STA302H1 Final Project Project Proposal (Part 1)

Economic, Social, and Health Determinants of Life Expectancy: A Cross-Country Comparison of Developing and Developed Nations

Mason Law, Kai Wang

Contributions

1. **Mason Law:**
    - Introduction:
        - ☑ ~~Research Question~~
        - ☑ ~~Why a Linear Model is Appropriate~~
        - ☑ ~~Who Benefits from This Model~~
    - Data Description:
        - ☑ ~~Statistical Summary~~
        - ☑ ~~Numerical / Graphical Summaries~~
        - ☑ ~~Justification of Interacted Predictor(s~~)
    - Ethics Discussion: ø
    - Preliminary Results:
        - ☑ ~~Preliminary Model Specifications~~
        - ☑ ~~Diagnostic Plots and Assumption Checks~~
        - ☑ ~~Model Summary Table~~
        - ☑ ~~Interpretation of Preliminary Results~~
    - Plan:
        - ☑ ~~Predictor Selection~~
        - ☑ ~~Model Assumptions and Diagnostics~~
        - ☑ ~~Model Refinement~~
        - ☑ ~~Project Timeline~~
    - Bibliography: ø
2. **Kai Wang:**
    - Introduction:
        - ☑ ~~Summary of Peer-Reviewed Research~~
    - Data Description:
        - ☑ ~~Data Origin~~
    - Ethics Discussion:
        - ☑ ~~Dataset Trustworthiness~~
        - ☑ ~~Data Collection Ethics~~
    - Preliminary Results: ø
    - Plan:
        - ☑ ~~Visual Project Summary~~
    - Bibliography:
        - ☑ ~~References~~

Introduction

1.1 Research Question

How do economic, social, and health-related factors influence life expectancy across countries, and do these effects differ between developing and developed nations?

1.2 Why a Linear Model is Appropriate

A multiple linear regression (MLR) model is appropriate since the response variable (life expectancy) is continuous and the predictors are both numerical and categorical. While simpler techniques (e.g., correlation or simple linear regression) can identify basic relationships, they cannot account for the combined effects of multiple factors. MLR provides better insights through:

- **Estimated coefficients (β):** Indicate how much life expectancy changes for a one-unit increase in each predictor, holding others constant..
- **P-values + confidence intervals**: Identify which predictors have statistically significant relationships with life expectancy.
- **Confidence intervals:** Quantify uncertainty.
- **Model predictions:** Estimate expected life expectancy given well-defined predictors.

Overall, MLR is an effective means for identifying which social, economic, and health factors most strongly influence life expectancy among different countries.

1.3 Summary of Peer-Reviewed Research

- While income is linked to health, disease (e.g., AIDS) disrupt life expectancy gains in low-income countries. Infant and adult mortality are influenced by GDP, which correlates positive with life expectancy, while disease has a negative effect (Soares, 2007).
- Education has a casual effect on reducing mortality, with each additional year of compulsory schooling linked to higher life expectancy (Lleras-Muney, 2006).
- Low education and socioeconomic status are associated with higher alcohol consumption and mortality, whereas light drinking (e.g., wine) may reduce cardiovascular and adult mortality. Alcohol use tends to decrease mortality in developed countries but increase in developing ones (Wakabayashi, 2025).

1.4 Who Benefits from This Model

- **Developing countries:** can benchmark their current expected life expectancy and determine which factors to prioritize to improve it.
- **Government agencies:** target investments (e.g., healthcare systems, education, etc.) in developing countries using model insights.
- **Public health policymakers:** use model findings to identify which interventions will yield the largest impact on life expectancy.

Data Description

2.1 Data Origin

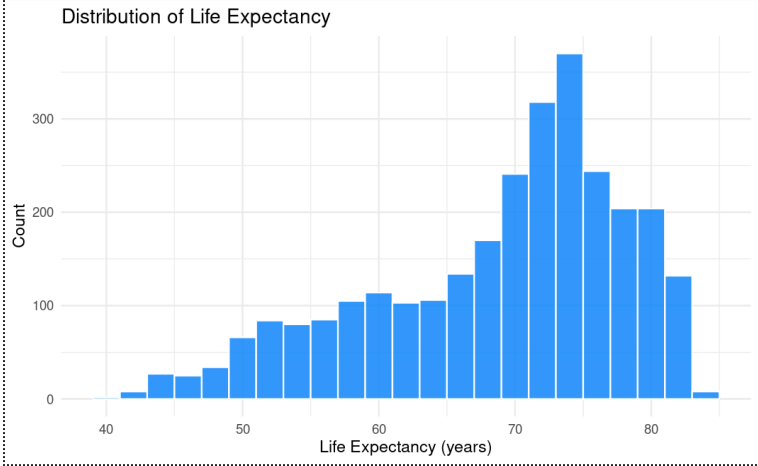
The dataset was found on Kaggle and was originally collected by the Global Health Observatory (GHO) under the World Health Organization (WHO). The data was originally collected to monitor and report on global progress in improving health. The curator of the dataset compiled tables from the openly available data from 2000 to 2015 and merged them into a single dataset.

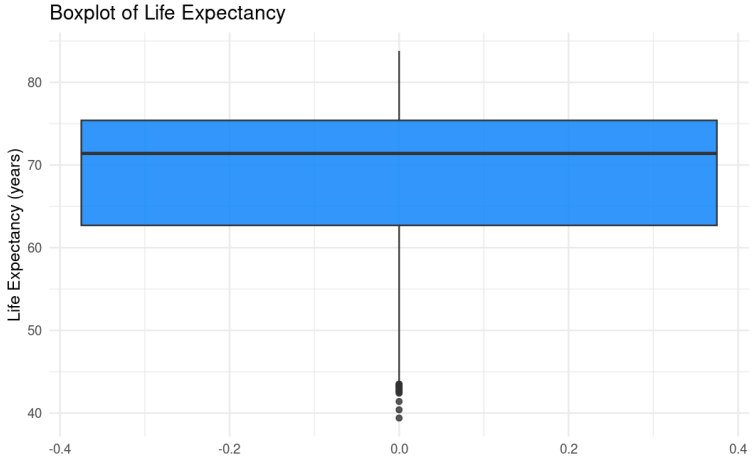
2.2 Statistical Summary of Response Variable

Life expectancy in years is given by:

- Range: **39.4-83.9**
- Mean: **68.86**
- Median: **71.4**
- Interquartile range (IQR): **12.7**
- Standard deviation: **9.41**

As shown in the plots below, the distribution of life expectancy is approximately unimodal with a slight left-skew.





Each observation corresponds to a unique country and year pair, so it's reasonable to treat them as independent. Life expectancy is continuous, and its range allows for the use of an MLR model.

2.3 Numerical / Graphical Summary of Predictor(s)

Numerical Summary of Preliminary Predictors

Predictor	Min	Q1	Median	Mean	Q3	Max
InfantDeathsPer1k	1.80	8.10	19.60	30.36	47.35	138.10
AdultMortalityPer1k	49.38	106.91	163.84	192.25	246.79	719.36
AlcoholLitersPerCapita	0.00	1.20	4.02	4.82	7.78	17.87
HepatitisBCoveragePercentage	12.00	78.00	89.00	84.29	96.00	99.00
GdpPerCapitaUsd	148.00	1,415.75	4,217.00	11,540.92	12,557.00	112,418.00
PopulationMillions	0.08	2.10	7.85	36.68	23.69	1,379.86
Thinness10To19Percentage	0.10	1.60	3.30	4.87	7.20	27.70
AvgSchoolingYears	1.10	5.10	7.80	7.63	10.30	14.10

Categorical Summary: Development Status

DevelopmentStatus	Count	Proportion
0	2272	79.3%
1	592	20.7%

Based on the numerical summary of our continuous and categorical predictors above, we notice the following characteristics:

- **Economic:**
    - GDP per capita is strongly right-skewed.
    - Population considerably skewed.
- **Social:**
    - Average schooling years show moderate variability.
    - Alcohol consumption is right-skewed.

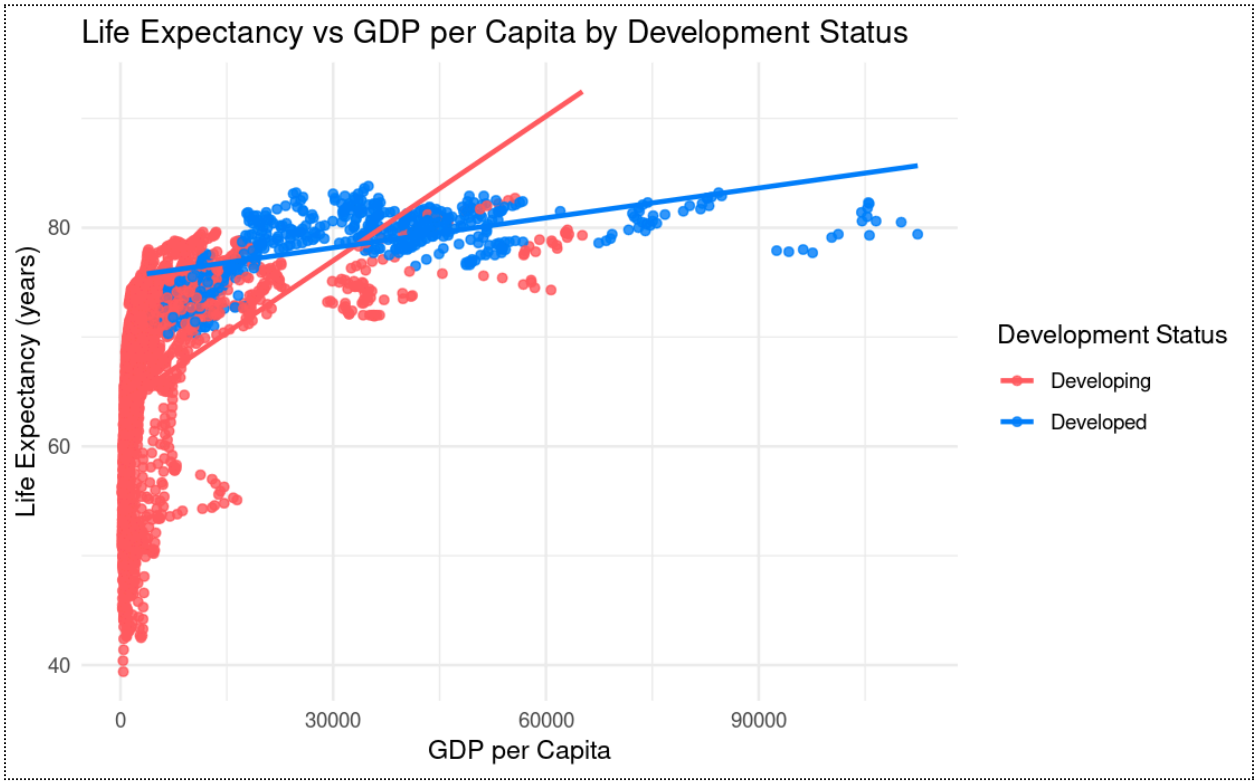
- **Health:**
    - Infant and adult mortality exhibit high variance and right-skew. Some countries have extreme values.
    - Hepatitis B coverage is mostly large, but the lower quartile indicates countries lag behind.
    - Adolescent thinness has a low median but extreme outliers, suggesting nutrition is uneven.
- **Categorical:**
    - Development status is imbalanced.

2.4 Justification of Interacted Predictor(s)

We will include an interaction between GDP and development status as the effect of GDP on life expectancy differs by country type:

- **Developing countries**: additional GDP greatly improves health outcomes as infrastructure improves.
- **Developed countries**: additional GDP has little effect as health infrastructure is (mostly) already established.

Visual inspection of the GDP vs. life expectancy scatterplot colored by development status (below) shows different slopes, supporting the inclusion of this interaction in our model.



Ethics Discussion

3.1 Dataset Trustworthiness

The dataset is trustworthy as it originates from reliable public health sources. The underlying data was obtained from the World Health Health Organization (WHO) based on Vital Registration (VR) data (Rajarshi, 2018). This dataset was then extended by filling in missing values for some countries using World Bank data, applying reasonable methods like averages from the previous three years (Gochiashvili, 2022). Both the original and extended datasets are highly rated and widely used on Kaggle, further supporting their credibility and transparency (Rajarshi, 2018; Gochiashvili, 2022).

3.2 Data Collection Ethics

From an ethical standpoint, the dataset we will be using respects privacy, consent, and ownership:

- **Privacy:** The data contains no personally identifiable information; all records are aggregated at the national level to protect individual privacy and minimize the risk of data misuse (World Health Organization, 2020).
- **Consent:** While individuals represented in the national registries used did not directly give consent for research use, the data is anonymized and collected by government agencies for statistical and policy purposes (World Health Organization, 2020). This makes secondary analysis through this project admissible.
- **Ownership:** The dataset's primary sources (i.e., WHO and the World Bank) are legitimate entities and data owners that have made their data publicly available for research and educational purposes. The curators who compiled and shared the dataset have provided appropriate attribution and transparency regarding their methods to supplement the data (Rajarshi, 2018; Gochiashvili, 2022).

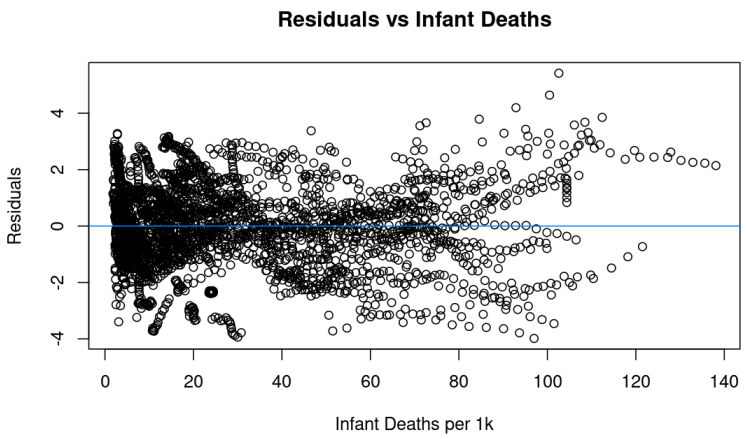
Overall, the dataset was collected and used in an ethical manner. It maintains public trust by ensuring accurate data, transparency, and respect for the aspects of privacy, consent, and ownership, all while supporting research into global health trends. Thus, our use of the data is ethical by extension on the above ethics discussion.

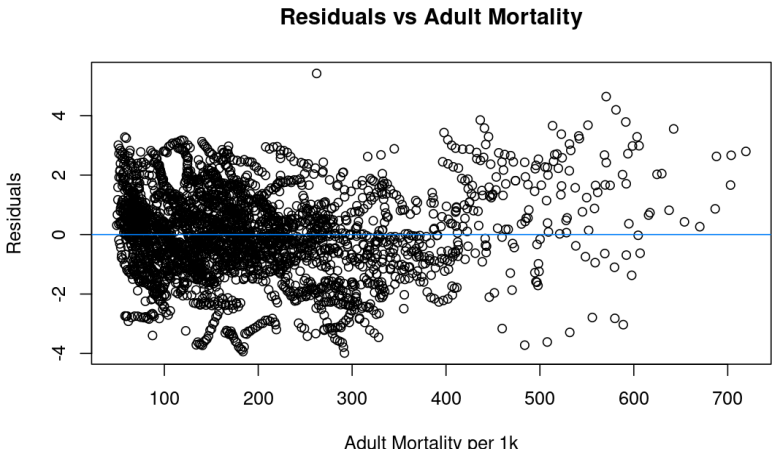
Preliminary Results

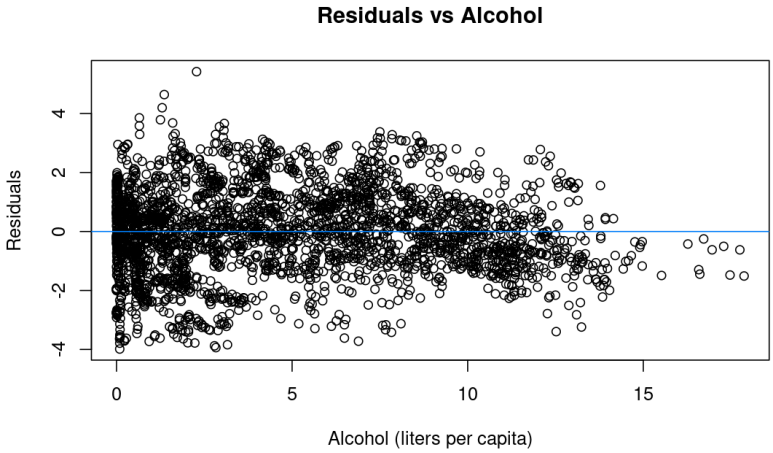
4.1 Preliminary Model Specification

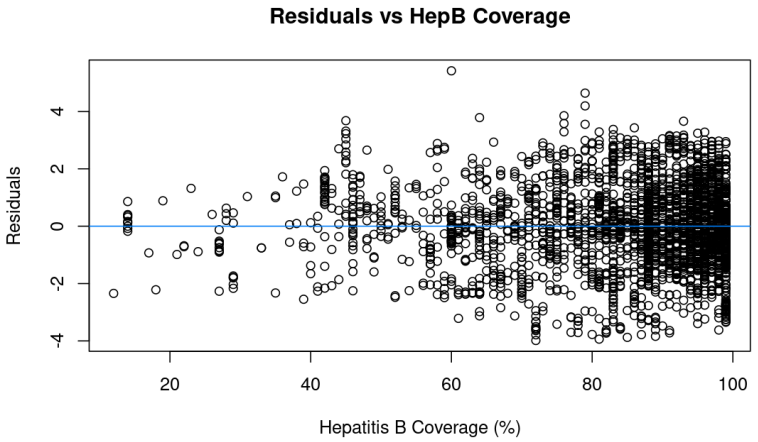
An MLR model was fitted to examine the relationship between life expectancy and nine predictors: infant mortality, adult mortality, alcohol consumption, hepatitis B coverage, GDP per capita, population size, adolescent thinness, average schooling years, and an interaction between GDP and development status.

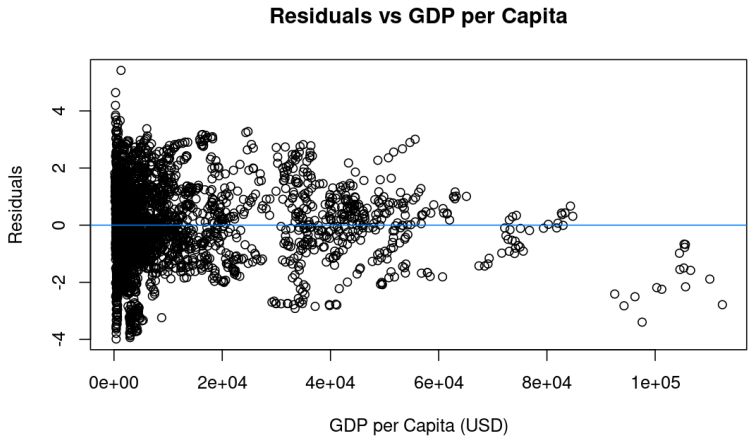
4.2 Diagnostic Plots and Assumption Checks

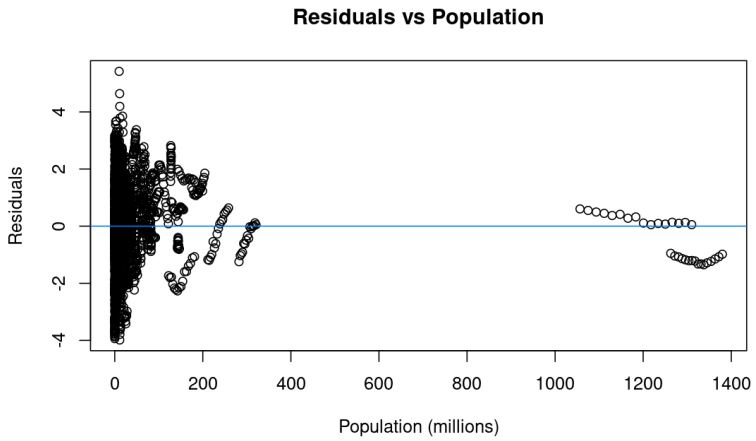


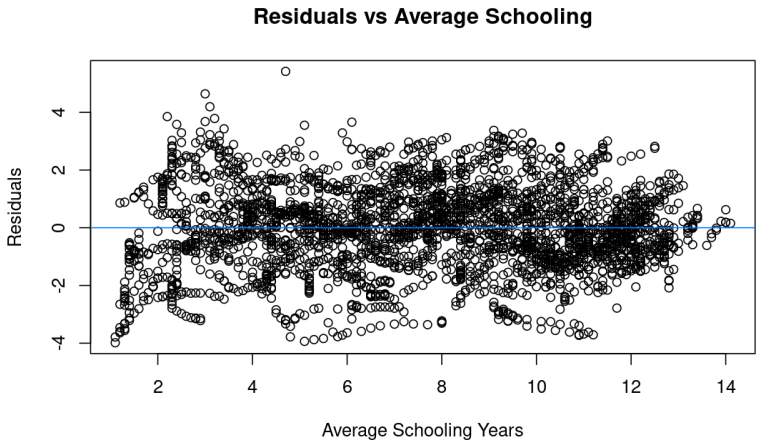




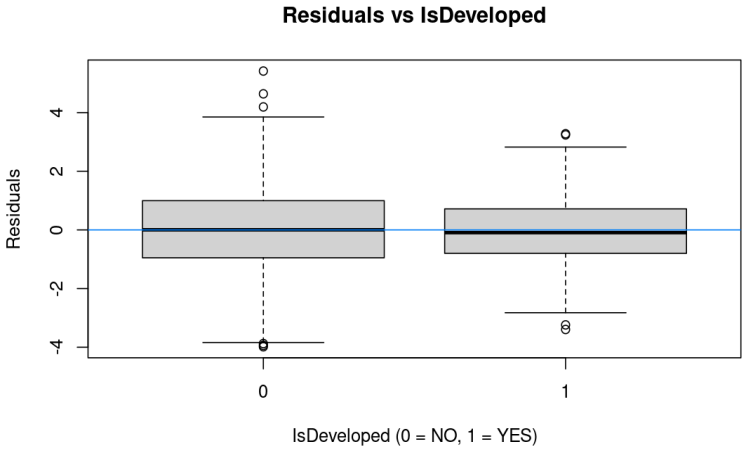


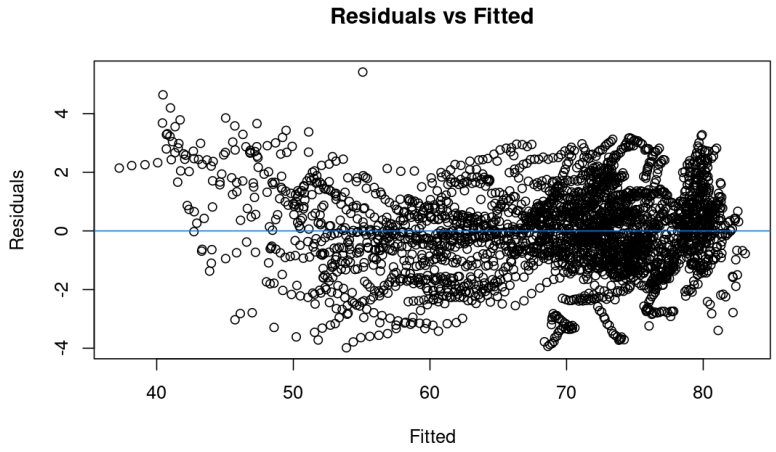


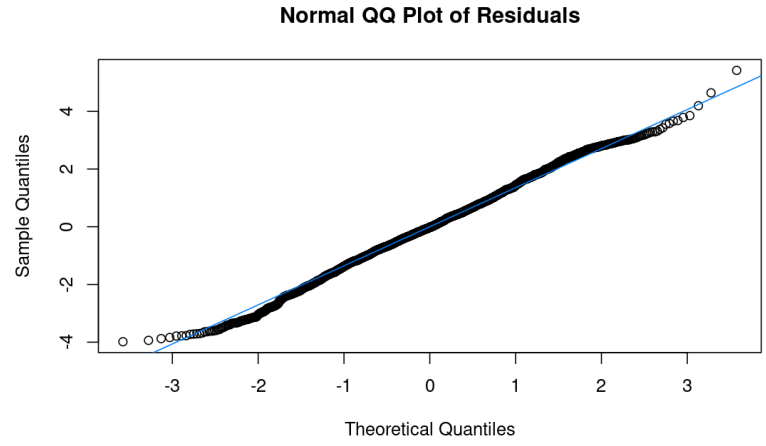


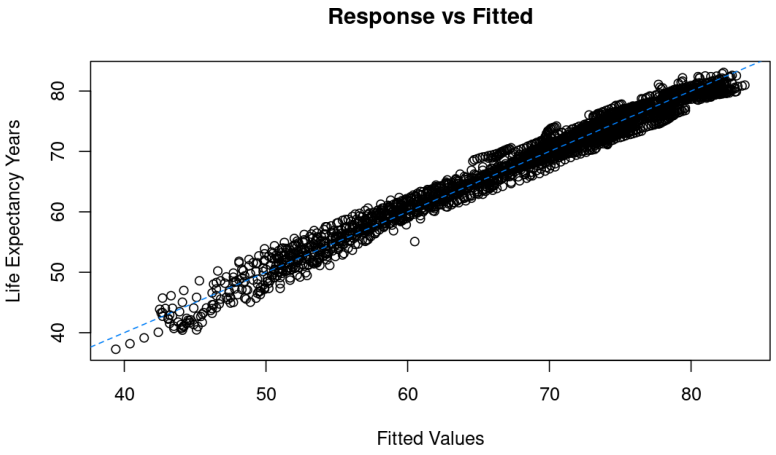

Residuals vs Average Schooling
Residuals
Average Schooling Years

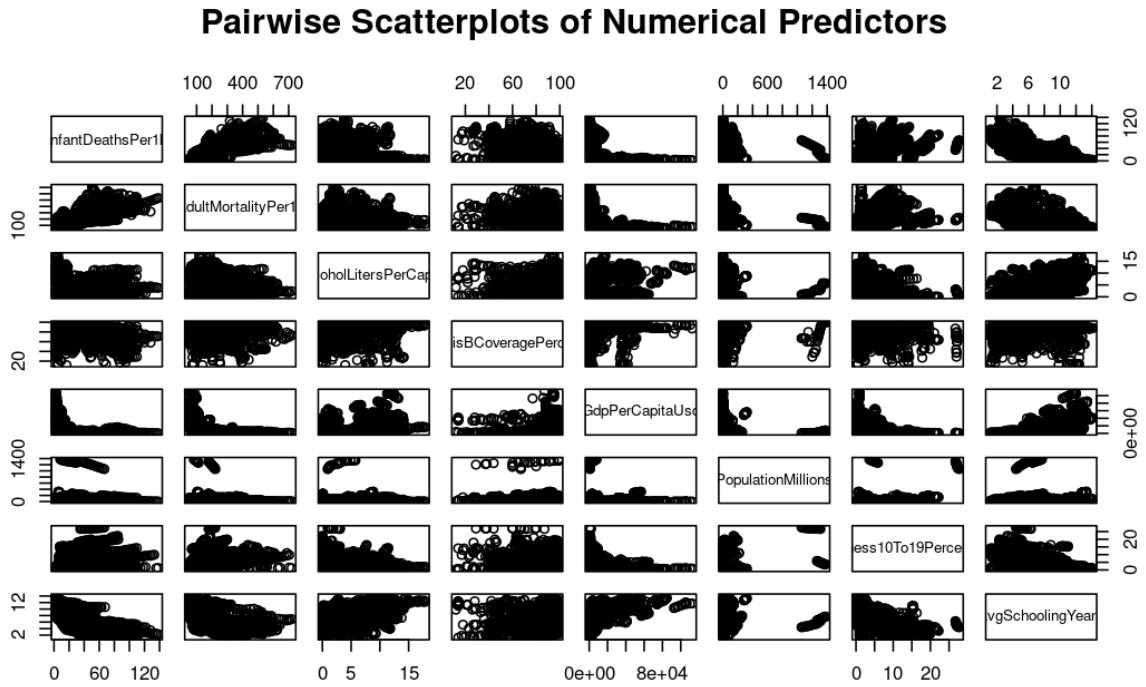

Residuals vs Thinness
Residuals
Adolescent Thinness (Ages 10-19) (%)


Residuals vs IsDeveloped
Residuals
IsDeveloped (0 = NO, 1 = YES)


Residuals vs Fitted
Residuals
Fitted


Normal QQ Plot of Residuals
Sample Quantiles
Theoretical Quantiles


Response vs Fitted
Life Expectancy Years
Fitted Values



Based on the above:

1. **Linearity:** Holds;no systematic patterns exist in the residuals-fitted or residuals-predictor plots.
2. **Uncorrelated errors:** Holds; no visual clustering in the residuals-fitted plot, aside from a minor right-side cluster likely due to dense observations with similar fitted values.
3. **Constant error variance:** Moderately violated with slight fanning in residuals-predictor plots (e.g., Hepatitis B, GDP, population, and more).
4. **Normality:** Holds; only minimal deviations at the QQ plot tails.

Additional conditions for MLR:

1. **Conditional mean response:** random diagonal scatter in response-fitted plot present; condition holds.
2. **Conditional mean predictor condition:** in all pairwise scatterplots, no observable trend or only a linear trend exists; condition holds.

4.3 Model Summary Table

Preliminary Linear Regression Results

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	80.999	0.255	317.161	0.000	80.498	81.500
InfantDeathsPer1k	−0.136	0.002	−57.832	0.000	−0.141	−0.131
AdultMortalityPer1k	−0.047	0.000	−112.751	0.000	−0.048	−0.046
AlcoholLitersPerCapita	0.055	0.010	5.349	0.000	0.035	0.075
HepatitisBCoveragePercentage	−0.004	0.002	−2.213	0.027	−0.008	−0.001
GdpPerCapitaUsd	6.73×10^{-6}	3.74×10^{-6}	1.798	0.072	-6.10×10^{-7}	1.41×10^{-5}
IsDeveloped	0.152	0.139	1.091	0.275	−0.121	0.425
PopulationMillions	4.48×10^{-5}	2.07×10^{-4}	0.216	0.829	-3.61×10^{-4}	4.51×10^{-4}
Thinness10To19Percentage	0.005	0.008	0.605	0.545	−0.011	0.020
AvgSchoolingYears	0.117	0.017	7.039	0.000	0.084	0.149
GdpPerCapitaUsd:IsDeveloped	2.36×10^{-5}	4.59×10^{-6}	5.142	0.000	1.46×10^{-5}	3.26×10^{-5}

4.4 Interpretation of Preliminary Results

Several predictors are significant in the expected directions:

- Infant (-0.136, p-value < 0.001) and adult mortality (-0.047, p-value < 0.001) have large negative effects.
- Average schooling years (0.117, p-value < 0.001) is strongly positive.
- Alcohol consumption (0.0055, p-value < 0.001) shows a small positive effect.

Our GDP and development interaction (2.36e-5, p-value < 0.001) confirms our expectation of the effect of GDP varying by development status. This aligns with previous findings that GDP positively correlates with life expectancy, particularly in high-income countries, but that this relationship can be disrupted in lower-income ones due to factors like disease (Soares, 2007). Also, our schooling coefficient (0.117) supports existing evidence that higher levels of education are associated with lower mortality and higher life expectancy (Lleras-Muney, 2006). The slight deviation from expectations in our GDP interaction may reflect nonlinearity not yet accounted for or other variables like infant/adult mortality, average schooling years, and others absorbing the effect of GDP in developing countries.

Plan

5.1 Predictor Selection

Our final model will be determined through a backward elimination process utilizing adjusted R^2 to identify weak predictors of life expectancy. Variables with significant theoretical weight (e.g., GDP, mortality, development status, etc.) will be kept because of their justified importance in prior research and interpretive value. For further predictor decision, we will also consider how each predictor influences model assumptions and the accuracy of subsequent inference.

5.2 Model Assumptions and Diagnostics

All assumptions (i.e., linearity, uncorrelated errors, homoscedasticity, and normality) will be re-checked after each adjustment. Diagnostic tools like residual-fitted, residual-predictor, and QQ plots will be utilized. Iterative checking will ensure improvements/violations are immediately dealt with as we progress. These checks will ensure that coefficient sampling distributions are valid. So hypothesis tests and confidence intervals are utilized appropriately.

5.3 Model Refinement

Given constant variance being slightly violated in our preliminary model, various transformations will be tested (e.g., log(Hepatitis B), log(GDP), log(population), and more). Box-Cox methods will also be tested to hopefully stabilize variance. Each transformation will be justified based on diagnostic improvements.

5.4 Project Timeline

Due to the size of our group (i.e., 2 people), we'll both work on the project according to this schedule:

- **Nov. 3-4:** Further variable screening and refinement
- **Nov 5-7:** Iterative model fitting and diagnostic checks
- **Nov 10-12:** Transformation testing + presentation/poster drafting
- **Nov 13-14:** Final model interpretation and visualization + presentation/poster drafting
- **Nov 17-21:** Refine poster/presentation + record final presentation

References

Gochiashvili, L. (2022). *Life Expectancy (WHO) Fixed*. Kaggle. https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated/data

Lleras-Muney, A (2006, July). *The Relationship between Education and Adult Mortality in the United States*. ResearchGate. https://www.researchgate.net/publication/228209433_The_Relationship_between_Education_and_Adult_Mortality_in_the_United_States_Adriana_Lleras-MuneyErratum

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

Soares, R. R. (2007, May 29). *On the Determinants of Mortality Reductions in the Developing World*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1111/j.1728-4457.2007.00169.x

Wakabayashi, I., & Groschner, K. (2025). *Alcohol and life expectancy*. Environmental health and preventive medicine. https://pmc.ncbi.nlm.nih.gov/articles/PMC12358757/

World Health Organization (2020, December). *WHO methods and data sources for life tables 1990-2019*. Department of Data and Analytics: Division of Data, Analytics and Delivery for Impact. WHO, Geneva. https://cdn.who.int/media/docs/default-source/gho-documents/global-health-estimates/ghe2019_life-table-methods.pdf?sfvrsn=c433c229_5