Modeling GDP Using Health and Socioeconomic Indicators

Erin Xu, Dora Dong, Shencen Cai, Sharon Lam

2025-06-12

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. From economic theory, GDP is influenced by components such as consumer spending, government expenditures, investment in capital goods, and net exports. Factors like human capital, infrastructure, technological innovation, and political stability are also vital.

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 the combination of continuous and categorical predictors used to explain the extent in which they affect GDP in countries around the world. Health spending, represented by percent of a country's expenditure and polio immunization coverage, has been shown to enhance productivity, and income composition and national development status reflect broader socioeconomic conditions. Education and population are also recognized as structural drivers of economic growth because educated workers increases human capital, research and innovation for better products, processes and overall economic advancement.

As economic theory suggests a positive relationship between GDP and improved development indicators, and estimating a linear model allows us to quantify the individual contribution of each predictor to GDP while controlling for the others, a positive relationship between GDP and indicated predictors can be expected. The focus of this analysis is on interpretability, to understand how each predictor 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.). The sample comprises over 1,600 complete observations, focusing on education, demographic, and socioeconomic indicators relevant to economic growth.

While the dataset was initially intended to examine factors affecting life expectancy, this project selects 7 of the original 22 variables that align with economic theory, which emphasizes the importance of education, health, and human capital in supporting sustained increases in GDP.

The preliminary model is prone to multiple violations of model assumption, but multiple linear regression is still an appropriate method for analysis, as the scatterplots of the response and each predictor show a

huge potential for linear association, constant error variance, and uncorrelated and normal errors, through diagnostic procedures like predictor transformations.

Table 1: Variables used in the model

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

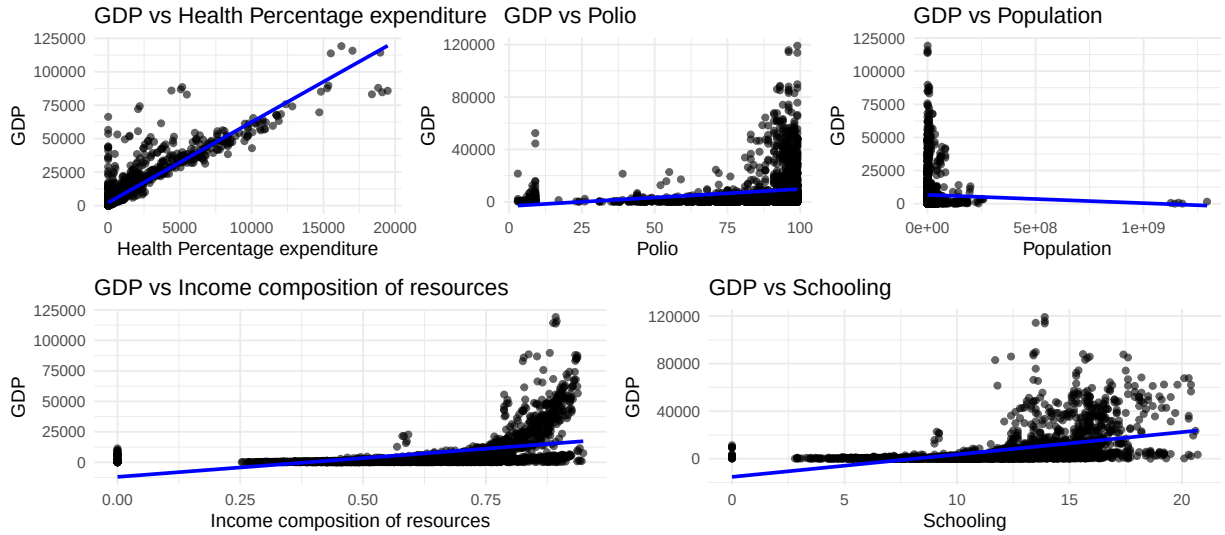
Table 2: Continuous variables summary

Variable	Mean	Std	Min	Q1	Median	Q3	Max
GDP	7483.16	14270.17	1.68	463.94	1766.95	5910.81	119172.74
Percentage expenditure	738.25	1987.91	0.01	4.69	64.91	441.53	19479.91
Polio	82.55	23.43	3.00	78.00	93.00	97.00	99.00
Population	1.28e+07	6.10e+07	34.00	1.96e+05	1.39e+06	7.42e+06	1.29e+09
Income composition of resources	0.63	0.21	0.00	0.49	0.68	0.78	0.95
Schooling	11.99	3.36	0.00	10.10	12.30	14.30	20.70

Table 3: Status (categorical variable) frequency

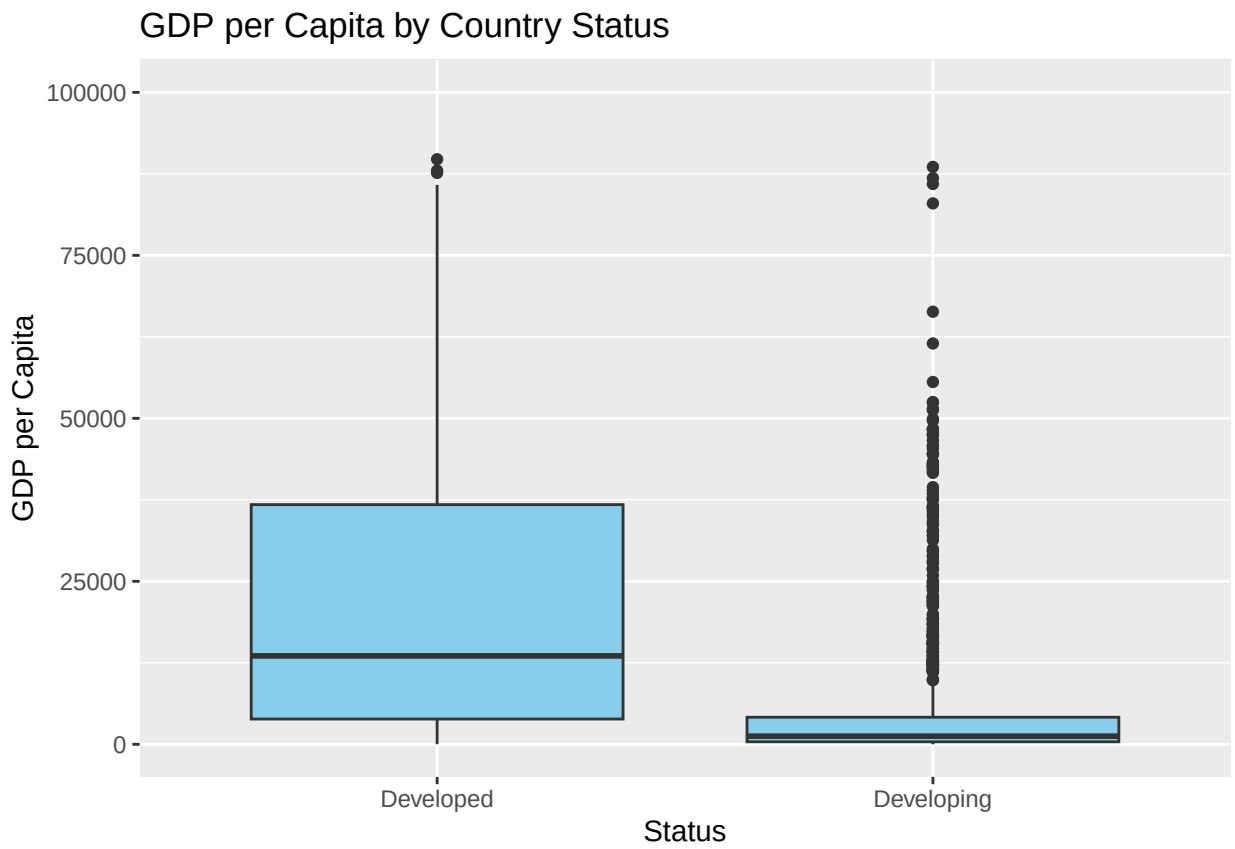
Status	Frequency
Developing	2426
Developed	512
Total	2938

Figure 1: Scatter plots of GDP against numeric predictors



GDP increases as health percentage expenditure does in a rather compelling linear manner, although the clustering near the lower ends of the domain is concerning due to outliers in countries that are experiencing geopolitical turmoil. GDP and polio as well as income composition of resources and schooling demonstrate weaker positive trends, looking more quadratic, most likely with leverage points at the tails. It's clear that there are bad leverage points in GDP and population. In the context of geography and the complexity of individual states, one can suspend their disbelief easily about certain leverage points, but nonetheless, these need to be dealt with to provide a more accurate prediction of GDP with the set predictors at hand.

Figure 2: Histogram of GDP & numeric predictors



GDP, percentage expenditure, and population are strongly right-skewed, with mostly low values. Schooling and income composition are slightly left-skewed, clustering at the high end. Income composition is also bimodal, although there is potential for a bell-curve-like shape. Polio rates are highly left-skewed. Since country status is a categorical predictor, the boxplot graph is better suited to evaluating the normality assumption. Developed countries have a higher median GDP per capital and wider IQR range, which indicates greater variability than developing countries, which violates a model assumption. Also, developing countries appear highly skewed to the right, with most countries clustered at low GDP per capital values. The long tail of outliers stretching upward indicate a few developing countries with relatively high GDPs. In context, these outlier countries like Malaysia, Mexico, or Turkey are often classified more deeply as upper-middle-income economies (World Bank) because they clearly act as outliers compared to the rest of the developing world.

Preliminary model results and diagnostics

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:

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

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 evidence that educational and economic factors influence GDP, though remaining issues with heteroscedasticity and non-normal residuals should be better resolved, as these violations weaken ordinary least squares (OLS) model accuracy. Employing further transformations like Box-Cox, investigating covariance, colinearity, the removal of bad leverage points and outliers, comparison of different models through F-test and analysis of variance (ANOVA), or even exploring weighted least squares (WLS) are key stratagem to confidently predict the expected GDP in a country.

Figure 3: Linearity and homoscedasticity graphs

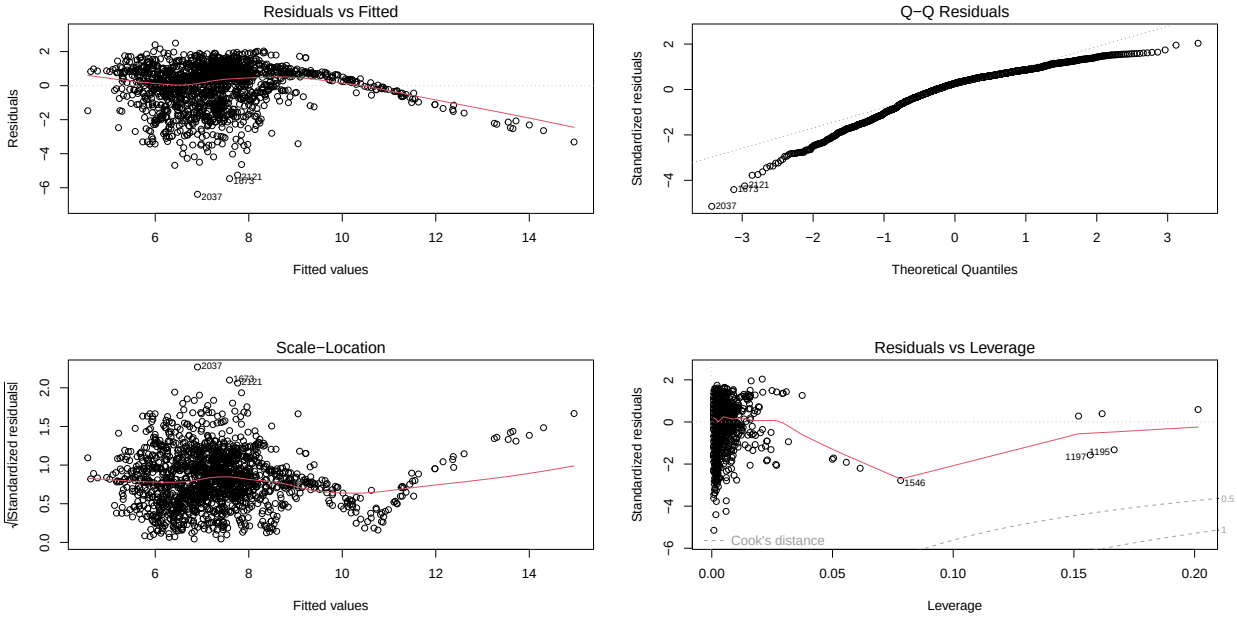


Figure 4: Residual vs each predictor

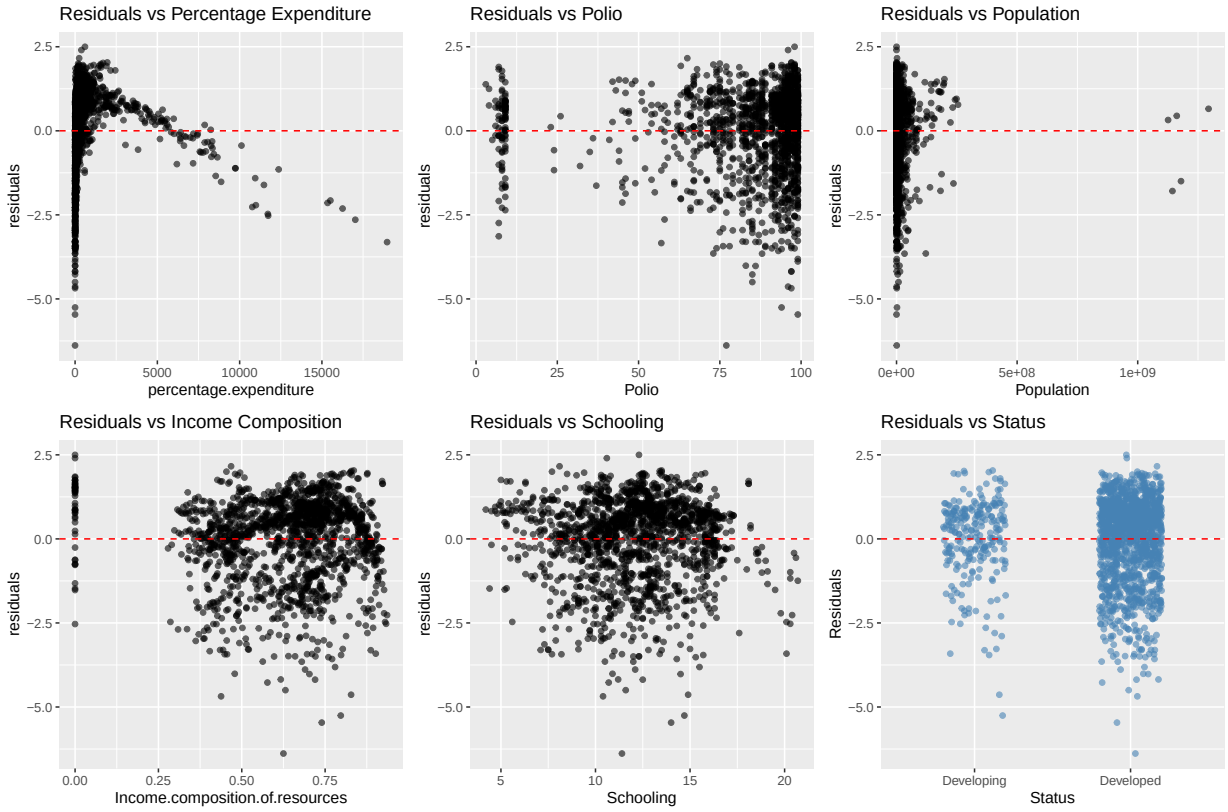
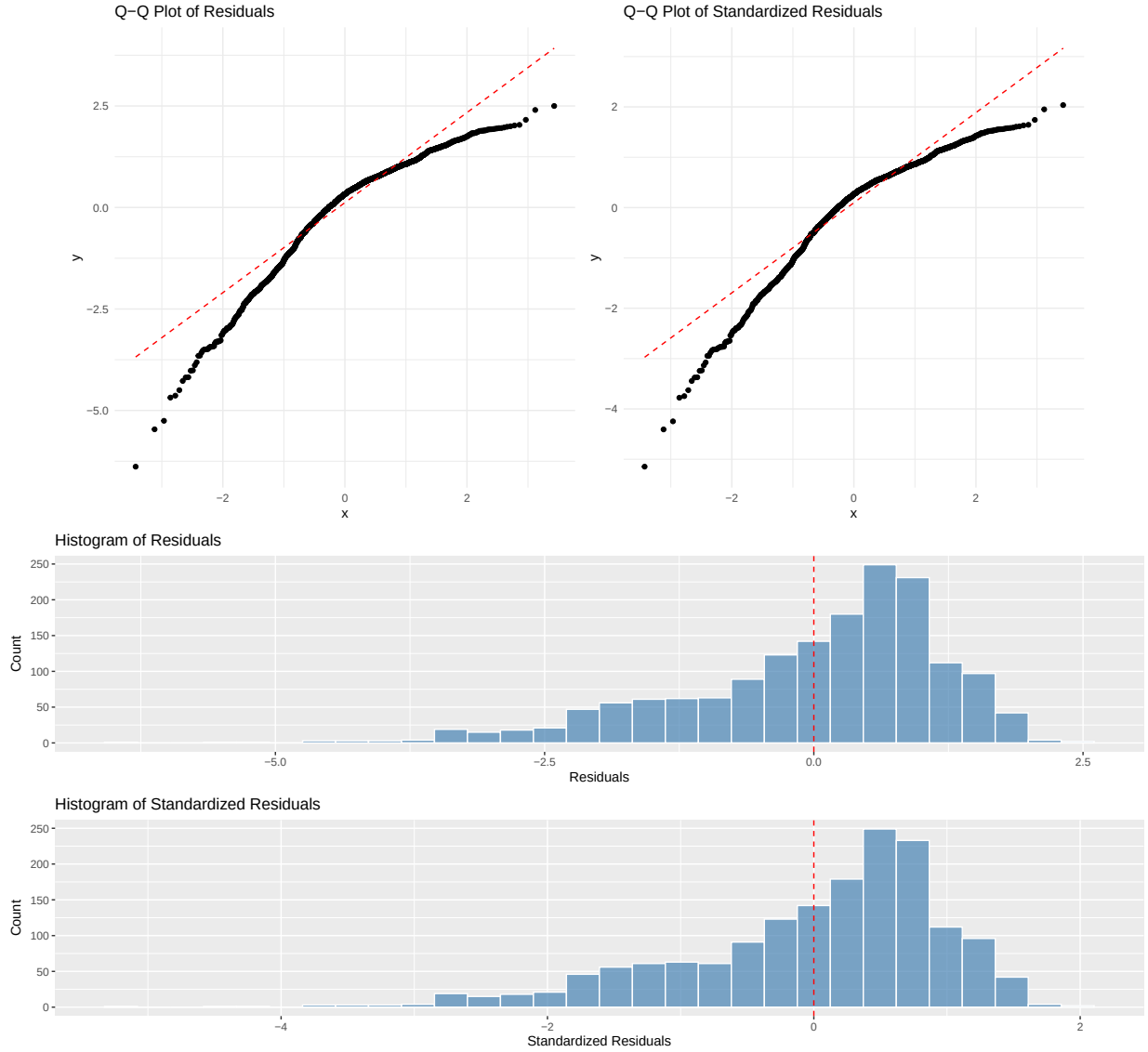


Figure 5: Normality graphs



Model selection

From the conclusion in preliminary model diagnostics, the model violated linear regression assumptions and would benefit from modifications.

The first attempt to fit to the population model was to fit all available predictors from the dataset without transformation using OLS. We then performed best subset selection and stepwise selection. Stepwise selection was done using *step*() based on Akaike Information Criterion (AIC) and Bayesian Information Critertion (BIC) which filtered the complexity and goodness of fit of the model. Best subset selection was done using *regsubsets*() from the **leaps** package, which tries all combinations of up to 10 predictors and evaluates them based on adjusted R^2. This is **Model 1**. The main issue here was the predictor choice that had been determined by the criterion: Both total expenditure on health and percentage expenditure on health were included in the model, however this is not meaningful as these predictors are just different perspectives of

health and GDP spending. Therefore, population replaced the percentage expenditure since population had previously a strong indication on its influence of GDP, and this became **Model 2**.

Then, since this model violated normality of residuals and heteroscedasticity, **Model 3** is the variable transformation of **Model 2**, where Box-Cox was applied to the response, with the result of maximized log-likelihood power parameter $\lambda = 0.101$. This significantly improved the normality and variance stability of the residuals, however, a curvature on the residual and fitted values plot was still observed.

In order to address the remaining violation, **Model 4** applies transformations on the predictors. Polio and population scatterplots from the introduction exploratory data analysis indicated a potential nonlinear relationship with GDP, therefore, a quadratic transformation was applied to these two predictors, although improvement on normality violations were not representative in diagnostic graphs.

Therefore, **Model 5** and **Model 6** are the combination of both response and predictor transformations. For these two final candidate models, multicollinearity in the predictors was also assessed through the Variance Inflation Factor (VIF). Both candidates returned VIF values of below 5, a strong indication that multicollinearity was not a concern in either model. In lieu of the research question, multicollinearity is unlikely to be a major concern as these predictors capture distinct phenomena that measure different social mechanisms. Also, predictors that were already expressed as a percentage of a population removes the common size factor that would otherwise cause co-linearity in cross-country data.

Model 6 is **Model 5** with an additional observation: Before transforming the population variable, the residual vs. leverage plot exhibited several points with high leverage. This would indicate influence on the fitted line to stray from the population line. As a result, a logarithmic transformation was applied on the population predictor here. This transformation substantially reduced the number and magnitude of bad high-leverage points and indicates that the model became less sensitive to extreme values in the predictors.

Table 4: Diagnostic metrics for Models 5 and 6

Metric	Model 5	Model 6
R^2	0.4209	0.4249
Adjusted R^2	0.4188	0.4288
AIC	8085.524	8074.323
BIC	8128.788	8117.587

Table 5: Explanation of transformed variables

Transformation	Variable(s)	Motivation
Box–Cox ($\lambda \approx 0.101$)	Response: `GDP`	Heavy right-skew and non-constant variance in raw GDP; log-like Box–Cox stabilizes variance and improves normality.
Quadratic term	`Polio`^{2}, `Schooling`^{2}, `Total.expenditure`^{2}	Scatterplots showed curvature; adding a squared term captures diminishing returns visible in the data.
Log scale	Population to `log(Population)`	Original residual-vs-leverage plot had extreme leverage for the largest countries; the log scale compresses the range and removes those high-influence points without discarding data.

All other predictors remained on their natural scale because residual plots and VIF values (< 5) did not indicate further non-linearity or multicollinearity problems. After removal of the clear outliers, diagnostics (Figure 3, 4) showed homoscedastic residuals and no further influential cases above conventional cut-offs. Dropped predictors were 'percentage.expenditure', because it duplicates health-spending information already captured by `Total.expenditure` and inflates collinearity. Also, other health, demographic, and mortality variables eliminated by stepwise AIC lacked statistical significance ($p > 0.05$) and raised VIFs > 5 when retained.

Table 6: ANOVA Table

Predictor	Df	Sum Sq	Mean Sq	F value	Pr($> F$)
Status	1	3194.3	3194.3	584.3467	2.2e-16
Total expenditure^2	1	308.6	308.6	52.9796	5.7e-13
Polio	1	682.4	682.4	117.1461	2.2e-16
log-population	1	10.3	10.3	1.7096	0.1836
Income composition	1	4114.7	4114.7	706.3467	2.2e-16
Schooling^2	1	75.9	75.9	13.0305	0.000317
Residuals	1416	8248.6	5.8	—	—

Table 7: Comparison of each model

Model	R^2	Adjusted R^2	AIC	BIC
Model 1	0.9262	0.9259	31226.955	31270.218
Model 2	0.3124	0.3099	34906.658	34949.922
Model 3	0.4209	0.4188	8085.575	8128.838
Model 4	0.3483	0.3459	34818.269	34861.533
Model 5	0.4209	0.4188	8085.524	8128.788
Model 6	0.4249	0.4288	8074.323	8117.587

In addition to examining diagnostic plots, overall fit and complexity of the final two models were evaluated with certain metrics. **Model 6** has a larger Adjusted R^2 value and smaller AIC and BIC values. To further identify which points were problematic, we used several standard diagnostic criteria: outliers were identified as observations with standardized residuals exceeding ± 4, high leverage points as those with hat values greater than $\frac{2(p+1)}{n}$, and influential points based on Cook's distance exceeding $\frac{4}{n}$ or DFFITS values greater than $2\sqrt{\frac{p+1}{n}}$. Based on this, we removed observations 126, 2121, 92, and 1233 as outliers, which improved the normality and homoscedasticity of the residuals, and the Q-Q plot more closely followed the 45-degree line. However, we chose not to remove high leverage or influential points unless they were also clear outliers. Removing these points led to curvature in the residuals vs fitted plot and deviations from normality in the Q-Q plot. This suggests that these points may contain genuine structure in the data necessary for the model to adequately capture relationships.

To test the overall significance of the model, we conducted an F-test with the following hypotheses:
H_0: All slope coefficients are equal to zero (the model has no explanatory power), $\beta_0 = \beta_1 = ... = \beta_k = 0, \forall k \in \mathbb{Z}^+$.

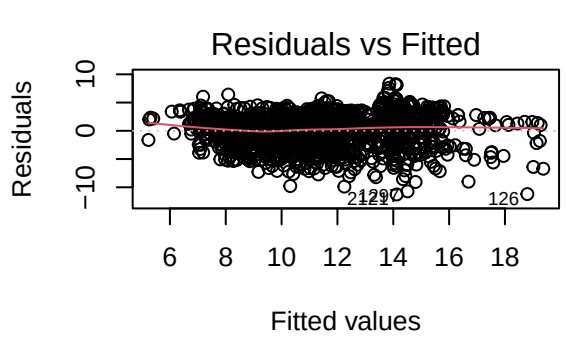
H_a: At least one slope coefficient is non-zero (the model explains some variation in the response), $\exists i \in \mathbb{Z}^+ st. \beta_i \neq 0$.

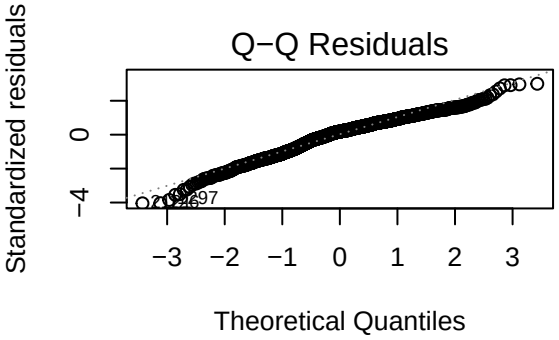
The overall F-statistic for **Model 6** is 209.8 with 6 and 1640 degrees of freedom, and a corresponding p-value of $< 2.2e-16$. This indicates that the model is statistically significant overall, meaning we can reject the null hypothesis. There is least one of the predictors contributes meaningfully to explaining variation in Box-Cox transformed GDP. The extremely small p-value provides strong evidence against the null hypothesis that all slope coefficients are zero.

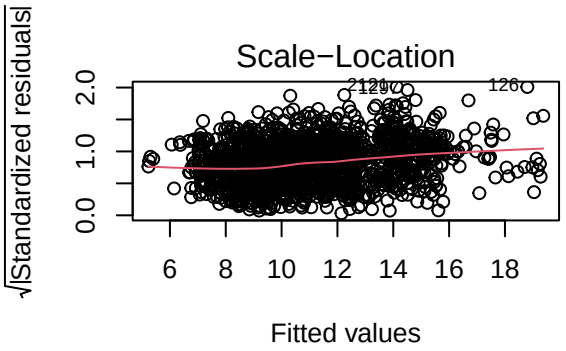
Based on best meeting these criteria, and solving violation assumptions, Model 6 is our final model for inference and interpretation of GDP by health and socioeconomic predictors.

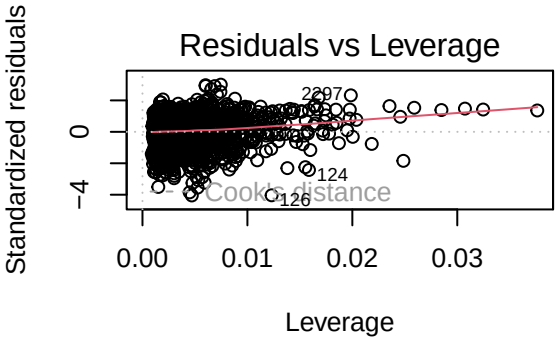
Figure 6: Basic diagnostics for Model 6

```
##
## Call:
## lm(formula = bc_GDP ~ Status + I(Total.expenditure^2) + Polio +
##     I(log(Population)) + Income.composition.of.resources + I(Schooling^2),
##     data = all_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2678  -1.8229   0.5069   1.9800   8.3447
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      6.227949   0.603082  10.327  < 2e-16 ***
## StatusDeveloping                -1.175649   0.235843  -4.985 6.86e-07 ***
## I(Total.expenditure^2)           0.003817   0.002451   1.557    0.120
## Polio                           -0.003145   0.003273  -0.961    0.337
## I(log(Population))              -0.001182   0.025241  -0.047    0.963
## Income.composition.of.resources  4.273179   0.596040   7.169 1.14e-12 ***
## I(Schooling^2)                   0.021669   0.001756  12.342  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.792 on 1642 degrees of freedom
## Multiple R-squared:  0.4249, Adjusted R-squared:  0.4228
## F-statistic: 202.2 on 6 and 1642 DF,  p-value: < 2.2e-16
```










```
##                            Status          I(Total.expenditure^2)
##                          1.473847                        1.102295
##                            Polio             I(log(Population))
##                          1.141679                        1.017460
## Income.composition.of.resources                I(Schooling^2)
##                          2.518552                        2.985993
```

```
##        df       AIC
## model1  8 31226.955
## model2  8 34906.658
## model3  8  8085.575
## model4  8 34818.269
## model5  8  8085.524
## model6  8  8074.323
```

```
##        df       BIC
## model1  8 31270.218
## model2  8 34949.922
## model3  8  8128.838
## model4  8 34861.533
## model5  8  8128.788
## model6  8  8117.587
```

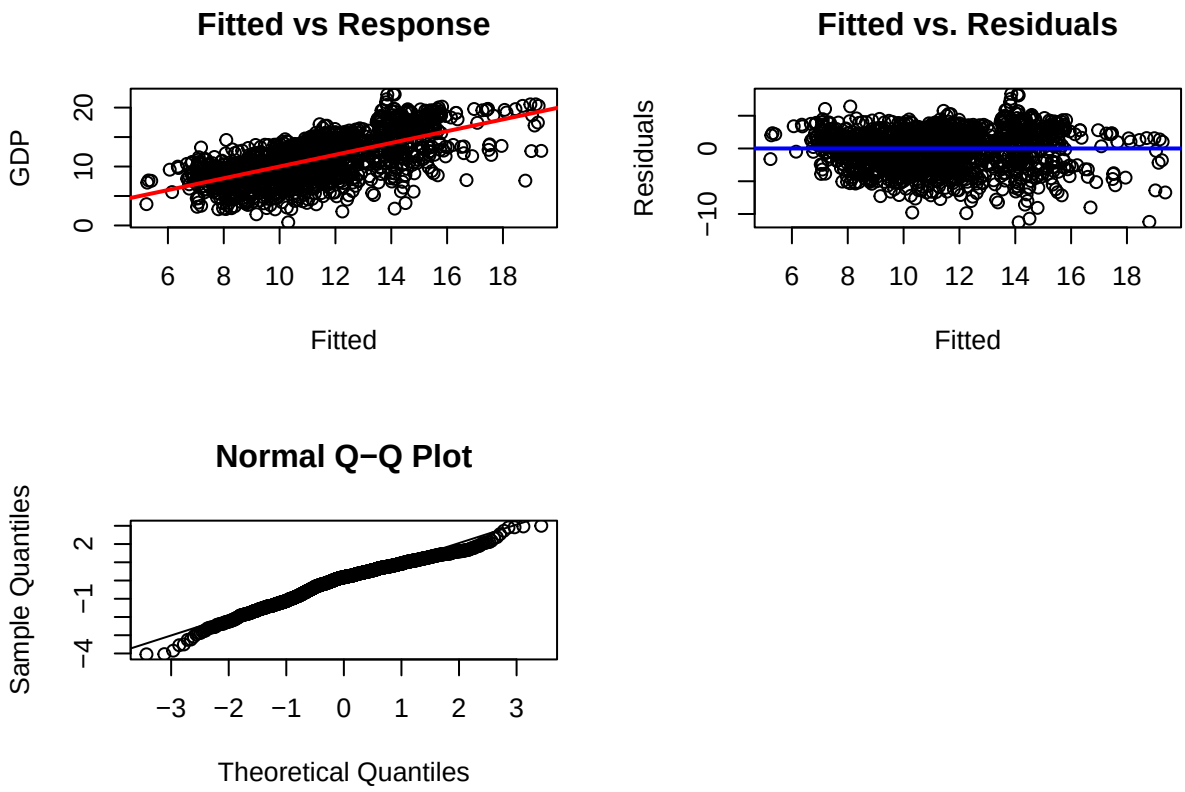
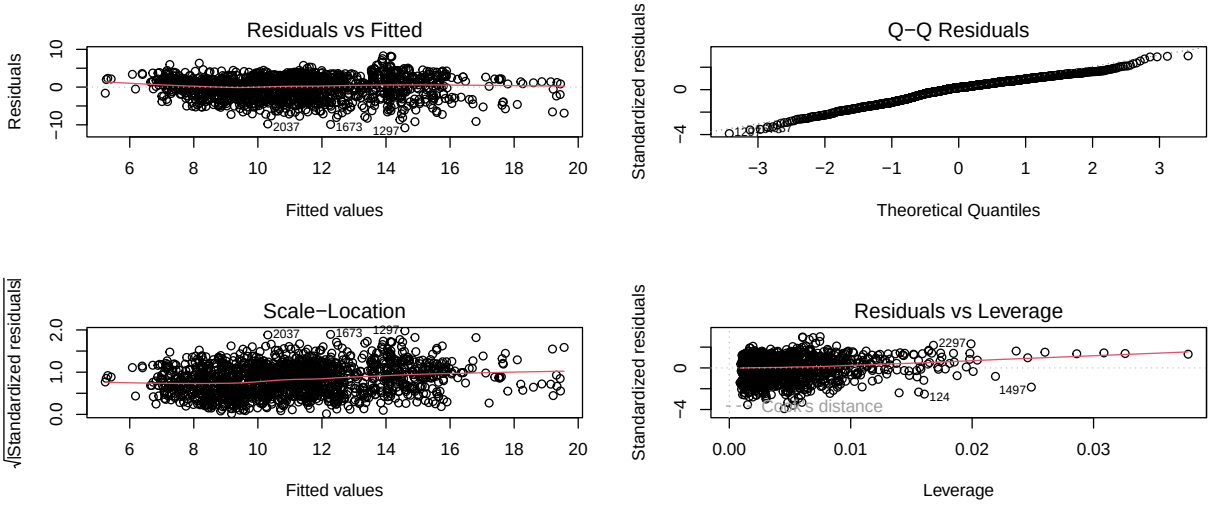


Figure 7: Diagnostics for Model 6

```
##
## Call:
## lm(formula = bc_GDP ~ Status + I(Total.expenditure^2) + Polio +
##     I(log(Population)) + Income.composition.of.resources + I(Schooling^2),
##     data = clean_data6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7965  -1.8298   0.4982   1.9687   8.2887
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      6.329789   0.597670  10.591  < 2e-16 ***
## StatusDeveloping                -1.245231   0.234032  -5.321 1.18e-07 ***
## I(Total.expenditure^2)           0.003577   0.002429   1.473    0.141
## Polio                           -0.003347   0.003242  -1.032    0.302
## I(log(Population))              -0.002211   0.025005  -0.088    0.930
## Income.composition.of.resources  4.154936   0.591220   7.028 3.07e-12 ***
## I(Schooling^2)                   0.022231   0.001746  12.733  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.765 on 1640 degrees of freedom
## Multiple R-squared:  0.4343, Adjusted R-squared:  0.4322
```

```
## F-statistic: 209.8 on 6 and 1640 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: bc_GDP
##                                  Df  Sum Sq Mean Sq  F value    Pr(>F)
## Status                            1  4376.3  4376.3 572.3355 < 2.2e-16 ***
## I(Total.expenditure^2)            1   271.9   271.9  35.5587 3.025e-09 ***
## Polio                             1   292.1   292.1  38.1981 8.048e-10 ***
## I(log(Population))                1     4.7     4.7   0.6124     0.434
## Income.composition.of.resources   1  3441.1  3441.1 450.0292 < 2.2e-16 ***
## I(Schooling^2)                    1  1239.6  1239.6 162.1216 < 2.2e-16 ***
## Residuals                      1640 12540.1     7.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Final model inference and results

The final regression model consists of six predictors which were selected and transformed in a way to accurately address the extent to which government spending on health and socioeconomic resources affects a country's GDP. Quadratic and logarithmic are examples of the transformations applied to the predictors for the final model. The response is also transformed using the Box-Cox transformation method, due to preliminary violations of the normality assumptions.

With Box-Cox transformation parameter $\lambda = 0.101$, the estimated model is:

$$\begin{aligned}\widehat{Y}^{(0.101)} &= \frac{\widehat{Y}^{0.101} - 1}{0.101} \\ &= \hat{\beta}_0 + hat\beta_1 X_{\text{Status}} + \hat{\beta}_2 X^2_{\text{TotalExpenditure}} + \hat{\beta}_3 X_{\text{Polio}} \\ &\quad + \hat{\beta}_4 \log(X_{\text{Population}}) + \hat{\beta}_5 X_{\text{IncomeCompositionofResources}} + \hat{\beta}_6 X^2_{\text{Schooling}} + \varepsilon\end{aligned}$$

Table 8: Final Model Summary Table at 0.05 Confidence Level

Predictor	Estimate	Std. Error	t value	$Pr(>\|t\|)$	2.5%	97.5%
(Intercept)	6.3298	0.5977	10.591	$<$ 2e-16	5.1575	7.5021
Status	-1.2452	0.2340	-5.321	1.18e-7	-1.7043	-0.7862
Total Expenditure	0.0036	0.0024	1.473	0.141	-0.0012	0.0083
Polio	-0.0033	0.0032	-1.032	0.302	-0.0097	0.0030
Population	-0.0022	0.0250	-0.088	0.930	-0.0513	0.0468
Income Composition	4.1549	0.5912	7.028	3.07e-12	2.9953	5.3146
Schooling	0.0222	0.0017	12.733	$<$ 2e-16	0.0188	0.0257

Intercept:

With all predictors at 0, the model's Box-Cox–transformed GDP (BC-GDP) is 6.33. Because predictors such as schooling or health spending cannot literally be zero, the intercept is a baseline constant rather than a quantity of direct policy interest.

Status:

Status is a categorical variable with **developed** as the reference. The coefficient –1.2452 indicates that developing countries have, on average, a 1.2452 lower BC-GDP than developed ones, holding other factors constant. This reflects the expected trend: developing countries often have lower investment in health and socioeconomic resources, resulting in reduced human capital and economic output.

Total Expenditure (squared):

The coefficient for total expenditure squared is 0.0036. Since there is no linear term, the relationship between health spending and BC-GDP is monotonically increasing and nonlinear—as total expenditure rises, BC-GDP increases at an accelerating rate. This suggests that higher health investment is linked to disproportionately larger economic benefits, likely through improved public health and productivity. However, the relationship is associative, not causal. It's also possible that countries with higher GDP can afford greater health spending. Still, the positive nonlinear trend supports the view that healthcare is a valuable investment, especially at moderate-to-high spending levels.

Polio:

Polio measures immunization coverage among 1-year-olds. The coefficient (–0.0033) suggests a slight negative relationship with BC-GDP, but it is not statistically significant ($p = 0.302$). This implies that, after adjusting for other variables, polio coverage does not have a meaningful impact on GDP. While total expenditure may influence immunization rates, it does not fully explain them.

Population (log):

The coefficient for log(Population) is –0.0022 and not statistically significant ($p = 0.930$). This suggests population size does not explain additional variation in GDP after accounting for other factors. The 95% confidence interval $[\tilde{}0.0513, 0.0468]$ includes 0, confirming no significant relationship at the 5% level.

Income Composition of Resources:

For every 1-unit increase in the income composition index (ranging from 0 to 1), BC-GDP increases by an estimated 4.1549 units, holding other factors constant. This suggests that countries with more equitable income distribution and greater access to economic opportunity tend to have significantly higher GDPs. The 95% confidence interval for the coefficient of Income Composition is $[2.9953, 5.3146]$. This means we are 95% confident that, after controlling for other factors, a one-unit increase in income composition of resources is associated with an increase in Box-Cox transformed GDP between approximately 2.99 and 5.31 units. Since the interval does not contain 0, the effect is statistically significant.

Schooling (squared):

The coefficient for schooling squared is 0.0222. Since only the squared term is included (with no linear component), the model implies a monotonic and accelerating positive relationship between years of schooling and BC-GDP. As the average number of school years increases, BC-GDP increases and does so more steeply at higher levels of education. This supports one of the strongest theories in economics that education builds human capital, and higher educational attainment translates into stronger economic output.

Model Evaluation Using Performance Metrics:

Using appropriate metrics to assess model performance is essential for evaluating the strength of the relationship between the predictors and the response variable and for selecting the best-fitting model.

One common metric is the **coefficient of determination**, R^2, which measures the proportion of variance in the response variable (on the transformed scale) that is explained by the model. It ranges from 0 to 1, where values closer to 1 indicate a stronger relationship and better explanatory power. However, R^2 always increases when more variables are added, even if those variables are not meaningful. Therefore, it is not always a reliable measure for comparing models, as our dataset contains cross-country economic and health data, which naturally includes a lot of heterogeneity. There are different regions, income levels, government systems, and data reliability.

To address this, we use **Adjusted** R^2, which adjusts for the number of predictors in the model and only increases when a new predictor improves the model more than expected by chance. This makes Adjusted R^2 more appropriate for comparing models, especially when using transformed responses, such as in our final model, which uses a Box-Cox transformation and many predictor transformations. While

In our final model **Model 6** (clean_model6 in code), $R^2 = 0.4343$ which means that about 43.4% (nearly half) of the variance in the transformed GDP can be explained by the model. The Adjusted $R^2 = 0.4322$ is a decently strong value, because modeling complex human and economic behavior with only 6 predictors is simple and simplistic when much of GDP variation may be driven by factors not observed in the dataset. For example, superfluous political institutions, international trade policies, or cultural factors.

While not exceptionally high, these values reflect a meaningful relationship given the diversity of countries, economies, and the inherent noise in cross-country economic data.

In addition to these, we used **likelihood-based criteria**, the **AIC (Akaike Information Criterion)** and **BIC (Bayesian Information Criterion)**. These penalize model complexity and are especially useful for comparing multiple competing models. Both AIC and BIC aim to identify the most parsimonious model that balances fit and complexity, where lower AIC/BIC values are preferred. Since our goal was not just prediction but interpretability, selecting a model with good AIC/BIC and trading off R^2 and Adjusted R^2 helps reduce overfitting of data. Our final model had the lowest values across all candidate models with AIC = 8074.323 and BIC = 8117.587.

Together, Adjusted R^2, AIC, and BIC all support the choice of **Model 6** (clean_model6 in code) as the best-performing model.

Discussion and conclusion

Our project aims to investigate the impact of health expenditure and socioeconomic factors on GDP per capita in countries. After data cleaning, exploratory analysis, and model selection, our final model is **Model 6** (clean_model6 in code), which uses the Box-Cox transformed GDP as the response variable, and quadratic and log transformations in some of the predictors. The six predictors are Status, Total Expenditure^2, Polio, log(Population), Income Composition, and Schooling^2.

From MLR diagnostic results, we have reason to believe that health and socioeconomic variables do contribute meaningfully to explaining international differences in GDP, though some variation remains beyond what can be captured by the selected predictors.

Final Model (Box-Cox Transformed GDP, $\lambda = 0.101$):

$$\begin{aligned}\hat{Y}^{(0.101)} = 6.3298 &- 1.2452 \cdot \text{Status}_{\text{Developing}} \\ &+ 0.0036 \cdot \text{TotalExpenditure}^2 - 0.0033 \cdot \text{Polio} \\ &- 0.0022 \cdot \log(\text{Population}) + 4.1549 \cdot \text{IncomeComposition} \\ &+ 0.0222 \cdot \text{Schooling}^2\end{aligned}$$

The analysis demonstrates that among all the variables considered, education and income composition are the most influential and statistically significant determinants of GDP. A country's economic performance is mainly affected by the level of education and the fairness of income distribution. In contrast, factors such as health expenditure, population size, and immunization rates have strong limited impact on economic outcomes and are generally not statistically significant. This conclusion is supported by our final regression model, which explains about 93% of the variation in GDP (adjusted R-squared ≈ 0.93). Specifically, the coefficient of schooling^2 is positive (0.0222) and highly significant ($p < 2e-16$), indicating that investment in education has a strong nonlinear positive impact on economic performance. Similarly, income composition has a coefficient of 4.1549 ($p = 3.07e-12$), highlighting the importance of inclusive growth and equitable resource allocation. Other factors, including health expenditure, population, and immunization coverage, do not show significant effects in the presence of the main predictors.

Limitations:

Although the final model demonstrates a high adjusted R-squared value and strong predictive power, several limitations should be acknowledged. Firstly, while our model identifies statistical correlations between the predictors and GDP per capita, the cross-sectional nature of our dataset prevents us from determining whether these factors are direct causes of economic growth. Because cross-sectional data only capture the circumstances of each country at a single point in time, we can assess associations but cannot establish causal relationships.

Secondly, the reliability of our estimates may be affected by measurement errors and omitted variable bias. Some key predictors, such as health expenditure and immunization coverage, may be subject to measurement errors or inconsistencies in reporting standards across countries, resulting in potential data quality issues. Furthermore, although we included major socioeconomic and health-related variables, our model does not account for all possible factors that may influence GDP, such as political stability, technological progress, or international trade. Omitting these important variables could bias the results and lead to inaccurate estimates of the effects of the included predictors.

Lastly, our findings are based on the specific countries and time period covered by our dataset, which may limit the generalizability of our results. The results of this study may not necessarily apply to countries with very different social, economic, or political backgrounds, or other periods marked by major global changes.

Recommendations:

Our analysis suggests that improving education and making income distribution fairer are two of the most effective ways for countries to boost their economic growth. We recommend that governments continue to invest in education, making sure that access and quality are improved for everyone, not just a small group. Creating policies that support more equal opportunities and reduce large gaps between the rich and the poor can also have a positive impact on the overall economy.

While spending on health and immunization did not show a strong direct effect on GDP in our results, these areas are still important for the well-being of a country's population and may influence growth in ways our model didn't capture. For this reason, we suggest not overlooking investments in public health. Lastly, we encourage policymakers to look at the bigger picture by considering a range of social and economic factors and to support better, more reliable data collection for future studies.

Improvements:

There are several ways our analysis could be improved in future studies. First, including more variables, for example, political stability, infrastructure, technological progress, or even cultural differences, could demonstrate a fuller picture of what drives economic growth. Another way is to model with different types of statistical models, not just linear regression, to see if they capture relationships between variables better. Multivariate, or vector-on-vector, regression could be useful to model and compare the GDP of specific buckets of countries. This is useful as the UN naturally groups certain countries together by income, and the habits of each groups remain very different. Underfitting our model because of this was a problem we faced in our discussion on parameter selection and outliers. This would explain GDP more contextually and provide a tailored analysis for different groups from certain predictors. Finally, ensuring datasets from different countries to be as consistent and accurate as possible will help strengthen model realizations.

In conclusion, our analysis demonstrates that education and income composition are the two most important factors associated with a country's GDP per capita. Countries with higher education levels and more equitable income distribution tend to have better economic performance. In contrast, when education and income are taken into account, factors such as health spending, population size, and immunization rates do not have a significant direct impact on GDP. While our results highlight the key role of human capital and equitable growth in economic development, they are subject to some limitations in the data and analysis. Future research with more comprehensive data and additional variables could help provide a deeper understanding of the complex factors that influence economic growth.

Author contributions

Sharon Lam: Introduction, Final Model Inference and Results

Shencen Cai: Data Description, Discussion and Conclusion

Erin Xu: Introduction, Preliminary Results, R Code, Model Selection, Bibliography

Dora Dong: Preliminary Results, R Code, Model Selection

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 Bank. 2023. "World Bank Country and Lending Groups." https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

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