Midterm Project

Yaquan Yang

2022-12-03

Abstract

This report focuses on using multilevel linear mixed models to examine the factors influencing human life expectancy at the global level and how that influence is affected by country and year differences.

Introduction

The main question of this project was to identify the main effective predictors of life expectancy. In short, we need to answer the question: if a health organization wants to increase life expectancy somewhere, what variables can they change in order to reach their goal?

The public dataset I used provides data for 193 countries from 2000 to 2015 and has a structure of 2938 rows (data points) divided into 22 columns (features). These features can be divided into two groups.

Health factors like "HIV", "Under Five Deaths", "Adult Mortality", "BMI" etc.

Economic factors like "GDP", "Income Composition of Resources", "Status" etc.

Since the data set is from the World Health Organization, we consider the data to be authentic and reliable. Most of the missing data are for population, hepatitis B, and GDP. The missing data came from less-known countries such as Vanuatu, Tonga, Togo, Cape Verde, etc. Finding all the data for these countries was difficult, so we decided to ignore the missing data

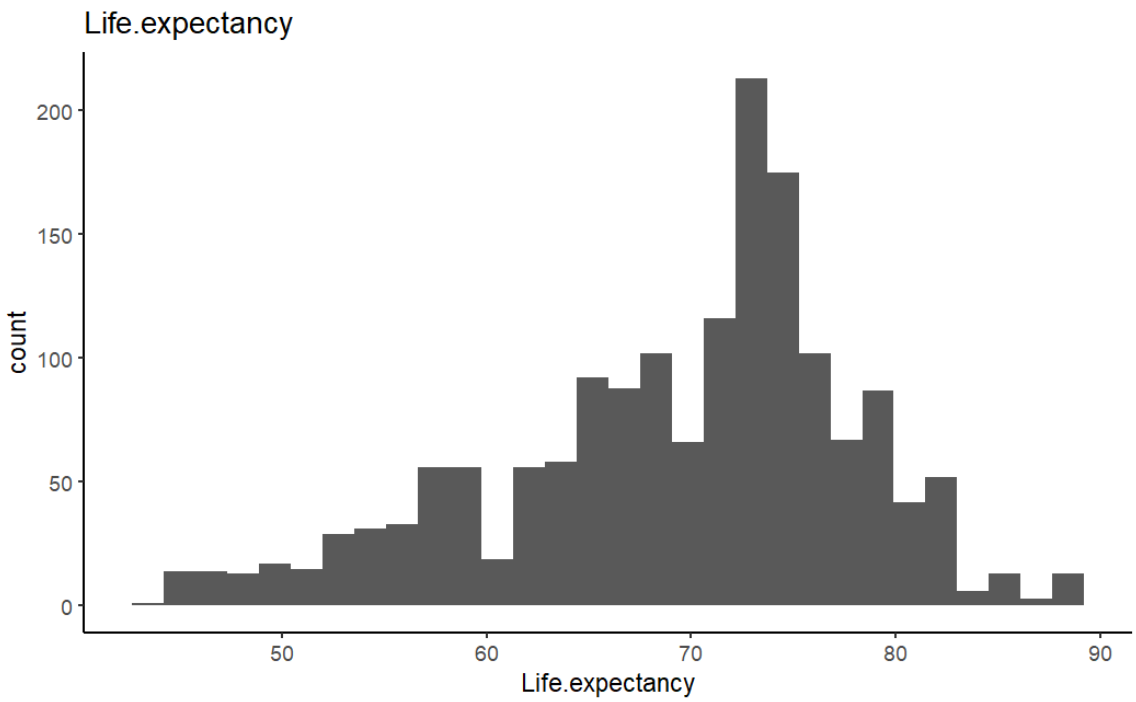
Exploratory Data Analysis

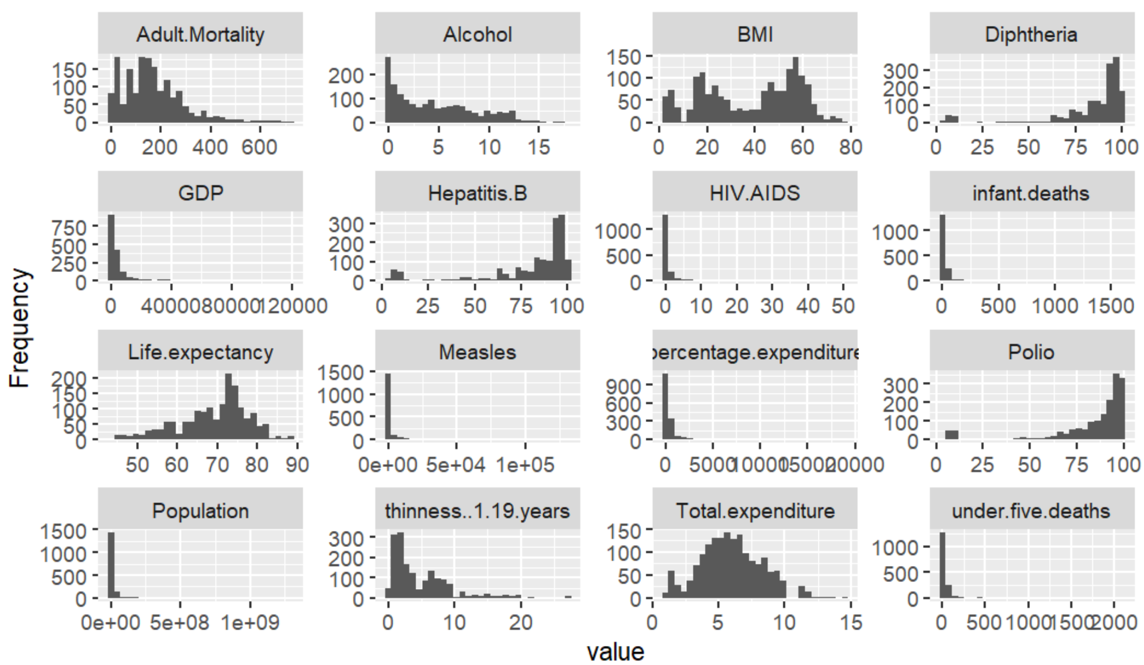
Data Source: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

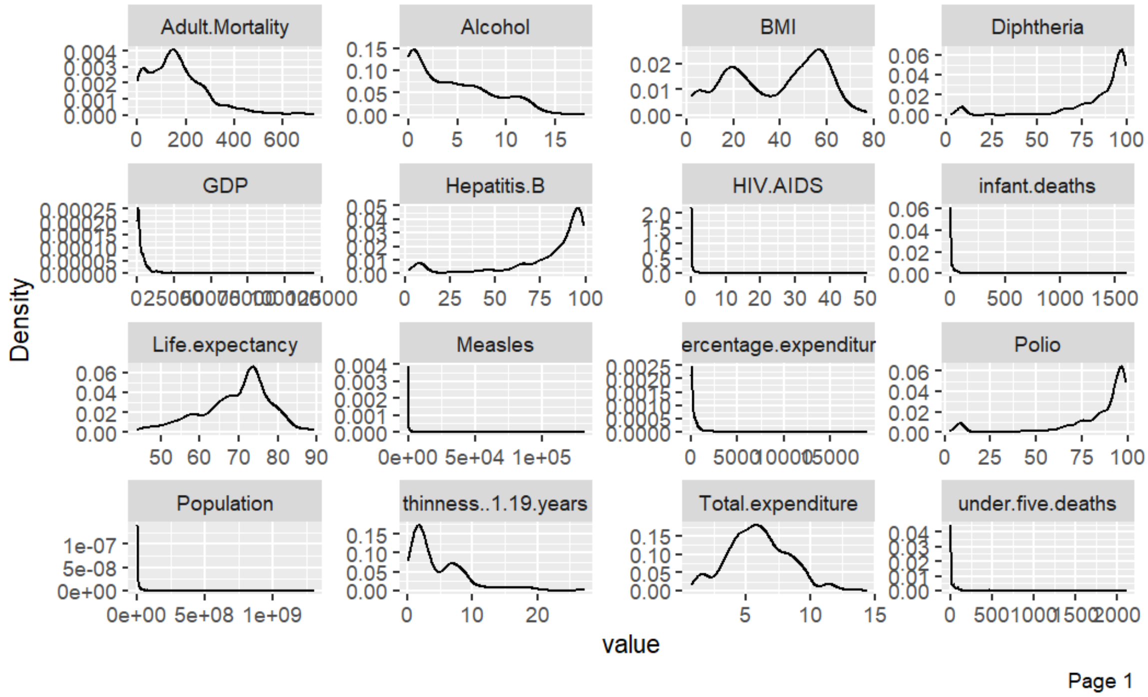
Data Description:

Column Name	Explanation
Country	Country
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
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 (%)
HIV/AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)
GDP	Gross Domestic Product per capita (in USD)
Schooling	Number of years of Schooling(years)
Income composition of resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
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(%)
Population	Population of the country

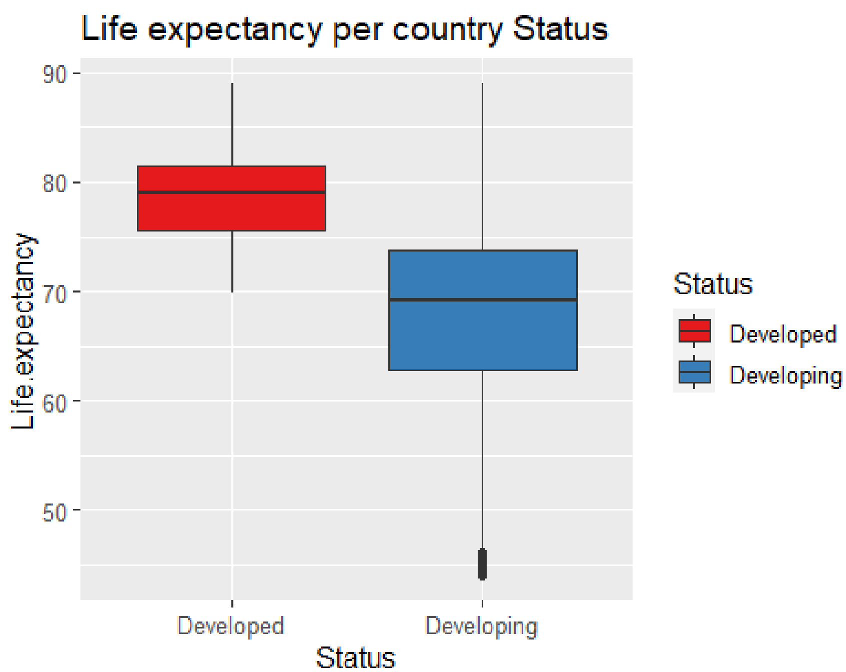
First, we did the data cleaning and processing, and do descriptive statistical analysis. Making graphs to show the distribution of each variable.


Life.expectancy
count
Life.expectancy


Adult.Mortality
Alcohol
BMI
Diphtheria
GDP
Hepatitis.B
HIV.AIDS
infant.deaths
Life.expectancy
Measles
percentage.expenditure
Polio
Population
thinness..1.19.years
Total.expenditure
under.five.deaths
Frequency
value



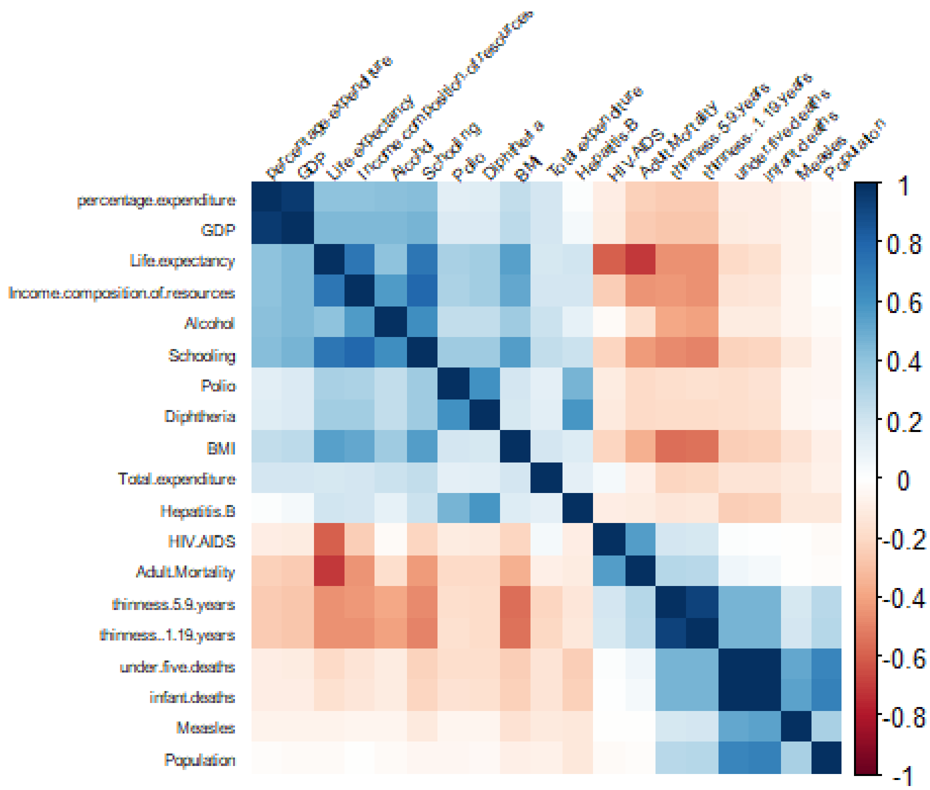
Categorical variables:



We see that life expectancy is higher in developed countries, which means that the categorical variables can be a good predictor for the model.

Correlation:

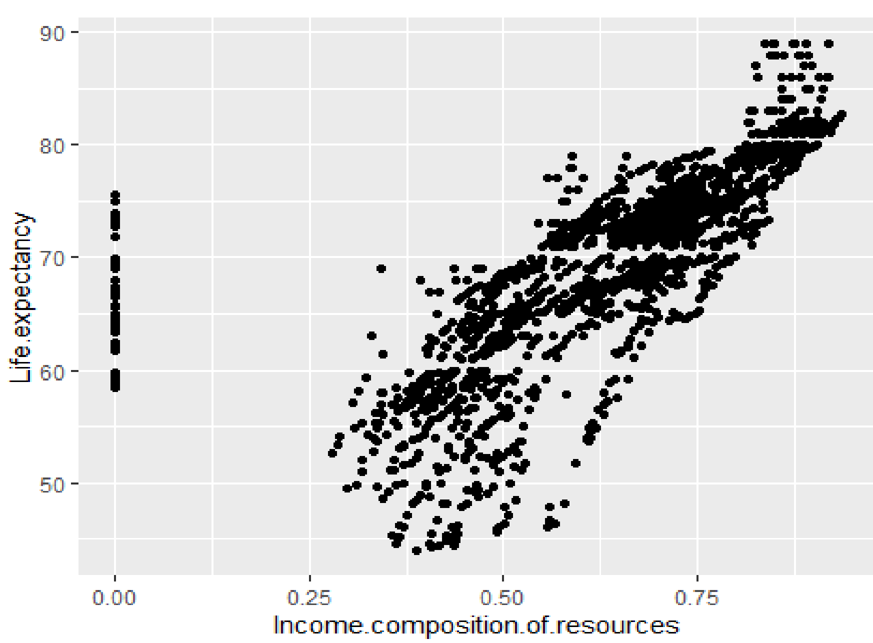
Next, we need to consider the correlation between different characteristic variables and life expectancy, and we make a plot of the correlation matrix to show the correlation between them, which facilitates our better selection of predictor variables.



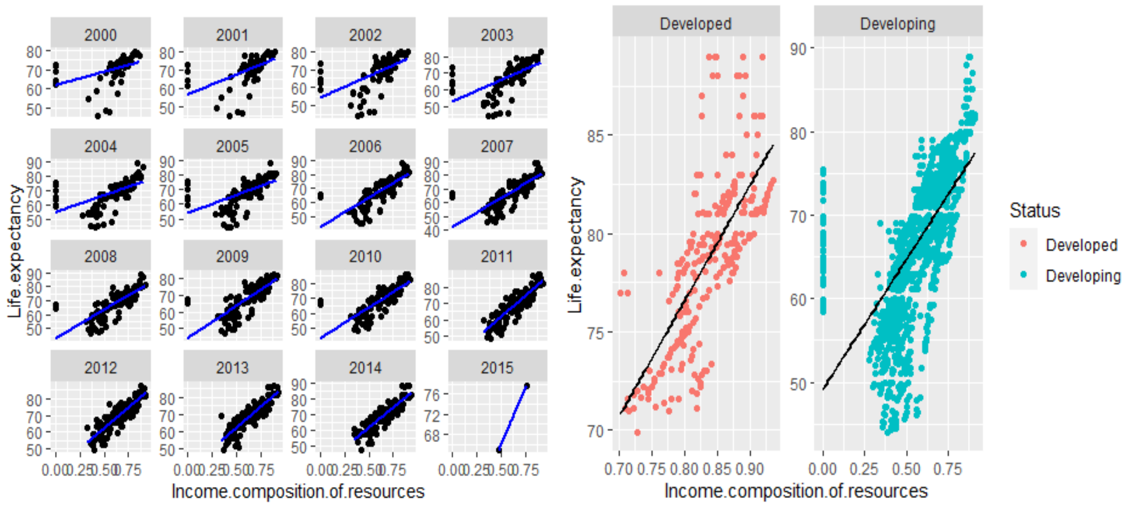
Moreover, two features with high correlation will directly show strong crosstalk, so we should try to avoid this in our model:

Infant deaths & Under-five deaths. GDP & Percentage expenditure. Thinness..1.19.years & Thinness.5.9.years.

For the above pairs, we thus need to eliminate one of the two when choosing our predictor.



I wanted to use “Income composition of resources” to predict Life expectancy, in the scatter plot of BMI and Life expectancy, I found that the linear relationship between them is not obvious, there seem to be multiple levels of linearity, so next, I plot the logarithm of the population against life expectancy by grouping the data by year.



From this set of plots we can feel that the difference in the effect of Income composition of resources on life expectancy with the year can be represented by the slope of the regression line.

Model fitting

We make use of the lme4 package for fitting mixed effects models, and some supplementary packages: *lmerTest* provides tools for obtaining p-values.

Model_1:

To start, we often fit an unconditional means model that provides us with information about within-group differences between years.

Fitting the unconditional means model:

```
m1 <- lmer( Life.expectancy ~ 1 + (1 | Year), mydata)
summary(m1)
```

Results:

```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5722 -0.6799  0.3197  0.6635  2.2410

Random effects:
 Groups   Name        Variance Std.Dev.
 Year     (Intercept)  2.34    1.530
 Residual             88.51    9.408
Number of obs: 2928, groups:  Year, 16

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  69.2249     0.4201 15.0000   164.8   <2e-16 ***
```

We extract the random effects with the VarCorr() function:

```
##  Groups   Name        Std.Dev.
##  Year     (Intercept) 0.045588
##  Residual             8.796724
```

The estimated correlation between the random effects is -1.00 The results are highly significant, indicating that our model fit is on the right path.

Model_2:

Now, let's add the predictor:

```
m2 <- lmer( Life.expectancy ~ BMI + Alcohol+ Total.expenditure + GDP + Schooling +
HIV.AIDS+infant.deaths+Adult.Mortality+Polio+Income.composition.of.resources+Status+Year+(1
+ Status | Country), mydata)
```

Results:

```
Scaled residuals:
     Min       1Q   Median       3Q      Max
-12.6067  -0.4657  -0.1295   0.2614   5.4643

Random effects:
 Groups   Name             Variance Std.Dev. Corr
 Country  (Intercept)       8.769   2.961
          StatusDeveloping 20.882   4.570    0.71
 Residual                   3.441   1.855
Number of obs: 2301, groups:  Country, 155

Fixed effects:
                                  Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                     -4.158e+02  2.485e+01  1.797e+03 -16.735  < 2e-16 ***
BMI                              1.766e-03  3.281e-03  2.157e+03   0.538  0.59040
Alcohol                         -8.762e-02  2.912e-02  2.225e+03  -3.009  0.00265 **
Total.expenditure               -5.089e-02  2.412e-02  2.160e+03  -2.110  0.03496 *
GDP                              2.623e-06  4.219e-06  2.164e+03   0.622  0.53423
Schooling                        2.458e-01  4.686e-02  2.228e+03   5.246 1.70e-07 ***
HIV.AIDS                        -3.248e-01  1.582e-02  2.194e+03 -20.532  < 2e-16 ***
infant.deaths                   -4.469e-03  1.519e-03  2.120e+03  -2.942  0.00330 **
Adult.Mortality                 -1.904e-03  4.789e-04  2.142e+03  -3.975 7.28e-05 ***
Polio                            4.487e-03  2.270e-03  2.137e+03   1.977  0.04818 *
Income.composition.of.resources  8.398e-01  4.771e-01  2.162e+03   1.760  0.07848 .
StatusDeveloping                -1.024e+01  8.916e-01  1.189e+02 -11.490  < 2e-16 ***
Year                             2.448e-01  1.259e-02  1.799e+03  19.441  < 2e-16 ***
```

Predictor variables that have a significant positive association with life expectancy are "Schooling", and "Year". Predictor variables that have a significant negative correlation are "HIV", "Adult Mortality" and "Alcohol".

Model_3:

From the fitting results of model-2 we can see that the effect of BMI and GDP on life expectancy is insignificant, so we decided to exclude these predictor variables and add both year and country as random effects to the model fitting to obtain model-3

```
m3 <- lmer( Life.expectancy ~ Alcohol + GDP + Schooling +  HIV.AIDS + infant.deaths +
Adult.Mortality + Income.composition.of.resources+Status + (1 + Status| Country) + (1 +
Status | Year), mydata)
```

Results:

```
Scaled residuals:
     Min       1Q   Median       3Q      Max
-12.5921  -0.4531  -0.1404   0.2737   5.5561

Random effects:
 Groups   Name              Variance  Std.Dev. Corr
 Country  (Intercept)        8.705294 2.95047
          StatusDeveloping 18.096819 4.25404  0.98
 Year     (Intercept)        1.375879 1.17298
          StatusDeveloping  0.001062 0.03259  -1.00
 Residual                    3.453732 1.85842
Number of obs: 2326, groups:  Country, 156; Year, 16

Fixed effects:
                                  Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                      7.679e+01  9.319e-01  1.587e+02  82.403  < 2e-16 ***
Alcohol                         -1.143e-01  3.047e-02  2.243e+03  -3.750 0.000181 ***
GDP                              2.184e-06  4.235e-06  2.151e+03   0.516 0.606217
Schooling                        2.056e-01  4.220e-02  2.236e+03   4.872 1.18e-06 ***
HIV.AIDS                        -3.243e-01  1.583e-02  2.182e+03 -20.480  < 2e-16 ***
infant.deaths                   -5.027e-03  1.522e-03  2.154e+03  -3.303 0.000971 ***
Adult.Mortality                 -1.862e-03  4.798e-04  2.152e+03  -3.880 0.000108 ***
Income.composition.of.resources  7.095e-01  4.749e-01  2.164e+03   1.494 0.135327
StatusDeveloping                -1.065e+01  8.934e-01  1.202e+02 -11.926  < 2e-16 ***

Correlation of Fixed Effects:
            (Intr) Alcohl GDP    Schlng HIV.AI infnt. Adlt.M Inc...
Alcohol     -0.247
GDP         -0.114 -0.029
Schooling   -0.522 -0.085  0.018
HIV.AIDS     0.075 -0.106 -0.037 -0.024
infant.dths -0.053  0.068 -0.021  0.065  0.011
Adlt.Mrtlty -0.069 -0.013  0.001  0.037 -0.154 -0.016
Incm.cmps.. -0.145 -0.013  0.021 -0.401 -0.037 -0.032  0.013
StatsDvlpng -0.580  0.192  0.090  0.139 -0.069 -0.041 -0.040  0.059
fit warnings:
Some predictor variables are on very different scales: consider rescaling
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.0021422 (tol = 0.002, component 1)
```

We use model 3 as an example for parameter interpretation:

Fixed Effects:

- (Intercept): When all predictor variables are zero, the average of life expectancy is 76.79.
- Alcohol: For every unit increase in alcohol consumption reduces life expectancy by 0.11, which is statistically significant.
- Schooling: For every unit increase in the number of years of Schooling, positive affect is expected to increase by 0.206, which is statistically significant.
- HIV.AIDS: Higher number of deaths per 1 000 live births HIV/AIDS tended to experience higher negative affect, 0.324, which is statistically significant.
- Infant Deaths: Higher number of Infant Deaths per 1000 population tended to experience a higher negative affect, 0.005, which is statistically significant.

- Adult Mortality: The higher probability of dying between 15 and 60 years per 1000 population, the lower the life expectancy(-0.02).
- Status Developing: Life expectancy in developing countries is 10.65 years less than in developed countries.

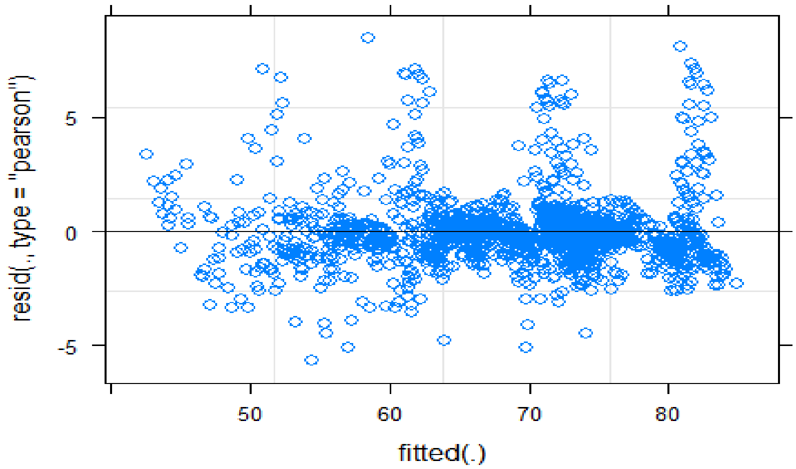
Random Effects:

- (Intercept): There are intergroup differences in life expectancy among populations in different countries(8.71) and years (1.38).
- Status Developing: In the relationship between StatusDeveloping and life expectancy, there are significant effects across countries(18.10).
- corr((Intercept), Country: The correlation between the random intercept and random slope was 0.98, which indicates that those countries that had higher intercepts for Life expectancy were also more likely to have greater (more positive) associations between Status and Life expectancy.

Model Check

We can also get confidence intervals for the fixed and random effects:

```
                                    2.5 %        97.5 %
.sig01                          2.100782e+00  4.082824e+00
.sig02                         -1.000000e+00  1.000000e+00
.sig03                          1.668902e+00  7.613170e+00
.sig04                          6.682648e-01  1.737829e+00
.sig05                         -1.000000e+00 -5.506819e-01
.sig06                          1.449827e-01  9.141368e-01
.sigma                          1.616137e+00  1.738597e+00
(Intercept)                     6.473424e+01  7.101473e+01
Alcohol                        -1.314470e-01 -7.453737e-03
GDP                            -4.260733e-06  1.852999e-05
Schooling                       4.624792e-01  7.761089e-01
HIV.AIDS                       -3.568544e-01 -2.943576e-01
infant.deaths                  -7.266685e-03  2.324472e-04
Adult.Mortality                -2.183802e-03 -3.728005e-05
Income.composition.of.resources 1.135112e+00  3.534635e+00
StatusDeveloping               -1.008579e+01 -6.120543e+00
```



Since the residual plots do not show a clear pattern and the distribution is relatively homogeneous, we consider the model fit to be better.

Result

By fitting a linear mixed model to analyze the data, we can conclude that the predictor variables that have a significant positive effect on life expectancy are “Schooling” and “Income composition of resources”. Both of them are economic factors. Moreover, the predictor variables that had a significant negative effect on life expectancy are “Alcohol”, “HIV”, “Infant deaths”, and “Adult Mortality”. All of them are health factors.

In addition, we have seen a worldwide trend of increasing life expectancy year by year, which may be related to technological advances, medical improvements, economic development and improved quality of life.

Discussion

In this report, a multilevel mixed model was used to calculate the relationship between life expectancy and economic and health factors for the global population. In addition, the model considers two group levels: country and year. In general, most of the predictors have a significant effect on life expectancy from a fixed effects perspective. From a random effects perspective, different countries and years also have an impact on the slope and intercept of each predictor variable. Finally, the validity of the model is supported by the model checks I did.

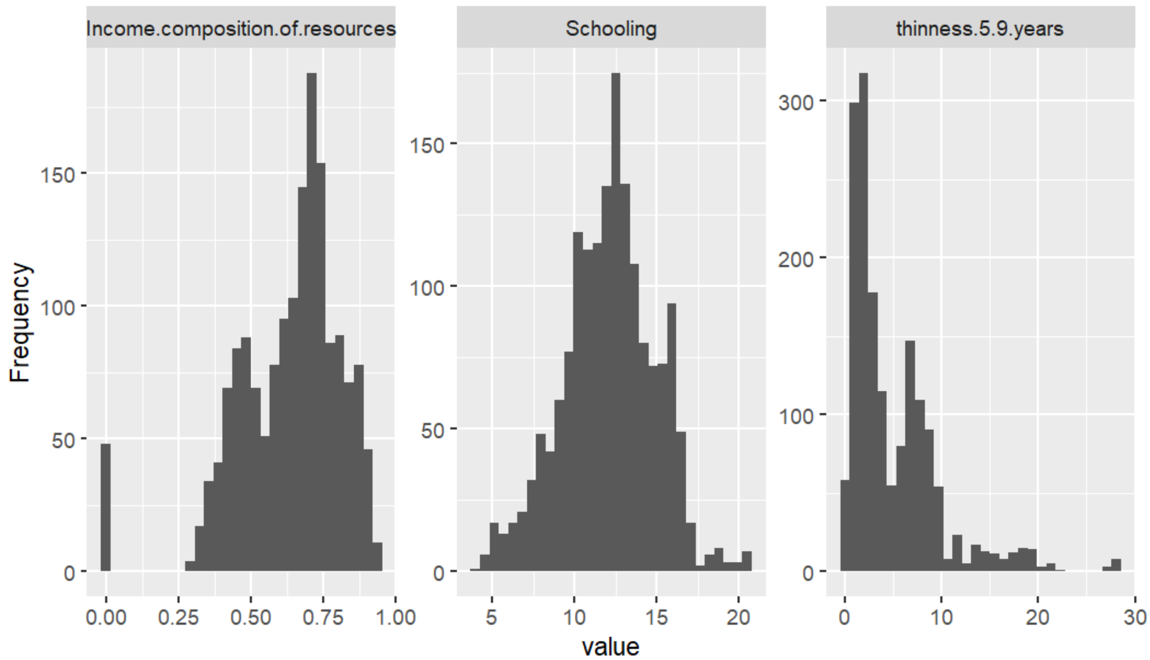
This report also has some limitations. Since the only publicly available datasets are from 2000 to 2015, this is more outdated for today in 2022, especially not considering the impact of the new crown pneumonia epidemic starting in 2020. This will undoubtedly have an impact on global population life expectancy.

Reference

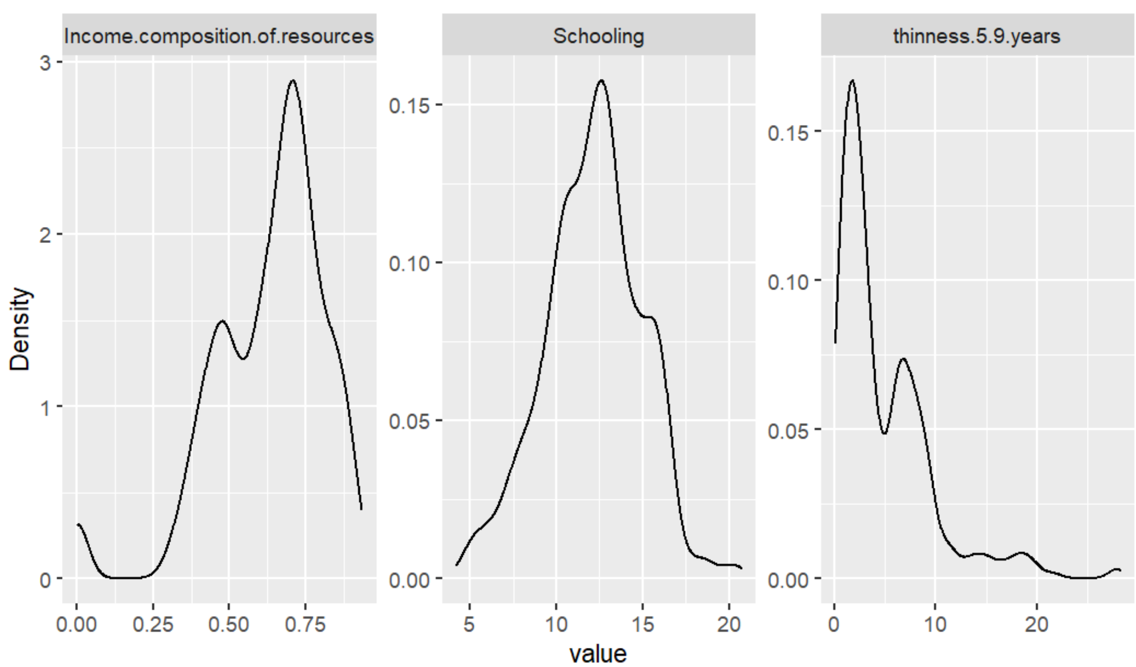
https://quantdev.ssri.psu.edu/tutorials/r-bootcamp-introduction-multilevel-model-and-interactions

https://www.kaggle.com/code/mohamedelsaadany/statistical-modeling-of-life-expectancy-data-r/notebook

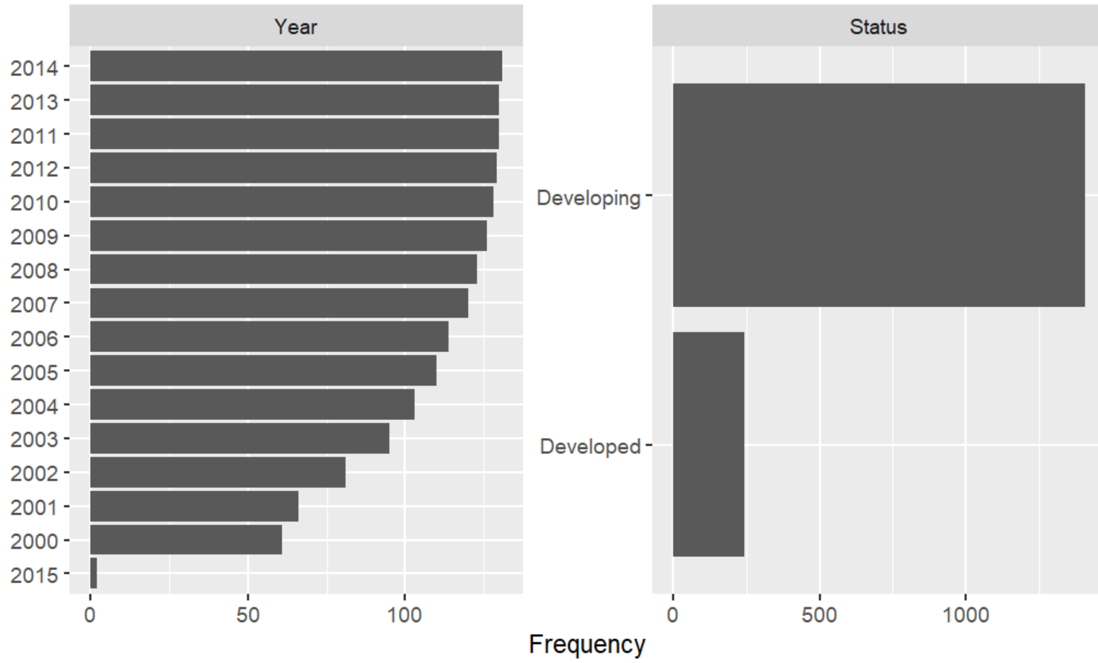
Appendix



Page 2



Page 2


Year
Status
2014
2013
2011
2012
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
2000
2015
0
50
100
Developing
Developed
0
500
1000
Frequency