Final Project

Zhaojin Zhu, Zhijia Ju, Yifei Wang, Zhiyi Xie

2020/11/29

github website

https://github.com/GGroup9/FinalProject.git https://github.com/GGroup9/FinalProject

Abstract

This project discusses the factors influencing life expectancy based on the dataset from 2000 to 2015 for 193 countries from the World Health Organization (WHO). This dataset focuses on immunization factors, mortality factors, economic factors, social factors and other health-related factors. Assumptions for regression analysis are diagnosed to make sure no violation is presented. Data is analyzed through different linear regression methods and the final model is obtained with proper model adequacy checking performed. Model validation is performed through cross-validation and the result is satisfied. We conclude that predictors like Diphtheria coverage, adult mortality rate, infant deaths rate, HIV/AIDS rate and expenditure on health percentage have a relationship with life expectancy, and they could be used to predict life expectancy. Furthermore, country status has an impact on lifespan as well. Developing countries have overall lower life expectancy than developed countries, which is as expected.

Introduction

Nowadays the living standards have improved significantly in many ways. Factors including economy, education and medical condition all have developed to a new stage after the Millennium. Previous researches did not include immunization factor and human development index when analyzing factors affecting life expectancy.That brought to our attention are immunization and human development index significant in influencing life expectancy. And, what other factors would actually affect life expectancy that were not included in the past studies. If immunization factor turns out to be significant in influencing life expectancy, could lifespan be improved if government spend more expenditure on healthcare? Furthermore, would personal habits and lifestyle have relationship with life expectancy? Would economical factor have an effect on life expectancy, that is, do developing countries have lower life expectancy than developed countries? What are the potential causes for a country to have a low life expectancy? We will describe the dataset by data

visualization and analyze the various factors through different regression models to find out the significant predictors that have impact on predicting the life expectancy.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

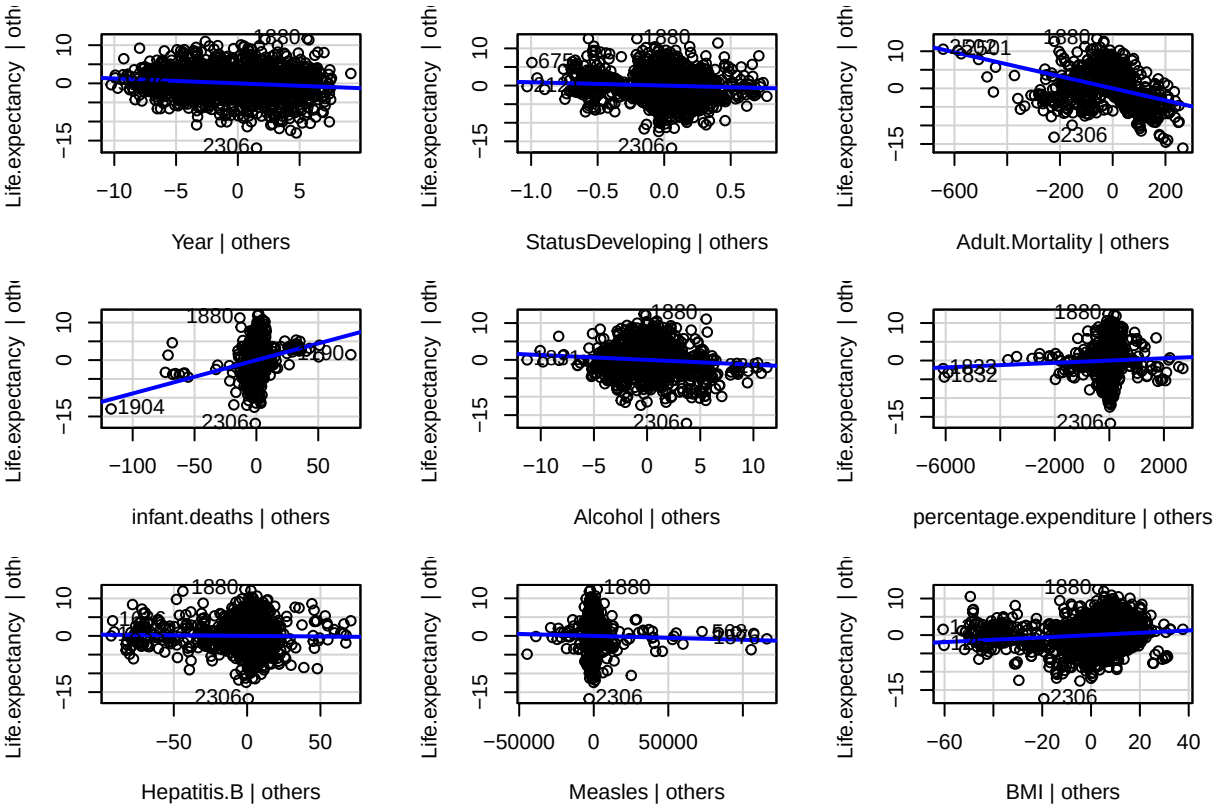
```
library(caret)
```

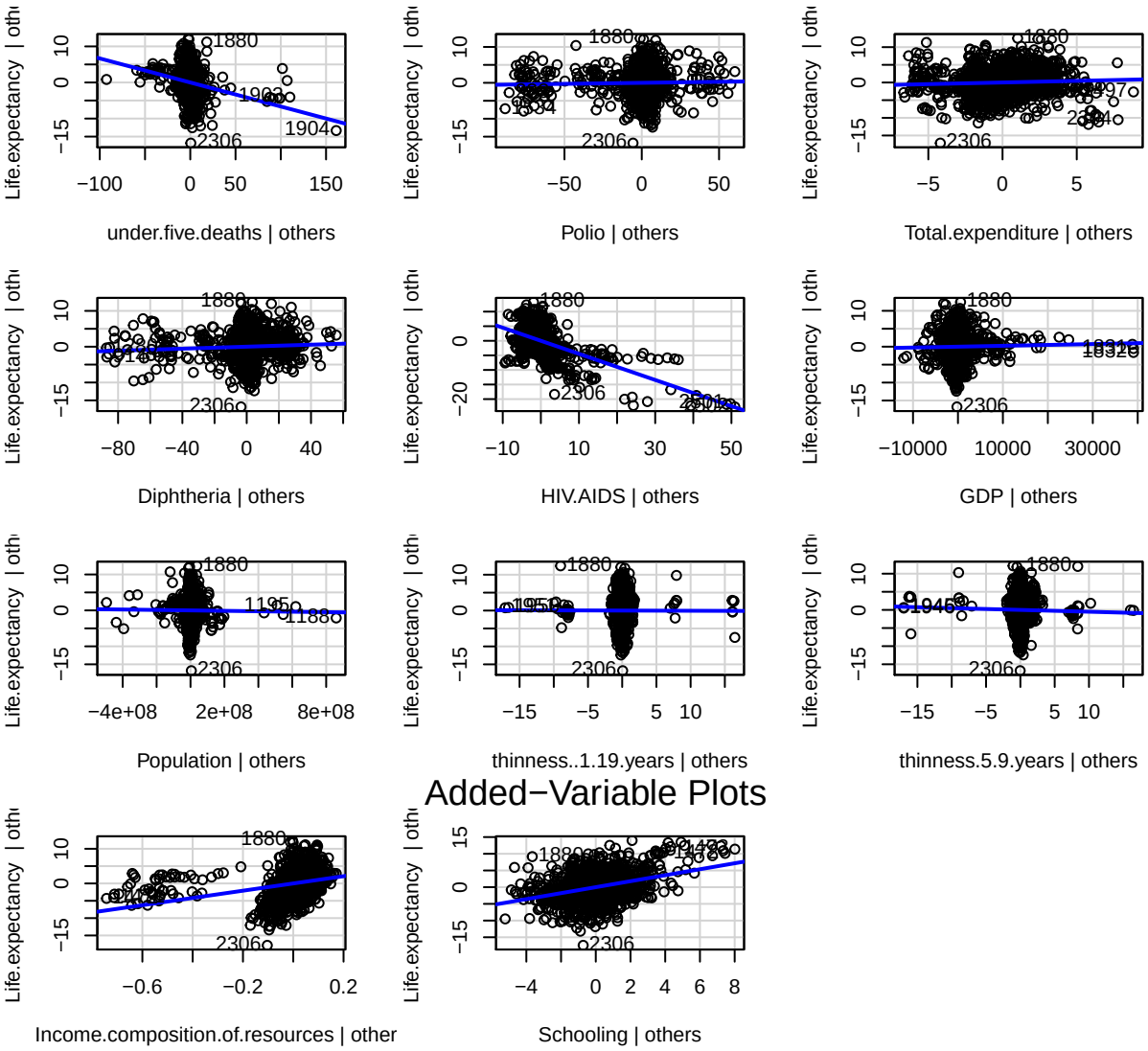
```
## Loading required package: lattice
```

```
##
```

```
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
who <- read.csv(file = "Life Expectancy Data.csv", header=TRUE)

attach(who)

#remove Country since we are not focusing on specific countries in this research
who <- who[,-c(1)]
# exclude missing values (Remove Rows with Missing Data)
who = who[complete.cases(who),]
# 1289 observations were deleted
# 1650 observations remaining

model <- lm(Life.expectancy~., data = who)

avPlots(model)
```

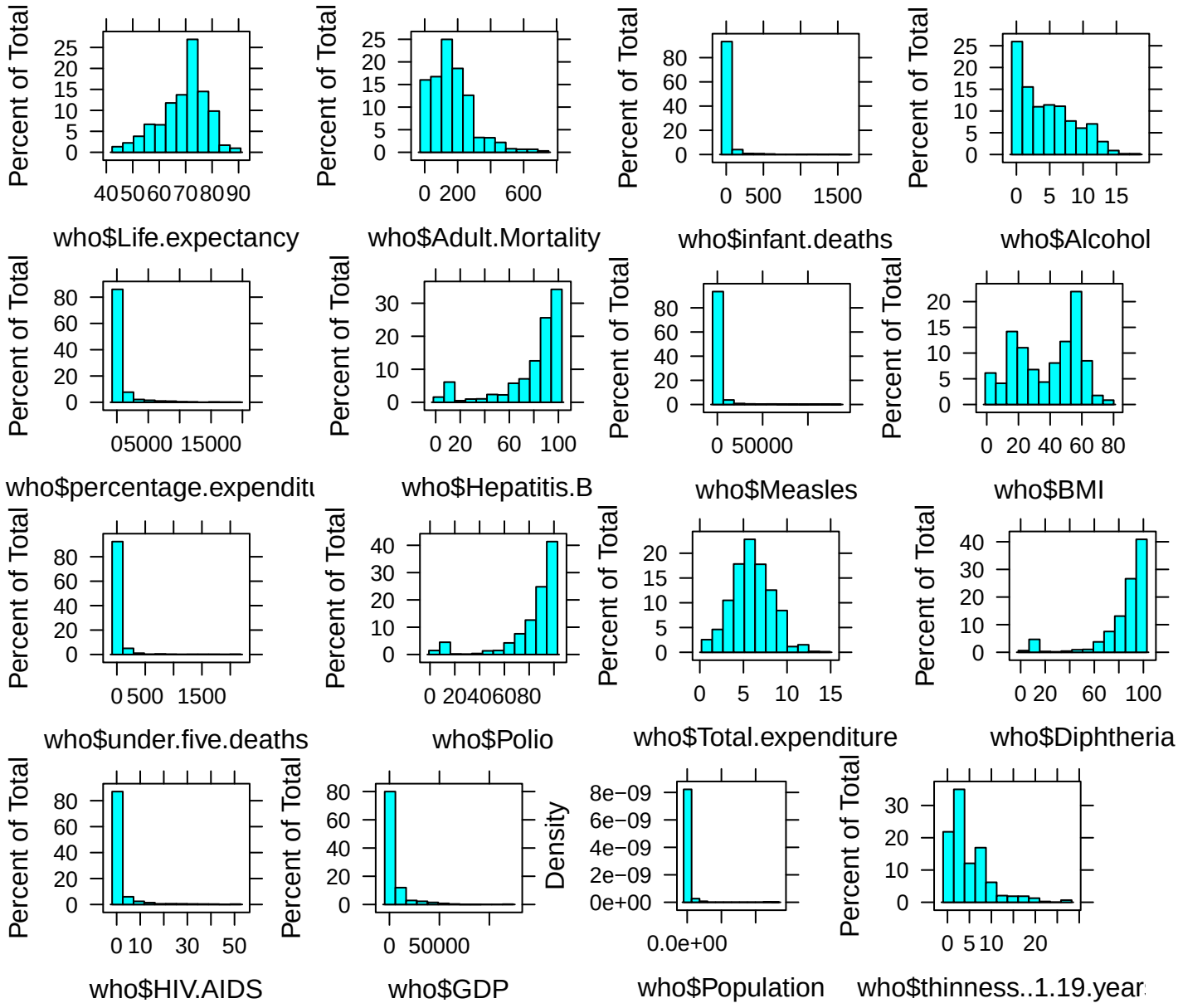


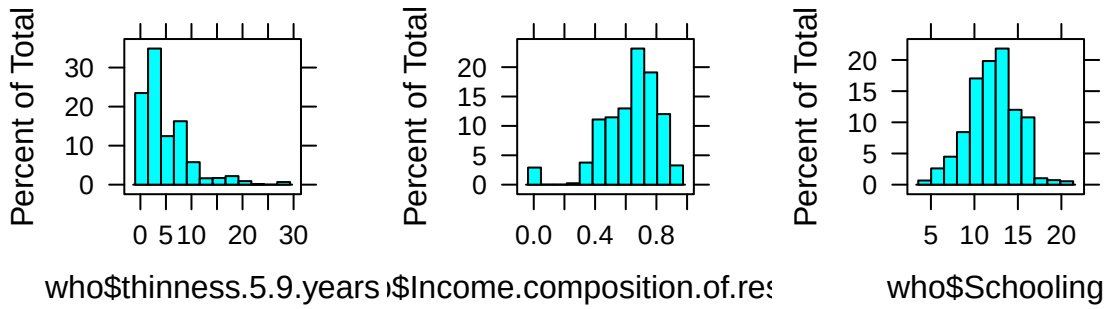



```
#check pairwise correlations
cor(who[,-c(1,2)])
```

```
par(mar = c(4, 4, .1, .1))
histogram(who$Life.expectancy)
histogram(who$Adult.Mortality)
histogram(who$infant.deaths)
histogram(who$Alcohol)
histogram(who$percentage.expenditure)
histogram(who$Hepatitis.B)
histogram(who$Measles)
histogram(who$BMI)
histogram(who$under.five.deaths)
```

```r
histogram(who$Polio)
histogram(who$Total.expenditure)
histogram(who$Diphtheria)
histogram(who$HIV.AIDS)
histogram(who$GDP)
histogram(who$Population)
histogram(who$thinness..1.19.years)
histogram(who$thinness.5.9.years)
histogram(who$Income.composition.of.resources)
histogram(who$Schooling)
```





Data Description

Our dataset consists of 22 Columns and 2938 rows, including 20 predictors that are divided into 4 broad categories: Immunization related factors, Mortality factors, Economical factors and Social factors.

Immunization related factors contains Hepatitis B (HepB immunization coverage among 1-year-old), Measles (number of reported cases per 1000 population), Polio (Pol3 immunization coverage among 1-year-old), Diphtheria (Diphtheria tetanus toxoid and pertussis, DTP3, immunization coverage among 1-year-old), HIV AIDS (Deaths per 1000 live births HIV/AIDS from 0 to 4 years old).

Mortality factors contains Adult Mortality (Adult Mortality Rates of both sexes), Infant Deaths (Number of Infant Deaths per 1000 population), Under Five Deaths (Number of under-five deaths per 1000 population).

Economical factors contains Country, Status (developing or developed), Percentage Expenditure (Expenditure on health as a percentage of Gross Domestic Product per capita), Total Expenditure (General government expenditure on health as a percentage of total government expenditure), GDP (Gross Domestic Product per capita), Population (Population of the country), Income Composition of Resources (Human Development Index in terms of income composition of resources).

Social factors contains Year (the year of the data), Alcohol (recorded per capita (15+) consumption), BMI (Average Body Mass Index of entire population), 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), Schooling (Number of years of Schooling).

If we consider the Country predictor as a detailed subset of the Status predictor, we could remove the Country predictor and only focus on the developing and developed status for constructing a general idea regardless of a specific country. However, if we want to know the life expectancy of a specific country, we could include the specific country value from the Country predictor.

We would like to introduce one new additional data point into our dataset. The values for this unique data point is as follows. Country:China, Year:2015, Status:Developing, Life expectancy:74.2625, Adult Mortality:73.75, infant deaths: 294.875, Alcohol:4.182, percentage.expenditure: 78.48934709, Hepatitis.B: 80.4375, Measles: 65857.9375, BMI: 21.80625, under.five.deaths: 350, Polio: 93.6875, Total.expenditure:

4.918, Diphtheria: 93.3125, HIV.AIDS: 0.1, GDP: 2345.303158, Population: 321812.0625, thinness..1.19.years: 4.6375, thinness.5.9.years: 4.025, Income.composition.of.resources: 0.66025, Schooling: 11.4375. These values are calculated based on the average values for each predictor from 2000 to 2015 for the country of China. Our modified dataset now consists of 22 Columns and 2939 rows.

From the added variable plots, we can see that predictors like infant.deaths, percentage.expenditure, Measles, under.five.deaths, GDP, Population, thinness..1.19.years and thinness.5.9.years seem to have no significant linear relationship with life expectancy. From the pairwise correlation data, there are several variables collinear with each other (they have correlation close to 1). These highly collinearly related variable pairs are under.five.deaths and infant.deaths, GDP and percentage.expenditure, thinness.5.9.years and thinness..1.19.years.

The histograms generated show that Life.expectancy, Total.expenditure, Income.composition.of.resources and Schooling are bell-shaped and distributed roughly equal from left to right. The distribution graphs for infant.deaths, percentage.expenditure, Measles, under.five.deaths, HIV.AIDS, GDP and Population have one bar that is dominant, indicate that the majority countries have similar conditions. Graphs for Hepatitis.B, Polio and Diphtherian show that the majority countries have a high coverage for immunization.

Method

To test if each predictor has linear relationship with life expectancy, we use t-test to investigate.

$H0 : beta_j = 0$
$H1 : beta_j != 0$

test statistics: $t0$
We reject H0 if $pvalue > alpha = 0.05$

From the OLS summary table, we see that the p-values for percentage.expenditure, Hepatitis.B, Measles, Polio, GDP, Population, thinness 1-19 years and thinness 5-9 years are higher than alpha=0.05, which means further investigation is needed to determine whether these predictors are significant or not in predicting the life expectancy. Will checked the diagonal of the hat matrix, only two observations have leverages higher than 0.2 , that may have the potential to be influential. However, given the large dataset, the rest of the observations have small leverages. Since none of these observations had large residuals, so we need to investigate a little more.

From stepwise regression, we get a final model that has 14 out of the 20 possible predictors. The figure sizes have been customized so that it's easily to put two images side-by-side.

Summary model

```
#checked the  model
summodel <-  summary(model)
#percentage.expenditure and Hepatitis.B and Measle and Polio and GDP and Population
#thinness..1.19.years  and thinness.5.9.years larger than 0.05
```

Unit length scaling

Due to different scaling in the dataset 'who' , it is difficult to compare regression coefficients because the magnitude of betaj reflects the units of measurement of the regressor xj. For this reason it is helpful to work with the scaled regressors and response variable. Here we will use Unit Length Scaling:

```
#Life.expectancy and GDP
gdp <- lm( Life.expectancy ~ GDP , data = who )
summary(gdp)

## standardize data using unit length scaling ##
#unit length scalings
s_Year<-sqrt(sum((who$Year-mean(who$Year))^2))
#s_Status<-sqrt(sum((Status-mean(Status))^2))
s_Adult.Mortality = sqrt(sum((who$Adult.Mortality-mean(who$Adult.Mortality))^2))
s_infant.deaths<-sqrt(sum((who$infant.deaths-mean(who$infant.deaths))^2))
s_Alcohol<-sqrt(sum((who$Alcohol-mean(who$Alcohol))^2))
s_percentage.expenditure<-sqrt(sum((who$percentage.expenditure-mean(who$percentage.expenditure))^2))
s_Hepatitis.B<-sqrt(sum((who$Hepatitis.B-mean(who$Hepatitis.B))^2))
s_Measles<-sqrt(sum((who$Measles-mean(who$Measles))^2))
s_BMI<-sqrt(sum((who$BMI-mean(who$BMI))^2))
s_under.five.deaths<-sqrt(sum((who$under.five.deaths-mean(who$under.five.deaths))^2))
s_Life.expectancy = sqrt(sum((who$Life.expectancy-mean(who$Life.expectancy))^2))

s12  = sqrt(sum((who$Polio - mean(who$Polio))^2))
s13  = sqrt(sum((who$Total.expenditure - mean(who$Total.expenditure))^2))
s14  = sqrt(sum((who$Diphtheria - mean(who$Diphtheria))^2))
s15  = sqrt(sum((who$HIV.AIDS - mean(who$HIV.AIDS))^2))
s16  = sqrt(sum((who$GDP - mean(who$GDP))^2))
s17  = sqrt(sum((who$Population - mean(who$Population))^2))
s18  = sqrt(sum((who$thinness..1.19.years - mean(who$thinness..1.19.years))^2))
```

```r
s19  = sqrt(sum((who$thinness.5.9.years - mean(who$thinness.5.9.years))^2))
s20  = sqrt(sum((who$Income.composition.of.resources -mean(who$Income.composition.of.resources))^2))
s21  = sqrt(sum((who$Schooling - mean(who$Schooling))^2))

z_Year<-(who$Year-mean(who$Year))/s_Year
#z_Status<-(Status-mean(Status))/s_Status
z_Adult.Mortality<-(who$Adult.Mortality-mean(who$Adult.Mortality))/s_Adult.Mortality
z_infant.deaths<-(who$infant.deaths-mean(who$infant.deaths))/s_infant.deaths
z_Alcohol<-(who$Alcohol-mean(who$Alcohol))/s_Alcohol
z_percentage.expenditure<-(who$percentage.expenditure-mean(who$percentage.expenditure))/s_percentage.ex
z_Hepatitis.B<-(who$Hepatitis.B-mean(who$Hepatitis.B))/s_Hepatitis.B
z_Measles<-(who$Measles-mean(who$Measles))/s_Measles
z_BMI<-(who$BMI-mean(who$BMI))/s_BMI
z_under.five.deaths<-(who$under.five.deaths-mean(who$under.five.deaths))/s_under.five.deaths
Life.expectancy_s<-(who$Life.expectancy-mean(who$Life.expectancy))/s_Life.expectancy

z12 = (who$Polio - mean(who$Polio))/s12
z13 = (who$Total.expenditure - mean(who$Total.expenditure))/s13
z14 = (who$Diphtheria - mean(who$Diphtheria))/s14
z15 = (who$HIV.AIDS - mean(who$HIV.AIDS))/s15
z16 = (who$GDP - mean(who$GDP))/s16
z17 = (who$Population - mean(who$Population))/s17
z18 = (who$thinness..1.19.years - mean(who$thinness..1.19.years))/s18
z19 = (who$thinness.5.9.years - mean(who$thinness.5.9.years))/s19
z20 = (who$Income.composition.of.resources - mean(who$Income.composition.of.resources))/s20
z21 = (who$Schooling - mean(who$Schooling))/s21

LifeSRModel<-lm(Life.expectancy_s~z_Year+z_Adult.Mortality+z_infant.deaths+z_Alcohol+z_percentage.expen

## variable selection ##
summary(LifeSRModel)
LifeSRModel$coefficients
```

After standardized and summary the model. We construct the hypothesis test:

F test :

2.2e-16 $<$ 0.05, reject H0, at least 1 predictor is useful.

T test :

for the predictors with p value > 0.05, we investigate them independently.

Based on the hypothesis test, we conclude that z_percentage.expenditure , z_Hepatitis.B , z_Measles , Polio , GDP , Population , thinness..1.19.years and thinness.5.9.years has p-value larger than 0.05.

SLM with life expectancy

Thus, we will use these 8 predictor to fit the SLM with Life.expectancy_s to see if they have relationships.

```
#Life.expectancy_s and percentage.expenditure
le <- lm( Life.expectancy_s ~ z_percentage.expenditure )
summary(le)

#Life.expectancy_s and Hepatitis.B
lh <- lm( Life.expectancy_s ~ z_Hepatitis.B )
summary(lh)

#Life.expectancy_s and Measles
lmea <- lm( Life.expectancy_s ~ z_Measles )
summary(lmea)

#Life.expectancy_s and Polio
lp <- lm( Life.expectancy_s ~ z12 )
summary(lp)

#Life.expectancy_s and GDP
lg <- lm( Life.expectancy_s ~ z16 )
summary(lg)

#Life.expectancy_s and Population
lpo <- lm( Life.expectancy_s ~ z17)
summary(lpo)

#Life.expectancy_s and thinness..1.19.years
lt1 <- lm( Life.expectancy_s ~ z18 )
summary(lt1)
```

```r
#Life.expectancy_s and thinness.5.9.years
lt5 <- lm( Life.expectancy_s ~ z19 )
summary(lt5)
```

Conclusion: Based on the model we fitted, the population has $p - value > 0.05$ when fitting simple linear regression line with life expectancy,so coefficient of population is close to 0. Thus population is not strongly linearly related with life expectancy. We can remove population from the model then.

Remove the population predictor from the model:

```r
#population remove from dataset 'who'
#with standardized data
LifeSRModel<-lm(Life.expectancy_s~z_Year+z_Adult.Mortality+z_infant.deaths+z_Alcohol+z_percentage.expend

summary(LifeSRModel)


#with original data
model0 <- lm(Life.expectancy~Year+Status+Adult.Mortality+infant.deaths+Alcohol+percentage.expenditure+H

summary(model0)
```

After remove the population predictor from the model, we find out the model contain the population and the model without population, their estimate coefficient and R-squared does not have significant changes. Thus, we conclude that population does not affect the Life expectancy.

Stepwise function

Here we use stepwise to select validate variable:

```r
##stepwise
#use stepwise method to complete variable selection
step(model, direction = 'both')


newmodel = lm(formula = Life.expectancy ~ Year + Status + Adult.Mortality +
    infant.deaths + Alcohol + percentage.expenditure + BMI +
    under.five.deaths + Total.expenditure + Diphtheria + HIV.AIDS +
    thinness.5.9.years + Income.composition.of.resources + Schooling,
    data = who)


summary(newmodel)
```

We removed 6 predictors from the model, since those predictors are not acting important roles in our model

fitting.

Confidence interval:

```
## confidence interval ##
confint( newmodel , level = 0.9 )
```

Residual analysis:

Then we do residual analysis on the new model generated.

```
## residual analysis ##


#Standardized Residuals vs. Index tut 4
plot( rstandard(newmodel))
```

```
#straight line around 0


n <- nrow( who )
plot( newmodel$residuals[1:(n-1)] , newmodel$residuals[2:n])
```

```
#straight line, linear relationship


plot(newmodel)
```

constant variance (if not good, use transformation of y) The first plot (Residuals vs fitted) has an approximately straight line at 0, while more data points appear on the right. Thus the model has generally met constant variance assumption.

linearity (if not good, use transformation of x) The second plot (Normal Q-Q plot) has an approximately straight line, thus the model generally met the linearity assumption.

Leverage and influence:

hat matrix: $H = X(X'X)^(- 1)X'$
hii:diagonal of hat matrix
If $hii > 2p/n$ then the ith observation can be considered a leverage point. Here we will use influence() function , using the model object as it's only argument.

```
#leverage and influence
2*ncol(who)/nrow(who)  #0.02545455
newmodel_influence <- influence(newmodel)
```

```r
sort(newmodel_influence$hat, decreasing = TRUE)
```

Based on the result, there are 68 observations have leverages higher than 0.02545455. Even though, many observation have larger leverage point than 0.02545455 but they are still very close to 0.02545455. Typically a point which has high leverage paired with a large residual it is likely to be influential. None of the observation had larger residuals, so we need to dig a little more.

Cook's distance:

$Di = (ri^2/p)(hii/(1 - hii))$
points with Di > 1 are considered to be influential.

```r
#cook's distance
newmodel_cook <- cooks.distance(newmodel)
sort(newmodel_cook, decreasing = TRUE)
```

We see that for our data and model we don't get any Di larger than 1.

outlier The observations 1880, 2306, 2300, 1901, 1902, 2503 are potential outliers. There's no point beyond cook's distance according to the leverage plot.

Our model contains 2 outliers, we will use robust regression to dampen the impact of highly influential observations and also violations in model assumptions.

Robust regression:

```r
#There is potential outlier, so we use robust
#robust
Rmodel = rlm(formula = Life.expectancy ~ Year + Status + Adult.Mortality +
    infant.deaths + Alcohol + percentage.expenditure + BMI +
    under.five.deaths + Total.expenditure + Diphtheria + HIV.AIDS +
    thinness.5.9.years + Income.composition.of.resources + Schooling,
    data = who, psi = psi.huber)

summary(Rmodel)


plot(Rmodel)
```

We want to investigate if there is a huge change in significance of regressors. percentage changed comparing with usual linear model fit: (use difference between coefficient of usual model and Robust model to divide coefficient of usual model) 0.14, 0.40, 0.19, 0.05, 0.03, 0.12, 0.12, 0.04, 0.22, 0.05, 0.06, 0.08, 0.02, 0.07 the

intercept has largest change, while the StatusDeveloping coefficient has a 40% decrease in magnitude of slope, which makes its significance decreases. Thus,the same variables do appear significant when compared to the usual linear model fit except for Status.

VIF:

```
##multicollinearity
#vif
vif(Rmodel)
```

We can see that infant.deaths and under.five.deaths have vif > 10, which means they are probably collinear.

Cross validation:

```
## model validation ##

#cross validation
set.seed(123)
nsamp = ceiling(0.8*length(who$Life.expectancy))
for (i in 1:5){
training_samps = sample(c(1:length(who$Life.expectancy)), nsamp)
training_samps = sort(training_samps)
train_data = who[training_samps, ]
test_data = who[-training_samps, ]



train_mdl = lm(formula = Life.expectancy ~ Year + Status + Adult.Mortality +
                 infant.deaths + Alcohol + percentage.expenditure + BMI +
                 under.five.deaths + Total.expenditure + Diphtheria + HIV.AIDS +
                 thinness.5.9.years + Income.composition.of.resources + Schooling,
               data = train_data)

preds = predict(train_mdl, test_data)

R.sq = R2(preds, test_data$Life.expectancy)
RMSPE = RMSE(preds, test_data$Life.expectancy)
MAPE = MAE(preds, test_data$Life.expectancy)
sd_RMSPE = RMSPE/sd(test_data$Life.expectancy)
```

```
print(c(R.sq,RMSPE,MAPE,sd_RMSPE))



}
```

We can see that R.sq is quite near 1, which means the model fitted is nice, above 80% data points error are explained by the model. The RMSPE after standardization is around 0.40, which is fine but not enough close to 0. Thus the model is generally valid.

Results

We compared the 8 potential predictors considered to have no linear relationship with life expectancy. Thus, we use the SLM to compare each predictors with life expectancy and we find out population is the only predictor that does not have a linear relationship with life expectancy. After removing the population predictor, we compared the full model with the reduced model and the result does not have significant changes. This indicated that the population predictor does not have a significant impact on our model. Moreover, since we want to fit a better regression model, we used stepwise function and the result shows that there are 6 predictors (thinness..1.19.years,Population,Hepatitis.B,GDP,Measles,Polio) that have no significant relationship with life expectancy. We performed residual analysis to find out that our new model (after removing the 6 predictors) has a constant variance and linearity. Therefore, we do not need transformation for our model. After the calculation on leverage and cook's distance, we figured out that the cook's distance are less than 1 for all observations. In the end, we tried robust regression to fit our final model to dampen the impact of highly influential observations and also violations in model assumptions.

Conclusion

From the analysis, we conclude that predictors like Diphtheria coverage, adult mortality rate, infant deaths rate, HIV/AIDS rate and expenditure on health percentage have a significant power in predicting life expectancy. Furthermore, country status would have an effect on lifespan as well. Life expectancy in developing countries is lower than Life expectancy in developed countries.

We removed Hepatitis.B and Polio from the model and we can conclude that immunization factor do not significantly influencing the life expectancy. However, the Diphtheria, adult mortality, infant deaths, HIV/AIDS and percentage.expenditure have relationship with life expectancy. Thus, we can conclude that lifespan can be improved if government spend more expenditure on healthcare. Besides, personal habits and lifestyle would have relationship with life expectancy, like the predictor Alcohol would have an impact on human's lifespan. Moreover, based on our model, we figured out that GDP does not have a significant relationship with life expectancy, but the predictor income composition do have a significant relationship

with life expectancy. Economical factor is also related to total expenditure and percentage.expenditure. In a word, we consider economical factors that are associated with government expenditure would have a significant effect on life expectancy. We also reckon education is a potential cause for a country to have a lower life expectancy, since a country with poor educational system would have less possibility to educate more excellent young individuals. Thus, there will be less professional doctors and researchers in the society, which would lead to a weak healthcare system.

Appendix

All data and R codes are attached below for reproducibility.

```
library(tidyverse)
library(car)
library(MASS)
library(caret)
who <- read.csv(file = "Life Expectancy Data.csv", header=TRUE)
#Looking at the  data
head(who)
attach(who)
```

```
## The following objects are masked from who (pos = 3):
##
##     Adult.Mortality, Alcohol, BMI, Country, Diphtheria, GDP,
##     Hepatitis.B, HIV.AIDS, Income.composition.of.resources,
##     infant.deaths, Life.expectancy, Measles, percentage.expenditure,
##     Polio, Population, Schooling, Status, thinness..1.19.years,
##     thinness.5.9.years, Total.expenditure, under.five.deaths, Year
```

```
# introduce one new additional data point into our dataset
china_new_data_point = data.frame("China", 2015, "Developing", 74.2625,73.75,   294.875 ,4.182,
 78.48934709    ,80.4375,   65857.9375, 21.80625,   350 ,93.6875,   4.918,  93.3125,    0.1,    2345.303
 321812.0625,   4.6375, 4.025   ,0.66025,   11.4375)

names(china_new_data_point) = c("Country","Year","Status","Life.expectancy","Adult.Mortality",
 "infant.deaths","Alcohol","percentage.expenditure","Hepatitis.B","Measles" , "BMI" ,
 "under.five.deaths" ,"Polio","Total.expenditure","Diphtheria" , "HIV.AIDS","GDP","Population",
 "thinness..1.19.years", "thinness.5.9.years","Income.composition.of.resources","Schooling")
```

```r
who = rbind(china_new_data_point,who)

#remove Country since we are not focusing on specific countries in this research
who <- who[,-c(1)]



# exclude missing values (Remove Rows with Missing Data)
who = who[complete.cases(who),]
# 1289 observations were deleted
# 1650 observations remaining

#fit regression model
model <- lm(Life.expectancy~., data = who)
summary(model)
avPlots(model)

#check pairwise correlations
cor(who[,-c(1,2)])

histogram(who$Life.expectancy)

histogram(who$Adult.Mortality)

histogram(who$infant.deaths)

histogram(who$Alcohol)

histogram(who$percentage.expenditure)

histogram(who$Hepatitis.B)

histogram(who$Measles)

histogram(who$BMI)

histogram(who$under.five.deaths)

histogram(who$Polio)

histogram(who$Total.expenditure)
```

```r
histogram(who$Diphtheria)
```

```r
histogram(who$HIV.AIDS)
```

```r
histogram(who$GDP)
```

```r
histogram(who$Population)
```

```r
histogram(who$thinness..1.19.years)
```

```r
histogram(who$thinness.5.9.years)
```

```r
histogram(who$Income.composition.of.resources)
```

```r
histogram(who$Schooling)
```

```r
#checked the  model
summodel <-  summary(model)
#percentage.expenditure and Hepatitis.B and Measle and Polio and GDP and Population
#thinness..1.19.years  and thinness.5.9.years larger than 0.05
```

```r
#Life.expectancy and GDP
gdp <- lm( Life.expectancy ~ GDP , data = who )
summary(gdp)

## standardize data using unit length scaling ##
#unit length scalings
s_Year<-sqrt(sum((who$Year-mean(who$Year))^2))
#s_Status<-sqrt(sum((Status-mean(Status))^2))
s_Adult.Mortality = sqrt(sum((who$Adult.Mortality-mean(who$Adult.Mortality))^2))
s_infant.deaths<-sqrt(sum((who$infant.deaths-mean(who$infant.deaths))^2))
s_Alcohol<-sqrt(sum((who$Alcohol-mean(who$Alcohol))^2))
s_percentage.expenditure<-sqrt(sum((who$percentage.expenditure-mean(who$percentage.expenditure))^2))
s_Hepatitis.B<-sqrt(sum((who$Hepatitis.B-mean(who$Hepatitis.B))^2))
s_Measles<-sqrt(sum((who$Measles-mean(who$Measles))^2))
s_BMI<-sqrt(sum((who$BMI-mean(who$BMI))^2))
s_under.five.deaths<-sqrt(sum((who$under.five.deaths-mean(who$under.five.deaths))^2))
s_Life.expectancy = sqrt(sum((who$Life.expectancy-mean(who$Life.expectancy))^2))

s12  = sqrt(sum((who$Polio - mean(who$Polio))^2))
```

```r
s13  = sqrt(sum((who$Total.expenditure - mean(who$Total.expenditure))^2))
s14  = sqrt(sum((who$Diphtheria - mean(who$Diphtheria))^2))
s15  = sqrt(sum((who$HIV.AIDS - mean(who$HIV.AIDS))^2))
s16  = sqrt(sum((who$GDP - mean(who$GDP))^2))
s17  = sqrt(sum((who$Population - mean(who$Population))^2))
s18  = sqrt(sum((who$thinness..1.19.years - mean(who$thinness..1.19.years))^2))
s19  = sqrt(sum((who$thinness.5.9.years - mean(who$thinness.5.9.years))^2))
s20  = sqrt(sum((who$Income.composition.of.resources -mean(who$Income.composition.of.resources))^2))
s21  = sqrt(sum((who$Schooling - mean(who$Schooling))^2))

z_Year<-(who$Year-mean(who$Year))/s_Year
#z_Status<-(Status-mean(Status))/s_Status
z_Adult.Mortality<-(who$Adult.Mortality-mean(who$Adult.Mortality))/s_Adult.Mortality
z_infant.deaths<-(who$infant.deaths-mean(who$infant.deaths))/s_infant.deaths
z_Alcohol<-(who$Alcohol-mean(who$Alcohol))/s_Alcohol
z_percentage.expenditure<-(who$percentage.expenditure-mean(who$percentage.expenditure))/s_percentage.exp
z_Hepatitis.B<-(who$Hepatitis.B-mean(who$Hepatitis.B))/s_Hepatitis.B
z_Measles<-(who$Measles-mean(who$Measles))/s_Measles
z_BMI<-(who$BMI-mean(who$BMI))/s_BMI
z_under.five.deaths<-(who$under.five.deaths-mean(who$under.five.deaths))/s_under.five.deaths
Life.expectancy_s<-(who$Life.expectancy-mean(who$Life.expectancy))/s_Life.expectancy

z12 = (who$Polio - mean(who$Polio))/s12
z13 = (who$Total.expenditure - mean(who$Total.expenditure))/s13
z14 = (who$Diphtheria - mean(who$Diphtheria))/s14
z15 = (who$HIV.AIDS - mean(who$HIV.AIDS))/s15
z16 = (who$GDP - mean(who$GDP))/s16
z17 = (who$Population - mean(who$Population))/s17
z18 = (who$thinness..1.19.years - mean(who$thinness..1.19.years))/s18
z19 = (who$thinness.5.9.years - mean(who$thinness.5.9.years))/s19
z20 = (who$Income.composition.of.resources - mean(who$Income.composition.of.resources))/s20
z21 = (who$Schooling - mean(who$Schooling))/s21

LifeSRModel<-lm(Life.expectancy_s~z_Year+z_Adult.Mortality+z_infant.deaths+z_Alcohol+z_percentage.expend

## variable selection ##
```

```r
summary(LifeSRModel)
LifeSRModel$coefficients
```

```r
#Life.expectancy_s and percentage.expenditure
le <- lm( Life.expectancy_s ~ z_percentage.expenditure )
summary(le)

#Life.expectancy_s and Hepatitis.B
lh <- lm( Life.expectancy_s ~ z_Hepatitis.B )
summary(lh)

#Life.expectancy_s and Measles
lmea <- lm( Life.expectancy_s ~ z_Measles )
summary(lmea)

#Life.expectancy_s and Polio
lp <- lm( Life.expectancy_s ~ z12 )
summary(lp)

#Life.expectancy_s and GDP
lg <- lm( Life.expectancy_s ~ z16 )
summary(lg)

#Life.expectancy_s and Population
lpo <- lm( Life.expectancy_s ~ z17)
summary(lpo)

#Life.expectancy_s and thinness..1.19.years
lt1 <- lm( Life.expectancy_s ~ z18 )
summary(lt1)

#Life.expectancy_s and thinness.5.9.years
lt5 <- lm( Life.expectancy_s ~ z19 )
summary(lt5)
```

```r
#population remove from dataset 'who'
#with standardized data
```

```r
LifeSRModel<-lm(Life.expectancy_s~z_Year+z_Adult.Mortality+z_infant.deaths+z_Alcohol+z_percentage.expend
summary(LifeSRModel)

#with original data
model0 <- lm(Life.expectancy~Year+Status+Adult.Mortality+infant.deaths+Alcohol+percentage.expenditure+He
summary(model0)
```

```r
##stepwise
#use stepwise method to complete variable selection
step(model, direction = 'both')

newmodel = lm(formula = Life.expectancy ~ Year + Status + Adult.Mortality +
    infant.deaths + Alcohol + percentage.expenditure + BMI +
    under.five.deaths + Total.expenditure + Diphtheria + HIV.AIDS +
    thinness.5.9.years + Income.composition.of.resources + Schooling,
    data = who)

summary(newmodel)
```

```r
## confidence interval ##
confint( newmodel , level = 0.9 )
```

```r
## residual analysis ##

#Standardized Residuals vs. Index tut 4
plot( rstandard(newmodel))
```

```r
#straight line around 0

n <- nrow( who )
plot( newmodel$residuals[1:(n-1)] , newmodel$residuals[2:n])
```

```r
#straight line, linear relationship

plot(newmodel)
```

```r
#leverage and influence
2*ncol(who)/nrow(who)  #0.02545455
newmodel_influence <- influence(newmodel)
```

```r
sort(newmodel_influence$hat, decreasing = TRUE)
```

```r
#cook's distance
newmodel_cook <- cooks.distance(newmodel)
sort(newmodel_cook, decreasing = TRUE)
```

```r
#There is potential outlier, so we use robust
#robust
Rmodel = rlm(formula = Life.expectancy ~ Year + Status + Adult.Mortality +
    infant.deaths + Alcohol + percentage.expenditure + BMI +
    under.five.deaths + Total.expenditure + Diphtheria + HIV.AIDS +
    thinness.5.9.years + Income.composition.of.resources + Schooling,
    data = who, psi = psi.huber)

summary(Rmodel)

plot(Rmodel)
```

```r
##multicollinearity
#vif
vif(Rmodel)
```

```r
## model validation ##

#cross validation
set.seed(123)
nsamp = ceiling(0.8*length(who$Life.expectancy))
for (i in 1:5){
training_samps = sample(c(1:length(who$Life.expectancy)), nsamp)
training_samps = sort(training_samps)
train_data = who[training_samps, ]
test_data = who[-training_samps, ]



train_mdl = lm(formula = Life.expectancy ~ Year + Status + Adult.Mortality +
                 infant.deaths + Alcohol + percentage.expenditure + BMI +
                 under.five.deaths + Total.expenditure + Diphtheria + HIV.AIDS +
                 thinness.5.9.years + Income.composition.of.resources + Schooling,
```

```
                data = train_data)

preds = predict(train_mdl, test_data)

R.sq = R2(preds, test_data$Life.expectancy)
RMSPE = RMSE(preds, test_data$Life.expectancy)
MAPE = MAE(preds, test_data$Life.expectancy)
sd_RMSPE = RMSPE/sd(test_data$Life.expectancy)
print(c(R.sq,RMSPE,MAPE,sd_RMSPE))



}
```