How economic, health and social factors affect life expectancy in developing countries?

Siddharth Gowda Syed Hasan Arshh Relan

April 2, 2025

Table of contents

1 Contributions

Siddharth Gowda: Cleaning data, generating tables and plots, plot interpretation, data description writing, preliminary model writing.

Syed Hasan: Research to find 3 peer reviewed articles, research question, Data description writing, Preliminary model writing and plot interpretation.

Arshh Relan: Introduction writing, Data Description (Predictor Variables), Preliminary Model Interpretation and Writing, Research Question.

2 Introduction

Life expectancy is a fundamental indicator of a country's overall health and development. It reflects the effectiveness of healthcare systems, economic stability, and social progress. In developing countries healthcare resources are often constrained and economic disparities persist. Here, understanding the factors that influence life expectancy is critical for shaping effective policies. This study aims to answer the question: How economic, health and social factors influence life expectancy in developing countries, collectively and individually?

A multiple linear regression model is used, incorporating GDP per capita, alcohol consumption (categorical variable), schooling, infant mortality, health expenditure (as a percentage of a GDP per capita), and HIV deaths as explanatory variables. Existing literature provides strong evidence supporting the relationship between these factors and life expectancy. Miladinov (2020) examined the link between socioeconomic development and life expectancy in EU accession candidate countries, finding that higher GDP per capita and lower infant mortality rates were associated with increased life expectancy. Our study focuses on developing nations facing similar economic and healthcare challenges.

Adebayo et al. (2024) analyzed data in the United States, finding higher health expenditure and GDP per capita positively impact life expectancy while higher infant mortality reduces it. Our study extends this analysis to developing countries, allowing us to compare how these determinants function in different economic contexts. Our study provides a broader perspective on additional health risks that impact life expectancy in low-income nations.

A third study on OECD countries (Roffia et al., 2022) found that education and economic conditions played a critical role in increasing life expectancy. This supports our inclusion of schooling as a key predictor, as education directly influences health outcomes and economic mobility.

Life expectancy is influenced by a combination of factors, making a univariate approach insufficient. Multiple linear regression (MLR) is well-suited for this study as it quantifies the relationship between life expectancy (continuous response variable) and multiple predictors. MLR allows for simultaneous control of confounding variables, ensuring each predictor's effect is estimated independently. The model provides interpretable coefficients allowing us to measure each socioeconomic and health factor's impact on a developing country's life expectancy. An interpretation of these results can help developing countries understand how to improve life expectancy.

3 Data Description

The dataset used was obtained from Kaggle (Rajarshi 2018). The data-authors compiled WHO and UN datasets which collect health, economic, and demographic data over multiple years from national agencies and censuses. The purpose was to inform policy-makers regarding individual variables impacting life expectancy in all countries. However, our paper aims to create a recipe for developing countries to improve life expectancy.

Before analysis, rows missing values were removed. Alcohol consumption was converted into a categorical variable due to legal restrictions in some countries. Using raw numerical values could be misleading as zero consumption may not indicate abstinence but a ban. The alcohol categories are: **Negligible** ($\leq$ 1.08 liters/capita) and **Consumed** (>1.08 liters/capita).

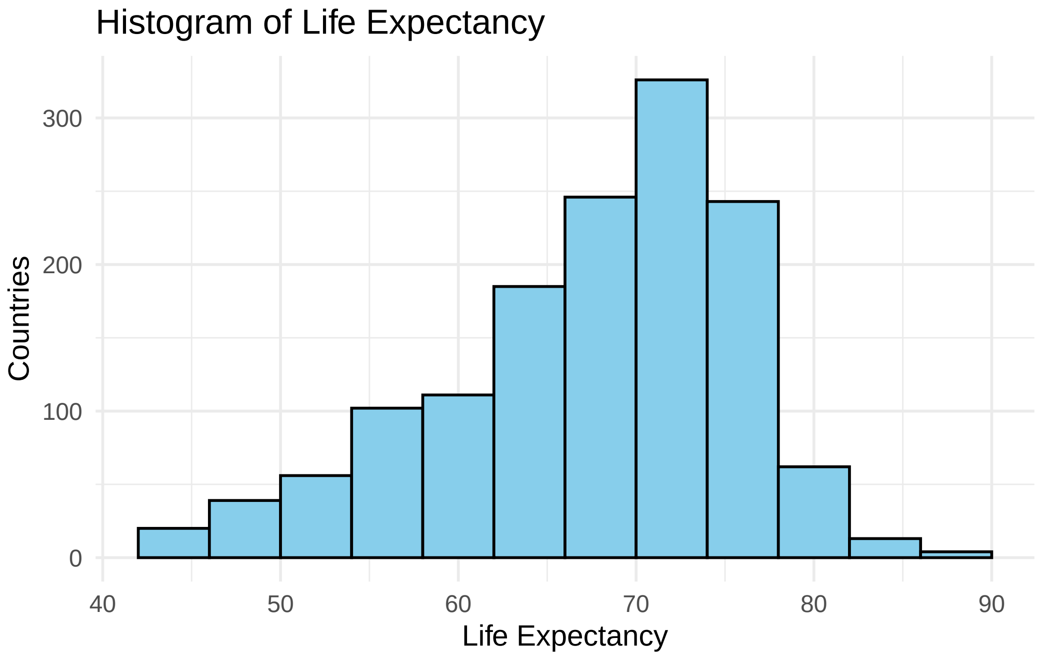


Figure 1

The response variable, life expectancy, represents the average years a newborn is expected to live under current conditions. It is continuous, making it suitable for multiple linear regression, as it is influenced by various economic and health-related factors. The distribution is slightly left-skewed, with a median of 69.2 and a standard deviation of 8.35.

3.1 Predictor Variables

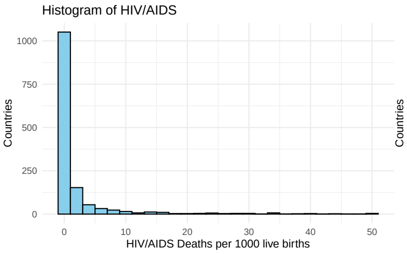


Figure 2

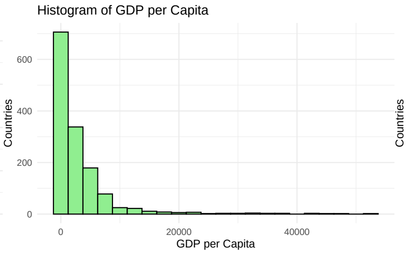


Figure 3

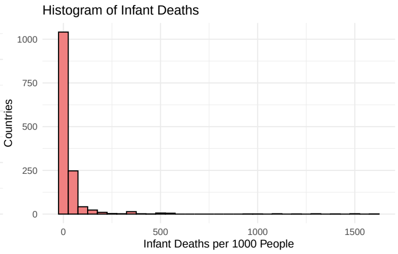


Figure 4

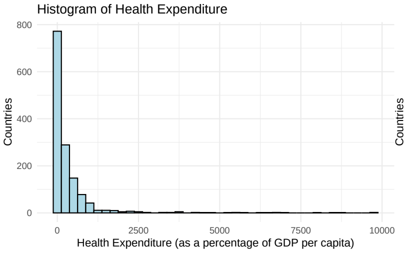


Figure 5

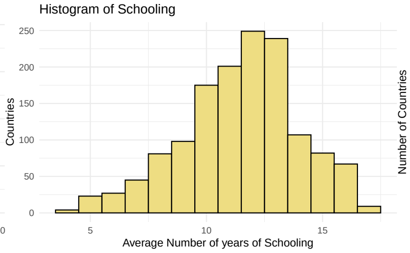


Figure 6

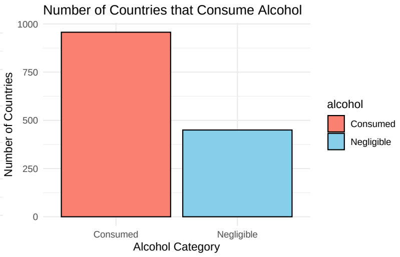


Figure 7

Most variables (HIV, GDP per capita, Infant Deaths, Health Expenditure) are heavily left-skewed with long outlier-tails, while schooling is slightly left-skewed, centered around 12 years. Most countries consume alcohol.

The selected predictor variables are supported by prior research. Miladinov (2020) found that higher GDP per capita is linked to better health and longer life expectancy (Adebayo et al., 2024). Health expenditure reflects a government's commitment to healthcare. Roffia et al. (2023) emphasized that investments in health and social care systems are key drivers of longevity. Infant mortality rate is a widely used indicator of healthcare quality and social well-being, where reductions contribute to longer life expectancy (Adebayo et al., 2024). Schooling represents human capital investment, influencing health awareness and healthcare access. HIV prevalence and alcohol consumption serve as public-health risk indicators, potentially reducing longevity (Miladinov 2020).

4 Preliminary Results

4.1 Preliminary Residual Plots

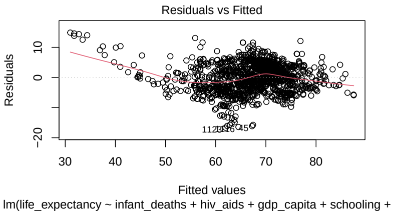


Figure 8

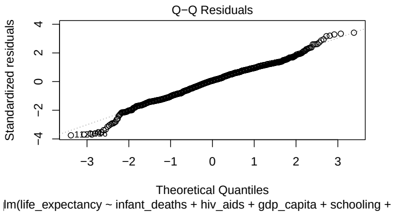


Figure 9

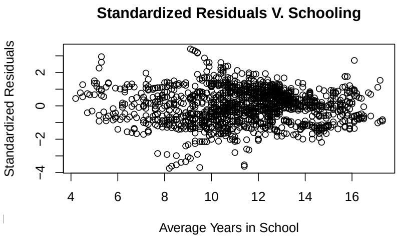


Figure 10

Residual analysis revealed several assumption violations. The **residual vs. fitted plot** showed a wave-like pattern, violating linearity, and demonstrated non-constant variance, particularly in the 60-70 year range. The **QQ plot** showed mostly normal residuals, but there was deviation at the tails, indicating minor normality violations. The **standardized residuals vs. schooling plot** showed no residual patterns and no significant violations of linearity or constant variance, suggesting schooling's relationship with life expectancy is appropriately modeled.

4.2 Preliminary Model Results

$$\begin{aligned}\widehat{\text{LifeExpectancy}} = 49.50 & \\ & - 0.0021 \times \text{InfantDeaths} \\ & - 0.6713 \times \text{HIVAIDSDeaths} \\ & + 7.542 \times 10^{-6} \times \text{GDPperCapita} \\ & + 1.701 \times \text{AverageSchoolingYears} \\ & + 0.00109 \times \text{HealthPercentageExpenditure} \\ & - 0.670 \times \text{AlcoholNegligible}\end{aligned}$$

Variable	Estimate	Std. Error	t Value	Pr(>|t|)
(Intercept)	49.5000	0.7476	66.216	< 2e-16
Infant Deaths	-0.0021	0.0009	-2.274	0.02309
HIV/AIDS	-0.6713	0.0192	-35.018	< 2e-16
GDP per Capita	0.0000	0.0001	0.137	0.89121
Schooling	1.7010	0.0613	27.757	< 2e-16
Percentage Expenditure	0.0011	0.0004	3.025	0.00253
Alcohol (Negligible)	-0.6700	0.2859	-2.343	0.01924

Our initial model examined five predictors of life expectancy, with schooling emerging as a significant predictor (coefficient = 1.701), indicating each additional year of schooling increases life expectancy by approximately 1.7 years, holding other variables constant. Surprisingly, GDP per capita showed no significant impact (coefficient ~ 0), possibly due to its effects being captured by other predictors such as education and healthcare expenditure. The **categorical predictor** of alcohol consumption yielded unexpected results, with negligible consumption associated with a 0.67-year decrease in life expectancy compared to regular consumption. This counterintuitive finding might correlate with higher socioeconomic status in some contexts, leading to better healthcare access and increased longevity.

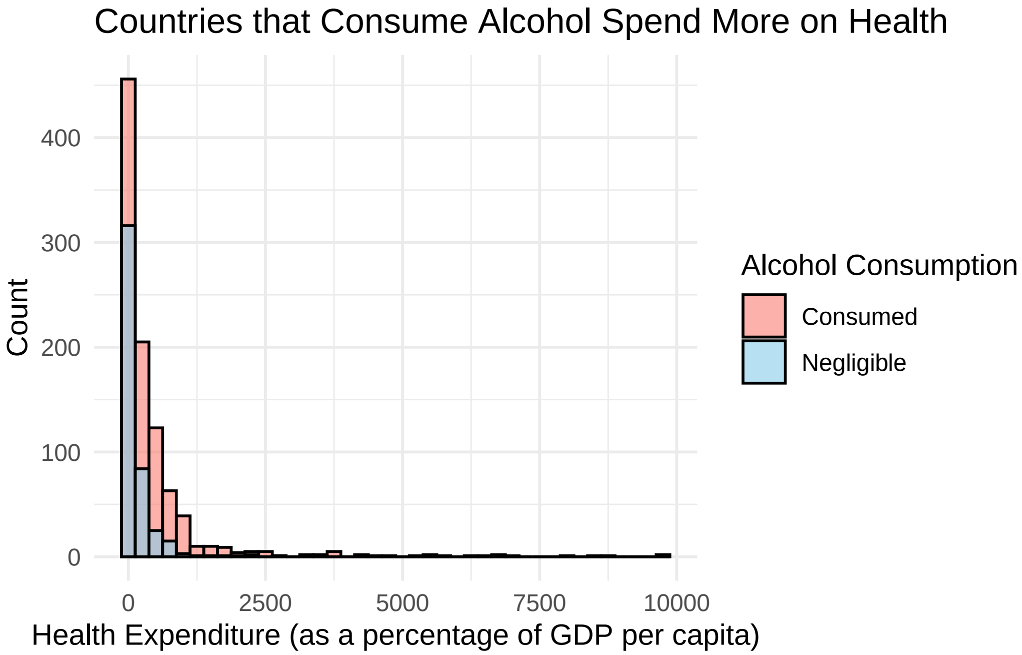


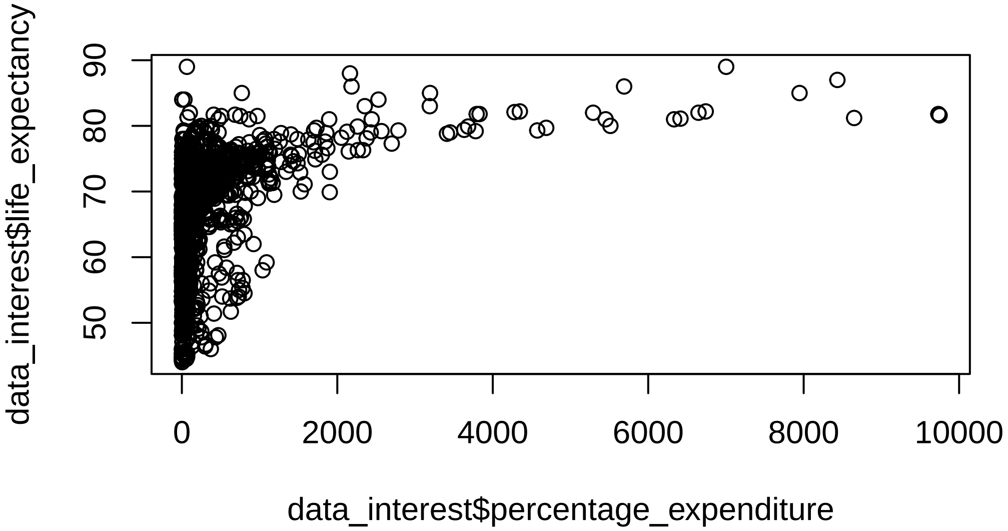
Figure 11

4.2.1 Literature Connection

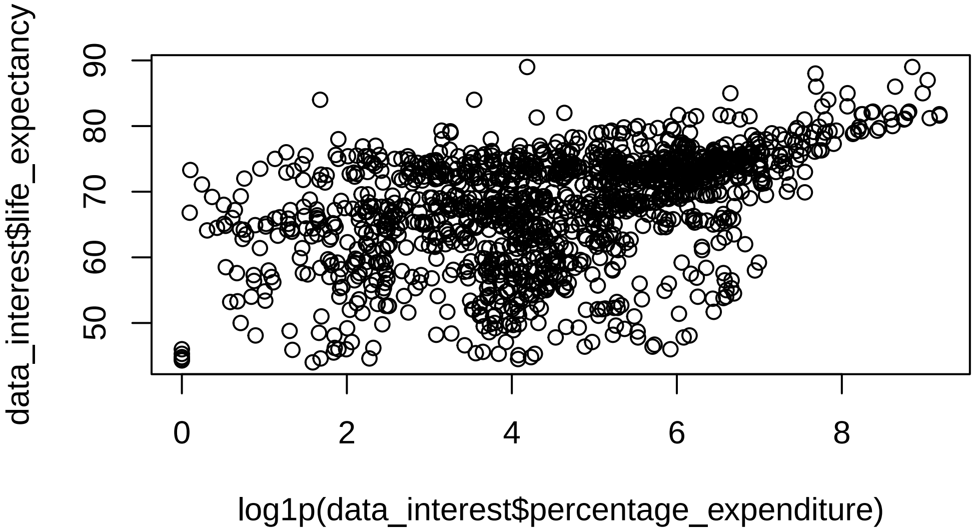
Our findings partially align with existing literature. Like Miladinov (2020), we found significant negative effects of infant mortality and HIV prevalence on life expectancy. Our results support the OECD study's emphasis on education's importance, showing a strong positive relationship between schooling and longevity. Adebayo et al. (2024) highlighted health expenditure's positive impact, which our model confirms.Our findings diverge regarding GDP per capita's significance. Both Miladinov (2020) and Adebayo et al. (2024) found GDP per capita to be a key predictor, contrary to our results. Additionally, our finding on alcohol consumption's relationship with life expectancy contradicts Miladinov's research, warranting further investigation.

5 Model Selection

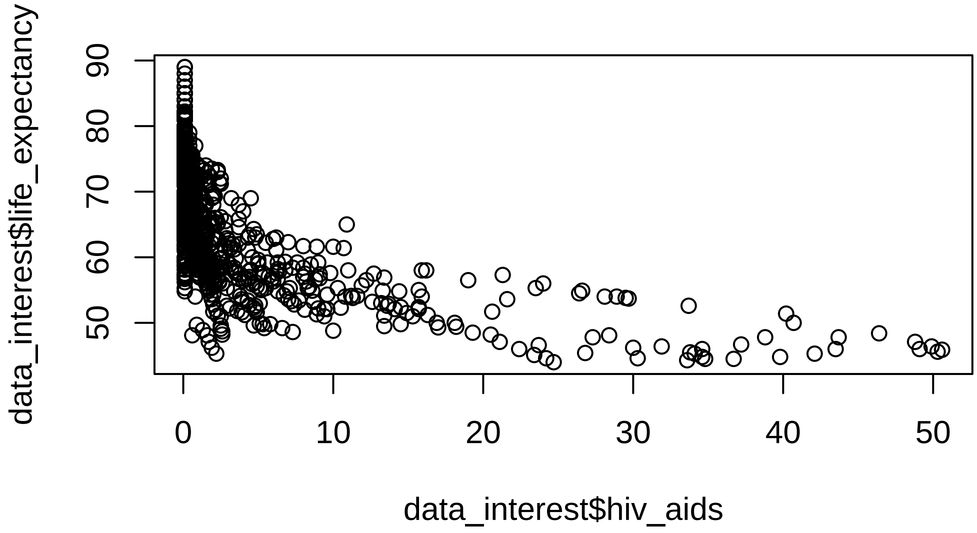
```
plot(data_interest$percentage_expenditure, data_interest$life_expectancy)
```



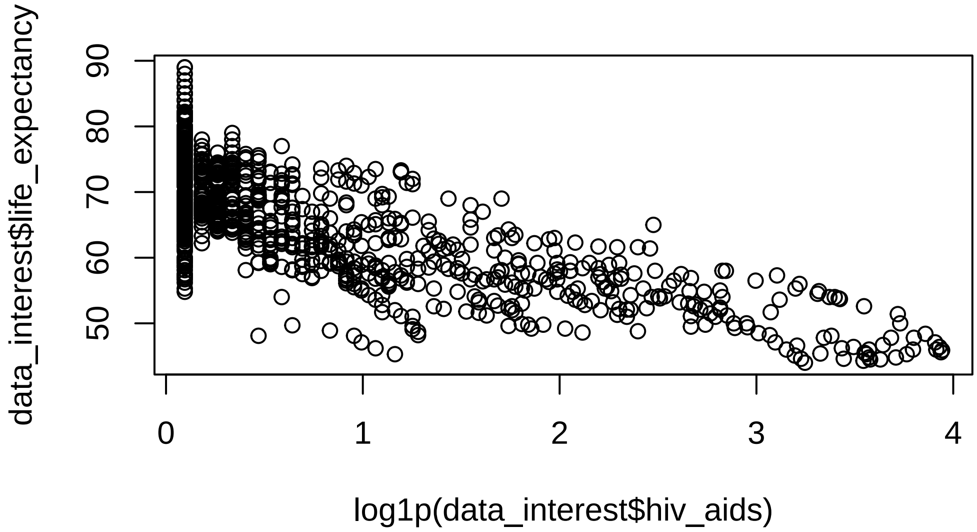
```
plot(log1p(data_interest$percentage_expenditure), data_interest$life_expectancy)
```



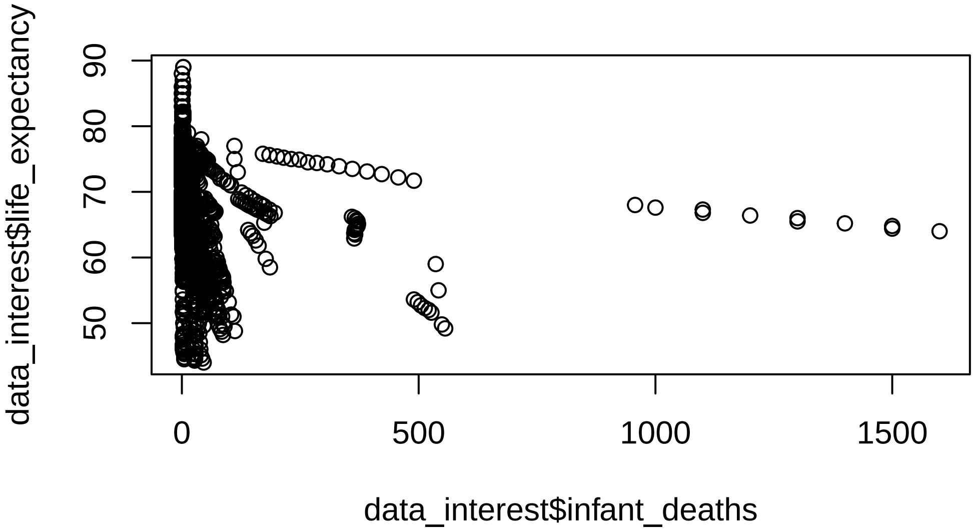
```
plot(data_interest$hiv_aids, data_interest$life_expectancy)
```



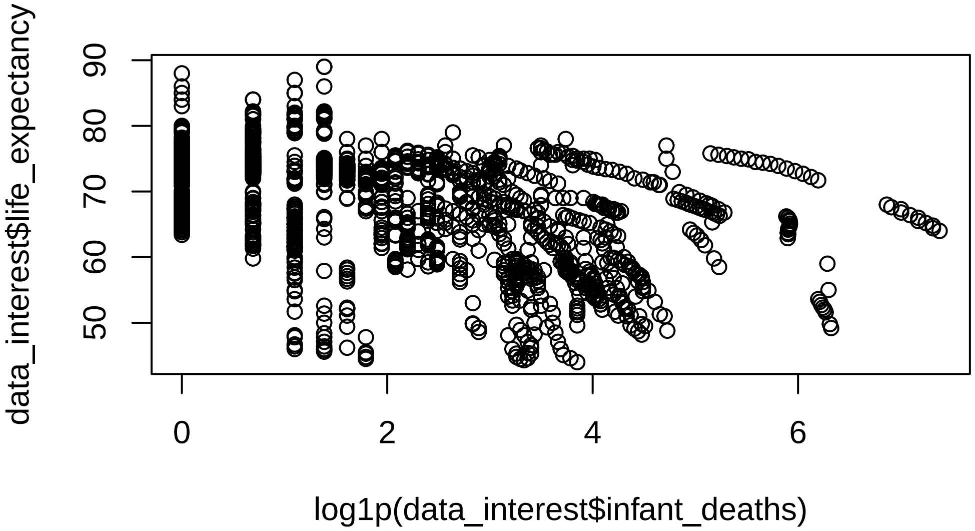
```
plot(log1p(data_interest$hiv_aids), data_interest$life_expectancy)
```



```
plot(data_interest$infant_deaths, data_interest$life_expectancy)
```



```
plot(log1p(data_interest$infant_deaths), data_interest$life_expectancy)
```



```
data_interest_final <- data_interest %>%
  select(-gdp_capita) %>%
  mutate(
    infant_deaths_lg1p = log1p(infant_deaths),
    infant_deaths_cat = case_when(
      infant_deaths_lg1p < 2 ~ "low",
      TRUE ~ "high"
    ),
    hiv_aids_lg1p = log1p(hiv_aids),
```

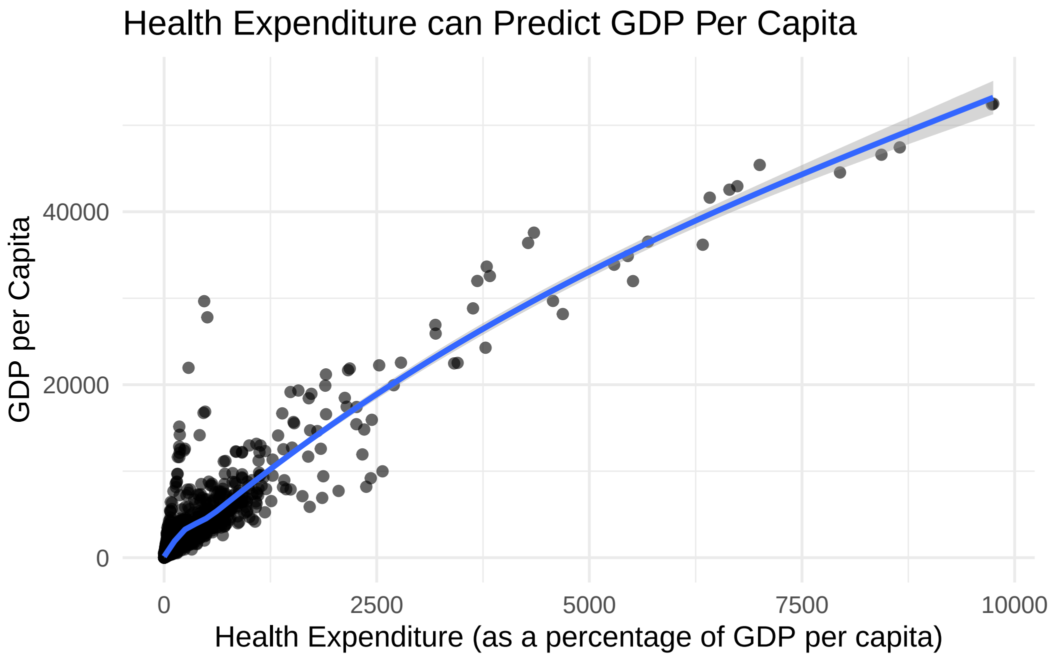


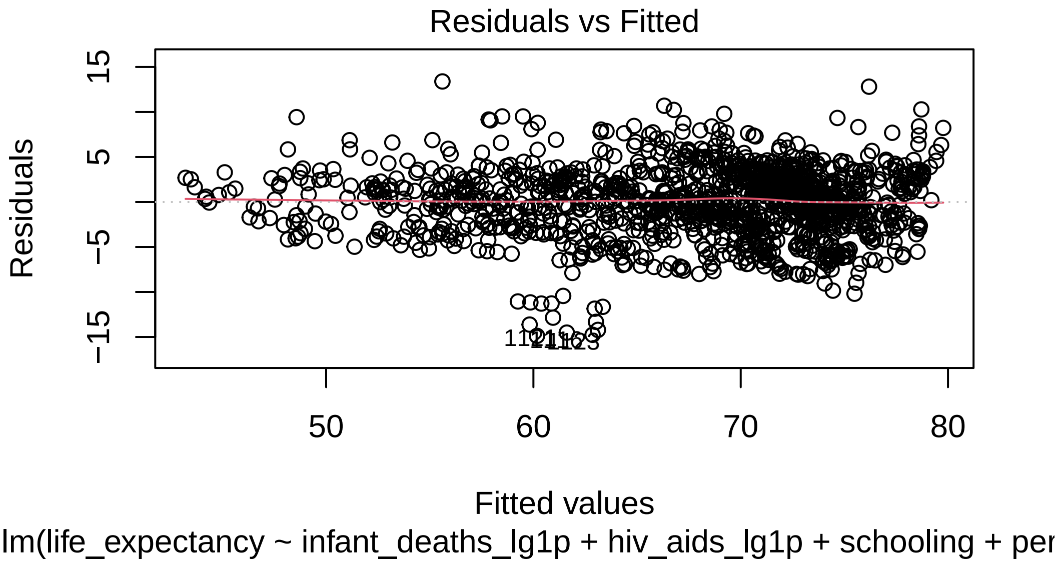
Figure 12

```
    percentage_expenditure_lg1p = log1p(percentage_expenditure)
    ) %>% select(life_expectancy, infant_deaths_lg1p, hiv_aids_lg1p,
                         schooling, percentage_expenditure_lg1p, alcohol)

secondary_model_infant_lg = lm(life_expectancy ~ infant_deaths_lg1p + hiv_aids_lg1p +
                         schooling + percentage_expenditure_lg1p + alcohol,
                       data=data_interest_final)
summary(secondary_model_infant_lg)
```

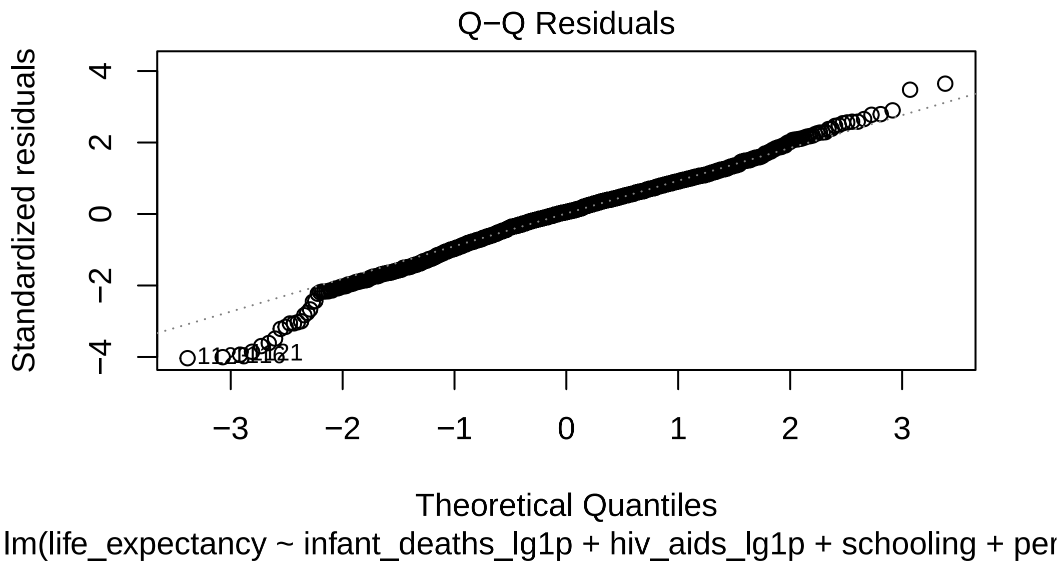
```
Call:
lm(formula = life_expectancy ~ infant_deaths_lg1p + hiv_aids_lg1p + 
    schooling + percentage_expenditure_lg1p + alcohol, data = data_interest_final)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8723  -2.2165   0.2026   2.3337  13.3893 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 57.86168    0.72966  79.300  < 2e-16 ***
```

```
infant_deaths_lg1p          -0.37360    0.06815  -5.482 4.99e-08 ***
hiv_aids_lg1p               -6.22788    0.13066 -47.663  < 2e-16 ***
schooling                    1.06694    0.05544  19.244  < 2e-16 ***
percentage_expenditure_lg1p  0.54056    0.06703   8.064 1.57e-15 ***
alcoholNegligible           -1.51114    0.23922  -6.317 3.57e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.693 on 1401 degrees of freedom
Multiple R-squared:  0.8053,    Adjusted R-squared:  0.8046 
F-statistic:  1159 on 5 and 1401 DF,  p-value: < 2.2e-16
```

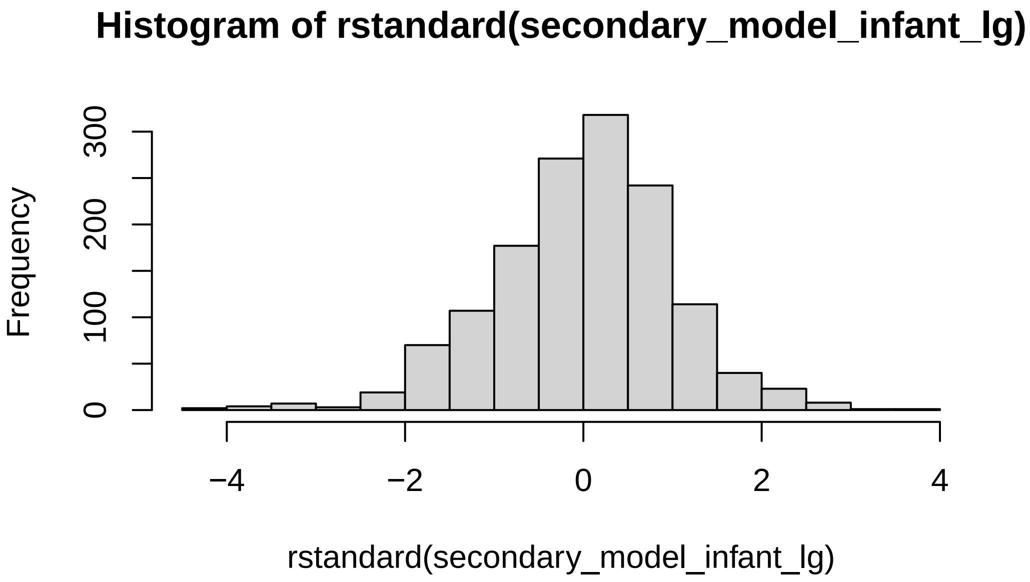
```r
plot(secondary_model_infant_lg, which = 1)
```



```r
plot(secondary_model_infant_lg, which = 2)
```



```
hist(rstandard(secondary_model_infant_lg))
```



```
# hist(rstandard(preliminary_model))
```

6 Final Model & Results

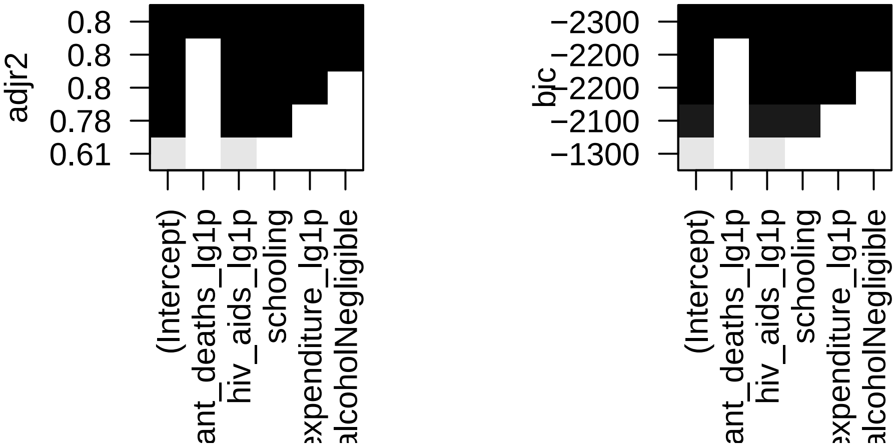
$$\begin{aligned}\widehat{\text{LifeExpectancy}} = 57.86 & \\ & - 0.374 \times \text{InfantDeaths\_lg1p} \\ & - 6.228 \times \text{HIV\_AIDS\_lg1p} \\ & + 1.067 \times \text{Schooling} \\ & + 0.541 \times \text{PercentageExpenditure\_lg1p} \\ & - 1.511 \times \text{AlcoholNegligible}\end{aligned}$$

```r
# Create the initial data frame with regression results
results <- data.frame(
  Term = c("(Intercept)", "infant_deaths_lg1p", "hiv_aids_lg1p", "schooling",
           "percentage_expenditure_lg", "alcoholNegligible"),
  Estimate = c(57.86168, -0.37360, -6.22788, 1.06694, 0.54056, -1.51114),
  Std_Error = c(0.72966, 0.06815, 0.13066, 0.05544, 0.06703, 0.23922),
  t_value = c(79.300, -5.482, -47.663, 19.244, 8.064, -6.317),
  p_val = c("< 2e-16", "4.99e-08", "< 2e-16", "< 2e-16", "1.57e-15", "3.57e-10")
)

# Calculating 95% confidence intervals
n <- 1407
p <- 5

df <- n - p - 1

alpha <- 0.05

t_crit <- qt(1 - alpha/2, df)

margin_of_error <- t_crit * results$Std_Error

lower_bound <- results$Estimate - margin_of_error
upper_bound <- results$Estimate + margin_of_error

results$UpperBound95 <- lower_bound
results$LowerBound95 <- upper_bound

kable(results, col.names = c("Term", "Estimate", "Std. Error", "t value", "Pr(>|t|)",
                             "95% CI Lower", "95% CI Upper"),
      caption = "Final Regression Results")
```

Table 2: Final Regression Results

Term	Estimate	Std. Error	t value	Pr(>|t|)	95% CI Lower	95% CI Upper
(Intercept)	57.86168	0.72966	79.300	< 2e-16	56.4303361	59.293024
infant_deaths_lg1p	-0.37360	0.06815	-5.482	4.99e-08	-0.5072870	-0.239913
hiv_aids_lg1p	-6.22788	0.13066	-47.663	< 2e-16	-6.4841903	-5.971570
schooling	1.06694	0.05544	19.244	< 2e-16	0.9581856	1.175694
percentage_expenditure_lg1p	0.54056	0.06703	8.064	1.57e-15	0.4090700	0.672050
alcoholNegligible	-1.51114	0.23922	-6.317	3.57e-10	-1.9804080	-1.041872

```
stepwise_regression <- regsubsets(life_expectancy ~ ., data = data_interest_final,
                                  nvmax = 10, nbest = 1, really.big = TRUE, method = "seqrep

par(mfrow = c(1,2))

plot(stepwise_regression, scale = "adjr2")
plot(stepwise_regression, scale = "bic")
```



7 Bibliography

Miladinov, G. Socioeconomic development and life expectancy relationship: evidence from the EU accession candidate countries. Genus 76, 2 (2020). https://doi.org/10.1186/s41118-019-0071-0

Adebayo, T. S., Nwosu, L. C., Alhassan, G. N., Uzun, B., Özkan, O., & Awosusi, A. A. (2024). Effects of health expenditure, death rate, and infant mortality rate on life expectancy: A case study of the United States. Energy & Environment, 0(0). https://doi-org.myaccess.library.utoronto.ca/10.1177/0958305X241281804

Roffia, P., Bucciol, A. & Hashlamoun, S. Determinants of life expectancy at birth: a longitudinal study on OECD countries. Int J Health Econ Manag. 23, 189–212 (2023). https://doi.org/10.1007/s10754-022-09338-5

Kumar Rajarshi. (2018, February 10). Life expectancy (WHO). Kaggle. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who