Final report, STAT306

Group 9
Konstantin Mestnikov
Yahan Cong
Venus Lee
Muke Wang

Introduction

The World Health Organization (WHO), through the Global Health Observatory (GHO), is responsible for keeping records of health-related factors for countries around the world. Collecting this type of information is valuable as it enables health data analyses to be performed. For this project, the life expectancy, health factors, and economic data of 193 countries were examined, and their individual datasets were combined together to form the final dataset. However, given the developments in the health sector in the recent decade, we have witnessed a drastic improvement in human mortality rates, especially in developing countries. Therefore, this project focused only on data from 2000-2015. Due to difficulties in gathering data in countries like Vanuatu, Tonga, Togo, Cabo Verde, etc., some population data were unavailable and thus excluded from the final dataset. As a result, the final dataset included 20 predicting variables which are then each further divided into Immunization related factors, Mortality factors, Economical factors and Social factors. Since the dataset provides rich grounds for exploring covariates affecting life expectancy across the globe, one research question is to determine the best predictors of life expectancy by developing two models -developed and developing for two groups of countries.

Analysis

Data description

The data was a subset of the initial data collected from WHO and the United Nations website, the data-set related to life expectancy, health factors for 193 countries have been collected from 2010 to 2015. The aim of WHO is to suggest a country which area should be given importance in order to efficiently improve the life expectancy of its population.

It is difficult to find missing data for countries like Vanuatu, Tonga, Togo, Cabo Verde, etc, so the missing data was handled in R software by using Missmap command.

Only those critical factors were chosen which are more representative in this study, so unlike other studies of life expectancy, this study will focus on immunization factors, mortality factors, economic factors, social factors in 2010, as we want to avoid dealing with time frames. Also, some highly correlated variables were removed to improve the stability of the model, and for those countries that have generally high life expectancy, they were removed from this study.

Explanatory variable:

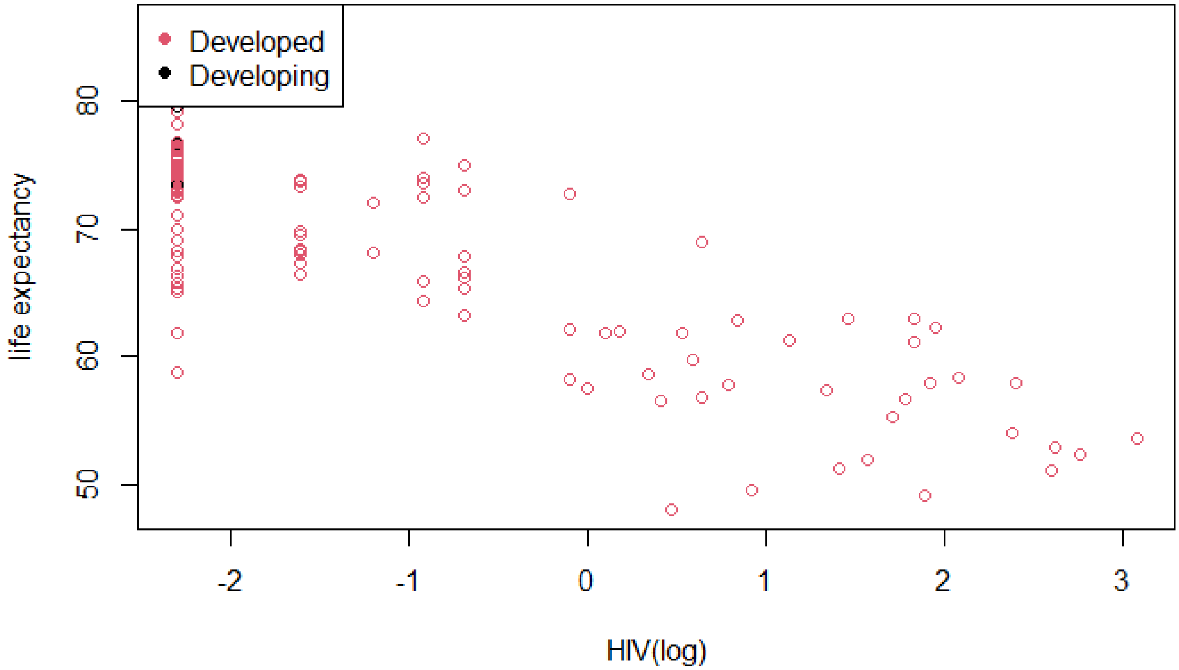
Status(Developed or Developing status)
Alcohol(recorded per capita (15+) consumption)
Infant deaths(Number of Infant Deaths per 1000 population)
Adult Mortality(Adult Mortality Rates of both sexes)
Hepatitis B(Hepatitis B (HepB) immunization coverage among 1-year-olds (%))
Measles(number of reported cases per 1000 population)
HIV(Deaths per 1 000 live births HIV/AIDS)
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)
under-five deaths(Number of under-five deaths per 1000 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 (%))
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(%))

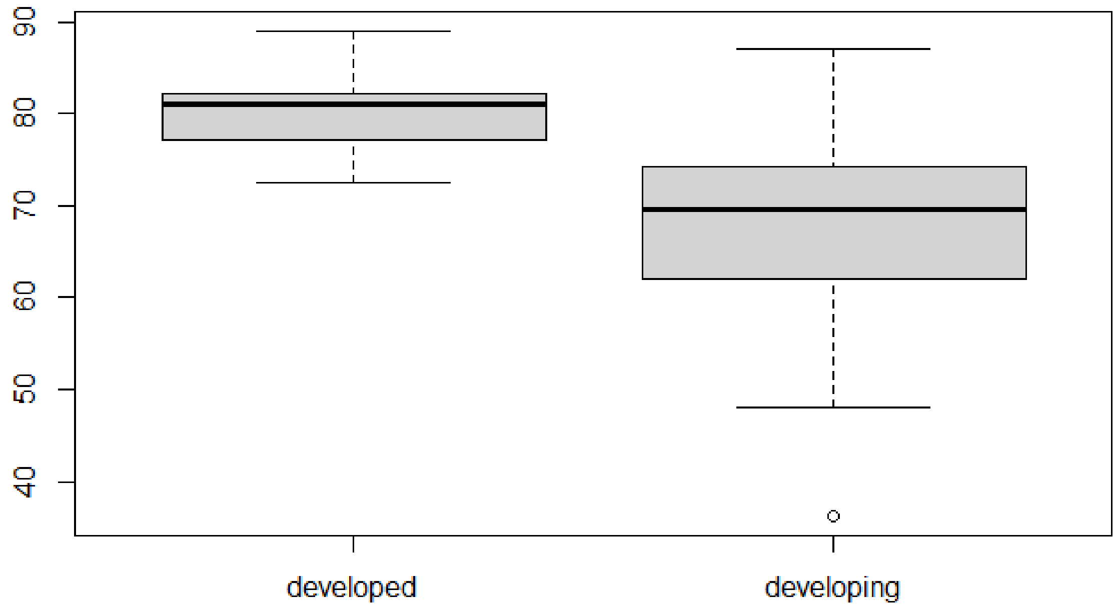
Income composition of resources(Human Development Index in terms of income composition of resources (index ranging from 0 to 1))

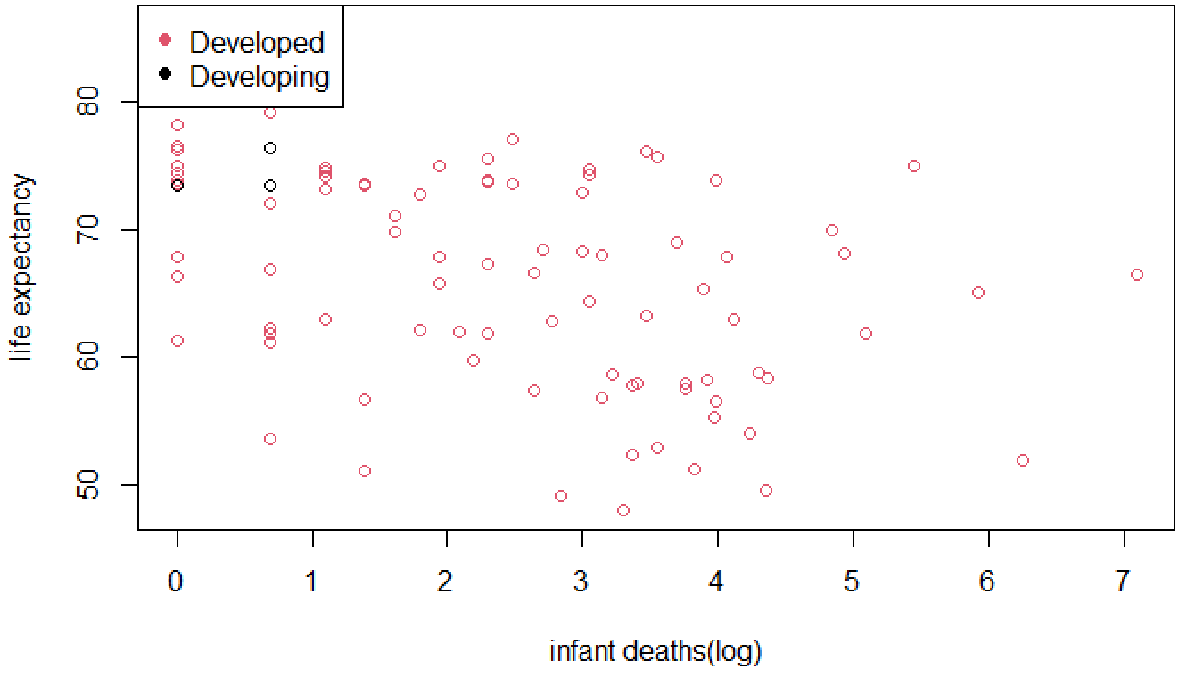
Schooling(Number of years of Schooling(years))

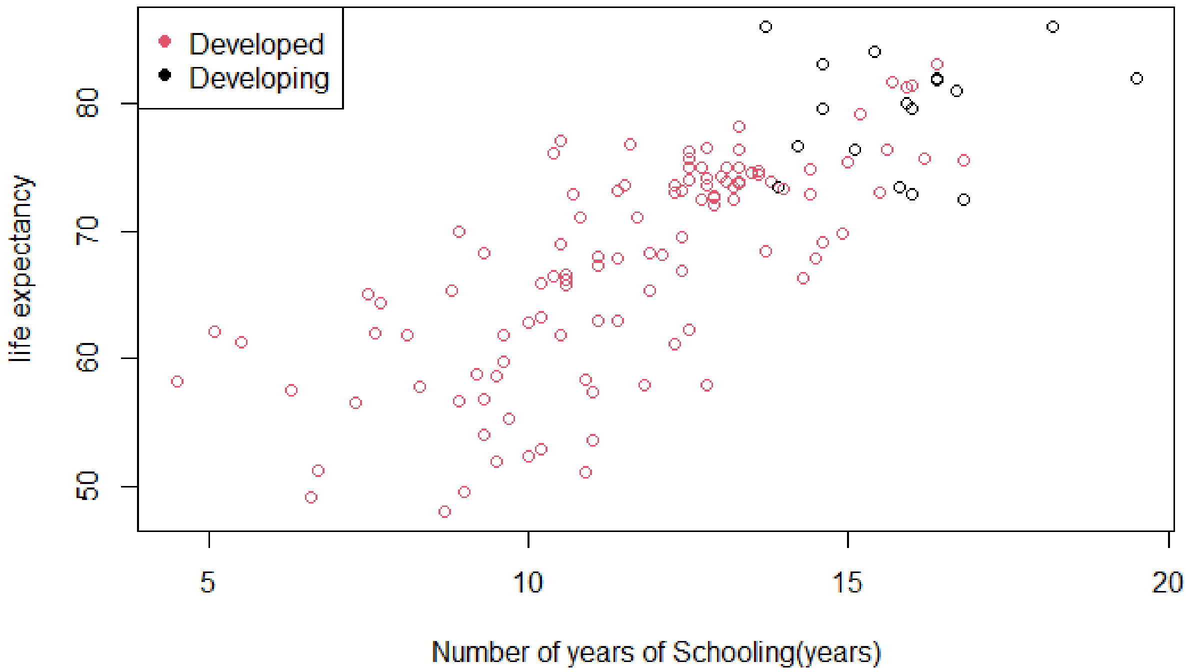
Response variable:
Life Expectancy(Life Expectancy in age)

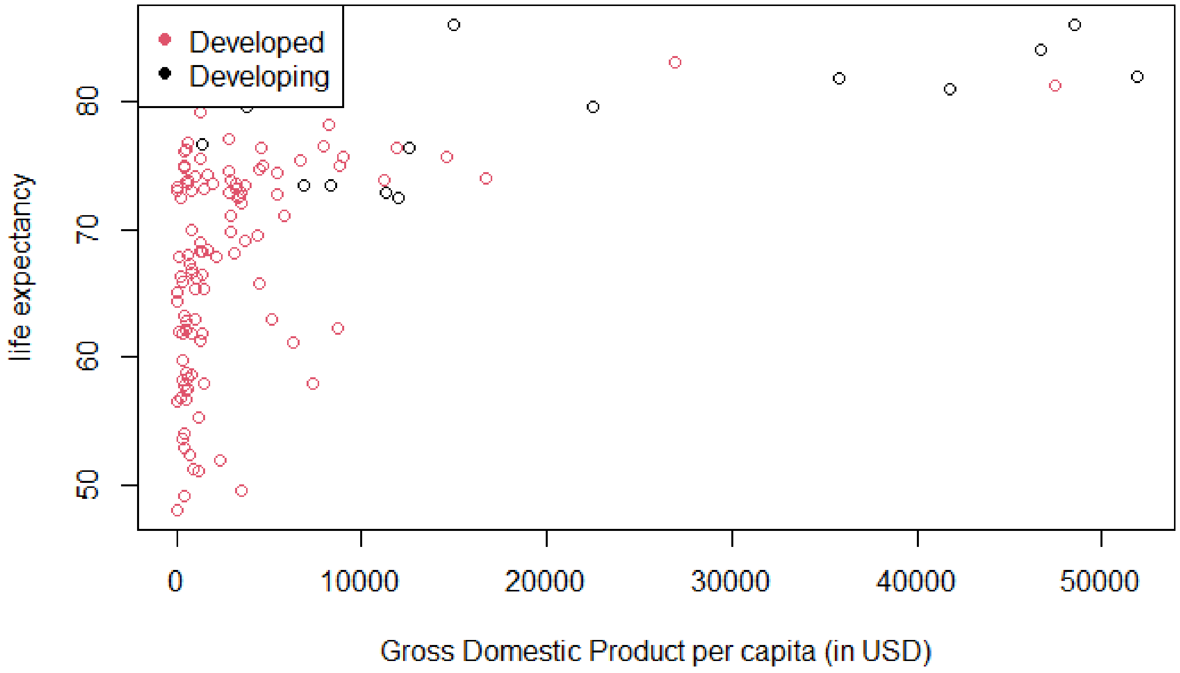
Visualizations:

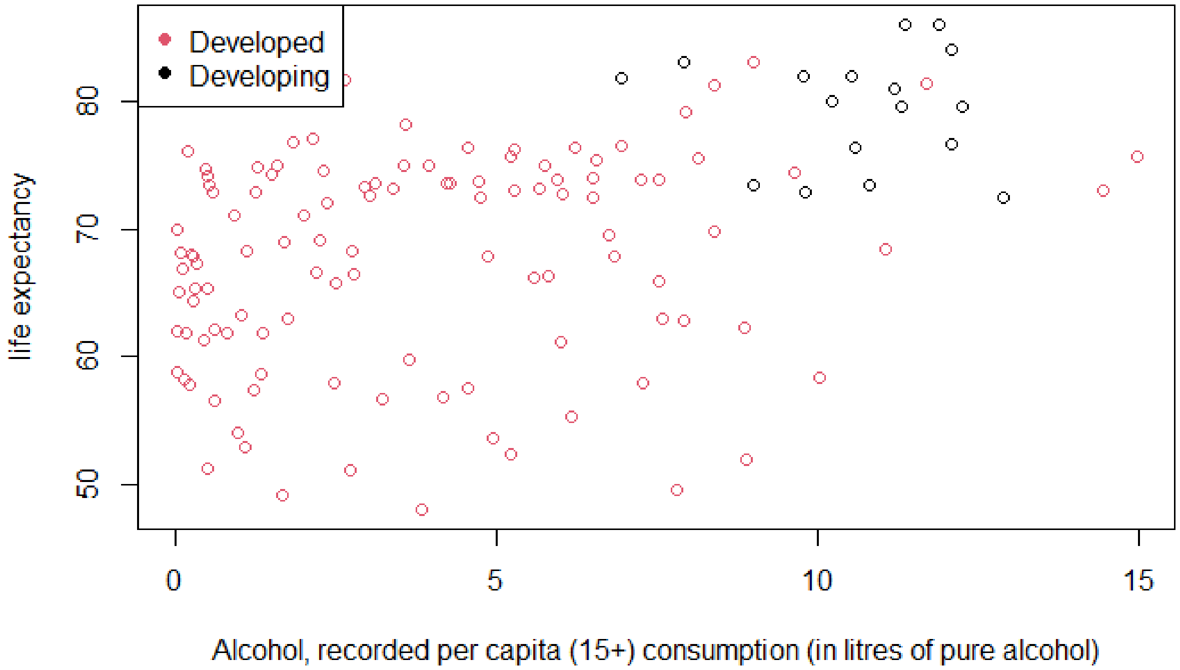


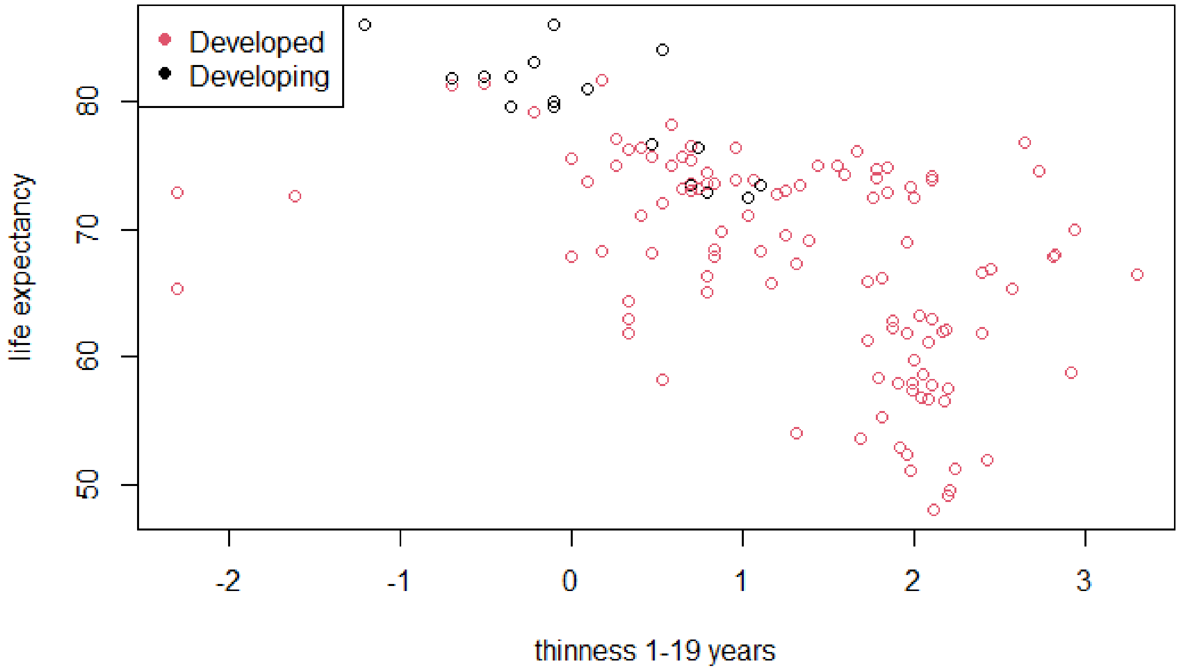

90
80
70
60
50
40
developed
developing


● Developed
● Developing
life expectancy
80
70
60
50
0
1
2
3
4
5
6
7
infant deaths(log)


● Developed
● Developing
life expectancy
50
60
70
80
5
10
15
20
Number of years of Schooling(years)


● Developed
● Developing
life expectancy
50
60
70
80
0
10000
20000
30000
40000
50000
Gross Domestic Product per capita (in USD)


● Developed
● Developing
life expectancy
50
60
70
80
0
5
10
15
Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)


● Developed
● Developing
life expectancy
50
60
70
80
-2
-1
0
1
2
3
thinness 1-19 years

A summary of any key features:

1.The distribution of life expectancy for developing countries is different from that of developed countries.

2. For the same amount of alcohol consumption, people in developing countries seem to have higher life expectancy.

3. There is a strong positive correlation between life expectancy and number of years of Schooling in years.

4. There is a strong negative correlation between life expectancy and HIV(Deaths per 1 000 live births HIV/AIDS).

5. Infant death rate seems to have no effect on life expectancy .

6. Underweight is linked to growth faltering and is associated with increased morbidity and mortality(Bovet, P., Kizirian, N., Madeleine, G., Blössner, M., & Chiolero, A. (2011). Prevalence of thinness in children and adolescents in the Seychelles: comparison of two international growth references. *Nutrition journal*, *10*(1), 1-6.), and our graph proves this idea: As thinness among children and for Age 10 to 19 increases, life expectancy decreases.

Methodology and Results:

The analyzed data are human life expectancy in 2010. All factors' values are collected from 183 countries around the world. Missing data is deleted in R software by na.omit. Since two variables: Adult Mortality and Income composition of resources are linearly dependent on other variables, we remove them to reduce collinearity.

1. Analysis the best predictors for global life expectancy in 2010

In order to find the best single predictors and optimal model for human's life expectancy in 2010 all around the world, we select the linear models including all explanatory variables in the processed data: *Status, Infant deaths, Alcohol, Percentage expenditure, Hepatitis B, Measles, BMI, Under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, GDP, Thinness 1-19 years, Thinness 5-9 years* and *Schooling* as the full model and make an exhaustive search by regsubset. The best single predictor is selected by comparing adjusted R^2 of simple linear models, and the optimal model is selected by comparison of Mallow's Cp and adjusted R^2.

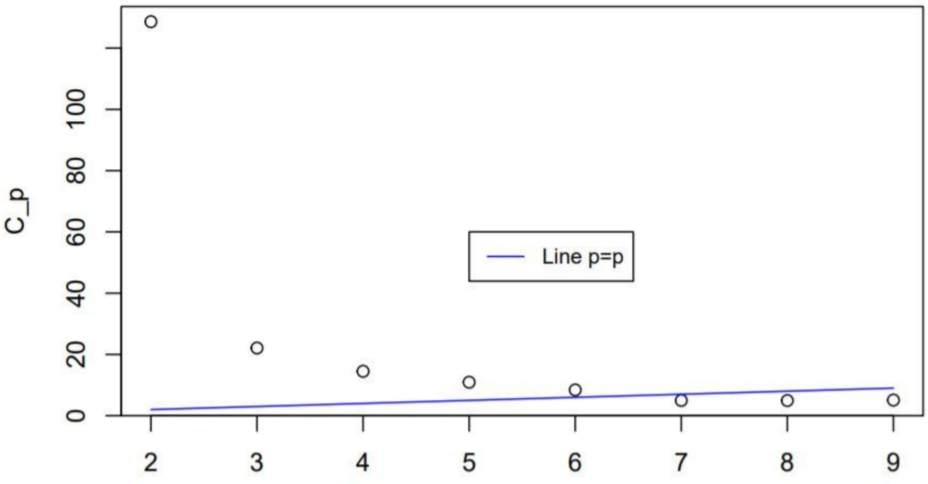
i) Best predictor for whole world life expectancy.

Using best subset selection, three variables: *BMI, Schooling and HIV.AIDS* are selected by regsubset. Since the adjusted R^2 of *Schooling* > adjusted $R^2$of *HIV.AIDS* > adjusted $R^2$of *BMI*, *Schooling* is the best predictor among all explanatory variables. It accounts for 58.18% variation in life expectancy of human all around the world along as seen from adjusted R^2.

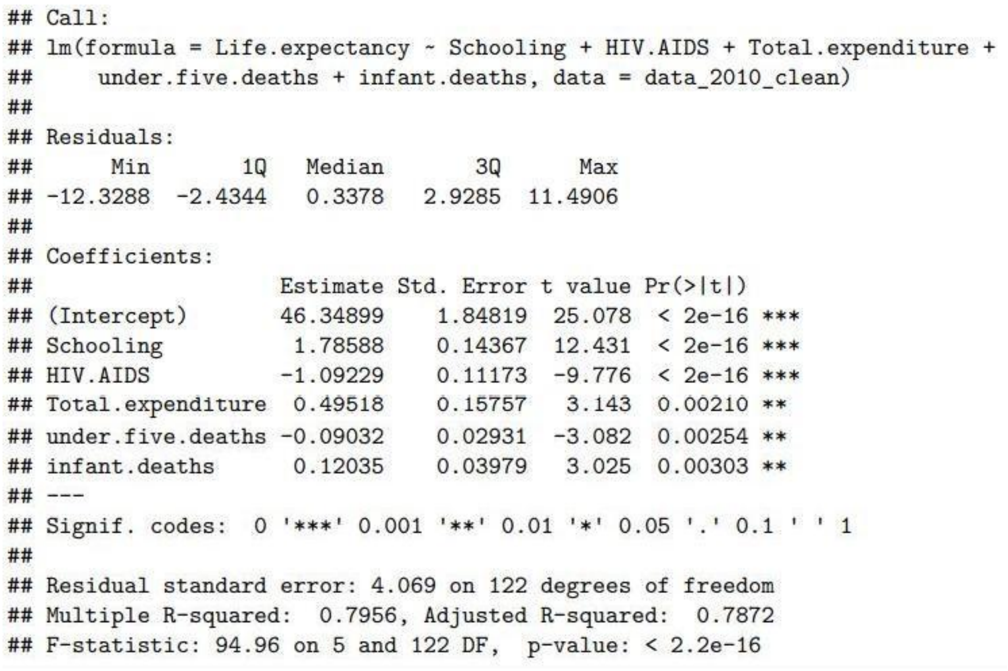
ii) Optimal model for whole world life expectancy.

To find the optimal for whole world life expectancy in 2010, we use regsubset to find the "best" linear models and compare their Mallow's Cp and adjusted R^2

Cp plot for the whole world life expectancy in 2010:

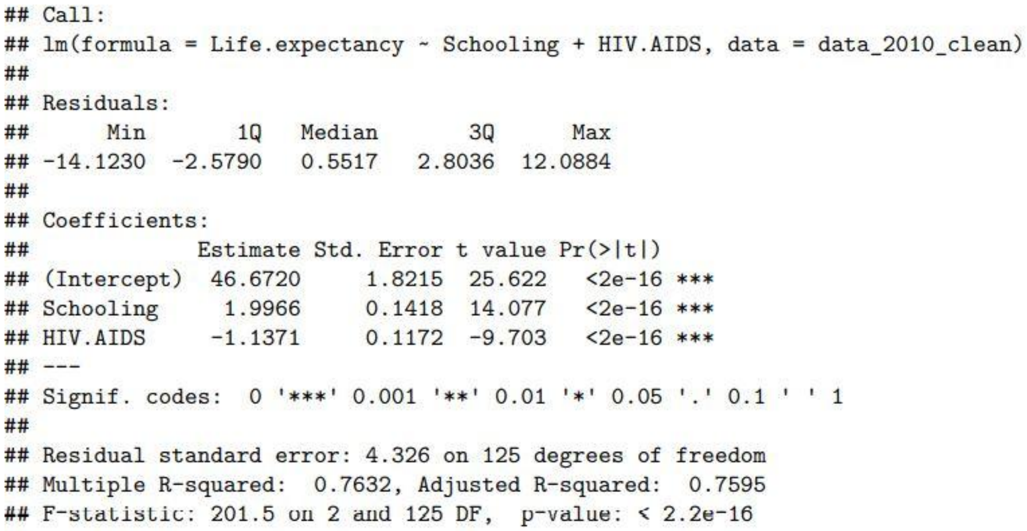


As shown in the whole world Cp plot, the model with 5 or 6 variables (6 or 7 parameters) can be taken as acceptable for the best model. Since their adjusted R^2 are very similar to each other, we choose the model with fewer variables for better explanation. The optimal model for whole world life expectancy in 2010 includes five explanatory variables: *Schooling, HIV.AIDS, Total expenditure, under five deaths* and *infant deaths.*



```
## Call:
## lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Total.expenditure +
##     under.five.deaths + infant.deaths, data = data_2010_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3288  -2.4344   0.3378   2.9285  11.4906
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       46.34899    1.84819  25.078  < 2e-16 ***
## Schooling          1.78588    0.14367  12.431  < 2e-16 ***
## HIV.AIDS          -1.09229    0.11173  -9.776  < 2e-16 ***
## Total.expenditure  0.49518    0.15757   3.143  0.00210 **
## under.five.deaths -0.09032    0.02931  -3.082  0.00254 **
## infant.deaths      0.12035    0.03979   3.025  0.00303 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.069 on 122 degrees of freedom
## Multiple R-squared:  0.7956, Adjusted R-squared:  0.7872
## F-statistic: 94.96 on 5 and 122 DF,  p-value: < 2.2e-16
```

Since *Schooling* and *HIV.AIDS* are the “best” predictors selected above, we compare the optimal model with the model with just these two top covariates. Based on each variables’ p_value in two chosen models, the three variables other than *Schooling* and *HIV.AIDS* provide only moderate additional predictive power. *Schooling* and *HIV.AIDS* alone account for 75.95% of variation of life expectancy.



```
## Call:
## lm(formula = Life.expectancy ~ Schooling + HIV.AIDS, data = data_2010_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1230  -2.5790   0.5517   2.8036  12.0884
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.6720     1.8215  25.622   <2e-16 ***
## Schooling     1.9966     0.1418  14.077   <2e-16 ***
## HIV.AIDS     -1.1371     0.1172  -9.703   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.326 on 125 degrees of freedom
## Multiple R-squared:  0.7632, Adjusted R-squared:  0.7595
## F-statistic: 201.5 on 2 and 125 DF,  p-value: < 2.2e-16
```

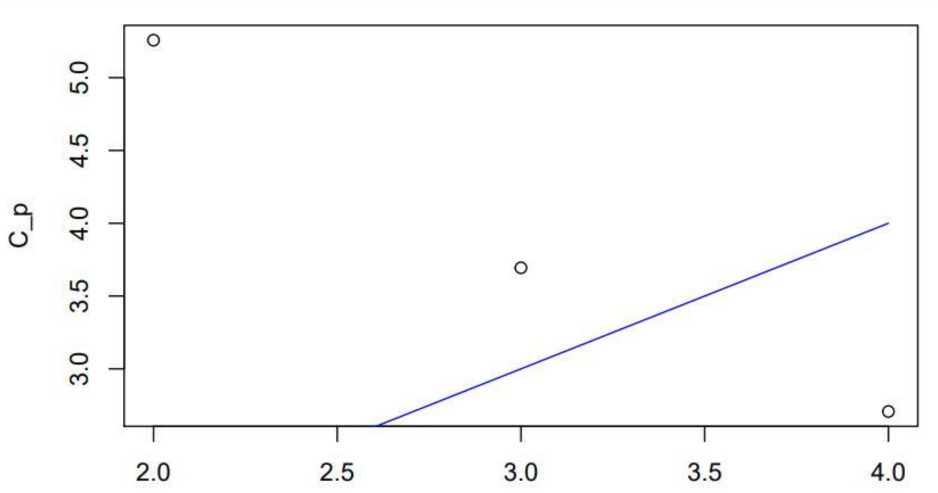
In the analysis of life expectancy of global human life expectancy, the best predictors are *Schooling, HIV.AIDS, Total expenditure, under five deaths* and *infant deaths*, and *Schooling* and *HIV.AIDS* are the two most important factors.

2. Developing vs Developed countries.

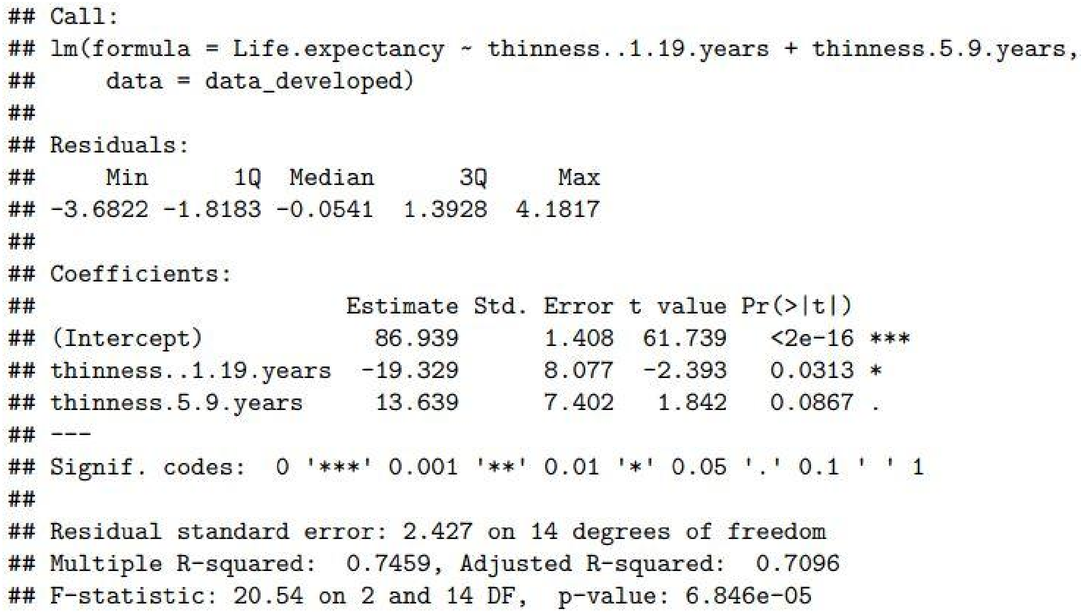
As the boxplot shown above, there is a distinct difference between life expectancy in developed countries and developing. In order to compare the best predictors for developed countries and developing countries, two datasets are divided by country *Status*. Using regsubsets, we find the best models for both two statuses.

1）Developed countries

Cp plot for the developed countries' life expectancy in 2010:



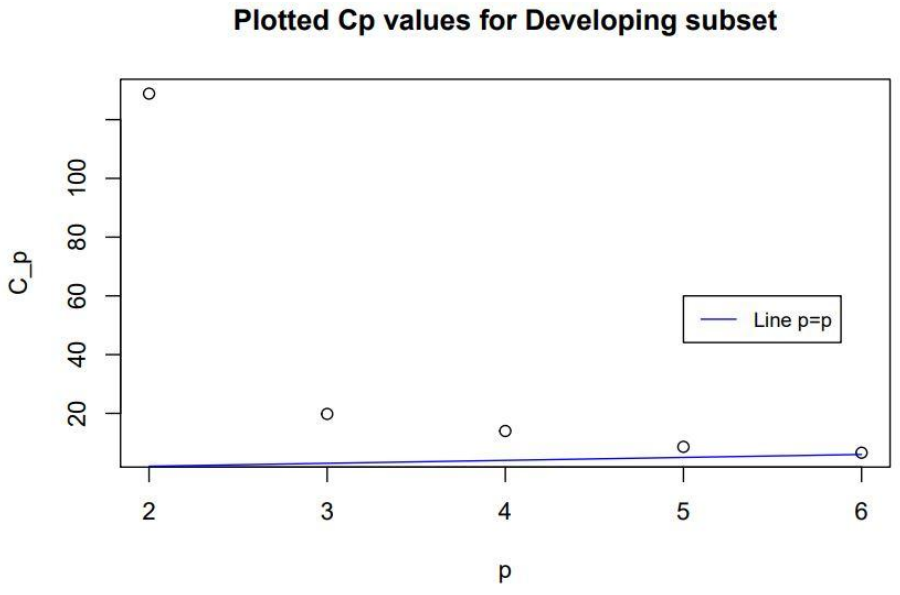
From the developed countries' Cp plot, we can see that model with 2 variables (3 parameters) is the optimal one since its Cp is closest to p. The best model for life expectancy in developed countries includes *thinness 1-19 years* and *thinness 5-9 years* as explanatory variables.



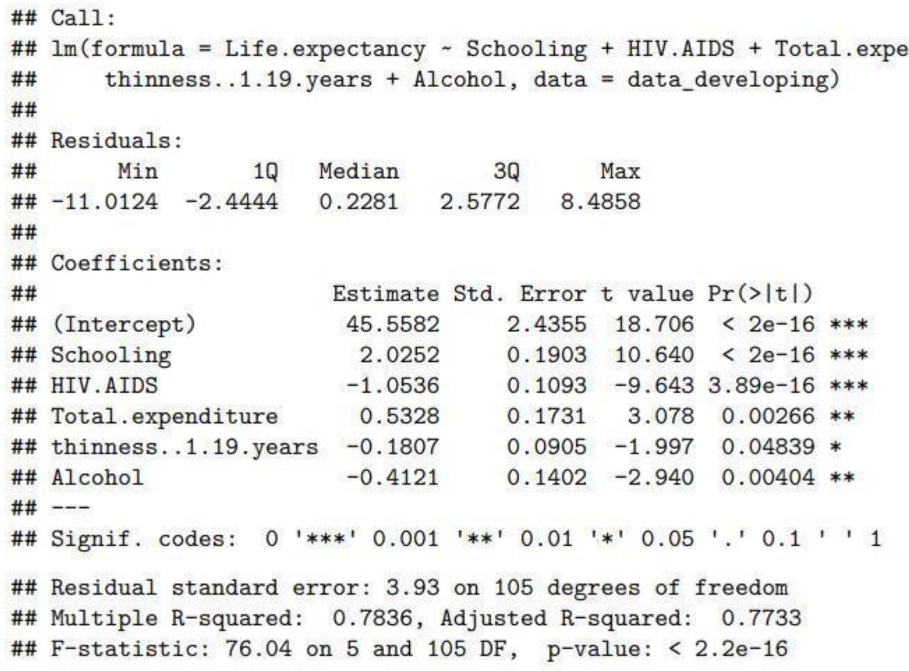
```
## Call:
## lm(formula = Life.expectancy ~ thinness..1.19.years + thinness.5.9.years,
##     data = data_developed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6822 -1.8183 -0.0541  1.3928  4.1817
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            86.939      1.408  61.739   <2e-16 ***
## thinness..1.19.years  -19.329      8.077  -2.393   0.0313 *
## thinness.5.9.years     13.639      7.402   1.842   0.0867 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.427 on 14 degrees of freedom
## Multiple R-squared:  0.7459, Adjusted R-squared:  0.7096
## F-statistic: 20.54 on 2 and 14 DF,  p-value: 6.846e-05
```

2) Developing countries

Cp plot for the developing countries' life expectancy in 2010:



From the developing countries' *Cp* plot, the model with 5 variables (6 parameters) is the optimal one in the developing case. The best linear model has *Schooling, HIV.AIDS, Total expenditure, thinness 1-19 years* and *alcohol* as covariates.



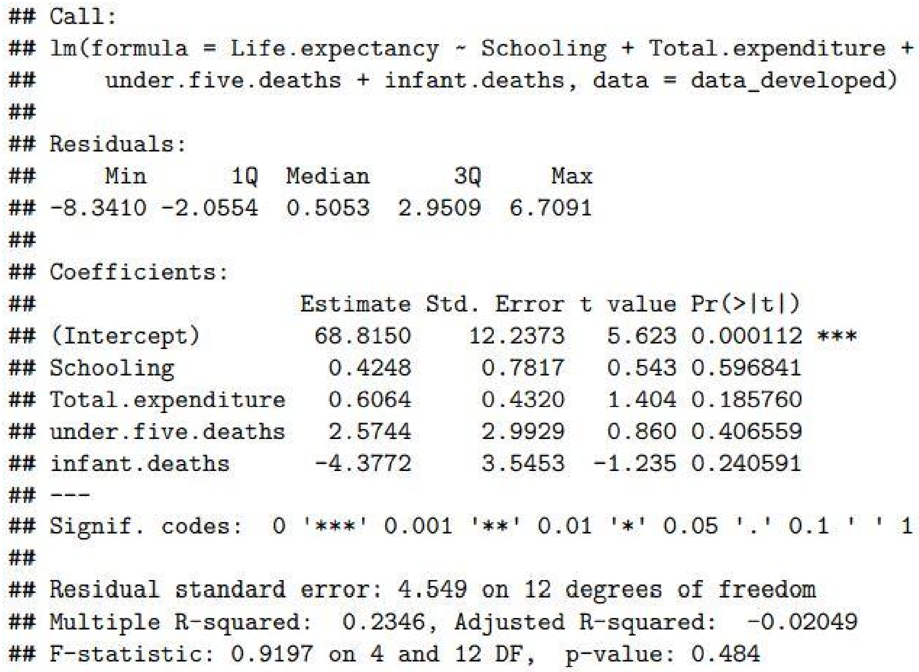
```
## Call:
## lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Total.expenditure +
##     thinness..1.19.years + Alcohol, data = data_developing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0124  -2.4444   0.2281   2.5772   8.4858
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           45.5582     2.4355  18.706  < 2e-16 ***
## Schooling              2.0252     0.1903  10.640  < 2e-16 ***
## HIV.AIDS              -1.0536     0.1093  -9.643 3.89e-16 ***
## Total.expenditure      0.5328     0.1731   3.078  0.00266 **
## thinness..1.19.years  -0.1807     0.0905  -1.997  0.04839 *
## Alcohol               -0.4121     0.1402  -2.940  0.00404 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 3.93 on 105 degrees of freedom
## Multiple R-squared:  0.7836, Adjusted R-squared:  0.7733
## F-statistic: 76.04 on 5 and 105 DF,  p-value: < 2.2e-16
```

The two “best” predictors for global life expectancy:*Schooling* and *HIV.AIDS* are included in the optimal model for developing countries, but not in the model for developed countries. A reasonable assumption for this phenomenon is since most of the data come from developing countries, some influential factors for developed countries also play an important role in global life expectancy explanation.

3. Developed countries vs Global

Since we have found that the “best” predictors for the global life expectancy model do not included in the optimal model, we did a test to check how the global model fitted data from developed countries:



```
## Call:
## lm(formula = Life.expectancy ~ Schooling + Total.expenditure +
##     under.five.deaths + infant.deaths, data = data_developed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3410 -2.0554  0.5053  2.9509  6.7091
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        68.8150    12.2373   5.623 0.000112 ***
## Schooling           0.4248     0.7817   0.543 0.596841
## Total.expenditure   0.6064     0.4320   1.404 0.185760
## under.five.deaths   2.5744     2.9929   0.860 0.406559
## infant.deaths      -4.3772     3.5453  -1.235 0.240591
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.549 on 12 degrees of freedom
## Multiple R-squared:  0.2346, Adjusted R-squared:  -0.02049
## F-statistic: 0.9197 on 4 and 12 DF,  p-value: 0.484
```

Except for the interception 0, all parameters are statistically insignificant in the chosen model, and the adjusted R^2 is close to 0, even negative. The optimal model for global data does not fit the developed countries’ situations at all.

Best predictors for global data: *Schooling, HIV.AIDS, Total expenditure, under.five.deaths* and *infant deaths*. And *Schooling* and *HIV.AIDS* play the most important role.

Best predictors for developing countries data: *Schooling, HIV.AIDS, Total expenditure, Thinness 1-19 years* and *Alcohol.*

Best predictors for developed countries data: *Thinness 1-19 years* and *Thinness 5-9 years.*

From the three optimal models for global, developed countries and developing countries’ life expectancy, we have found that the model for developing countries is similar to the global model. The best predictors for the global model also appear in developing countries' optimal model though the top predictors are different. In both developing countries and global environments, the number of years of schooling and HIV.AIDS plays a non-negligible role in both positive and negative aspects respectively. Moreover, compared to global and developing countries, developed countries have their unique top predictors. Their life expectancy data does not fit the optimal global model at all.

Conclusions

According to the box plot, there is an apparent difference in life expectancy between developed and developing countries. We found that the covariates of the developing model is similar to the covariates of the worlds model. It might be that there is a large number of developing countries in the data. Developing countries might be more poor compared to developed countries. Therefore, it might be harder for children to get schooling or receive an education. This is why schooling is the best predictor among all explanatory variables for whole world life expectancy. The number of kids who go to school in a country can determine whether the country is poor or not.

If we were to put developed countries data into global data, this would make the important variables not significant in the global model. Therefore, the developing countries are more similar to the global model. The developed countries include thinness 1-99 years and thinness 5-99 years as explanatory variables. Schooling and HIV.AIDS are the two most important explanatory variables for developing countries. However, Schooling and HIV.AIDS are not important factors for developed countries. This may be due to the wealthiness of the country. Developed countries will be more likely to be wealthier. Therefore, people will be able to go to school or see doctors. Hence, there are more complete education and medication systems in a developed country compared to a developing country.