Global Life Expectancy Due to Socioeconomic Factors

Introduction

Motivation and Research Question

Life expectancy is a key measure of a nation's health as it reflects the nation's healthcare systems and overall development. Understanding the factors that influence life expectancy is crucial for policymakers aiming to improve public health. Among various factors, gross domestic product (GDP) per capita and health expenditure are often highlighted as significant contributors to life expectancy. Higher GDP per capita may provide individuals with better access to nutrition, education, and healthcare services, while increased health expenditure can enhance the quality and availability of medical care. This leads us to our central research question:

How does Health expenditure and GDP affect life expectancy across countries in 2014?

The Dataset

The chosen dataset measures several factors that affect life expectancy, with the health factors for 193 countries from 2000 to 2015 taken from the World Health Organization (WHO) website and the corresponding economic factors taken from the United Nations (UN) website. The dataset consists of 2938 rows and 22 variables, including immunisation-related factors, mortality factors, and economic factors. However, since our research question primarily focuses on the impact of economic factors on life expectancy, we decided to eliminate the health-based variables. Therefore, the main variables explored in this report are the following:

Response Variable:

- **Life Expectancy**: The life expectancy in years

Main Economic Exploratory and Potentially Confounding Variables:

- **Status**: Whether the country is classified as developing or developed
- **Percent_expend**: Health expenditure as a percentage of GDP per capita
- **Total_expend**: Health expenditure as a percentage of total government spending
- **GDP**: Gross domestic product per capita (in USD)
- **Income_comp:** Human development index component related to income (ranging from 0 to 1)
- **Schooling**: Average number of years of education
- **Pop:** Total population of the country

Additional explanatory variables used in the dataset are listed in the *Appendix A*.

Analysis

Data Cleaning

The original dataset contained missing values as well as inappropriate headers. To prepare our dataset for further analysis, we got rid of all rows containing missing values and renamed several columns to more appropriate headers shown above. Furthermore, since a specific year of focus was chosen (2014), we filtered out all rows for that year, getting rid of the year column entirely. This ensures that every row in our cleaned dataset corresponds to a distinct country, which means we can eliminate the country column as well. Additionally, since we focus on economic covariates, we decided to get rid of health factors at this stage. These modifications cut down our number of observations to 131.

Exploratory Data Analysis and Primary Selection of Covariates

Before we did any further analysis, we visualised the distribution of life expectancy (see *Figure 1*). From this figure, the distribution for life expectancy is approximately normal, which indicates that linear regression may be appropriate for this analysis.

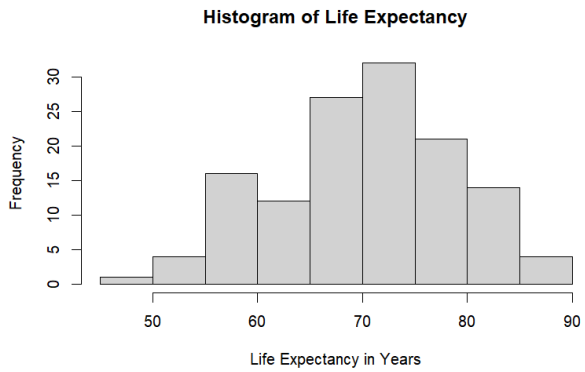

Figure 1: *Distribution of life expectancy*

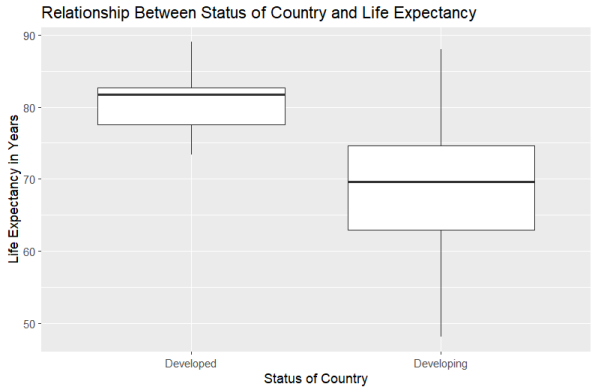

Figure 2*: Boxplot of response based on status of country*

Since status is a categorical variable, we decided to take a boxplot approach to identify its significance on life expectancy. As we can see in *Figure 2* above, there is a clear difference in the median between the levels of status, indicating that the status of a country could be significant in determining life expectancy, and it should remain in our dataset. Moreover, from this figure, we can infer that developed countries seem to have a higher life expectancy than developing countries.

To get an initial idea of significant variables, we began by looking at the correlation between the numerical variables in our dataset. As our dataset has 20 variables, a heatmap was produced for the numerical ones (see *Appendix B*). From this, the life_expect row was extracted and correlations for economic factors are seen in *Figure 3*.

Variable	Correlation	Relationship
percent_expend	0.41	Positive and weak
total_expend	0.32	Positive and weak
gdp	0.44	Positive and weak
income_comp	0.89	Positive and strong
schooling	0.80	Positive and strong
pop	-0.04	Negative and weak

***Figure 3**: Correlation between numeric covariates and response*

Data Transformation

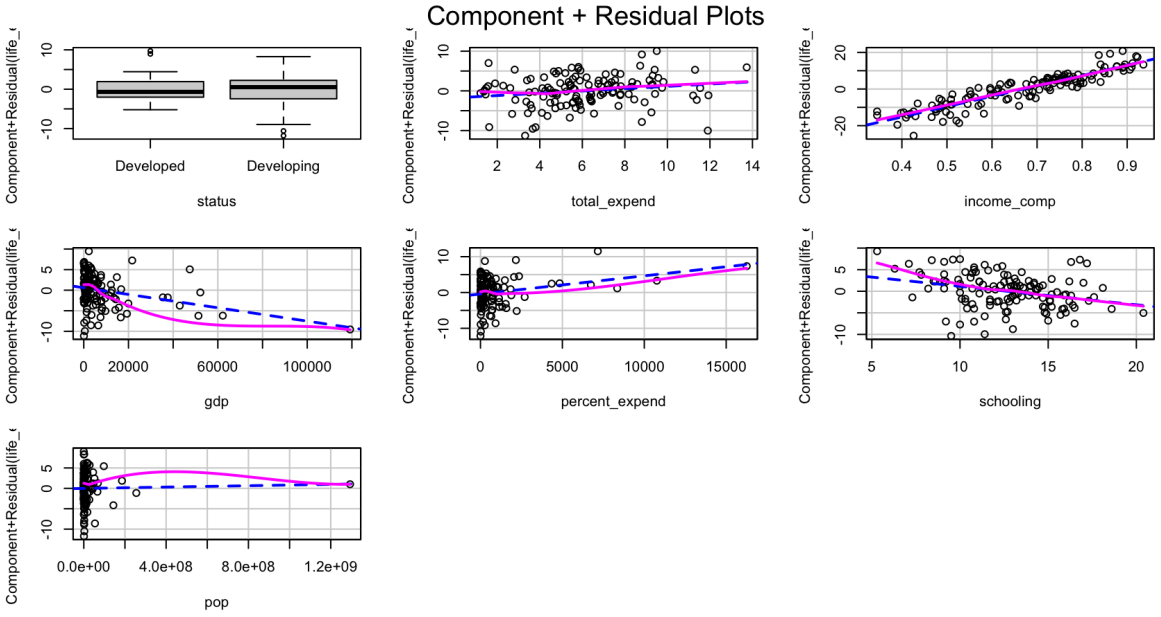

***Figure 4:** CR plot between covariates and response*

We start with fitting a full linear model with the 7 covariates we have selected (see *Figure 4*). Then, we plot the residual of each covariate. It is obvious that some covariates are heavily skewed right with a few outliers on the far left side. This clearly violates the assumption of linearity, and the validity of the fitted model will be affected. Therefore, we implemented $\log(x_i+1)$ to *gdp*, *pop*, and *percent_expend* to fix the distribution and handle zeros.

The transformed covariates are:

- status
- log_percent_expend
- schooling
- total_expend
- income_comp
- log_pop
- log_gdp

Model Selection

We implemented the stepwise selection to find the best fitted model and the table below shows the AIC of some of the models fitted in the process

Variables	**AIC**
Total_expend + income_comp + log_gdp + log_percent_expend + schooling + status	731.1362
Total_expend + income_comp + log_gdp +log_percent_expend+schooling	729.2596
Total_expend+ log_gdp + income_comp+log_percent_exp	729.0661
total_expend + income_comp + log_gdp	730.0338

Figure 5: *AIC values for different models*

From the table we can see that the following variables produce the lowest AIC and are thus selected to used for out final model

- total_expend
- income_comp
- log_gdp
- log_percent_expend

Afterwards we compared several statistics of our selected model, such as the residual standard error, R-squared, adjusted R-squared, AIC, and Mallows' C_p, against the original full model (see *Figure 6*). Although the RSE is lower in the full model, the adjusted R-squared for our selected model is higher and the AIC for our model smaller, indicating our model is a slightly better fit for our data. Since Mallows' C_p statistic is meaningless for the full model, the C_p statistic for our model is quite close to 5, meaning that it is an acceptable model.

Model	**RSE**	**R^2**	**Adj. R^2**	**AIC**	**Mallows' C_p**
Full Model	3.809	0.8138	0.8032	732.4860	8.0000
Best Selection Model	3.818	0.8101	0.8041	729.0661	4.4466

Figure 6: *Important statistic for model comparison*

Looking at *Figure 7*, we see that both variables' VIF values are below 10, this means that multicollinearity between the variable and life expectancy is not an issue. Additionally, we see that *total_expend* is not significantly associated with life expectancy at a 5% significance level.

Variable	Estimate	Standard Error	p-value	VIF
total_expend	0.2645	0.1410	0.0630	1.1431
log_gdp	-1.1222	0.5410	0.0401	9.1316

Figure 7: *Statistics for our variables of interest*

Selected Fitted Model

Life expectancy = 38.5576 + 0.2645 * (Total Expenditure) + 51.4139 * (Income_comp) - 1.1222 * log(GDP) + 0.8571 * log(Percentage Expenditure)

Model Interpretation

Intercept: 38.46

- When all covariates are at zero (which is unrealistic), the baseline life expectancy is expected to be **38.56 years**.

Total Expenditure: 0.2645

- One unit increase in total health expenditure is associated with an **expected increase of 0.26 years** in life expectancy.

Income composition: 51.41

- One unit increase in the income composition index is associated with an **expected increase of 51.41 years** in life expectancy.

Log(GDP): 1.12

- One unit increase in log(GDP) is associated with an **expected decrease of 1.12 years** in life expectancy

Log(percentage expenditure): 0.86

- One unit increase in log(percentage expenditure) is associated with an **expected increase of 0.86 years** in life expectancy

Model Diagnostic

We conducted model diagnostics on the selected model by plotting its residual plot and Q-Q plot. In the residual plot (see *Figure 8*), we can see the residuals spread roughly around the horizontal line at 0 with no clear patterns. This indicates that the relationship between the covariates and the response variable is approximately linear. Therefore, the linearity assumption is satisfied. However, the plot is slightly fan shaped, the residuals appearing a little bit more frequently as the fitted values increase. Although this might suggest potential heteroscedasticity, we still consider that the homoscedasticity assumption was met.

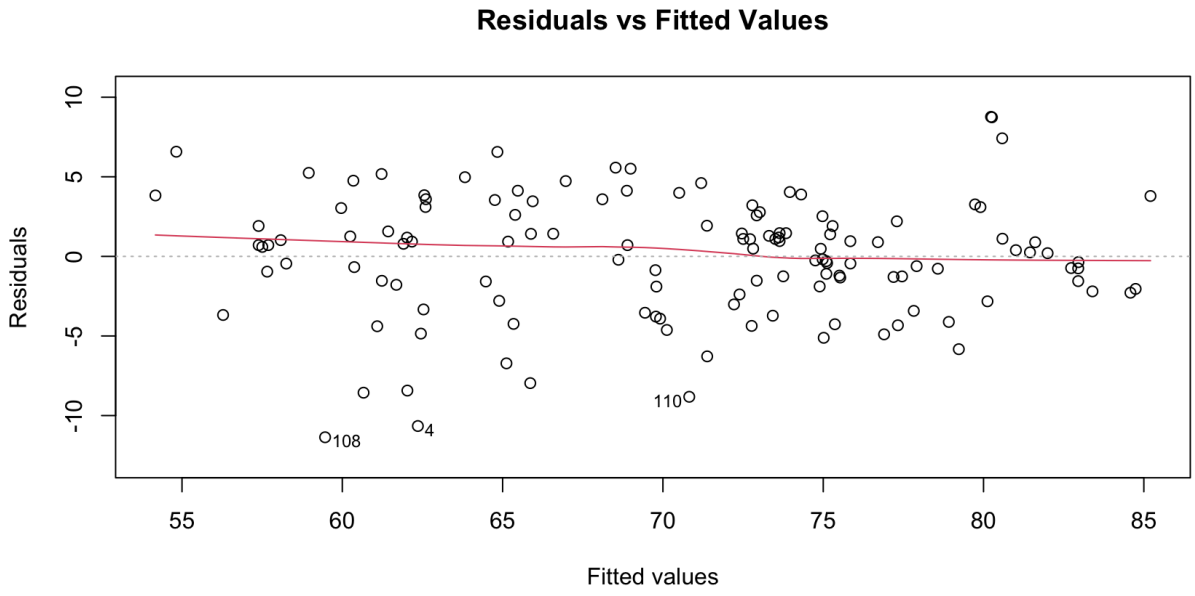

***Figure 8**: Residual plot of final model*

In the Normal Q-Q plot (see *Figure 9*), most of the points lie on the diagonal reference line. A few deviations at both of the tails, this suggests some light-tailed behaviour. However, these deviations are reasonable and unlikely to have a huge impact on the validity of the model. Therefore, we consider the residuals are approximately normally distributed, the assumption of normality of residuals satisfied. Overall, together with the independence assumption we have made priorly, the assumptions of linear regression were sufficiently met.

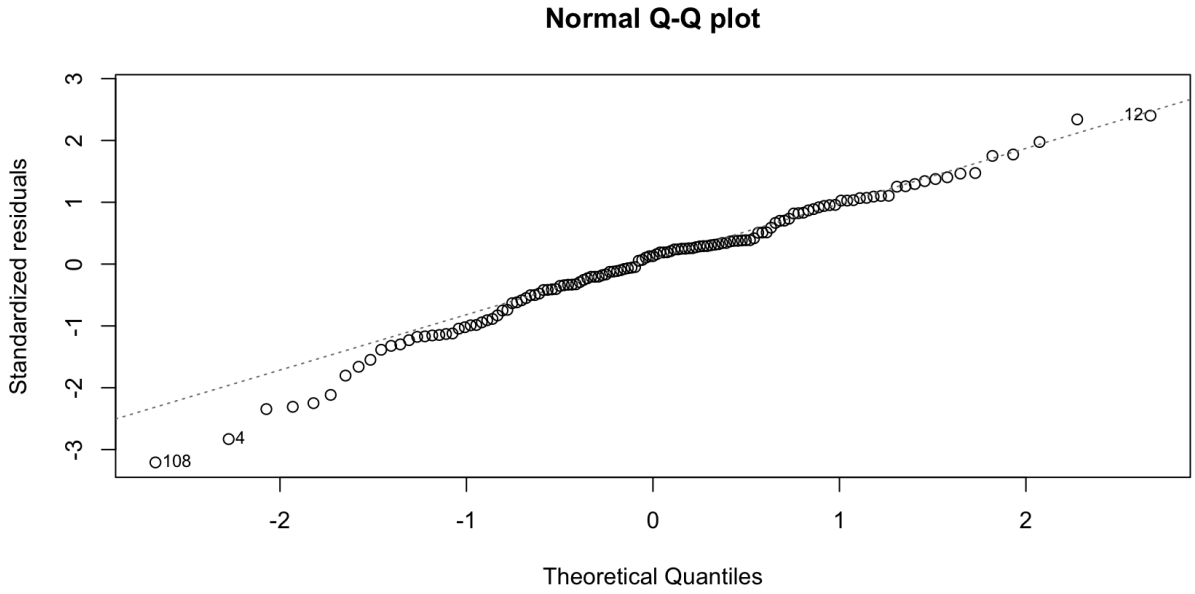

***Figure 9**: Q-Q plot for final model*

Conclusion

Results and Interpretation

We aim to investigate how total expenditure and GDP affect life expectancy across countries in 2014. The multiple linear regression model suggests that both total expenditure and GDP (log-transformed) have statistically significant effects on life expectancy across countries in 2014. Based on the multiple linear regression model, GDP (log-transformed) has a significant negative association with life expectancy ($\beta = -1.1222$, $p = 0.0401$). Expenditure on health as a percentage of total government spending shows positive effects over life expectancy, but all slightly insignificant ($\beta = 0.2645$, $p = 0.0630$). Expenditure on health as a percentage of GDP per capita(log-transformed) also has a positive effect on life expectancy with marginally insignificant, ($\beta = 0.8571$, $p = 0.0918$). The strongest predictor is income composition, with a large positive impact and very high statistical significance ($\beta = 51.4139$, $p < 2e{-}16$). The model demonstrates strong explanatory power with an adjusted R^2 of 0.804, indicating that over 80% of the variability in life expectancy is explained by the model.

Discussion

In this study, we have a very interesting and counter-intuitive result: GDP has a significant negative association with life expectancy. This could be due to several reasons:

1. In very high-GDP countries, increase in GDP may no longer correspond to health improvements. These countries might face problems like aging populations, or lifestyle-related diseases (e.g., obesity, alcohol, stress).

2. GDP doesn't necessarily equal health related investment. GDP measures the size of the economy, not how money is spent, and countries with high GDP may not prioritize public health. Factors like expenditure on health as a percentage of total government spending and expenditure on health as a percentage of GDP per capita may better reflect investment in health.

Limitations

1. **Observational Data, Not Causal**
   This study relies on observational data from 2014, which limits the ability to make causal inferences. While associations are identified, we cannot determine whether these variables directly cause changes in life expectancy.

   Proposed solution: The use of a more complex study that uses data collected over many years. Instead of looking at just one year, this type of data would let us see how changes in GDP or health spending over time affect life expectancy in each country and provide stronger arguments as to whether they have an effect.

2. **Ambiguity in Interpretation of Health Expenditure**
   Total health expenditure is used to represent a country's investment in public health, it may mask important differences in how that money is allocated. For example, countries that spend more on accessible public healthcare infrastructure (e.g., primary care, preventive services) may see greater gains in life expectancy than those that spend similar amounts on specialized research

   Proposed solution:
   A more detailed breakdown of spending categories (e.g., % on preventive care, staff wages, medical technology, or public outreach) would provide a clearer understanding of which aspects of health investment most strongly influence life expectancy.

3. **Outdated Data**
   Since the dataset is from 2014, the relationships identified may not reflect current global health and economic conditions. Significant changes over the past decade (e.g., COVID-19, economic shifts) may affect the validity of these findings today.

   Proposed solution: The study could be repeated using more recent data (e.g., post-2020) to validate whether the relationships observed still hold true under present conditions.

Potential Future Research

Future research could explore how other factors influence life expectancy over time. For instance, we could study how health factors affect life expectancy or introduce interaction terms to see how our analysis changes.

Appendix

A. Full list of explanatory variables

- country: The country from which the data was collected
- year: The year from which the data was collected
- a_mortal: Mortality rates for both sexes aged 15-60 per 1000 population
- i_mortal: Number of infant deaths per 1000 population
- under5_mortal: Number of deaths under-five per 1000 population
- alcohol: Alcohol consumption per capita in litres of pure alcohol
- hepatitis: Percentage of Hepatitis B immunization coverage among 1-year-olds
- polio: Percentage of Polio immunization coverage among 1-year-olds
- diphtheria: Percentage of Diphtheria immunization coverage among 1-year-olds
- measles: Number of reported cases of measles per 1000 population
- bmi: Average body mass index of entire population
- hiv: Deaths of children aged 0-4 due to HIV/AIDS per 1000 live births
- pop: Population of the country
- thin1_19: Percentage of thinness among those aged 10-19
- thin5_9: Percentage of thinness among children aged 5-9
- schooling: Number of years of schooling

B. Heatmap for all numerical variables + Correlation between covariates and response