# Socioeconomic, Health, and Demographic Predictors of Life Expectancy

Ema, Eileen, Elsie

## Introduction

Life expectancy is consistently regarded as one of the main indicators of the overall health of a population (Health Status | US EPA, 2025). Life expectancy reflects the combined influence of economic development, healthcare access, disease burden, education, and demographic conditions, making it a powerful indicator of population well-being (Irandoust et al., 2025; Mondal & Shitan, 2013). A major part of evaluating life expectancy is understanding which factors have an association with it, and examining how life expectancy varies in relation to a range of social and health measures. Understanding what is associated with differences in life expectancy across countries is a central question in global health research. Global life expectancy has increased steadily over the past several decades across the world, although large disparities persist between regions, income groups, and levels of healthcare access (Dattani et al., 2023). Public health research has consistently identified major determinants of life expectancy, including economic development, education, infectious disease burden, vaccination coverage, nutrition, and child and adult mortality rates as key contributors to differences in population survival (Irandoust et al., 2025). Understanding how these determinants vary across countries and over time is important for informing conversations about where improvements in health systems might be prioritized.

Previous research has identified several variables strongly associated with life expectancy. Economic factors such as GDP per capita, social determinants like education and living conditions, and access to healthcare (e.g., physician density, hospital resources) are associated with longevity (Irandoust et al., 2025; Mondal & Shitan, 2013). Nutrition, vaccination coverage, and sanitation also shape population health (Dattani et al., 2023). Our study uses a dataset containing several of these variables to examine their association with life expectancy, building on prior evidence of how social and health factors can reflect differences in population survival (Daniel et al., 2018).

The primary goal of this project is to address the research question: Which health, economic, and demographic factors are significantly associated with life expectancy? To answer this, we employ a reproducible statistical analysis along with appropriate regression modeling. Rather than conducting individual analyses of every variable, we focus on a targeted set of meaningful predictors, such as income level, region, vaccination rates, nutritional indicators, and schooling, to evaluate their associations with life expectancy in a clear and interpretable way.

I. Dataset

To address this question, we use an updated global health dataset covering 179 countries from 2000 to 2015 to investigate which factors most strongly predict life expectancy worldwide. The dataset includes measures of vaccination coverage, HIV incidence, BMI, alcohol consumption, mortality rates, schooling, population, GDP, and Gross National Income (GNI) classifications, with missing values carefully addressed through regional and temporal imputation methods, such as such as filling in missing country-year values with the average of the surrounding three years or, if a country had missing values across all years, using the average of its region. Countries with excessive missingness were removed to ensure high data quality, and economic categories were aligned with World Bank standards for comparability. The dataset used in this project came originally from Kaggle, and was initially compiled from WHO, World Bank, and Our World in Data sources. Each source collects its data as follows: The World Bank collects population, GDP, life expectancy, and GNI per capita data from national statistical agencies and international reporting systems; WHO collects health-related data, such as vaccination coverage, HIV incidence, BMI, alcohol consumption, mortality rates, and thinness, from national health reports and surveillance systems across countries; Our World in Data, a University of Oxford project, compiles schooling data from national and international educational statistics. Each row represents a country and all the data collected from one year in that country across all the other variables. Even though the dataset comes from multiple sources, they are all said to be adjusted and standardized. The observational unit was a country-year. All the dataset variables are listed below:

**Response Variable:**

Life_expectancy: The average life expectancy for both genders across different years, from 2000 to 2015.

**Predictor Variables:**

Economic and Demographic –

1. GDP_per_capita: Gross Domestic Product per capita in current US Dollars.
2. Population_mln: Total population of a country in millions.
3. Schooling: Average years individuals aged 25 and over have spent in formal education.
4. Economy_status_Developed: A binary indicator (0 or 1) denoting whether a country is classified as 'Developed'.
5. Economy_status_Developing: A binary indicator (0 or 1) denoting whether a country is classified as 'Developing'.

Lifestyle –

6. Alcohol_consumption: Records alcohol consumption in litres of pure alcohol per capita for individuals aged 15 years and over.
7. BMI: Body Mass Index, a measure of nutritional status in adults (defined as a person's weight in kilograms divided by the square of that person's height in meters).
8. Thinness_ten_nineteen_years: Prevalence of thinness among adolescents aged 10-19 years (specifically, BMI < -2 standard deviations below the median).
9. Thinness_five_nine_years: Prevalence of thinness among children aged 5-9 years (specifically, BMI < -2 standard deviations below the median).

Mortality and Disease –

10. Infant_deaths: Represents the number of infant deaths per 1,000 population.
11. Under_five_deaths: Represents the number of deaths of children under five years old per 1,000 population.
12. Adult_mortality: Represents the number of deaths of adults per 1,000 population.
13. Hepatitis_B: Represents the percentage of coverage for Hepatitis B (HepB3) immunization among 1-year-olds.
14. Measles: Represents the percentage of coverage for Measles containing vaccine first dose (MCV1) immunization among 1-year-olds.
15. Polio: Represents the percentage of coverage for Polio (Pol3) immunisation among 1-year-olds.
16. Diphtheria: Represents the percentage of coverage for Diphtheria tetanus toxoid and pertussis (DTP3) immunization among 1-year-olds.
17. Incidents_HIV: Represents the incidents of HIV per 1,000 population aged 15-49.

Observational Units –

18. Country: A list of 179 distinct countries included in the dataset.
19. Region: Categorizes the 179 countries into 9 geographical regions, such as Africa, Asia, Oceania, and the European Union.
20. Year: The observed year, ranging from 2000 to 2015.

II. Rationale

The purpose of this data analysis is to help identify which health and economic variables are most strongly associated with differences in life expectancy. It also helps shows how different diseases / vaccinations and child and adult mortality can contribute to the overall health of a nation. This data analysis can also help us make conclusions on what types of health policy can be recommended or suggested in order to target preventable health risks in certain countries. This analysis contributes to the existing body of literature by examining multiple determinants simultaneously, allowing us to identify which predictors remain associated with life expectancy after accounting for other factors.

III. Exploratory Data Analysis

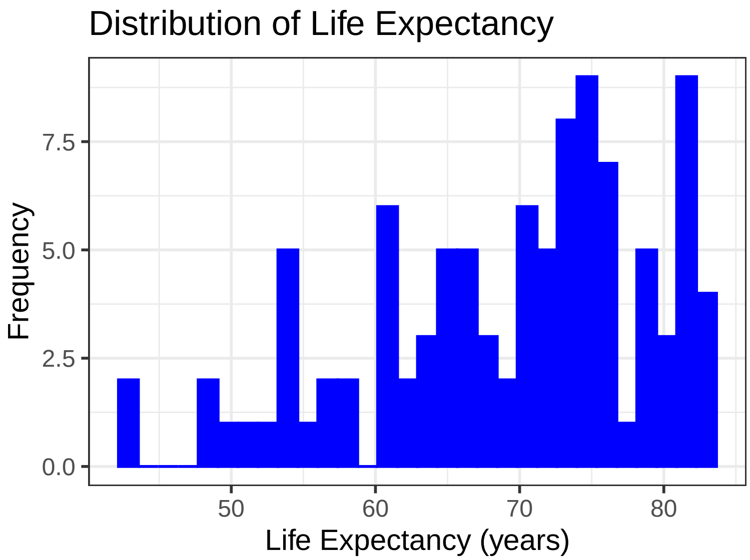


**Figure 1.** Distribution of Life Expectancy

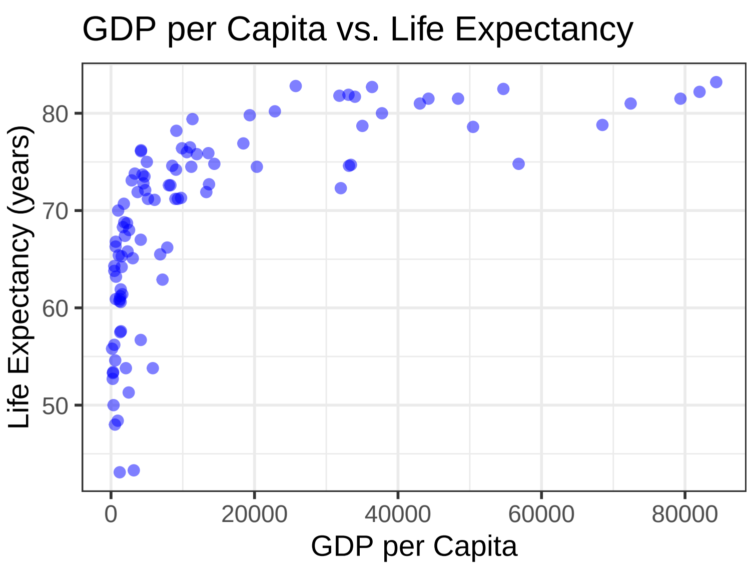


**Figure 2.** GDP per Capita vs. Life Expectancy

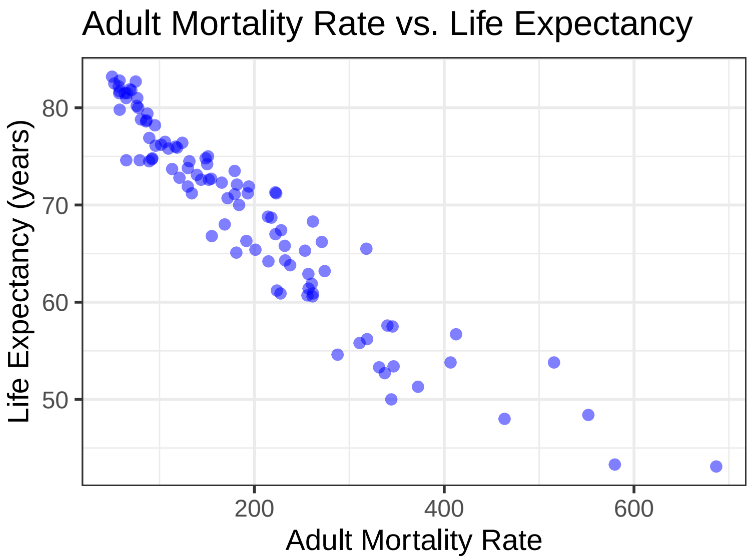


**Figure 3.** Adult Mortality Rate vs. Life Expectancy

The EDA helps us visualize some of the aspects of this dataset. It helps look at life expectancy as a whole across all the countries, and allows us to see if it is normally distributed or left skewed. The scatterplots provide an initial look at how life expectancy varies with GDP per capita and adult mortality, suggesting potential associations between economic and health indicators and life expectancy that we will investigate more formally in the regression analysis. Additional exploratory figures, including visualizations of BMI and other predictors, are provided in Appendix A for completeness. These exploratory patterns help guide which variables may plausibly relate to life expectancy and justify their inclusion in the regression model.

## Methodology

The goal of this analysis was to assess how the variables in the dataset are associated with life expectancy. We selected a linear regression model over other models because the outcome variable (life expectancy) is continuous and the research question focuses on estimating relationships between predictors and mean life expectancy, while holding other variables constant. A linear regression is appropriate for this dataset also because the predictors are a mix of numerical and categorical variables, and the goal is to estimate how each variable relates to life expectancy. Other methods such as logistic regression, ANOVA, and t-tests, were not appropriate here because they are designed for categorical outcomes or for comparing a small number of group means, and T-tests are already conducted by this model for each variable. Therefore, a linear regression appeared to be the most efficient and appropriate method that fits our research question.

The linear regression uses the equation shown below as a guide, with life expectancy as the output, and the other variables as its predictors (simplified due to the large amount of variables): Life Expectancy = intercept + (output coefficient × each quantitative predictor) + (output coefficient × each categorical predictor) + error. When replicated, it should be ensured the dataset is loaded into R, and that when inputting this model into R with the lm() function, that life expectancy is the output, and that all other variables are evaluated quantitatively except Economy Status Developed and Region, which are categorical and therefore dummy variables in this linear regression. Economy Status Developed was changed into a factor variable, 1 for developed, 0 for not. Results are shown using the summary() function. P-values and slope/beta terms were assessed for the goals of this project.

The residual plot is symmetrically distributed around the horizontal axis, showing that our data satisfied the assumption of linearity. The residuals appear to be roughly vertically evenly spaced along the y axis, satisfying the assumption of constant variance of the errors. The histogram of the residuals also appears to be roughly symmetrical, satisfying the assumption of a normal distribution of errors. Finally, for independence in this dataset, each row represents a specific country in a specific year. Since each country operates independently with its own health system, economy, and demographic conditions, the error for one observation can be assumed to be independent of the error for another. Because the same country can appear multiple times (for different years), it's possible that a country's values might be somewhat related over time. However, the model treats each country-year as its own observation, and we also included variables like Region and Economy Status, which already help capture similarities between countries that might otherwise create dependence. Because of this, even though small within-country patterns over time could exist, assuming that the errors are independent overall is still a reasonable and practical assumption for this analysis. Plots for the assumptions are shown in Appendix B.

## Results

```
# A tibble: 25 x 3
   Variable                          Estimate P_value
   <chr>                                <dbl>   <dbl>
 1 (Intercept)                       -162.      0.032
 2 Year                                 0.126   0.001
 3 Infant_deaths                       -0.127   0.03
 4 Under_five_deaths                   -0.014   0.702
 5 Adult_mortality                     -0.044   0
 6 Alcohol_consumption                 -0.041   0.588
 7 Hepatitis_B                         -0.022   0.162
 8 Measles                              0.01    0.315
 9 BMI                                 -0.3     0.035
10 Polio                                0.056   0.407
```

| | | | |
|---|---|---|---|
| 11 | Diphtheria | -0.079 | 0.225 |
| 12 | Incidents_HIV | -0.018 | 0.859 |
| 13 | GDP_per_capita | 0 | 0.346 |
| 14 | Population_mln | 0.001 | 0.597 |
| 15 | Thinness_ten_nineteen_years | -0.304 | 0.001 |
| 16 | Thinness_five_nine_years | 0.21 | 0.009 |
| 17 | Schooling | 0.066 | 0.545 |
| 18 | Economy_status_Developed1 | 2.59 | 0.003 |
| 19 | RegionAsia | 1.42 | 0.028 |
| 20 | RegionCentral America and Caribbean | 3.33 | 0 |
| 21 | RegionEuropean Union | 0.48 | 0.638 |
| 22 | RegionMiddle East | 1.48 | 0.079 |
| 23 | RegionOceania | 0.246 | 0.825 |
| 24 | RegionRest of Europe | 1.06 | 0.19 |
| 25 | RegionSouth America | 3.10 | 0.001 |

**Table 1.** Linear regression output for life expectancy and its predictors

Results were analyzed with emphasis on the P-values resulting from the T-tests that were run on each individual variable, while holding other variables constant. The null hypothesis for each variable was that the slope, or beta value, was equal to 0 (there is no relationship between the variable and life expectancy). The alternative hypothesis is that the slope/beta is not equal to 0 (there is a relationship between the variable and life expectancy). Each T-test had a T distribution under the null hypothesis with 1 degree of freedom. These P-values were analyzed with a significance level of alpha equals .05. The variables where the null hypothesis was rejected (P-value below .05) and we had enough evidence, based on the data sourced from Kaggle, to say that there is a relationship between said variable and life expectancy are as follows: Year, Infant_deaths, Adult_mortality, BMI, Thinness_ten_nineteen_years, Thinness_five_nine_years, Economy_status_Developed1, RegionAsia, RegionCentral America and Caribbean, and RegionSouth America.

## Discussion

The results of our analysis provide clear insights into the socioeconomic, health, and demographic factors associated with life expectancy across countries, and supports the conclusion drawn from the corresponding T-tests performed. Parameters including Year, Infant_deaths, Adult_mortality, BMI, Thinness in adolescents and children, Economy_status_Developed, Regions (Asia, Central America and Caribbean, South America), all had statistically significant data supporting their association to life expectancy, which was calculated through T-tests on the individual predictors while holding the other variables constant, and of which had p-values that were smaller than the significance value of alpha equals 0.05. This indicates that both health and economic factors shown by the BMI, thinness, and economic status, and region exhibit a relationship with life expectancy across populations. This aligns with our goal of identifying the predictors of life expectancy, and supports prior research that emphasize the role of social determinants, healthcare access, and economic health and development in shaping population health outcomes. For instance, the association of BMI with life expectancy may reflect the overall nutrition and healthcare of a country, while differences in adult and infant mortality rates can indicate variations in population health, as captured by life expectancy.

While our methodology of using a multiple linear regression model was appropriate for modeling a continuous outcome and estimating the relationship between specific parameters and the life expectancy, there are some limitations to this approach. Firstly, the assumption for independence in this scenario may be partially violated as addressed earlier because countries are observed across time, which can potentially introduce some correlation across years. An alternative way of observing the countries could be observing the countries at random intercepts, which could take into account the repeated measures across similar time frames, and ultimately improve the reliability of the estimates. Secondly, although the dataset we used included many diverse parameters to reduce the effect of confounding variables, some parameters, such as healthcare access, lifestyle factors beyond alcohol and BMI, or environmental exposures/status weren't available, limiting the extensiveness of the analysis. Including more multifaceted aspects of variables related to life expectancy would help improve the model's comprehensiveness and usefulness. However, the data is relatively reliable and valid, as the inclusion of many data points helps mitigate the impact of outliers. Additionally, the data was obtained through reliable sources (i.e. the World Bank), which supports the reliability of the data, and data points with too many missing values (for instance, missing four columns of values) were omitted to increase the usefulness of the data. Imputation methods were also applied to address missing values. However, if the missingness is not completely random, it may also introduce bias into the dataset. This reduces the validity of our dataset as the values used weren't the true ones obtained from reliable sources.

Future analyses could be strengthened by incorporating longitudinal modeling techniques, such as looking at trends specific to a country. In addition, finding ways to capture nonlinear relationships between predictors and life expectancy could strengthen the model and analyses as well, by providing more complex insight on how different parameters relate and influence life expectancy. Including variables such as access to clean water, sanitation, or disease prevalence could provide a more nuanced understanding of the factors influencing life expectancy across countries. Furthermore, the model's performance could be tested more thoroughly to assess it's ability to predict the true values, which can help improve its applicability and reliability in real world applications across different countries.

Overall, our findings address the research question by showing that year, infant and adult mortality, BMI, thinness, economic status, and certain regions listed above have significant relationships with life expectancy. Because this is an observational dataset, the associations we identify should not be interpreted as causal relationships. While our methodology effectively addressed the research question, future analyses using more advanced modeling techniques and comprehensive datasets could strengthen the results and broaden their applicability, including for policy recommendations.

## References


Daniel, H., Bornstein, S. S., & Kane, G. C. (2018). Addressing social determinants to improve patient care and promote health equity: An American College of Physicians Position Paper. Annals of Internal Medicine, 168(8), 577–578. https://doi.org/10.7326/m17-2441

Dattani, S., Rodés-Guirao, L., Ritchie, H., Ortiz-Ospina, E., & Roser, M. (2023, November 28). Life expectancy. Our World in Data. https://ourworldindata.org/life-expectancy?insight=life-expectancy-has-increased-across-the-world#key-insights

Health Status | US EPA. (2025, July 25). US EPA. https://www.epa.gov/report-environment/health-status#:~:text=While%20no%20single%20set%20of,and%20quality%20of%20health%20care.

Irandoust, K., Daroudi, R., Tajvar, M., & Yaseri, M. (2025). Global and regional impact of health determinants on life expectancy and health-adjusted life expectancy, 2000–2018: an econometric analysis based on the Global Burden of Disease study 2019. Frontiers in Public Health, 13, 1566469. https://doi.org/10.3389/fpubh.2025.1566469

Life Expectancy & Socio-Economic Determinants Dataset CSV download free | Open Data Marketplace. (2025, July 24). https://www.opendatabay.com/data/ai-ml/0bdbea7e-f40b-4c41-b010-37537d03a723?utm_source

Mondal, M. N. I., & Shitan, M. (2013). Relative importance of demographic, socioeconomic and health factors on life expectancy in low- and Lower-Middle-Income countries. Journal of Epidemiology, 24(2), 117–124. https://doi.org/10.2188/jea.je20130059

Healthy People 2030, U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion. Retrieved [2025, December 2], from https://health.gov/healthypeople/objectives-and-data/social-determinants-health

## Appendix A

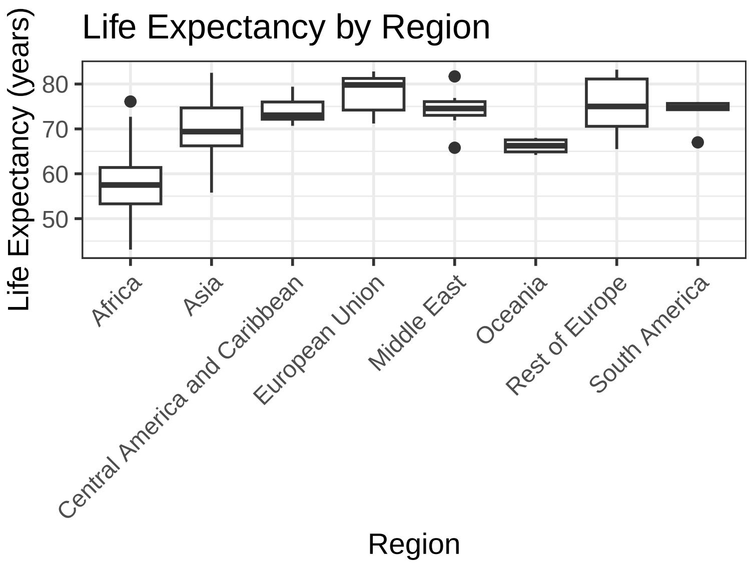


**Figure 4.** Life Expectancy by Region

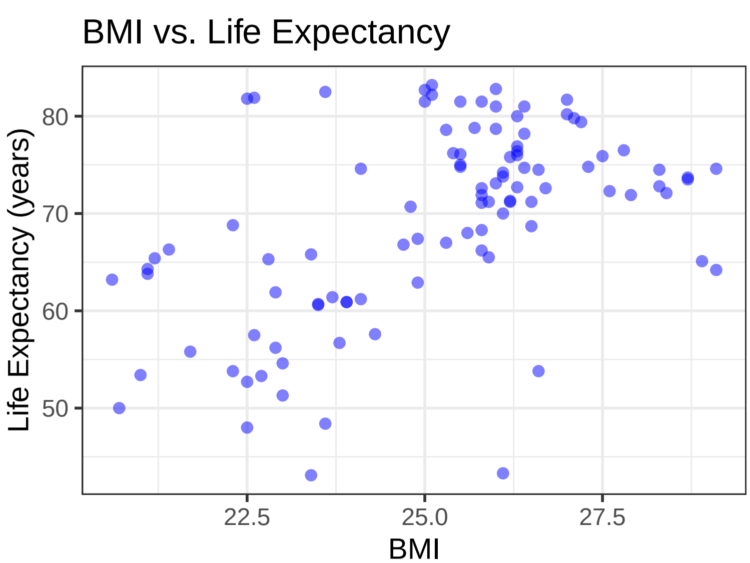


**Figure 5.** BMI vs. Life Expectancy

## Appendix B

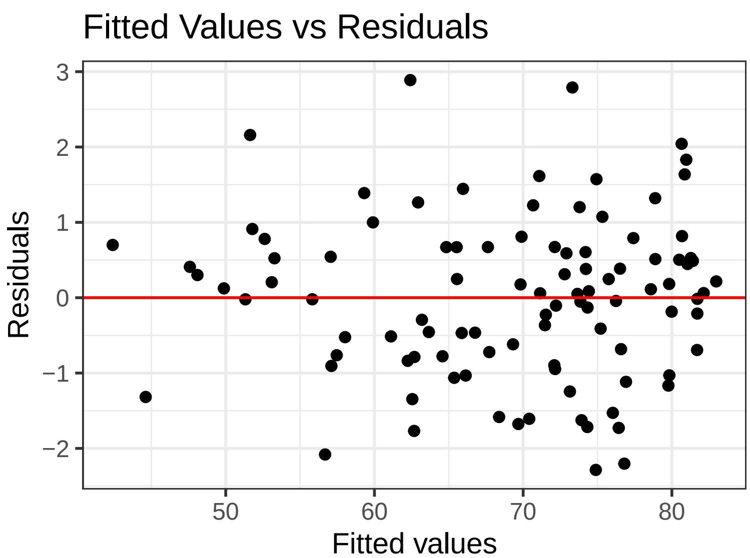


**Figure 6.** Residual plot

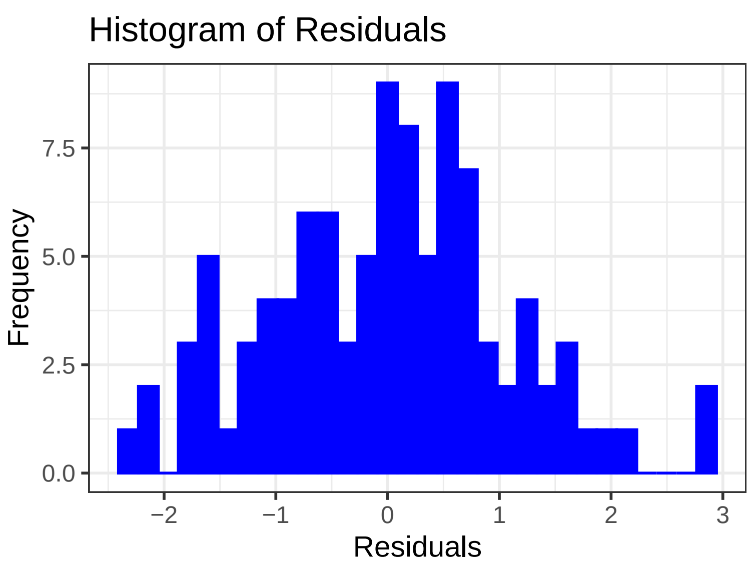


**Figure 7.** Histogram of residuals

## AI Statement

We did not use AI in the completion of this assignment.