

# Modeling GDP Using Health and Socioeconomic Indicators

Erin Xu, Dora Dong, Shencen Cai, Sharon Lam

2025-05-19

## Contributions

**Introduction:** Sharon Lam

**Data Description:** Shencen Cai

**Preliminary Results:** Erin Xu, Dora Dong

**Bibliography:** Everyone

## Introduction

Gross domestic product (GDP) is a widely used measure of a country's economic output, representing the total market value of goods and services produced within its borders over a specified period. It serves as a key indicator of national economic performance and enables comparison across countries and time periods. By standard economic theory, GDP is influenced by components such as consumer spending, government expenditures, investment in capital goods, and net exports. In addition, factors like human capital, infrastructure, technological innovation, and political stability are considered critical for long-term economic growth (Solow, 1956).

This project applies multiple linear regression (MLR) to investigate the extent to which health-related and socioeconomic factors are associated with GDP, with the research question being: *To what extent do government spending on health and socioeconomic resources affect a country's GDP?* Specifically, country status (developed vs. developing), percentage expenditure on health, polio immunization coverage, income composition of resources, years of schooling, and population are used as predictors. These variables, comprising both continuous and categorical data, are examined for their ability to explain cross-country variation in GDP. Health spending, represented by expenditure and immunization coverage, has been shown to enhance productivity, while income composition and national development status reflect broader socioeconomic conditions. Education and population are also recognized as structural drivers of economic capacity.

Estimating a linear model facilitates the quantification of each predictor's contribution to GDP while accounting for the influence of other variables. As economic theory suggests a positive relationship between GDP and improved development indicators, the application of MLR is appropriate for this context. The focus of the analysis is on interpretability, aiming to understand how each factor relates to economic output and to support evidence-based approaches to development and policy planning.

## Data Description

The dataset used in this project is titled *Life Expectancy* (WHO), sourced from *Kaggle* (Kumar, 2018). Its primary usage is for health data analysis. Data collectors combined publicly available data from the *World*

*Health Organization* (WHO) and the *United Nations* (UN), which were gathered through national health departments, structured questionnaires, and annual statistical submissions by participating countries (World Health Organization, n.d.; United Nations, n.d.).

While the dataset was initially intended to examine factors affecting life expectancy, this project selects 7 of the original 22 variables. The sample comprises over 1,600 complete observations, focusing on education, demographic, and socioeconomic indicators relevant to economic growth. These variables align with economic theory, which emphasizes the importance of education, health, and human capital in supporting sustained increases in productivity and GDP.

Multiple linear regression is an appropriate method for analysis, as the dataset consists of independent observations and the model assumes normally distributed residuals, which can be evaluated through diagnostic procedures.

Table 1: Variables used in the model

Variables	Description
GDP	Gross Domestic Product per capita (USD)
Status	Developed or Developing status
Percentage expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita (%)
Polio	Polio immunization coverage among 1-year-olds (%)
Population	Population of the country
Income composition of resources	Human Development Index in terms of income composition (index from 0 to 1)
Schooling	Number of years of schooling (years)

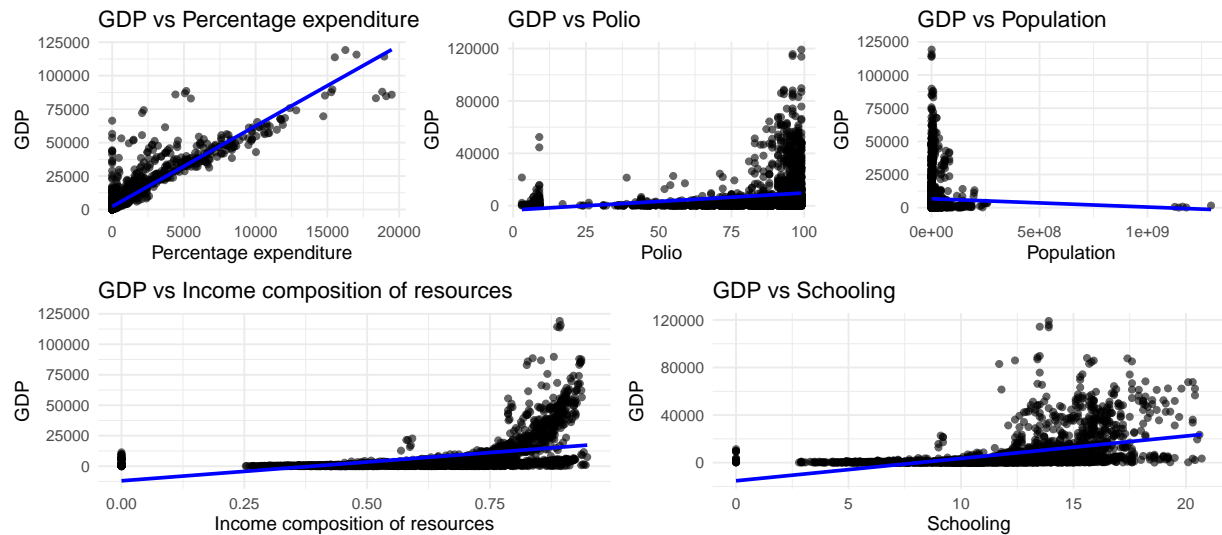
Table 2: Continuous variables summary

Variable	Mean	Std	Min	Q1	Median	Q3	Max
GDP	7284.31	14027.92	1.68	1400.69	2654.32	6891.00	119172.74
Percentage expenditure	4.28	3.76	0.01	1.32	3.45	6.33	87.60
Polio	82.53	23.58	0.00	78.00	91.00	97.00	100.00
Population	3.38e+07	1.19e+08	366.00	3.35e+06	1.26e+07	3.88e+07	1.36e+09
Income composition of resources	0.63	0.15	0.00	0.54	0.67	0.76	0.95
Schooling	10.30	2.98	0.00	8.00	10.30	12.43	20.00

Table 3: Status (categorical variable) frequency

Status	Frequency
Developing	2426
Developed	512
<b>Total</b>	<b>2938</b>

Figure 1: Scatter plots of GDP against numeric predictors



## Preliminary Results

Load data and check if needed to clean

```
all_data <- read.csv("life_expectancy.csv", na.strings = c("", "NA"))
anyNA(all_data)
```

```
## [1] TRUE
```

```
sum(is.na(all_data))
```

```
## [1] 2563
```

```
colSums(is.na(all_data))
```

```
##          Country          Year
##          0          0
##          Status      Life.expectancy
##          0          10
##          Adult.Mortality      infant.deaths
##          10          0
##          Alcohol      percentage.expenditure
##          194          0
##          Hepatitis.B          Measles
##          553          0
##          BMI          under.five.deaths
##          34          0
##          Polio          Total.expenditure
##          19          226
##          Diphtheria          HIV.AIDS
```

```
##              19              0
##              GDP              Population
##              448              652
##      thinness..1.19.years      thinness.5.9.years
##              34              34
## Income.composition.of.resources      Schooling
##              167              163
```

## Decide predictors

```
all_data <- read.csv("life_expectancy.csv")
all_data <- na.omit(all_data)

all_data$Status <- as.factor(all_data$Status)

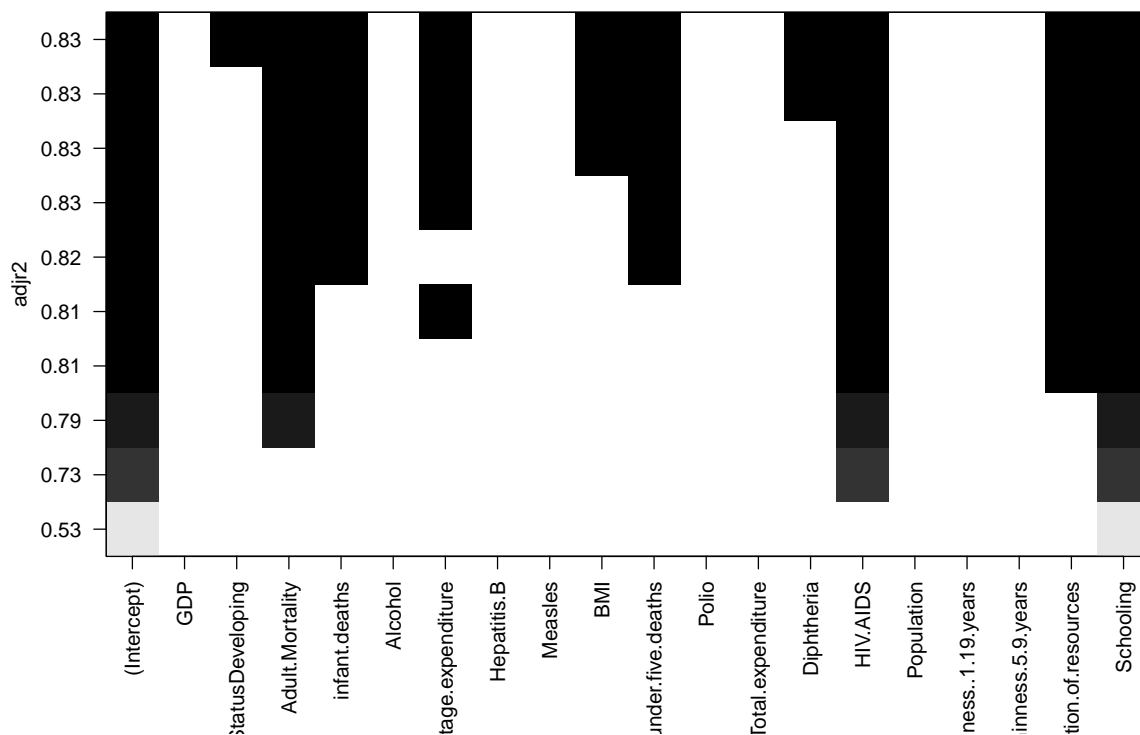
# Clean model
full_model <- lm(GDP ~ Life.expectancy + Status + Adult.Mortality +
  infant.deaths + Alcohol + percentage.expenditure +
  Hepatitis.B + Measles + BMI + under.five.deaths + Polio +
  Total.expenditure + Diphtheria + HIV.AIDS + Population +
  thinness..1.19.years + thinness.5.9.years +
  Income.composition.of.resources + Schooling,
  data = all_data)
summary(full_model)

step_model <- step(full_model, direction = "both")
summary(step_model)
```

## Best subset selection

```
best_model <- regsubsets(Life.expectancy ~ GDP + Status + Adult.Mortality +
  infant.deaths + Alcohol +
  percentage.expenditure + Hepatitis.B + Measles +
  BMI + under.five.deaths + Polio +
  Total.expenditure + Diphtheria + HIV.AIDS +
  Population + thinness..1.19.years +
  thinness.5.9.years + Income.composition.of.resources
  + Schooling, data = all_data, nvmax = 10)

# Summary and plot
best_summary <- summary(best_model)
plot(best_model, scale = "adjr2")
```



We propose the following multiple linear regression model:

$$GDP = E(\log(GDP)) + e =$$

$$b_0 + b_1 \text{PercentageExpenditure} + b_2 \text{Polio} + b_3 \text{Population} + b_4 \text{IncomeCompositionofResources} + b_5 \text{Schooling} + b_6 \text{Status}$$

Get the response and predictors.

```
all_data <- read.csv("life_expectancy.csv")
all_data <- na.omit(all_data)
all_data$log_GDP <- log(all_data$GDP)

response <- all_data$log_GDP
x0 <- all_data$percentage.expenditure
x1 <- all_data$Polio
x2 <- all_data$Population
x3 <- all_data$Income.composition.of.resources
x4 <- all_data$Schooling

all_data$Status <- as.factor(all_data$Status)

model <- lm(response ~ Status + x0 + x1 + x2 + x3 + x4, data = all_data)
summary(model)
```

##

```
## Call:
## lm(formula = response ~ Status + x0 + x1 + x2 + x3 + x4, data = all_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.3838 -0.6281  0.3300  0.8680  2.4999
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.656e+00  2.280e-01  16.037 < 2e-16 ***
## StatusDeveloping 5.719e-02  1.064e-01   0.538  0.591
## x0            3.862e-04  2.030e-05  19.019 < 2e-16 ***
## x1           -7.981e-04  1.459e-03  -0.547  0.585
## x2           -1.866e-10  4.348e-10  -0.429  0.668
## x3            1.398e+00  2.733e-01   5.115 3.51e-07 ***
## x4            2.086e-01  1.871e-02  11.152 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.241 on 1642 degrees of freedom
## Multiple R-squared:  0.5001, Adjusted R-squared:  0.4983
## F-statistic: 273.8 on 6 and 1642 DF,  p-value: < 2.2e-16
```

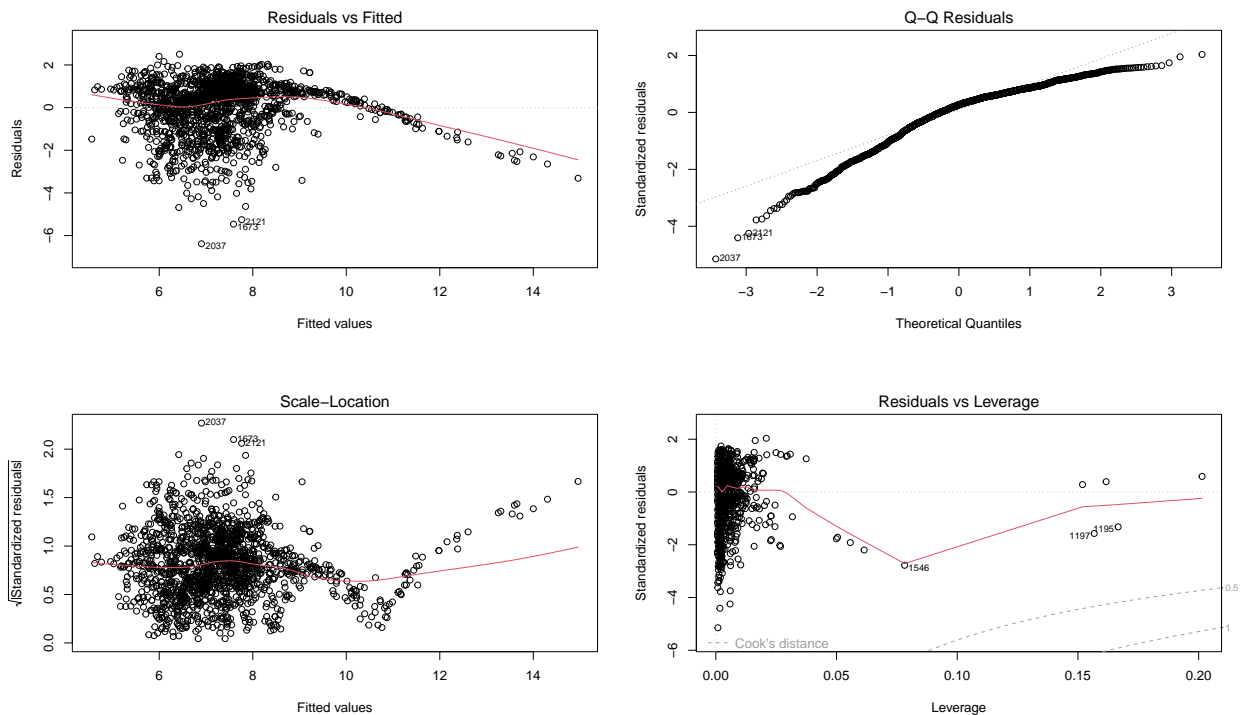
We estimate the deterministic model as

$$\begin{aligned} \hat{GDP} = \exp(\hat{b}_0 + \hat{b}_1 \text{PercentageExpenditure} + \hat{b}_2 \text{Polio} + \\ \hat{b}_3 \text{Population} + \hat{b}_4 \text{IncomeCompositionofResources} + \hat{b}_5 \text{Schooling} + \hat{b}_6 \text{Status}) \end{aligned}$$

by using the *lm* function to find the values of the coefficients that minimize the RSS.

## Residual plots

```
par(mfrow = c(2, 2))
plot(model)
```



## Residual VS each predictor (Regression assumptions)

```
all_data$residuals <- resid(model)
all_data$fitted <- fitted(model)

p1 <- ggplot(all_data, aes(x = `percentage.expenditure`, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Percentage Expenditure")

p2 <- ggplot(all_data, aes(x = Polio, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Polio")

p3 <- ggplot(all_data, aes(x = Population, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Population")

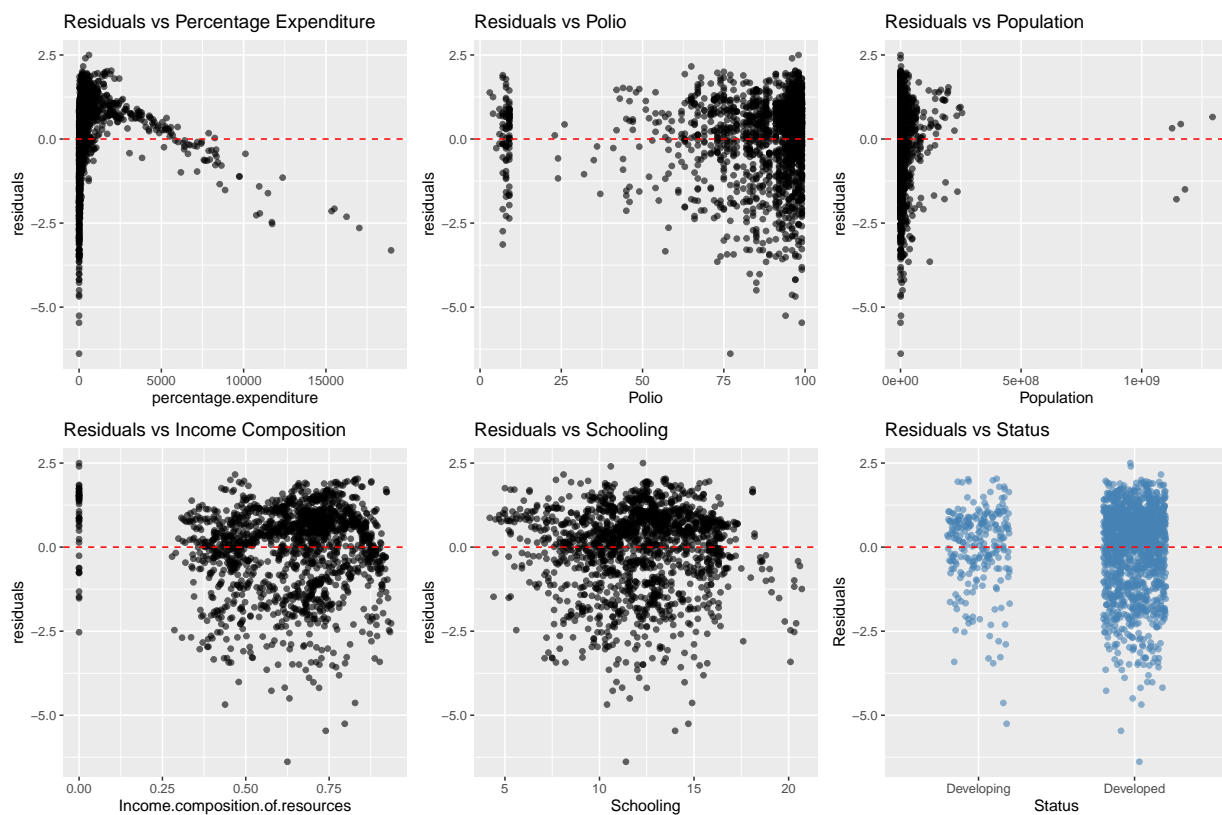
p4 <- ggplot(all_data, aes(x = `Income.composition.of.resources`,
  y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Income Composition")

p5 <- ggplot(all_data, aes(x = Schooling, y = residuals)) +
```

```
geom_point(alpha = 0.6) +
geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
labs(title = "Residuals vs Schooling")
```

```
p6 <- ggplot(all_data, aes(x = as.factor(Status), y = residuals)) +
  geom_jitter(width = 0.2, alpha = 0.6, color = "steelblue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  scale_x_discrete(labels = c("Developing", "Developed")) +
  labs(title = "Residuals vs Status",
       x = "Status",
       y = "Residuals")
```

```
(p1 | p2 | p3) / (p4 | p5 | p6)
```



## Residual VS Fitted plot (Regression assumptions)

```
# Store residuals only once (already done earlier)
all_data$residuals <- resid(model)
all_data$std_residuals <- rstandard(model)

# Q-Q Plot
qq_plot <- ggplot(all_data, aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line(color = "red", linetype = "dashed") +
```



```

labs(title = "Q-Q Plot of Residuals") +
theme_minimal() + # Optional: adds a cleaner theme that may help with rendering
theme(aspect.ratio = 1) # Makes the plot square, which often helps QQ plots display correctly

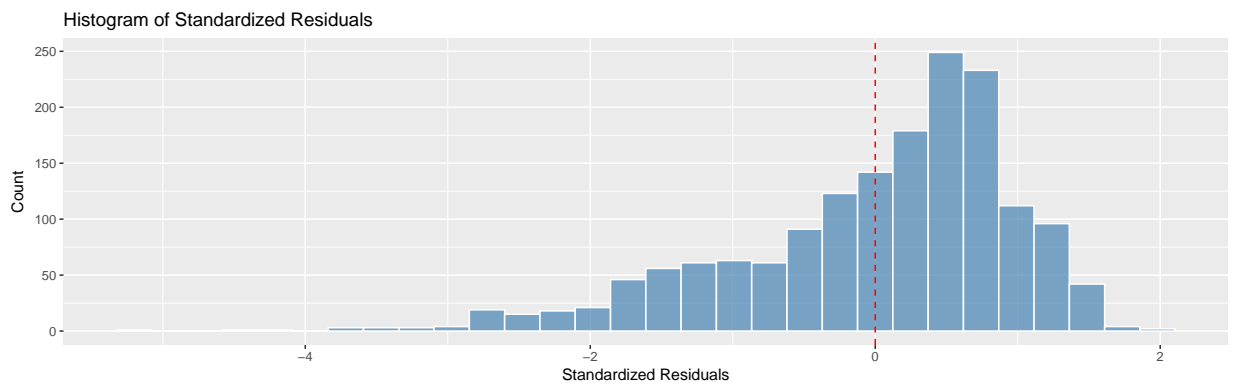
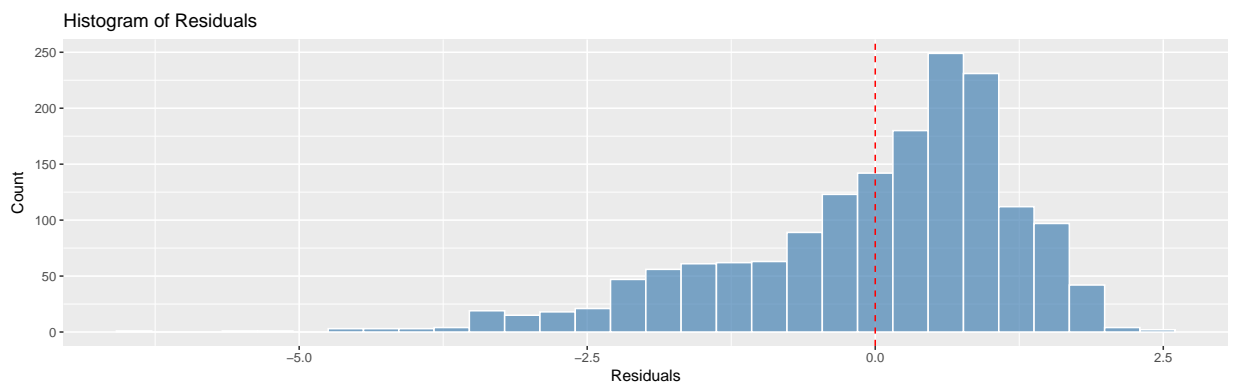
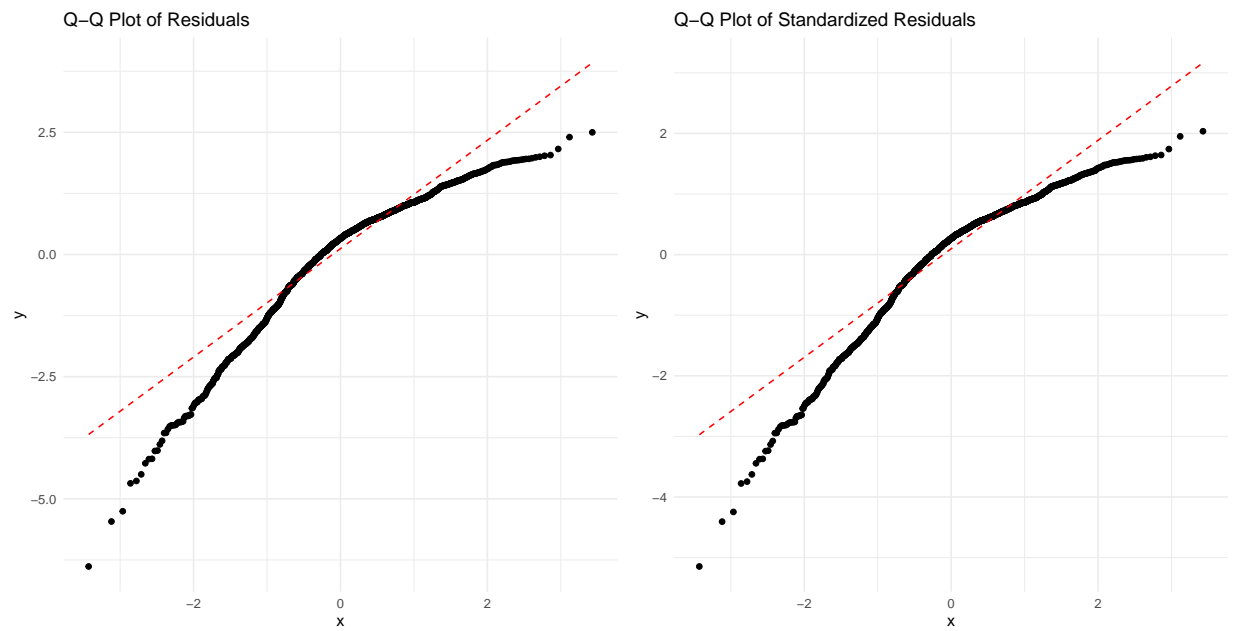
# Standardized Q-Q Plot
std_qq_plot <- ggplot(all_data, aes(sample = std_residuals)) +
  stat_qq() +
  stat_qq_line(color = "red", linetype = "dashed") +
  labs(title = "Q-Q Plot of Standardized Residuals") +
  theme_minimal() +
  theme(aspect.ratio = 1)

# Histogram
hist_plot <- ggplot(all_data, aes(x = residuals)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white", alpha = 0.7) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Count")

std_hist_plot <- ggplot(all_data, aes(x = std_residuals)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white", alpha = 0.7) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Histogram of Standardized Residuals",
       x = "Standardized Residuals",
       y = "Count")

(qq_plot | std_qq_plot) / (hist_plot / std_hist_plot)

```



# Bibliography