Project Report

Indiana University, Indianapolis

INFO- B18- Applied Statistics in Biomedical Informatics

April 30, 2024

Examining the Impact of Health Indicators, Socioeconomic Factors, and Education on Global Life Expectancy

Sree Uma Maheshwar Vangapaty

Part 1:Introduction

Analysis of Life Expectancy Data (2000-2015)
The dataset from the WHO's Global Health Observatory covers health-related metrics for 193 countries from 2000-2015. It includes 22 columns and 2,938 rows, focusing on immunization, mortality, economic, and social factors. Life expectancy, a crucial indicator of a nation's overall well-being, is influenced by a multitude of factors ranging from health conditions to socioeconomic determinants (Life Expectancy (WHO), 2018). This study delves into the intricate interplay between life expectancy and variables such as health indicators, economic status, and educational attainment, using a comprehensive dataset encompassing 193 countries spanning the years 2000 to 2015. By employing rigorous statistical analyses, including linear regression modeling, the research aims to uncover the primary drivers of life expectancy disparities across nations (Life Expectancy (WHO), 2018).

Part2: Data Collection

Link**:** https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
Variables:
Country, Year, Status, Life expectancy, Adult Mortality, infant deaths, Alcohol, percentage expenditure, Hepatitis B, Measles, BMI, under five deaths, Polio, Total expenditure, Diphtheria, HIV.AIDS, GDP, Population, thinness1-19 years, thinness 5-9 years, Income composition of resources, Schooling(Life Expectancy (WHO), 2018).
Only country and status columns are categorical, rest of the columns are numerical.

Parameters of interest:

For the research questions:

1. "What are the primary health indicators and socio-economic factors that significantly influence life expectancy across different countries?"
2. "How does the level of education impact life expectancy when controlling economic status and health indicators?"

Status, Life expectancy, Adult Mortality, infant deaths, Alcohol, percentage expenditure, Hepatitis B, Measles, BMI, under five deaths, Polio, Total expenditure, Diphtheria, HIV AIDS, GDP, thinness1-19 years, thinness 5-9 years, Income composition of resources, Schooling are the relevant variables. These variables cover various health indicators related to mortality rates, immunization coverage, and disease prevalence, as well as socio-economic factors like GDP, and income composition. Analyzing these variables can provide insights into how different health and socio-economic factors influence life expectancy across countries.

Part3: Data cleaning

We have found many missing values in the several columns, and we have omitted the rows with these missing values resulting in 1853 rows. Spaces in the column names are removed.

Part4: Exploratory Data Analysis

We initiated our analysis by generating scatter plots and boxplots for each dependent variable

against the life expectancy measure. This allowed us to observe the interactions between each explanatory variable and the response variable, as well as to identify any outliers that could significantly affect the regression models, potentially skewing the analysis. For continuous variables like "Life.expectancy", "Adult.Mortality", "infant.deaths", and others, the minimum and maximum values, as well as the quantiles (1st quartile, median, and 3rd quartile) are measured. These statistics give insights into the central tendency and spread of the data. For categorical variables like "Status", we displayed the class and unique values or modes present in the data. This information helps understand the nature and characteristics of the variables, which is crucial for data exploration and analysis. We defined a function to count outliers in each numeric variable. Outliers were identified using the interquartile range (IQR) method, where values falling below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR were considered outliers. We then applied this function to each numeric column in our dataset, selected_data1, and calculated the total number of outliers per column. To visually inspect the distribution of data and identify outliers, we created boxplots for each numeric variable. We divided the plots into batches for better visualization, allowing us to handle multiple variables efficiently by displaying up to nine boxplots per page. This step helped in visually assessing the spread and central tendency of the data, as well as spotting any extreme values. We implemented a function to cap outliers, replacing them with the nearest non-outlier values. This capping was done by setting values below the lower boundary to the lower boundary value and values above the upper boundary to the upper boundary value. This modification was applied to all numeric columns in the dataset, thus reducing the impact of extreme values on the analysis. After capping the outliers, we generated a new set of boxplots to verify the effect of the adjustments. These plots were also organized into batches, like step 2, but this time the plots were colored differently to differentiate them from the initial plots. This final visualization step confirmed that the data's extreme values had been adequately managed, ensuring that our subsequent analyses would be less likely to be skewed by such anomalies.

The scatter plot analyses reveal distinct linear relationships between life expectancy and various health, demographic, and socio-economic variables. For instance, a significant negative linear correlation is observed between life expectancy and both adult and under-five mortality rates, indicating that higher mortality rates adversely affect life expectancy. Conversely, positive linear relationships are evident with factors such as GDP, schooling, and health expenditure, where increases in these variables are associated with improvements in life expectancy. Specifically, a stronger GDP per capita and higher levels of education substantially boost life expectancy, underscoring the importance of economic and educational advancements in enhancing health outcomes. Additionally, the plots also show weaker but positive relationships with variables like alcohol consumption and hepatitis B immunization coverage, suggesting that these factors, while influential, are less potent predictors of life expectancy compared to economic and educational indicators. However, the relationship between life expectancy and measles incidence appears relatively flat, suggesting a minimal impact of measles incidence on life expectancy within the scope of the studied variables. This array of relationships highlights the complex interplay between socio-economic conditions, health expenditures, and demographic factors in determining the life expectancy of populations. We have done the Spearman's rank correlation to identify the correlation between the selected variables.

Part5: Methodology

A stepwise multi-variable linear regression model was implemented to derive the optimal model for predicting life expectancy using the explanatory variables. Initially, all variables were included in the model, which then systematically removed the non-significant ones. This refinement resulted in a model defined only by significant variables, which achieved the highest possible adjusted R2 value. This methodology is highly effective in identifying significant correlations between life expectancy and the examined variables. The analysis tests two hypotheses: the null hypothesis (H0) states that the variables do not significantly impact life expectancy, while the alternative hypothesis (Halt) contends that the variables are significant determinants of life expectancy. The p-value for each variable's coefficient is crucial for determining whether there is sufficient evidence to reject the null hypothesis and retain the variable in the final model.

Part6: Results

The initial comprehensive model included all the variables deemed potentially influential on life expectancy. This model was refined to identify and retain only the statistically significant variables, enhancing the predictive accuracy and relevance of the resultant model. Although all initial variables were considered important, not all were retained; instead, the process distilled them down to the ones that genuinely impacted life expectancy. This refinement led to the development of a more streamlined model which consisted of a select group of significant predictors such as the status of development, adult mortality, infant deaths, health expenditure as a percentage of GDP, Hepatitis B coverage, under-five deaths, Diphtheria immunization coverage, HIV/AIDS prevalence, income composition of resources, thinness in children aged 5-9 years, and schooling.

These significant variables were significant at a 0.05 level, which effectively allowed for the rejection of the null hypothesis for these variables, confirming their substantial impact on life expectancy. The p-values, notably small for predictors like adult mortality and HIV/AIDS, provided strong statistical evidence against the null hypothesis, further reinforcing their importance in the model. The significant predictors showcased varied effects, with some like income composition of resources showing a highly positive influence on life expectancy, whereas others like HIV/AIDS demonstrated a severe negative impact.

The refined model achieved a high adjusted R-squared value of 0.8319, indicating that approximately 83.19% of the variability in life expectancy could be explained by the selected variables. This robust model not only underscores the critical role of socio-economic and health-related factors in shaping life expectancy but also highlights the precision of using statistical methods to distill meaningful insights from complex data sets. Such a model provides a valuable tool for policymakers and health professionals aiming to identify and prioritize interventions that can significantly improve life outcomes.

Table

Variable	Coefficient	p-value
Intercept	64.3279	< 2e-16
Status	-1.1437	2.67e-05
Adult Mortality	-0.01786	< 2e-16
Infant Deaths	0.2085	0.000119
Percentage Expenditure	0.001509	1.19e-12
Hepatitis B	-0.01911	0.006249
Under Five Deaths	-0.2130	1.20e-06
Diphtheria	0.05745	1.45e-07
HIV/AIDS	-6.6334	< 2e-16
Income Composition of Resources	10.0672	< 2e-16
Thinness 5-9 Years	-0.1735	4.87e-11
Schooling	0.2262	2.87e-05

This regression model provides a detailed quantification of how various factors impact life expectancy. The intercept indicates that if all other variables are held at zero, the expected life expectancy is approximately 64.33 years, a hypothetical scenario given the nature of the variables. For example, the model suggests that being in a developing status (compared to a developed one) decreases life expectancy by approximately 1.14 years. Similarly, each unit increase in adult mortality rate (per 1,000 people) decreases life expectancy by about 0.018 years, illustrating a significant adverse impact. In contrast, some variables show a positive influence on life expectancy. Each additional year of schooling is associated with an increase in life expectancy of about 0.226 years, highlighting the importance of education. Similarly, for every unit increase in the income composition of resources, which may reflect socio-economic status, life expectancy increases by a substantial 10.07 years, underscoring the strong effect of economic factors on health outcomes. Moreover, health-related factors such as infant and under-five mortality rates show that higher mortality rates are linked to significantly lower life expectancy, with coefficients of 0.209 and -0.213, respectively. This indicates that improvements in child health could lead to substantial increases in life expectancy. Notably, the model also captures the drastic negative impact of HIV/AIDS, where each unit increase in the prevalence rate significantly reduces life expectancy by about 6.63 years. It demonstrates the complex interplay of socio-economic, educational, and health variables in determining life expectancy, providing valuable insights for policymakers to prioritize interventions in areas that could yield significant improvements in public health outcomes.

Part7: Model Diagnostics

Histogram of Residuals with Normal Curve

The histogram of residuals superimposed with a normal curve is a crucial diagnostic tool. It

visually assesses the normality of residuals, which is an important assumption in linear regression. The histogram shown indicates that the residuals of the model are approximately normally distributed, as the shape of the histogram follows the bell curve closely, albeit with some deviations, particularly in the tails.

Residual Plot

The residuals are scattered around the zero line without forming any discernible patterns or funnels. This spread indicates that the residuals are well distributed, suggesting that the homoscedasticity assumption (constant variance of error terms) is reasonably met. The absence of clear patterns or trends also suggests that the linear model is a good fit for the data.

QQ Plot

The majority of points in the QQ plot adhere closely to the reference line (red line), which indicates that the residuals closely follow a normal distribution. So, we concluded that the residuals were nearly normal.

Part8: Conclusion

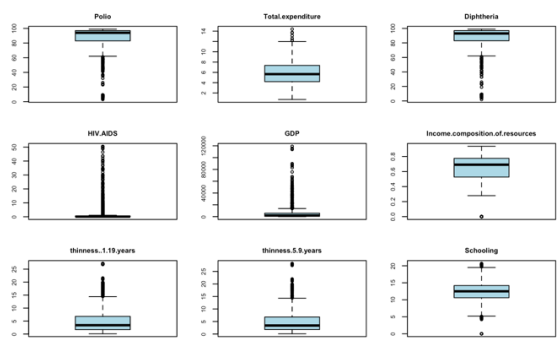
From the statistical analysis it can be concluded that the best model for determining the life expectancy is a linear regression model with the status of development, adult mortality, infant deaths, health expenditure as a percentage of GDP, Hepatitis B coverage, under-five deaths, Diphtheria immunization coverage, HIV/AIDS prevalence, income composition of resources, thinness in children aged 5-9 years, and schooling as significant explanatory variables.
For the first research question, small p-values indicate that the observed relationships between these variables and life expectancy are highly unlikely to have occurred by chance, the evidence strongly suggests that these health indicators and socio-economic factors are indeed significant determinants of life expectancy across different countries. For second research question, the level of education, represented by the "Schooling" variable, has a significant positive impact on life expectancy, even after controlling for economic status and health indicators. The p-value of 2.87e-05 for the "Schooling" variable indicates strong evidence against the null hypothesis, suggesting that education plays a crucial role in determining life expectancy, beyond the effects of other factors.

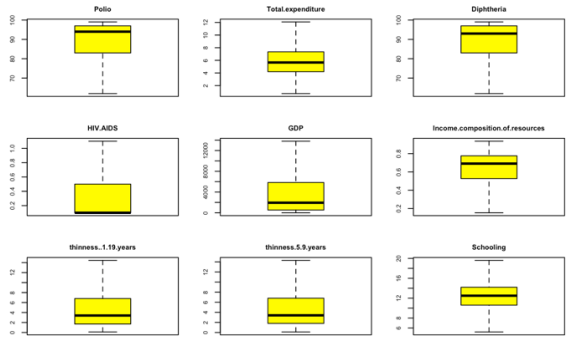
References

Life expectancy (WHO). (2018, February 10). Kaggle.

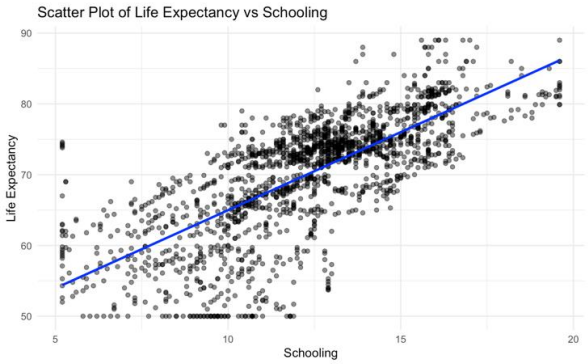
https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

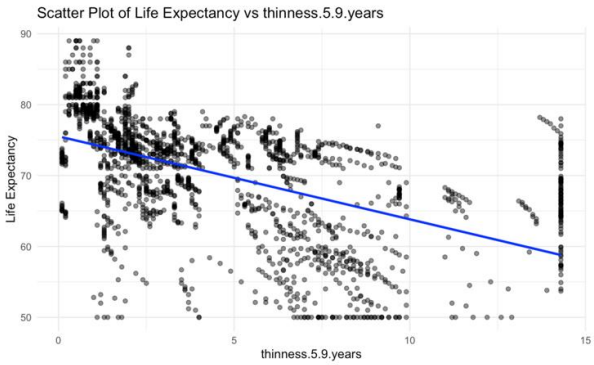
Appendix

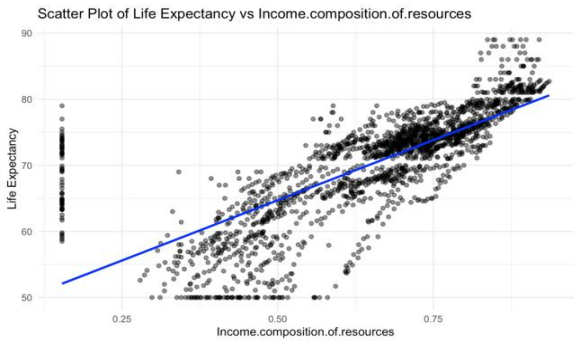


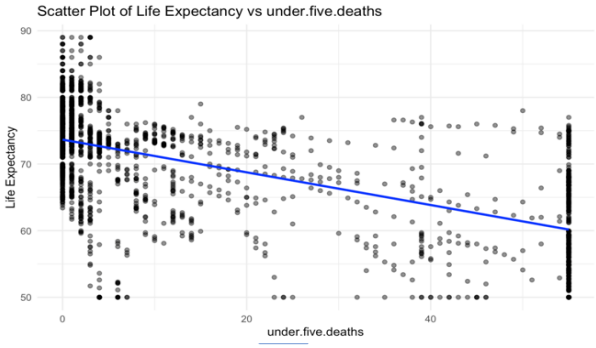


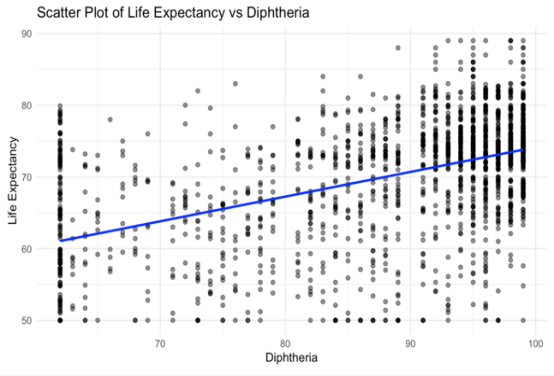
Detection and capping the outliers

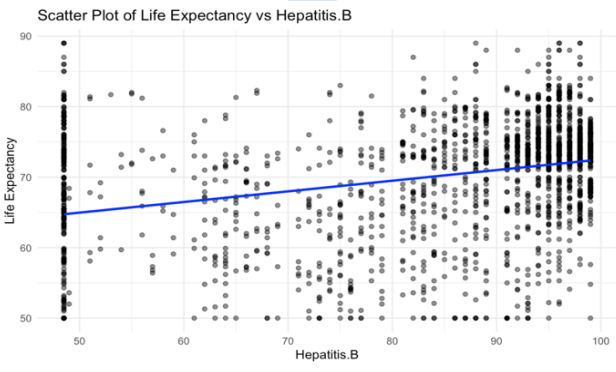


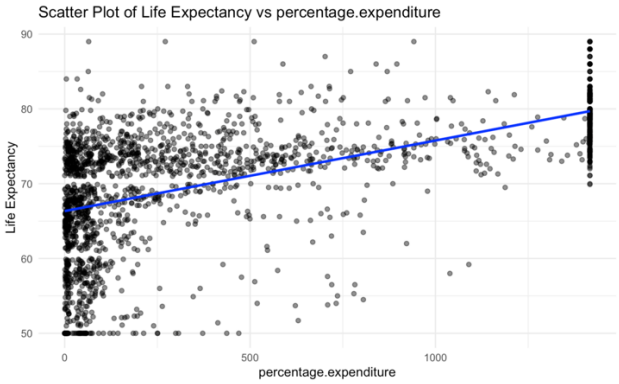


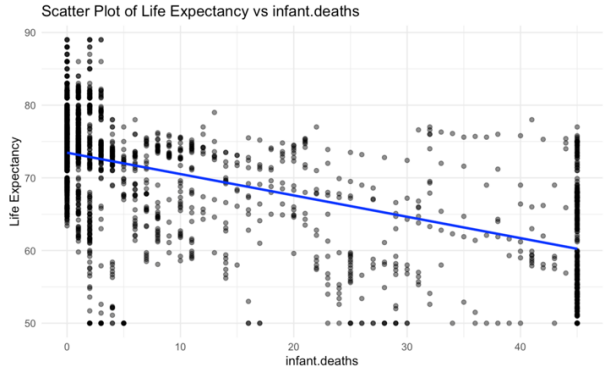


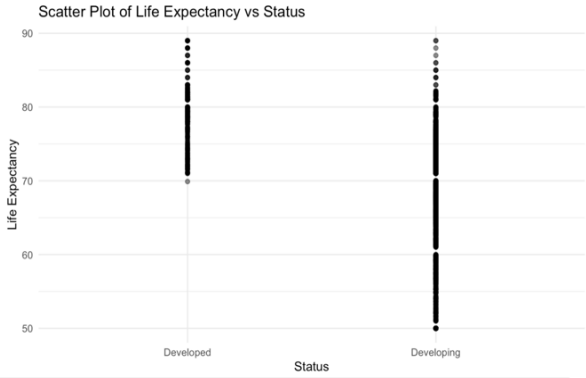


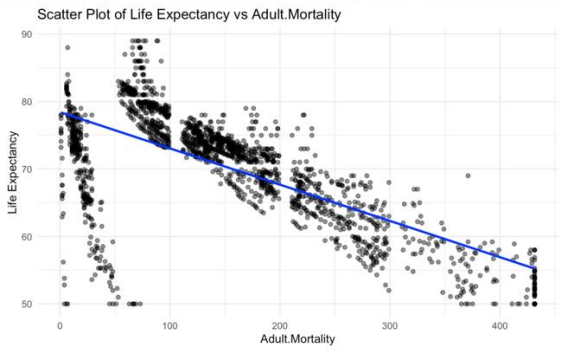


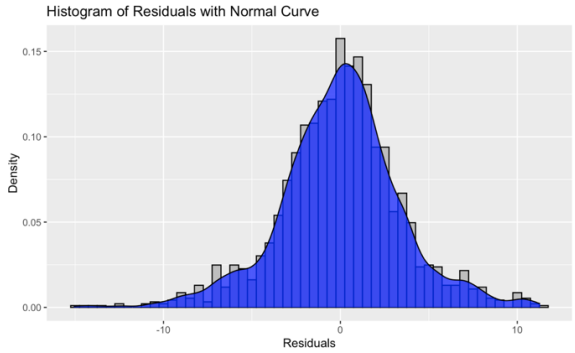




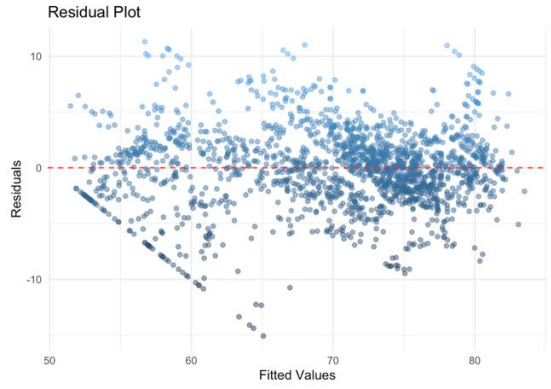


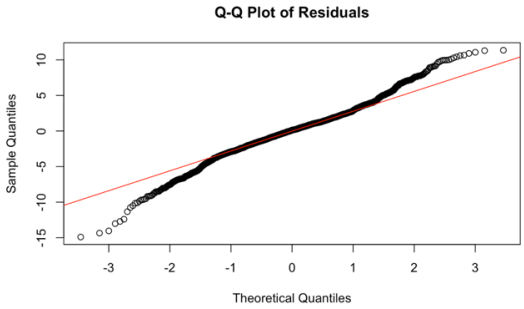


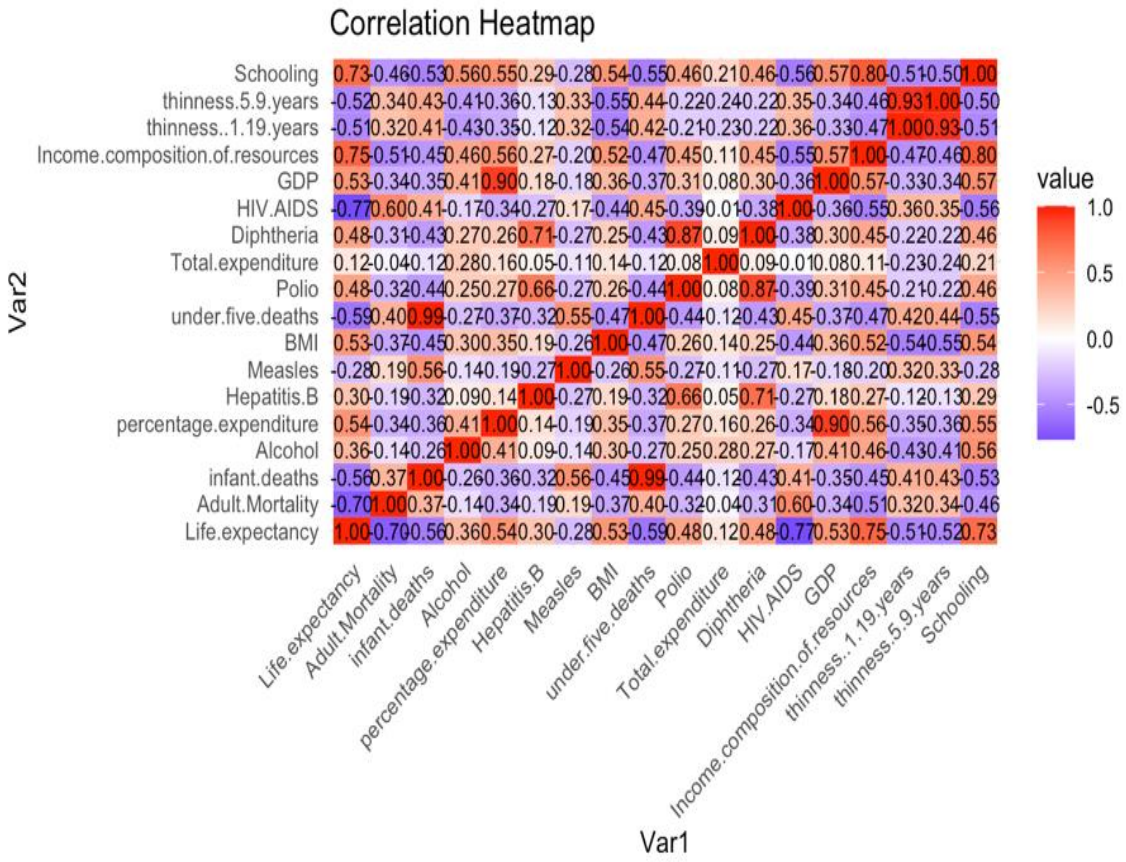




Scatter plots for the significant predictors







Spearman's rank Correlation heat map