Investigating the Predictive Factors of Life Expectancy Across Countries

Lingjie Chen, Yidong Wu, Yuyin Yang

2023-12-16

Introduction

In recent years, the relationship between socio-economic factors and health has attracted significant attention. Understanding these dynamics is crucial for policymakers and health organizations aiming to improve life expectancy and overall health in various regions. This project focuses on examining how various factors, such as economic conditions, immunization rates, education levels and health-related metrics, influence life expectancy across different countries. It spans a period from 2000 to 2015, covering a diverse range of countries, thereby offering a comprehensive view of global health trends.

The primary objective of this project is to construct a linear model that effectively utilizes a range of explanatory variables to estimate life expectancy. By identifying and quantifying the impact of various socio-economic and health indicators, this project aims to shed light on the crucial factors of life expectancy across nations. This, in turn, can inform targeted interventions and policies to enhance health situations globally.

Data Description

Data Source and Collection

The dataset used in this project was initially sourced from Kaggle. It covered a wide range of variables related to life expectancy: health metrics, immunization rates, and socio-economic indicators from 179 countries, covering the years 2000 to 2015. However, the dataset was plagued with severe inaccuracies and missing values.

Data Update and Validation

Key demographic and economic data such as population, GDP, and life expectancy figures were updated in line with the World Bank's authoritative datasets. Health-related information, including vaccination rates for diseases like Measles, Hepatitis B, Polio, and Diphtheria, along with data on alcohol consumption, Body Mass Index (BMI), HIV incidents, mortality rates, and prevalence of thinness, was carefully compiled from the WHO's public datasets. Additionally, educational metrics, specifically schooling data, were sourced from 'Our World in Data', a project affiliated with the University of Oxford. The updated dataset can be downloaded from here.

Data Structure

We first load the data and change the names of columns for simplicity. Then we take a glimpse of our data.

```
## Rows: 2,864
## Columns: 21
## $ country                     <chr> "Afghanistan", "Afghanistan", "Afghanistan~
## $ region                      <chr> "Asia", "Asia", "Asia", "Asia", "Asia", "A~
## $ year                        <int> 2000, 2001, 2002, 2003, 2004, 2005, 2006, ~
## $ infant_deaths               <dbl> 90.5, 87.9, 85.3, 82.7, 80.0, 77.3, 74.6, ~
## $ under_five_deaths           <dbl> 129.2, 125.2, 121.1, 116.9, 112.6, 108.4, ~
## $ adult_mortality             <dbl> 310.8305, 304.8580, 298.8855, 292.0365, 28~
## $ alcohol                     <dbl> 0.020, 0.020, 0.020, 0.020, 0.020, 0.016, ~
## $ hepatitis_B                 <int> 62, 63, 64, 65, 67, 66, 64, 63, 64, 63, 66~
## $ measles                     <int> 12, 13, 14, 15, 16, 17, 18, 21, 23, 24, 29~
## $ bmi                         <dbl> 21.7, 21.8, 21.9, 22.0, 22.1, 22.2, 22.3, ~
## $ polio                       <int> 24, 35, 36, 41, 50, 58, 58, 63, 64, 63, 66~
## $ diphtheria                  <int> 24, 33, 36, 41, 50, 58, 58, 63, 64, 63, 66~
## $ hiv                         <dbl> 0.02, 0.02, 0.02, 0.02, 0.02, 0.02, 0.02, ~
## $ gdp                         <int> 148, 163, 320, 332, 323, 346, 354, 393, 39~
## $ population                  <dbl> 20.78, 21.61, 22.60, 23.68, 24.73, 25.65, ~
## $ thinness_ten_nineteen_years <dbl> 2.3, 2.1, 19.9, 19.7, 19.5, 19.3, 19.2, 19~
## $ thinness_five_nine_years    <dbl> 2.5, 2.4, 2.2, 19.9, 19.7, 19.5, 19.3, 19.~
## $ school                      <dbl> 2.2, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.9, 3.~
## $ developed                   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ developing                  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ life                        <dbl> 55.8, 56.3, 56.8, 57.3, 57.8, 58.3, 58.8, ~
```

The dataset comprises 2,864 rows and 21 columns, representing data from 179 countries for the years 2000 to 2015. It includes 19 numeric variables - 17 quantitative and 2 categorical - providing a well-rounded foundation for analyzing life expectancy determinants. Note that there are two columns named "developed" and "developing", and we will keep one of them when performing the regression.

Table 1: Description of Variables in the Dataset

Variable	Description
country	List of the 179 countries
region	179 countries are distributed in 9 regions. E.g. Africa, Asia, Oceania, European Union, Rest of Europe and etc.
year	Years observed from 2000 to 2015
infant_deaths	Represents infant deaths per 1000 population
under_five_deaths	Represents deaths of children under five years old per 1000 population
adult_mortality	Represents deaths of adults per 1000 population
alcohol	Represents alcohol consumption that is recorded in liters of pure alcohol per capita with 15+ years old
hepatitis_B	Represents % of coverage of Hepatitis B (HepB3) immunization among 1-year-olds
measles	Represents % of coverage of Measles containing vaccine first dose (MCV1) immunization among 1-year-olds
bmi	BMI is a measure of nutritional status in adults. It is defined as a person's weight in kilograms divided by the square of height in meters
polio	Represents % of coverage of Polio (Pol3) immunization among 1-year-olds
diphtheria	Represents % of coverage of Diphtheria tetanus toxoid and pertussis (DTP3) immunization among 1-year-olds
hiv	Incidents of HIV per 1000 population aged 15-49
gdp	GDP per capita in current USD
population	Total population in millions
thinness_ten_nineteen_years	Prevalence of thinness among adolescents aged 10-19 years. BMI < -2 standard deviations below the median
thinness_five	

Variable	Description
_nine_years	Prevalence of thinness among children aged 5-9 years. BMI < -2 standard deviations below the median
school	Average years that people aged 25+ spent in formal education
developed	Developed country
developing	Developing county
life	Average life expectancy of both genders in different years from 2000 to 2015

Missing values

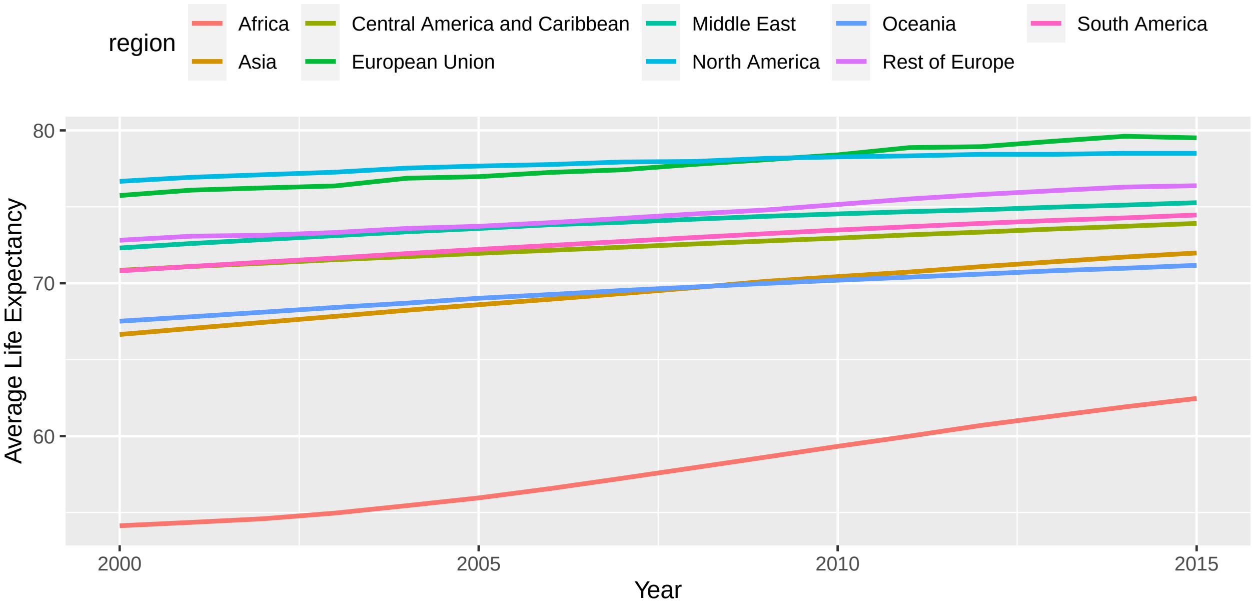
```
##   country region year infant_deaths under_five_deaths adult_mortality alcohol
## 1       0      0    0             0                 0               0       0
##   hepatitis_B measles bmi polio diphtheria hiv gdp population
## 1           0       0   0     0          0   0   0          0
##   thinness_ten_nineteen_years thinness_five_nine_years school developed
## 1                           0                        0      0         0
##   developing life
## 1          0    0
```

Great! There are no missing values in this dataset.

Exploratory Data Analysis

In this part, we will do several visualizations to help us understand the data better.

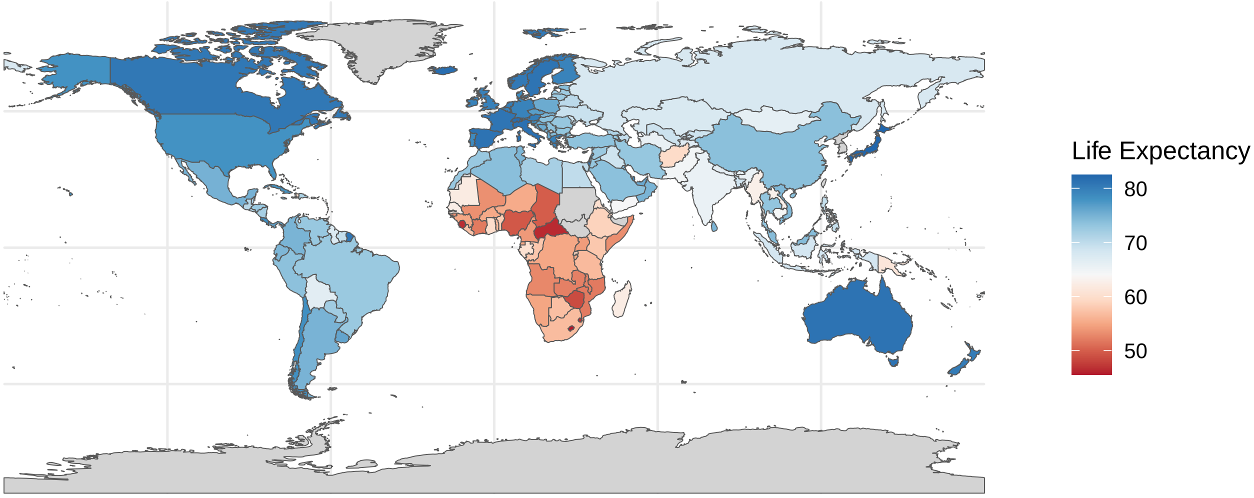
Average Life Expectancy by Region Over Time



This line graph illustrates the trend in average life expectancy by region from the year 2000 to 2015. It reveals that all regions had an increase in life expectancy over the given period, with Africa showing a remarkable improvement despite starting from the lowest baseline. Developed regions, such as North America and the European Union, exhibit high life expectancies that appear to plateau, possibly indicating a saturation point where further increases are hard to achieve. The variations between regions suggest a complex interaction of factors affecting life expectancy, including medical care, social policies, and economic conditions.

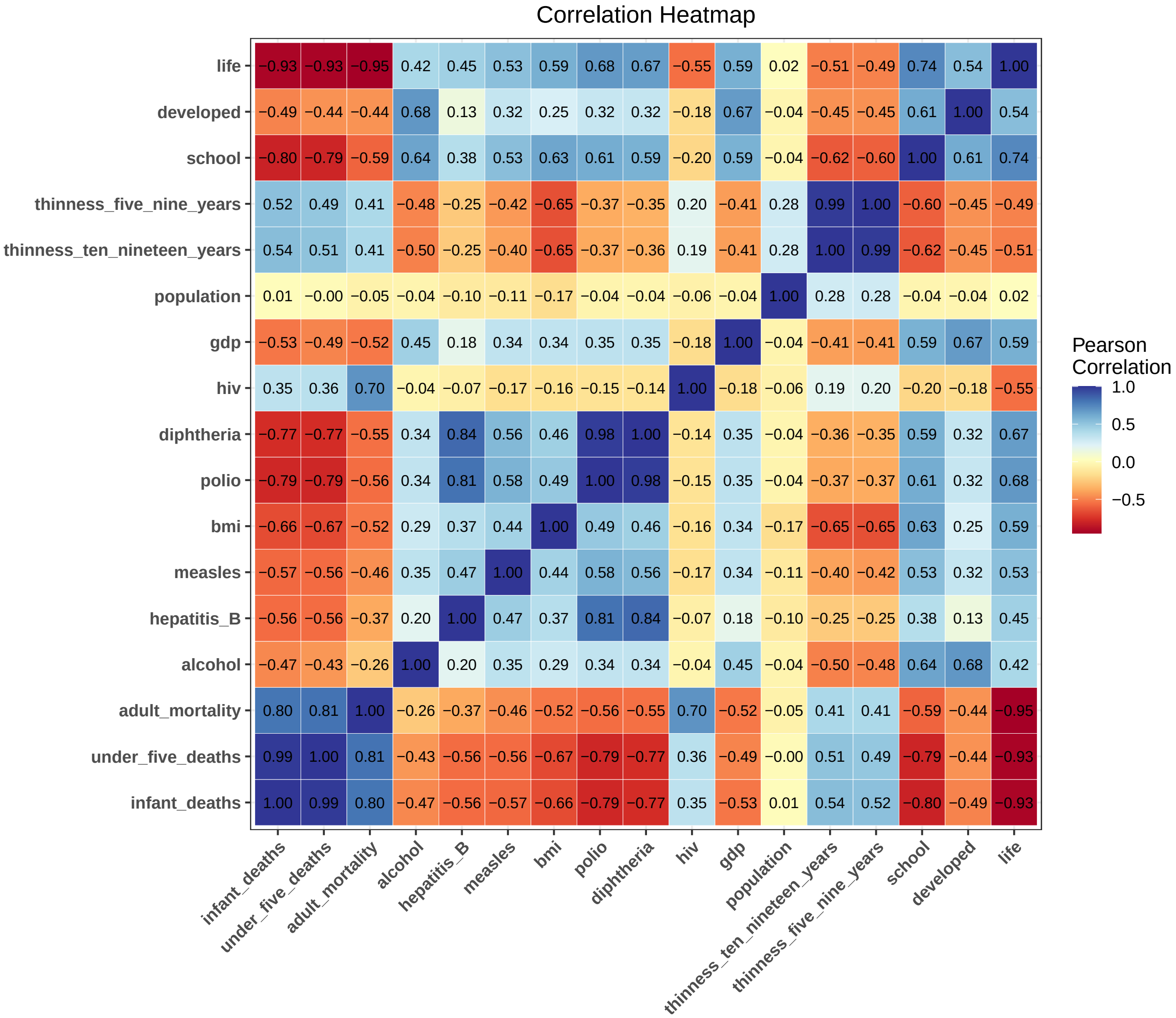
At this point, we encounter a specific challenge: the data for each country over different years forms a time series with **autocorrelation**, conflicting with the assumption of independence between observations in a linear model. To address this issue, we decide to **average the data across these years for each country**.

World Map with Life Expectancy



The map illustrates stark contrasts in life expectancy across the world. High life expectancy in North America, Europe, and Australia likely reflects robust healthcare, higher living standards, and effective public health policies. In contrast, Africa's lower life expectancy may point to healthcare access challenges, prevalent diseases, and social and economic issues. This visualization highlights the uneven distribution of life expectancy across different countries, revealing potential disparities in wealth, medical resources, nutrition, and governance.

Correlation Heatmap



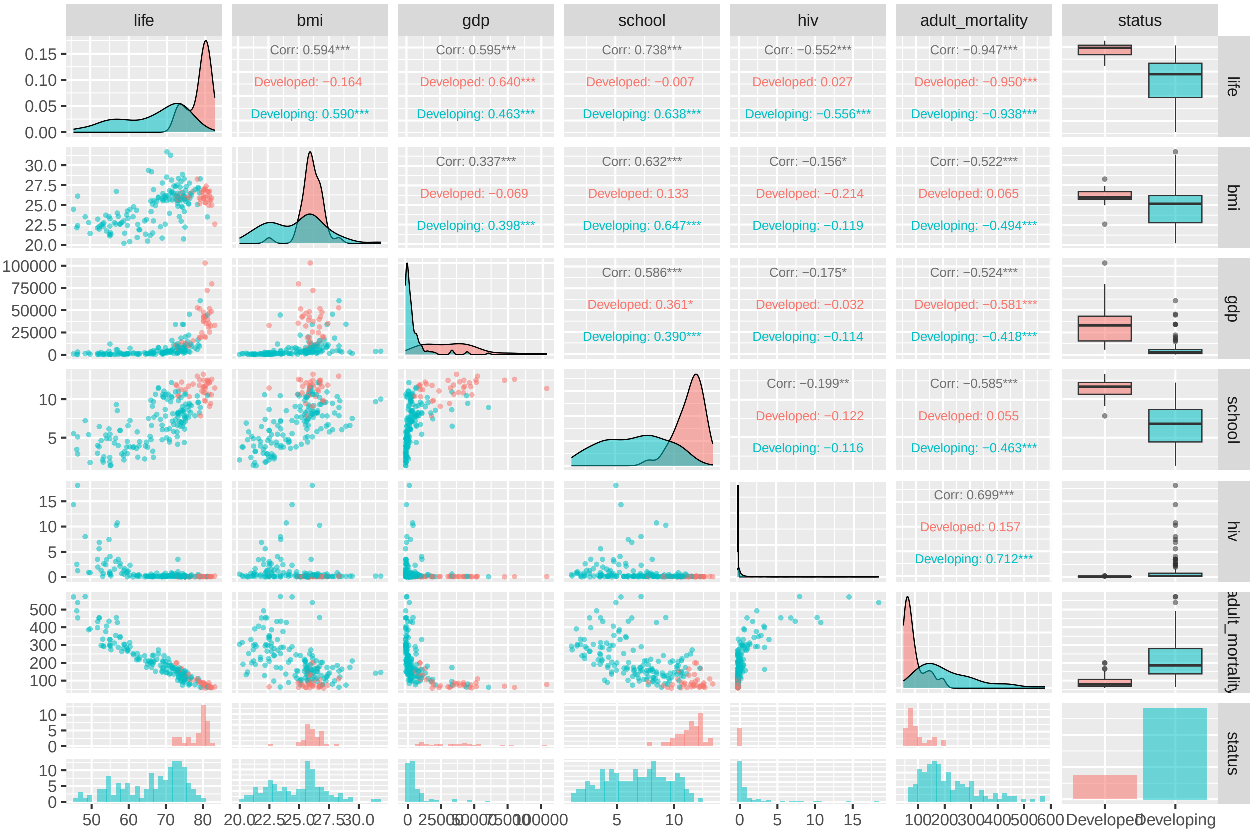
From the correlation heatmap, we can observe several key factors that potentially influence life expectancy. A strong negative correlation is present with variables such as *adult_mortality*, *infant_deaths*, *under_five_deaths*. These correlations are intuitive as higher mortality rates directly indicate a shorter average lifespan. High infant and under-five death rates often reflect poor health conditions, inadequate healthcare services, and a higher prevalence of infectious diseases, all of which can decrease life expectancy.

In contrast, *school*, *gdp*, *polio* and *diphtheria* exhibit a strong positive correlation with life expectancy. The high vaccination rate can prevent some diseases well, which is directly reflected in the increase of life expectancy. Increased schooling suggests better education and potentially better access to healthcare and health-related information. Higher GDP often reflects a country's wealth, which may afford its citizens better healthcare services, nutrition, and living conditions, contributing to a longer lifespan.

The heatmap also reveals instances of **multicollinearity**. There is a positive correlation between *school* and *gdp*, suggesting that higher levels of education within a population are often found in countries with higher GDP. In contrast, there is a negative correlation between *school* and *hiv*. This may indicate that higher levels of education lead to greater awareness of HIV and understanding of how to prevent it. Educated individuals might have better access to information on safe practices, which can reduce the rate of HIV transmission. This observation suggests that these factors may lead to larger variance on the estimated coefficients in the model. We will employ some techniques to reduce dimensionality and apply regularization to alleviate the effects of multicollinearity, thereby enhancing the model's accuracy and interpretability. But here, we can observe that the correlation between *infant_deaths* and *under_five_deaths* is so high (0.99), and they represent nearly the same thing. So we decide to **remove** *under_five_deaths*.

Pair plot

Since there are too many variables, it's hard to plot all of them. Let's pick some variables that we are interested in.



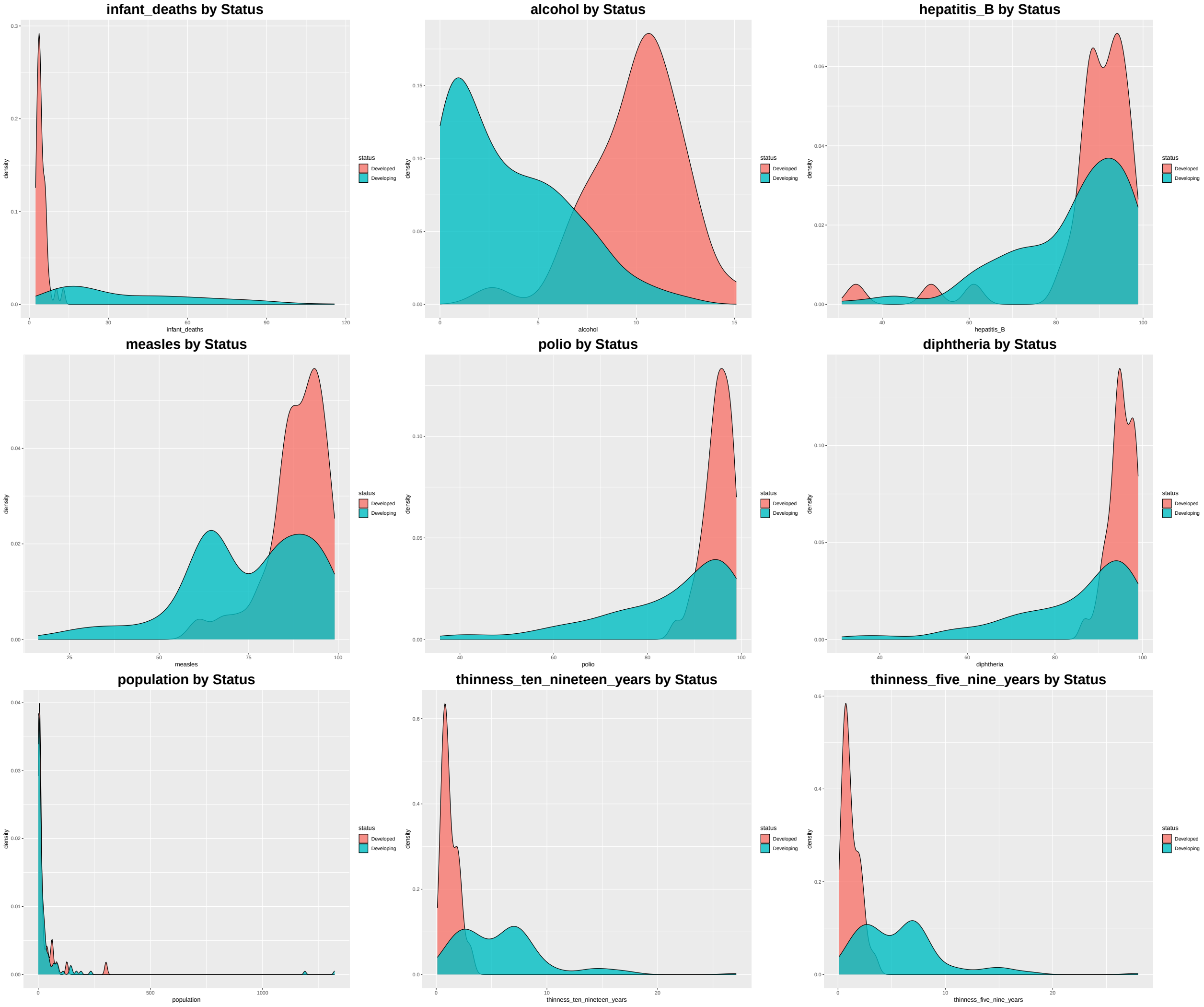
The pair plot reveals significant correlations among various explanatory variables, supporting the relationships previously discussed, such as the positive correlation between *gdp* and *life_expectancy*, and the negative correlation between *hiv* and *schooling*.

Focusing on the distributions, we can observe that life expectancy is right-skewed for developing countries, indicating that while many individuals have lower life expectancies, there is a tail of the population that lives longer (may be China). For developed countries, the distribution is more symmetric, suggesting a more uniform life expectancy across the population. We can also see the *school* distributions are quite different for developed and developing countries, implying that developed countries have higher levels of education,

The distribution patterns of *gdp* and *hiv* are predominantly right-skewed, suggesting that a majority of countries fall into the lower GDP and HIV rate categories, with notable exceptions at the higher end of the spectrum. Such skewness in the data may require transformations for more accurate regression analysis. It is also crucial to address outliers in the HIV data judiciously, as they have the potential to distort the results of statistical models.

The boxplots at the end compare developed and developing countries directly across these indicators. Developed countries tend to have higher life expectancy, higher BMI, higher GDP, more schooling, lower HIV prevalence, and lower adult mortality, emphasizing the disparities between developed and developing nations. These differences accentuate the need for policy interventions that are specifically tailored to the distinct socio-economic and health environments of developed and developing countries.

Distribution of the Rest Variables



We can also see some skewness and outliers. Let's handle them in the next part.

Data Preprocessing

Transforming Skewness

The following statistic can be used to determine if we need to transform the data:

$$S = \frac{U - M}{M - L}$$

This uses the fact that in a symmetric distribution, upper and lower quantiles are equidistant from the median.
For each numeric column in the dataset, with the exception of "year," "developed," and "developing," we will implement the Box-Cox transformation. The Box-Cox transformation is defined as:

$$X^{(p)} = \begin{cases} \frac{X^p - 1}{p} & p \neq 0 \\ \ln(X) & p = 0 \end{cases}$$

For columns where $S > 1$, we will explore transformation parameters $p \in \{0.5, 0, -0.5, -1, -1.5, -2\}$. Our goal is to identify the parameter that adjusts S closest to 1. Conversely, for columns with $S < 1$, we will evaluate parameters $p \in \{1.5, 2, 2.5, 3, 3.5, 4\}$ again aiming to achieve S as close to 1 as possible. This approach ensures a more normalized distribution in our dataset, enhancing the robustness and reliability of subsequent analyses.

Original S before transformation:

```
##               infant_deaths             adult_mortality
##                   2.6398104                   1.4479532
##                     alcohol                 hepatitis_B
##                   1.2567816                   0.6517572
##                     measles                         bmi
##                   0.4868421                   0.3195876
##                       polio                  diphtheria
##                   0.3113456                   0.3305556
##                         hiv                         gdp
##                   4.1703704                   2.5512442
##                  population thinness_ten_nineteen_years
##                   2.7167380                   2.0052083
##    thinness_five_nine_years                      school
##                   1.6792453                   0.8786566
##                        life
##                   0.3728353
```

Results after transformation:

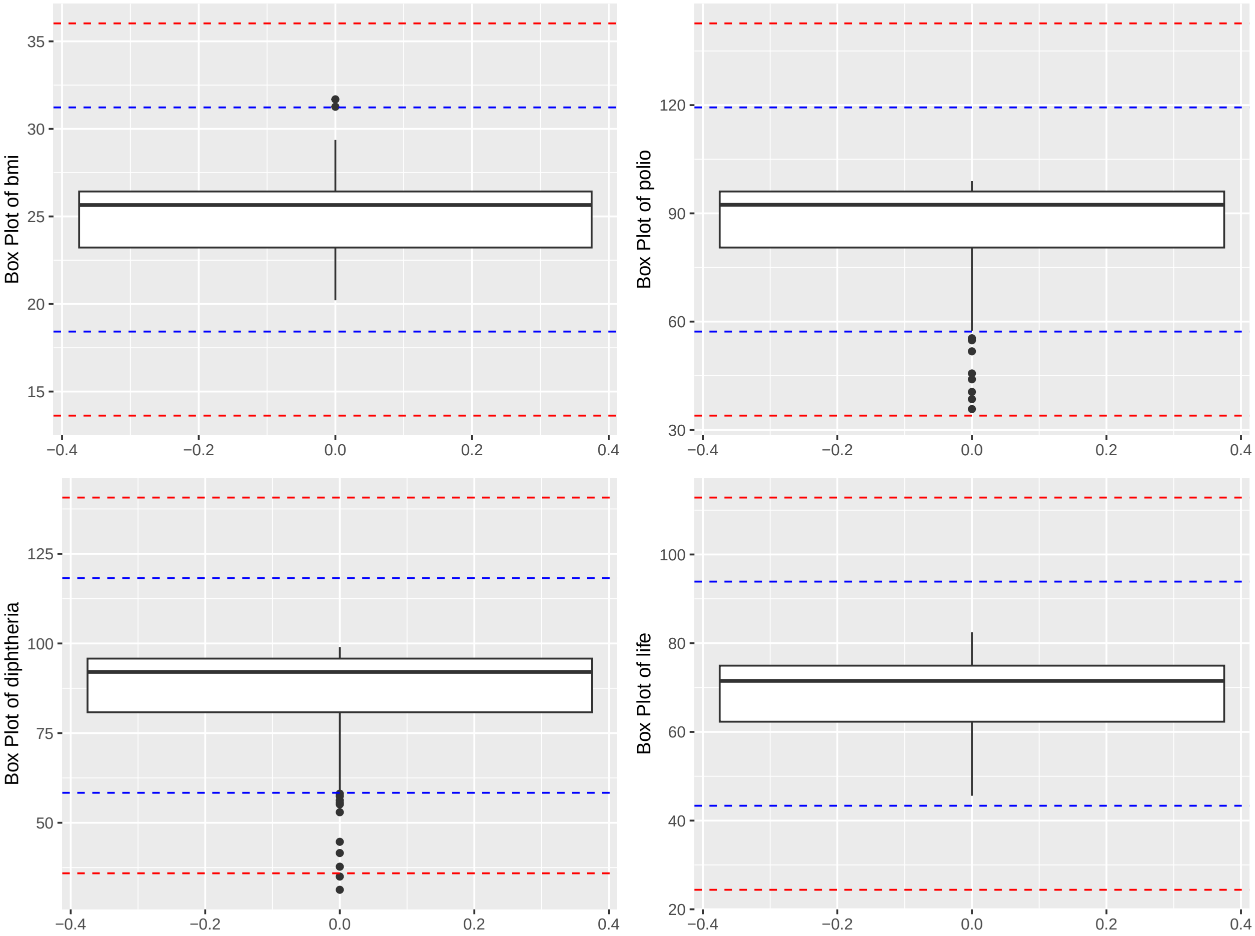
```
##                          Column P_Value   Final_S
## 1                 infant_deaths     0.0 1.0724849
## 2               adult_mortality     0.0 0.9527580
## 3                       alcohol     0.5 0.8286179
## 4                   hepatitis_B     4.0 0.8589772
## 5                       measles     4.0 0.8136292
## 6                           bmi     4.0 0.3857206
## 7                         polio     4.0 0.4012981
## 8                    diphtheria     4.0 0.4228843
## 9                           hiv    -0.5 1.0022944
## 10                          gdp     0.0 0.8833444
## 11                   population     0.0 0.8428431
## 12 thinness_ten_nineteen_years     0.0 0.9929600
## 13    thinness_five_nine_years     0.0 0.8377794
## 14                       school     1.5 1.0499465
## 15                         life     4.0 0.4867205
```

We noted that for right-skewed columns ($S < 1$), $p = 0$ is generally effective, with the exception of the *hiv* column which requires $p = -0.5$. For simplicity, we will apply $p = 0$ to all right-skewed columns.

However, we encountered challenges with several left-skewed columns such as *bmi*, *polio*, *diphtheria*, and *life*. Despite applying larger transformation parameters, S of these columns did not approach 1 as intended. This deviation might be attributed to the presence of outliers in these data sets. Our next step take a close look to these outliers to ensure more accurate and representative data analysis.

Univariate Outliers

We first check the box-plot of these columns in the dataset before the transformation: *bmi*, *polio*, *diphtheria*, and *life*.



The univariate outliers observed across the datasets indicate significant deviations. Their impact is considerable to potentially distort the overall analysis. However, we must be cautious in removing any outliers, as they may not be regression outliers and could hold special significance. For instance, outliers in 'bmi' might represent some developed countries, while outliers in other variables could represent some poorer nations. Therefore, we merely acknowledge that these points **might** affect the regression results, but we do not remove them. We will review them again during regression diagnostics.

Standardizing the data

Standardizing ensures that all features contribute equally to the model, preventing variables with larger scales from disproportionately influencing the results. It also improves the **interpretability** of coefficients, as they can be compared on the same scale. Additionally, standardizing is crucial for PCA to work well later.

Now, we can perform the MLR. Let's take a look.

```
##
## Call:
## lm(formula = life ~ ., data = data_clean_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58193 -0.08767 -0.00071  0.09376  0.52496
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.037965   0.017452   2.175 0.031038 *
## infant_deaths               -0.328354   0.046827  -7.012 5.93e-11 ***
## adult_mortality             -0.608691   0.036444 -16.702  < 2e-16 ***
## alcohol                      0.079413   0.021196   3.747 0.000248 ***
## hepatitis_B                 -0.033910   0.027297  -1.242 0.215928
## measles                     -0.048013   0.019873  -2.416 0.016797 *
## bmi                          0.008474   0.023281   0.364 0.716329
## polio                       -0.037685   0.084329  -0.447 0.655553
## diphtheria                   0.128314   0.085753   1.496 0.136503
## hiv                         -0.019689   0.022141  -0.889 0.375181
## gdp                          0.039622   0.032984   1.201 0.231397
## population                   0.036001   0.014484   2.486 0.013945 *
## thinness_ten_nineteen_years -0.030059   0.076481  -0.393 0.694815
## thinness_five_nine_years     0.006434   0.076891   0.084 0.933412
## school                       0.017132   0.028463   0.602 0.548076
## developed                   -0.183669   0.057849  -3.175 0.001792 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1701 on 163 degrees of freedom
## Multiple R-squared:  0.9735, Adjusted R-squared:  0.9711
## F-statistic: 399.4 on 15 and 163 DF,  p-value: < 2.2e-16
```

Model Selection

Main Effect Variable selection

First, we carry out the main effect variable selection using `regsubset()` function.

```
##   (Intercept) infant_deaths adult_mortality alcohol hepatitis_B measles   bmi
## 1        TRUE         FALSE            TRUE   FALSE       FALSE   FALSE FALSE
## 2        TRUE          TRUE            TRUE   FALSE       FALSE   FALSE FALSE
## 3        TRUE          TRUE            TRUE    TRUE       FALSE   FALSE FALSE
## 4        TRUE          TRUE            TRUE    TRUE       FALSE   FALSE FALSE
## 5        TRUE          TRUE            TRUE    TRUE       FALSE   FALSE FALSE
## 6        TRUE          TRUE            TRUE    TRUE       FALSE   FALSE FALSE
## 7        TRUE          TRUE            TRUE    TRUE       FALSE    TRUE FALSE
## 8        TRUE          TRUE            TRUE    TRUE       FALSE    TRUE FALSE
##   polio diphtheria   hiv   gdp population thinness_ten_nineteen_years
## 1 FALSE      FALSE FALSE FALSE      FALSE                       FALSE
## 2 FALSE      FALSE FALSE FALSE      FALSE                       FALSE
## 3 FALSE      FALSE FALSE FALSE      FALSE                       FALSE
## 4 FALSE      FALSE FALSE FALSE      FALSE                       FALSE
## 5 FALSE      FALSE FALSE FALSE       TRUE                       FALSE
## 6 FALSE       TRUE FALSE FALSE       TRUE                       FALSE
## 7 FALSE       TRUE FALSE FALSE       TRUE                       FALSE
## 8 FALSE       TRUE FALSE FALSE       TRUE                        TRUE
##   thinness_five_nine_years school developed
## 1                    FALSE  FALSE     FALSE
## 2                    FALSE  FALSE     FALSE
## 3                    FALSE  FALSE     FALSE
## 4                    FALSE  FALSE      TRUE
## 5                    FALSE  FALSE      TRUE
## 6                    FALSE  FALSE      TRUE
## 7                    FALSE  FALSE      TRUE
## 8                    FALSE  FALSE      TRUE
```

Then, we calculate the AIC and BIC for all these 8 models for reference and choose the best one among them.

```
## Best model by AIC includes:
##  infant_deaths, adult_mortality, alcohol, measles, diphtheria, population, thinness_ten_nineteen_yea
```

```
## Best model by BIC includes:
##  infant_deaths, adult_mortality, alcohol, measles, diphtheria, population, developed
```

```
## [1] "The best model from AIC metric is: 8"
```

```
## [1] "The best model from BIC metric is: 7"
```

```
## [1] "The best model's AIC value is: -116.819270272969"
```

```
## [1] "The best model's BIC value is: -86.6059608645577"
```

However, the `regsubset()` function inherently possesses computational complexity constraints. As a result, this method limits the variable combinations to a maximum of eight, which could potentially be insufficient for our comprehensive variable selection needs. The limitation in the scope of variable combination could lead to overlooking some vital variables that might be crucial for the model.

Considering a critical factor in our decision was the performance of the models in terms of their AIC and BIC values. So that we carry out an exhaustive research using AIC and BIC as evaluation and select the best variable set. Conceptually, it will serve as a more robust and effective approach for our variable selection process.

```
## Best model by AIC includes:
##  infant_deaths, adult_mortality, alcohol, measles, diphtheria, gdp, population, thinness_ten_nineteen
```

```
## The best model's AIC value is: -116.8559
```

```
## Best model by BIC includes:
##  infant_deaths, adult_mortality, alcohol, measles, diphtheria, population, thinness_ten_nineteen_year
```

```
## The best model's BIC value is: -84.94541
```

In conclusion, the model selected by AIC has 9 explanatory variables, namely infant_deaths, adult_mortality, alcohol, measles, diphtheria, gdp, population, thinness_ten_nineteen_years and the dummy variable developed. The model selected by BIC drop `gdp` and `thinness_ten_nineteen_years` due to larger penalty on variable numbers. To decide which model to choose as our final model, we do F test on `gdp` and ‘`thinness_ten_nineteen_years`, checking whether they are significant.

```
## Analysis of Variance Table
##
## Model 1: life ~ infant_deaths + adult_mortality + alcohol + measles +
##     diphtheria + gdp + population + developed
## Model 2: life ~ infant_deaths + adult_mortality + alcohol + measles +
##     diphtheria + gdp + population + thinness_ten_nineteen_years +
##     developed
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    170 4.9253
## 2    169 4.8248  1   0.10048 3.5196 0.06237 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: life ~ infant_deaths + adult_mortality + alcohol + measles +
##     diphtheria + population + thinness_ten_nineteen_years + developed
## Model 2: life ~ infant_deaths + adult_mortality + alcohol + measles +
##     diphtheria + gdp + population + thinness_ten_nineteen_years +
##     developed
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    170 4.8800
## 2    169 4.8248  1  0.055208 1.9338 0.1662
```

```
## Analysis of Variance Table
##
## Model 1: life ~ infant_deaths + adult_mortality + alcohol + measles +
##     diphtheria + population + developed
## Model 2: life ~ infant_deaths + adult_mortality + alcohol + measles +
##     diphtheria + gdp + population + thinness_ten_nineteen_years +
##     developed
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    171 4.9772
## 2    169 4.8248  2   0.15231 2.6675 0.07235 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-values for H0: $\beta_{gdp} = 0$ and $\beta_{thin} = 0$ are 0.06 and 0.16 repectively. The The p-value for H0: $\beta_{gdp} = \beta_{thin} = 0$ is 0.07, indicating that we can't reject the null hypothesis. So we exclude both `gdp` and `thinness_ten_nineteen_years` in our model. Above all, we choose the variables selected by the BIC searching method with 7 explanatory variables: infant_deaths, adult_mortality, alcohol, measles, diphtheria, population, developed.

Interaction Terms Selection

In this section, we evaluate the significance of various interaction terms derived from the main effect variables identified in earlier analyses. Our methodology does not incorporate forward or backward selection techniques; instead, we aim to directly assess the explanatory power of each interaction term with respect to the response variable. We select the interaction terms with its p-values in the ANOVA, using the threshold of 0.05. We select *adult_mortality:developed* and *alcohol:developed* as our additional interaction variables. This approach is selected over **backward selection** to avoid the potential instability and model volatility often associated with **one-step** selection methods. By focusing on the interaction terms individually, we intend to determine their individual contributions to the model's predictive capacity, ensuring a robust and theoretically sound model construction and not nelegcting any important explanatary variable.(Check the Rmd file for exact F-test output)

Interpretation for the Interaction Terms

- **Adult Mortality and Development Status:**

The interaction between adult mortality and development status is significant, with a p-value of 0.04618. This suggests that the effect of adult mortality on life expectancy is significantly different between developed and developing countries, with a strong indication that reducing adult mortality could have a different effect on increasing life expectancy in one group compared to the other. Which is intuitive since the baseline of *adult_mortality* between developing and developed countries aren't the same.

- **Alcohol and Development Status:**

The interaction is also significant for alcohol and development status, with a p-value of 0.01878. This highlights the varying impact of alcohol on life expectancy between developed and developing countries, implying that having access to alcohol has different meaning in different countries. Its effect on *life* may be shown by different ways in developed and developing countries

- **Non-Significant Interaction Terms:**

The interactions involving *infant_deaths*, *measles*, *diphtheria* and *population* with development status, though not statistically insignificant, have higher p-values compared to the aforementioned factors. This implies that these factors impact life expectancy more uniformly across different development statuses and might not need as differentiated an approach as adult mortality or under-five deaths.

- **Implications for Public Health Policy:**

The analysis implies that public health interventions should not adopt a one-size-fits-all strategy. Instead, they should be tailored to address the specific challenges faced by developed and developing countries. For instance:

1. In developing countries, a focused strategy on reducing adult mortality and decreasing the malnutrition youth's proportion could lead to significant improvements in life expectancy.
2. The significant interaction with alcohol suggests that developing countries may need to prioritize alcohol access in life. expectancy.
3. Factors like population infectious diseases may require global health policies that are effective across different developmental contexts, as their impact seems to be more uniform across countries.

Regression Diagnostics

In this part, we first regress on our selected main variables to test its performance and set as a benchmark for following refined models.

```
##
## Call:
## lm(formula = best_main_formula, data = data_clean_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56840 -0.09958  0.00013  0.09583  0.51623
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.03291    0.01662   1.981  0.04920 *
## adult_mortality -0.64071    0.02867 -22.345  < 2e-16 ***
## infant_deaths   -0.37404    0.04015  -9.316  < 2e-16 ***
## alcohol          0.09041    0.01914   4.723 4.81e-06 ***
## measles         -0.04659    0.01825  -2.553  0.01157 *
## diphtheria       0.05582    0.02009   2.778  0.00608 **
## population       0.03402    0.01321   2.575  0.01088 *
## developed       -0.15923    0.05153  -3.090  0.00234 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1706 on 171 degrees of freedom
## Multiple R-squared:  0.972,  Adjusted R-squared:  0.9709
## F-statistic: 849.2 on 7 and 171 DF,  p-value: < 2.2e-16
```

Examination of the coefficients' p-values reveals that all but one variable demonstrate statistical significance within this model. Notably, compared to our initial model which included the entire set of variables, the current model exhibits substantially enhanced interpretability. This improvement significantly augments our capacity to conduct a detailed analysis of the dataset, thereby allowing for the extraction of more valuable insights.

Next, we incorporate the interaction terms into the model through the **one-step-backward selection** process and check its performance.

```
##
## Call:
## lm(formula = best_formula, data = data_clean_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56507 -0.09454  0.01030  0.09497  0.49791
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.03736    0.01654   2.259  0.02519 *
## adult_mortality           -0.65268    0.03010 -21.684  < 2e-16 ***
## infant_deaths             -0.37483    0.03974  -9.432  < 2e-16 ***
## alcohol                    0.09914    0.01927   5.145 7.35e-07 ***
## measles                   -0.05699    0.01878  -3.035  0.00278 **
## diphtheria                 0.05486    0.01993   2.753  0.00655 **
## population                 0.03419    0.01307   2.617  0.00967 **
## developed                  0.15796    0.13914   1.135  0.25785
## alcohol:developed         -0.18612    0.08269  -2.251  0.02569 *
## adult_mortality:developed  0.10954    0.05710   1.919  0.05673 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1686 on 169 degrees of freedom
## Multiple R-squared:  0.973,  Adjusted R-squared:  0.9716
## F-statistic: 676.6 on 9 and 169 DF,  p-value: < 2.2e-16
```

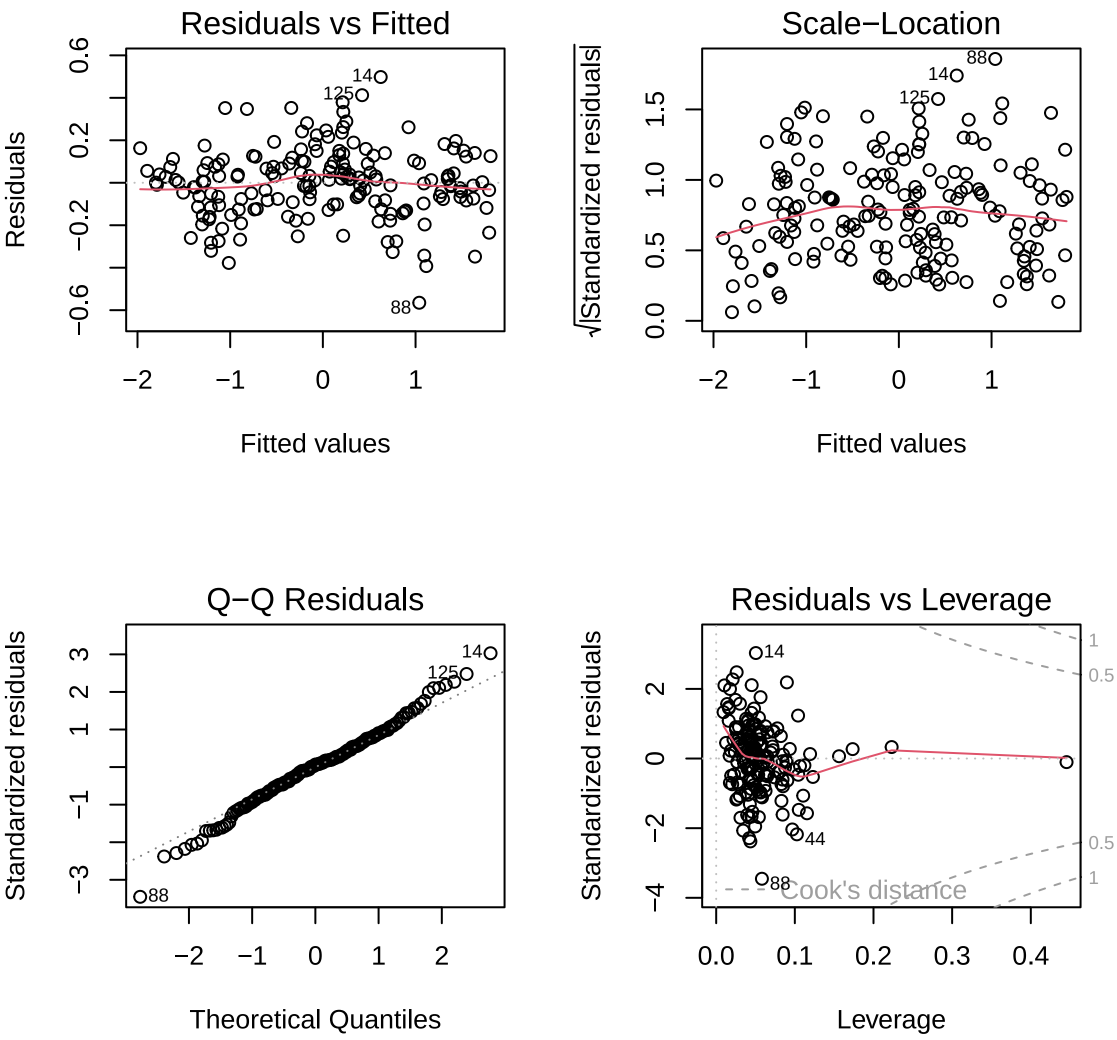
The coefficients obtained align closely with our initial assumptions and the results of our selection criteria. Notably, the p-value associated with the variable *developed* is relatively high, suggesting it does not hold statistical significance in this particular model.

However, its inclusion remains justified based on the **Principle of Marginality**, which stipulates that main effects should be retained in a model when interaction terms are present. Furthermore, the apparent insignificance of *developed* after the inclusion of interaction terms may be attributed to the interaction terms themselves encapsulating the main effect of *developed* on life expectancy. This observation potentially highlights the differential impact of various variables on life expectancy between developing and developed countries, underscoring the nuanced nature of these relationships within the model's framework.

Furthermore, the model summary highlights several key findings:

- The coefficients for *adult_mortality*, *infant_deaths*, *alcohol*, *measles*, *diphtheria*, and *population* are statistically significant, with *adult_mortality* showing the most substantial negative impact on life expectancy.
- Interaction terms *alcohol:developed* and *adult_mortality:developed* are significant, suggesting that the effect of *alcohol* and *adult_mortality* on life expectancy differs between developed and developing countries.
- The overall model has a high Multiple R-squared value, indicating that it accounts for a large proportion of the variability in life expectancy. And the model's F-statistic is highly significant, which provides strong evidence that the model is a good fit for the data.

Then we generate some diagnostic plots:



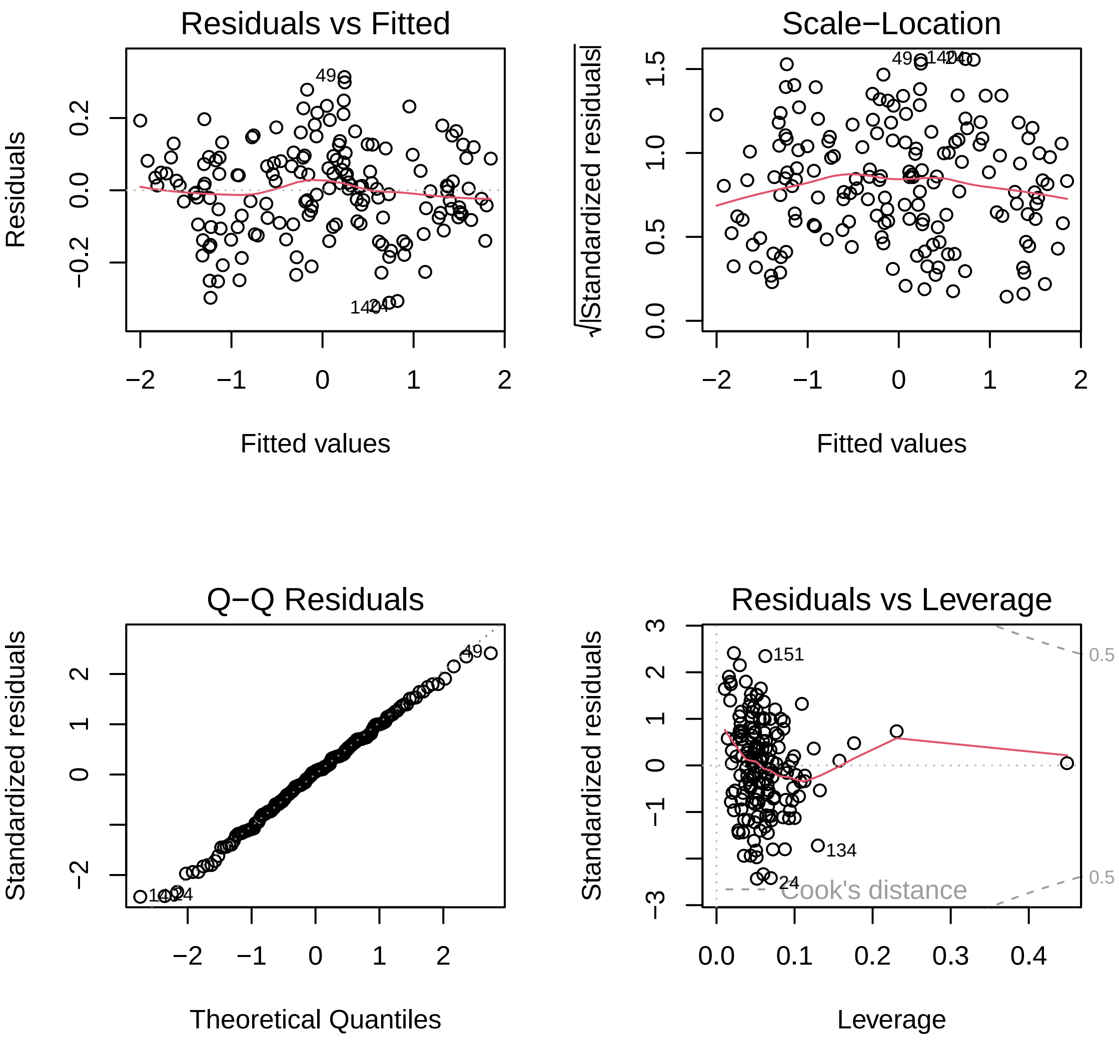
From the plots we can conclude that:

The uniformly distributed residuals can ensure the validity of the linear model assumptions. Since there is no apparent pattern, which suggests that the variance of the residuals is constant, and the linearity assumption is reasonable. And the Scale-Location plot shows that residuals spread equally along the ranges of predictors. This is indicative of homoscedasticity which aligns with the residual plot's result.

Furthermore, the Normal Q-Q plot displays how well the residuals match a normal distribution. The points largely follow the reference line, suggesting that the residuals are normally distributed, but there are still some deviations in the tails, which could be outliers and need further treatment.

Outlier elimination

In this part, we focus on eliminating the **regression outliers** from the dataset. The rubric we are using here is the standard residuals and Cook's distance. Then we refit the model and start further diagnostics.



```
##
## Call:
## lm(formula = formula(best_model), data = data_clean)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.311412 -0.087009  0.007616  0.088022  0.313538
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.05060    0.01417   3.571 0.000474 ***
## adult_mortality           -0.65316    0.02508 -26.043  < 2e-16 ***
## infant_deaths             -0.39909    0.03380 -11.808  < 2e-16 ***
## alcohol                    0.09370    0.01646   5.694 6.12e-08 ***
## measles                   -0.05343    0.01550  -3.448 0.000728 ***
## diphtheria                 0.05541    0.01624   3.412 0.000823 ***
## population                 0.04405    0.01090   4.041 8.36e-05 ***
## developed                  0.06523    0.11131   0.586 0.558720
## alcohol:developed         -0.15427    0.06589  -2.341 0.020504 *
## adult_mortality:developed  0.09170    0.04633   1.979 0.049548 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1314 on 154 degrees of freedom
## Multiple R-squared:  0.9838, Adjusted R-squared:  0.9829
## F-statistic:  1041 on 9 and 154 DF,  p-value: < 2.2e-16
```

Improvement

The diagnostic plots for the refined model suggest substantive improvements following the elimination of outliers. Specifically:

- The **Residuals vs Fitted** plot exhibits a more randomized dispersion of residuals about the horizontal axis, indicative of enhanced homoscedasticity and adherence to linearity assumptions.
- The **Scale-Location** plot displays a uniform spread of standardized residuals across the range of fitted values, signaling a reduction in heteroscedasticity.
- The **Normal Q-Q** plot demonstrates that residuals more closely follow the theoretical quantiles, suggesting better normality.
- The **Residuals vs Leverage** plot shows that points with high leverage have diminished influence on the model, with Cook's distances remaining within acceptable limits.

Collectively, these diagnostic improvements suggest that the model's assumptions are better satisfied post-outlier removal.

The model's high Multiple R-squared and significant F-statistic post-refinement underscore its improved fit and interpretability, which is much larger than the R-square of the model including all variables, meaning our model has extract the structure and main component from the data. Despite the non-significance of *developed*, the model's stability and predictive power are enhanced, making it a more robust tool for predicting life expectancy.

Final Model

$$\begin{aligned}
\text{LifeExpectance}^4 = & -0.54918 \\
& - 2.61264 \times \log(\text{adult\_mortality}) \\
& - 1.59636 \times \log(\text{infant\_deaths}) \\
& + 0.7496 \times \text{alcohol}^{0.5} \\
& - 0.05343 \times \text{measles}^4 \\
& + 0.05541 \times \text{diphtheria}^4 \\
& + 0.1762 \times \log(\text{population}) \\
& + 1.49508 \times \text{developed} \\
& + 0.3668 \times \log(\text{adult\_mortality}) \times \text{developed} \\
& - 1.23416 \times \text{alcohol}^{0.5} \times \text{developed}
\end{aligned}$$

The variables included in the final regression model have been standardized, which precludes the model's immediate application for predictive purposes in its current form. It should be noted that the primary objective of this project is to examine the interrelations among the variables through regression analysis rather than to emphasize predictive accuracy.

Analysis of Regression Coefficients

- **Adult Mortality:**

The coefficient for the log-transformed adult mortality is negative (-0.63080), indicating that as adult mortality rates increase, life expectancy decreases. The logarithmic transformation suggests this relationship is nonlinear, with diminishing impact as mortality rates rise.

- **Alcohol Consumption:**

Alcohol consumption has a positive coefficient (0.13244) when transformed with a square root, which may reflect a complex relationship where moderate levels of alcohol consumption correlate with higher life expectancy, perhaps due to social and lifestyle factors.

- **Development Status:**

The binary variable for development status shows a significant positive coefficient (0.34341), implying that life expectancy is higher in developed countries.

- **Interaction Terms:**

1. The interaction between the log-transformed adult mortality and development status is positive (0.10206), suggesting that the negative impact of adult mortality on life expectancy may be lessened in developed countries.
2. The interaction term for square-root-transformed alcohol consumption and development status is negative (-0.27348), which could indicate that the positive relationship between alcohol and life expectancy in the general model is not as strong or is reversed in developed countries.

Additional Work

Using `Region` As a Dummy Variable in Regression

In the previous discussion, we calculated the average of various quantitative variables for each country over years as the explanatory variables and employed average life expectancy over years as the response variable. We solely incorporated `developed` as a dummy variable while overlooking the variable `Region`. The reason we excluded `Region` in MLR is that the number of regions is too large and including this dummy variable may result in a large number of variables, leading to unstable estimates of coefficients. Our underlying assumption is that the variations of life expenctency between regions could be accounted for by disparities among countries in other variables.

In this section, we introduce `Region` as an additional dummy variable revisit the earlier discussion. We will compare this model with the best model obtained in the preceding analysis.

Model Selection

We use the same methods in previous work for model selection found the model with the smallest BIC is `lm(life ~ infant_deaths + adult_mortality + population + school + regionCentral America and Caribbean + regionSouth America)` and the smallest AIC model is `lm(life ~ infant_deaths + adult_mortality + population + measles +  diphtheria + school + isCentralAmerica + isSouthAmerica)`. We use similar F-test method to determine whether we drop `measles` and `diphtheria` in our model.

The results show that we can reject all the three null hypothesis with the p value of 0.02028, 0.01243 and 0.0173 respectively. So our selected model is `lm(life ~ infant_deaths + adult_mortality + population + measles +  diphtheria_school + isCentralAmerica + isSouthAmerica)`.

Interaction selection

We evaluate the significance of various interaction terms and choose the only significant term: `adult_mortality:isCentralAmerica`, with a p value of 0.002349.

Regression Diagnostics

With this final model, we get the model of `life ~ infant_deaths + adult_mortality + population + measles +  diphtheria + school + isCentralAmerica + isSouthAmerica + adult_mortality:isCentralAmerica`. We do model diagnose as before and remove outliers.

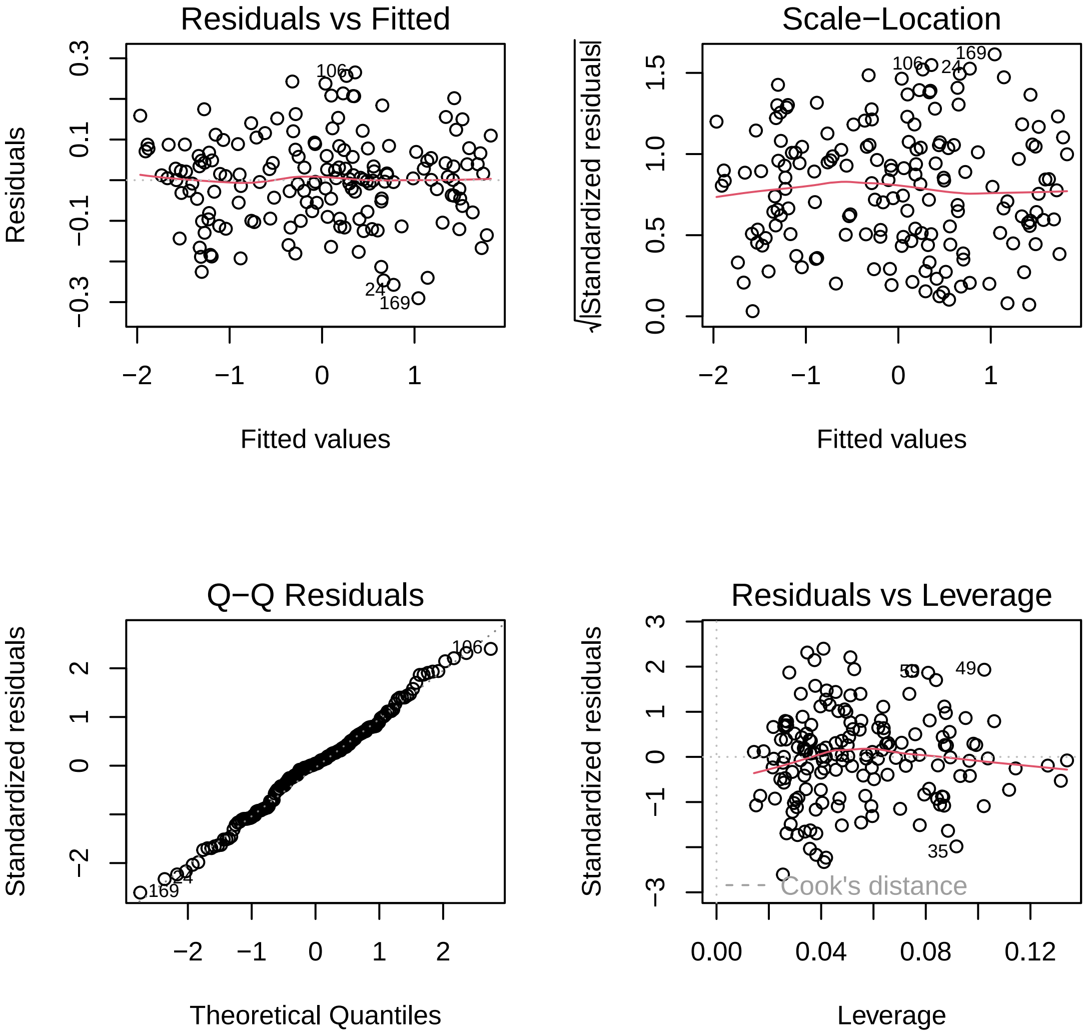
The result of the model on the cleaned dataset is as below.

```
##
## Call:
## lm(formula = formula(best_model_region), data = data_clean_region)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.288866 -0.075947  0.004192  0.069528  0.265392
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -0.050314   0.009645  -5.216 5.72e-07 ***
## infant_deaths                   -0.349403   0.025880 -13.501  < 2e-16 ***
## adult_mortality                 -0.612043   0.017539 -34.897  < 2e-16 ***
## population                       0.046897   0.009456   4.959 1.83e-06 ***
## measles                         -0.042158   0.012989  -3.246  0.00143 **
## diphtheria                       0.043394   0.013961   3.108  0.00224 **
## school                           0.076875   0.016996   4.523 1.20e-05 ***
## isCentralAmerica                 0.340774   0.033160  10.277  < 2e-16 ***
## isSouthAmerica                   0.225410   0.035544   6.342 2.34e-09 ***
## adult_mortality:isCentralAmerica -0.130183   0.088925  -1.464  0.14522
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1125 on 156 degrees of freedom
## Multiple R-squared:  0.9884, Adjusted R-squared:  0.9877
## F-statistic:  1476 on 9 and 156 DF,  p-value: < 2.2e-16
```

The interaction term become not significant when we fit the model on the cleaned dataset, indicating that the previous low p value is probably caused by outliers. So we exclude the interaction term in our regression and get the final formula: `life ~ infant_deaths + adult_mortality + population + measles +  diphtheria + school + isCentralAmerica + isSouthAmerica`.

The results for this model are as below:

```
##
## Call:
## lm(formula = life ~ infant_deaths + adult_mortality + population +
##     measles + diphtheria + school + isCentralAmerica + isSouthAmerica,
##     data = data_clean_region)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.290538 -0.077304  0.004576  0.069124  0.265184
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.050320   0.009680  -5.198 6.19e-07 ***
## infant_deaths    -0.349429   0.025974 -13.453  < 2e-16 ***
## adult_mortality  -0.614192   0.017541 -35.015  < 2e-16 ***
## population        0.046873   0.009491   4.939 1.99e-06 ***
## measles          -0.042356   0.013035  -3.249  0.00142 **
## diphtheria        0.042304   0.013992   3.023  0.00292 **
## school            0.076602   0.017057   4.491 1.37e-05 ***
## isCentralAmerica  0.349402   0.032751  10.669  < 2e-16 ***
## isSouthAmerica    0.225231   0.035673   6.314 2.67e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1129 on 157 degrees of freedom
## Multiple R-squared:  0.9882, Adjusted R-squared:  0.9876
## F-statistic:  1648 on 8 and 157 DF,  p-value: < 2.2e-16
```



Model Comparison and Analysis

Final Model

$$
\begin{aligned}
\text{LifeExpectance}^4 = & -0.4055 \\
& - 2.456768 \times \log(\text{adult\_mortality}) \\
& - 1.397716 \times \log(\text{infant\_deaths}) \\
& - 0.042356 \times \text{measles}^4 \\
& + 0.042304 \times \text{diphtheria}^4 \\
& + 0.187492 \times \log(\text{population}) \\
& + 0.204272 \times \text{school}^{1.5} \\
& + 1.397608 \times \text{isCentralAmerica} \\
& + 0.900924 \times \text{isSouthAmerica}
\end{aligned}
$$

Selected Model Variables Comparison We compare the model with region information with our previous model and find that apart from the region dummy variable, the two model lay emphasis on different explanatory variables. The new model excluded *alcohol* and the dummy variable *developed*, including *school* and regional dummy variables instead. The exclusion of the `developed` variable may be explain by the introduction of regional dummy variables. It indicates that regional factor can better explain the variance between different countries than the developed status. Also, the countries within the same region tend to share the same developed or developing status, alcohol consumption, so there may be no need to include those variables again when we use regional dummy variables.

Coefficient Comparison and Interpretation The estimate for the coefficient of the variables are quite similar comparing the two model. The new model chooses two regional dummy variables: *isCentralAmerica* and *isSouthAmerica*, indicating whether the country is in central America and Caribbean or in South America. It's surprising that the model didn't include indicators for the country located in Africa (where the life expectancy is lowest) or North America (where the life expectancy is highest). The model shows that holding all the other variables constant, the country in Central America have 0.349 higher average life expectancy and the South American countries witness a 0.225 rise. These rise may be caused by variables outside the data.

Performance Comparison Comparing the two model, the model with regional information has a slightly higher adjusted R square value (0.9876 vs 0.9829) with less explanatory variables. Also in the new model, we can see that all the variables are highly significant, proving that the model is strongly convincing. However, the model didn't include the information of development status and its interaction with alcohol assumption and adult mortality. Since the objective of this project is to examine the interrelations among the variables through regression analysis and to get insight for the influence of variables on life expectancy, we may lose those precious information. However, if the reader would like to choose a model for more accurate prediction performance, we suggest using the model with regional information since it has a higher adjusted R squared value.

Dimensionality reduction with PCR

Another way to avoid multicollinearity and mitigate overfitting is to instead use principal components regression, which finds M linear combinations (known as "principal components") of the original p predictors and then uses least squares to fit a linear regression model using the principal components as predictors. We set `scale=FALSE` since we've already scaled the data. We use k-fold cross-validation to evaluate the performance of the model. (k = 10 by default)

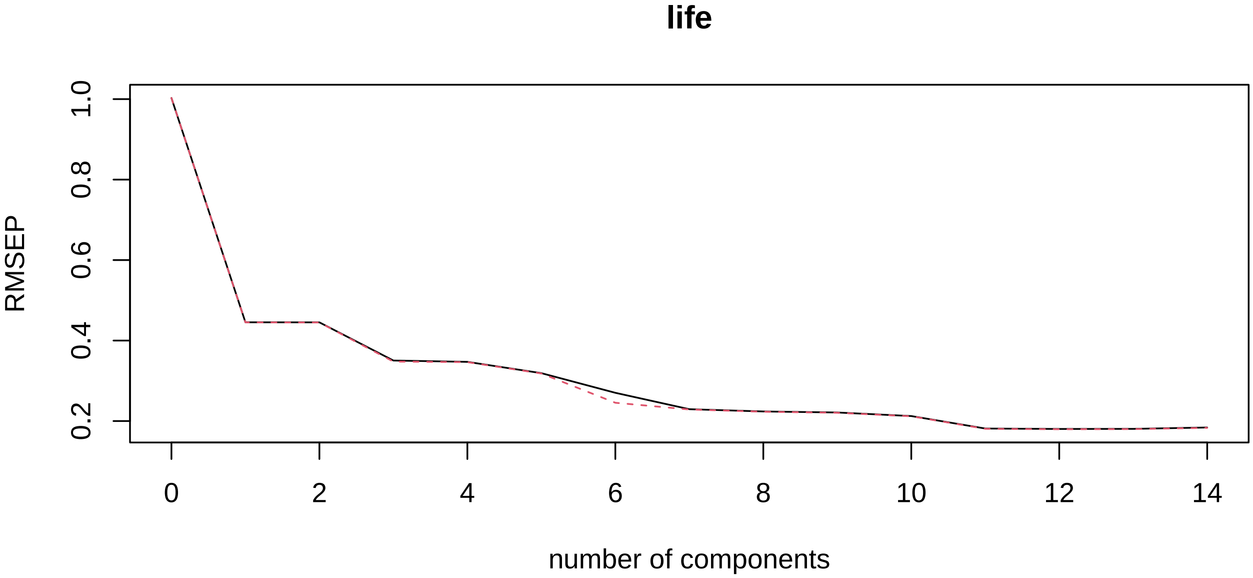
```
## Data:    X dimension: 179 14
##  Y dimension: 179 1
## Fit method: svdpc
## Number of components considered: 14
```

```
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           1.003   0.4454   0.4452   0.3504   0.3471   0.3190   0.2702
## adjCV        1.003   0.4452   0.4450   0.3478   0.3464   0.3182   0.2455
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV      0.2294   0.2238   0.2213    0.2124    0.1814    0.1803    0.1806
## adjCV   0.2288   0.2231   0.2206    0.2118    0.1807    0.1796    0.1800
##        14 comps
## CV        0.184
## adjCV     0.183
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X       53.06    66.75    76.13     83.6    87.98    91.23    94.37    96.38
## life    80.51    80.82    88.63     89.1    90.91    95.01    95.14    95.45
##       9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
## X       97.72     98.73     99.38     99.81     99.92    100.00
## life    95.57     95.98     97.10     97.16     97.17     97.19
```

Principal Component Number Selection

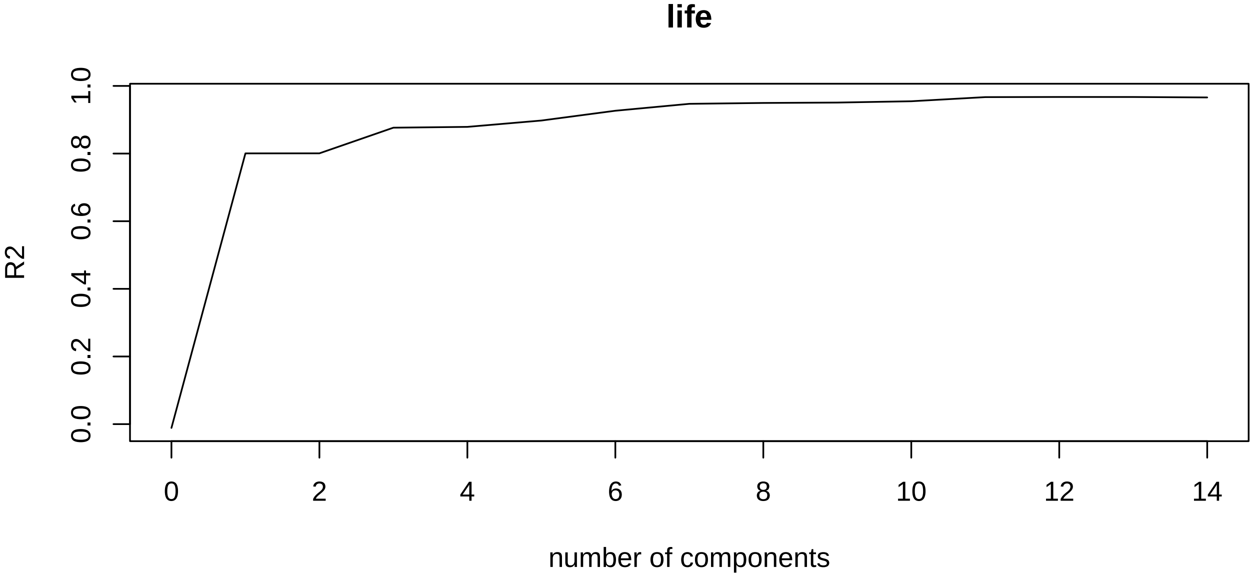
Once we've fit the model, we need to determine the number of principal components worth keeping. The way to do so is by looking at the test root mean squared error (test RMSE) and calculated by the k-fold cross-validation and the percentange of variance explained

1. RMSE This table tells us the test RMSE calculated by the k-fold cross validation. We can see the following: If we only use the intercept term in the model, the test RMSE is 0.3646. If we add in the first principal component, the test RMSE drops to 0.2363. If we add in the second principal component, the test RMSE drops to 0.2270. We can see that adding additional principal components actually leads to an decrease in test RMSE. However, after the number of components reaches 6, the improvement is slight.



2. Percentange of Variance Explained This table tells us the percentage of the variance in the response variable explained by the principal components. We'll be able to explain more variance by using more principal components. By adding in the top 6 principal component, we can explain over 90% (91.23%) of

the variation in the response variable, and we can see that adding in more than 6 principal components doesn't actually increase the percentage of explained variance by much.



Model comparison and Analysis

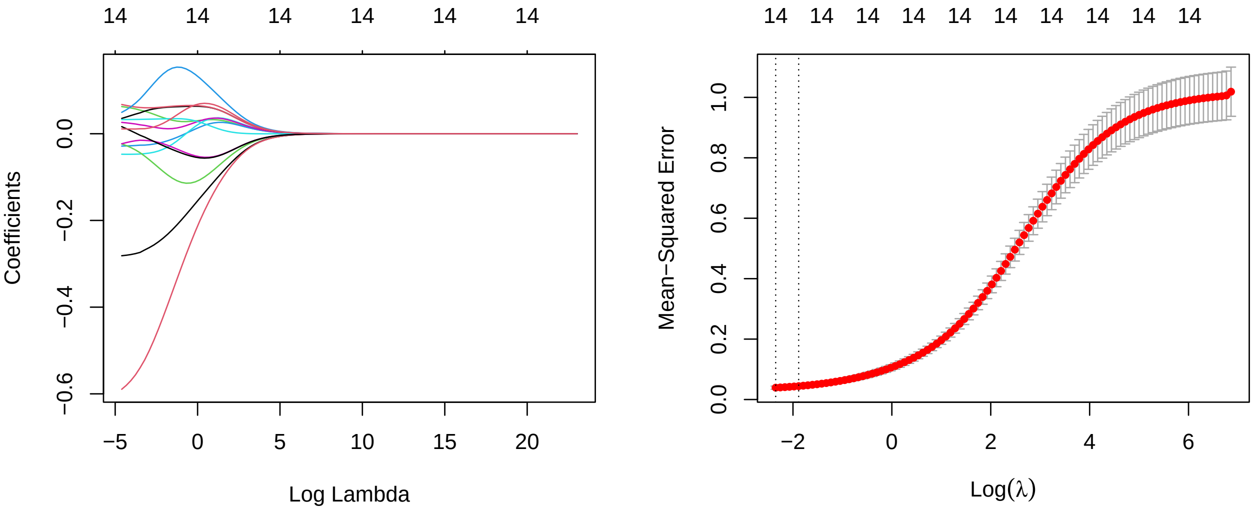
We exclude all dummy variables in PCA since it's harder to interprete the model once these dummy variables are included. However, we can add the principal components we choose with selected dummy variables in our final model to boost the model predictive performance.

```
## Best model by AIC includes:
##  isSouthAmerica, isCentralAmerica, PC1, PC2, PC3, PC4, PC5, PC6

## Best model by BIC includes:
##  isSouthAmerica, isCentralAmerica, PC1, PC2, PC3, PC4, PC5, PC6

## [1] "The best model from AIC metric is: 8"

## [1] "The best model from BIC metric is: 8"

## [1] "The best model's AIC value is: -64.678779753436"

## [1] "The best model's BIC value is: -32.8049216950285"

##
## Call:
## lm(formula = full_model_formula, data = data_pca_6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43863 -0.11412 -0.00862  0.12207  0.61508
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.054658   0.016286  -3.356 0.000975 ***
## isSouthAmerica    0.239276   0.060743   3.939 0.000119 ***
## isCentralAmerica  0.363817   0.051095   7.120 2.90e-11 ***
## PC1               0.324556   0.005424  59.833  < 2e-16 ***
## PC2              -0.038995   0.010683  -3.650 0.000348 ***
## PC3               0.276136   0.013552  20.376  < 2e-16 ***
## PC4               -0.068708   0.014393  -4.774 3.88e-06 ***
## PC5               -0.176971   0.019110  -9.260  < 2e-16 ***
## PC6                0.276017   0.022020  12.535  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.196 on 170 degrees of freedom
## Multiple R-squared:  0.9633, Adjusted R-squared:  0.9616
## F-statistic:   558 on 8 and 170 DF,  p-value: < 2.2e-16
```

PCA can effectively reduce multicollinearity in the dataset by transforming correlated variables into a set of linearly uncorrelated components. However, the adjusted R-squared of the PCR model is 0.9616, lower than the previous model we selected. This might be caused by the fact that we've already mitigated the problem of multicollinearity by model selection. As is shown in the hot plot before, we can see that the variables we choose are hardly linearly related. Also, PCA loses interpretation ability since its components are orthogonal vectors generated by the data.

Variance Reduction with Ridge and Lasso

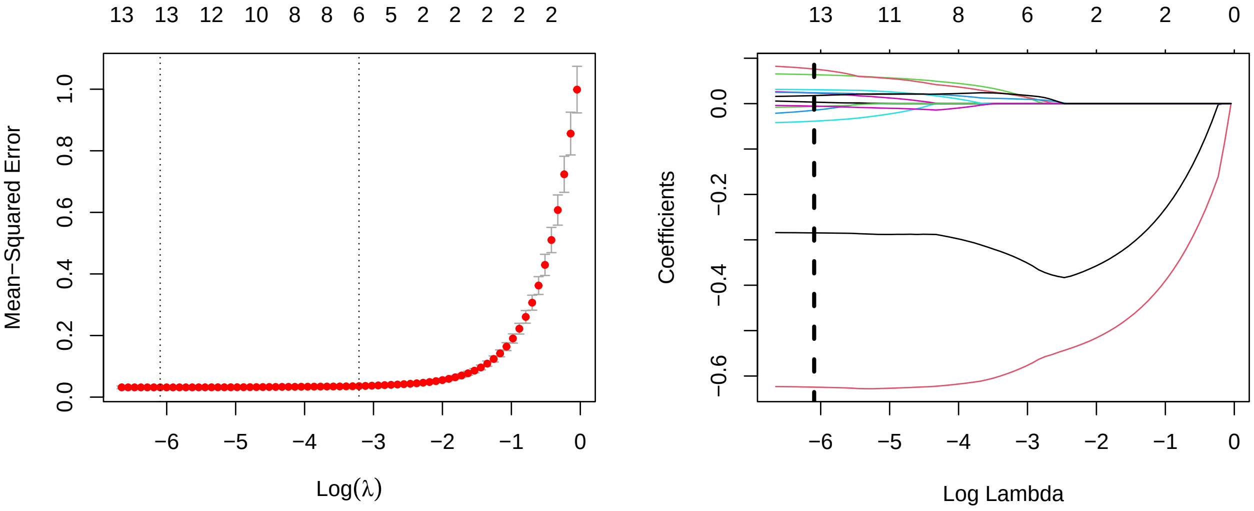
Ridge



```
## [1] "The best ridge lambda selected from the CV MSE is: 0.0953729236282161"
```

```
## [1] "The lowest CV MSE of the ridge model is: 0.0391759484887538"
```

Lasso



```
## [1] "The best lasso lambda selected from the CV MSE is: 0.00225508268171956"
```

```
## [1] "The lowest CV MSE of the lasso model is: 0.0316578462201936"
```

```
## Variables selected by Lasso includes:
##  infant_deaths, adult_mortality, alcohol, hepatitis_B, measles, bmi, polio, diphtheria, hiv, gdp, pop
```

Lasso and Ridge also use the stratefy of compromising bias for variance. Unlike AIC and BIC, which penalize on the number of variables, they directly penalize on the coefficients of the variables. The cross validation MSE achieved by lasso is much lower than ridge model. Also, Lasso model can be used as model selection strategy. Compared with AIC and BIC, Lasso includes more variables in the model selection. It is also worth noticing that the variables it chooses includes all the variables selected in previous models, proving that these variables are significant.

Conclusion

Objective

The project investigates the influence of factors like economic status, vaccination coverage, educational attainment, and health indicators on worldwide life expectancy using linear model's superiosity in inference and intepretability. Our goal doesn't focus on creating a linear model that precisely forecasts life expectancy based on several variables; it pinpoints and measures the effects of different social, economic, and health-related factors. The insights gained from this study could guide specific interventions and policy-making to improve health conditions around the world. To accomplish our objectives, employing diverse strategies learned in class, we developed two linear models. The first delivers precise predictions with a high adjusted R-squared value, while the second offers valuable insights from a different viewpoint.

Model Construction Procedure

Our approach to building the model consists of four main phases: data preprocessing, model selection, regression diagnostics, and model comparison. During data preprocessing, we first examine the data and make up for the missing values. Considering the dependece between yearly data, our solution is averaging the data over years. Then, we employ box-cox transformations to reduce skewness. In the model selection phase, we use AIC and BIC metrics to guide our choices and apply F-tests for selecting interaction terms. Ultimately, we develop two distinct linear models by incorporating different dummy variables for regression focusing on differentiating the developing status of countries and the other on countries' regions. Additionally, we explored other techniques like PCA, Lasso, and Ridge regression to decrease the model's variance and address multicollinearity. We also the effectiveness of these models.

Analysis and Insights from Model Coefficients

In our first model, we exclude the regional information and regress life expectency on Infant deaths, adult mortality, alcohol consumption, measles vaccine coverage, diphtheria vaccine coverage, population and development status and the interaction between development status and adult mortality and alcohol. Our second model include the regional factor and regress life expectency on infant deaths, adult mortality, population, measles vaccine coverage, diphtheria vaccine coverage, years in school and two dummy variables, indicating whether the country is in central America or south America.

Both models reveal that higher adult mortality and infant deaths negatively impact life expectancy, a finding that aligns with common understanding. A notable result is the positive correlation between diphtheria vaccine coverage and life expectancy, highlighting the significance of social healthcare services. Interestingly, a larger population seems to correlate with increased life expectancy, though the causality of this relationship is unclear; it might be that higher life expectancy leads to larger populations, or other confounding factors could be at play.

The first model unexpectedly suggests a positive correlation between alcohol consumption and life expectancy, a counterintuitive finding that merits further exploration. The second model shows the positive impact of education, indicated by the 'school' variable's positive coefficient.

The interaction terms in our first model merit further examination. Specifically, the positive interaction between adult mortality and development status implies that in developed countries, the detrimental effect of adult mortality on life expectancy might be mitigated. Conversely, the negative interaction involving alcohol consumption and development status suggests that the positive link between alcohol and life expectancy observed in the overall model may weaken or even invert in developed nations. This might account for the unexpected positive coefficient of 'alcohol': in developed regions, the overall impact of alcohol appears negative, potentially means alcohol is detrimental to one's health if abused. However, in developing countries, the access to alcohol means people's life condition is comparatively better, leading to higher life expectancy.

Contrary to expectations, both models indicate a negative correlation between measles vaccine coverage and life expectancy. Closer examination revealed that many measles data points were imputed with average regional figures or averaged over years due to missing original data, casting doubt on the reliability of the data and this variable's coefficient, necessitating further investigation.

Model Comparison

In this project, we developed two distinct models, each with a unique objective. Both models demonstrate high efficacy in prediction, as evidenced by their adjusted R-squared values exceeding 0.98 while using fewer than 10 variables. This indicates the effectiveness of our models in forecasting with the existing dataset. Additionally, the valuable insights derived from our analysis have the potential to inform targeted strategies and policy decisions aimed at enhancing global health conditions.

Discussion

In this project, we still have some limitations that can be further refined and imporved:

Data Reliability

As highlighted in the **Conclusion** section, the coefficient for *measles* presents a conceptual challenge, potentially arising from our approach to handling missing values in the dataset. This issue extends beyond *measles*, as similar imputation methods were applied to other variables with missing data. Enhancing the accuracy and completeness of our dataset could substantially improve the model's predictive power and reliability.

Time Series Analysis

Our dataset encompasses time series data, for which we employed an averaging method to mitigate dependency issues, based on observed linear trends over time. However, alternative approaches to handling time series data might yield more insightful results. We anticipate exploring these methods more thoroughly, particularly after systematically studying Time Series analysis, to enrich our understanding and model performance.

Outlier Elimination

Regression analysis involved identifying and removing outliers to enhance the model's performance. We relied on standard residuals and Cook's distance, utilizing default parameters for outlier detection, such as a threshold of two times the standard residual. This generalized approach, without a detailed examination of each outlier, might introduce inaccuracies. Future work will involve a more nuanced analysis of outliers to refine our model's accuracy and robustness.

Reference

- Lasha. Life Expectancy (WHO) - Updated. Kaggle. 2023. [Online]. Available: https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated.
- Fox, John. Applied regression analysis and generalized linear models. Sage Publications, 2015.
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning. Vol. 112. New York: Springer, 2013.