678 Final project

Weixiao Li

2022-12-09

Abstract

In the context of continuous economic development, people are more and more concerned about their physical health while pursuing material improvement, especially the topic of life expectancy, which has become the focus in recent years. I will first group the variables and then examine the factors that affect life expectancy. I choose to use Multilevel Regression model to fit data. After reading this report, you will have a comprehensive understanding of the factors that influence life expectancy.

Introduction

Life expectancy is a statistical measure of the average time an organism is expected to live. My project will focus on the relationship between life expectancy and other factors. For example, I will an analysis of which factors have a positive and which have a negative impact on increasing life expectancy and find variables that have a significant effect, so individuals and governments can pay more attention to improving that aspects. The project will consist of the following parts: Abstract, Introduction, Method, Result, Discussion, Appendix, and Supplement.

Method

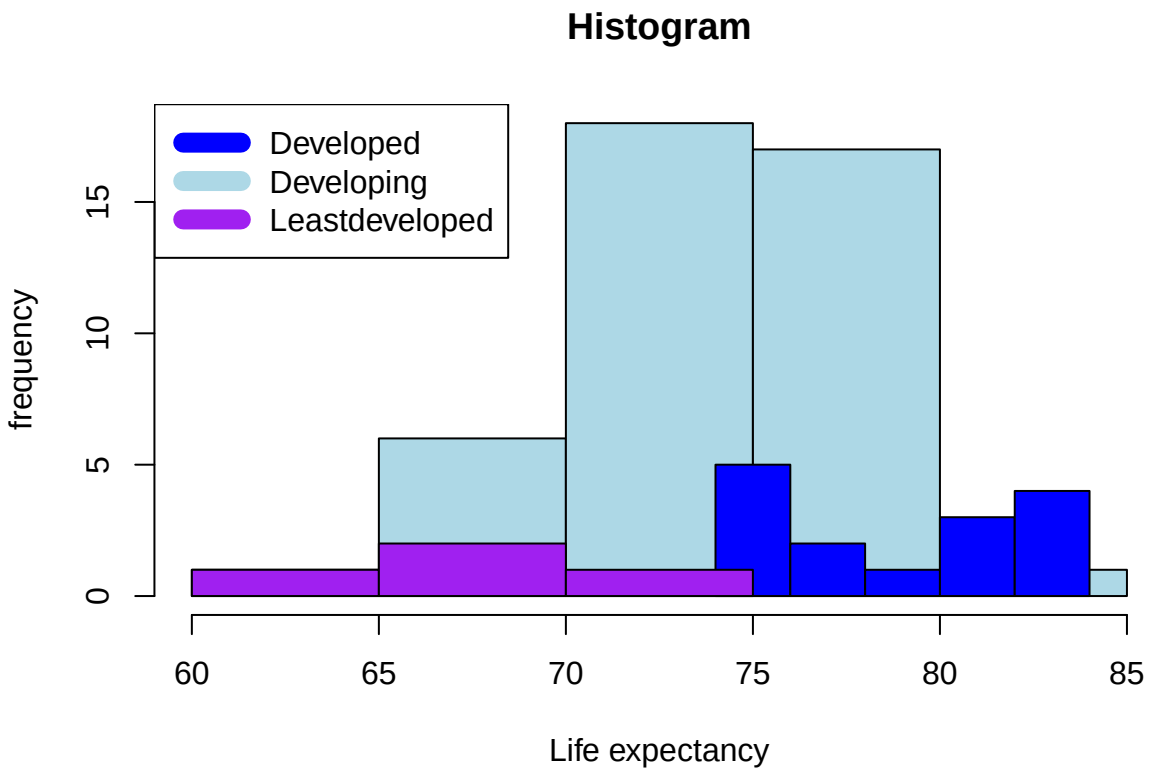
The dataset comes from the website Kaggle. There are nineteen variables included in the dataset originally. I will select several factors to represent the aspects of economic, social, and culture to have a deep analysis. Because the conditions of each country vary greatly, I will classify each country into five continents,and then group by continents and developing status, so that we can more intuitively and clearly see the differences between regions and status in the subsequent classification. Because Healthexppercapita, Electricity and Gdppercapita are very large, so I take a log of these items in the following analysis.

Column names	Explanation
Group	Grouped by continent and level of development
Literacy rate	Literacy rate, adult total (% of people ages 15 and above)
Homicid	Homicidies per 100k people
Electricity	Electric power consumption (kWh per capita)
Status	Economic development status of country
Wateraccess	Access to improved water sources (% of total population with access)
Tuberculosis	Incidence of tuberculosis (per 100,000 people)
Inflation	Inflation, consumer prices (annual %)

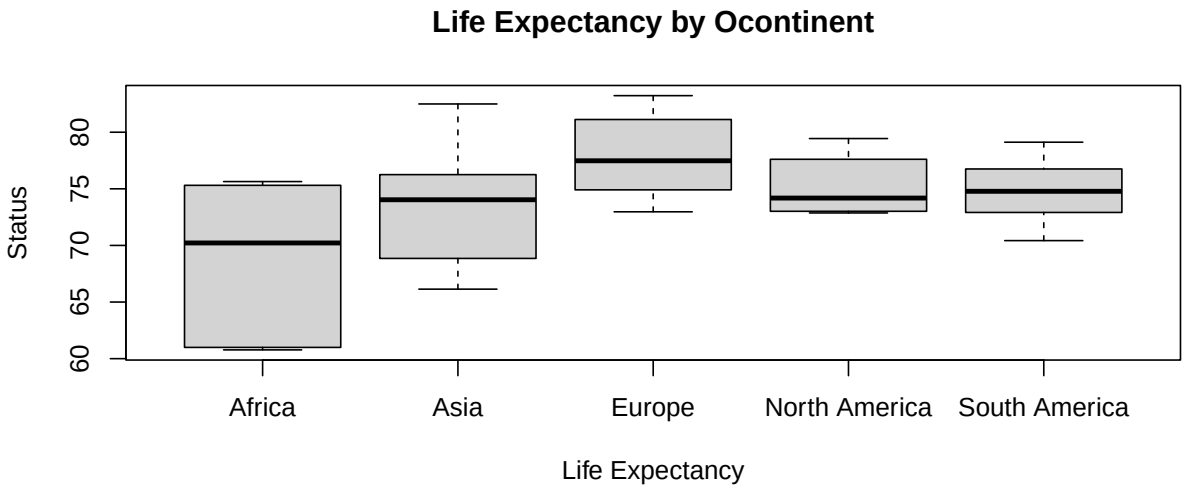
Column names	Explanation
Healthexppercapita	Average health expenditure per capita
Schooling	Number of years of Schooling (years)
HIV.AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)
Fertilityrate	Fertility rate, total (births per woman)
Lifeexp	Life expectancy
Gdppercapita	GDP per capita, PPP (current international $)
CO2	Average CO2 emissions (metric tons per capita)
Forest	Forest area (% of land area)
Urbanpop	Urban population
Urbanpopgrowth	Average urban population growth (annual %)

Exploratory Data Analysis

Grouping



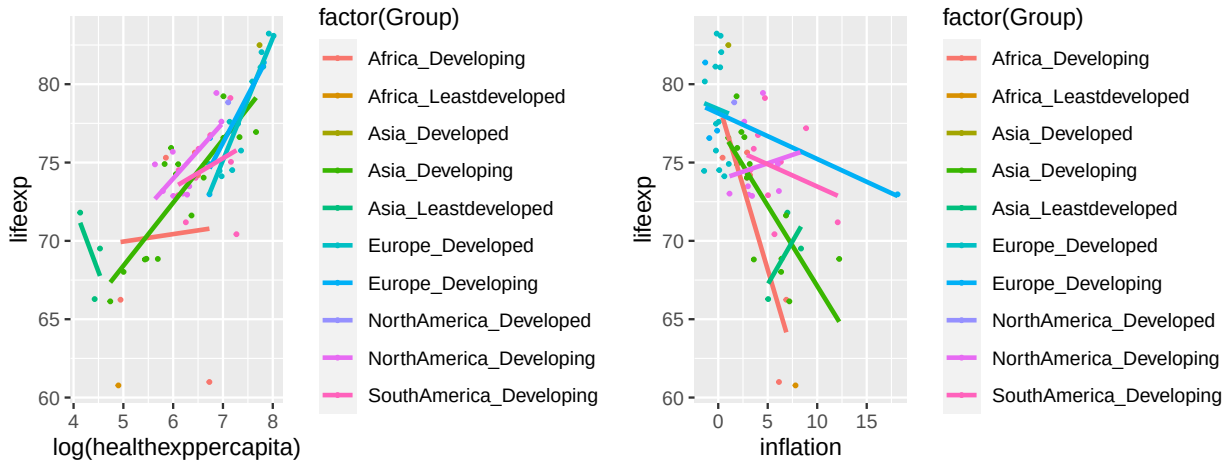
The plot shows the frequency distribution of life expectancy. As we can see that life expectancy varies a lot among different status countries. In developing countries is distributed across all ages, with life expectancy in developed countries at the higher end and life expectancy in countries that are lagging behind at the lower end. Thus, I decide to analyze what factors influence life expectancy in countries that in different status of development.



From the boxplot we can see that life expectancy varies a lot among continents, thus I want to group by status and continent and there will be ten groups.

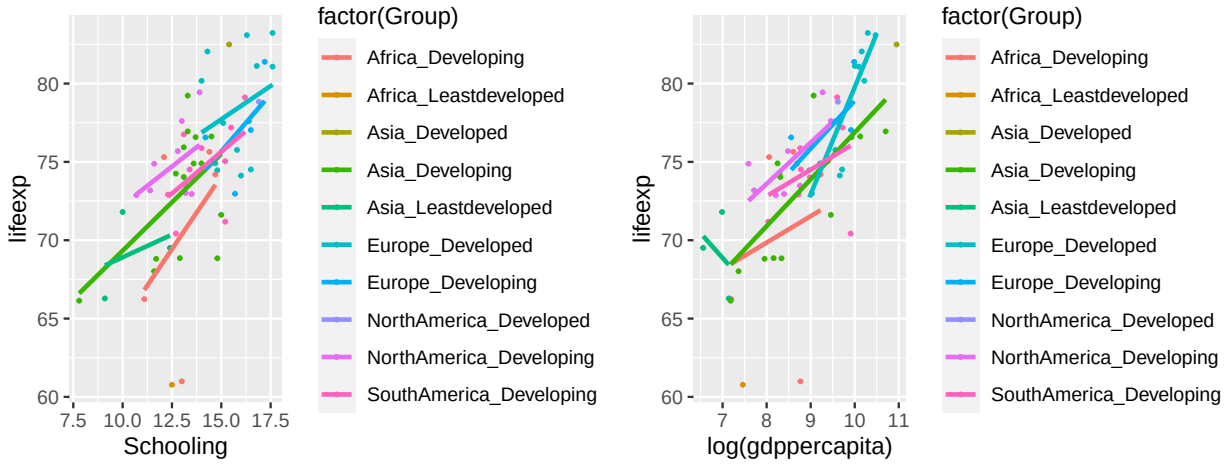
Selection of variables

a



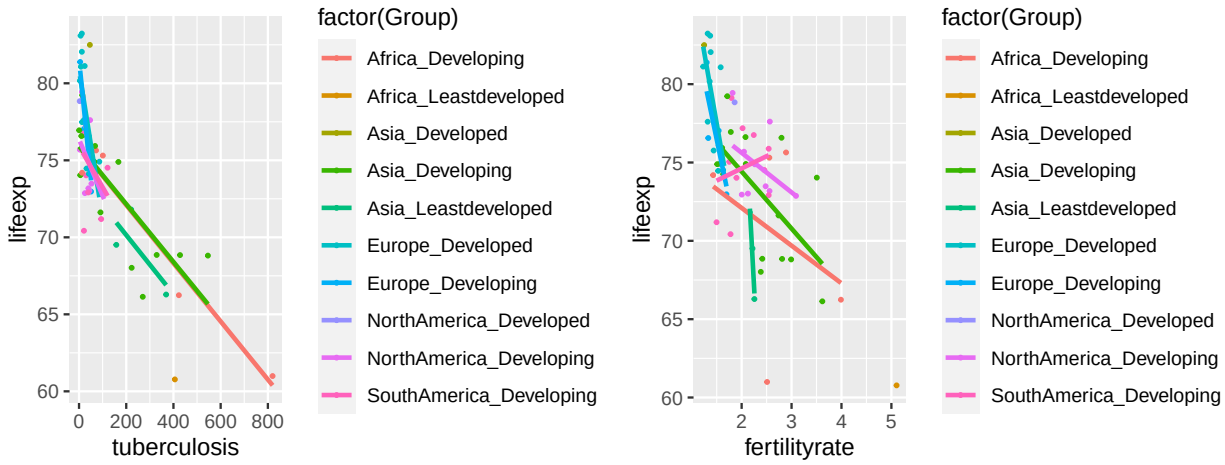
The left plot shows the relationship between life expectancy and health expenditure per capital. We can see that the two variables are in a positive related while in least developed Asia countries are negative. And the right plot illustrates that as inflation rises, person's life expectancy goes down in most groups. The trend remains consistent across the groups. Thus, I will keep these two variables in model to have a deep analysis.

b



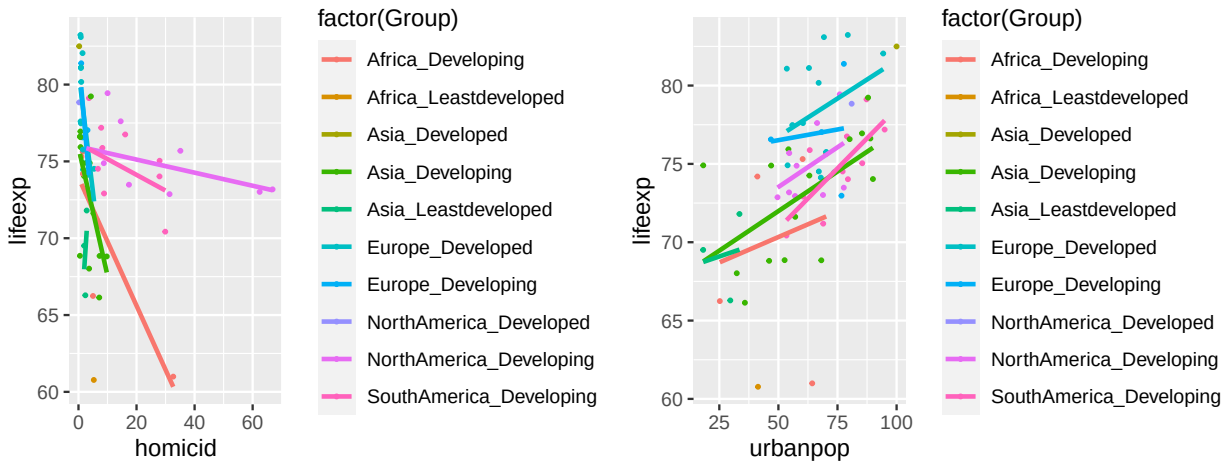
From the plot we can see that Schooling and log(gdppercapita) are in a positive relationship with life expectancy, even though GDP per capita in developing Africa is negatively correlate. The difference in each group is not big,therefore, I will remain the two items in the following analysis.

c



These two plots present that lower fertility rate and lower Incidence of tuberculosis in every group contribute to higher person's life expectancy, this trend is consistent across groups.

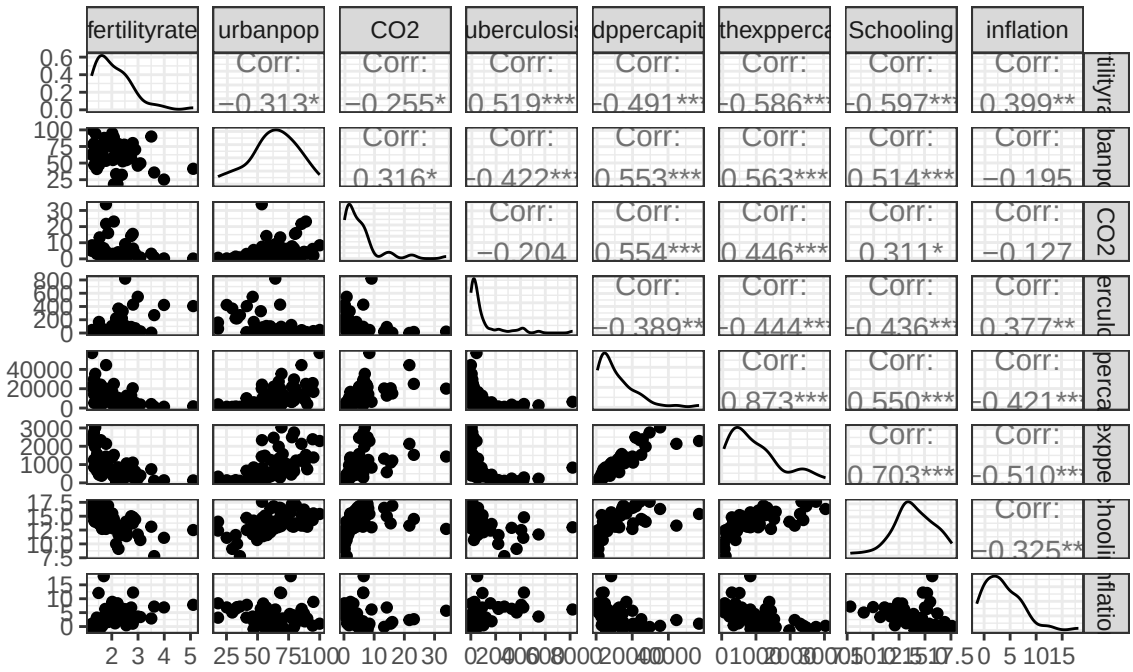
d



These plot illustrate that homicidies in a negative relationship with life expectancy while urban population in positive relationship with life expectancy in every group. For other variables, they are not suitable for multivariate linear analysis, thus I put them in appendix.

Model Fitting

Because the selecting of variables is important for the model fitting,I want to further screen the variables that are put into the model. Therefore, I tested the correlation between variables and the two variables that are highly correlated are substitutable for each other. So, the correlation plot will help me to streamline the variables improve the model accuracy

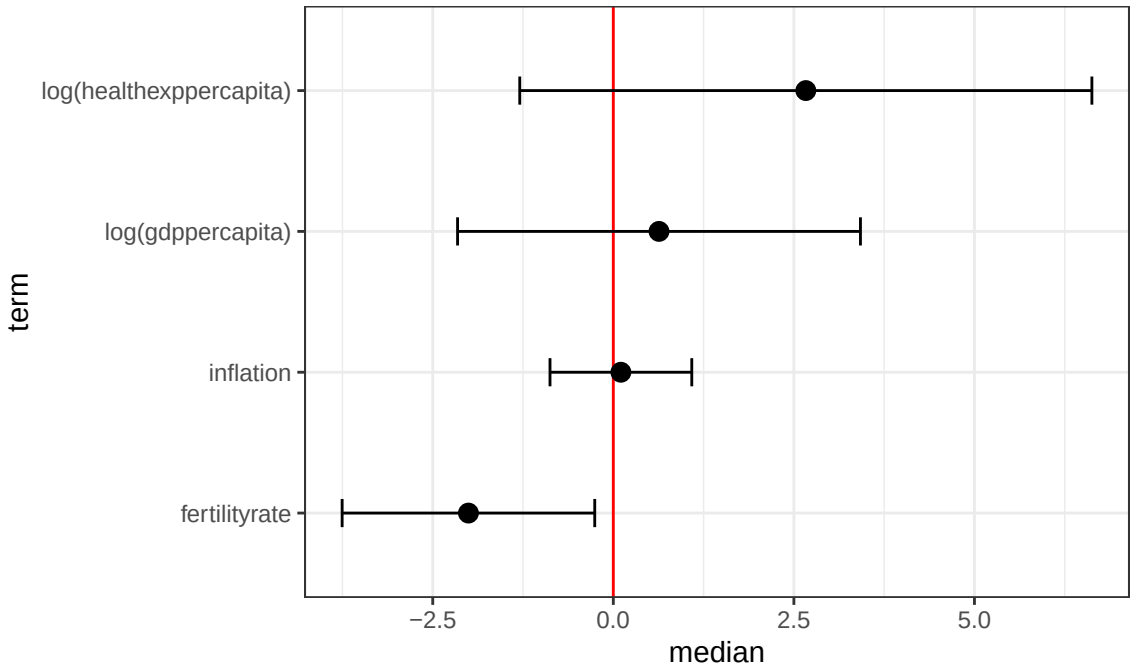


Let's look at the correlation plot. The correlation coefficients of the majority of variables are below 0.5, indicating that they are not highly correlated and can be put into the model at the same time. However, the correlation between schooling and healthexppercapita are very high, over 0.7.Because I am more interested in

expenditure, so I will keep the variable health consumption expenditure per capita in the model.Moreover, the coefficients of tuberculosis and fertility rate are relatively higher, thus I will choose fertility rate to analysis for it is a common perception that the incidence of disease is negatively correlated with population life expectancy. So I prefer to study the unclear relationship of fertility rate and life expectancy. In the end, I keep Fertility rate. log(healthexppercaptia), inflation, and log(gdppercapita) four economically as well as socially relevant variables for the study.

Here is the output of model:

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: lifeexp ~ fertilityrate + log(healthexppercapita) + inflation +
##     log(gdppercapita) + (1 + fertilityrate + log(healthexppercapita) +
##     inflation + log(gdppercapita) | Group)
##    Data: lifeexp
##
## REML criterion at convergence: 276.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2062 -0.4628  0.0199  0.6235  1.7120
##
## Random effects:
##  Groups   Name                    Variance Std.Dev. Corr
##  Group    (Intercept)             55.317   7.438
##           fertilityrate            3.388   1.841     0.67
##           log(healthexppercapita) 21.704   4.659    -0.54 -0.62
##           inflation                1.462   1.209    -0.83 -0.90  0.85
##           log(gdppercapita)        9.133   3.022     0.33  0.47 -0.97 -0.70
##  Residual                          3.459   1.860
## Number of obs: 62, groups:  Group, 10
##
## Fixed effects:
##                          Estimate Std. Error t value
## (Intercept)              54.71440    5.05152  10.831
## fertilityrate            -1.93084    0.82293  -2.346
## log(healthexppercapita)   2.62174    1.98231   1.323
## inflation                 0.06542    0.46769   0.140
## log(gdppercapita)         0.77848    1.39701   0.557
##
## Correlation of Fixed Effects:
##             (Intr) frtlty lg(hl) infltn
## fertilityrt  0.016
## lg(hlthxpp) -0.266 -0.384
## inflation   -0.549 -0.696  0.761
## lg(gdpprcp) -0.077  0.316 -0.937 -0.559
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```



Result

For the fixed effects: lifeexp~54.71-1.93*(fertilityrate+1) + 2.62*log(healthexppercapita+1)+0.065*(inflation+1)+ *0.78*(log(gdppercapita)). I choose Asia_developing, Europe developed and Africa Leastdeveloped groups as an example to interpret.Random effects you can see in the Appendix. The fixed effects added random effects. For Asia developing coubtries:lifeexp~60.49-1.27*(fertilityrate+1) + 3.17*log(healthexppercapita+1)-0.365*(inflation+1)- 0.27*(log(gdppercapita)) For Europe developed countries:lifeexp~46.67-3.66*(fertilityrate+1) + 6.4*log(healthexppercapita+1)+1.375*(inflation+1)-1.02*(log(gdppercapita)) For Africa Leastdeveloped countries:lifeexp~52,28-0.5*(fertilityrate+1) + 3.98*log(healthexppercapita+1)+ 0.745*(inflation+1)+0.2*(log(gdppercapita))

Interpretation

For Asia developing countries, we can conclude that fertility rate, inflation and gdp per capital have negative relationship with life expectancy. For fertility rate and inflation rise 1 point with others fixed, the life expectancy will decrease 1.27 years and 0.365 years respectively, and for gdp per capital increase 1%, the life expectancy will decrease 0.27 years on average. Healthy expenditure per capital have a positive relationship with life expectancy, every one percent increase in healthexppercapita, life expectancy will increase 3.17 years on average. And for Europe developed countries, the coefficient of fertility rate, healthy expenditure per capital, gdp per capital and inflation are all larger, which means that the effect of these variables on life expectancy is greater in Europe developed regions than in Asia developing regions. And the intercept is smaller. Similarly, We can conclude that the effect of inflation and gdp per capital is least in Africa Leastdeveloped countries, and Healthy expenditure per capital has the greatest impact on the African region compared to the other two regions.

Model Checking

From the residual plot we can see that most points depicted are randomly scattered above and below the line with 0 as the horizontal axis,in the range of[-2,2], while several points that deviate more from the zero value.

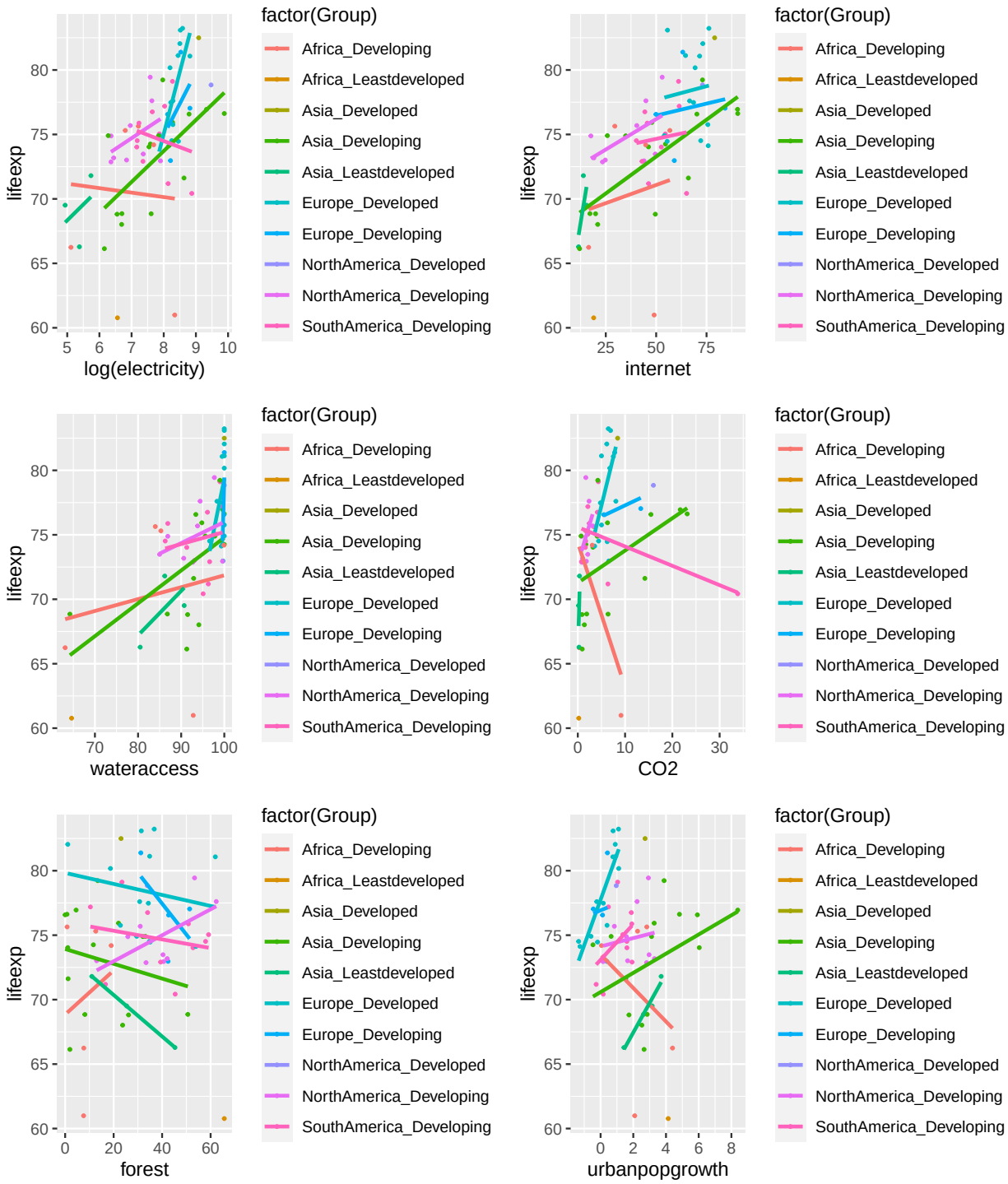
Discussion

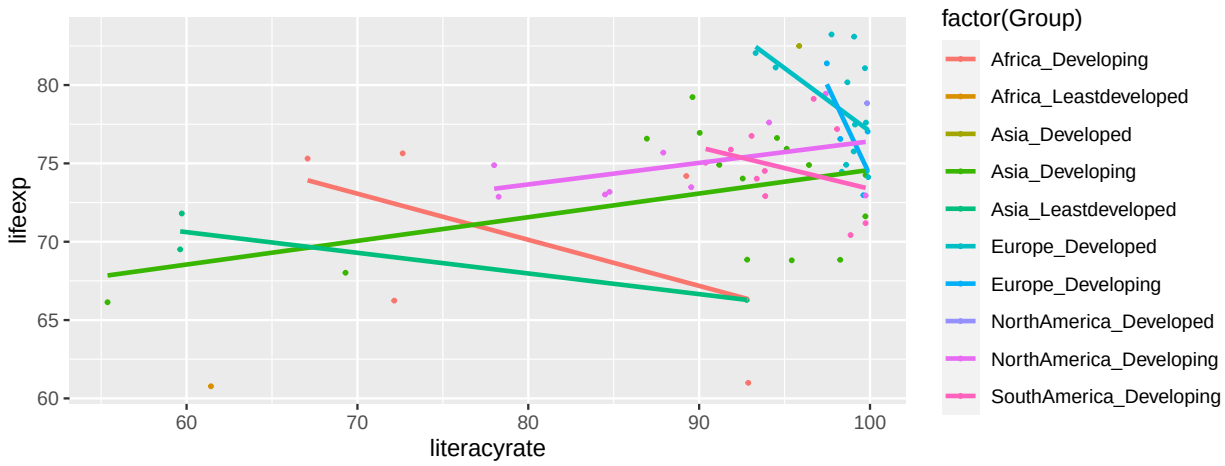
Using multivariate linear model allows me to see the relationship between each variable and life expectancy under different group categories. And which variables such as per capita health consumption expenditure, remaining positively correlated with life expectancy in each group, as well as some variables that have different relationships with life expectancy in different group, for example, inflation is negatively correlated with life expectancy in developing countries in each region and positively correlated in other development status.

However, there are some shortcomings, the limited selection of variables and the fact that some groups contain fewer countries , which can cause a large error. As well as there are several variables with low linear correlation.So in the next steps, I will add some variables that can represent each convenient aspect, such as cultural, social, economic. The variables I put in the model are more related to the economy now.

Appendix

```
##                          (Intercept) fertilityrate log(healthexppercapita)
## Africa_Developing               9.16          2.87                    -9.24
## Africa_Leastdeveloped          -2.43         -1.43                     1.36
## Asia_Developed                 -1.18         -0.30                    -0.52
## Asia_Developing                 5.78          0.66                     0.55
## Asia_Leastdeveloped            -4.59         -1.43                    -0.07
## Europe_Developed               -8.04         -1.73                     3.78
## Europe_Developing               1.57          0.49                    -0.30
## NorthAmerica_Developed         -0.68         -0.19                     0.82
## NorthAmerica_Developing        -2.75         -0.22                     2.94
## SouthAmerica_Developing         3.16          1.27                     0.69
##                          inflation log(gdppercapita)
## Africa_Developing            -2.27              5.62
## Africa_Leastdeveloped         0.68             -0.58
## Asia_Developed                0.09              0.55
## Asia_Developing              -0.43             -1.05
## Asia_Leastdeveloped           0.64              0.69
## Europe_Developed              1.31             -1.80
## Europe_Developing            -0.25              0.00
## NorthAmerica_Developed        0.18             -0.52
## NorthAmerica_Developing       0.52             -1.90
## SouthAmerica_Developing      -0.45             -1.00
```


lifeexp
log(electricity)
internet
wateraccess
CO2
forest
urbanpopgrowth
factor(Group)
Africa_Developing
Africa_Leastdeveloped
Asia_Developed
Asia_Developing
Asia_Leastdeveloped
Europe_Developed
Europe_Developing
NorthAmerica_Developed
NorthAmerica_Developing
SouthAmerica_Developing


lifeexp
80
75
70
65
60
60
70
80
90
100
literacyrate
factor(Group)
Africa_Developing
Africa_Leastdeveloped
Asia_Developed
Asia_Developing
Asia_Leastdeveloped
Europe_Developed
Europe_Developing
NorthAmerica_Developed
NorthAmerica_Developing
SouthAmerica_Developing