Group Project Report

Group C1 - Jerry Gai, Makayla Hastings, Subin An, and Sunny Lau

Introduction

The final dataset was collected from the World Health Organization (WHO) and the United Nations websites. Kumar Rajarshi (2015) merged the two data sets and posted the final dataset on Kaggle.com in 2018.

The data was collected between the years of 2000 to 2015 for 193 countries. Rajarshi removed missing data using the Missmap command in R and saw no evident errors. The majority of missing data came from population, hepatitis B, and GDP. Additionally, less known countries (Vanuatu, Tonga, Togo, Cabo Verde etc.) with a lot of missing data were excluded from the final dataset.

The dataset consists of 22 Columns and 2938 rows. Life expectancy is our response variable. Table 1 gives insight to the description of the variables. They are colour coded to denote four broad categories: immunization related factors, mortality factors, economical factors, social factors and other health related factors.

Variable	**Description/Units**
Country	Country
Life expectancy	Life Expectancy in age
Year	Year
Population	Population of the country
Hepatitis B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
Status	Developed or Developing status
GDP	Gross Domestic Product per capita (in USD)
Income composition of resources	Income composition of resources
percentage expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita (%)

Total expenditure	General government expenditure on health as a percentage of total government expenditure (%)
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
under-five deaths	Number of under-five deaths per 1000 population
HIV/AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)
Alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
Schooling	Number of years of Schooling (years)
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 (%)
Measles	Measles - number of reported cases per 1000 population
BMI	Average Body Mass Index of entire population

It has been observed that in the past 15 years, there has been a huge development in the health sector resulting in an improvement in human mortality rates, especially in developing countries in comparison to the past 30 years. We would like to investigate how life expectancy changes with respect to factors related to immunization in both developed and developing countries. We will also further investigate if certain variables within the immunization related factors have a larger effect on the improvements in life expectancy for developing countries so that those variables should be given priorities.

Analysis

We aim to examine the changes in life expectancy in terms of immunization factors. For this purpose, we have chosen variables hepatitis B, polio, and diphtheria, year, and status, to investigate the disparities in life expectancy between developing and developed countries over time. We will be performing our analysis using R studio.

Figure 1. Correlation Matrix

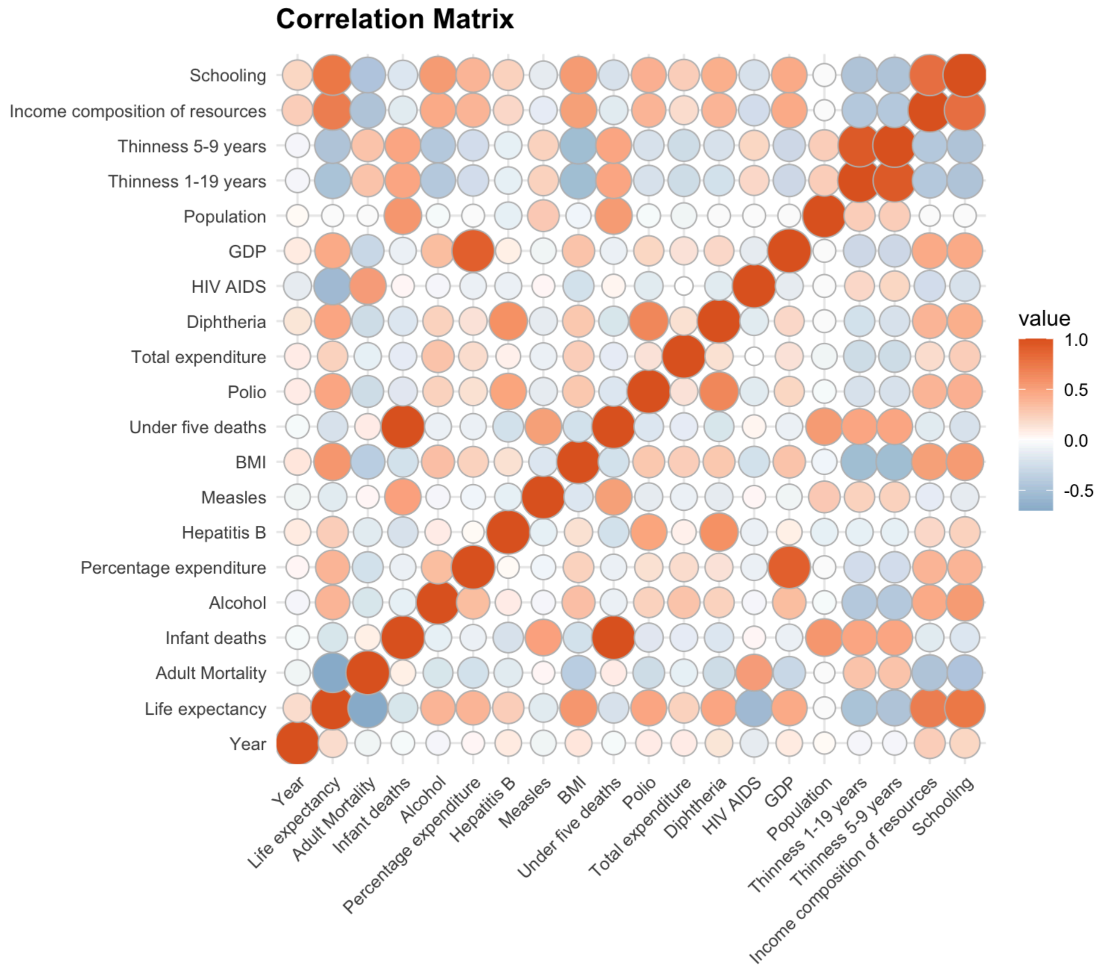
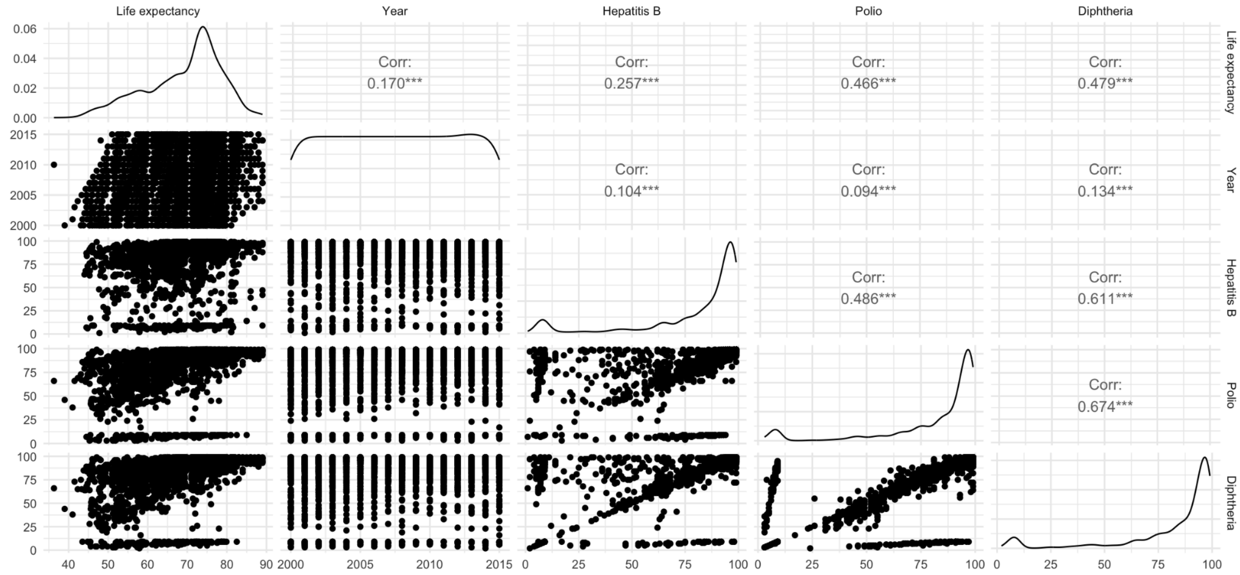


Figure 1 depicts the correlation matrix encompassing all variables in our dataset. Shades closer to orange or blue indicates stronger correlations between two variables. Notably, adult mortality, HIV AIDS, schooling, and income composition of resources have the highest correlation with life expectancy. Additionally, immunization related factors (hepatitis B, polio, diphtheria) display moderately strong correlations with life expectancy. However, these immunization factors may pose collinearity issues in linear regression, we may need to remove some of the highly correlated variables.

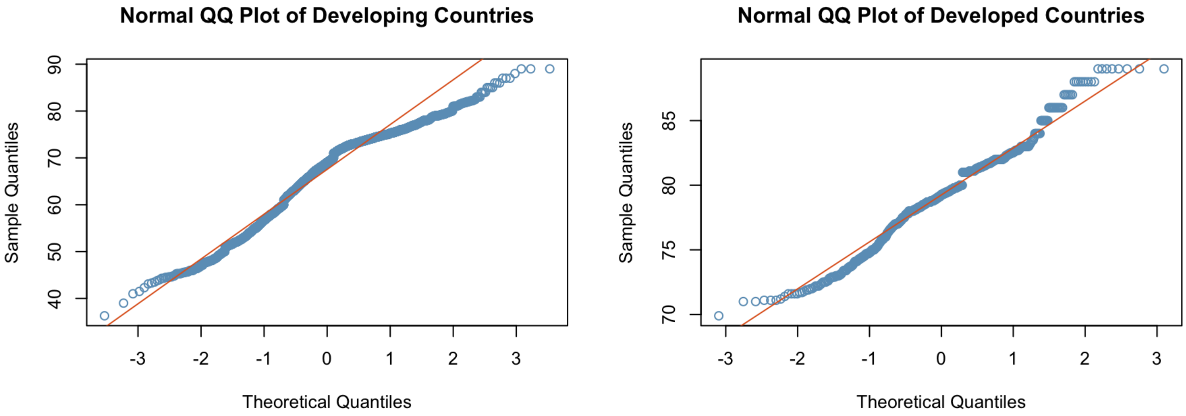
Figure 2. Pairs of Immunization Factors, Year and Life Expectancy



We plotted the pairs of immunization factors and year with life expectancy to visualize the data distribution and correlation coefficients. The moderate correlation between immunization factors and life expectancy implies we should further explore how immunization coverage influences life expectancy. Figure 2 also displays a moderate correlation between immunization factors which raises concerns regarding collinearity issues among these variables.

We will use two sample t-tests to determine if there is a significant difference in life expectancy of developed and developing countries. We are assuming observations to be independent of one another. Before performing the t-test, we checked if the data in both groups were approximately normally distributed using the Normal QQ plots in Figure 3. It can be seen that both of the plots approximately follow the QQ lines, which indicates that our data points meet the Normality assumptions for two sample t-tests.

Figure 3. Normal QQ Plots for Life Expectancy of Developing and Developed Countries.

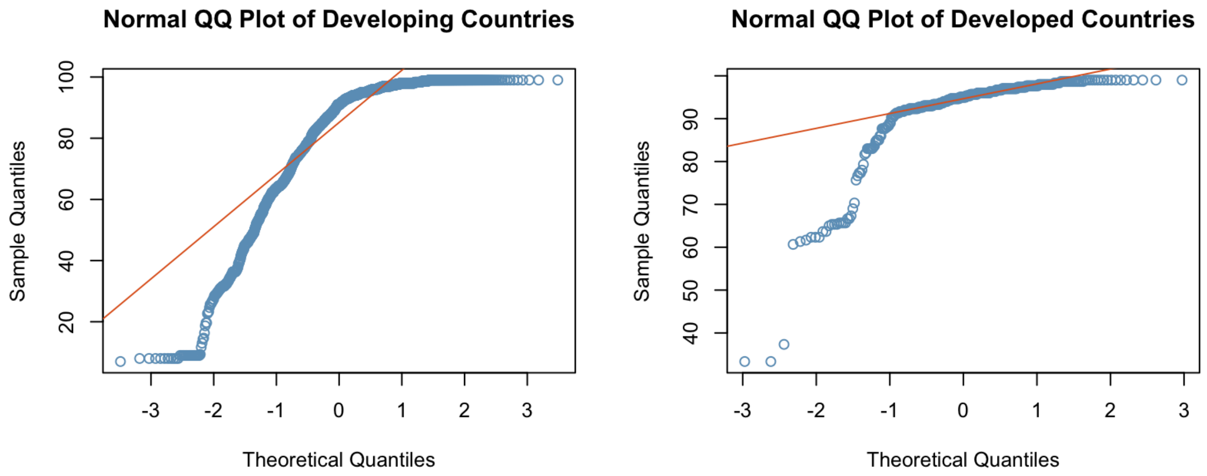


Although we expected the variances to be similar regardless of status, the developed countries had a variance of 15.4523 and the developing countries had a variance of 81.10969. Thus, we used Welch's two sample t-test instead of the standard independent samples t-test to account for the different variances.

The Welch two sample t-test comparing life expectancy between developed and developing countries generated a test statistic of 47.868 for 1807 degrees of freedom and a p-value less than 2.2 x 10^{-16}. These values indicate a statistically significant relationship in life expectancy between developed and developing countries, with developed countries having a higher mean life expectancy (79.19 years) than developing countries (67.11 years).

Next, we performed another two sample t-test to compare immunization coverage of hepatitis B, polio, and diphtheria with country status. Since we just want to see a general picture of the immunization differences between developing and developed countries, we decided to take the average coverage of the three factors to compare developing and developed countries. Figure 4 shows the data in both groups are not normally distributed, however due to the sample size being large and the central limit theorem, we can assume this will have little effect on our analysis.

Figure 4. Normal QQ Plots for Immunization Factors of Developing and Developed Countries



Similarly to the life expectancy t-test, when comparing immunization factors of developing and developed countries we used the Welch two sample t-test due to the variances being different. The variance of developed countries was found to be 97.54891 whereas the variance of developing countries was found to be 397.3399.

The Welch two sample t-test comparing immunization factors between developed and developing countries generated a test statistic of 14.255 for 881.79 degrees of freedom and a p-value less than 2.2 x 10^{-16}. These values indicate a statistically significant difference in immunization coverage between developed and developing countries, with developed countries having a higher mean immunization coverage (92.03%) than developing countries (82.14%).

Due to the significant differences between life expectancy of developed and developing countries and between immunization coverage factors of developing and developed countries, the following model was built allowing interactions between status and different immunization factors, where statusdeveloped = 1 for developed countries, and statusdeveloped = 0 for developing countries:

$$Life\,Expectancy =- 158.3 + 0.003927(Hepatitis.B) + 0.06907(Polio) + 0.07918(Diphtheria) + 0.1066(year) \\ + (- 0.02777(Hepatitis.B) - 0.05162(Polio) - 0.1052(Diphtheria) + 0.2056(year))statusdeveloped \\ - 387(statusdeveloped)$$

The summary of the model indicated only the intercept, polio, diphtheria, year, status,the interaction term between year and status and the interaction term between diphtheria and status were significant. We employed stepwise backward selection to determine a final model. The backward selection reduces the number of unnecessary variables by keeping only significant predictors in the model. By removing the least significant predictors, it improves the accuracy of the final model.

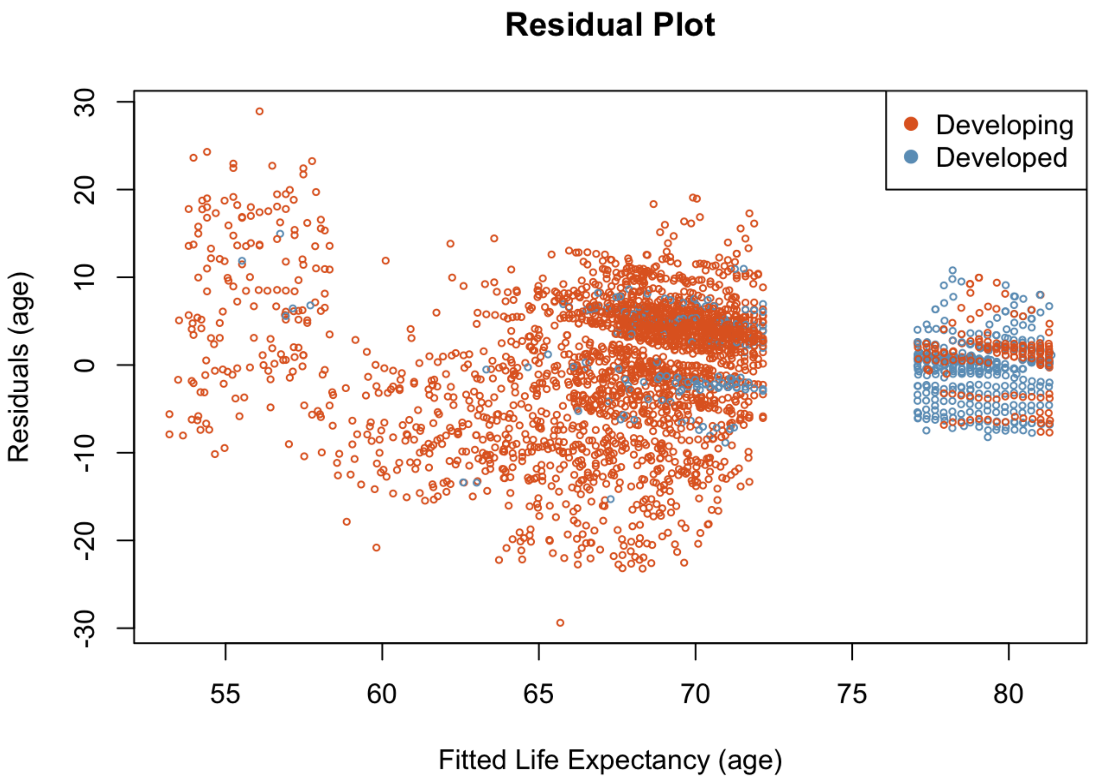
Based on the stepwise backward selection, we chose a model having the following eight parameters: intercept, polio, diphtheria, year, status, and the interaction between polio and status, diphtheria and status, and year and status.

To assess collinearity we found the variance inflation factors (VIF). Since there is a categorical variable Status, and interaction terms of it with other variables, the scaled generalized variance inflation factors (GVIF) is used. GVIF of polio coverage (GVIF = 8.312) and diphtheria coverage (GVIF = 8.316) are too high. These values indicate significant collinearity between both polio and diphtheria coverage. To fix this problem, we try dropping the diphtheria term due to it having the highest GVIF. We then conducted another stepwise backward selection on the model after dropping the diphtheria term. The new model's parameters are all significant with low GVIF values for polio, year and status, leaving us with the following as our final model, where statusdeveloped = 1 for developed countries, and statusdeveloped = 0 for developing countries:

$$Life\,Expectancy =- 506.1 + 0.1537(polio) + 0.2794(year) + 24.74(statusdeveloped) - 0.1578(polio * statusdeveloped)$$

Note that the coefficient of polio variable is 0.1537, indicating that polio coverage has a significant positive impact on life expectancy for developing countries. However, the coefficient of the interaction term of polio and status is -0.1578, which indicates that increasing polio coverage may not have that positive effect on life expectancy of developed countries. This further implies that developing countries should give immunization coverage more priority in order to catch up with developed countries.

Figure 5. Residual Plot of Final Model



The residual plot in figure 5 has two distinct groupings with a gap surrounding 75 years old. The residuals plotted for the younger group are primarily from developing countries, whereas the older group is primarily developed countries. This is consistent with the significant difference in life expectancy mean values we determined earlier in the analysis. We also noted the developing countries have a larger range of life expectancies and variance than the developed countries. Since each grouping primarily contained developing or developed countries only, we should assess the quality of the residual plot by looking at each grouping separately. Within each grouping, we can see that there is almost no obvious pattern for the residual points and all the points are scattering roughly evenly around the zero line. Hence, we can confirm that our final model did a good job to describe the relationships between life expectancy and other variables we chose.

Initially, we planned to conduct further analysis on which variables within the immunization related factors have a larger effect on the improvements in life expectancy for developing countries. However, since our final model dropped most of them and left only one of them, the polio coverage, it would be difficult to determine which one is more important as they all have high collinearity problems.

Conclusion

Through our analysis we found significant differences in life expectancy and immunization coverage between developed and developing countries. Developed countries generally have higher life expectancies and greater immunization coverage than developing countries. Due to the high collinearity between immunization factors, we conducted model selection techniques to select our best model and our final model for predicting life expectancy was based on status, polio immunization coverage, and year.

The results of our analysis contributes to our understanding of the positive relationship between life expectancy and immunizations in regards to global health disparities. We found that increasing immunization coverage will have a significant positive effect on life expectancy of developing countries, but the same effect may not be that huge on developed countries given they already have high enough coverage and life expectancy. Hence, immunizations against polio and other diseases should be considered a priority when addressing inequality between developed and developing countries.

The original dataset poses some limitations to our analysis due to covering only 15 years. Data from before 2000 and after 2015 would provide us with a better understanding of how immunization coverage's effects on life expectancy has changed as technology has become more advanced. Additionally, the assumption that all observations are independent of each other could have an effect on our analysis.

Future directions of study could encompass how the covid-19 pandemic and immunization coverage has affected life expectancy or if the inclusion of other vaccines, such as influenza, play a significant role. The goals of decreasing global health inequalities should be at the forefront of future analysis to improve population health outcomes.

References

KumarRajarshi. (2015). *Life Expectancy (WHO)*. Kaggle.com. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data