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 data set consists of independent observations and the model assumes normally distributed residuals, which can be evaluated through diagnostic procedures.

GDP increases with higher schooling, income composition, and population, though with some spread. Percentage expenditure and polio demonstrate weaker positive trends. Education and income equality appear more strongly linked to economic growth than health spending.

GDP, percentage expenditure, and population are right-skewed, with mostly low values and a few extreme highs. Schooling and income composition are left-skewed, clustering at the high end. Polio rates are highly left-skewed. These patterns demonstrate how the predictors vary and help explain differences in GDP.

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
Total	2938

Figure 1: Scatter plots of GDP against numeric predictors

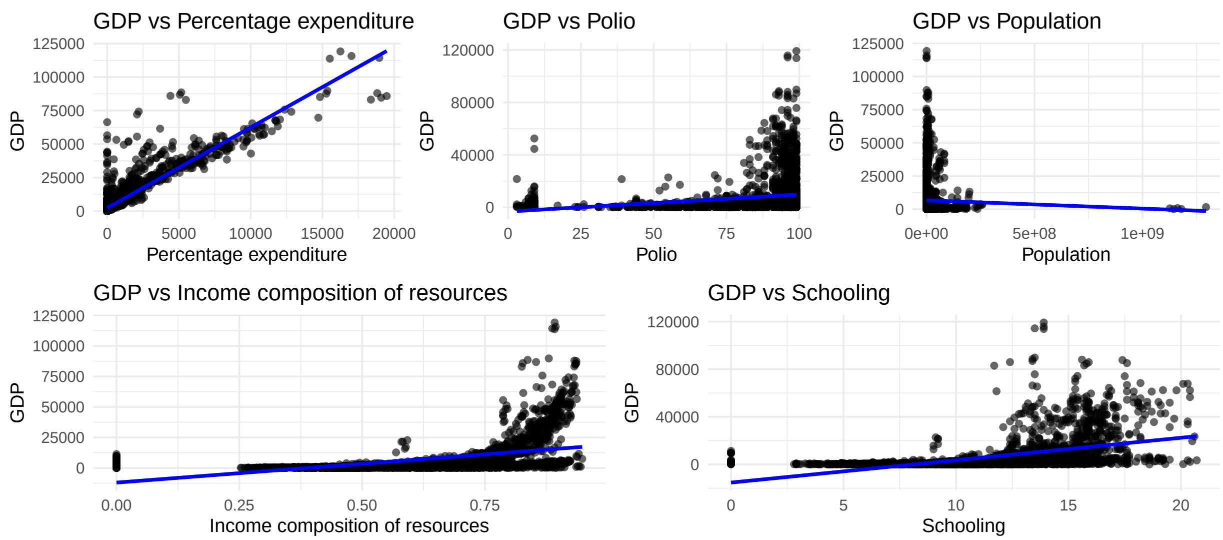
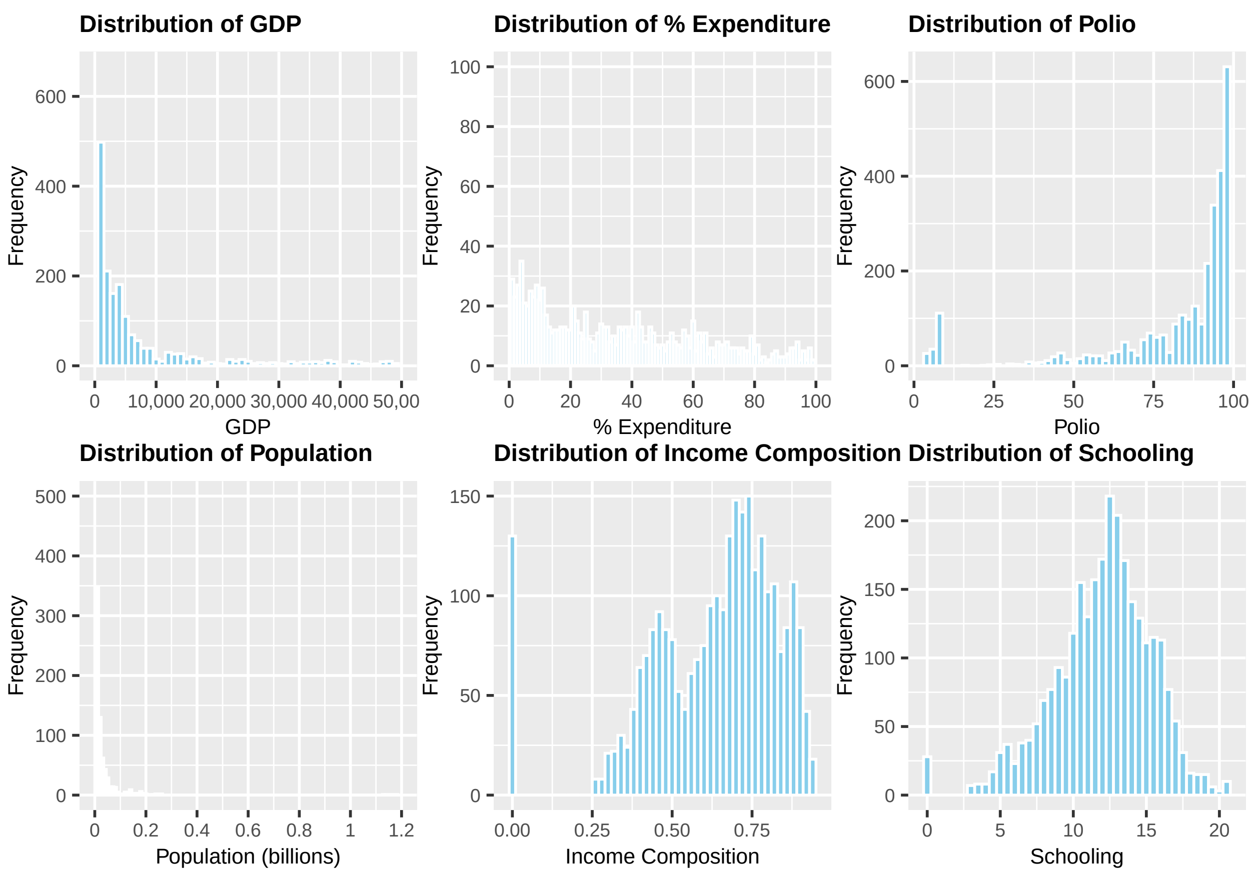


Figure 2: Histogram of GDP & numeric predictors



Preliminary Results

We propose the following multiple linear regression model:

$$\begin{aligned}
GDP &= \mathbb{E}[\log(GDP)] + e \\
&= b_0 + b_1 \cdot \text{PercentageExpenditure} + b_2 \cdot \text{Polio} + b_3 \cdot \text{Population} \\
&\quad + b_4 \cdot \text{IncomeCompositionOfResources} + b_5 \cdot \text{Schooling} + b_6 \cdot \text{Status}
\end{aligned}$$

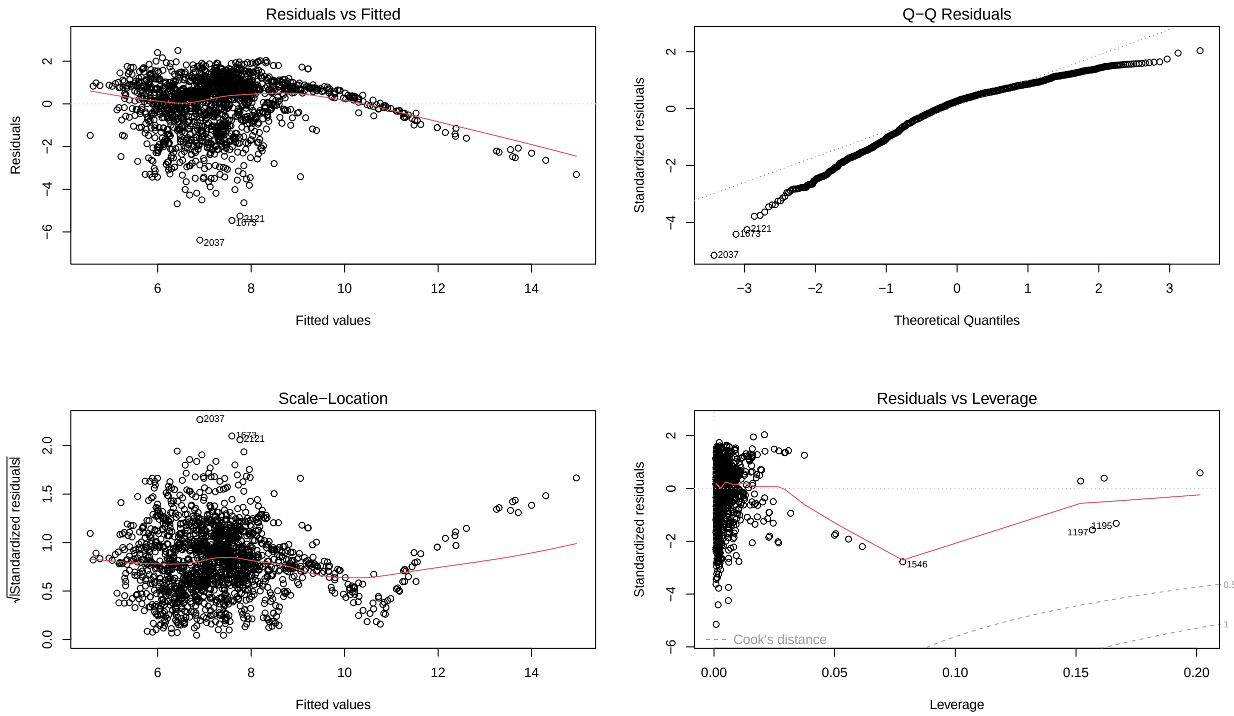
Get the response and predictors:

```
##
## 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:

$$\begin{aligned}
G\hat{D}P = \exp(&\hat{b_0} + \hat{b_1} \cdot \text{PercentageExpenditure} + \hat{b_2} \cdot \text{Polio} + \hat{b_3} \cdot \text{Population} \\
&+ \hat{b_4} \cdot \text{IncomeCompositionOfResources} + \hat{b_5} \cdot \text{Schooling} + \hat{b_6} \cdot \text{Status})
\end{aligned}$$

Figure 3: Linearity and homoscedasticity graphs



Initially, the distribution of GDP was heavily right-skewed due to a small number of countries with disproportionately large economies. Residual plots also showed signs of heteroscedasticity, violating regression assumptions. To address this, we applied a log transformation to the response variable, which preserved the interpretability of a linear model while improving the spread of residuals.

The adjusted R-squared of the transformed model was 0.9258, indicating that 92.6% of the variation in GDP is explained by the model. Among numerical predictors, percentage expenditure on health and schooling were both statistically significant ($p < 0.001$), suggesting strong positive relationships with GDP. This supports the idea that education and health investment enhance human capital and economic productivity (Radcliffe, Raghupathi). For the categorical predictor Status (Developed vs. Developing), the p-value was approximately 0.0038, indicating that developed countries tend to have significantly higher GDPs after accounting for other variables. In contrast, polio immunization and population size were not statistically significant, implying weaker associations.

The residual plots assess linearity and constant variance assumptions. Residuals were mostly centered around zero, but a slight V-shape indicates some remaining heteroscedasticity, especially at the lower and higher ends of fitted GDP. This suggests more stable residuals in middle-income countries, while richer and poorer countries show more unpredictable patterns. For example, the U.S., Qatar, and Luxembourg all have high GDPs, but for very different reasons (tech, gas, or tax policy). Similarly, lower-income countries may have inconsistent or less reliable data.

The Q-Q plot showed that residuals were somewhat normal, but skewed left. Residuals versus individual predictors showed random scatter for income composition and schooling, but some structure for expenditure, population, and polio, suggesting potential nonlinearities.

Overall, the model shows strong evidence that educational and economic factors influence GDP, though future improvements could explore nonlinear modeling or methods to address residual heteroscedasticity.

Figure 4: Normality graphs

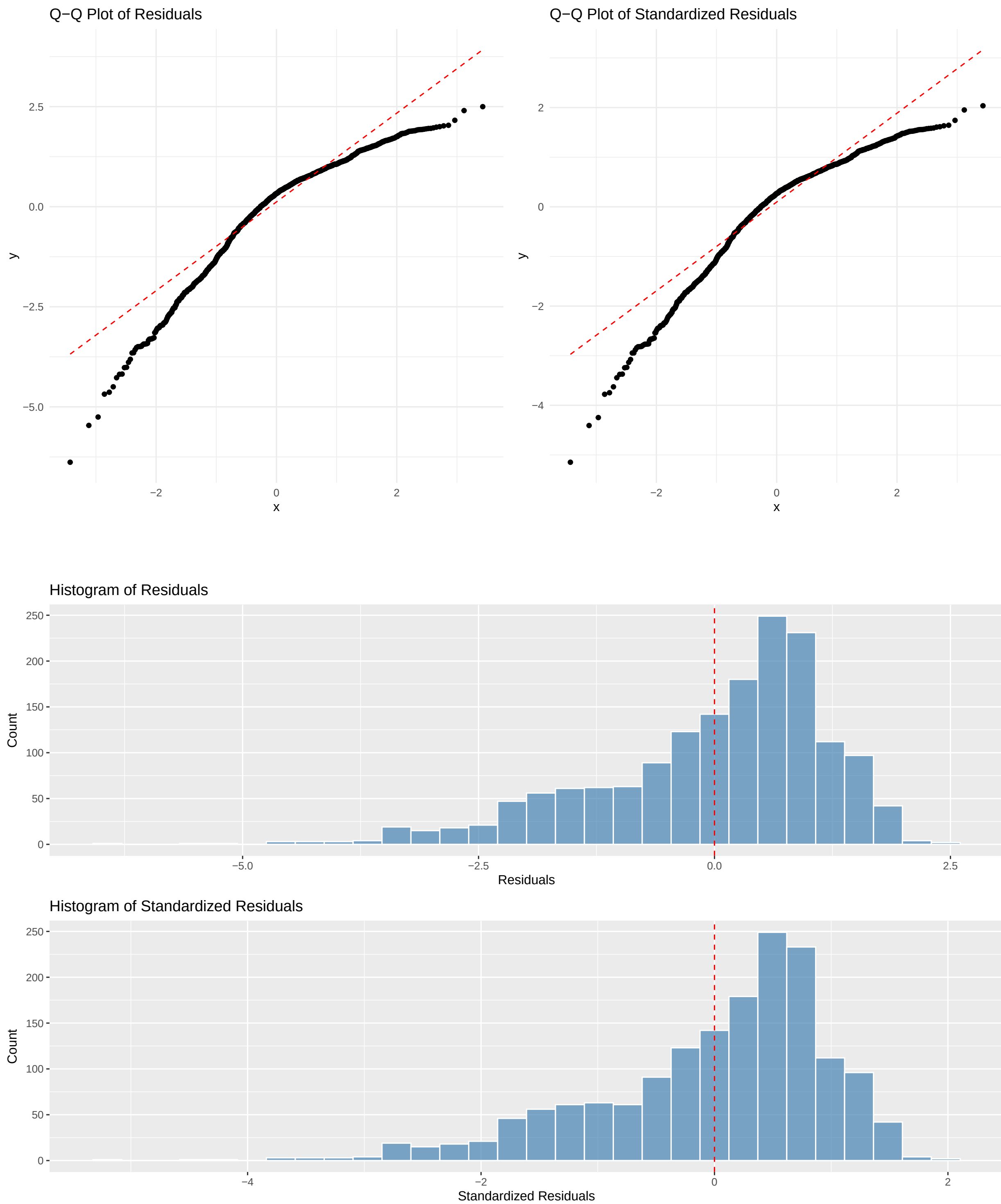
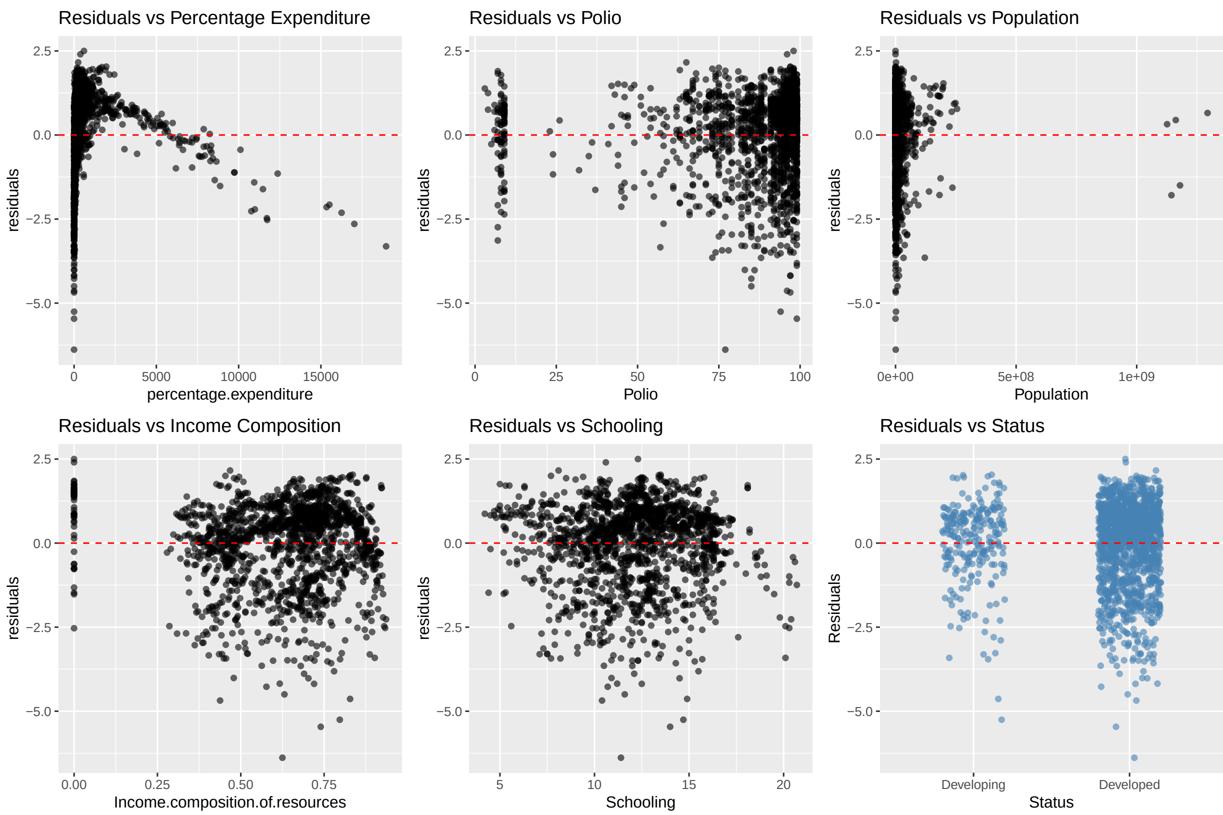


Figure 5: Residual vs each predictor



References

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

Organisation for Economic Co-operation and Development. n.d. "Nominal Gross Domestic Product (GDP)." https://www.oecd.org/en/data/indicators/nominal-gross-domestic-product-gdp.html?oecdcontrol-d7f68dbeee-var3=2023.

Radcliffe, Brent. n.d. "How Education and Training Affect the Economy." https://www.investopedia.com/articles/economics/09/education-training-advantages.asp.

Raghupathi, Viju, and Wullianallur Raghupathi. 2020. "Healthcare Expenditure and Economic Performance: Insights from the United States Data." *Frontiers in Public Health* 8: 156. https://doi.org/10.3389/fpubh.2020.00156.

Solow, Robert M. 1956. "A Contribution to the Theory of Economic Growth." *The Quarterly Journal of Economics* 70 (1): 65–94. https://doi.org/10.2307/1884513.

United Nations. n.d. "UN Data." https://data.un.org.

World Health Organization. n.d. "Global Health Observatory (GHO) Data Repository." https://www.who.int/data/gho.