Regression and Analysis of Variance STAT 3340 / MATH 3340

# Abstract

This document performs a complete regression analysis of factors that influence life expectancy. The data comes from the WHO. Resources can be found at <https://www.kaggle.com/kumarajarshi/life-expectancy-who>. In the beginning, a full model is built; however, we observe that not all regressors pass the coefficient significance test. This indicates that a better model is needed. Hence, a backward elimination method is used to re-build the model. It is worth mentioning that a new data point is added to validate the correctness of our model. When testing the multicollinearity of the model, we conclude that two of the regressor have the problem of multicollinearity. In conclusion, we observe that developed countries have much higher life expectancy when comparing with developing counties. Also, life expectancy increases year by year in both developing and developed countries.

# Introduction

People tend to pay close attention to life expectancy since many of us want to have a higher life expectancy. There is a major difference between this dataset with older datasets. This dataset concerns with immunization factors like Hepatitis B which are not included in the older datasets. There are 22 columns and 2938 rows in the dataset, and 20 of the columns represent the predicting variables. 193 countries including both developing and developed counties are included. Furthermore, economic, and social factors are in the dataset as well. By carefully building the correct and accurate model, we want to be able to predict the life expectancy accurately.

# Data Description

An additional data point is added in later section. The modeling is built on data with no missing values, we will select a piece of data from the unused dataset for model verification.

First carry out statistics of data dimensions.

dat<-read.csv("~/Desktop/3340 Final project/Life Expectancy Data.csv", header=TRUE)  
dim(dat)

## [1] 2938 22

A total of 2938 pieces of data with 22 data dimensions.

Identification of missing values.

# total na  
sum(is.na(dat))

## [1] 2563

# numbers of na in each columns  
  
apply(dat, 2, function(x){   
 return (sum(is.na(x)))})

## Country Year   
## 0 0   
## Status Life.expectancy   
## 0 10   
## Adult.Mortality infant.deaths   
## 10 0   
## Alcohol percentage.expenditure   
## 194 0   
## Hepatitis.B Measles   
## 553 0   
## BMI under.five.deaths   
## 34 0   
## Polio Total.expenditure   
## 19 226   
## Diphtheria HIV.AIDS   
## 19 0   
## GDP Population   
## 448 652   
## thinness..1.19.years thinness.5.9.years   
## 34 34   
## Income.composition.of.resources Schooling   
## 167 163

The number of missing variables in all data is 2563, and the number of missing values in each column is displayed.Delete rows containing missing values.

dat\_new<-na.omit(dat)  
dim(dat\_new)

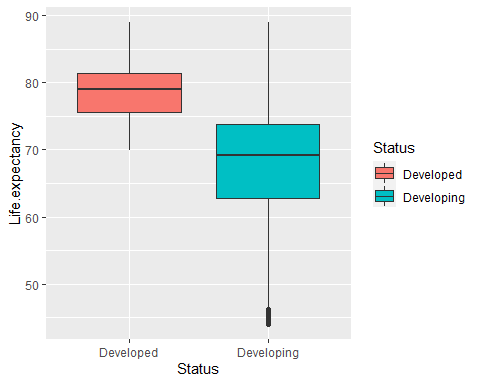
## [1] 1649 22

The remaining data volume after deleting the missing values is 1649 rows.

## the distribution of Life expectancy

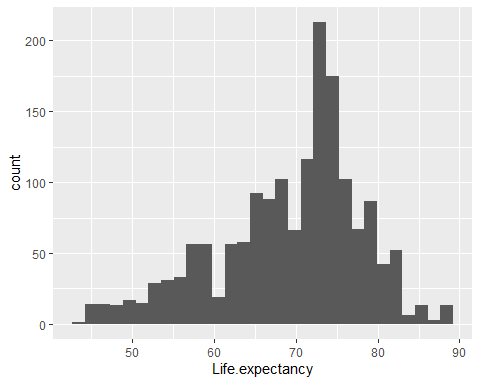
Show the distribution of Life.expectancy in different countries (developing or non-developing countries)

library(ggplot2)  
ggplot(data = dat\_new,aes(x=Status,y=Life.expectancy))+  
 geom\_boxplot(aes(fill = Status))

 It can be seen from the box plot that the Life.expectancy of developed countries is significantly higher than that of developing countries, and the variance is smaller.

ggplot(data = dat\_new,aes(x=Life.expectancy))+geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 From the distribution histogram of Life.expectancy, it can be seen that Life.expectancy is mainly concentrated in the age range of 70-80, with a slight left skew distribution.

Show the average value of Life.expectancy in different years.

aggregate(dat\_new$Life.expectancy,list(dat\_new$Year),mean)

## Group.1 x  
## 1 2000 70.20000  
## 2 2001 70.25758  
## 3 2002 68.58395  
## 4 2003 68.34421  
## 5 2004 68.43398  
## 6 2005 68.47273  
## 7 2006 68.90439  
## 8 2007 68.67417  
## 9 2008 68.66667  
## 10 2009 68.77937  
## 11 2010 69.12734  
## 12 2011 69.86308  
## 13 2012 70.24496  
## 14 2013 70.38769  
## 15 2014 70.51985  
## 16 2015 71.40000

With the improvement of medical and health conditions, we can see that the average value of Life.expectancy has shown an increasing trend year by year.

## Correlation between variables

Choose 5 variables to perform correlation analysis with Life.expectancy, display the correlation coefficient, and visualize it.

cor(dat\_new[,4:8])

## Life.expectancy Adult.Mortality infant.deaths Alcohol  
## Life.expectancy 1.0000000 -0.70252306 -0.16907380 0.4027183  
## Adult.Mortality -0.7025231 1.00000000 0.04245024 -0.1755351  
## infant.deaths -0.1690738 0.04245024 1.00000000 -0.1062169  
## Alcohol 0.4027183 -0.17553509 -0.10621692 1.0000000  
## percentage.expenditure 0.4096308 -0.23760989 -0.09076463 0.4170474  
## percentage.expenditure  
## Life.expectancy 0.40963082  
## Adult.Mortality -0.23760989  
## infant.deaths -0.09076463  
## Alcohol 0.41704736  
## percentage.expenditure 1.00000000

{r,fig.show='hold' ,out.width='40%', fig.align='center', fig.cap='1.3 Correlation between variables'} pairs(dat\_new[,4:8])

A picture containing text, screenshot

Description automatically generated

There is a strong negative correlation between Life.expectancy and Adult.Mortality, the correlation coefficient is -0.7025231, and the correlation between Life.expectancy and other variables is very weak.

# Methods

## full model

The dependent variable is Life.expectancy, because Life.expectancy has a significant relationship with Status (country type). First, the Status variable is factorized, and then all variables except the country are used as independent variables to build the full model.

dat\_new$Status<-as.factor(dat\_new$Status)  
  
lm\_full<-lm(Life.expectancy~.,data = dat\_new[,-1])  
summary(lm\_full)

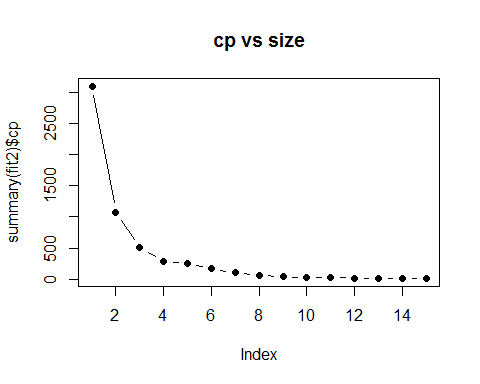
##   
## Call:  
## lm(formula = Life.expectancy ~ ., data = dat\_new[, -1])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.7681 -2.1427 0.0273 2.1776 12.4242   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.090e+02 4.621e+01 6.687 3.12e-11 \*\*\*  
## Year -1.272e-01 2.308e-02 -5.510 4.18e-08 \*\*\*  
## StatusDeveloping -8.865e-01 3.353e-01 -2.644 0.00827 \*\*   
## Adult.Mortality -1.621e-02 9.441e-04 -17.171 < 2e-16 \*\*\*  
## infant.deaths 8.873e-02 1.059e-02 8.376 < 2e-16 \*\*\*  
## Alcohol -1.313e-01 3.366e-02 -3.901 9.95e-05 \*\*\*  
## percentage.expenditure 3.026e-04 1.789e-04 1.691 0.09096 .   
## Hepatitis.B -3.258e-03 4.449e-03 -0.732 0.46413   
## Measles -1.033e-05 1.070e-05 -0.966 0.33439   
## BMI 3.183e-02 5.955e-03 5.345 1.03e-07 \*\*\*  
## under.five.deaths -6.662e-02 7.673e-03 -8.682 < 2e-16 \*\*\*  
## Polio 5.797e-03 5.121e-03 1.132 0.25776   
## Total.expenditure 9.220e-02 4.042e-02 2.281 0.02268 \*   
## Diphtheria 1.403e-02 5.877e-03 2.387 0.01712 \*   
## HIV.AIDS -4.481e-01 1.780e-02 -25.174 < 2e-16 \*\*\*  
## GDP 2.451e-05 2.826e-05 0.867 0.38594   
## Population -6.085e-10 1.733e-09 -0.351 0.72558   
## thinness..1.19.years -5.815e-03 5.254e-02 -0.111 0.91189   
## thinness.5.9.years -5.010e-02 5.185e-02 -0.966 0.33412   
## Income.composition.of.resources 1.045e+01 8.327e-01 12.549 < 2e-16 \*\*\*  
## Schooling 8.949e-01 5.910e-02 15.142 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.556 on 1628 degrees of freedom  
## Multiple R-squared: 0.8386, Adjusted R-squared: 0.8366   
## F-statistic: 422.9 on 20 and 1628 DF, p-value: < 2.2e-16

It can be seen that many variables like Polio have not passed the coefficient significance test, and the next step of model optimization is needed.

## Variable screening

According to the $c\_p $ index to optimize the model of the whole subset, the direction is set to “backward”, and the minimum cp corresponding to the number of real different variables is shown in the figure below.

library(leaps)  
  
fit2 <- regsubsets(Life.expectancy~., nvmax=15,data=dat\_new[,-1],method = "backward")  
  
plot(summary(fit2)$cp, pch=16,type='b',main='cp vs size')



2.1 cp vs size

It can be seen that after the number of variables is 7, as the number of variables increases, the decrease in $c\_p $value is very weak. Finally, the number of variables used is 7 and the variable names and coefficients are displayed.

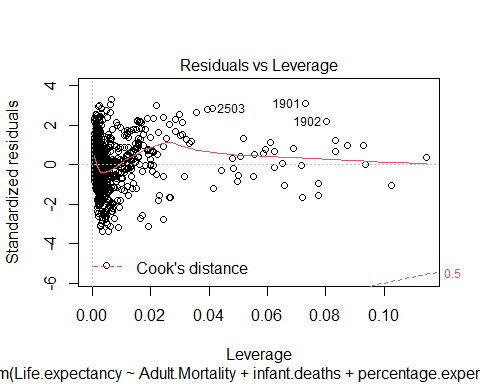
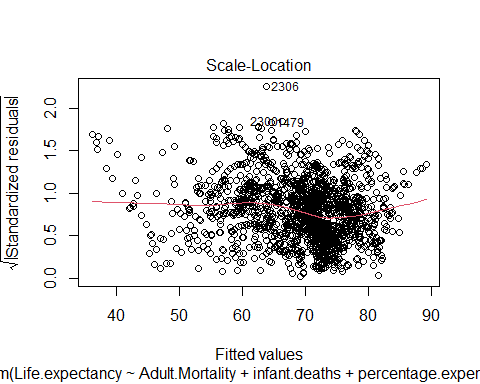
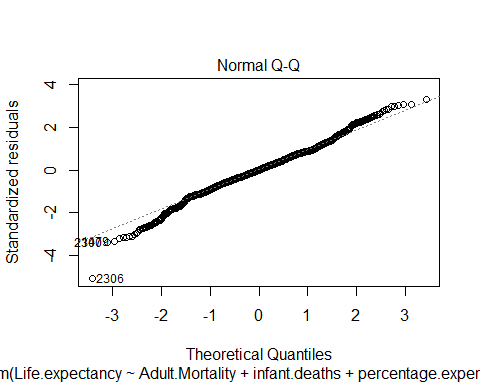
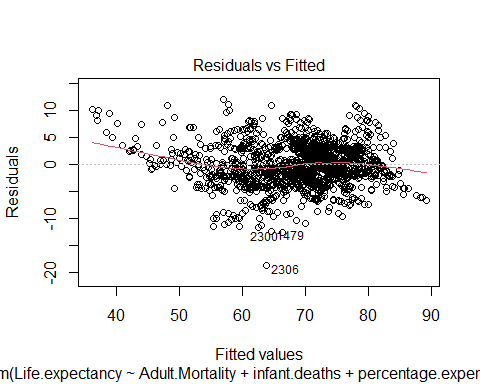
coef(fit2,7)

## (Intercept) Adult.Mortality   
## 54.2571934448 -0.0179239734   
## infant.deaths percentage.expenditure   
## 0.0952264116 0.0004525107   
## under.five.deaths HIV.AIDS   
## -0.0735808494 -0.4437872056   
## Income.composition.of.resources Schooling   
## 10.7764519764 0.9878209348

Finally, seven variables including Adult.Mortality and infant.deaths were selected as independent variables to establish a linear regression model.

## Modification of the model

lm\_final<-lm(Life.expectancy~Adult.Mortality+infant.deaths+percentage.expenditure+  
 under.five.deaths+HIV.AIDS+Income.composition.of.resources+Schooling,data=dat\_new)  
  
plot(lm\_final)



Most of the scattered points of the norm qq graph are distributed along a 45-degree diagonal. The lower left part of the data has a left skewed distribution, and the upper right part has a licensing distribution, which satisfies the residual normality test as a whole;

Residuals vs fitted is a graph for testing the linear relationship between the dependent variable and the independent variable. Since the dependent variable is linearly related to the independent variable, it can be seen that the residuals have no relationship with the fitted values, which is reasonable.

The third graph shows that the standardized residuals have no obvious contraction and expansion trend with the increase of the fitted value, and the residuals basically satisfy the uniformity of variance.

The last picture shows that there are individual strong leverage points, but the cook distance is not very large, which has little effect on the overall and is within an acceptable range.

## addition data point

dat\_incomplete<-dat[!complete.cases(dat),]  
  
head(na.omit(dat\_incomplete[,c("Adult.Mortality","infant.deaths","percentage.expenditure",  
 "under.five.deaths","HIV.AIDS","Income.composition.of.resources","Schooling",'Life.expectancy')]))

## Adult.Mortality infant.deaths percentage.expenditure under.five.deaths  
## 33 19 21 0.00000 24  
## 45 146 20 25.01852 23  
## 46 145 20 148.51198 23  
## 47 145 20 147.98607 24  
## 48 145 21 154.45594 25  
## 49 335 66 0.00000 98  
## HIV.AIDS Income.composition.of.resources Schooling Life.expectancy  
## 33 0.1 0.743 14.4 75.6  
## 45 0.1 0.663 11.5 71.7  
## 46 0.1 0.653 11.1 71.6  
## 47 0.1 0.644 10.9 71.4  
## 48 0.1 0.636 10.7 71.3  
## 49 1.9 0.531 11.4 52.4

## additional data point  
  
add\_data<-na.omit(dat\_incomplete[,c("Adult.Mortality","infant.deaths","percentage.expenditure",  
 "under.five.deaths","HIV.AIDS","Income.composition.of.resources","Schooling",'Life.expectancy')])[1,]  
  
# Model prediction  
  
# sd of Life.expectancy  
sd(dat$Life.expectancy,na.rm = T)

## [1] 9.523867

predict(lm\_final,newdata = add\_data)

## 33   
## 76.3376

add\_data$Life.expectancy

## [1] 75.6

The Standard deviation of the overall average life span is 9.5238. The actual value of our additional data point is 75.6 and the predicted value is 76.3376, which is enough to show that the model establishment is relatively successful.

## Multicollinearity analysis

library(car)

## Warning: package 'car' was built under R version 4.0.2

## Loading required package: carData

vif(lm\_final)

## Adult.Mortality infant.deaths   
## 1.747734 172.529817   
## percentage.expenditure under.five.deaths   
## 1.247782 173.143971   
## HIV.AIDS Income.composition.of.resources   
## 1.442091 2.765619   
## Schooling   
## 2.832519

According to the criterion of vif<10, the variables infant.deaths and under.five.deaths may have multicollinearity.

# Results

Show the final model

lm\_final<-lm(Life.expectancy~Adult.Mortality+infant.deaths+percentage.expenditure+  
 under.five.deaths+HIV.AIDS+Income.composition.of.resources+Schooling,data=dat\_new)  
  
anova(lm\_final)

## Analysis of Variance Table  
##   
## Response: Life.expectancy  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Adult.Mortality 1 62941 62941 4697.86 < 2.2e-16 \*\*\*  
## infant.deaths 1 2477 2477 184.91 < 2.2e-16 \*\*\*  
## percentage.expenditure 1 7291 7291 544.18 < 2.2e-16 \*\*\*  
## under.five.deaths 1 3024 3024 225.72 < 2.2e-16 \*\*\*  
## HIV.AIDS 1 8294 8294 619.05 < 2.2e-16 \*\*\*  
## Income.composition.of.resources 1 17080 17080 1274.88 < 2.2e-16 \*\*\*  
## Schooling 1 4436 4436 331.13 < 2.2e-16 \*\*\*  
## Residuals 1641 21986 13   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm\_final)

##   
## Call:  
## lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +   
## percentage.expenditure + under.five.deaths + HIV.AIDS + Income.composition.of.resources +   
## Schooling, data = dat\_new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.5927 -2.1488 0.0179 2.4021 12.0359   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.426e+01 5.477e-01 99.069 < 2e-16 \*\*\*  
## Adult.Mortality -1.792e-02 9.512e-04 -18.843 < 2e-16 \*\*\*  
## infant.deaths 9.523e-02 9.800e-03 9.717 < 2e-16 \*\*\*  
## percentage.expenditure 4.525e-04 5.725e-05 7.904 4.92e-15 \*\*\*  
## under.five.deaths -7.358e-02 7.283e-03 -10.103 < 2e-16 \*\*\*  
## HIV.AIDS -4.438e-01 1.795e-02 -24.725 < 2e-16 \*\*\*  
## Income.composition.of.resources 1.078e+01 8.190e-01 13.158 < 2e-16 \*\*\*  
## Schooling 9.878e-01 5.429e-02 18.197 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.66 on 1641 degrees of freedom  
## Multiple R-squared: 0.8276, Adjusted R-squared: 0.8269   
## F-statistic: 1125 on 7 and 1641 DF, p-value: < 2.2e-16

The F value of the overall model test is 1125, p-value: <2.2e-16, and the overall model is significant.

Adjusted R-squared is 0.8269, 82.69% of the variance of the dependent variable can be explained by the regression model, and the goodness of fit of the model is relatively high.

The p values corresponding to the significance t test of the independent variables were all less than 0.05. All independent variables passed the significance test, and there is a strong linear correlation between the independent variables and the dependent variables.

From the point of view of the coefficient, HIV.AIDS has a strong impact on a country’s Life expectancy, and this impact is negative. Increasing Income composition of resources and Schooling can also effectively increase Life expectancy.

In summary, the establishment of the model is relatively successful.

# conclusion

Through the overall analysis of this article, descriptive statistics and the establishment of regression models, the following key conclusions are finally obtained:

The life expectancy of developed countries is significantly higher than that of developing countries. On the whole, the average life expectancy is increasing year by year.

Since it was first reported in the United States in 1981, HIV has been like a “ghost” that has swept all corners of the world. There were about 38.0 million people living with HIV at the end of 2019 (WHO,2020). Because of the gap of the HIV services,690000people died from HIV-related causes in 2019(WHO,2020). The impact of AIDS on average life expectancy is very significant. From the regression model established above, we can see that if other factors remain unchanged, for every unit increase in the HIV/AIDS coefficient, average life expectancy decreases by 0.4438 units on average.

Income composition of resources is the most basic source of income in a country or region. It is closely related to the production and life of residents. According to the regression model, it can be seen that if other factors remain unchanged, for every unit increase in the Income composition of resources, average life expectancy increases by 10.78 units on average. Therefore, we must attach importance to the use and protection of natural resources.

**Appendix**

All data and R markdown files are posted in the Github with this report.

**Reference**

WHO.(n.d).HIV/AIDS.Retrieved from: https://www.who.int/news-room/fact-sheets/detail/hiv-aids

Rajarshi.K.(2017).Retrieved from: https://www.kaggle.com/kumarajarshi/life-expectancy-who