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

Siddharth Gowda Syed Hasan Arshh Relan

April 10, 2025

Table of contents

1 Contributions [TODO: UPDATE THIS]

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).

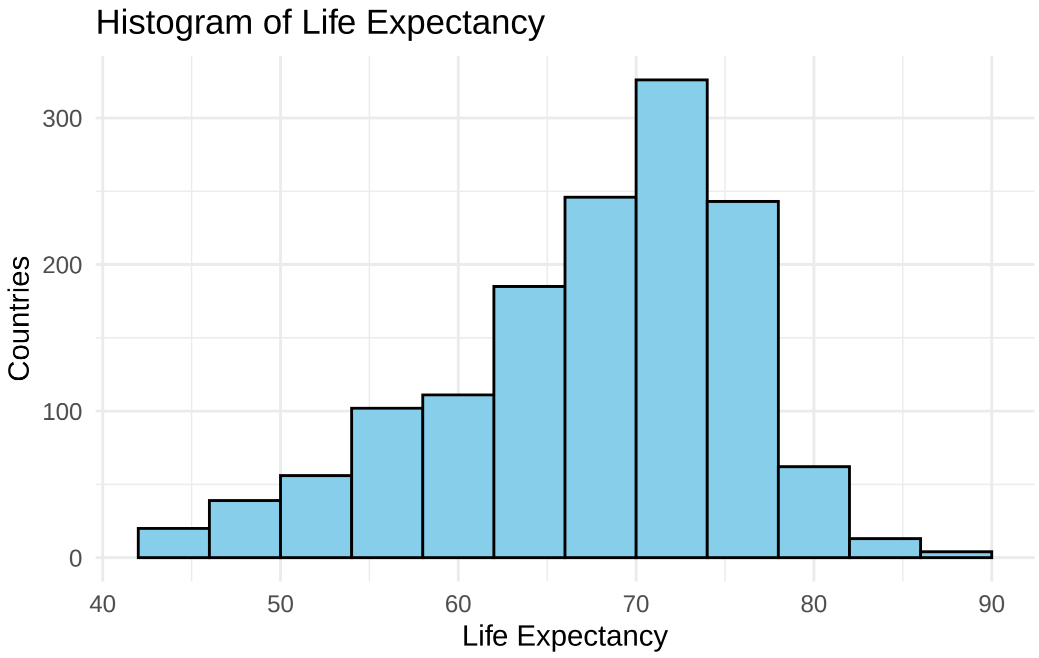


Figure 1

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

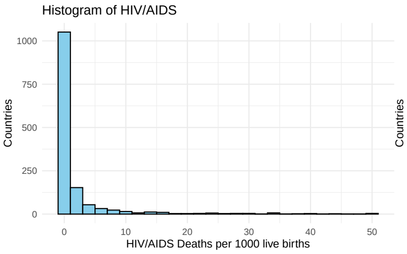


Figure 2

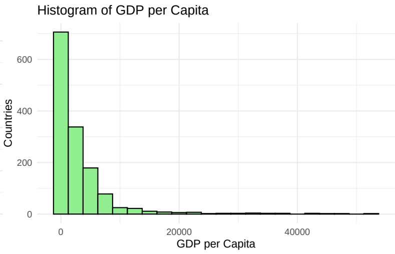


Figure 3

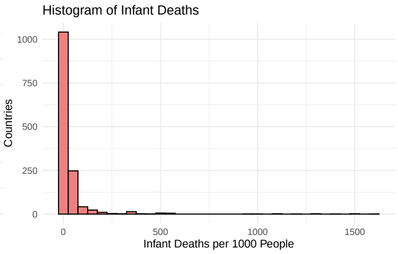


Figure 4

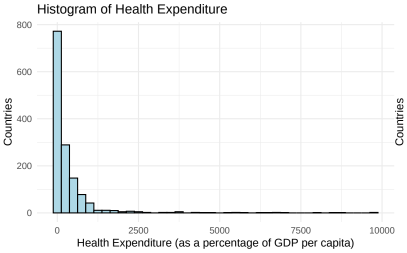


Figure 5

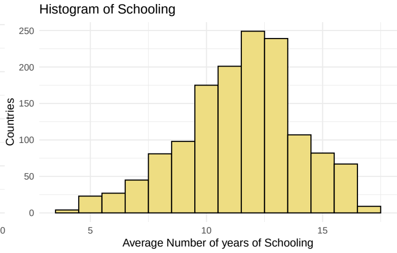


Figure 6

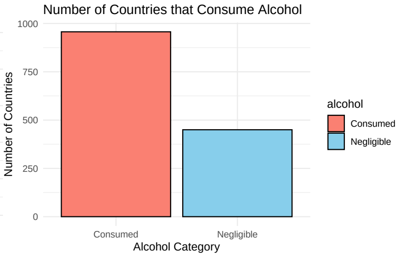


Figure 7

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

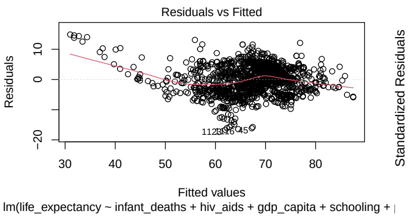


Figure 8

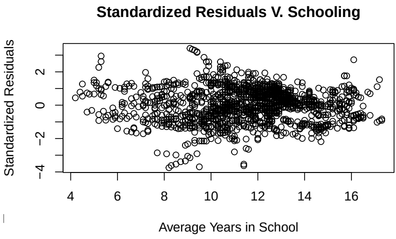


Figure 9

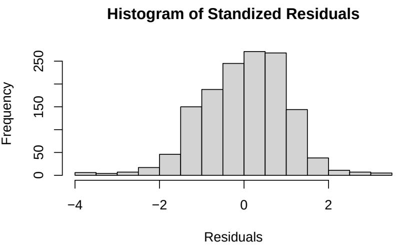


Figure 10

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 **residual histogram plot** showed a slightly slightly skewed left distribution, which is a minor violation of the normal errors assumption. 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

Variable	Estimate	Std. Error	t Value	Pr(>|t|)
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.

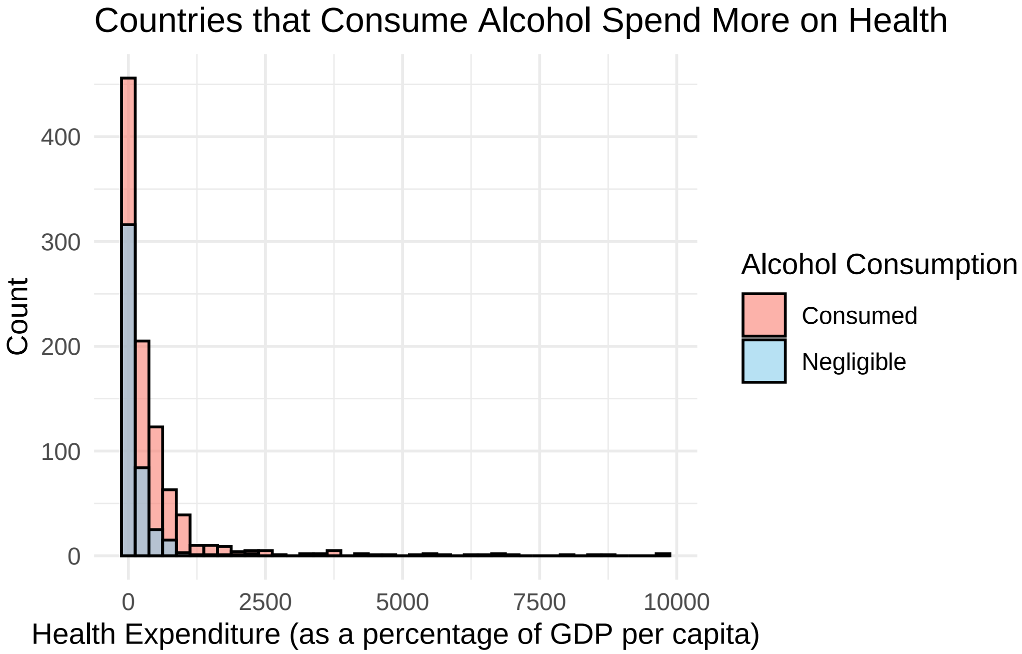


Figure 11

4.2.1 Literature Connection

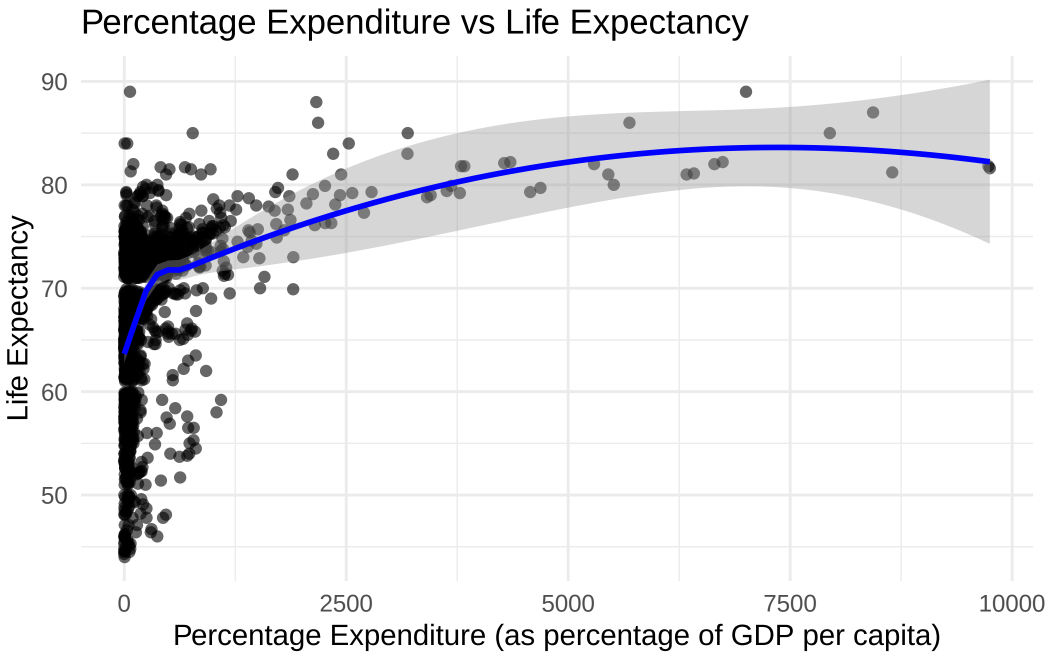
Our findings partially align with existing literature. Like Miladinov (2020), we found significant negative effects of infant mortality and HIV prevalence on life expectancy. Our results support the OECD study's emphasis on education's importance, showing a strong positive relationship between schooling 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

From the preliminary results section, we discovered that a simple linear model that predicts life expectancy for developing countries using the variables of interest had invalid predicting power. The biggest reason for this was because the preliminary model violated linearity. Therefore, we had to make numerous transformations.

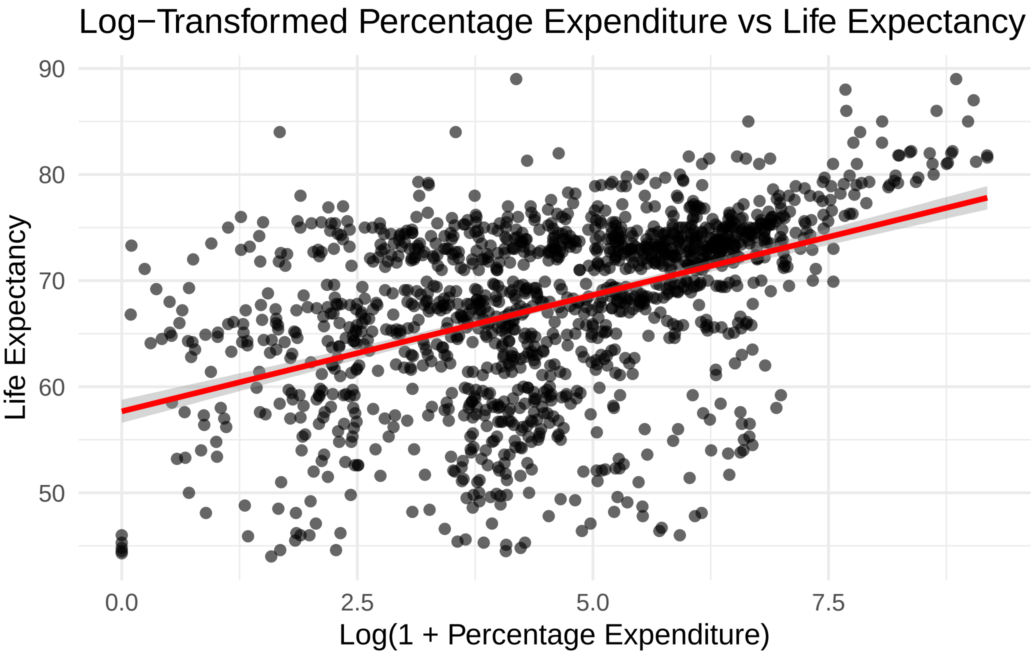
```r
ggplot(data_interest, aes(x = percentage_expenditure, y = life_expectancy)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "loess", se = TRUE, color = "blue") +
  labs(x = "Percentage Expenditure (as percentage of GDP per capita)", y = "Life Expectancy"
       title = "Percentage Expenditure vs Life Expectancy") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```
ggplot(data_interest, aes(x = log1p(percentage_expenditure), y = life_expectancy)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(x = "Log(1 + Percentage Expenditure)", y = "Life Expectancy",
       title = "Log-Transformed Percentage Expenditure vs Life Expectancy") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

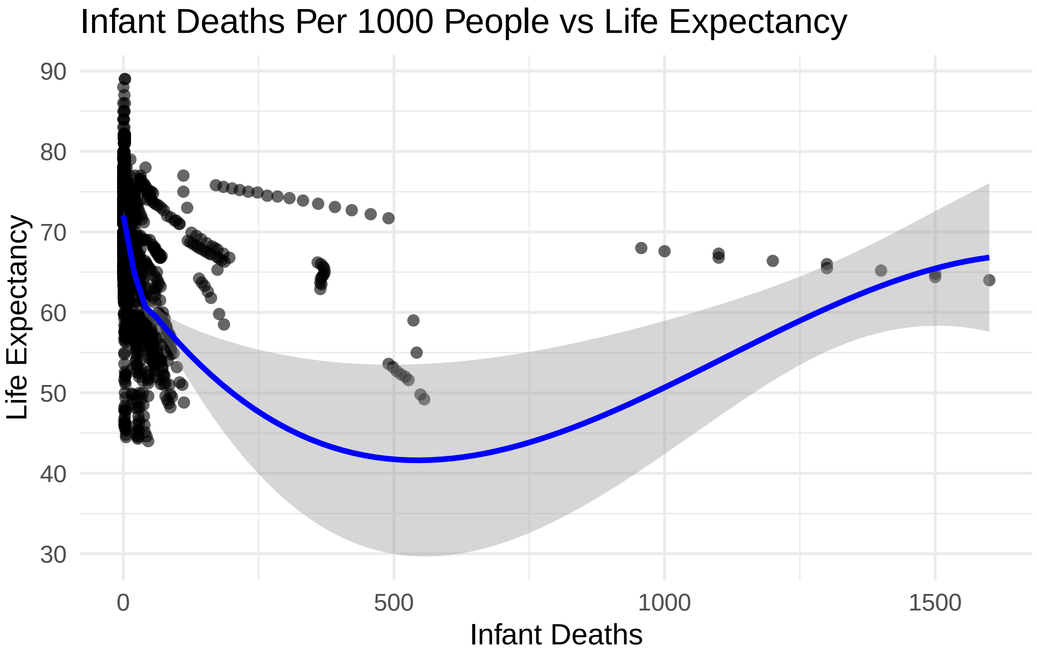


The first transformation we had was with health expenditure as a percentage of GDP per capita. For health expenditure, the figure above reveals a non-linear, positive exponential relationship with life expectancy. To linearize this relationship, we applied a $log(1 + HealthExpenditure)$ transformation as the log function is in the inverse of exponential. We used $log(1 + x)$ rather than $log(x)$ to accommodate the theoretical possibility of zero expenditure values since $log(0)$ is undefined.

The transformation figure demonstrates a moderately strong positive linear relationship between transformed health expenditure and life expectancy, though we note that variance decreases slightly as percentage expenditure increases, indicating some minor constant variance violations.

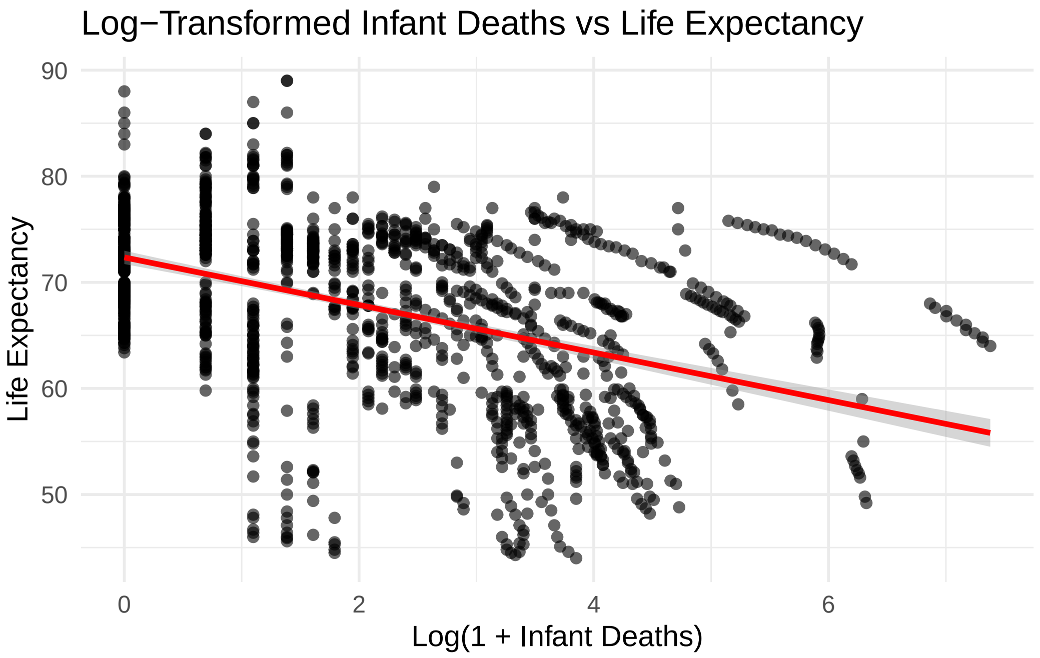
```
ggplot(data_interest, aes(x = infant_deaths, y = life_expectancy)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "loess", se = TRUE, color = "blue") +
  labs(x = "Infant Deaths", y = "Life Expectancy",
       title = "Infant Deaths Per 1000 People vs Life Expectancy") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```
ggplot(data_interest, aes(x = log1p(infant_deaths), y = life_expectancy)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(x = "Log(1 + Infant Deaths)", y = "Life Expectancy",
       title = "Log-Transformed Infant Deaths vs Life Expectancy") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

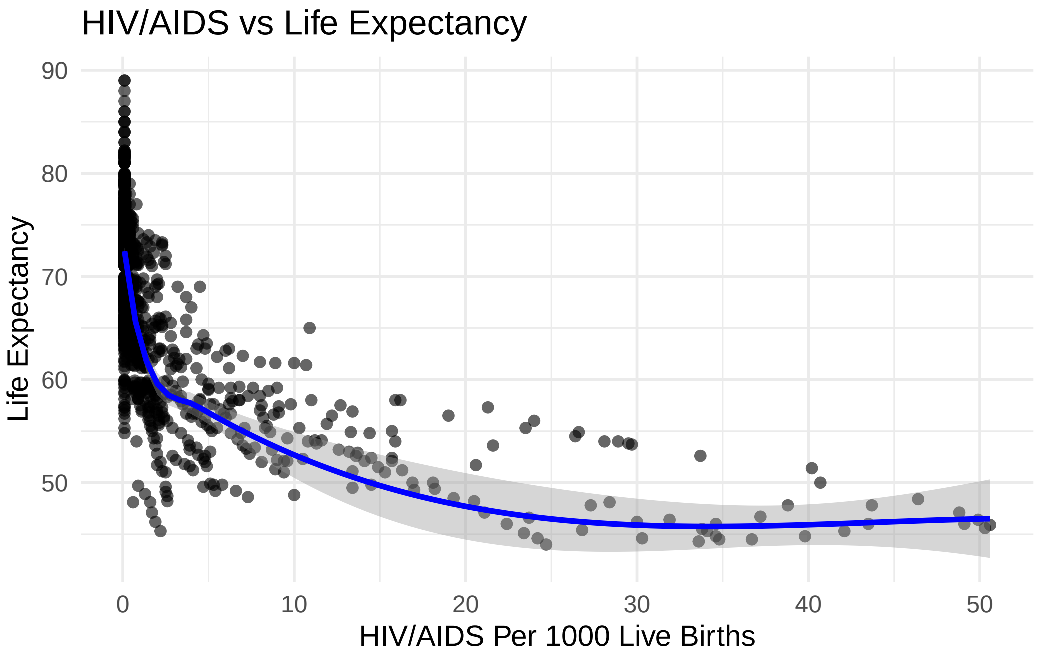


For infant deaths, we applied a $log(1 + \text{infant deaths})$ transformation based on the weak exponential relationship observed in the figure. We used $log(1 + x)$ to maintain a valid domain for potential zero values. The transformed data reveals a generally negative relationship with life expectancy, though with two distinct patterns: higher variability and occasional uniformity between 0-2, and a curved downward slope between 2-6. After testing multiple transformations, $log(1 + x)$ yielded the best results, despite not resulting in a perfectly linear relationship.

We also considered removing outlier countries with over 900 infant deaths per 100 people as shown in the figure. This was done to potentially create a stronger exponential relationship between infant deaths and life expectancy. However, after removing the outliers, the transformation still showed no improved linearity after removal. Furthermore, the final model results remained largely unchanged without these outliers, indicating they were not influential. Therefore, we kept these data points in the final model dataset.

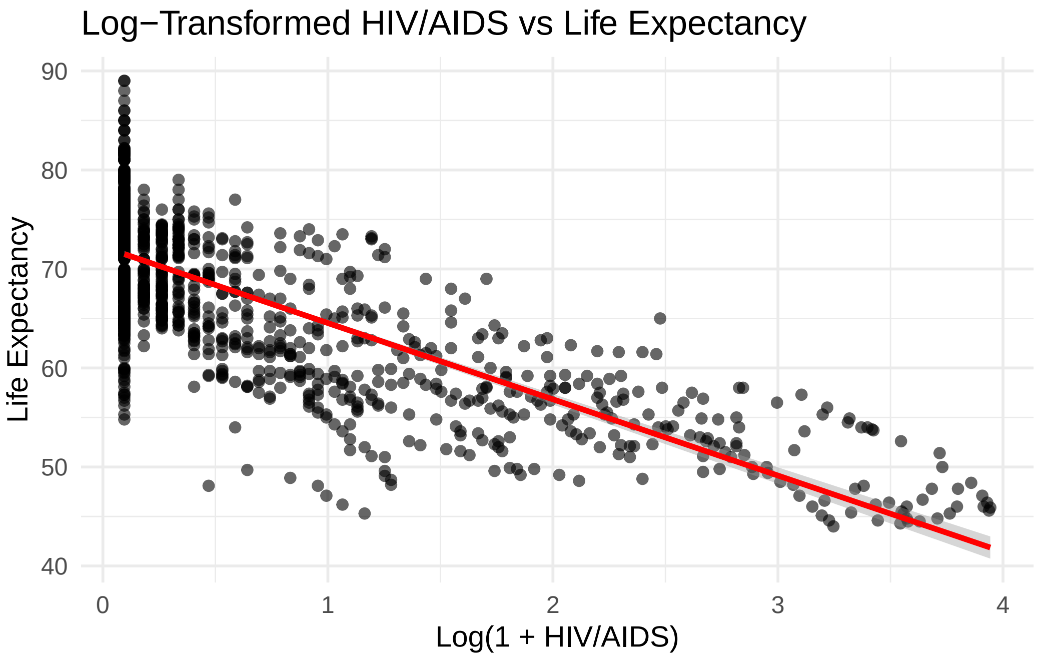
```r
ggplot(data_interest, aes(x = hiv_aids, y = life_expectancy)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "loess", se = TRUE, color = "blue") +
  labs(x = "HIV/AIDS Per 1000 Live Births", y = "Life Expectancy",
       title = "HIV/AIDS vs Life Expectancy") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```
ggplot(data_interest, aes(x = log1p(hiv_aids), y = life_expectancy)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(x = "Log(1 + HIV/AIDS)", y = "Life Expectancy",
       title = "Log-Transformed HIV/AIDS vs Life Expectancy") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



For the third and final transformation, we decided to do a $log(1+\text{HIV/AIDS Deaths})$ transformation. From the figure above, the relationship between aids and life expectancy is negative and exponential so a log transformation is valid. Similarly, $log(1+x)$ is used to ensure a valid domain. From the transformation plot above, $log(1+\text{HIV/AIDS Deaths})$ and life expectancy appear to have a moderately strong negative linear relationship.

In terms of variables removed, we decided to remove GDP per capita because in the preliminary model, its coefficient had a p-value of 0.891, making it redundant. After doing further analysis, we determined that GDP Per Capita is highly correlated with a country's health expenditure (as seen in the figure above). Therefore, it is unnecessary to include GDP per capita in our final model.

6 Final Model Inference & Results

```
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"
    ),
```

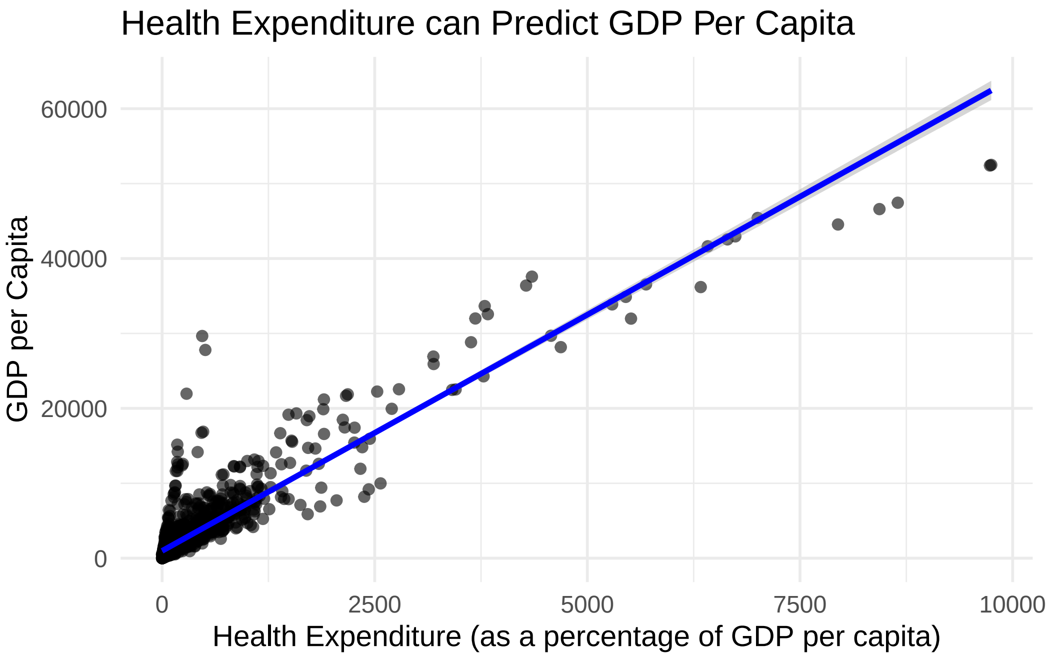


Figure 12

```
    hiv_aids_lg1p = log1p(hiv_aids),
    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
```

The final multiple linear regression model constructed to address the research question: How economic, health and social factors influence life expectancy in developing countries, collectively and individually? It revealed nuanced insights into these relationships.

6.1 Coefficent and Model Interpretation

The regression equation developed from our analysis was as follows:

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

Each regression coefficient offers significant interpretive power concerning how various factors influence life expectancy in the specific context of developing countries.

The intercept (57.86 years) provides a baseline measure of predicted life expectancy, indicating that under conditions of minimal education, health expenditure, negligible mortality from infants or HIV/AIDS, life expectancy would approximate 57.9 years. This baseline highlights the intrinsic life expectancy potential in developing countries, notwithstanding the prevailing adverse conditions.

Among the health-related factors, infant mortality emerges as a crucial determinant of life expectancy. The negative coefficient of -0.374 indicates that each 1 unit increase in infant deaths (after log transformation) results in a decrease of approximately 0.374 years in life expectancy, assuming all other variables are held constant. This finding underscores the critical role infant mortality plays as a reflection of healthcare quality, maternal health services, and broader socioeconomic conditions. This result is consistent with prior literature; Miladinov (2020) similarly identified infant mortality as significantly detrimental to life expectancy within EU accession candidate countries, emphasizing the universal implications of reducing infant mortality as a public health priority. Similarly, Adebayo et al. (2024) found that increased infant mortality significantly diminishes life expectancy in the United States, aligning with the global public health consensus on the relationship between infant health outcomes and longevity.

The factor with the most severe negative impact in our model was HIV/AIDS prevalence, with a coefficient of -6.228, indicating that a 1 unit rise in HIV/AIDS deaths (log-transformed) would significantly lower life expectancy by approximately 6.23 years, assuming all other variables are held constant. This substantial reduction emphasizes the profound effect HIV/AIDS continues to have in developing countries, both through direct mortality and broader health system burdens. The literature corroborates this finding, as prior research consistently highlights HIV/AIDS as a severe constraint on longevity and overall development in low-income settings (Miladinov, 2020).

Education, measured as years of schooling, exhibited a strong positive correlation with life expectancy. Specifically, the coefficient of 1.067 suggests that each additional year of schooling corresponds to an increase of approximately 1.067 years in life expectancy, assuming all other variables are held constant. This significant positive impact is consistent with prior research, such as the findings by Roffia et al. (2023), who reported education as a robust predictor of increased longevity across OECD countries. Enhanced educational attainment likely translates to improved health literacy, better access to healthcare, higher income levels, and healthier lifestyle choices, all collectively enhancing life expectancy. This result demonstrates the critical role educational investment plays in improving health outcomes within developing countries.

Health expenditure, measured as a percentage of GDP per capita, was also positively associated with increased life expectancy. A 1 unit increase in health expenditure as a percentage of GDP per capita (log-transformed) correlates with an approximately 0.54-year increase in life expectancy, assuming all other variables are held constant. This finding aligns with previous studies highlighting healthcare spending as a central determinant of health outcomes and longevity. Both Adebayo et al. (2024) and Roffia et al. (2023) confirm that investments in healthcare infrastructure, preventive medicine, and medical treatment are pivotal in improving overall health conditions and life expectancy. This underscores the importance for policymakers in developing nations to prioritize healthcare expenditure strategically to maximize public health benefits.

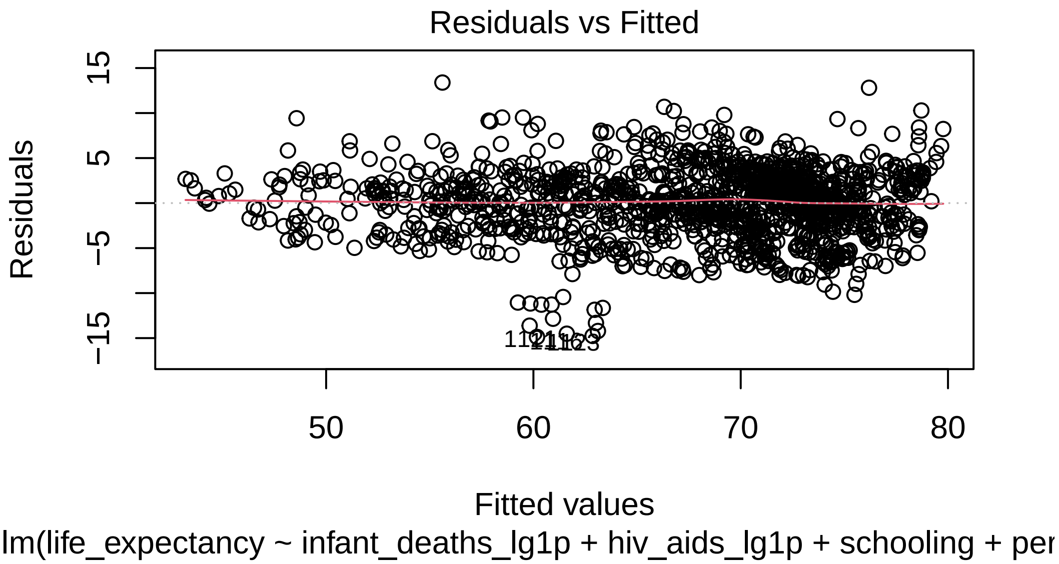
```r
# Create the initial data frame with regression results
results <- data.frame(
  Term = c("(Intercept)", "infant_deaths_lg1p", "hiv_aids_lg1p", "schooling",
           "percentage_expenditure_lg", "alcoholNegligible"),
  Estimate = c(57.86168, -0.37360, -6.22788, 1.06694, 0.54056, -1.51114),
  Std_Error = 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

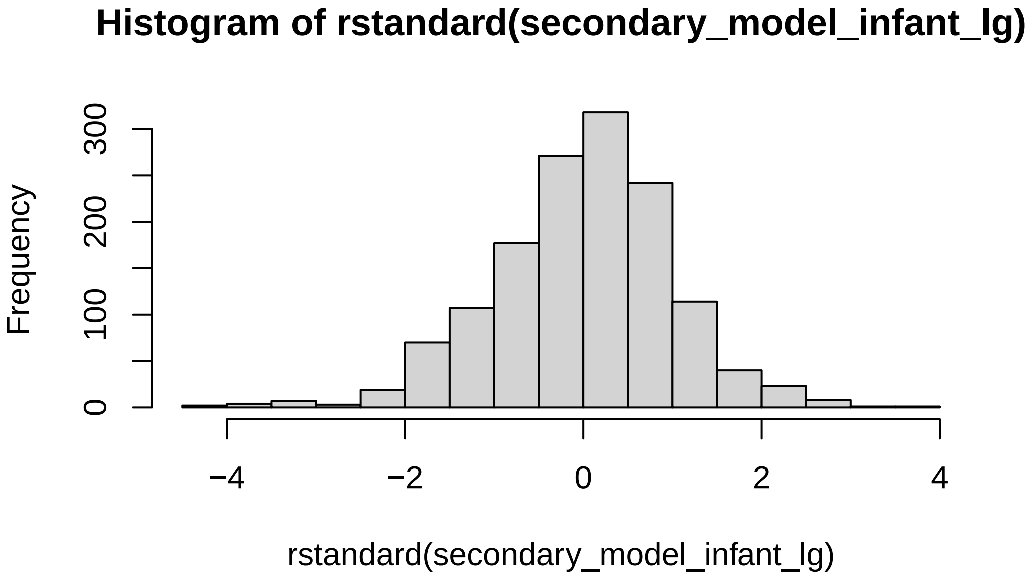
Term	Estimate	Std. Error	t value	Pr(>|t|)	95% CI Lower	95% CI Upper
(Intercept)	57.86168	0.72966	79.300	< 2e-16	56.4303361	59.293024
infant_deaths_lg1p	-0.37360	0.06815	-5.482	4.99e-08	-0.5072870	-0.239913
hiv_aids_lg1p	-6.22788	0.13066	-47.663	< 2e-16	-6.4841903	-5.971570
schooling	1.06694	0.05544	19.244	< 2e-16	0.9581856	1.175694
percentage_expenditure	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

6.2 Model Assumptions

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



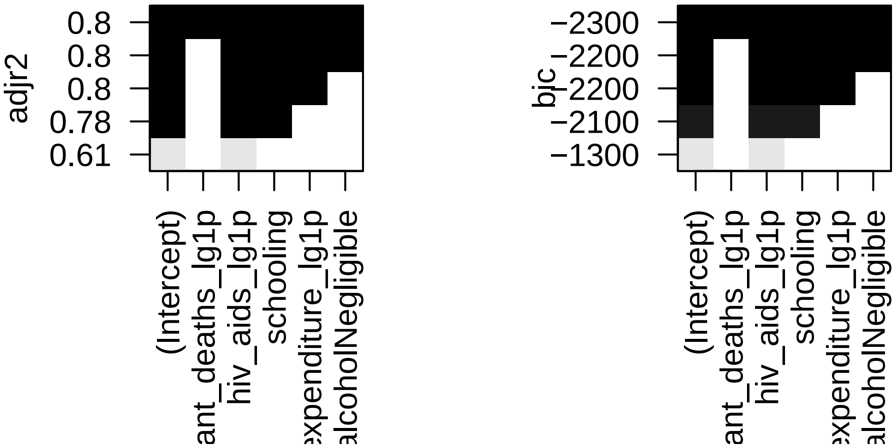
```
hist(rstandard(secondary_model_infant_lg))
```



6.2.1 TODO: Add A Couple Sentences about how the model fits the assumptions now

6.3 Model Performance

```
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")
```



In assessing the robustness and explanatory power of our model, several key metrics were considered. The multiple R-squared value of 0.8053 indicates that approximately 80.5% of the variation in life expectancy across developing nations is explained by our model. This high value implies strong predictive capability and robustness of our chosen explanatory variables. The adjusted R-squared value of 0.8046, closely aligning with the R-squared, reinforces our model's reliability, suggesting minimal overfitting. Additionally, model selection guided by Bayesian Information Criterion (BIC) reinforced the chosen variables' optimal balance between complexity and explanatory power. Together, these metrics demonstrate a well-performing model suitable for policy analysis and strategic decision-making.

In summary, our final model provides clear, statistically significant insights into how various economic, health, and social factors collectively and individually shape life expectancy in developing countries. The results confirm established hypotheses regarding infant mortality, HIV/AIDS prevalence, education, and health expenditure. They also illuminate intriguing questions concerning the complex role of alcohol consumption and economic indicators in public health contexts.

7 Discussion and Conclusion

7.1 Summary of Study

In this study we explored factors that influence life expectancy in developing countries using multiple linear regression. We obtained our dataset from Kaggle (Life Expectancy WHO dataset), which compiles health, demographic, and economic indicators from the World Health Organization and United Nations. After cleaning the dataset and removing missing values, we focused on five predictors: schooling, alcohol consumption (categorical), infant deaths, HIV/AIDS prevalence, and health expenditure (as a percentage of GDP per capita). Because the relationships for infant deaths, HIV/AIDS, and health expenditure with life expectancy were all exponential-like, log-transformations were applied to stabilize variance and improve model fit. Residual analysis confirmed no major violations of linear regression model assumptions, while model selection using regsubsets confirmed that our final set of predictors balanced model simplicity and predictive strength. Our goal was to determine which factors most significantly affect life expectancy in developing nations and to derive policy recommendations based on our findings.

7.2 Key Findings

From the model results, we have determined that school has a significant and sizable impact on life expectancy for developing countries. The impact is significant because the t-value is less than 2×10^{-16}, meaning there is an extremely small probability of the impact of schooling on life expectancy solely by chance. The reason the impact is sizable is because in terms of absolute value, school has the second largest estimate (behind alcohol being negligible) with an increase of 1.06 in life expectancy for every additional year in schooling.

This finding also makes sense from a theoretical standpoint. The more educated an individual is, they will probably have more skills in adulthood. Therefore, they most likely will get better jobs compared to people with less education. As a result, those individuals will have more money for basic needs and healthcare increasing their overall life expectancy. Additionally, people who are more educated may also be more likely to be aware of health risks, which also increases their life expectancy. Hence, we recommend countries to prioritize increasing their education system compared to most of the other variables in the model.

However, it is important to note that schooling's relationship with life expectancy could be confounding. For instance, richer countries might in general have higher life expectancy and also have stronger education systems. In this case, the wealth of the country is driving both the higher and life expectation and stronger education system. However, we doubt that schooling is confounding in this manner since weather countries probably spend more on health care, so if schooling was confounding is should be correlated with health expenditure, but this is not true since both variables's coefficients have p-values of 2×10^{-16}.

Infant mortality was found to have a significant negative relationship with life expectancy. After log-transforming the infant deaths variable, the estimated coefficient was -3.241 with a p-value of <2e-16. This strong negative coefficient indicates that even a small increase in the logged number of infant deaths significantly lowers predicted life expectancy. This showed that as infant deaths per 1000 live births increase, life expectancy declines significantly. This finding is intuitive because high infant mortality rates reflect poor healthcare infrastructure and limited access to essential resources.

HIV/AIDS prevalence was another strong negative predictor of life expectancy. After log-transformation, the estimated coefficient for HIV/AIDS was -1.787 with a p-value of < 2e-16. The negative sign indicates that countries with higher rates of HIV/AIDS experience lower life expectancy, consistent with the devastating impact of HIV/AIDS on national health systems. HIV/AIDS directly weakens healthcare systems and contributes to shorter life expectancy.

Health expenditure (as a percentage of GDP per capita) was positively associated with life expectancy. It was log-transformed and has a coefficient of 1.425 and a p value of 0.00087. Although the p-value is slightly higher than the other predictors, it is still highly statistically significant.Its effect is still smaller compared to schooling and infant deaths but remains statistically significant. This suggests that even moderate increases in healthcare investment can meaningfully improve population health. This aligns with expectations: better-funded health systems tend to provide wider access to vaccinations, early disease detection and quality treatments. All these factors enhance longevity.

Reducing infant mortality, controlling HIV/AIDS and increasing healthcare expenditure are interconnected strategies that collectively raise life expectancy for developing countries. Our findings show that public health initiatives are crucial for improving infant healthcare, access to vaccinations, food banks and diseases prevention programs. Countries which aim to improve life expectancy should adopt a more comprehensive approach that combines investment in health care infrastructure which will have a trickle effect on all these other variables. Together, these efforts would counteract the drivers of infant mortality, HIV/AIDs and overall low life expectancy.

Alcohol consumption produced an interesting result. Countries under negligible alcohol consumption (1.08 liters per capita) had a life expectancy 0.676 years lower on average than countries where alcohol consumption was present. This was based on the coefficient for negligible alcohol (-0.676 and p-value = 0.0123). Although this finding is statistically significant, it is counterintuitive. Alcohol is typically associated with negative health effects. A likely explanation is that there may be a confounding variable: alcohol consumption might correlate with broader indicators of economic development, such as better healthcare or nutrition, rather than causing increased life expectancy directly. Future research regarding recreational drug use and cultural factors could help clarify this relationship.

7.3 Future Steps

In the future, this analysis could be expanded on in numerous manners. For instance, the analysis could be done for both semi-developed and developed countries. This would allow us to compare and contrast what variables impact life expectancy for each category. As a result, we could create a step-by-step guide for how developing countries can increase life expectancy even after those countries have developed into semi-developed or developed countries.

Additionally, our analysis could have benefited from a wider range of variables. In our final model, four of our five variables were health related (infant deaths, hiv/aids cases, percent of GDP per capita spent on healthcare, if alcohol is consumed). In the future, we would look at other data sources, and try to get other types of variables, especially from economic and financial data.

Moreover, the biggest potential improvement for this study has to do with the alcohol consumed variable. The issue with the variable is that it is a bit too specific. The point is that alcohol consumption is merely one type of recreational drug use. In the future, we would get data that would track information related to numerous types of recreational drug use. This is useful because different countries and different cultures tend to use different types of drugs for recreational purposes (if at all). Therefore capturing all this data would allow us to have a better understanding of how drug use affects life expectancy in developing countries.

8 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