BIC",xlab="# of predictors")

index<-which(bics==min(bics))

points(index,bics[index],col="red",pch=10)

adjr2<-summary(reg.bwd)$adjr2

plot(1:15,adjr2,type="l",ylab="Adjusted R-squared",xlab="# of predictors")

index<-which(adjr2==max(adjr2))

points(index,adjr2[index],col="red",pch=10)

rss<-summary(reg.bwd)$rss

plot(1:15,rss,type="l",ylab="train RSS",xlab="# of predictors")

index<-which(rss==min(rss))

points(index,rss[index],col="red",pch=10)

summary(reg.bwd)$bic[which(bics==min(bics))]

summary(reg.bwd)$adjr2[which(adjr2==max(adjr2))]

summary(reg.bwd)$rss[which(rss==min(rss))]/100

```

# Pick up the model backward

```{r }

m1<-coef(reg.bwd,2)

m1

```

#Backward model with AIC, BIC, R-square

```{r }

backward<-lm(Life.expectancy~Adult.Mortality+Income.composition.of.resources, data=completeData2)

summary(backward)

AIC(backward)

BIC(backward)

summary(backward)$adj.r.squared

```

# Model selection "stepwise"

```{r }

completeData2$"log(GDPPperCapita)"<-log(completeData2$GDPPerCapita)

completeData2$"thinness..1.19.years"<-log(completeData2$thinness..1.19.years)

completeData2$"thinness.5.9.years"<-log(completeData2$thinness.5.9.years)

completeData2 = completeData1 %>% select(-c(Country, Status, Total.expenditurelevel))

reg.stp=regsubsets(Life.expectancy~., data=completeData2, method="seqrep", nvmax=20)

reg.stp

```

# Plot ASE for stepwise model

```{r }

set.seed(1234)

index<-sample(1:dim(completeData2)[1], 100, replace=F)

stptrain<-completeData2[index,]

stptest<-completeData2[-index,]

reg.stp=regsubsets(Life.expectancy~., data=stptrain, method="seqrep", nvmax=15)

reg.fwd

predict.regsubsets =function (object , newdata ,id ,...){

form=as.formula (object$call [[2]])

mat=model.matrix(form ,newdata )

coefi=coef(object ,id=id)

xvars=names(coefi)

mat[,xvars]%\*%coefi

}

stptestASE<-c()

#note my index is to 20 since that what I set it in regsubsets

for (i in 1:15){

predictions<-predict.regsubsets(object=reg.stp,newdata=stptest,id=i)

stptestASE[i]<-mean((stptest$Life.expectancy-predictions)^2)

}

par(mfrow=c(1,1))

plot(1:15,stptestASE,type="l",xlab="iteration #",ylab="test vs train ASE", ylim=c(5, 50))

index<-which(stptestASE==min(stptestASE))

points(index,stptestASE[index],col="red",pch=10)

stprss<-summary(reg.stp)$rss

lines(1:15,stprss/100,lty=3,col="blue")

```

### BIC, adj R-squared and RSS plots for stepwise selection

```{r }

par(mfrow=c(1,3))

bics<-summary(reg.stp)$bic

plot(1:15,bics,type="l",ylab="BIC",xlab="iteration #")

index<-which(bics==min(bics))

points(index,bics[index],col="red",pch=10)

adjr2<-summary(reg.stp)$adjr2

plot(1:15,adjr2,type="l",ylab="Adjusted R-squared",xlab="iteration #")

index<-which(adjr2==max(adjr2))

points(index,adjr2[index],col="red",pch=10)

rss<-summary(reg.stp)$rss

plot(1:15,rss,type="l",ylab="train RSS",xlab="#iteration #")

index<-which(rss==min(rss))

points(index,rss[index],col="red",pch=10)

summary(reg.stp)$bic[which(bics==min(bics))]

summary(reg.stp)$adjr2[which(adjr2==max(adjr2))]

summary(reg.stp)$rss[which(rss==min(rss))]/100

```

# model for stepwise

```{r }

ms<-coef(reg.stp,3)

ms

```

# model stepwise with AIC,BIC,R-square

```{r }

stepwise<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+Income.composition.of.resources, data=completeData2)

summary(stepwise)

confint(stepwise)

AIC(stepwise)

BIC(stepwise)

summary(stepwise)$adj.r.squared

```

#VIF for stepwise model and plot

```{r }

vif(stepwise)

plot(stepwise)

```

#Trying Stepwise with interaction

```{r }

# with interactions, the interaction is not significant

stepwiseI<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+Income.composition.of.resources+Income.composition.of.resources\*Total.expenditure+Adult.Mortality\*Total.expenditure+Adult.Mortality\*Income.composition.of.resources, data=completeData2)

summary(stepwiseI)

AIC(stepwiseI)

BIC(stepwiseI)

summary(stepwiseI)$adj.r.squared

vif(stepwiseI)

```

# LASSO

```{r,echo=T}

set.seed(1234)

index<-sample(1:dim(completeData2)[1],100,replace=F)

train<-completeData2[index,]

test<-completeData2[-index,]

#reg.fwd=regsubsets(log(AvgWinnings)~.,data=train,method="forward",nvmax=20)

```

```{r }

require("glmnet")

#Formatting data for GLM net

set.seed(1234)

x=model.matrix(Life.expectancy~Adult.Mortality+Hepatitis.B+BMI+I(BMI^2)+I(BMI^3)+Polio+Total.expenditure+Diphtheria+GDP+Population

+Income.composition.of.resources+thinness..1.19.years + I(thinness..1.19.years^2)+thinness.5.9.years+I(thinness.5.9.years^2)+Schooling+HIV.AIDS

+log(GDPPerCapita),train)[,-1]

y=train$Life.expectancy

xtest<-model.matrix(Life.expectancy~Adult.Mortality+Hepatitis.B+BMI+I(BMI^2)+I(BMI^3)+Polio+Total.expenditure+Diphtheria+GDP+Population

+Income.composition.of.resources+thinness..1.19.years + I(thinness..1.19.years^2)+thinness.5.9.years+I(thinness.5.9.years)+Schooling+HIV.AIDS

+log(GDPPerCapita),test)[,-1]

ytest<-test$Life.expectancy

grid=10^seq(10,-2, length =100)

lasso.mod=glmnet(x,y,alpha=1, lambda =grid)

cv.out=cv.glmnet(x,y,alpha=1) #alpha=1 performs LASSO

plot(cv.out)

bestlambda<-cv.out$lambda.min #Optimal penalty parameter. You can make this call visually.

lasso.pred=predict (lasso.mod ,s=bestlambda ,newx=xtest)

testMSE\_LASSO<-mean((ytest-lasso.pred)^2)

testMSE\_LASSO

```

#Refit the laso model

```{r }

require("glmnet")

#Formatting data for GLM net

set.seed(1234)

x=model.matrix(Life.expectancy~Adult.Mortality+Total.expenditure+HIV.AIDS+Income.composition.of.resources+ BMI + Diphtheria + Population+thinness.5.9.years+I(thinness.5.9.years^2)+Schooling+log(GDPPerCapita)+ Hepatitis.B+I(BMI^2)+I(BMI^3)+Polio+GDP+thinness..1.19.years+I(thinness..1.19.years^2),train)[,-1]

y=train$Life.expectancy

xtest<-model.matrix(Life.expectancy~Adult.Mortality+Total.expenditure+HIV.AIDS+Income.composition.of.resources+ BMI + Diphtheria + Population+thinness.5.9.years+I(thinness.5.9.years^2)+Schooling+log(GDPPerCapita)+ Hepatitis.B+I(BMI^2)+I(BMI^3)+Polio+GDP+thinness..1.19.years+I(thinness..1.19.years^2),test)[,-1]

ytest<-test$Life.expectancy

grid=10^seq(10,-2, length =100)

lasso.modr=glmnet(x,y,alpha=1, lambda =grid)

cv.out=cv.glmnet(x,y,alpha=1) #alpha=1 performs LASSO

plot(cv.out)

bestlambda<-cv.out$lambda.min #Optimal penalty parameter. You can make this call visually.

lasso.predr=predict (lasso.modr ,s=bestlambda ,newx=xtest)

testMSE\_LASSO<-mean((ytest-lasso.predr)^2)

testMSE\_LASSO

```

# Best lambda for lasso

```{r}

coef(lasso.mod,s=bestlambda)

```

# Best lambda for lasso refit

```{r}

coef(lasso.modr,s=bestlambda)

```

#Lasso AIC, BIC, R-square, VIF, summary and plot

```{r }

#Run the lasso model

lasso.model<-lm(Life.expectancy~Adult.Mortality+Total.expenditure+HIV.AIDS+Income.composition.of.resources+ BMI + Diphtheria + Population+thinness.5.9.years+I(thinness.5.9.years^2)+Schooling+log(GDPPerCapita)+ Hepatitis.B+I(BMI^2)+I(BMI^3)+Polio+GDP+thinness..1.19.years+I(thinness..1.19.years^2), data=completeData2)

summary(lasso.model)

plot(lasso.model)

confint(lasso.model)

vif(lasso.model)

AIC(lasso.model)

BIC(lasso.model)

summary(lasso.model)$adj.r.squared

```

######################################################################################################################################################################## 2-Way ANOVA ###################################################################################

# Prepare data for 2-way ANOVA

```{r}

lifeanovadata = completeData1 %>% select(c(Status, Total.expenditurelevel, Life.expectancy))

str(lifeanovadata)

ggplot(lifeanovadata, aes(x=Total.expenditurelevel, y=Life.expectancy, fill = Status)) +

geom\_boxplot() + geom\_smooth(method=lm) +

xlab("Total Expenditure") + ylab("Life Expectancy") + ggtitle("Total Expenditure vs Life Expectancy")

ggplot(lifeanovadata, aes(x=Total.expenditurelevel, y=log(Life.expectancy), fill = Status)) +

geom\_boxplot() + geom\_smooth(method=lm) +

xlab("Total Expenditure") + ylab("Log Life Expectancy") + ggtitle("Total Expenditure vs Log Life Expectancy")

```

# compute Mean SD

```{r}

mysummary1<-function(x){

result<-c(length(x),mean(x),sd(x), min(x), max(x))

names(result)<-c("N","Mean","SD", "Minimum", "maximum")

return(result)

}

sumstats1<-aggregate(Life.expectancy~Total.expenditurelevel\*Status,data=lifeanovadata,mysummary1)

sumstats1<-cbind(sumstats1[,1:2],sumstats1[,-(1:2)])

sumstats1

```

# Means plot using the standard deviations (SD)

```{r}

ggplot(sumstats1,aes(x=Total.expenditurelevel,y=Mean,group=Status,colour=Status))+

ylab("Life expectancy")+

geom\_line()+

geom\_point()+

geom\_errorbar(aes(ymin=Mean-SD,ymax=Mean+SD),width=.1)

```

# ANOVA model

```{r}

model.fit<-aov(log(Life.expectancy)~Status+Total.expenditurelevel+Status:Total.expenditurelevel,data=lifeanovadata)

summary(model.fit)

par(mfrow=c(1,2))

plot(exp(model.fit$fitted.values),model.fit$residuals,ylab="Resdiduals",xlab="Fitted")

qqnorm(model.fit$residuals)

```

```{r}

model.fit1<-aov(log(Life.expectancy)~Status+Total.expenditurelevel+Status:Total.expenditurelevel,data=lifeanovadata)

Anova(model.fit1,type=3)

```

```{r}

model.fit2<-aov(log(Life.expectancy)~Status+Total.expenditurelevel,data=lifeanovadata)

Anova(model.fit2,type=3)

```

```{r}

library(car)

TukeyHSD(model.fit2,"Status",conf.level=.95)

TukeyHSD(model.fit2,"Total.expenditurelevel",conf.level=.95)

```