Project 1: Life Expectancy (WHO)

Statistical Analysis on factors influencing Life Expectancy

By Suchismita Moharana and Jamie Vo

## Introduction

Immunization has received national attention concerning the risks and dangers of receiving vaccines. The World Health Organization, and many other institutions, studied the factors that would affect life expectancy, with the exception of including immunizations. By including three vaccines (Hepatitis B, Polio, and Diphtheria), the observational study determines whether the variables play a significant role in reducing or increasing life expectancy.

## Data Description

The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The datasets are made available to the public for the purpose of health data analysis. The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from the United Nation website. The dataset used for this project is available on [kaggle’s site](https://www.kaggle.com/kumarajarshi/life-expectancy-who?select=Life+Expectancy+Data.csv).

The original dataset has 2,938 observations and 22 variables. For an in depth explanation of each variable, please visit the [appendix](#_i7vdx7emixjg). We have considered observations for 2014 for analysis of this project which has 183 records. The data has been reduced to only 2014 for consistency considering health climates have the ability to change year over year. The final dataset utilized in the analysis mapped below contains 183 observations with 21 variables, omitting year and country. Country is omitted due to it being unique for each observation.

## Exploratory Data Analysis (EDA)

Initially, the data has many missing data points, primarily from the population and GDP columns. Removing all observations with missing data resulted in ~30% loss of the dataset, hence the replacement of nulls with the mean of each variable is opted for.

The EDA caused questions to arise concerning the data collection methods for specific variables. Percentage expenditure (expenditure on health as a percentage of Gross Domestic Product per capita(%) ) has values ranging from 0 to 19480. Since this is a percentage value, it should not go beyond 100. And in practical scenarios, no country spends 100% of its GDP on healthcare.

Population is a right skewed field and removed because it has a high number of missing values and it does not appear to have any relationship with life expectancy. Adult mortality rates appear to have strong negative correlation with life expectancy whereas schooling and income composition show a clear positive linear trend for life expectancy. Developed countries have a higher mean of life expectancy than a developing country.

In analyzing the immunizations (Hepatitis.B, Diphtheria, and Polio), there appears to be no linear relationship with life expectancy. Initially appearing to have a quadratic relationship. Despite this more precise model, the ability to explain the data is less than 15%. Due to this, the expectation of immunizations having a significant impact on life expectancy is low.

## Objective 1

### Restatement of Problem and the overall approach to solve it

The primary goal of the analysis is to determine key variables that attribute to life expectancy. For model one, in addition to determining what key factors affect life expectancy, interpretability is kept in mind. Model 2 focuses specifically on predictability, with the concern for explanation to be reduced. Finally, model 3 utilized a nonparametric model to determine statistically significant variables.

After completing the EDA for the data set, forward, backward, stepwise, and lasso techniques were used to narrow down the variables and determine significance. Stepwise and backward resulted in the same variables, hence, moving forward, only stepwise will be referred to. Lasso returned a number of variables that were not in the linear regression stepwise. It was ultimately determined that Lasso’s only significant return variable is income and composition of resources, considering that it trends towards zero.

|  |  |
| --- | --- |
| Stepwise/Backward | Lasso |
| * StatusDeveloping * Adult.Mortality * infant.deaths * under.five.deaths * Total.expenditure * Diphtheria * HIV.AIDS * Income.composition.of.resources | * Income.composition.of.resources * StatusDeveloping * Adult.Mortality * Alcohol * under.five.deaths * Total.expenditure * Diphtheria * thinness.5.9.years * Income.composition.of.resources |

In running the selections above, collinearity is eliminated and not a concern with the remaining models.

The higher complexity model consisted of three phases:

1. Adding in additional variables that were considered to be significant from Lasso and stepwise, but did not make it into the final model
2. Adding in interaction terms

In running the different models and comparing the metrics, the model with the greatest ability to predict data was selected as the final model.

Non-parametric models were tested (KNN, decision tree, random forest). KNN is dependent on the value of selected K and results are not consistent. Because this is dependent on the data set split (Train / Test) and results would be dependent on the values of nearest neighbours. While it can give surprisingly accurate results, the error rate can be high as well sometimes. Decision Tree methods need tree pruning to provide the optimal result whereas a random forest model handles it on its own. Random forest was selected as the method due to KNN proving to be an inappropriate model. Rather than selecting only decision trees over random forest, random forest does not require pruning and accounts for the potential of overfitting.

### Model Selection

#### Type of Selection

The selection types tested are stepwise, forward, lasso, and manual. [Lasso](#_byqg19g0r1gs), followed by [stepwise](#_5vn06fofiobq) and [forward](#_dgdjknx3qe0f) selection had the lowest to highest number of variables (respectively), that were considered to be significant. As anticipated, forward resulted in the highest number of variables since it continues to add variables on top of the existing model, stepwise has less variables as it adds and removes variables as they become insignificant, and lasso results in the lowest amount due to its trend towards zero. Lasso ultimately was narrowed down to only one variable considering the coefficients of the variable are less than 0.10.

#### Checking Assumptions

Linear regression requires 4 assumptions to be met in order for regression to be applied properly. Available in the [appendix](#_4eme23w3he1y) linearity, normality, and constant variance are met for all 3 models. In terms of independence, since all models were subjected to a selection method, any collinearity has been eliminated in the process. This is due to the fact that any collinearity would cause all but one of the variables to be insignificant, resulting from being eliminated from the final model.

Outliers were present in the data, but not to an extent where the models were greatly skewed. Reviewing the [Cook’s D plots](#_4eme23w3he1y) for all 3 models, none exceed the scale of 10. All three plots have Cook’s D below 1, causing the concern for outliers to be eliminated for linear regression.

#### Compare Competing Models

Lasso was run to determine the most significant variables, which was determined to only be income and composition of resources. Stepwise resulted in a number of variables, followed by forward which had a significantly greater number.

All models were run to collect metrics in table 1, below. In addition, they were trained on a training set before using the model to predict a test set that was split 80/20. Reviewing the scores MSE, CV, AIC, and BIC, the trend follows lasso, forward and stepwise, respectively in decreasing order. This indicates that the stepwise selection model is considered to be the most accurate model of the three, and lasso being the least accurate.

The adjusted R^2 indicates that the stepwise and forward selection models are able to explain similar degrees of the data (0.86, 0.85, respectively), while lasso has a significant drop (0.73). After testing with the split data set, forward and stepwise do share similar accuracy metrics, with only 0.01 difference. Reviewing the accuracy score for lasso is higher than anticipated when looking at the R^2 models, having only a 0.03 difference to stepwise.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **MSE** | **CV** | **AIC** | **BIC** | **AdjR2** | **Accuracy** |
| forward | 3.35 | 14.14 | 464.26 | 534.87 | 0.85 | 0.78 |
| stepwise | 3.29 | 11.71 | 447.13 | 482.44 | 0.86 | 0.79 |
| lasso | 4.34 | 19.84 | 549.09 | 558.72 | 0.73 | 0.76 |

Table 1: Linear regression metrics for forward, stepwise, and lasso selections.

This indicates that although stepwise resulted in more promising metrics on the training dataset, lasso has similar abilities to predict accurately. As the focus is on interpretability, and there is little difference in predicting ability, lasso is selected to be the primary model.

#### Parameter Interpretation (Simple model only)

The model can be interpreted as a linear relationship between life expectancy and income/composition of resources. Note that income is on an index scale from 0 to 1. One being the max income/composition of resources, this indicates that the greatest average life expectancy is anticipated to be 87.31 years (p-value <0.001). In the case of no income or composition of resources, the mean life expectancy is anticipated to be 37.09 (p-value <0.001). An example of this case could be unemployment, homeless, etc…

*Life expectancy = 37.09 + 50.22 (income)*

The 95% confidence interval for the life expectancy with no income ranges from 33.82 to 40.36 years and 45.63 to 54.84 years for the addition of income and composition of resources.

|  |  |  |
| --- | --- | --- |
| **Variable** | **CI 2.5 %** | **CI 97.5 %** |
| (Intercept) | 33.82 | 40.36 |
| Income.composition.of.resources | 45.63 | 54.84 |

Table 2: Linear regression coefficient confidence intervals

#### Conclusion

Life expectancy has a linear relationship with income/composition of resources that can explain 73% of the data. Income is indexed on a scale of 0 to 1. It is suggested that the average maximum life expectancy is 87.31 years (p-value < 0.001). If there is no income or composition of resources, the mean life expectancy is anticipated to be 37.09 (p-value <0.001). The 95% confidence interval for the life expectancy with no income ranges from 33.82 to 40.36 years and 45.63 to 54.84 years for the addition of income and composition of resources.

#### Scope of inference

This is an observational study, hence no cause-and-effect can be established. It is regardless interesting to observe the impact that income and composition of resources has on the life expectancy in the dataset.

### Complex Model

#### Model Selection

The focus of model 2 is predictability, hence interpretability is no longer a concern. Variables that were determined to be significant, but had low coefficients from the lasso selection were added and removed manually to determine significance. This included total expenditure, adult mortality, etc...

After selecting the final variables, interaction terms were tested to determine significance. The [model](#_hiofvmkeritl) determined that income, adult mortality, HIV/AIDs, and the interaction term between HIV/AIDS and income to be significant.

*Life.expectancy = 53.00 + 33.73\*Income.composition.of.resources - 0.03\*Adult.Mortality -9.07\*(Income.composition.of.resources\*HIV.AIDS)*

#### Assumptions

All assumptions for linearity, normality, independence, and constant variance are met, noted in the [appendix](#_kuxoocofxrsc). For outliers, the cook’s d plot scale is well below 0.5, and no outliers are determined to be significant.

#### Compare Models

The complex model exceeds the simple model in all metrics. The model is able to account for 84% of the explanatory variables for life expectancy, as opposed to 73% with the simple model. The accuracy score also increases significantly with the test, train sets from 76% to 85%.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Models** | **CV** | **AIC** | **BIC** | **AdjR2** | **Accuracy** |
| Model 1 | 19.84 | 549.09 | 558.72 | 0.73 | 0.76 |
| Model 2 | 12.74 | 373.07 | 390.97 | 0.84 | 0.85 |

Table 3: Linear regression metrics comparison for the simple and complex models

Model 2 increases predictive ability and while adding decent complexity (Table3). Overall, due to the metrics being more beneficial, model 2 is able to explain a greater portion of the data in relation to life expectancy and predict with greater accuracy.

## Objective 2

### Selection Method

KNN and regression trees were tested as non-parametric methods to predict and model life expectancy. Random forest was selected as KNN proved to be an inappropriate method for the data, as well as no requirement for pruning, and addresses the potential for overfitting.

Random forest iterates through the variables and generates randomly selected trees (by default 500). It then runs each tree for classification, in which it then takes all the outcomes, ultimately determining the final classification with the ensemble of all trees. In doing so, this allows the model to account for errors in which a single tree could potentially not consider (Diagram1).

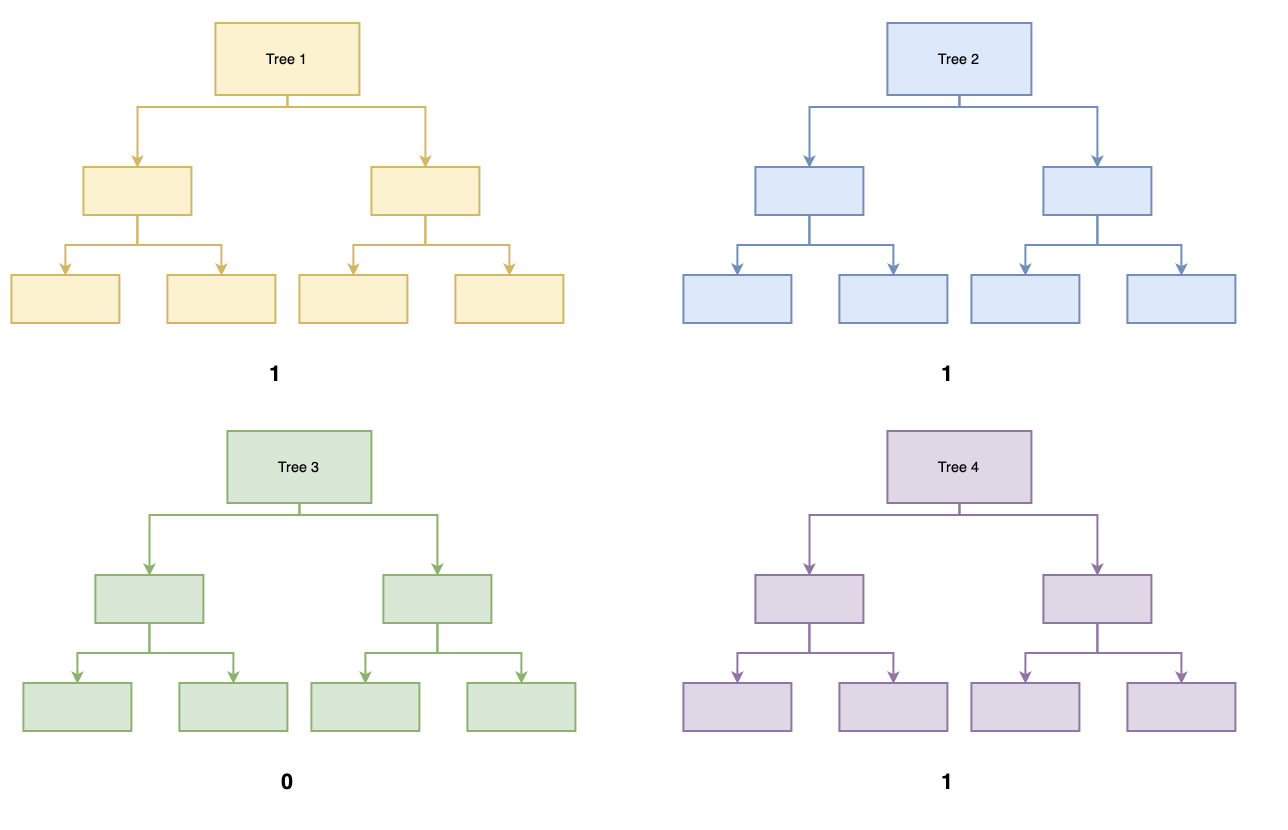


Diagram 1: Example of a random forest. Three of the four are classified to 1, hence the data in question will be classified as group 1.

Two models were run, one with [all variables](#_opvhzrt27i3c) included, and another with the [limited variables](#_rnz0j16oxb0j) from model 2. The full model iterated over 500 trees, with 6 variables per tree. The limited model also iterated over 500 trees, but only accounted for a single variable per tree due to the limited number of variables inputted. All three variables for the limited model (HIV/AIDs, income, and adult mortality) were all determined to be significant variables, while the full model considered in addition schooling to be a significant factor.

Reviewing the plots, the full vs. limited models were able to explain 90% and 88%, respectively, of life expectancy. Notice that the increase from 84% to 88% is significant considering the variables were not changed. This difference can be accounted to be due to the random forest algorithm.

### Comparing Competing Models

As the model complexity increases from simple linear regression, to regression with interaction terms, and finally random forest, the model’s ability to explain life expectancy increases from 73% to 90%. While there is an increase in the mean squared error for model 2 and 3 (3.50 to 7.51, respectively), the increase in the R2 is significant. The increase in residual error for the random forest may be due to the algorithm’s nature of iterating through all variables and the 500 trees.

|  |  |  |
| --- | --- | --- |
| **Models** | **MSE** | **AdjR2** |
| Model 1 | 4.34 | 0.73 |
| Model 2 | 3.50 | 0.84 |
| Model 3 | 7.51 | 0.90 |

Table 4: Regression metrics comparison for the three models (simple, complex, and random forest).

Model 1 is the simplest of the three, with an intermediate mean squared error (MSE) and decent ability to explain life expectancy data. Model 2 has the lowest MSE at 3.5, with the ability to explain 84% of the data. While model 2 is an increase in complexity from model 1, the additional variables and interaction terms result in a 0.09 increase in R2. Finally, model 3 has the highest MSE, and also the highest explanatory ability of 90%, but is the most complex of all three models. Arguably, model 2 may be the most balanced of the three when considering interpretability, MSE, explanatory ability, and weights of bias and precision.

### Conclusion

All three models are capable of explaining greater than 70% of the life expectancy data. Model 1 has the greatest interpretability, with a moderate MSE (4.34) by including only income and composition of resources in a linear regression. Model 2 increases complexity by introducing additional variables (adult mortality, HIV/AIDs) and an interaction term (where income and HIV/AIDs are dependent on one another). The model has the lowest MSE (3.5) of all three models, while having the explanatory ability of 84%, also through linear regression. Model 3 is the most complex of all the models, using the random forest algorithm. While maintaining the highest explanatory ability of 90%, model 3 also has the highest MSE of 7.51 when including all variables.

Model 2 is the recommended model for all metrics (interpretability, error, etc…). Model 2 is relatively simple in comparison to random forest, and has an explanatory model that is considered sufficient. The model also is not as biased as model 1 (with only one variable) and more precise than model 3 (all variables imputed to the algorithm), providing a balance between the two.

The study is observational, hence no cause-and-effect can be determined. A concern with the data set is the missing information. There appears to be missing columns concerning smoking and other factors that may have been beneficial to the analysis. In addition, the data is only analyzed for the year 2014. Additional analysis on each year may provide additional insights.

# 

# Appendix

Well commented SAS/R Code

Graphics and summary tables (Can be placed in the appendix or in the written report itself.)

## Data

### Data Definitions

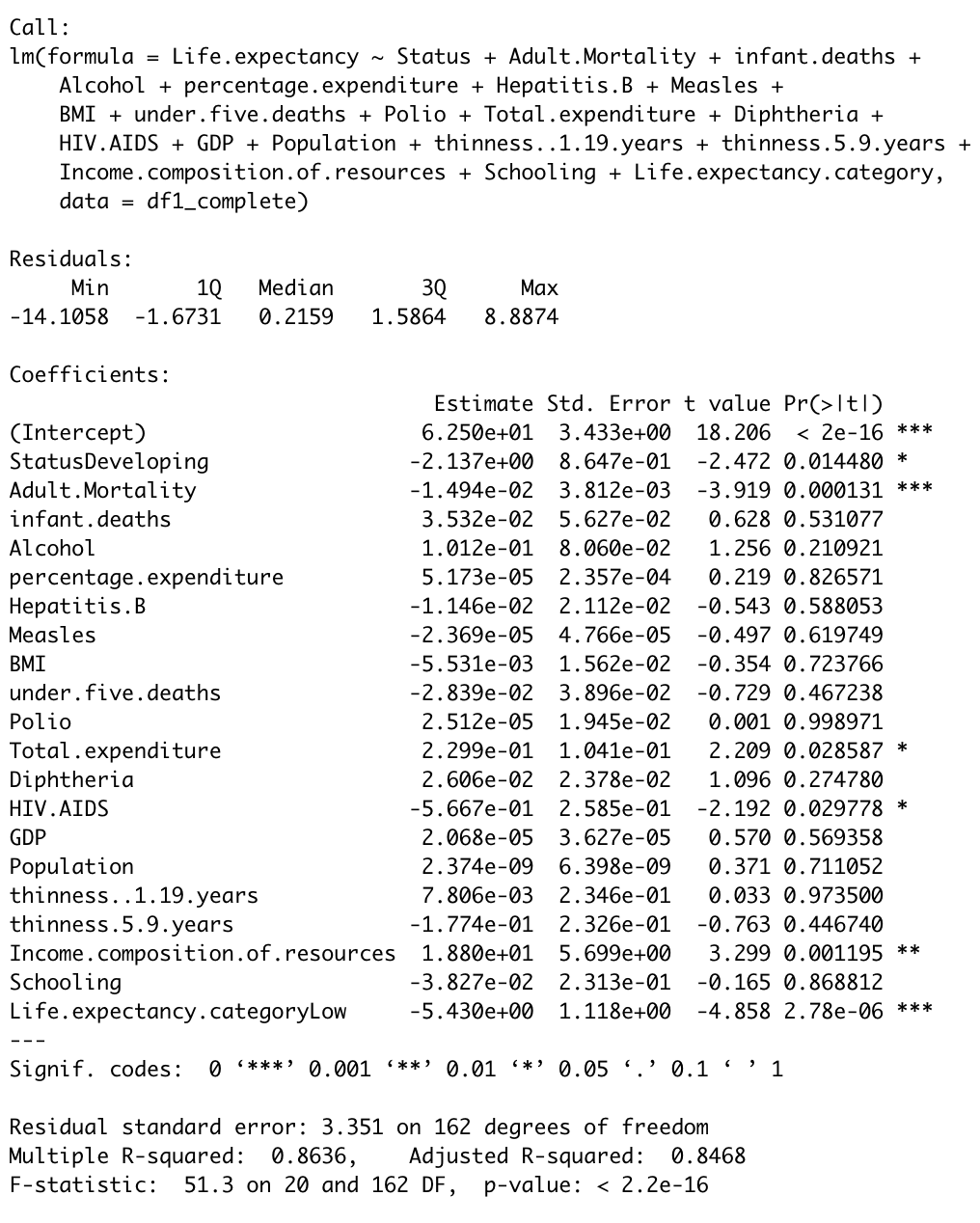
|  |  |
| --- | --- |
| **Variable** | **Definition** |
| Year | Year |
| Status | Developed or Developing status |
| Life expectancy | Life expectancy in Age |
| Adult Mortality | Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population) |
| Infant deaths | Number of Infant Deaths per 1000 population |
| Alcohol | Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol) |
| percentage expenditure | Expenditure on health as a percentage of Gross Domestic Product per capita(%) |
| Hepatitis B | Hepatitis B (HepB) immunization coverage among 1-year-olds (%) |
| Measles | Measles - number of reported cases per 1000 population |
| BMI | Average Body Mass Index of entire population |
| Polio | Polio (Pol3) immunization coverage among 1-year-olds (%) |
| Total expenditure | General government expenditure on health as a percentage of total government expenditure (%) |
| Diphtheria | Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%) |
| HIV/AIDS | Deaths per 1 000 live births HIV/AIDS (0-4 years) |
| GDP | Gross Domestic Product per capita (in USD) |
| Population | Population of the country |
| thinness 1-19 years | Prevalence of thinness among children and adolescents for Age 10 to 19 (%) |
| thinness 5-9 years | Prevalence of thinness among children for Age 5 to 9(%) |
| income composition of resources | Human Development Index in terms of income composition of resources (index ranging from 0 to 1) |
| schooling | Number of years of Schooling(years) |

## Objective 1

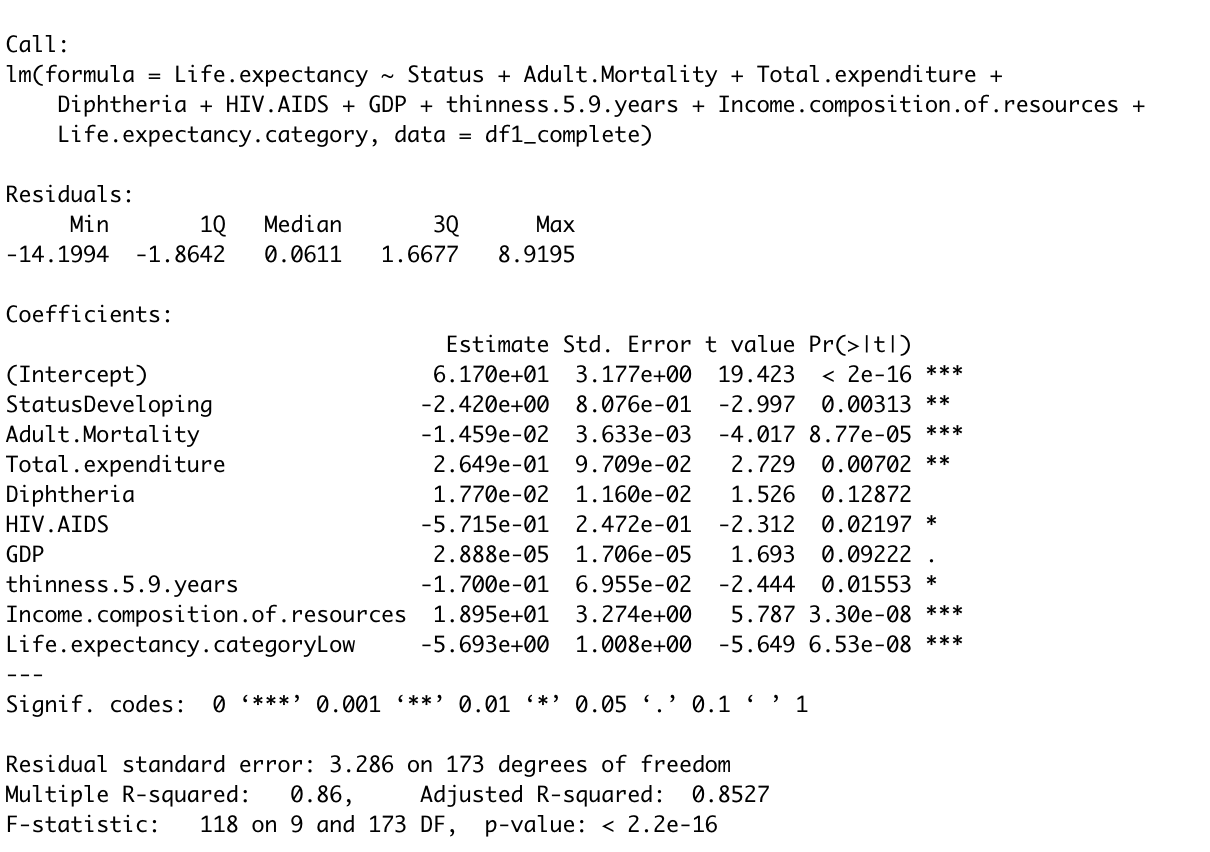
### Model 1

#### Model Selections

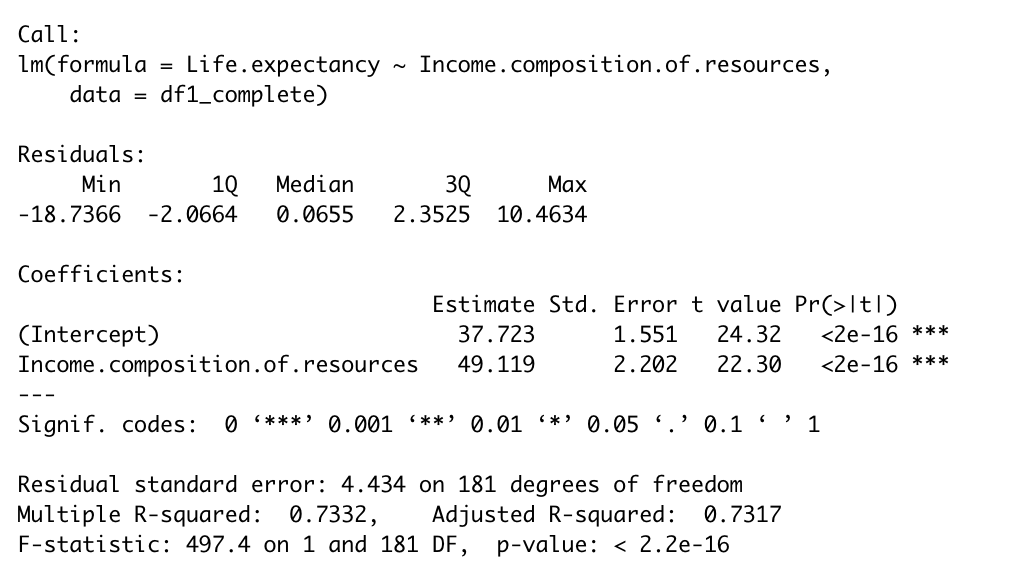
##### Forward



##### Stepwise



##### Lasso

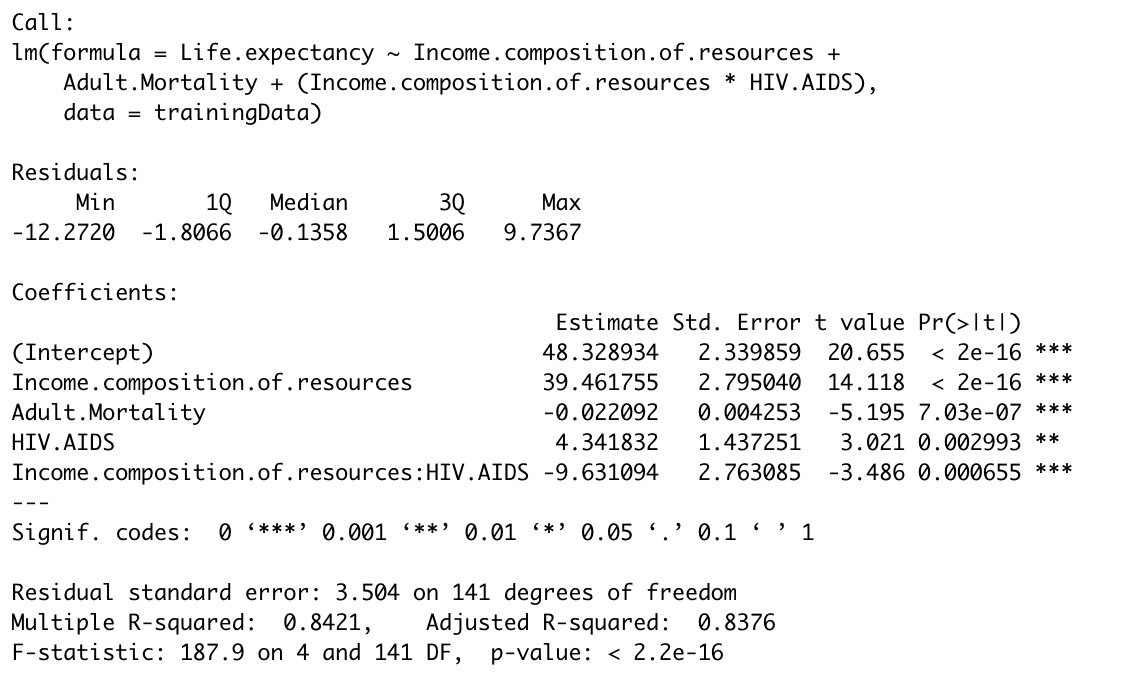


#### Assumptions

|  |  |  |  |
| --- | --- | --- | --- |
| **Models/Graphs** | **Forward** | **Backward/Stepwise** | **Lasso** |
| Linearity |  | | |
| Constant Variance |  |  |  |
| Normality |  |  |  |
| Outliers |  |  |  |
| Independence | No evidence against | No evidence against | No evidence against |

### Model 2

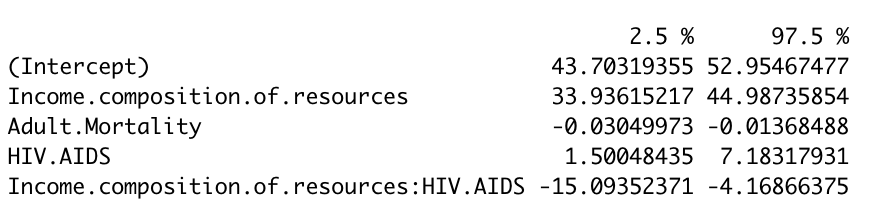
#### Interaction Term Model



#### Assumptions

|  |  |
| --- | --- |
| Assumption | Graph |
| Linearity |  |
| Normality |  |
| Constant Variance |  |
| Independence | No evidence against |
| Cook’s D |  |

### Confidence Intervals

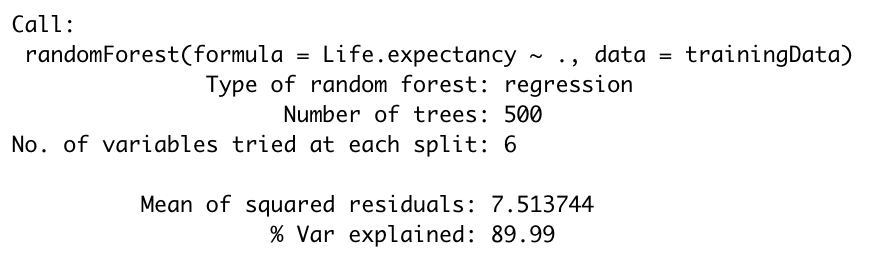


## Objective 2

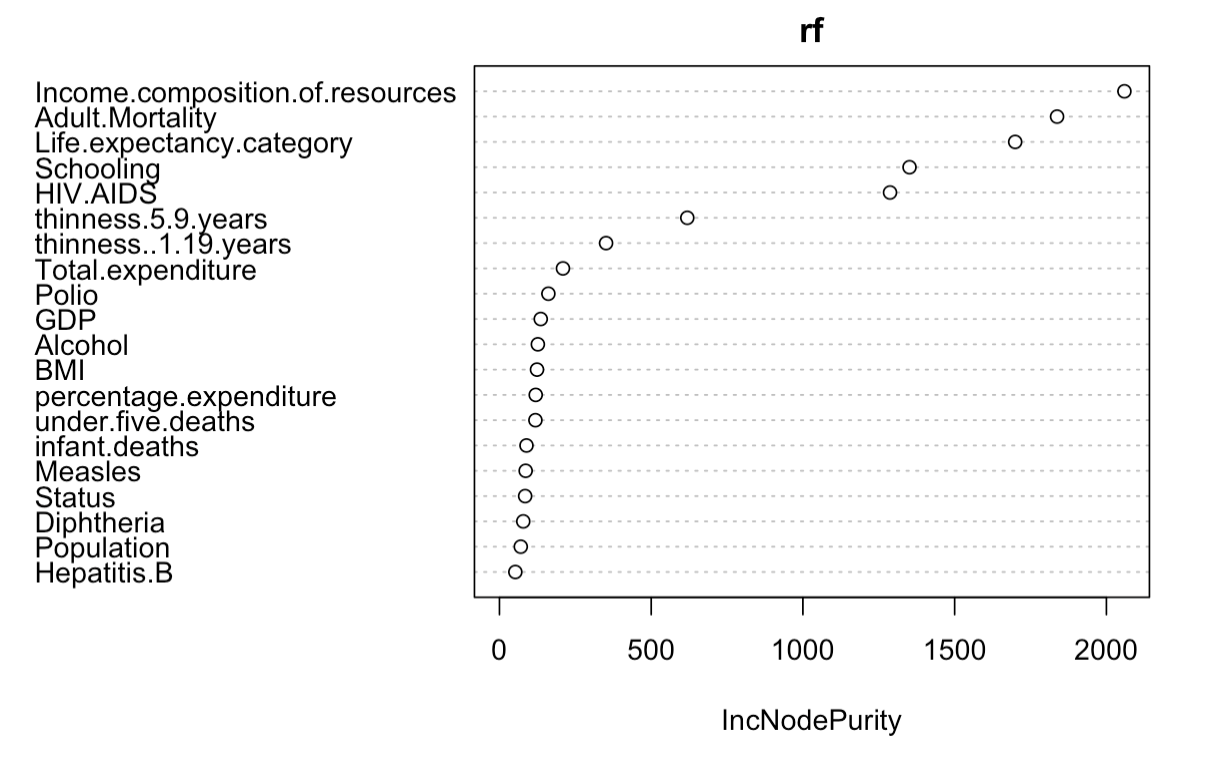
### Random Forest Regression Tree

#### Full Model

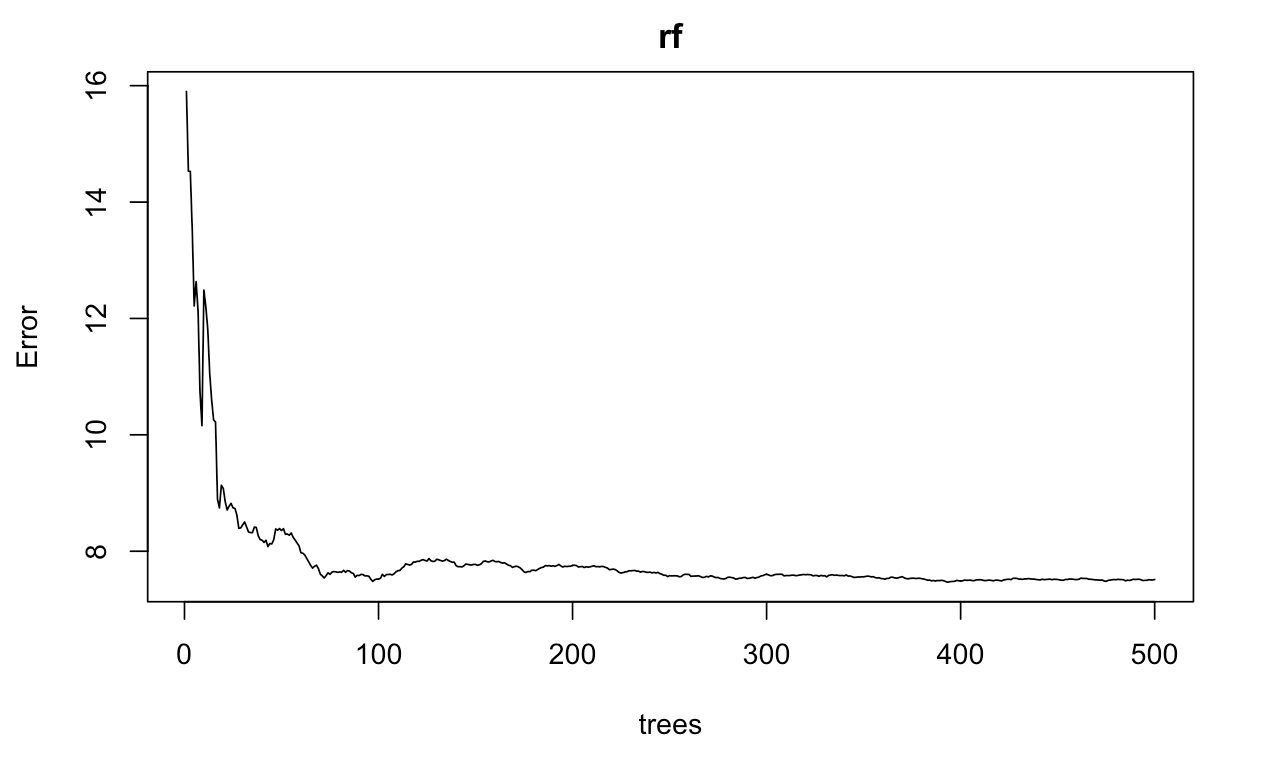
##### Metrics



##### Variable Analysis

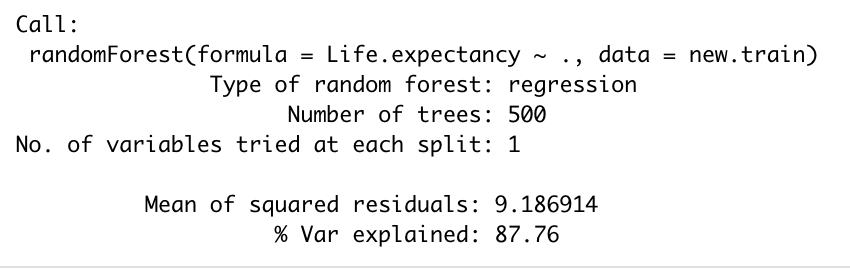


##### Error Trend

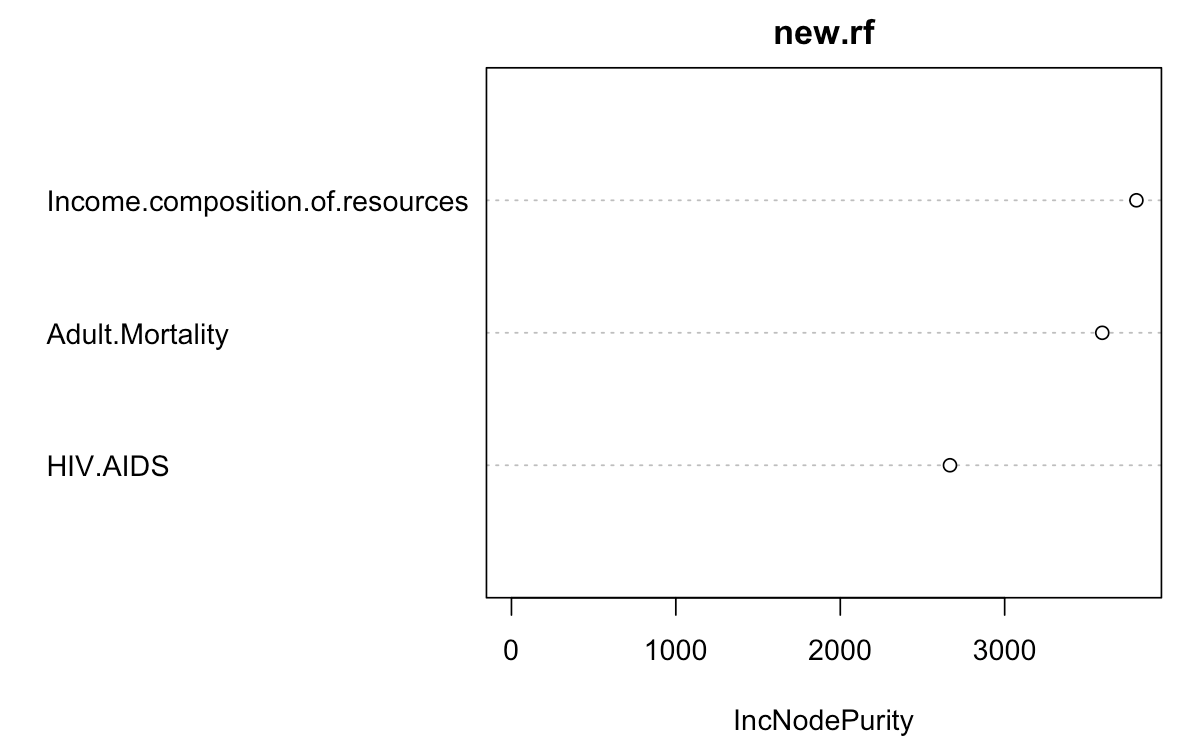


#### Limited Variables Model

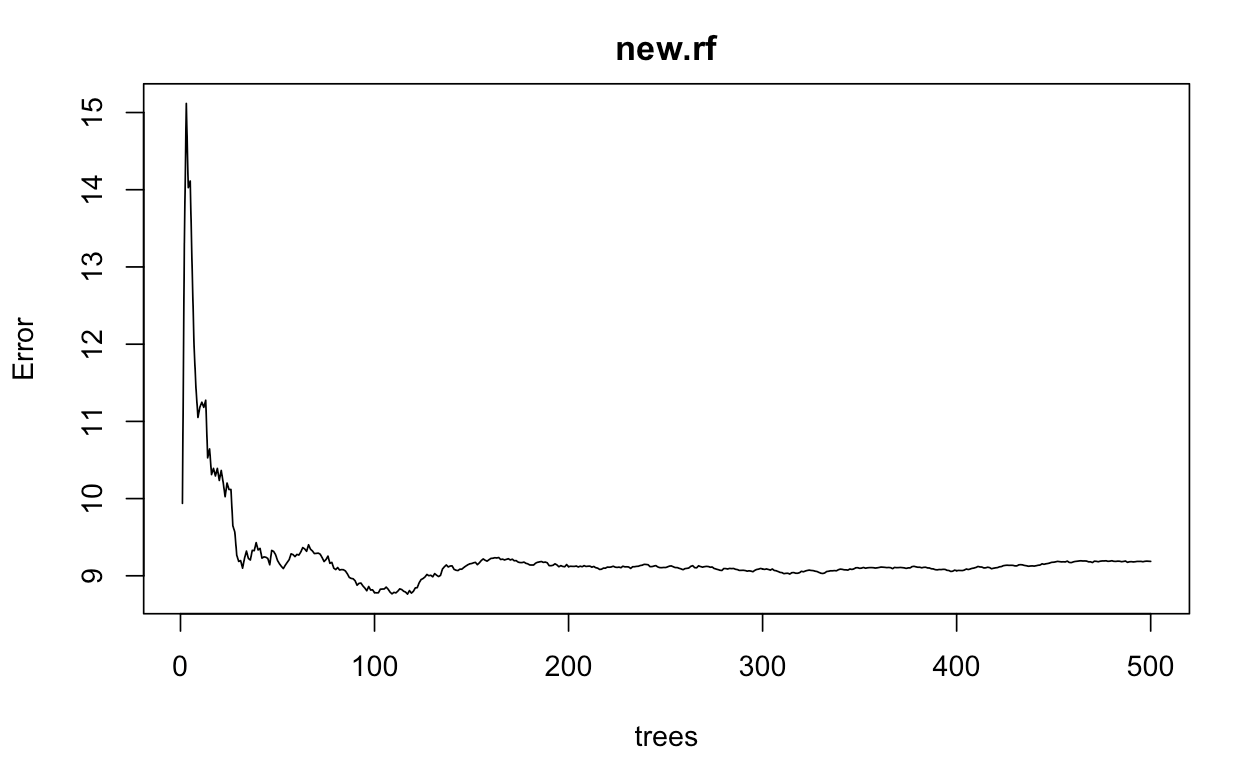
##### Metrics



##### Variable Analysis



##### Error Trend



## Code

---

title: "DS 6372 Project 1"

output: html\_document

---

# World Health Organization

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)

library(ggplot2)

library(plyr)

library(dplyr)

library(GGally)

library(tidyverse)

library(naniar)

library(zoom)

library(MASS)

library(plotly)

library(ggpubr)

library(onewaytests)

library(glmnet)

library(randomForest)

library(DAAG)

library(forecast)

```

## Context

Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that affect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

\*Column Names: Meanings\*

- Year: Year

- Status: Developed or Developing status

- Life expectancy: Life Expectancy in age

- Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)

- infant deaths: Number of Infant Deaths per 1000 population

- Alcohol: Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)

- percentage expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita(%)

- Hepatitis B: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)

- Measles:Measles - number of reported cases per 1000 population

- BMI: Average Body Mass Index of entire population

- Polio: Polio (Pol3) immunization coverage among 1-year-olds (%)

- Total expenditure: General government expenditure on health as a percentage of total government expenditure (%)

- Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)

- HIV/AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years)

- GDP: Gross Domestic Product per capita (in USD)

- Population: Population of the country

- thinness 1-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19 (%)

- thinness 5-9 years: Prevalence of thinness among children for Age 5 to 9(%)

- income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

- schooling: Number of years of Schooling(years)

```{r}

# creates all required resisual plots

residual.plots <- function (model) {

par(mfrow = c(2, 3))

p1 <- plot(model)

p2 <- plot(model, 3)

p3 <- plot(model, 4)

return(list(p1, p2, p3))

}

# Compute R^2 from true and predicted values

eval\_results <- function(true, predicted, df) {

SSE <- sum((predicted - true)^2)

SST <- sum((true - mean(true))^2)

R\_square <- 1 - SSE / SST

RMSE = sqrt(SSE/nrow(df))

# Model performance metrics

data.frame(

RMSE = RMSE,

Rsquare = R\_square

)

}

```

#### ETL

Prior to running the models, the data was manipulated to remove nulls, adjusted for inconsistencies in the data, and eliminating columns that were not required.

1. Removing Nan values

This proved to have higher R^2 values and lower number of variables that were considered to be significant. While the model appears to have a better fit, this is not the ideal scenario since ~30% of the data was lost in this process.

2. Filter out percentage.expenditure greather than 100.

This caused additional values to be considered in significant. This is filtered considering that a population cannot spend more than 100% of their GDP on health care.

3. Replace Nan with means

This allows for all data to be conserved.

```{r}

# Reading datafile

Life\_Expectancy\_Df <- read.csv("../data\ /Life\ Expectancy\ Data.csv", header=TRUE)

Life\_Expectancy\_Df\_2014 <- Life\_Expectancy\_Df %>% filter(Year == 2014) # Keeping only 2014 data as per requirement

gg\_miss\_var(Life\_Expectancy\_Df\_2014)

# remove Nan rows

# # dummy code (developing = 1, developed = 0)

# values <- c(0, 1)

# Life\_Expectancy\_Df\_2014$Status\_dc <- values[Life\_Expectancy\_Df\_2014$Status]

# Life\_Expectancy\_Df\_2014\_dc <- Life\_Expectancy\_Df\_2014[,c(4:23)]

# remove nulls

# df1\_no.nan <- na.omit(Life\_Expectancy\_Df\_2014\_dc)

# drop population column

# df1\_drop <- subset(Life\_Expectancy\_Df\_2014\_dc, select = -c(Population, GDP) )

# df1\_drop <- na.omit(df1\_drop)

# replace missing data wtih averages

df\_means <- Life\_Expectancy\_Df\_2014

for(i in 1:ncol(df\_means)){

df\_means[is.na(df\_means[,i]), i] <- mean(df\_means[,i], na.rm = TRUE)

}

sum(is.na(df\_means))

df1\_complete <- df\_means[,c(3:22)]

#df\_means.condense <- df\_means.condense %>% filter(df\_means.condense$percentage.expenditure < 100)

# Adding a column to categorize Life.expectancy as >= 65 and < 65 for comparison

Life.expectancy.category=ifelse(df1\_complete$Life.expectancy >= 65.0,"High","Low")

df1\_complete =data.frame(df1\_complete ,Life.expectancy.category)

```

#### 1. Does various predicting factors which has been chosen initially really affect the Life expectancy? What are the predicting variables actually affecting the life expectancy?

After running multiple regression models (forward, backward, and stepwise), it was determined that the following variables are the predictors that are statically significant in regards to life expectancy, where stepwise and backward had the same results.

Variables

StatusDeveloping -1.923054

Adult.Mortality -0.016917

infant.deaths 0.086154

under.five.deaths -0.067124

Total.expenditure 0.250440

Diphtheria 0.028511

HIV.AIDS -1.020423

Income.composition.of.resources 27.870308

adjusted R^2: 0.83

AIC: 992.018

While the other variables intuitively may appear significant, the variables above are the predictors that have significant impact according to the regression models. interestingly enough, all models resulted in the same variables being returned.

```{r}

# Fit the full model

full.model <- lm(Life.expectancy ~., data = df1\_complete)

# Stepwise regression model

step.model <- stepAIC(full.model, direction = "both",

trace = FALSE)

summary(step.model)

AIC(step.model)

residual.plots(step.model)

plot(Life.expectancy ~., data = df1\_complete)

abline(step.model)

# forward regression model

foward.model <- stepAIC(full.model, direction = "forward",

trace = FALSE)

summary(foward.model)

residual.plots(foward.model)

# backward regression model

backward.model <- stepAIC(full.model, direction = "backward",

trace = FALSE)

summary(backward.model)

CV(backward.model)

residual.plots(backward.model)

# assumptions testing

par(mfrow = c(2, 2))

residual.plots(full.model)

```

#### 2.Should a country having a lower life expectancy value(<65) increase its healthcare expenditure in order to improve its average lifespan?

We assume that the collected data set is not accurate for percentage expenditure column.

If the definition of percentage expenditure is Expenditure on health as a percentage of Gross Domestic Product per capita(%), then it can never be more than 100%.

Also, it is higly unlikely that any country would spend 100% of its GDP on healthcare.

So, ignoring percentage expenditure column in analysis. We are using only Total expenditure for finding out if it has an impact on life expectancy value.

From EDA and 2-sample t-test, we see that life expectancy value does not have statisctical significance on healthcare expenditure.

t = 1.9583,

df = 181,

p-value = 0.05173

95 percent confidence interval: -0.00710969, 1.88675914

Since 0 is one of the plausible values, so we can say that effect of Total expenditure on life expectancy value greater than 65 and less than 65 is not statistically significant.

```{r}

#2-sample t-test for Life Expectancy value:

res.total.expenditure <- t.test(Total.expenditure ~ Life.expectancy.category , data = df1\_complete, var.equal = TRUE)

res.total.expenditure

# Fitting the tree model for classifying life expectancy value based on healthcare expenditure

df1\_complete$Life.expectancy.category <- as.factor(df1\_complete$Life.expectancy.category)

healthcare.expenditure.tree <- tree(formula = Life.expectancy.category ~ Total.expenditure, data = df1\_complete)

# summary statistics of healthcare.expenditure.model for life expectancy

summary(healthcare.expenditure.tree)

plot(healthcare.expenditure.tree, pretty =0)

```

#### 3.How does Infant and Adult mortality rates affect life expectancy?

From EDA, effect of Infant death does not look significant on life expectancy. But we see that Adult.Mortality rate is negatively coreelated Life Expectancy,

The relationship between Adult.Mortality and life expectancy can be modeled by the regression equation below:

life expectancy = 80.64428 + -0.06125 (Adult Mortality)

We notice that as Adult.Mortality value increase by an one, life expectancy is expected to decrese by 0.06125 years. With no Adult.Mortality, a life expectancy is exepcted to be 80.64428 years.

Adjusted R^2: 0.5732

And when we used Adult.Mortality and Infant death together, it did not meet linear regression model assumption.

```{r}

#Fitting Adult.Mortality model

Adult.Mortality.model <- lm(Life.expectancy ~Adult.Mortality, data = df1\_complete)

summary(Adult.Mortality.model)

# graph the regession

p.Adult.Mortality <- ggplot(df1\_complete, aes(x = Adult.Mortality , y = Life.expectancy )) + geom\_point() + stat\_smooth(method = "lm", col = "red")

ggplotly(p.Adult.Mortality)

# assumptions testing

par(mfrow = c(2, 2))

plot(Adult.Mortality.model)

#Fitting model for both Adult.Mortality and Infant death

Adult.infant.model <- lm(Life.expectancy ~Adult.Mortality+infant.deaths, data = df1\_complete)

summary(Adult.infant.model)

# assumptions testing

par(mfrow = c(2, 2))

plot(Adult.infant.model)

```

#### 4.Does Life Expectancy has positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol etc.

We assume that BMI rate is related to lifestyle, eating habits, exercise. Since we see a positive co-relation of BMI with Life Expectancy, we assume that with good eating habits, ample exercice and healthy lifestyle life expectancy would be more.

The summary statistics shows high significants of alpha less than 0.001. The relationship can be modeled by the regression equation below:

life expectancy = 63.57 + 0.19423 (BMI)

We notice that as BMI values increase by an unit, life expectancy is expected to increase by 0.19423 years.

With no BMI, a life expectancy is exepcted to be 63.57 years (which practially does not make sense).

Adjusted R^2: 0.2226

Note: Alcohol effect is covered in question: 6.

```{r}

#Fitting BMI model

BMI.model <- lm(Life.expectancy ~BMI, data = df1\_complete)

summary(BMI.model)

# graph the regession

p.BMI <- ggplot(df1\_complete, aes(x = BMI , y = Life.expectancy )) + geom\_point() + stat\_smooth(method = "lm", col = "red")

ggplotly(p.BMI)

# assumptions testing

par(mfrow = c(2, 2))

plot(BMI.model)

```

#### 5.What is the impact of schooling on the lifespan of humans?

Life expectancy and schooling have a positive linear relationship. Despite not being significant in the full model, schooling is a significant indicator when its modeled as a singly linear regression.

The summary statistics shows high significants of alpha less than 0.001. The relationship can be modeled by the regression below:

life expectancy = 41.42 + 2.34 (schooling)

We notice that as schooling increases by a year, life expectancy is increased by 2.5 years, with no schooling having a life expectancy of 38.72 years. Adjusted R^2: 0.5949

```{r}

# linear model

school.model <- lm(Life.expectancy ~Schooling, data = df1\_complete)

# summary statistics of schooling vs. life expectancy

summary(school.model)

# graph the regession

p.school <- ggplot(Life\_Expectancy\_Df\_2014, aes(x = Schooling , y = Life.expectancy )) + geom\_point() + stat\_smooth(method = "lm", col = "red")

ggplotly(p.school)

# assumptions testing

par(mfrow = c(2, 2))

plot(school.model)

```

#### 6.Does Life Expectancy have positive or negative relationship with drinking alcohol?

Life expectancy has a slight positive increase with alcohol consumption.

life expectancy = 68.03 + 1.07(Alcohol)

Notice that as alcohol consumption increases, life expectancy incrases by 1.07 years, starting at 68.01 years expected if no alcohol is consumed.

This is a bit counter intuitive considering the knowledge that alcohol is not considered to be healthy and many studies suggest that alcohol could shorten life spans. Keeping this in coonsideration, additional studies may need to be conducted.

In addition, the model does not meet the require assumptions for linear regression. Specifically, the model lacks constant variance, hence it is not correct.

```{r}

# linear model

alcohol.model <- lm(Life.expectancy ~Alcohol, data = df1\_complete)

# summary statistics of schooling vs. life expectancy

summary(alcohol.model)

# graph the regession

p.alcohol <- ggplot(Life\_Expectancy\_Df\_2014, aes(x = Alcohol , y = Life.expectancy )) + geom\_point() + stat\_smooth(method = "lm", col = "red")

ggplotly(p.alcohol)

par(mfrow = c(2, 2))

plot(alcohol.model)

```

#### 7.Do densely populated countries tend to have lower life expectancy?

Initally, there is a significant outlier that skews the data dramatically. Once removed, we noticed the sloped droped by nearly 50%. Considering that there is the possibility of having significanlt high populations, it has been concluded to keep the data point in.

life expectancy = 71.71 - 1.12e-8(Population)

Notice that as the population in a country increases, life expectancy decreases by 2.65 years, starting at 70.58 years expected if there is no population. In this scenario, the intercept independent of the slope has no logical reasoning considering that no population would result in no life expectancy.

We notice again that the model fails constant variance, hence, the model cannot be utilized.

```{r}

# linear model

popultion.model <- lm(Life.expectancy ~Population, data = df1\_complete)

# summary statistics of schooling vs. life expectancy

summary(popultion.model)

# graph the regession

p.population <- ggplot(Life\_Expectancy\_Df\_2014, aes(x = Population , y = Life.expectancy )) + geom\_point() + stat\_smooth(method = "lm", col = "red")

ggplotly(p.population)

# assumptions testing

par(mfrow = c(2, 2))

plot(popultion.model)

# remove the outlier

attach(df1\_complete)

# sort by mpg

newdata <- df1\_complete %>% arrange(desc(Population))

# drop outlier

newdata = newdata[-1,]

# linear model

popultion.model2 <- lm(Life.expectancy ~Population, data = newdata)

# summary statistics of schooling vs. life expectancy

summary(popultion.model2)

# graph the regession

p.population2 <- ggplot(newdata, aes(x = Population , y = Life.expectancy )) + geom\_point() + stat\_smooth(method = "lm", col = "red")

ggplotly(p.population2)

# assumptions testing

par(mfrow = c(2, 2))

plot(popultion.model2)

```

#### 8.What is the impact of Immunization coverage on life Expectancy?

Reviewing the graphs, there does not appear to be a significant relationship between life expectancy and immunization.

1. linear regession is not applicable due to the fact that the data doesn't match the required assumptions.

- Fails constant variance

2. log transformations also did not satisfy the required assumptions for regression

- Fails constant variance

3. polynomial regression matched the required assumptions with a low R^2 (0.14) and high AIC (1280)

Life.expectancy = 68.85 - 0.27 Hepatitis.B + 0.003 I(Hepatitis.B^2)

```{r}

# 1. LINEAR REGRESSION

# list of immunizations

df\_immunizations <- df1\_complete %>% select("Life.expectancy", "Hepatitis.B", "Polio", "Diphtheria")

# linear model

immunization.model <- lm(Life.expectancy ~., data = df\_immunizations)

# summary statistics of schooling vs. life expectancy

summary(immunization.model)

# graph the regession

require(gridExtra)

p1 <- ggplot(Life\_Expectancy\_Df\_2014, aes(x = Hepatitis.B , y = Life.expectancy )) + geom\_point()

p2 <- ggplot(Life\_Expectancy\_Df\_2014, aes(x = Polio , y = Life.expectancy )) + geom\_point()

p3 <- ggplot(Life\_Expectancy\_Df\_2014, aes(x = Diphtheria , y = Life.expectancy )) + geom\_point()

grid.arrange(p1, p2, p3, ncol=2)

# assumptions testing

residual.plots(immunization.model)

# 2. LOG TRANSFORMATIONS

# create logged transformations

df\_immunizations$log.Life.expectancy <- log(df\_immunizations$Life.expectancy)

df\_immunizations$log.Hepatitis.B <- log(df\_immunizations$Hepatitis.B)

# log-log linear model

log.immunization.model <- lm(log.Life.expectancy ~log.Hepatitis.B, data = df\_immunizations)

# summary statistics

summary(log.immunization.model)

# graph regression

ggplot(df\_immunizations, aes(x = log.Hepatitis.B , y = log.Life.expectancy )) + geom\_point()

# residual plots

residual.plots(log.immunization.model)

# 3. POLYNOMIAL REGRESSION TESTING

# linear model

ploy.immunization.model <- lm(Life.expectancy ~Hepatitis.B+I(Hepatitis.B^2), data = df\_immunizations)

# summary statistics

summary(ploy.immunization.model)

AIC(ploy.immunization.model)

BIC(ploy.immunization.model)

# graph regression

ggplot(df\_immunizations, aes(x = Hepatitis.B , y = Life.expectancy )) + geom\_point() + stat\_smooth(method = "loess", formula = y ~ x, size = 1) + ggtitle("Life Expectancy vs. Heptatitis B with trend line")

ggplot(df\_immunizations, aes(x = Hepatitis.B , y = Life.expectancy )) + geom\_point() + stat\_smooth(method = "lm", formula = y ~ x + I(x^2), size = 1) + ggtitle("Life Expectancy vs. Heptatitis B with quadratic regression line")

# residual plots

residual.plots(ploy.immunization.model)

```

## Objective 1

Display the ability to build regression models using the skills and discussions from Unit 1 and 2 with the purpose of identifying key relationships, interpreting those relationships, and making good predictions.

\*Reminder, key here is to tell a good story.\*

### Build Model 1

- Identify key relationships

- Ensure interpretability

Unit 2 Objectives:

- bias vs. variance

- complexity

- LASSO/LARS, CV (cross validation)

1. Perform regression analysis

- LASSO

Using the lasso regression method, the following were determined to be significant

Income.composition.of.resources

While there were other variables that made the model (total expenditure, HIV.AIDs, and BMI), their coefficients are extremely small and not considered to be significant.

Life expectancy = 37.72 + 49.11 (income)

R^2: 0.7317

CV: 9.896252

Notice that there is a drop in R^2 as opposed to the linear model, but there is significantly less complexity.

```{r}

# LASSO

## set the seed to make your partition reproducible

smp\_size <- floor(0.5 \* nrow(df1\_complete))

set.seed(1234)

index <- sample(seq\_len(nrow(df1\_complete)), size = smp\_size)

train<-df1\_complete[index,]

test<-df1\_complete[-index,]

#Formatting data for GLM net

x=model.matrix(Life.expectancy~.,train)[,-1]

y=(train$Life.expectancy)

xtest<-model.matrix(Life.expectancy~.,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

# review coefficients

coef(lasso.mod,s=bestlambda)

# plot variable analysis

plot(lasso.mod,xvar="lambda",label=TRUE)

# review R^2 results

eval\_results(ytest, lasso.pred, test)

# Fit linear regresion to model

lasso.linear.model <- lm(Life.expectancy ~Income.composition.of.resources, data = df1\_complete)

summary(lasso.linear.model)

residual.plots(lasso.linear.model)

ggplot(df1\_complete, aes(x = Income.composition.of.resources , y = Life.expectancy )) + geom\_point() + stat\_smooth(method = "loess", formula = y ~ x, size = 1) + ggtitle("Life Expectancy vs. Income with trend line")

ggplot(df1\_complete, aes(x = Income.composition.of.resources , y = Life.expectancy )) + geom\_point() + stat\_smooth(method = "lm", formula = y ~ x, size = 1) + ggtitle("Life Expectancy vs. Income with regression line")

```

2. Report predictive ability

a. Test/train set

b. CV data

```{r}

# Create Training and Test data -

set.seed(100) # setting seed to reproduce results of random sampling

trainingRowIndex <- sample(1:nrow(df1\_complete), 0.8\*nrow(df1\_complete)) # row indices for training data

trainingData <- df1\_complete[trainingRowIndex, ] # model training data

testData <- df1\_complete[-trainingRowIndex, ] # test data

# Build the model on training data -

lmMod <- lm(Life.expectancy ~Income.composition.of.resources, data=trainingData) # build the model

distPred.confidence <- predict(lmMod, testData, interval = 'confidence', level=0.95) # predict distance

distPred.prediction <- predict(lmMod, testData, interval = 'prediction') # predict distance

confint(lmMod)

summary (lmMod) # model summary

# accuracy statistics

actuals\_preds <- data.frame(cbind(actuals=testData$Life.expectancy, predicteds=distPred.confidence)) # make actuals\_predicteds dataframe.

correlation\_accuracy <- cor(actuals\_preds) # 82.7%

AIC(lmMod)

BIC(lmMod)

CV(lmMod)

predict(lmMod, testData, interval="confidence")

```

Comparing the models

```{r}

#train/test all models

train.full.model <- lm(Life.expectancy ~., data = trainingData)

# Stepwise regression model

trian.step.model <- stepAIC(train.full.model, direction = "both",

trace = FALSE)

distPred <- predict(trian.step.model, testData)

actuals\_preds <- data.frame(cbind(actuals=testData$Life.expectancy, predicteds=distPred)) # make actuals\_predicteds dataframe.

correlation\_accuracy <- cor(actuals\_preds)

# Forward regression model

trian.for.model <- stepAIC(train.full.model, direction = "forward",

trace = FALSE)

distPred.for <- predict(trian.for.model, testData)

actuals\_preds.for <- data.frame(cbind(actuals=testData$Life.expectancy, predicteds=distPred.for)) # make actuals\_predicteds dataframe.

correlation\_accuracy.for <- cor(actuals\_preds.for)

summary(lasso.linear.model)

# dataframe to hold all metrics

df.models1 <- data.frame(Models=c('forward', 'stepwise', 'lasso'))

df.models1$CV <- c(CV(foward.model)[[1]],CV(step.model)[[1]], CV(lasso.linear.model)[[1]])

df.models1$AIC <- c(CV(foward.model)[[2]],CV(step.model)[[2]], CV(lasso.linear.model)[[2]])

df.models1$BIC <- c(CV(foward.model)[[4]],CV(step.model)[[4]], CV(lasso.linear.model)[[4]])

df.models1$AdjR2 <- c(CV(foward.model)[[5]],CV(step.model)[[5]], CV(lasso.linear.model)[[5]])

df.models1$Accuracy <- c(0.78,0.79, 0.76)

```

4. Interpret the coefficients

Life expectancy has a linear relationship iwth income/composition of resource. where as the income index increases, life expectancy increases by 50.2. It should be noted that income/composition of resources ranges from 0 to 1, where the maximum life expectancy if 87.3. This is not fully realistic considering that many may live past this age. In addition, if there is no income or composition of resources, it is expected that the life expectancy is 37.1. While this also may not be applicable, this could pertain to those who are considered unemployed without any income source.

5. Confidence intervals

```{r}

distPred.confidence <- predict(lmMod, testData, interval = 'confidence', level=0.95) # predict distance

distPred.prediction <- predict(lmMod, testData, interval = 'prediction') # predict distance

confint(lmMod)

```

6. Practical and statistical significance

The income index is the most significant predictor for life expectancy, explaining more than 70% of the data.

### Model 2

- Product the best predictions as possible

- Interpretation is no longer required, hence complexity is no longer an issue

1. Feature selection to avoid overfitting

A. Linear Regression

- model a: linear regression

life expectancy = 36.55 + 50.73(income)

Adjusted R^2 = 0.79

- model b: linear regression + adult mortality

life expectancy = 48.5 + 38.77(income) - 0.025 (adult mortality)

Adjusted R^2 = 0.84

- model c: linear regression + adult mortality + HIV.AIDS

life expectancy = 49.8 + 36.05 (income) - 0.016 (adult mortality) - 0.95 (HIV/AIDS)

Adjusted R^2 = 0.85

B. Interaction Terms

- model d: linear regression + adult mortality + HIV.AIDS

```{r}

# Lasso model lasso.mod determined that income the primary predictor for life expectancy

# SIMPLE MODEL

simple.model <- lm(Life.expectancy ~Income.composition.of.resources, data = df1\_complete)

# Stepwise regression model

simple.step.model <- stepAIC(simple.model, direction = "both",

trace = FALSE)

summary(simple.step.model)

# ADDITION OF VARIABLES MODEL

# adult mortality

simple.model <- lm(Life.expectancy ~Income.composition.of.resources + Adult.Mortality , data = df1\_complete)

summary(simple.model)

# Thinness

df.expenditure <- df1\_complete %>% filter(df1\_complete$percentage.expenditure < 100)

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

summary(simple.model)

# AIDS

simple.model <- lm(Life.expectancy ~Income.composition.of.resources + Adult.Mortality + HIV.AIDS, data = df1\_complete)

summary(simple.model)

# Interaction terms

simple.model <- lm(Life.expectancy ~Income.composition.of.resources + Adult.Mortality + HIV.AIDS + (Adult.Mortality\*HIV.AIDS) + (Income.composition.of.resources\*Adult.Mortality) +(Income.composition.of.resources\*HIV.AIDS), data = df1\_complete)

summary(simple.model)

```

2. Create the model

```{r}

complex.model <- lm(Life.expectancy ~Income.composition.of.resources + Adult.Mortality +(Income.composition.of.resources\*HIV.AIDS), data = trainingData)

summary(complex.model)

residual.plots(complex.model)

```

3. Compare model 1 vs. model 2

```{r}

distPred.complex <- predict(complex.model, testData)

actuals\_preds.complex <- data.frame(cbind(actuals=testData$Life.expectancy, predicteds=distPred.complex)) # make actuals\_predicteds dataframe.

correlation\_accuracy.complex <- cor(actuals\_preds.complex)

confint(complex.model)

df.models <- data.frame(Models=c('model1', 'model2'))

df.models$CV <- c(CV(lasso.linear.model)[[1]], CV(complex.model)[[1]])

df.models$AIC <- c( CV(lasso.linear.model)[[2]], CV(complex.model)[[2]])

df.models$BIC <- c(CV(lasso.linear.model)[[4]], CV(complex.model)[[4]])

df.models$AdjR2 <- c(CV(lasso.linear.model)[[5]], CV(complex.model)[[5]])

df.models$Accuracy <- c( 0.76, 0.85)

print(df.models)

```

4. Comment on the differences of the models and whether model 2 brings any benefit

We notice that model 2 (the more complex of the two models) has a lower CV PRESS and higher adjusted R2. While model 1 is simple to comprehend, model 2 has higher predictability powers.

## Objective 2

- Nonparametric technique

- kNN or regression trees (select one)

Set of predictors from previous regression: (fill this out)

1. Model

```{r}

# RANDOM FOREST

# all variables

rf <- randomForest(

Life.expectancy ~ .,

data=trainingData

)

varImpPlot(rf)

pred = predict(rf, newdata=testData)

pred.df <- as.data.frame(pred)

pred.df$true <- testData$Life.expectancy

pred.df$idu <- as.numeric(row.names(pred.df))

ggplot() + geom\_point(data=pred.df, aes(x = idu, y = true ), color='blue')+ geom\_point(data=pred.df, aes(x = idu, y = pred ), color='red')

plot(rf)

print(rf)

# limited variables

new.train <- trainingData %>% select(Adult.Mortality,HIV.AIDS, Income.composition.of.resources, Life.expectancy)

new.test <- testData %>% select(Adult.Mortality,HIV.AIDS, Income.composition.of.resources, Life.expectancy)

new.rf <- randomForest(

Life.expectancy ~ .,

data=new.train

)

varImpPlot(new.rf)

new.pred <- predict(new.rf, newdata=new.test)

new.pred.df <- as.data.frame(new.pred)

new.pred.df$true <- new.test$Life.expectancy

new.pred.df$idu <- as.numeric(row.names(new.pred.df))

ggplot() + geom\_point(data=new.pred.df, aes(x = idu, y = true ), color='blue')+ geom\_point(data=new.pred.df, aes(x = idu, y = pred ), color='red')

plot(new.rf)

print(new.rf)

new.rf$confusion

class.error

```