Title: Time Series Analysis and Optimization of Socio-Economic Indicators Across Global Economies

Instructor: **Dr. Dinesh Ekanayake**

Class: **MATH489 (SP25)**

Date: **04/28/25**

By: **Kinshuk Mangal**

Abstract

This study investigates key socio-economic indicators across five representative countries—India, Brazil, Nigeria, United States, and China—from 1980 to 2023. The analysis focuses on the Gini Index, GDP per capita, Unemployment Rate, and Education Expenditure as target variables. The methodology is structured into three phases: (1) exploratory data analysis (EDA) and preliminary feature selection using Pearson correlation and LASSO regression; (2) multi-country optimization of predictor variables through Bayesian Information Criterion (BIC); and (3) time series modeling using Vector Autoregression (VAR) to forecast the selected indicators. The study uses diagnostic tools to validate model assumptions, including normality, stationarity, and heteroskedasticity, and evaluates forecasting performance using MAE and RMSE. The findings provide a data-driven framework for understanding inequality and economic dynamics across varied developmental contexts.

1. Introduction

Socio-economic indicators like income inequality, national income, education investment, and unemployment offer insights into a country's development trajectory. However, modeling these variables over time is complex due to high interdependence and structural differences across countries. This study integrates optimization-based statistical techniques and multivariate time series models to identify critical predictors and forecast future trends in socio-economic development.

We select five countries representing diverse geographical and developmental contexts:

- **India** (emerging, Asia)
- **Brazil** (emerging, Latin America)
- **Nigeria** (developing, Africa)
- **United States** (developed, North America)
- **China** (emerging, Asia)

Four primary indicators are analyzed:

GDP per Capita: Captures overall economic performance and living standards. It helps explain how national income levels relate to inequality.

Unemployment Rate: Reflects access to income and job opportunities. Higher unemployment often leads to greater income disparities.

Education Expenditure (% of GDP): Indicates investment in human capital. More spending on education can reduce inequality by improving access to opportunities.

Gini Index: This is the dependent variable being predicted — a direct measure of income inequality.

2. Methodology

2.1 Data Collection and Preprocessing

Data was sourced from the World Bank Open Data Repository for the years 1980–2023.

Indicators Used:

- **Target Variables**: Gini Index, GDP per Capita, Unemployment Rate, Education Expenditure
- **Potential Predictors**: Inflation Rate, Trade Balance, FDI, Literacy Rate, Life Expectancy, Urban Population %, Poverty Rate, Population Growth

Steps Taken:

- **Imputed missing values using linear interpolation and contextual knowledge**
- **Ensured temporal alignment of time series across countries**

2.2 Feature Selection

Pearson Correlation Analysis:

The Pearson Correlation Coefficient measures the strength and direction of the linear relationship between two continuous numerical variables. It ranges from:

- **+1** → perfect positive linear correlation
- **0** → no linear correlation
- **–1** → perfect negative linear correlation

Mathematically, it's defined as:

$$r = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where:

- Cov(X,Y) is the covariance between variables XXX and YYY
- σ_x and σ_y are their standard deviations

In this study, Pearson correlation was used to:

- **Explore relationships** between key socio-economic indicators
- **Identify redundant or highly related variables** (e.g., multicollinearity)
- **Support feature selection** for modeling (e.g., which variables are worth keeping for Gini, GDP, etc.)
- **Compare relationships across countries**, each with unique development patterns

Lasso Regression:

Lasso regression was valuable with high-dimensional socio-economic data, in selecting important predictors across a wide range of countries. When studying socio-economic indicators like GDP, Gini index, education levels, unemployment rates, health expenditure, and population growth across multiple countries and time periods, the number of potential explanatory variables can be large and highly correlated. In such settings, lasso regression can be used to identify the most influential socio-economic variables affecting a particular outcome—such as income inequality, poverty rates, or economic growth—while automatically excluding less relevant features.

Mathematically, lasso solves the following optimization problem:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - X_i^\top \beta\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where X contains the socio-economic variables for all countries (e.g., X_{ij} could be the GDP of country i at time j), and y might represent the target variable such as the Gini index. By tuning the regularization parameter λ, researchers can control the sparsity of the model, allowing lasso to automatically select a parsimonious subset of variables that are most predictive of the target. This not only improves model interpretability but also provides policy-relevant insights into which socio-economic factors are most impactful across diverse national contexts.

2.3 Multi-country optimization using BIC

Dataset Construction:

To construct a country-level panel data structure, we begin by loading a multi-country socio-economic dataset from the World Bank, containing variables like Gini Index, GDP, population, and other indicators across multiple years. The dataset is filtered iteratively for each of the top 30 most populated countries, isolating their respective time series data. Within each country's data slice, missing values are linearly interpolated to fill internal gaps, and non-essential or highly incomplete variables are removed to ensure data quality. This cleaning process standardizes the structure across countries, ensuring each has a consistent set of time-indexed observations. After processing, these individual country datasets can be aligned and combined into a larger panel format—where each row corresponds to a specific country-year observation, allowing for comparative analysis across both time and countries. This structure enables robust time series and panel data modeling while preserving the unique temporal dynamics of each country.

BIC optimization

The Bayesian Information Criterion (BIC) is a model selection criterion that balances model fit and model complexity. It is defined as:

$$\text{BIC} = -2 \cdot \ln(\hat{L}) + k \cdot \ln(n)$$

Where:

- L is the likelihood of the model given the data.
- k is the number of parameters (including the intercept).
- N is the number of observations.

BIC penalizes models with more parameters to avoid overfitting. A lower BIC value indicates a better model, assuming all models are fit to the same data.

Using BIC for Best Subset Selection:

In our project, we aim to predict the 4 variables for each of the 5 countries using multiple socio-economic indicators. However, not all indicators are equally informative, and including too many may lead to overfitting. **To identify the most relevant predictors for each country, we use best subset selection optimized by BIC:**

1. For each country, we first clean and preprocess the data, ensuring that time series are aligned and missing values are interpolated.

2. We isolate the 4 variables as the dependent variable and identify all other available variables as potential predictors.

3. For each subset size k (i.e., using exactly k predictors), we evaluate all possible combinations of k variables.

4. For each combination, we fit an Ordinary Least Squares (OLS) regression model and compute its BIC.

5. We record the combination with the lowest BIC for each k, and finally select the overall best model with the lowest BIC across all k.

This process is applied separately to each of the five countries to find the best local model per country, and then can also be applied to the combined panel dataset to identify globally consistent predictors across all countries.

2.4. Time Series Analysis (VAR Model)

Introduction to VAR Modeling

What is a VAR Model?:

A Vector Autoregression (VAR) model is a statistical model used to capture the linear interdependencies among multiple time series. In a VAR model, each variable is a function of its own past values and the past values of all other variables in the system.

Mathematical Definition (VAR(p)):

Let Y_t be a vector of k endogenous variables observed at time t, then a VAR model of order p (denoted as VAR(p)) is written as:

$$Y_t = c + A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \varepsilon_t$$

Where:

- Y_t is a column vector of variables (e.g., Gini Index, GDP per Capita, etc.)
- C is a vector of constants (intercepts)
- A_i are coefficient matrices for each lag i
- E_t is a vector of white noise error terms (with zero mean and no autocorrelation)

This framework allows all variables to be treated as simultaneously interdependent — no variable is purely exogenous.

Why was VAR used for This Project?:

VAR is ideal for socio-economic modeling because:

- Variables like GDP, education, inequality, and unemployment naturally influence each other over time.
- It allows us to model dynamic feedback effects, like how education investment affects unemployment, which in turn may influence inequality.
- Forecasting performance is generally good in the short to medium term.
- Tools like Impulse Response Functions (IRFs) and Forecast Error Variance Decomposition (FEVD) allow deeper interpretation of how shocks propagate over time.

Understanding Co-evolving Variables

The concept of co-evolving variables refers to time series that are not independent but move together, each influencing the others over time. In this project's context:

- An increase in education expenditure might reduce unemployment, which could lead to higher GDP per capita and potentially reduce inequality.

- VAR captures this systemic interdependence without requiring a predefined causal order.

Stationarity Requirement

VAR models assume that all variables are stationary — meaning their:

- Mean, variance, and autocovariance do not change over time.

We test this using the Augmented Dickey-Fuller (ADF) test.

- If non-stationary, we take first differences of the variables to stabilize them before fitting the VAR.

Selecting the Optimal Lag Length

Choosing the right lag order ppp is critical. Too few lags → underfitting; too many lags → overfitting and loss of degrees of freedom.

We determine the optimal number of lags using Bayesian Information Criterion (BIC):

$$BIC(p) = \log|\Sigma_p| + \frac{k^2 p \log T}{T}$$

Where:

- Σ_p is the residual covariance matrix of the VAR(p) model
- T is the number of observations
- k is the number of variables

We fit models with increasing lags and select the one with the lowest BIC — balancing goodness of fit with model simplicity.

2.5 Country-Level VAR Modeling: Short-Term and Long-Term Forecasts

Why Two Representative Countries?:

To deeply analyze the dynamic interdependencies of our four socio-economic indicators — Gini Index, GDP per Capita, Unemployment Rate, and Education Expenditure — we focus on two diverse and representative countries: Brazil and India. These countries differ in terms of geography, development patterns, inequality, and policy direction, making them ideal for contrast.

While the full dataset includes 30 countries, and 5 are explored in detail, Brazil and India are selected for full VAR forecasting and diagnostic analysis. Patterns observed here are later compared to other countries to identify broader trends and exceptions.

To model the time-series, we first cleaned and prepared the dataset by selecting relevant variables and interpolating missing values using linear interpolation. Columns with excessive missing data (more than five nulls) were dropped. Each country's dataset was filtered individually to ensure that the VAR model was built on consistent and continuous data without gaps. After cleaning, the data was subset to include four main indicators: Gini Index, GDP per capita (GDPC), Unemployment Rate, and Education Expenditure (% of GDP).

The optimal lag length was selected automatically using the Bayesian Information Criterion (BIC), allowing the model to account for time dependencies between variables without overfitting.

To assess short-term predictive power, the last 20% of the data was reserved for validation. The VAR model forecasted the variables over this period, and predictions were compared with actual observed values. Following this, a long-term forecast was generated for seven additional years (up to 2030). A 95% confidence interval was also visualized based on the standard deviation of the short-term prediction error.

Train-Test Forecasting Strategy

We split each country's data into:

- Training set: First 80% of years (to build the VAR model)
- Test set: Last 20% of years (to validate forecast accuracy)

Impulse Response Functions (IRFs)

What is an Impulse Response Function (IRF)?

An Impulse Response Function (IRF) is a tool used in time-series econometrics, particularly in Vector Autoregression (VAR) models, to analyze how a shock or sudden change to one variable affects other variables in the system over time.

Mathematically, in a VAR model where all variables are interdependent, a one-time "impulse" (shock) to one variable propagates through the system, affecting both itself and the other variables across multiple future time periods.

Why IRFs Are Useful in Socio-Economic Analysis

IRFs allow us to:

- **Simulate policy interventions** — e.g., increasing education expenditure
- **Understand causality patterns over time**
- **Visualize ripple effects** of economic shocks or structural changes
- **Evaluate the dynamic stability** of an economy or social system

2.6 Diagnostic Testing and Model Credibility

To better understand the credibility of each model's forecasts, I conducted three statistical diagnostic tests on the residuals of the VAR models for each country. These tests helped assess whether key model assumptions were met.

Diagnostic Tests Used:

1. **Ljung-Box Test (Autocorrelation)**
    - **Purpose:** To check if residuals from the model were autocorrelated.
    - **Why It Matters:** Autocorrelation indicates that the model missed some patterns, which would make the forecast less reliable.

- **Interpretation:** A p-value > 0.05 suggests the residuals are uncorrelated — which is desirable.

2. **Jarque-Bera Test (Normality)**

   - **Purpose:** To test if the residuals followed a normal distribution**.**
   - **Why It Matters:** Normal residuals are an assumption for valid confidence intervals and better inference in VAR models.
   - **Interpretation:** A p-value > 0.05 means the residuals are close to normal — ideal for model reliability.

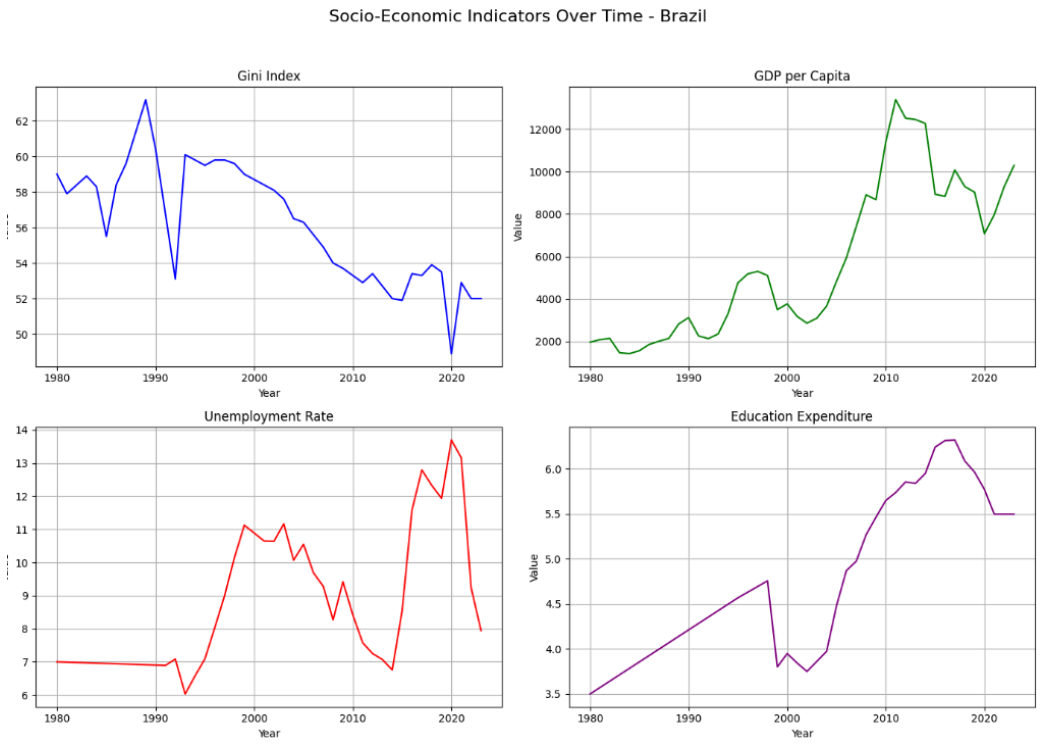
3. **Break-Variance (Heteroskedasticity) Test**

   - **Purpose:** To check if the residuals had consistent variance over time.
   - **Why It Matters:** Heteroskedasticity (changing variance) suggests instability or the presence of structural breaks in the time series.
   - **Interpretation:** A p-value > 0.05 means the variance is consistent — which supports long-term stability.

3. Analysis/Results of Visualizations

Time Series Line Plot:

Brazil

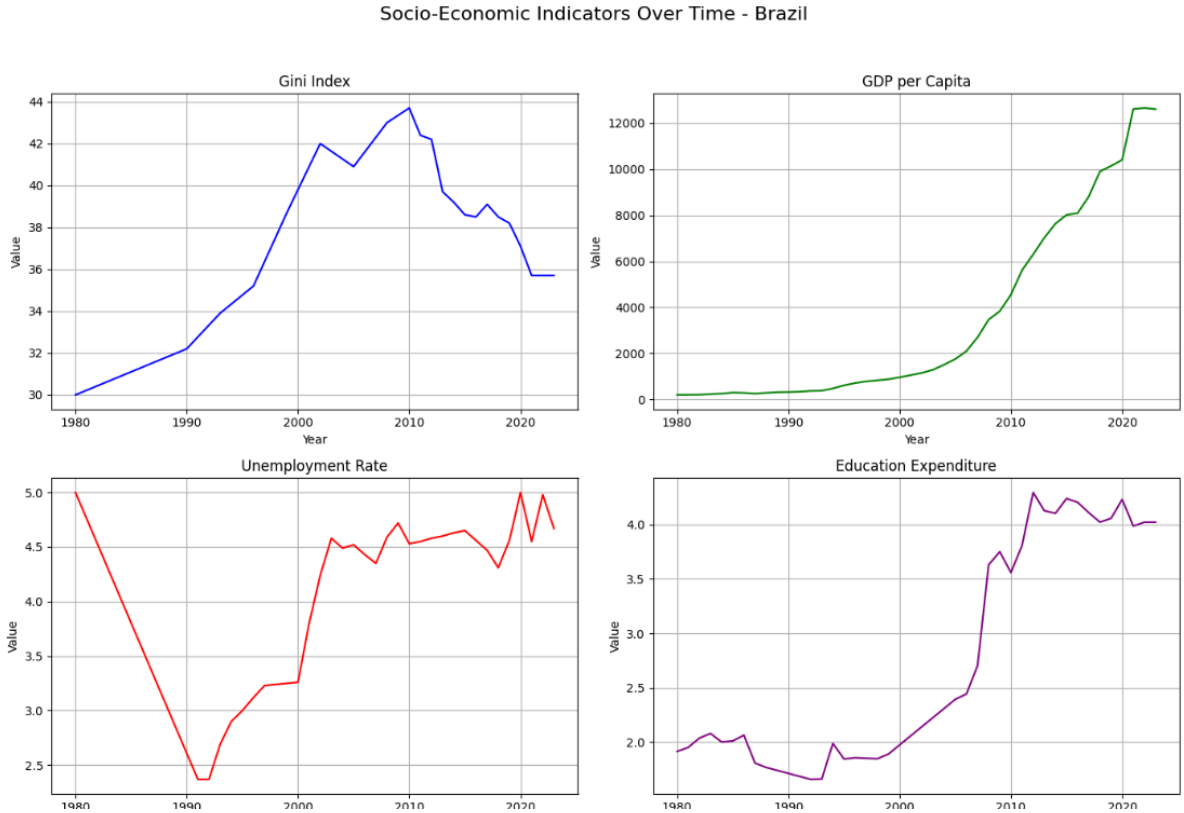


- **Gini Index**: Income inequality steadily declined from a high of ~63 in 1990 to ~52 in 2023.
- **GDP per Capita**: Experienced strong growth post-2004 with major dips in 2015 and 2020.
- **Unemployment Rate**: Rose sharply in the 2000s and again after 2015, peaking above 13%.
- **Education Expenditure**: Increased to over 6% of GDP by 2016 but declined slightly in recent years.

Insight

Brazil has made meaningful strides in reducing inequality, likely supported by social programs and increased education spending, but rising unemployment and recent cuts to education may threaten continued progress.

China



Gini Index:

- Rising sharply until around 2010, then declining, indicating early-stage widening inequality followed by recent moderation—possibly due to social policies and rural development efforts.

GDP per Capita:

- Exponential growth, especially from 2000 onward. Clear evidence of economic transformation and urbanization.

Unemployment Rate:

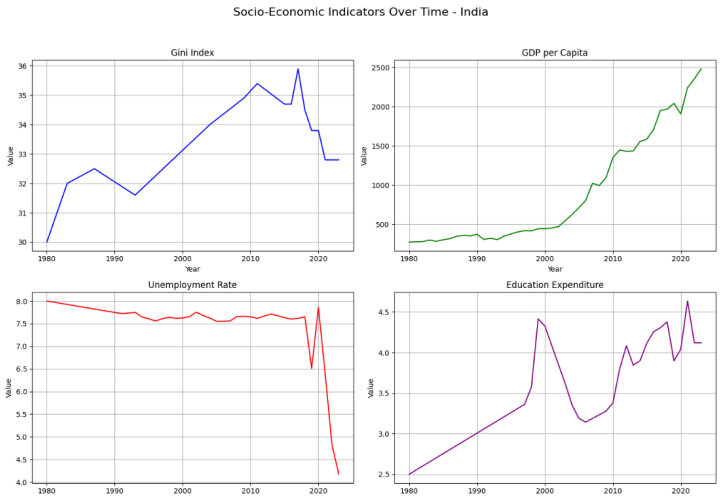
- Gradual increase over time with some stability post-2015. Indicates transitional labor market adjustments.

Education Expenditure:

- Increasing from the early 2000s, stabilizing post-2010. Investment appears strong, supporting long-term human capital growth.

Insight: China shows a classic case of rapid economic development followed by deliberate efforts to control inequality and invest in social infrastructure.

India



Gini Index:

- A noticeable increase in inequality from 1990s to 2015, followed by a recent decline.

GDP per Capita:

- Steady increase, particularly strong post-2005, reflecting liberalization and services growth.

Unemployment Rate:

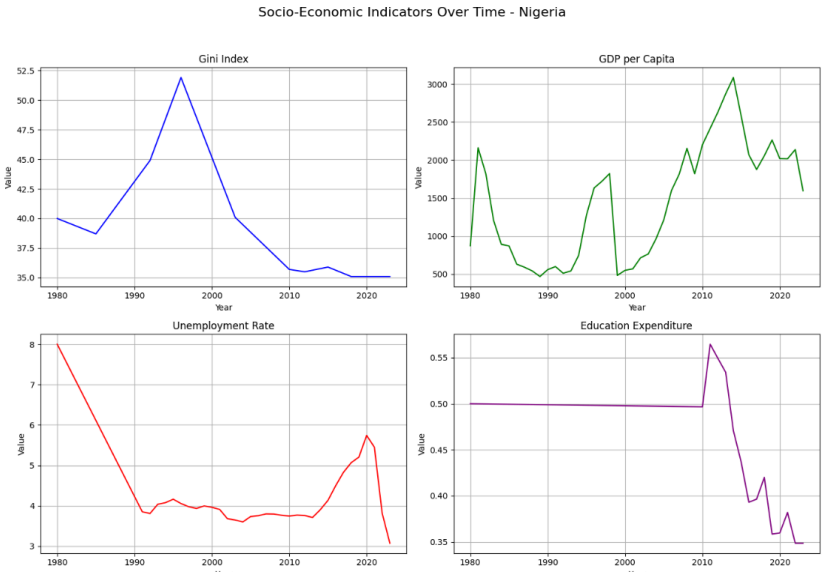
- Largely stable but a sharp dip post-2020, which is likely due to COVID-19 data distortion or misreporting.

Education Expenditure:

- Rising trend with some volatility; peaked around 2020, then dipped. Indicates growing but inconsistent prioritization of education.

Insight: India is growing economically but must stabilize education funding and address recent labor market disruptions.

Nigeria



Gini Index:

- Peaked in the mid-90s, then declined and stabilized, indicating slightly reduced inequality in recent years.

GDP per Capita:

- Highly volatile, suggesting dependency on oil and external shocks. Spikes and drops reflect economic fragility.

Unemployment Rate:

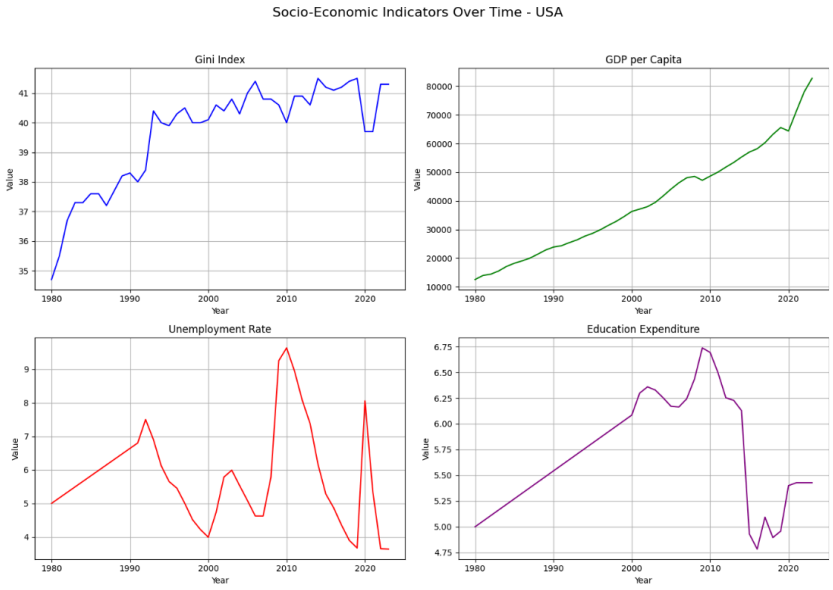
- General downward trend, with some increase post-2015. This could reflect informal sector growth or reporting inconsistencies.

Education Expenditure:

- Very low and declining—below 1% of GDP. This is a major concern for long-term development.

Insight: Nigeria faces structural volatility with underinvestment in education and an economy prone to shocks, even as inequality slightly declines.

USA



Gini Index:

- Steadily rising since 1980s, peaking around 2020. Clear sign of growing income inequality.

GDP per Capita:

- Strong, steady growth over 40 years. COVID-19 causes a brief dip, but recovery is evident.

Unemployment Rate:

- Cyclical pattern with spikes during economic crises (early 1990s, 2008, 2020), then recoveries.

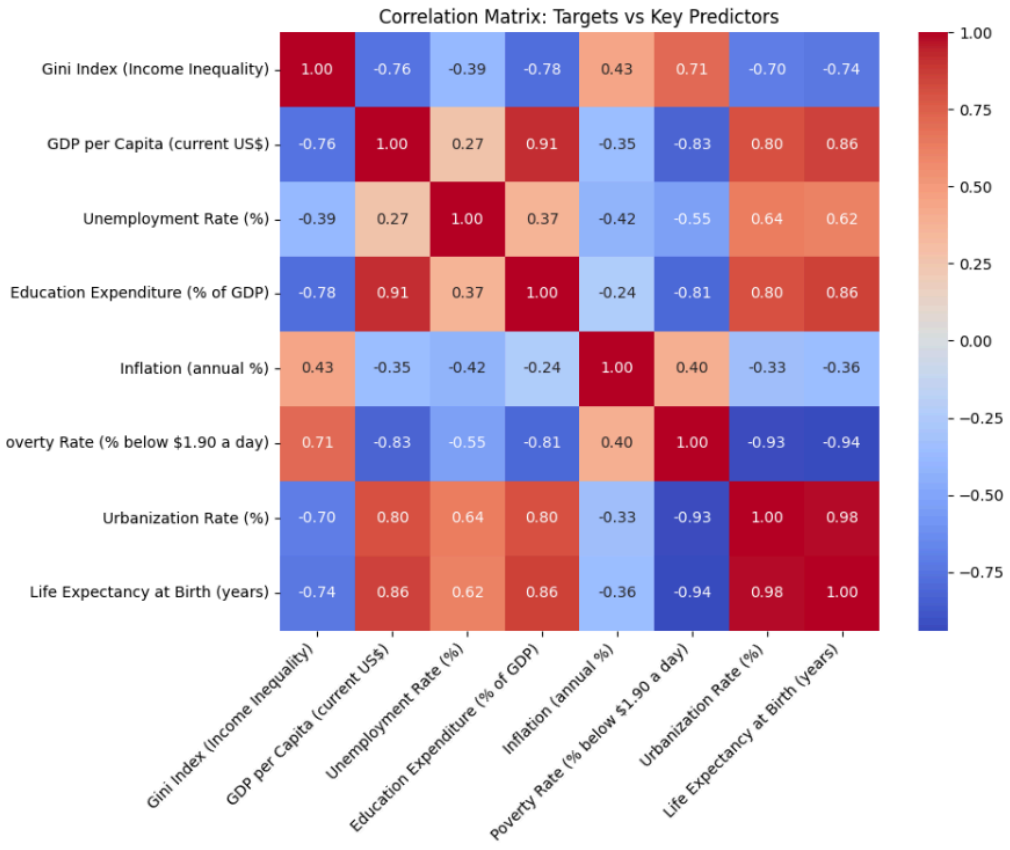
Education Expenditure:

- Peaked in 2010, but declined significantly post-2015, potentially due to budget reallocation.

Insight: Despite strong economic performance, the U.S. faces persistent inequality and recent cutbacks in education funding.

Correlation Heatmaps:

Brazil



Observations:

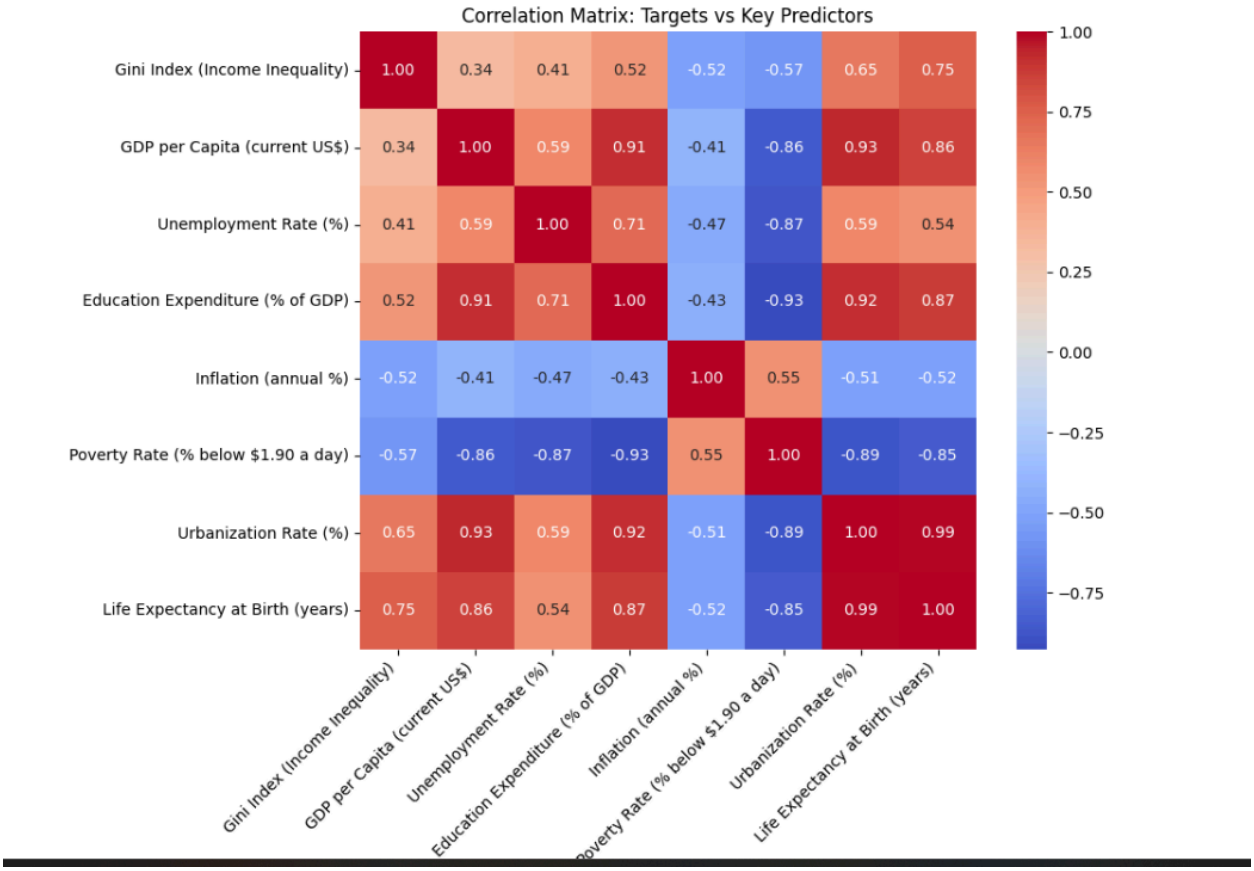
- **Gini Index negatively correlates with GDP per Capita (-0.76) and Education Expenditure (-0.78): economic growth and social investment reduce inequality.**
- **Poverty Rate strongly increases inequality (+0.71), as expected.**

- **GDP per Capita shows very strong positive correlation with Education Expenditure (+0.91) and Life Expectancy (+0.86).**

Insight:

Brazil's inequality is closely linked to poverty and education. As education spending rises, inequality tends to fall, affirming the role of inclusive social investment.

China



Observations:

- **Gini Index is positively correlated with GDP, unemployment, and education (moderately), suggesting that rapid growth may coexist with rising inequality.**

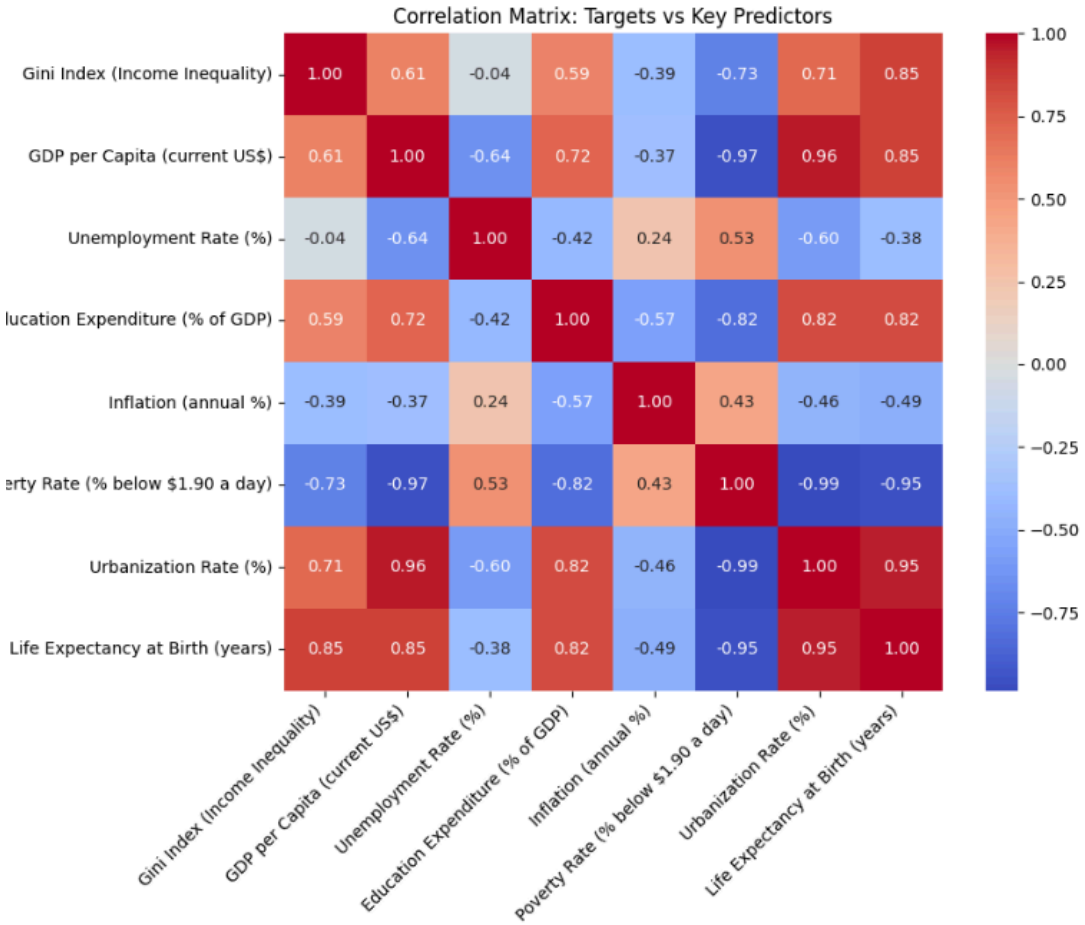
- **Poverty Rate has strong negative correlations with GDP (−0.86), education (−0.93), and life expectancy (−0.85), highlighting the transformative effect of development.**

- **Urbanization and life expectancy have extremely high correlations with GDP (>0.9).**

Insight:

In China, inequality is not as tightly linked to poverty reduction. While economic growth has lowered poverty, the benefits may be unevenly distributed.

India



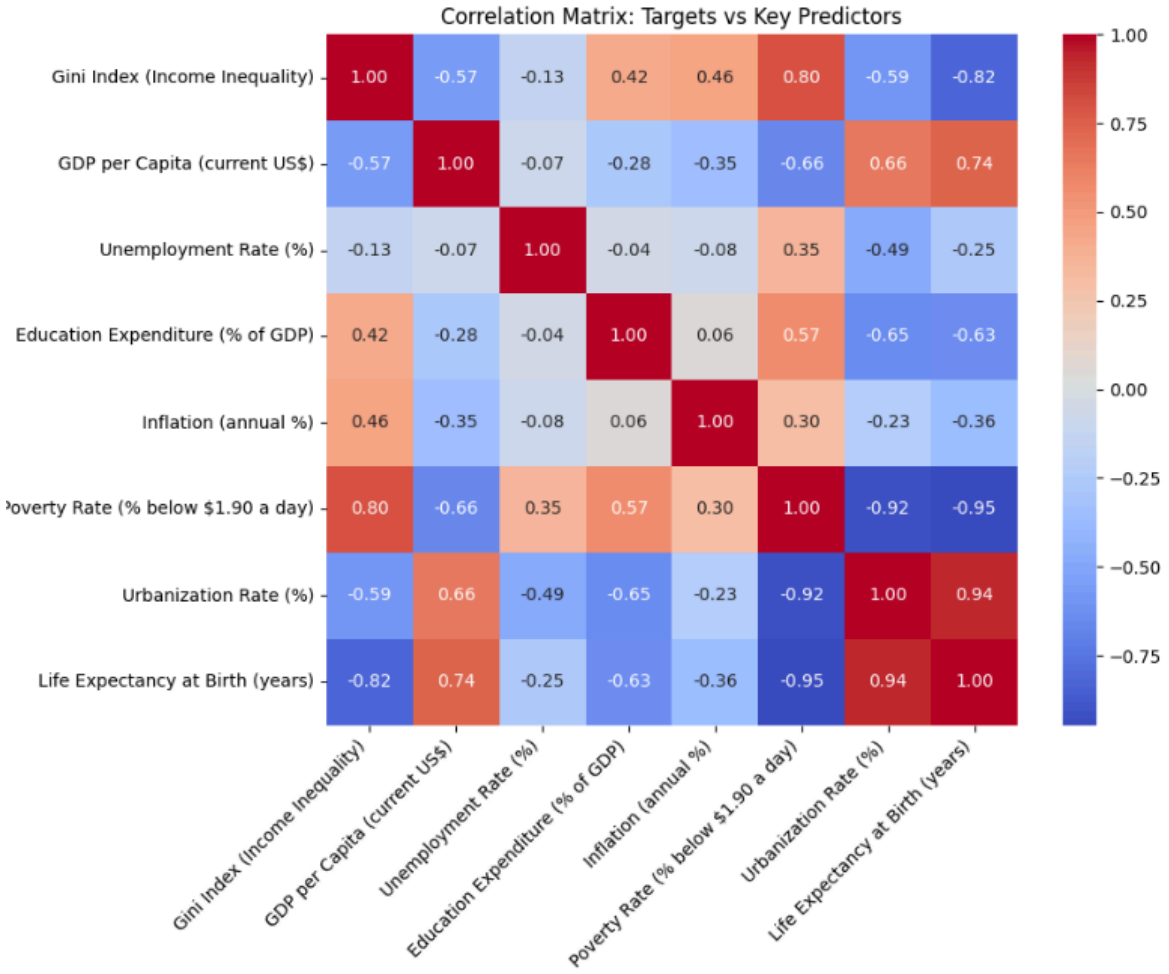
Observations:

- **Gini Index has moderate positive correlations with GDP (+0.61) and education (+0.59), hinting that growth and investment may not yet be equitably distributed.**
- **Poverty Rate strongly correlates negatively with GDP (−0.97) and urbanization (−0.99), showing development's impact on poverty reduction.**
- **Life Expectancy correlates highly with almost everything (especially GDP and urbanization).**

Insight:

India shows strong structural relationships: growth reduces poverty and raises life expectancy, but inequality persists — likely driven by regional or class disparities.

Nigeria



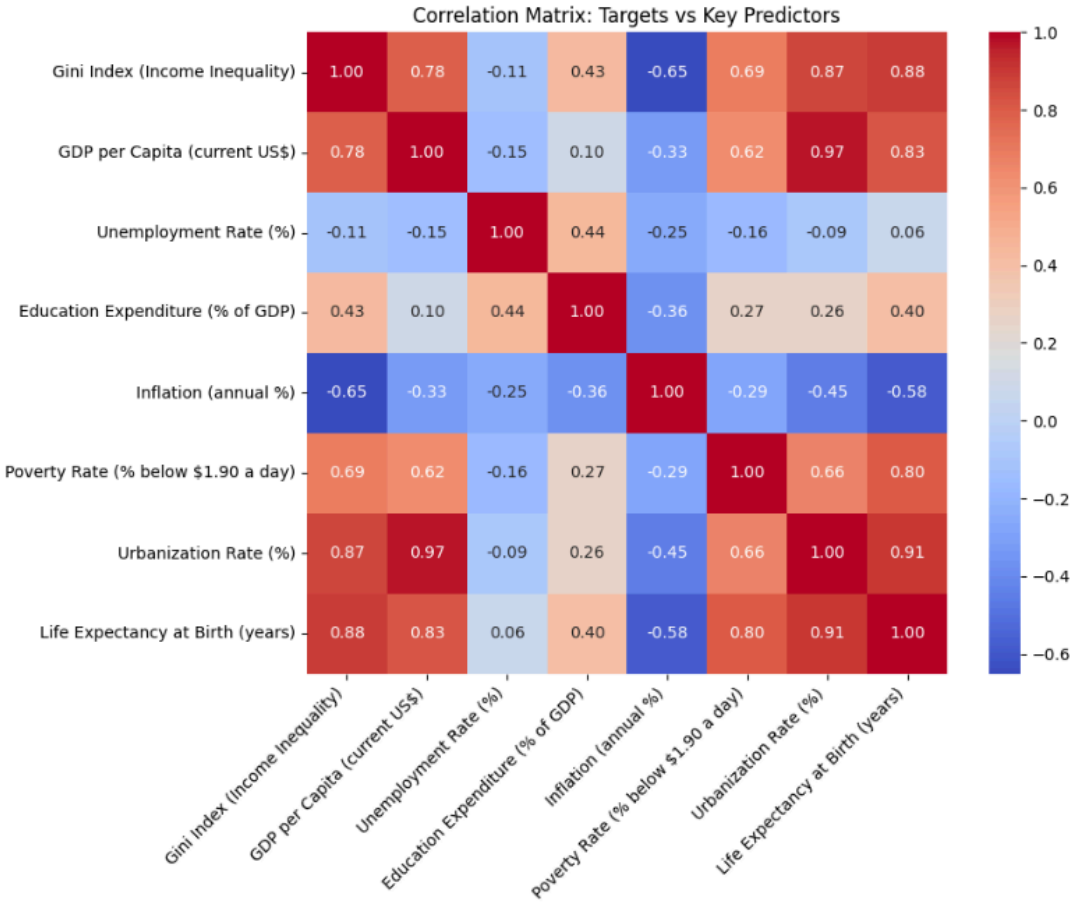
Observations:

- **Gini Index shows weak correlations overall, even with GDP (−0.57) and education (+0.42), indicating complex or unstable drivers of inequality.**
- **Poverty Rate is highly correlated with urbanization (−0.92) and life expectancy (−0.95).**
- **Education and GDP are modestly linked (+0.28), suggesting under-leveraged human capital.**

Insight:

Nigeria's socio-economic indicators suggest weak integration: growth and education are not yet translating clearly to reduced inequality or stable employment.

USA



Observations:

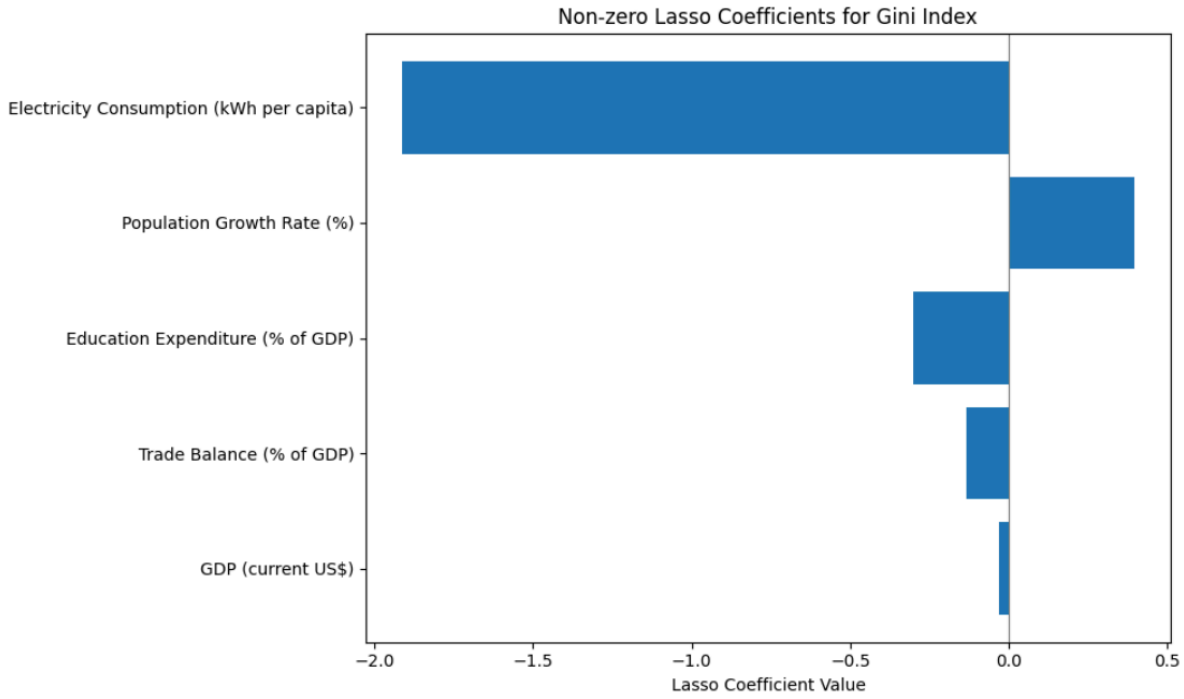
- **Gini Index is positively correlated with GDP (+0.78), education (+0.43), and poverty (+0.69), showing that inequality persists despite wealth.**
- **Urbanization and GDP per Capita have an extremely high correlation (+0.97).**
- **Unemployment has weak correlations with all other variables, showing limited short-term linkage.**

Insight:

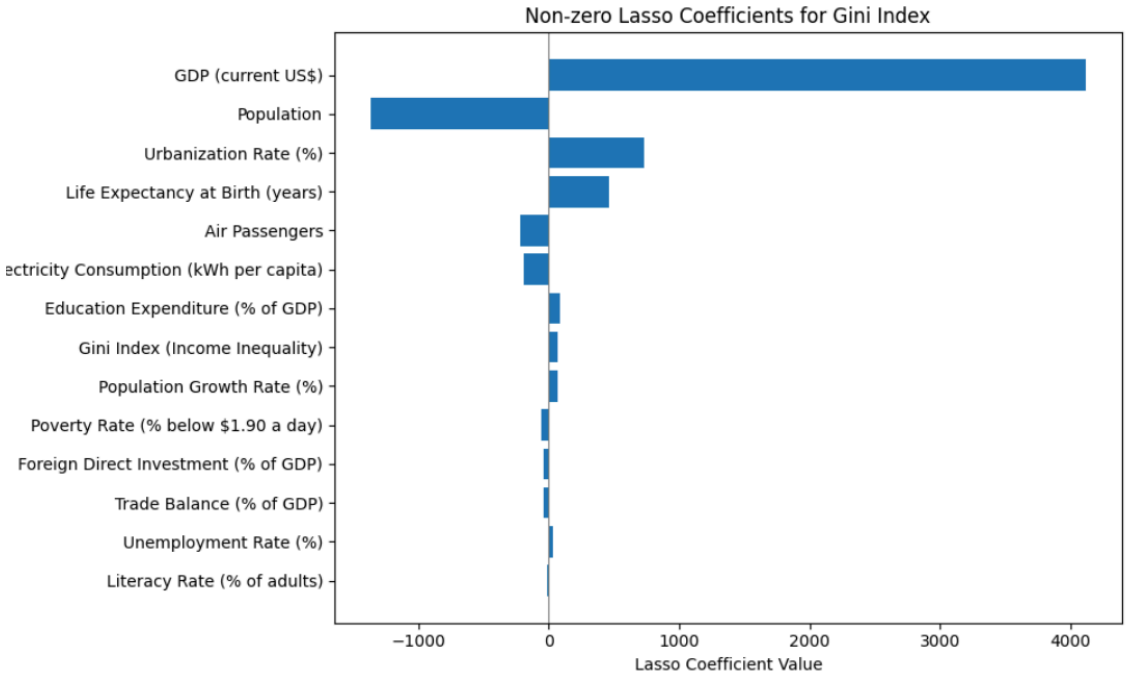
In the U.S., inequality rises alongside prosperity, suggesting structural income disparities. Growth and urbanization benefit the economy broadly, but not equitably.

Brazil:

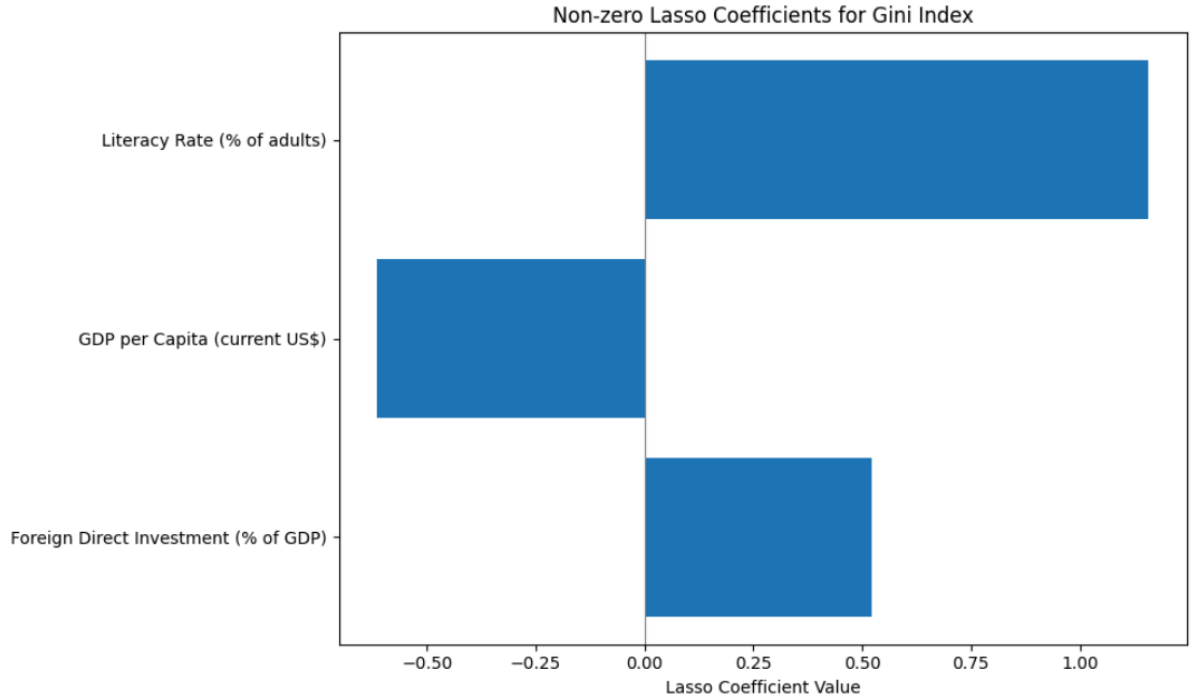
Gini Index:



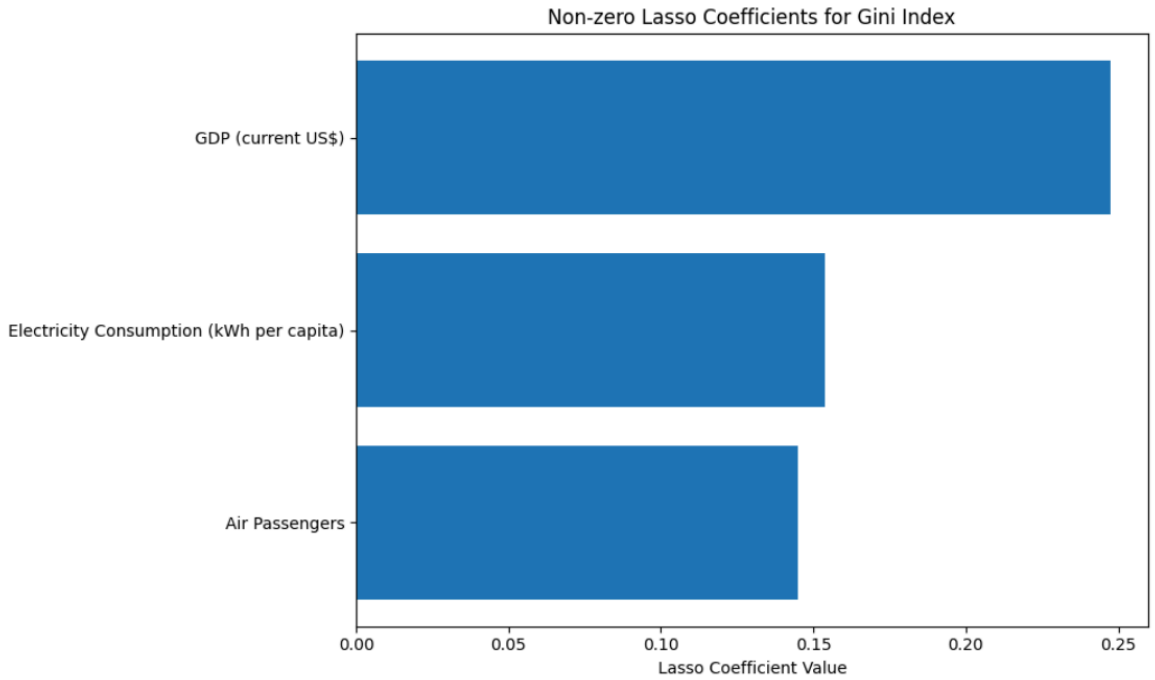
GDP per capita:



Unemployment rate:



Education Expenditure:

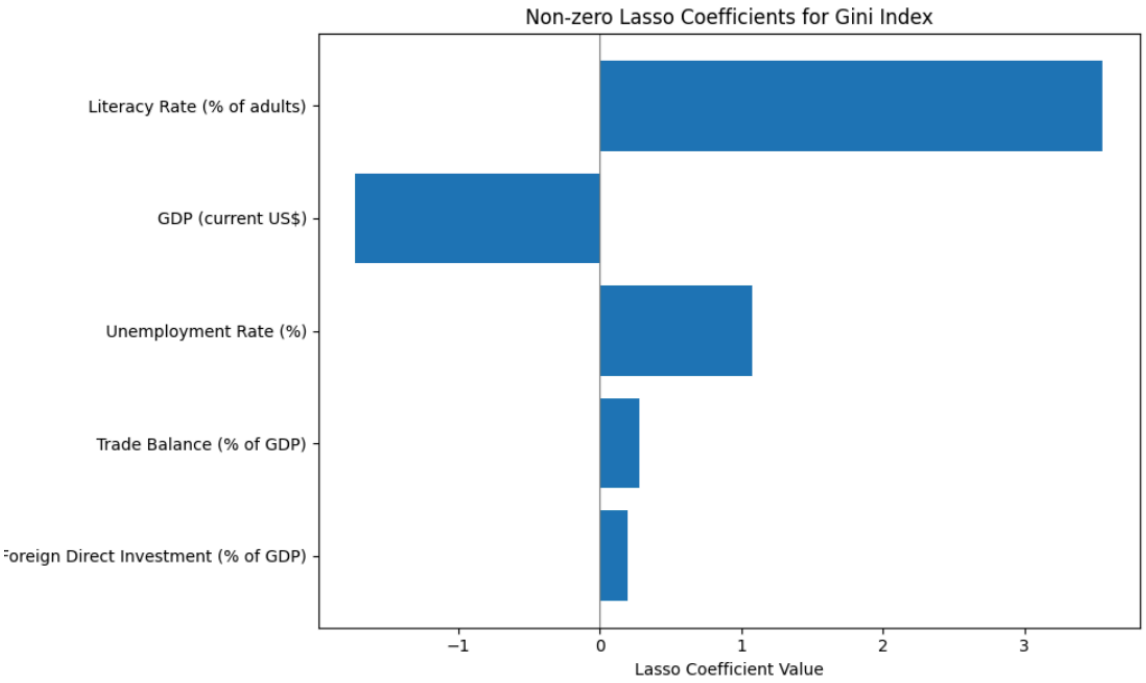


Insight: Lasso regression applied to Brazil's socio-economic data highlights key drivers for each outcome by selecting only the most relevant variables. For the Gini Index, factors like education expenditure and electricity consumption show strong negative associations, suggesting that investment in education and infrastructure can reduce income inequality. In the model for GDP per capita, lasso identifies a broad set of influential features—including GDP, urbanization, life expectancy, and poverty rate—indicating its multifaceted nature. The unemployment rate is negatively associated with GDP per capita, while a surprising positive link with literacy rate may reflect job-skill mismatches.

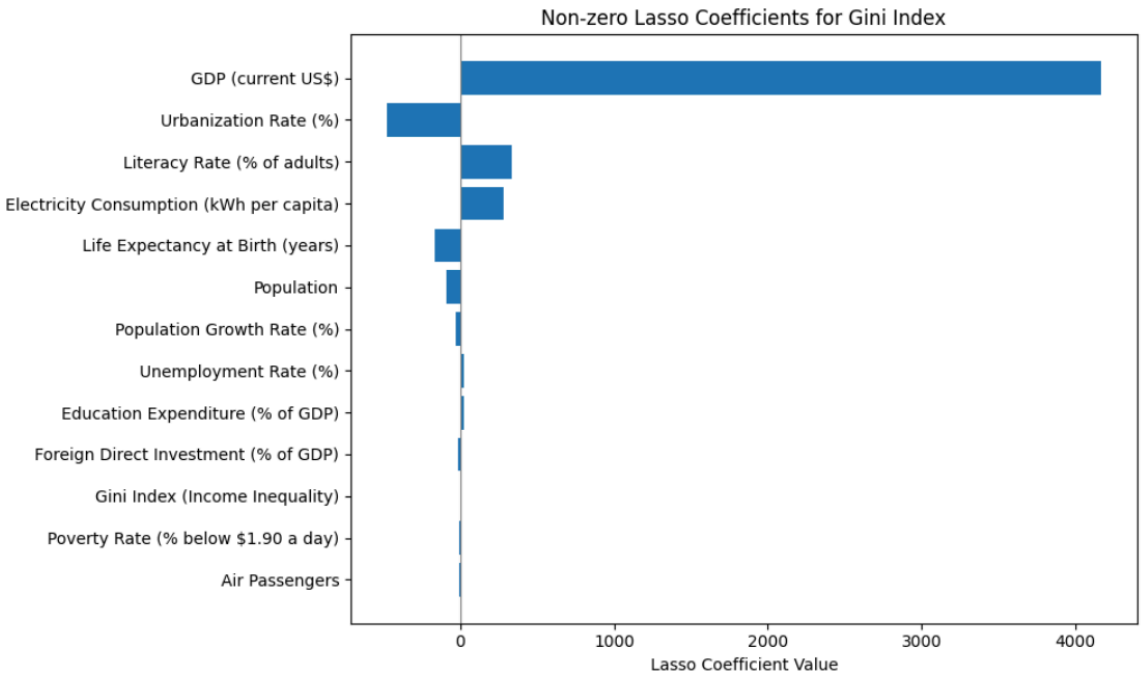
Finally, education expenditure is positively influenced by indicators of development such as GDP, electricity use, and air travel, reinforcing the role of national capacity in education investment. Overall, lasso effectively narrows down complex relationships, offering interpretable insights into Brazil's development dynamics.

China:

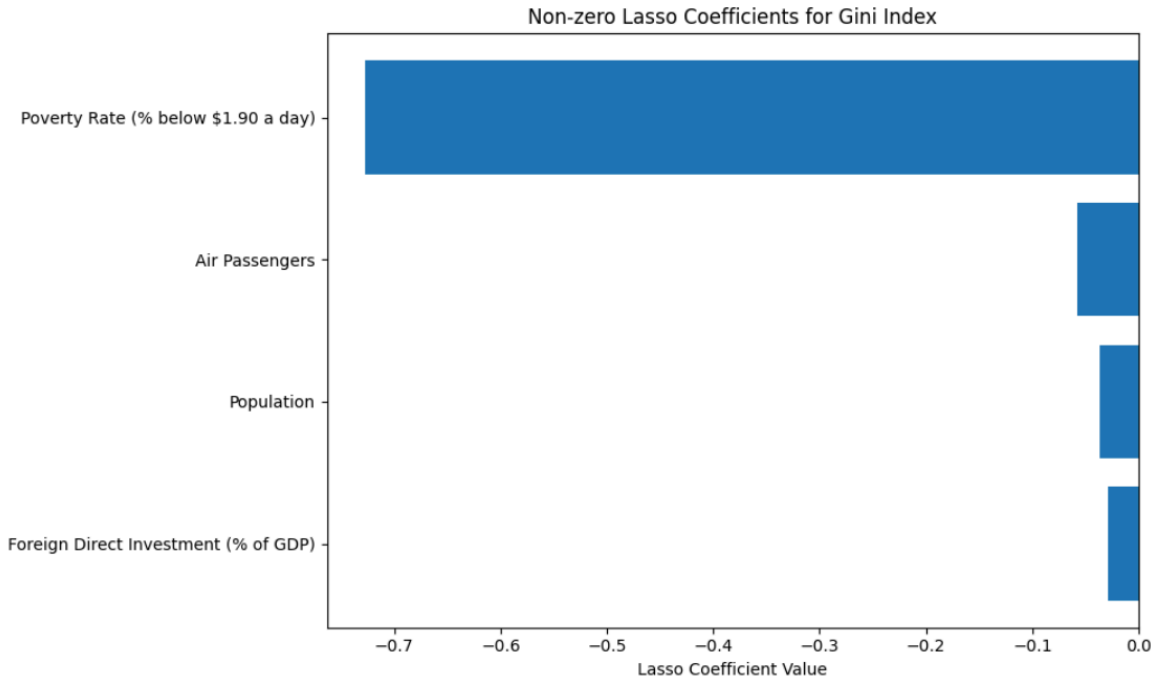
Gini Index:



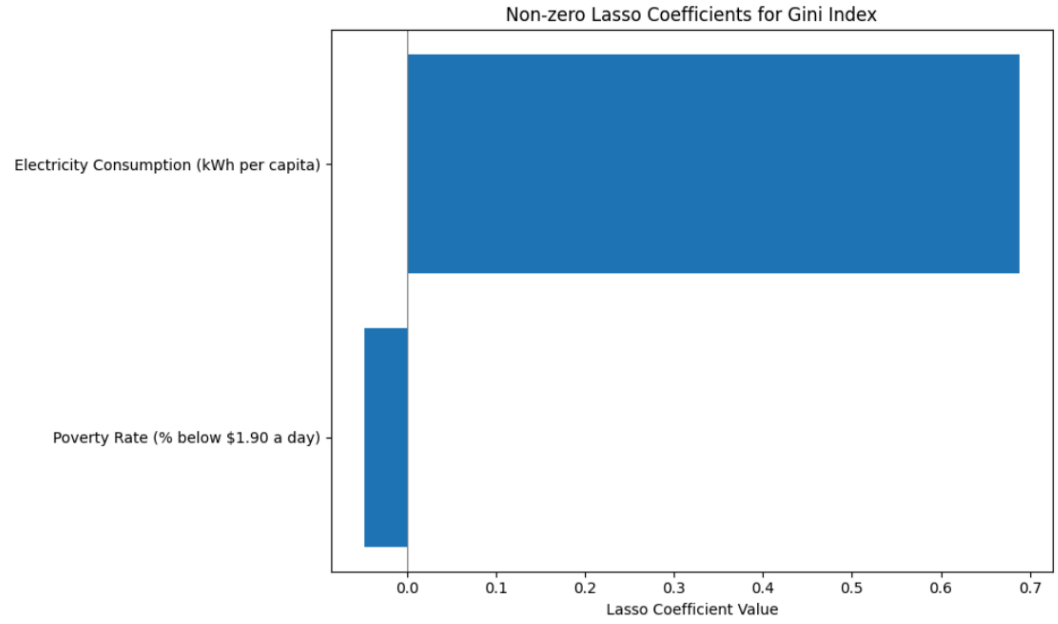
GDP per capita:



Unemployment:



Education Expenditure:

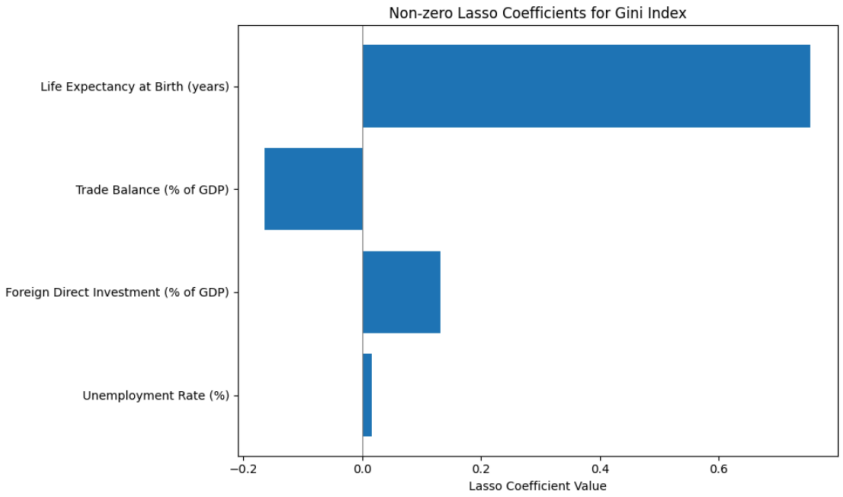


Insight: Lasso regression results for China reveal several insightful relationships among socio-economic indicators. For the Gini Index, higher literacy rates, foreign direct investment, trade balance, and unemployment are positively associated with greater inequality, while GDP has a strong negative effect, suggesting that economic growth may help reduce income disparities. In the model for GDP per capita, key positive contributors include GDP (current US$), electricity consumption, education expenditure, and

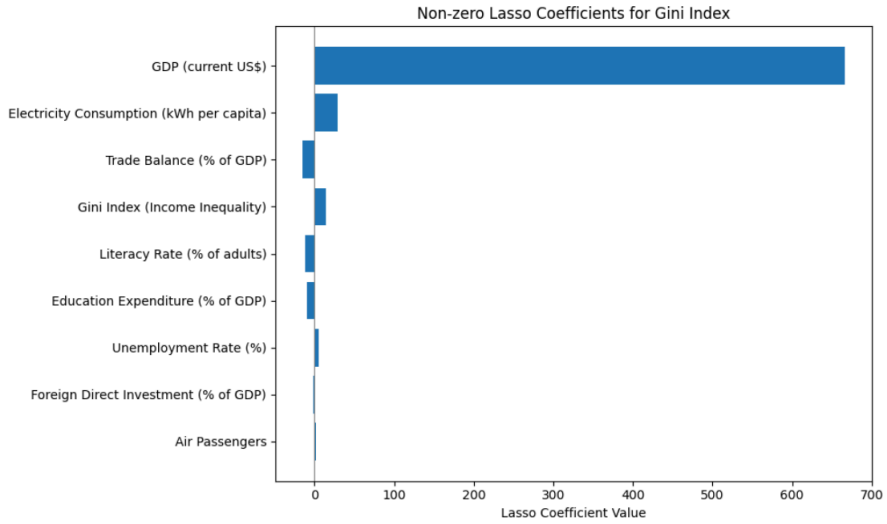
literacy rate, reflecting the role of development, infrastructure, and human capital. However, urbanization, population growth, and life expectancy surprisingly show strong negative coefficients—possibly indicating structural or demographic inefficiencies. Unemployment is negatively influenced by poverty rate, implying that reductions in extreme poverty may slightly reduce unemployment. Lastly, education expenditure increases with electricity consumption, while slightly decreasing with poverty, suggesting that improvements in infrastructure promote investment in education, while higher poverty may limit it. Overall, lasso highlights both expected and nuanced relationships, offering a focused view of how various socio-economic factors shape China's development outcomes.

India:

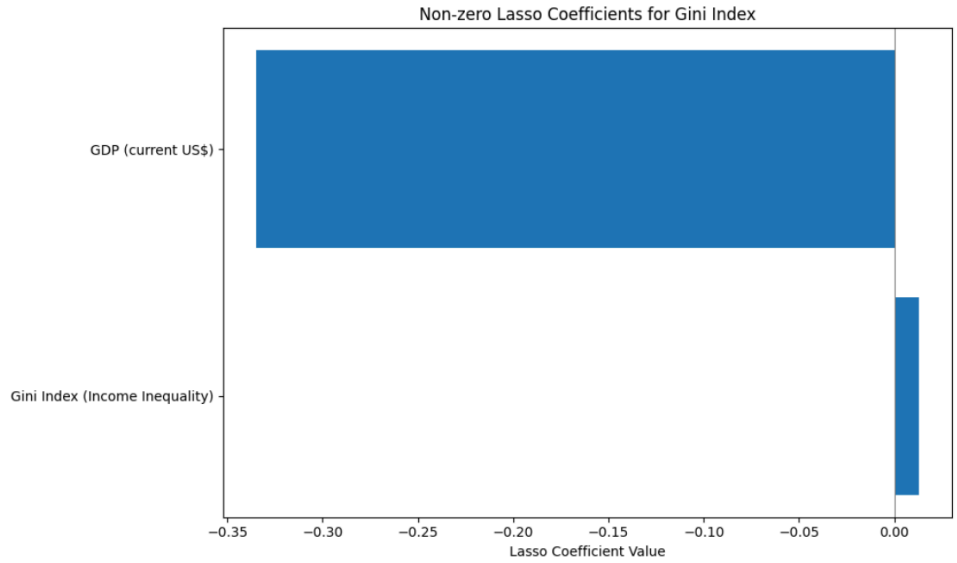
Gini Index:



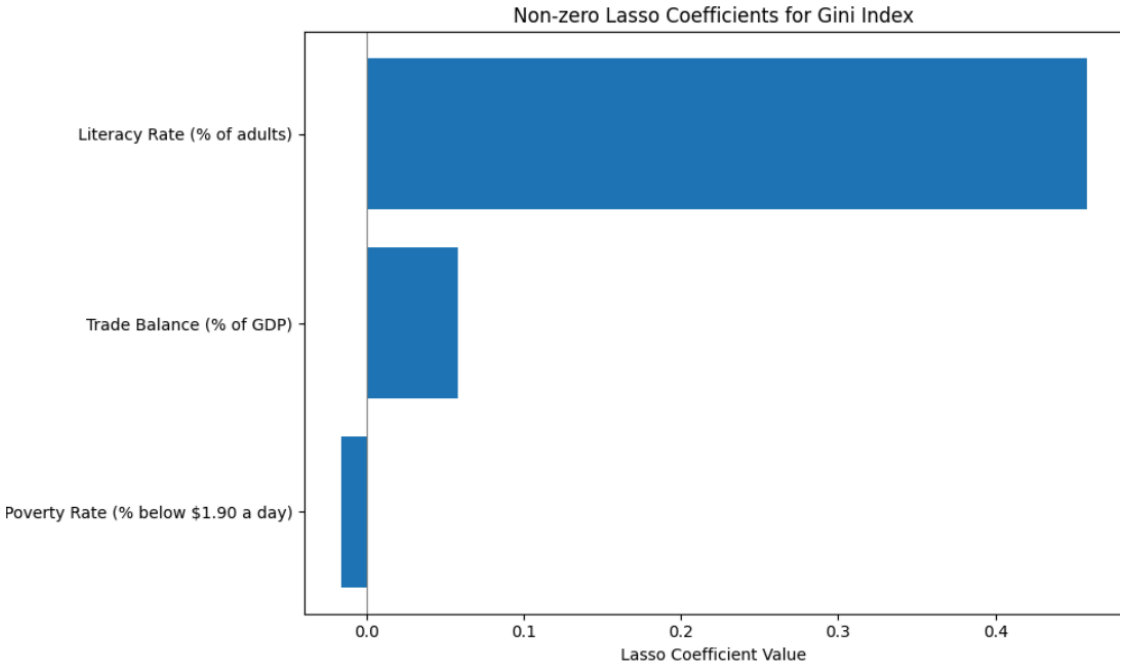
GDP per capita:



Unemployment:



Education expenditure:

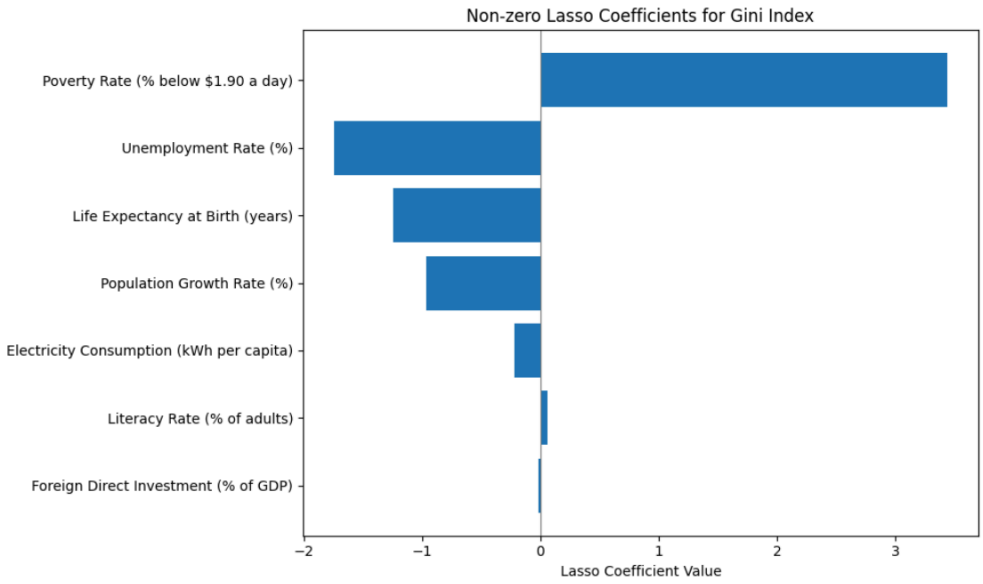


Insight: Lasso regression results for India highlight meaningful yet complex socio-economic dynamics. For the Gini Index, positive coefficients on FDI, life expectancy, and unemployment suggest that improvements in economic and health indicators don't necessarily reduce inequality—perhaps due to uneven distribution of gains. Interestingly, a negative coefficient on trade balance implies that trade

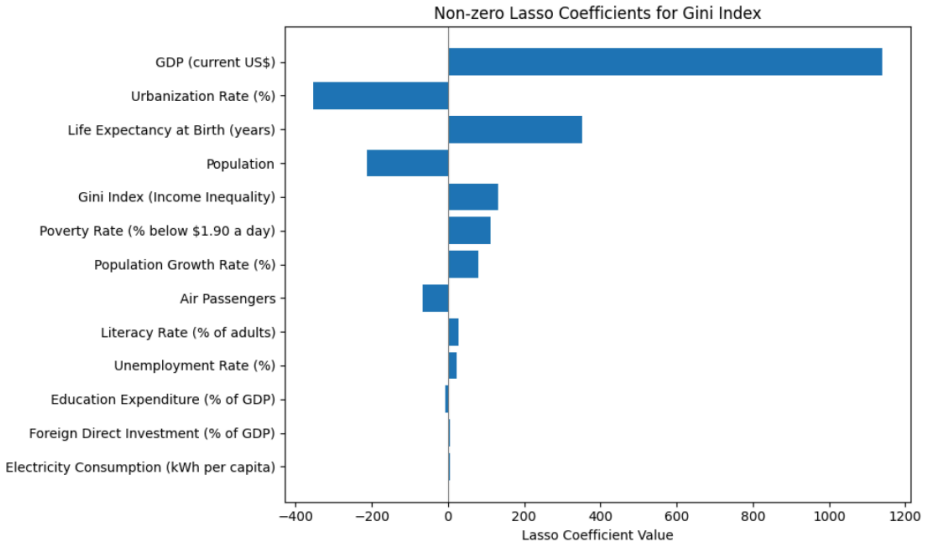
deficits might slightly help reduce inequality, possibly through increased import access benefiting lower-income groups. For GDP per capita, strong positive influence from GDP itself, air travel, and electricity consumption reflects standard development drivers, while a surprisingly high positive coefficient for the Gini Index indicates that rising income levels may be accompanied by widening inequality. Negative relationships with education spending, literacy, and trade balance raise questions about the efficiency or targeting of social investments. Unemployment decreases slightly with higher GDP, reflecting expected economic trends, while education expenditure increases with literacy, suggesting that as educational attainment rises, so does investment—potentially due to policy response or societal demand. Overall, lasso reveals that while India's growth is powered by infrastructure and economic scale, inclusive development remains a key challenge.

Nigeria:

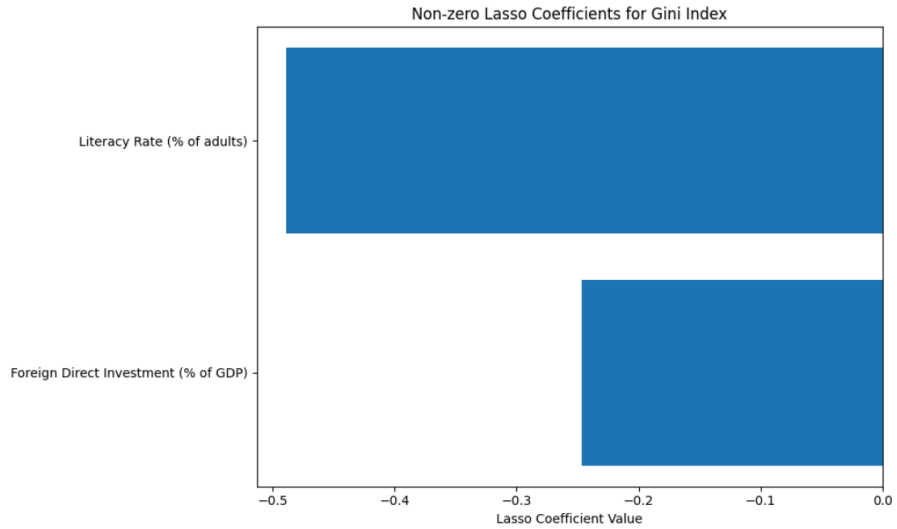
Gini Index:



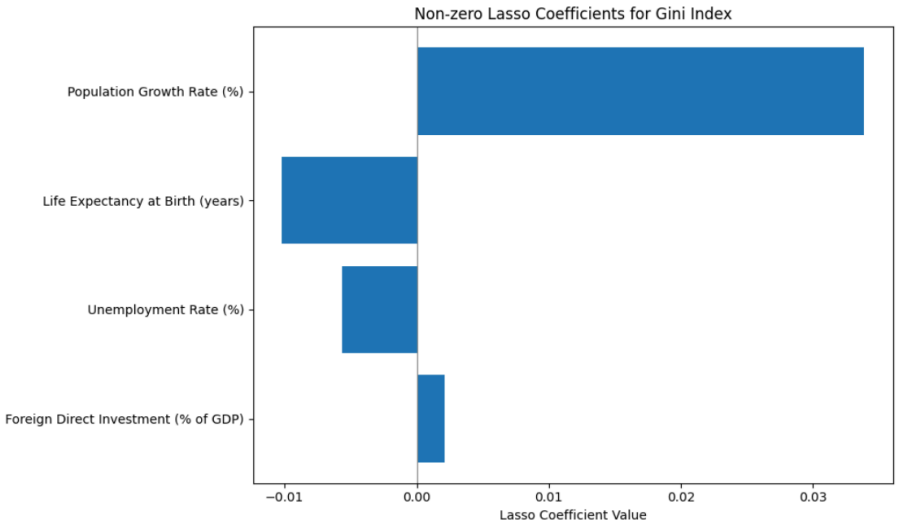
GDP per Capita:



Unemployment:



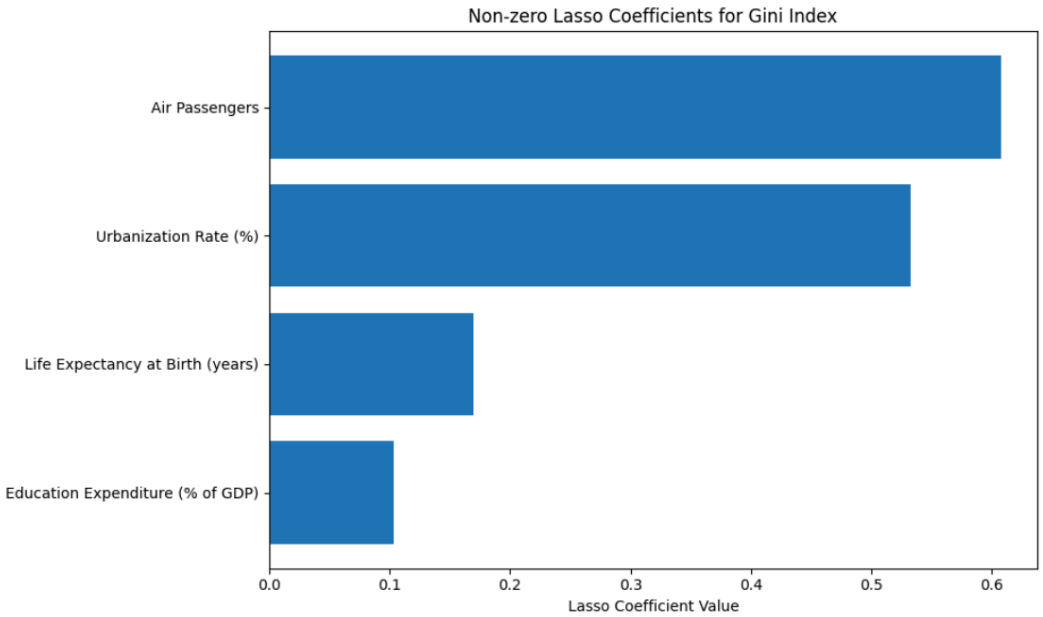
Education Expenditure:



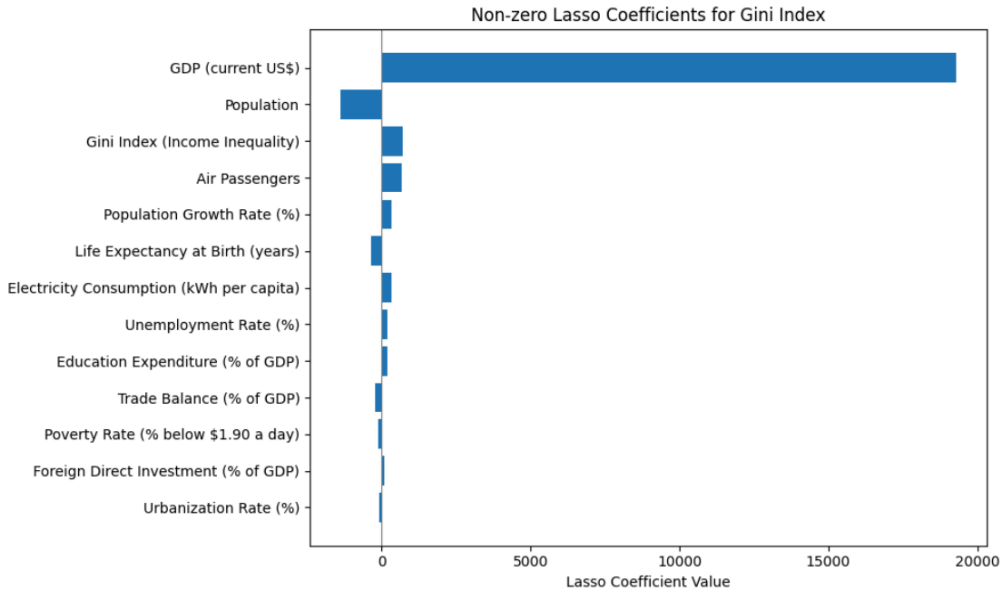
Insight: Lasso regression for Nigeria reveals that inequality (Gini Index) decreases with improvements in electricity use, life expectancy, population growth, and unemployment, while poverty strongly increases inequality—highlighting poverty as a central challenge. For GDP per capita, strong positive effects from GDP, life expectancy, poverty, and inequality suggest that economic growth and longevity coexist with deep disparities. Negative coefficients for education spending, urbanization, and population imply that growth hasn't translated into equitable development. Unemployment is slightly reduced by higher FDI and literacy, signaling modest benefits from investment and education. Finally, education expenditure shows weak and scattered relationships, reflecting potential disconnects between spending and broader development outcomes. Overall, lasso points to Nigeria’s structural imbalance—growth without shared prosperity.

USA:

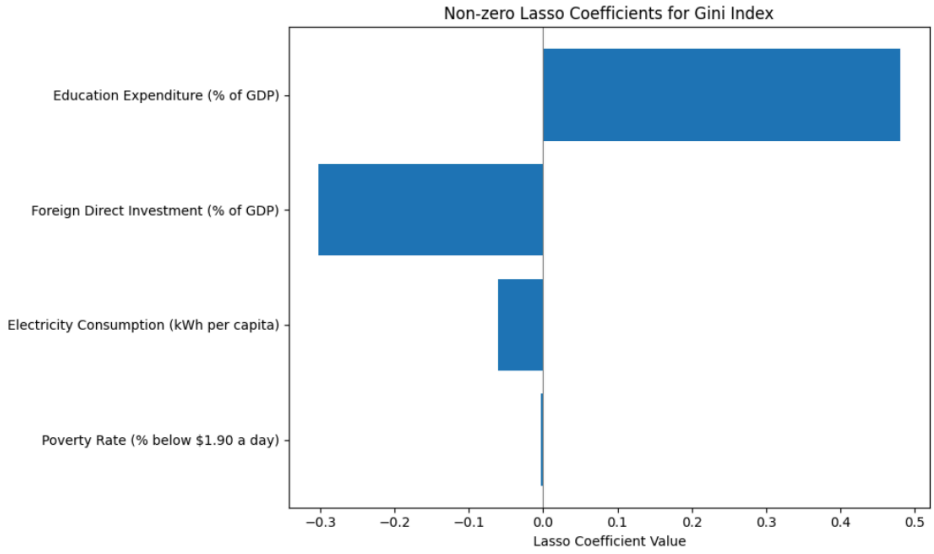
Gini Index:



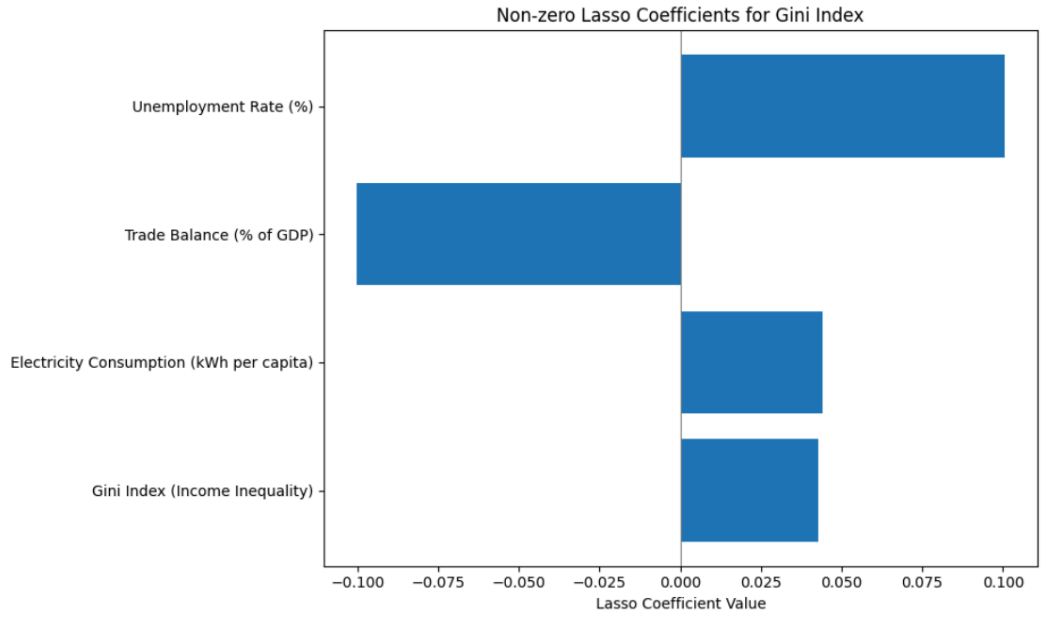
GDP per capita:



Unemployment:



Education expenditure:

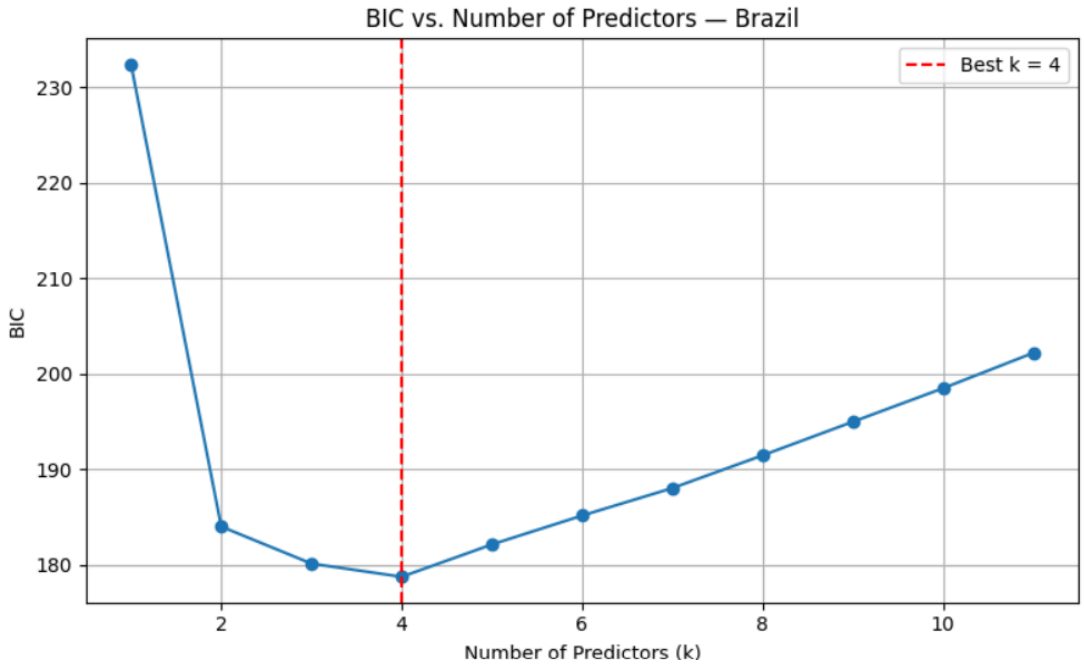


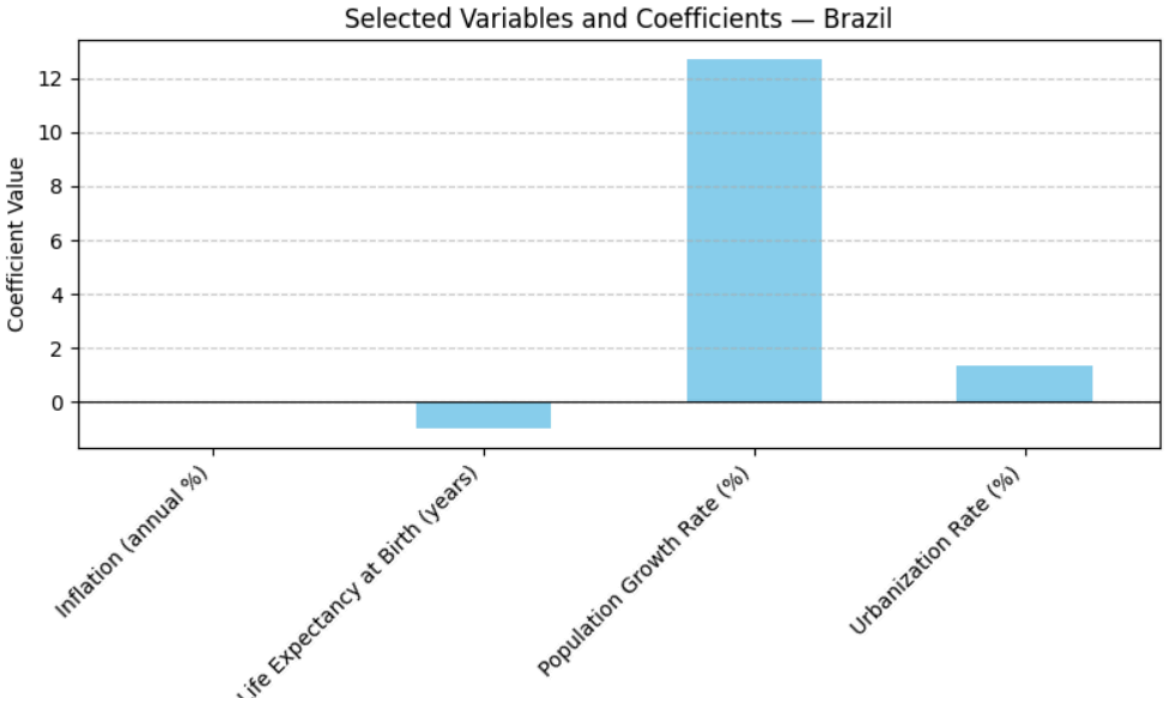
Insight: Lasso regression for the USA shows that rising inequality is positively linked to development indicators like urbanization, air travel, and education spending, suggesting that economic growth isn't evenly distributed. For GDP per capita, strong contributions from GDP, electricity use, FDI, and inequality highlight robust economic drivers, while negative effects from population, life expectancy, and trade balance point to underlying structural pressures. Unemployment is slightly reduced by FDI and electricity

use, but increases with education spending, indicating possible inefficiencies in job-market alignment. Education expenditure itself is modestly influenced by inequality and unemployment, reflecting a reactive rather than strategic investment pattern. Overall, the model reveals a high-growth economy grappling with persistent inequality and uneven policy impacts.

Brazil:

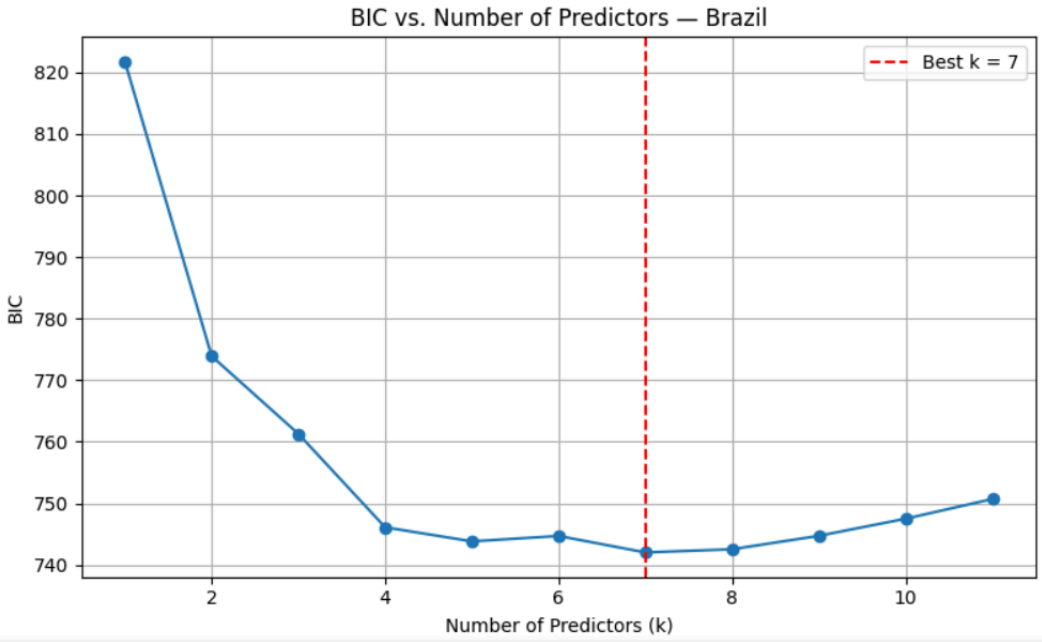
Gini Index:

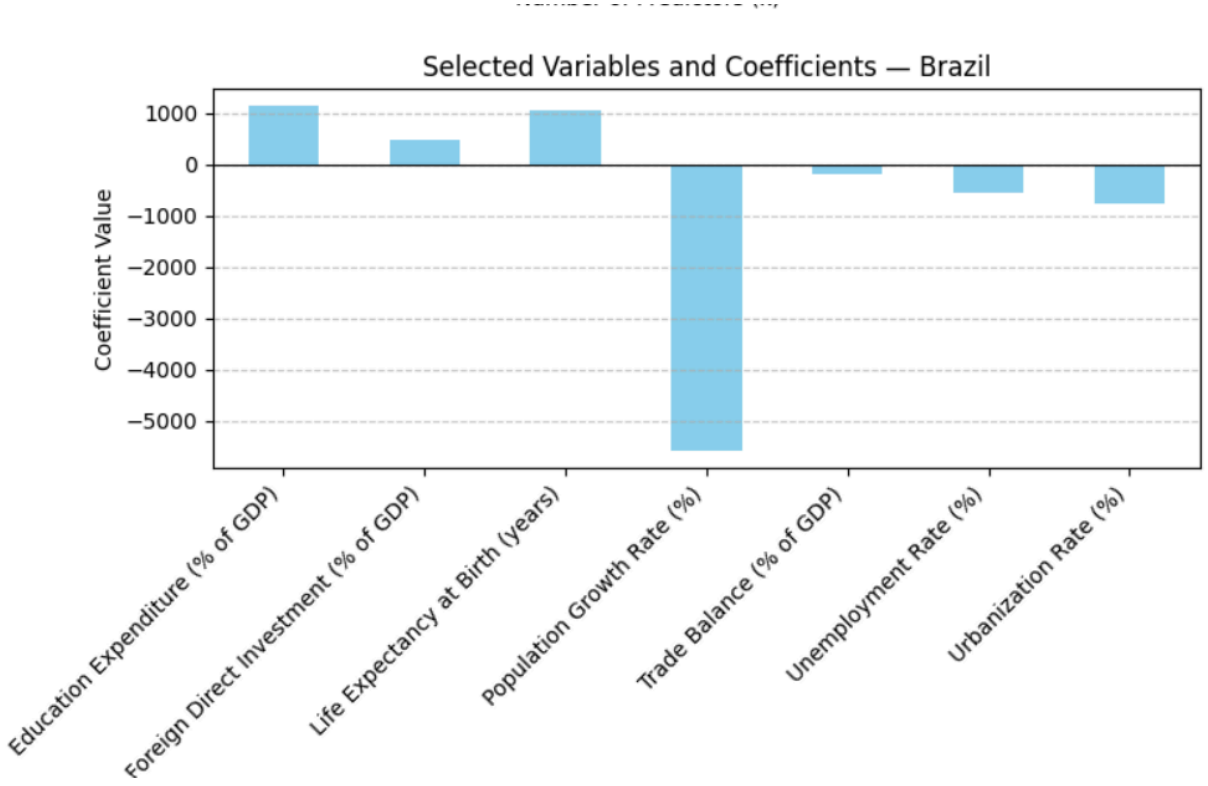




Adjusted R-squared: 0.9996

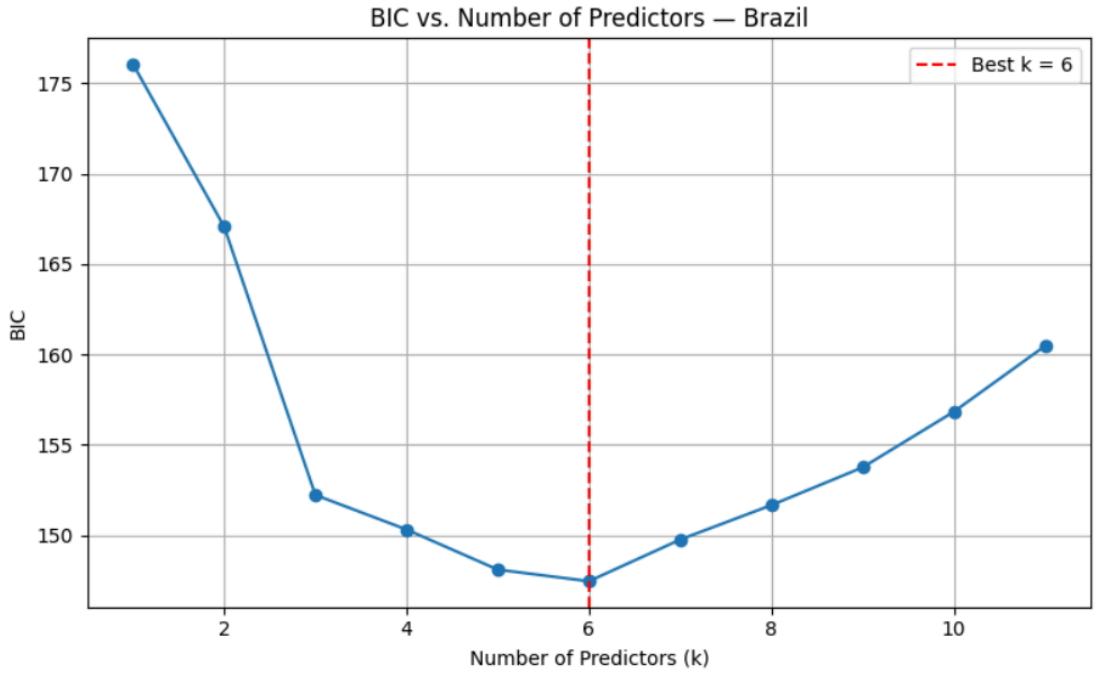
GDP per Capita:

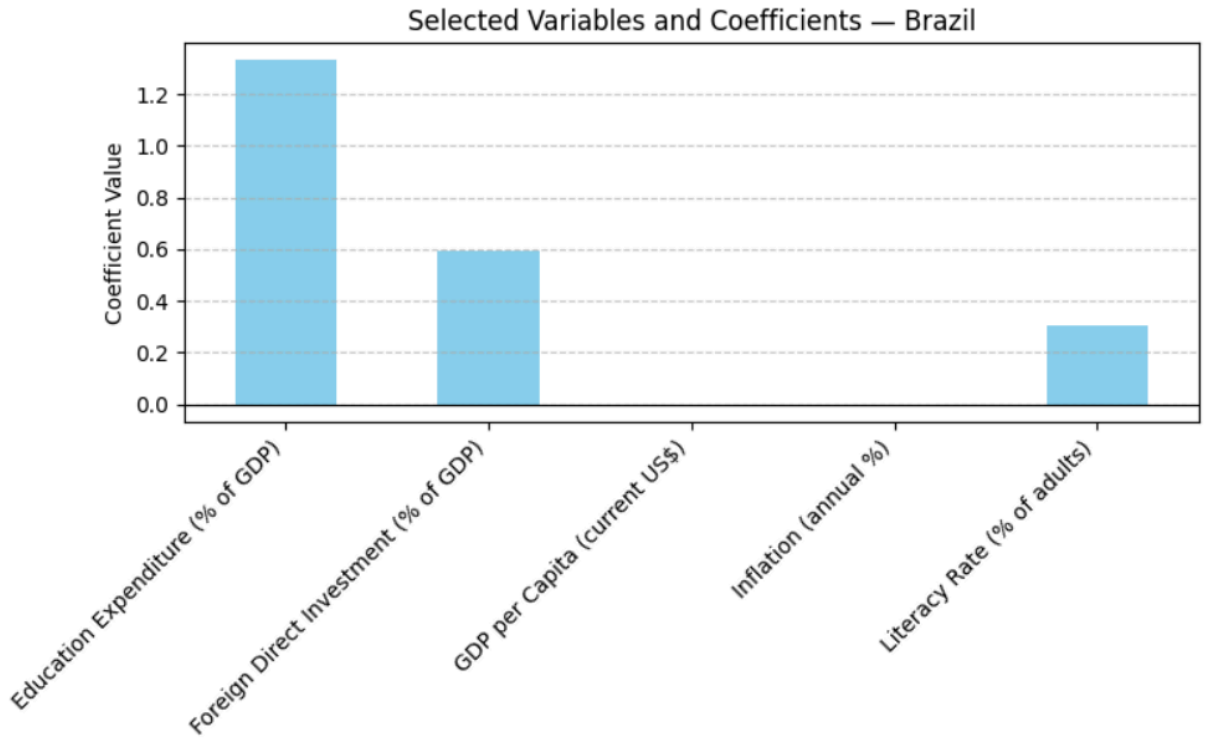




Adjusted R-squared: 0.98248

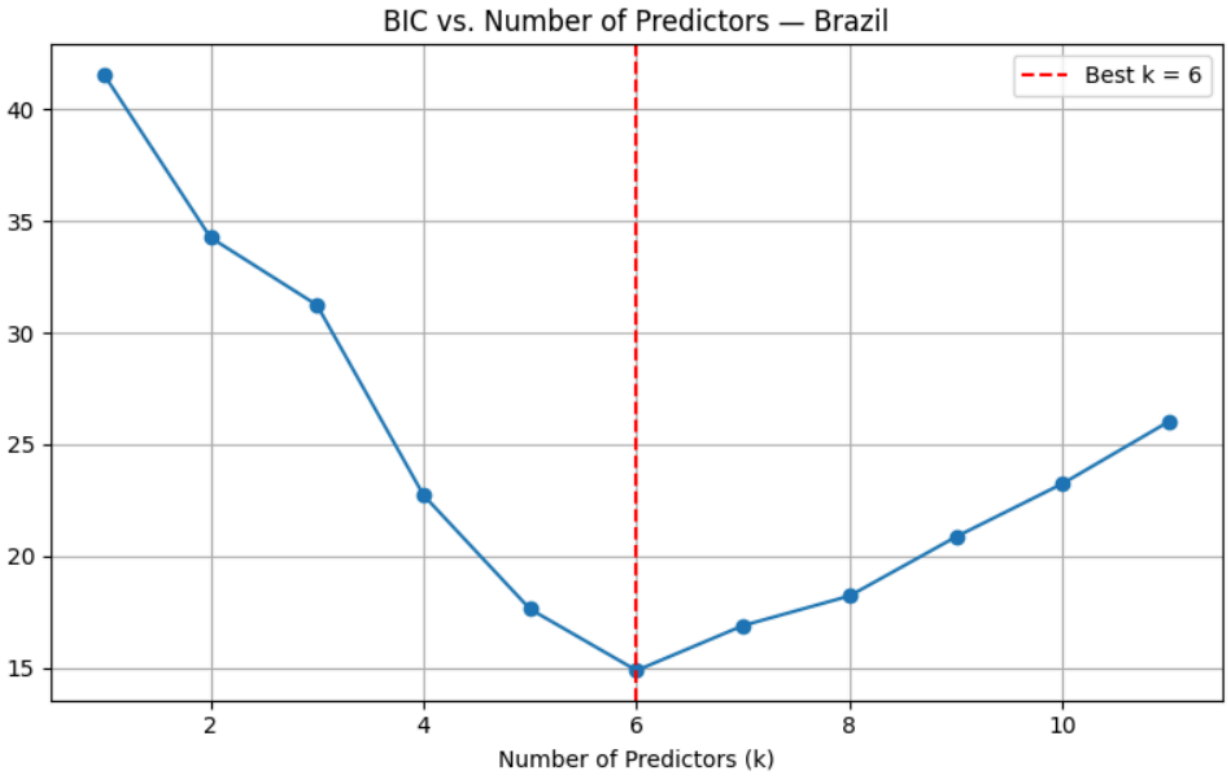
Unemployment Rate:

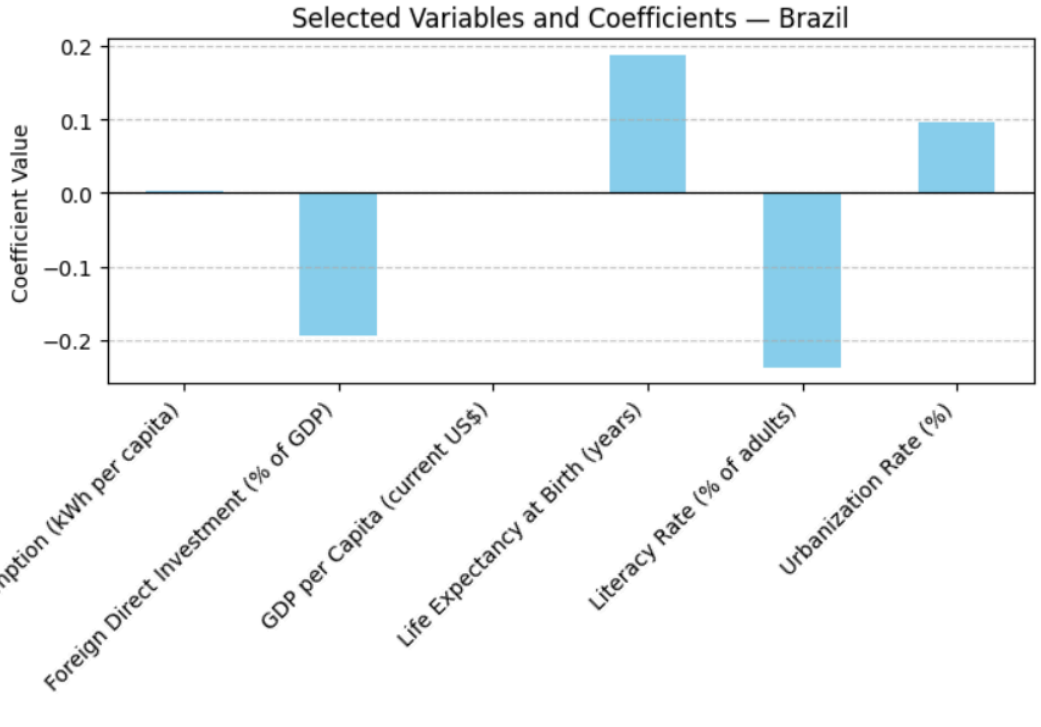




Adjusted R-squared: 0.7411237592440194

Education Expenditure:

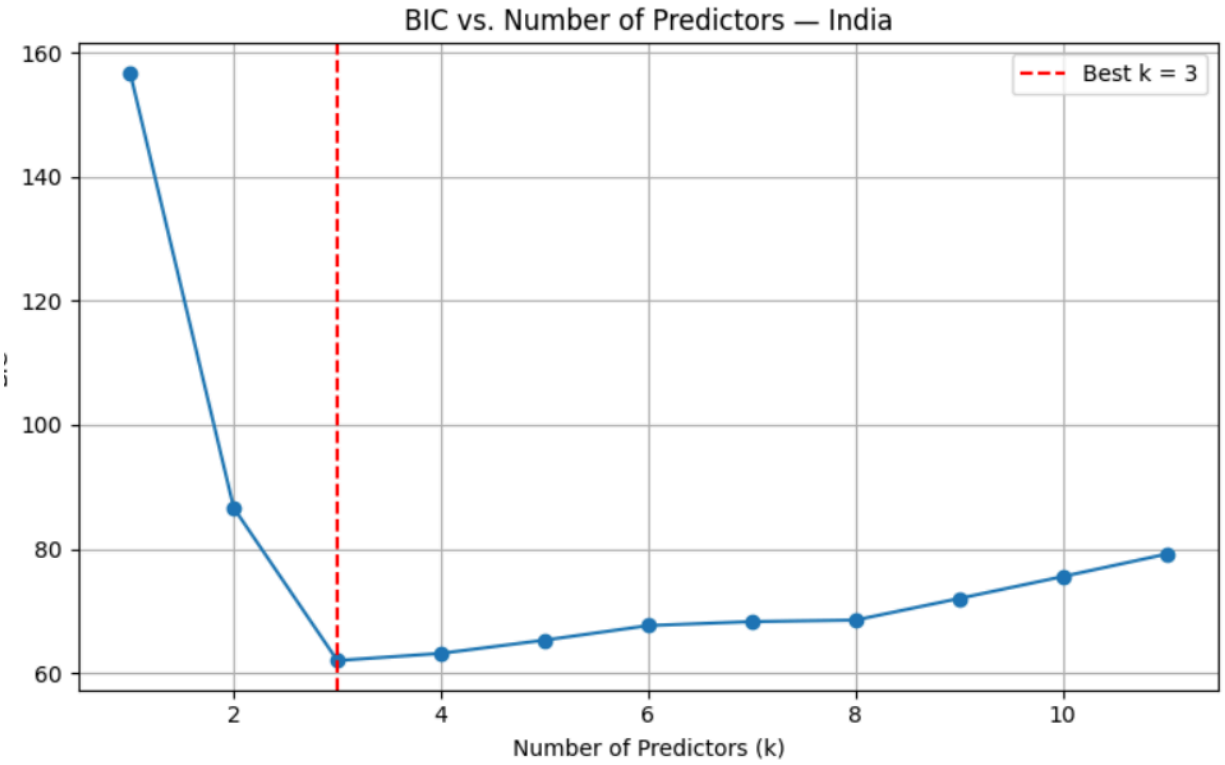


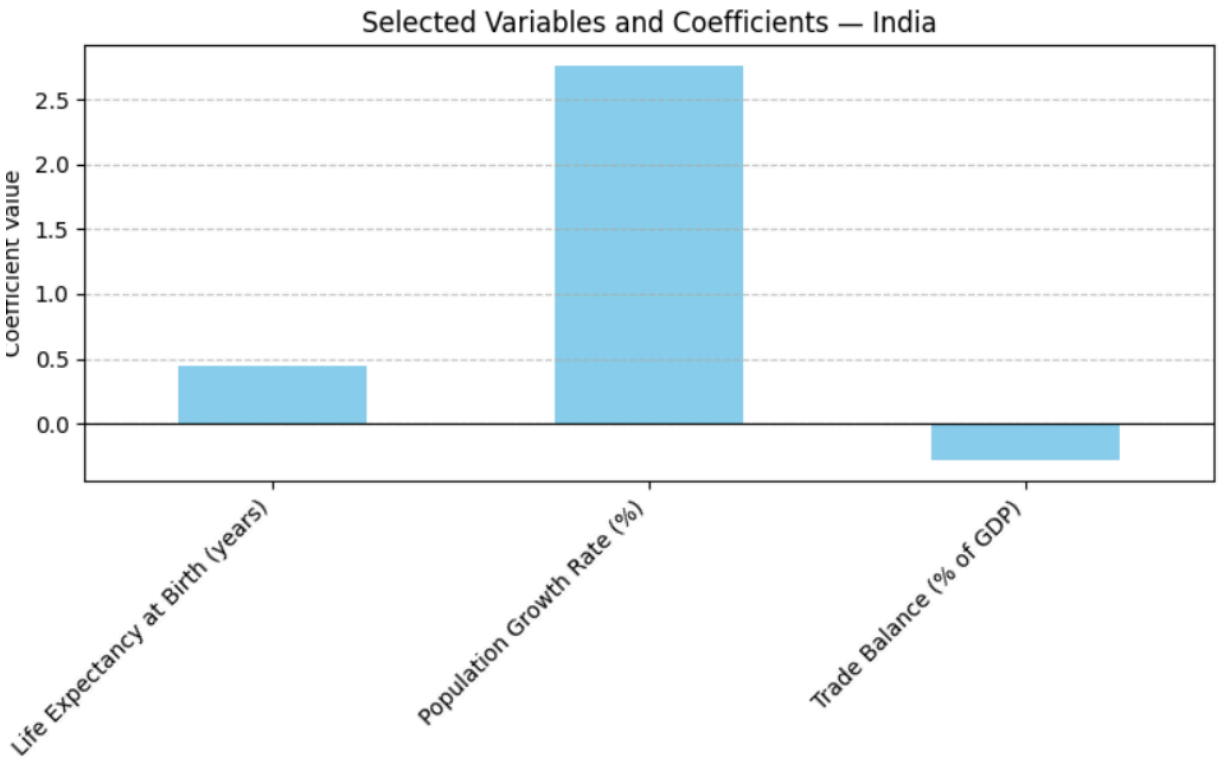


Adjusted R-squared: 0.9975711152758852

India:

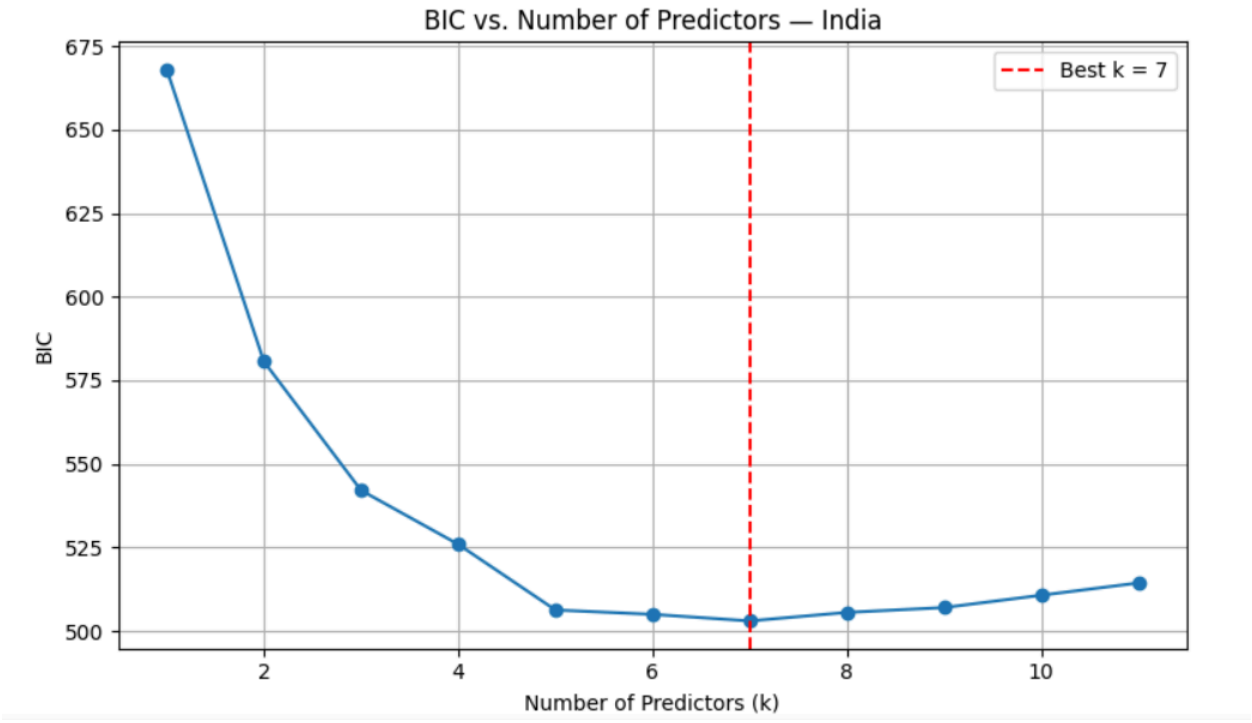
Gini Index:

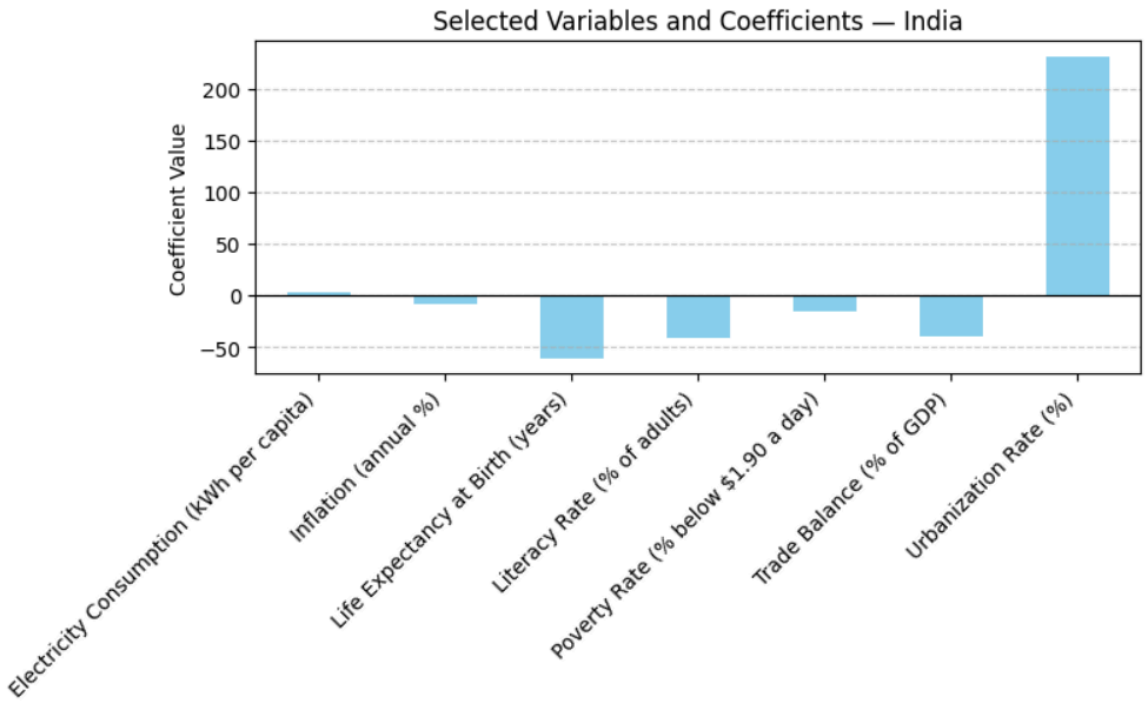




Adjusted R-squared: 0.99982

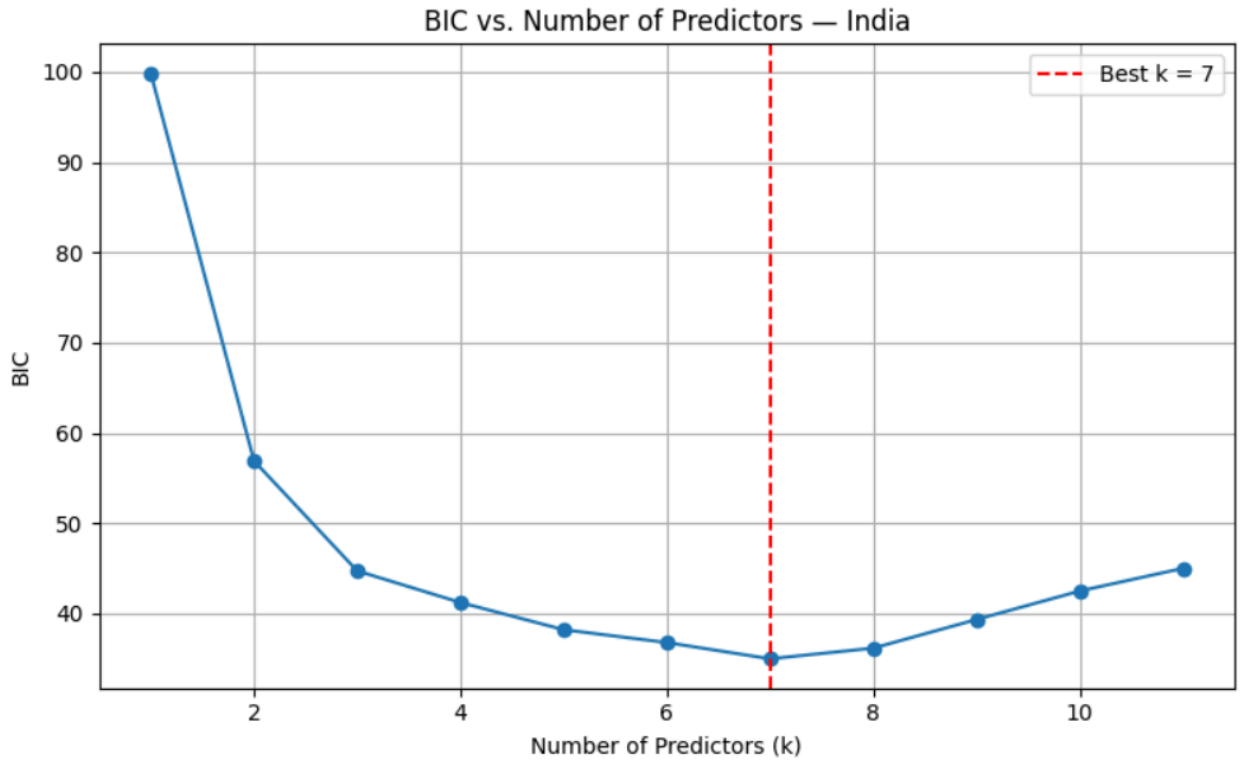
GDP per capita:

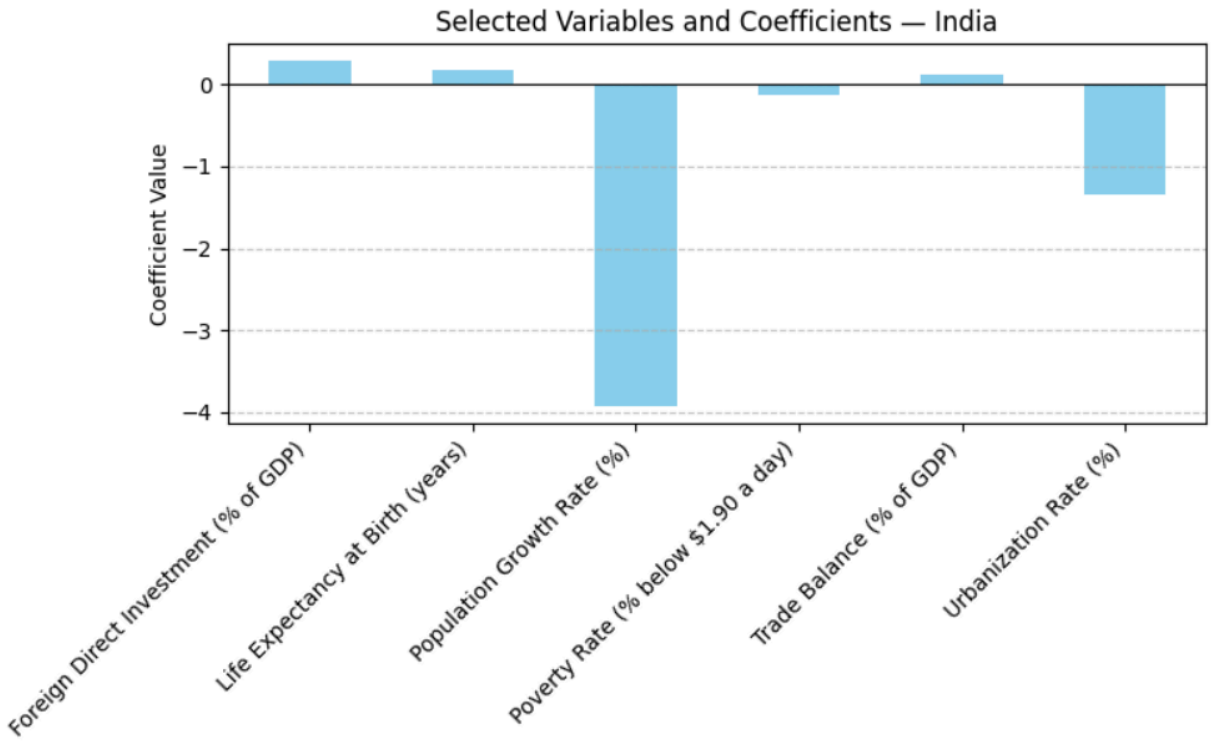




Adjusted R-squared: 0.99723

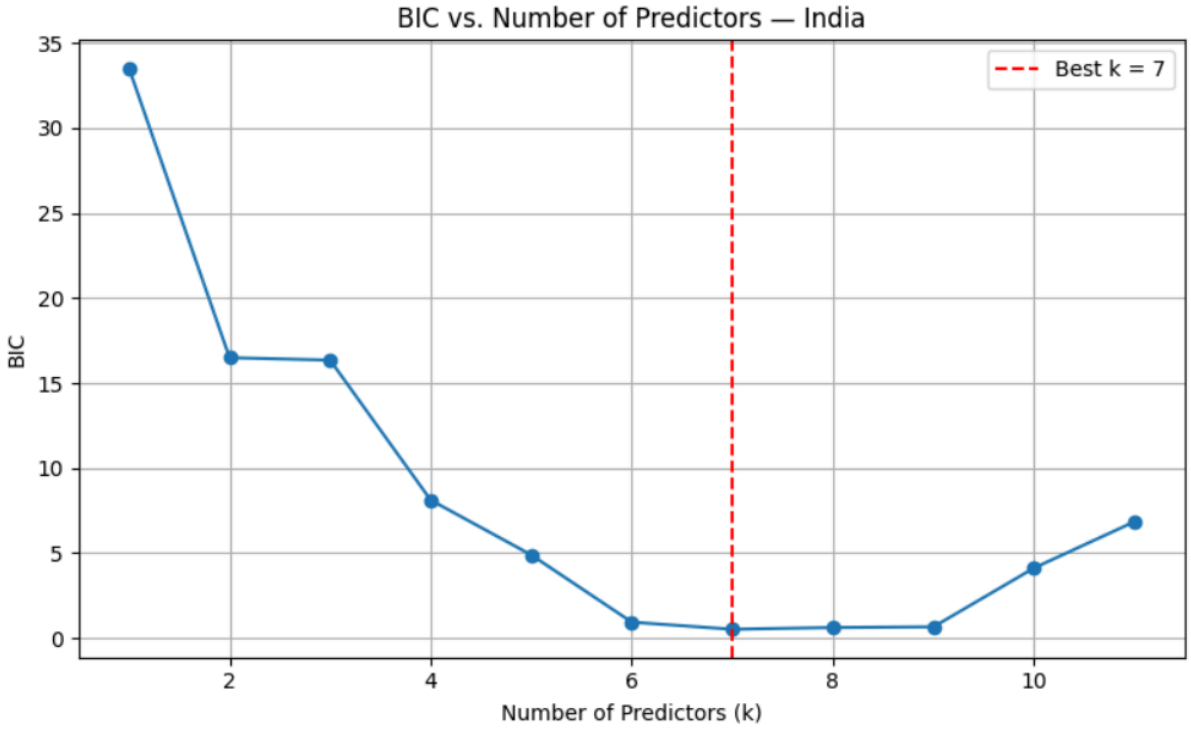
Unemployment Rate:

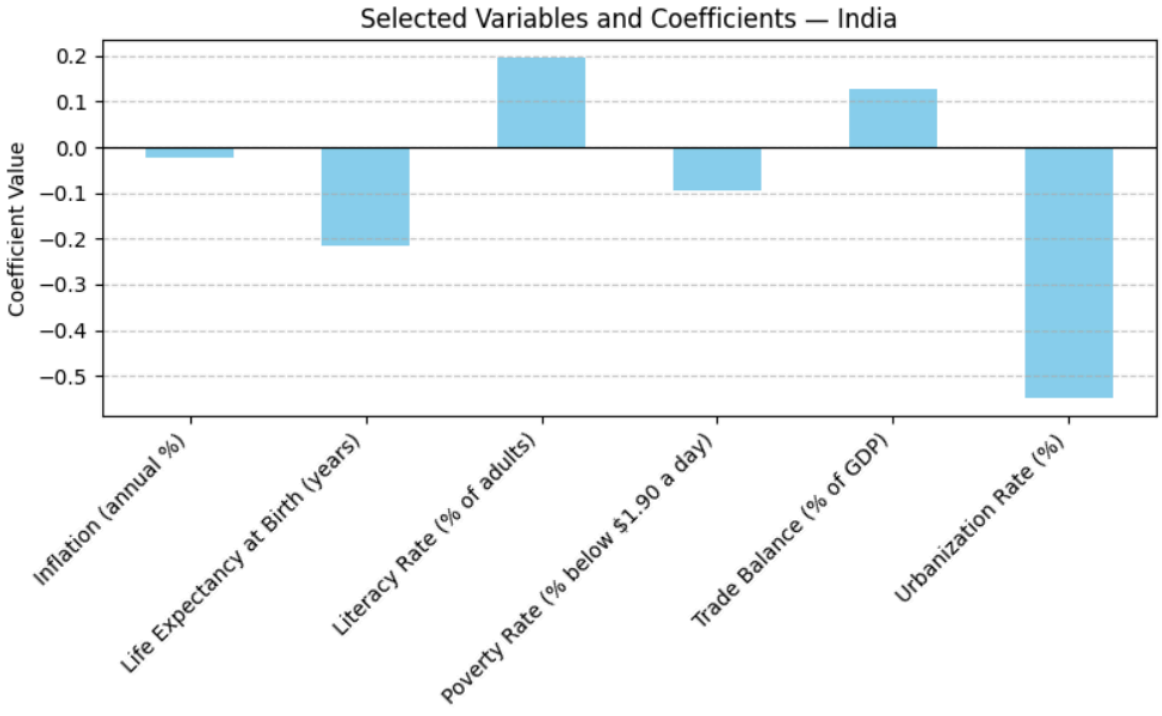




Adjusted R-squared: 0.84120

Education Expenditure:

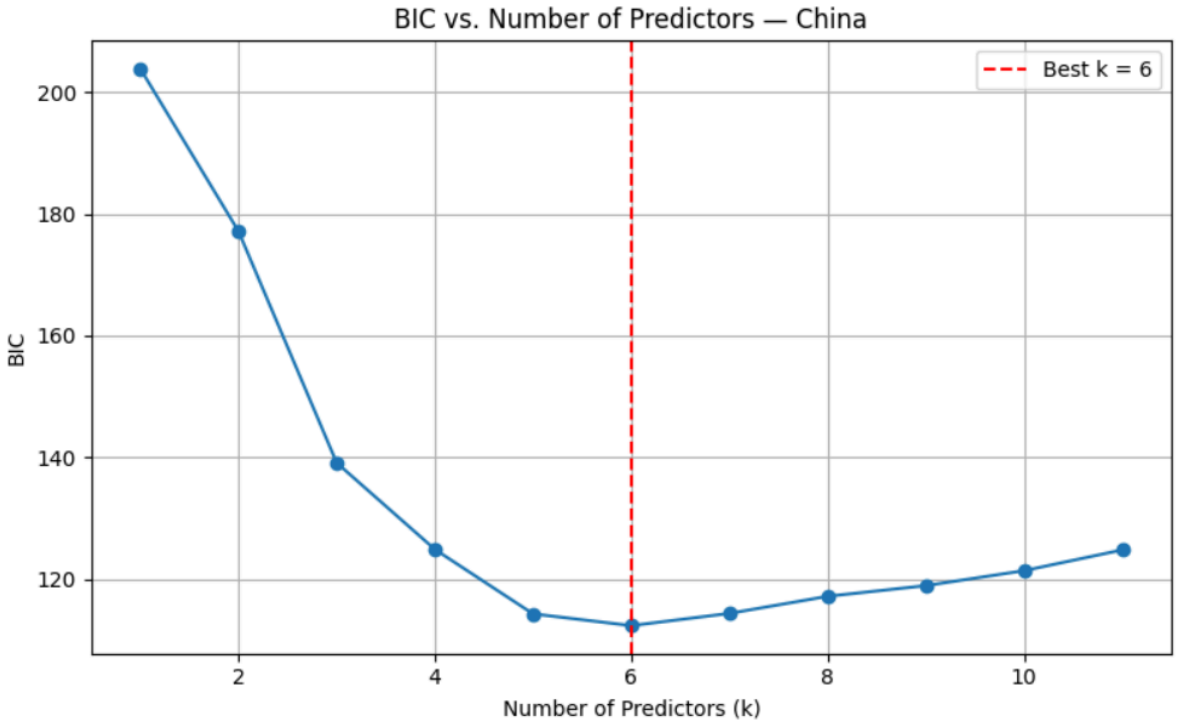




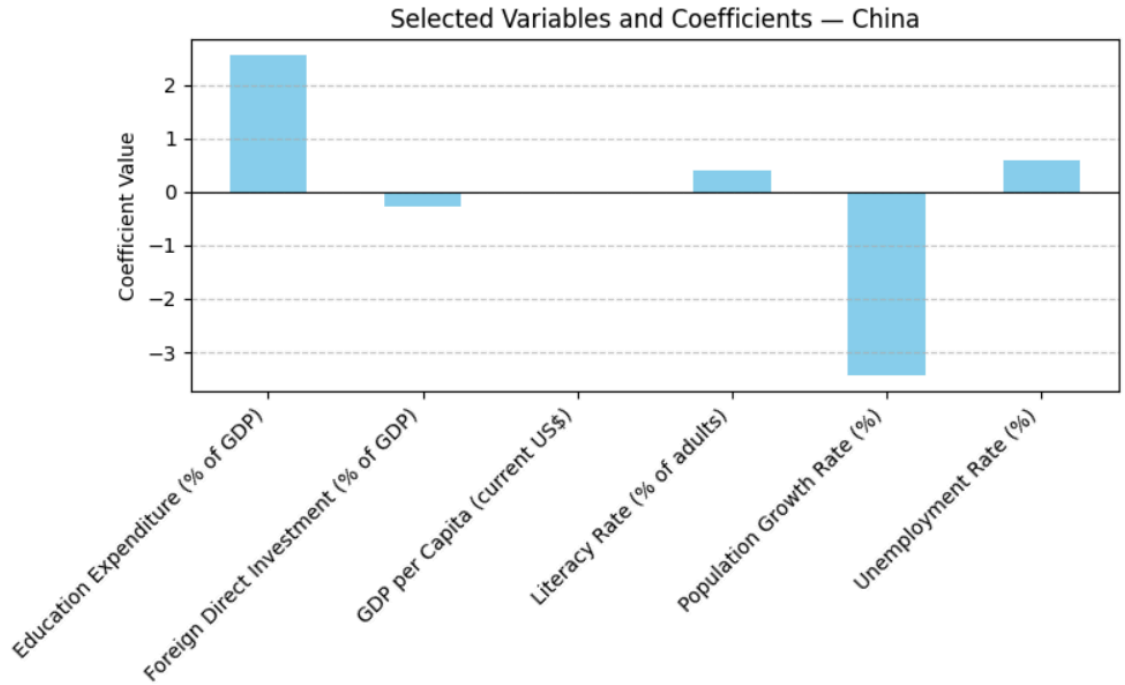
Adjusted R-squared: 0.8902357700468326

China:

Gini Index:

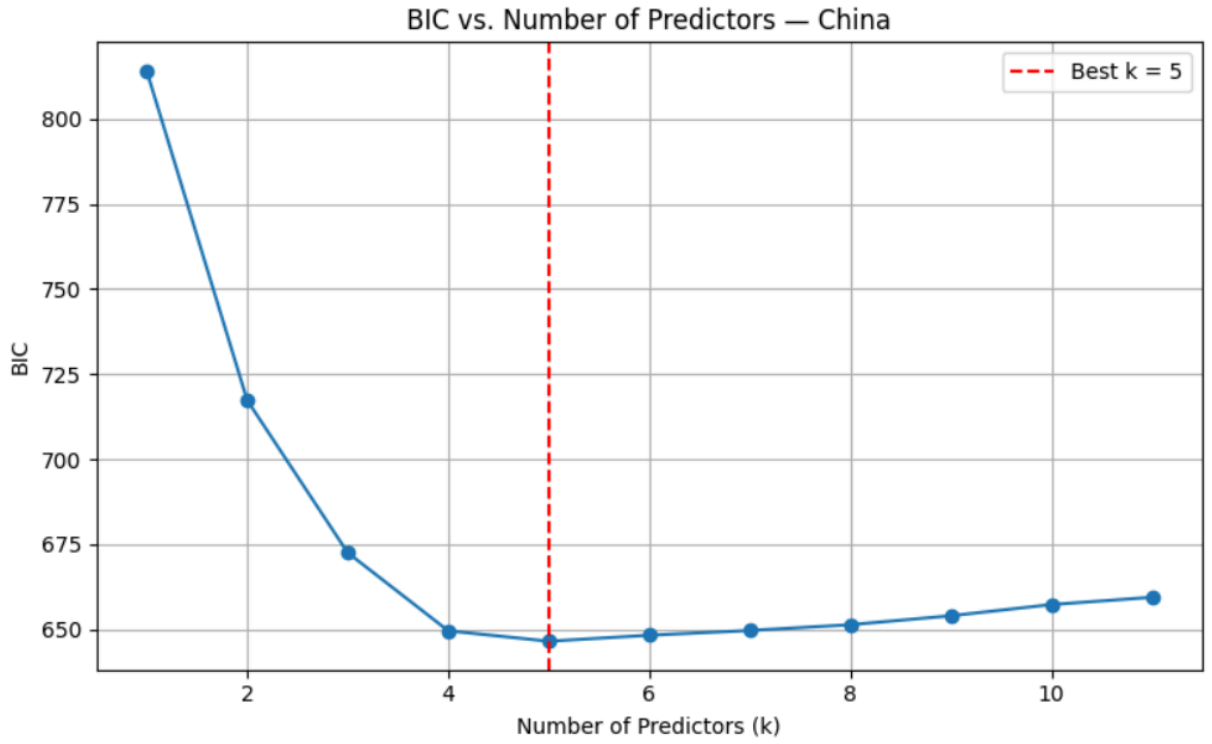


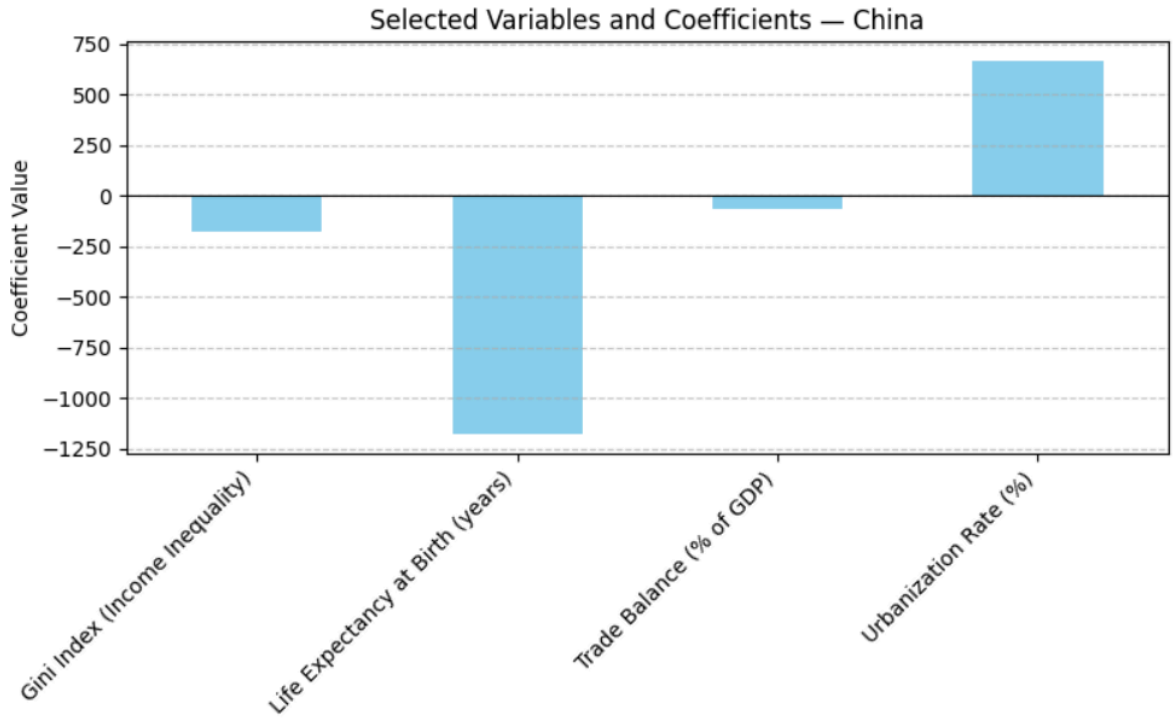
Selected Variables and Coefficients — China



Adjusted R-squared: 0.99962

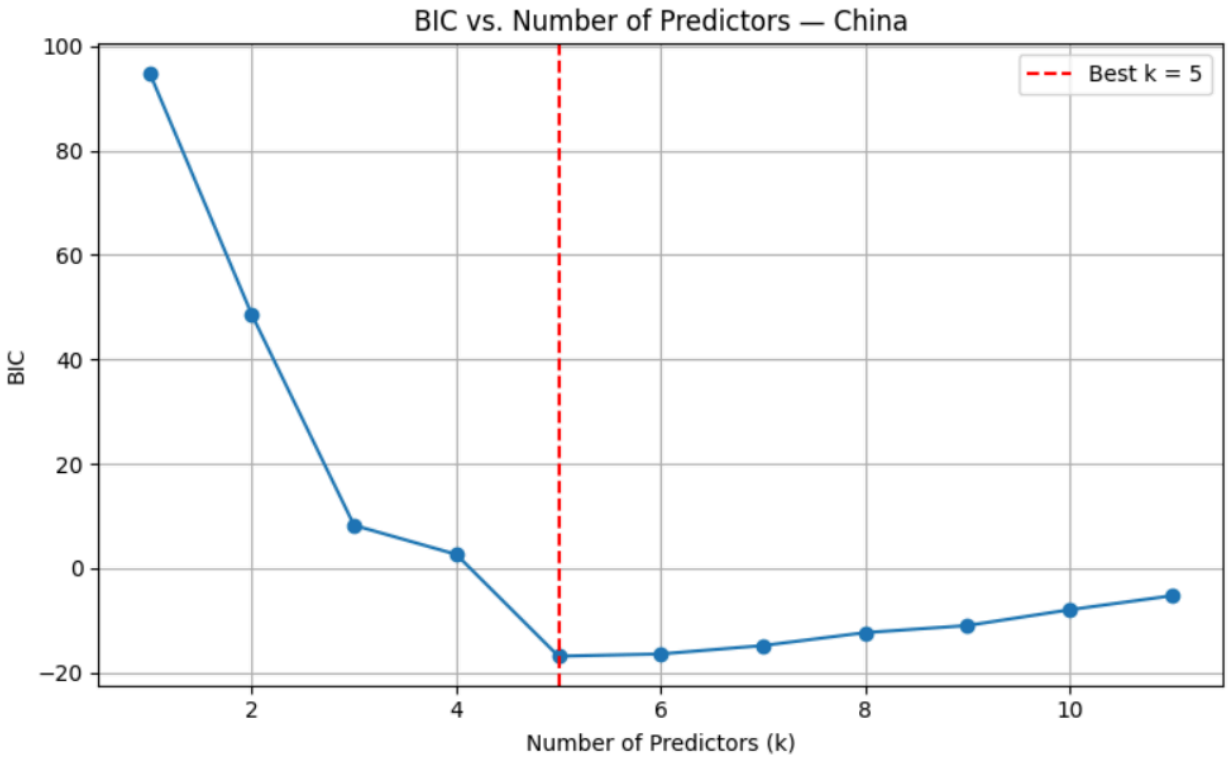
GDP per capita:

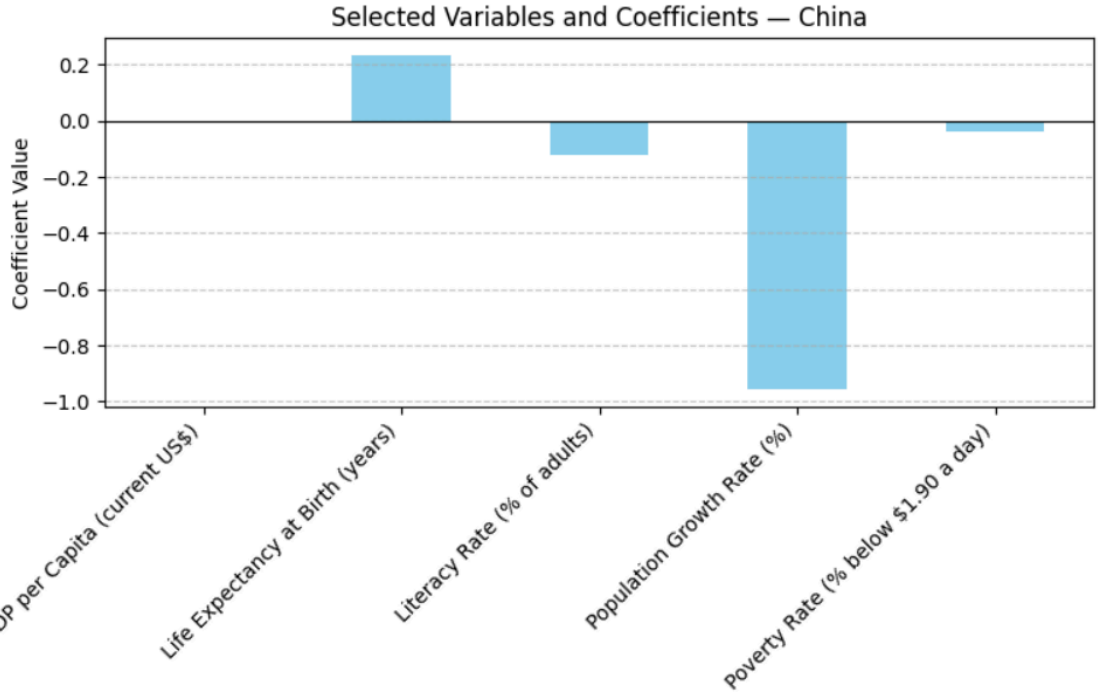




Adjusted R-squared: 0.99382

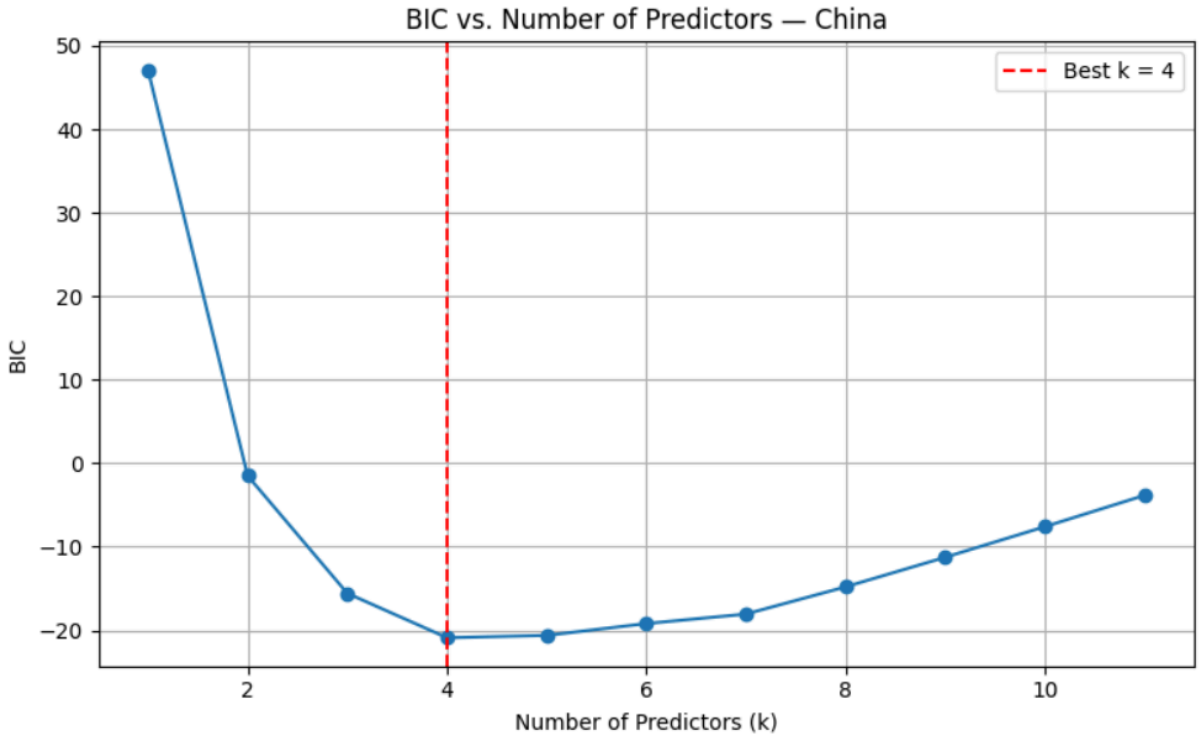
Unemployment Rate:

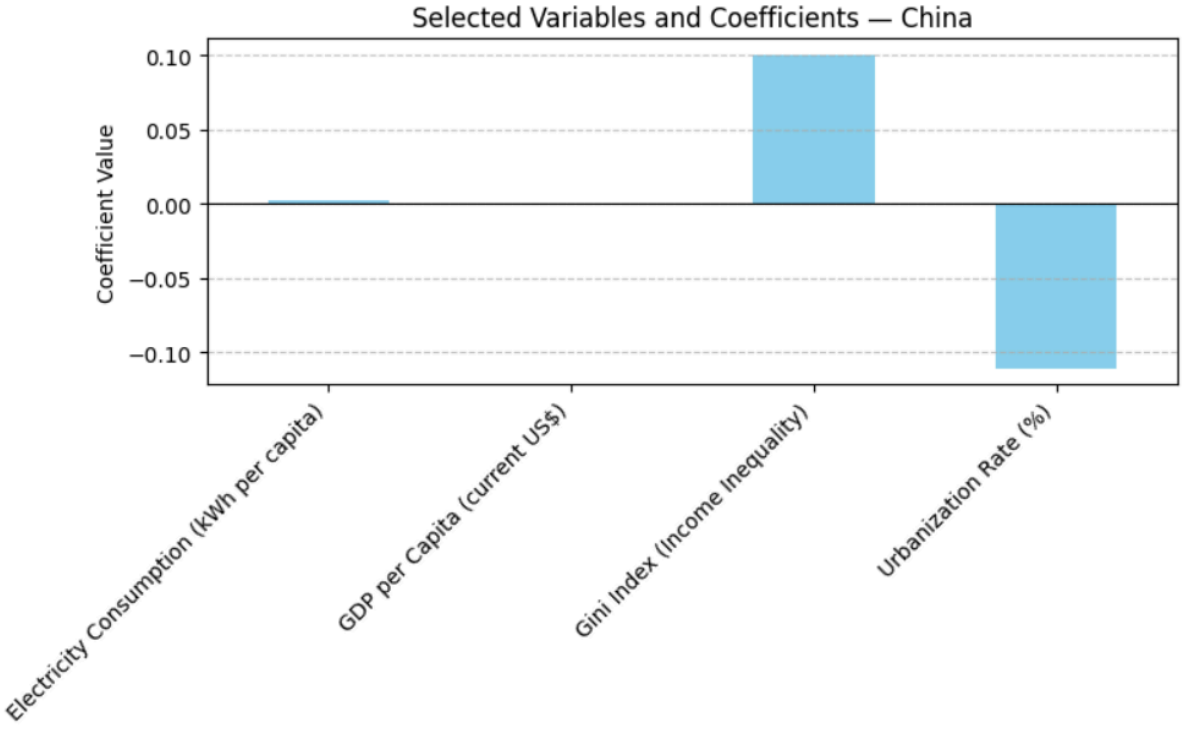




Adjusted R-squared: 0.99822

Education Expenditure:

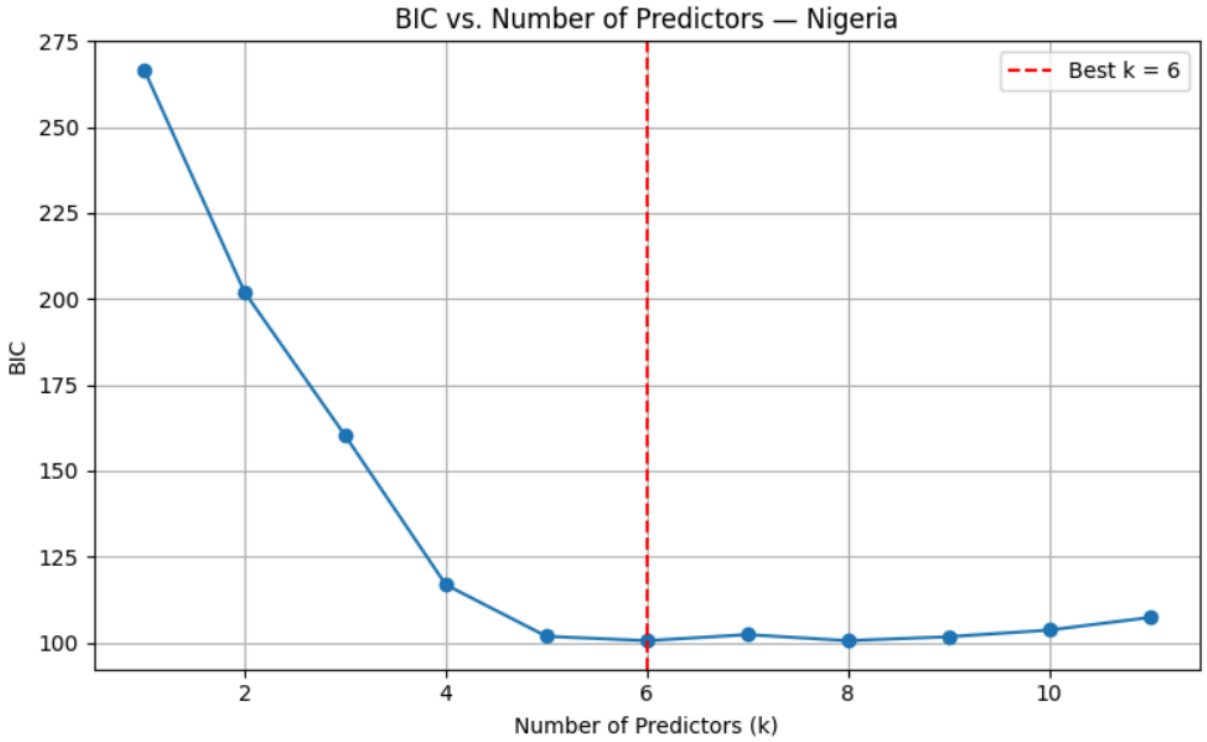


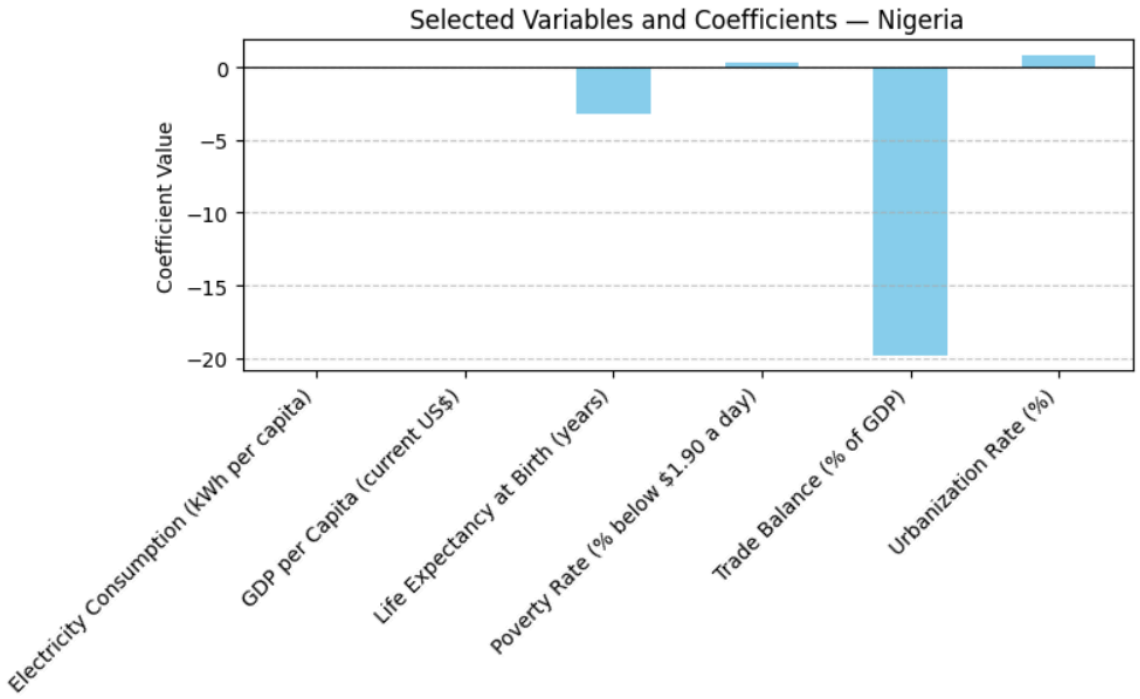


Adjusted R-squared: 0.99662

Nigeria:

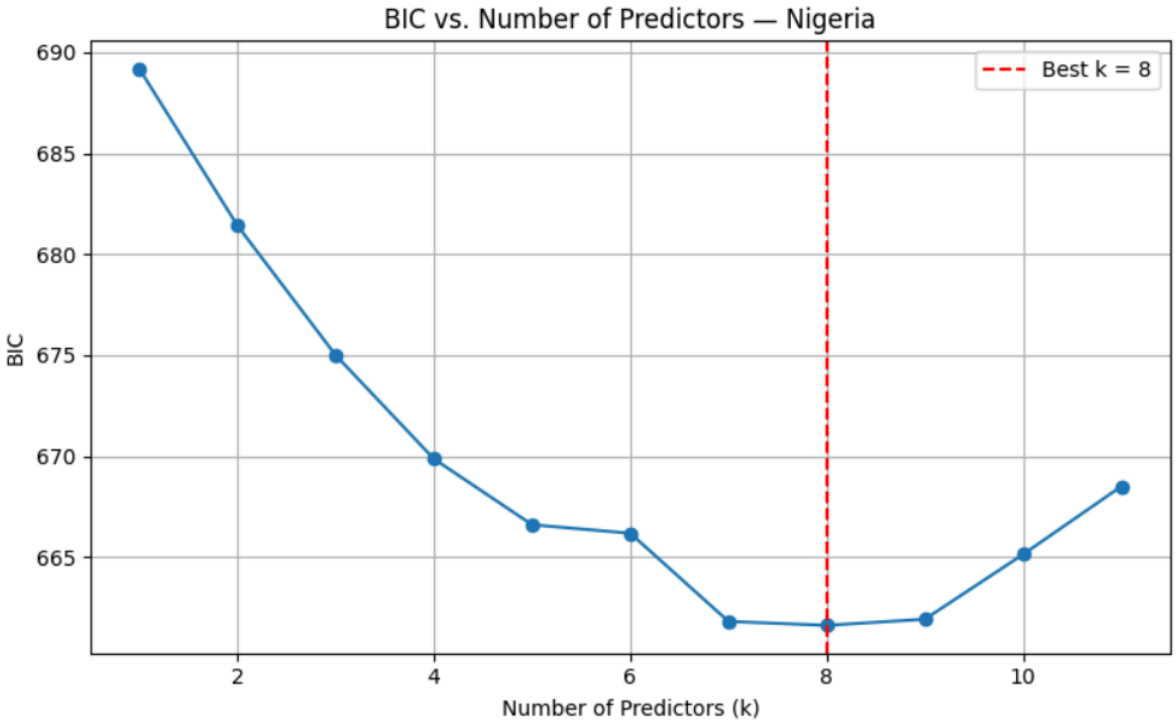
Gini Index:

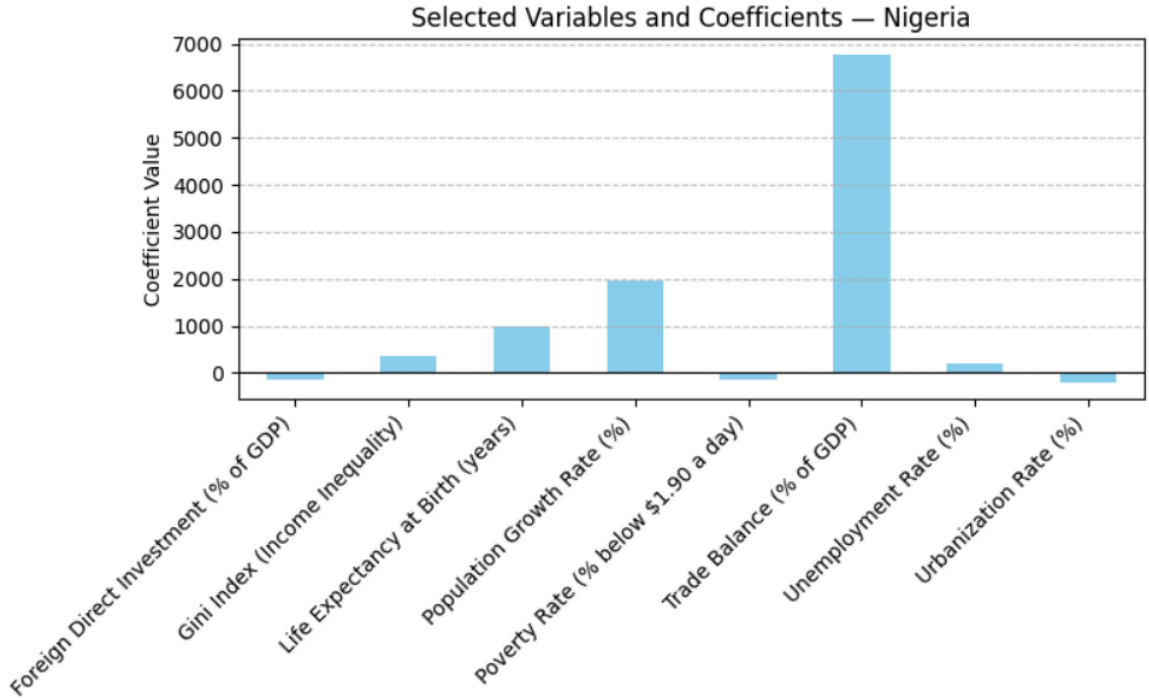




Adjusted R-squared: 0.98315

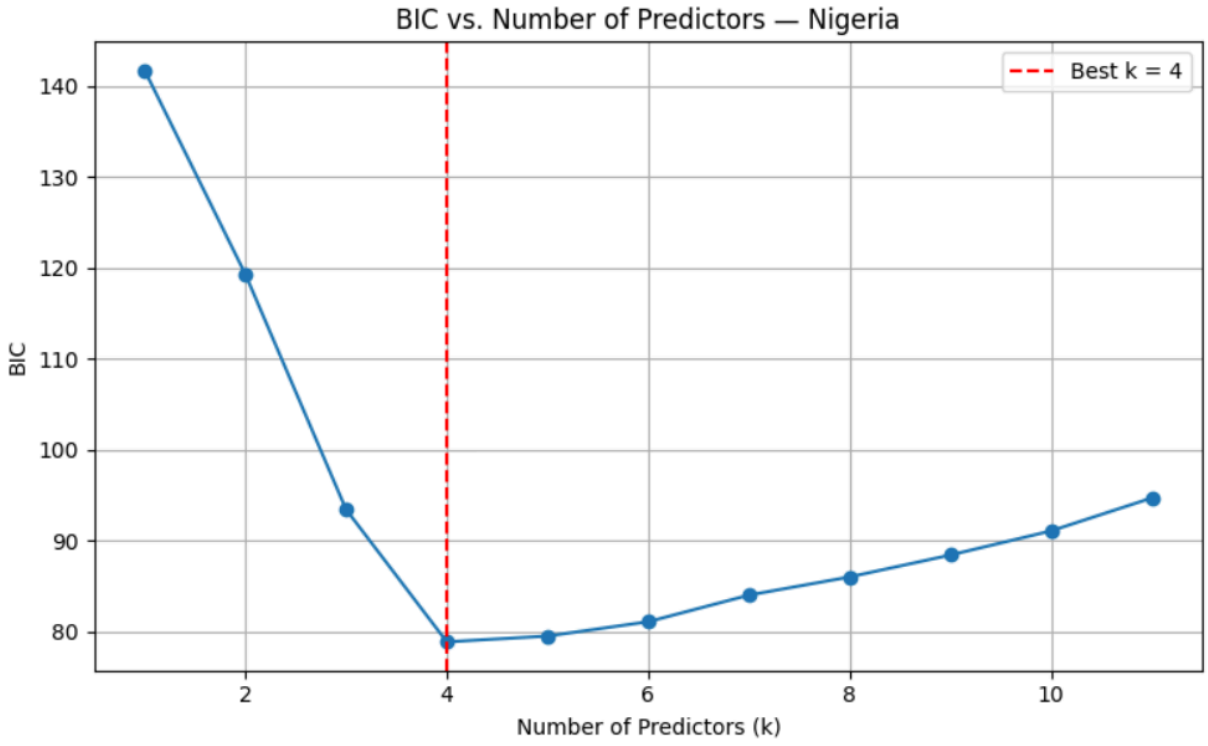
GDP per capita:

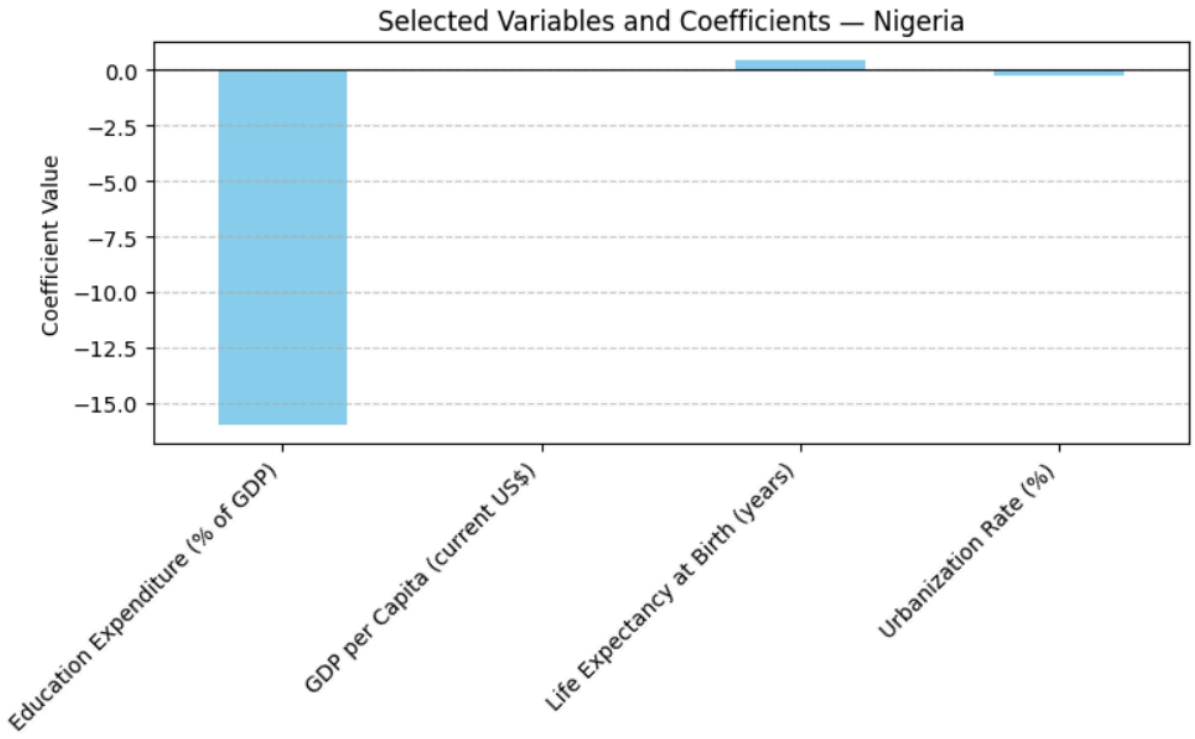




Adjusted R-squared: 0.791928

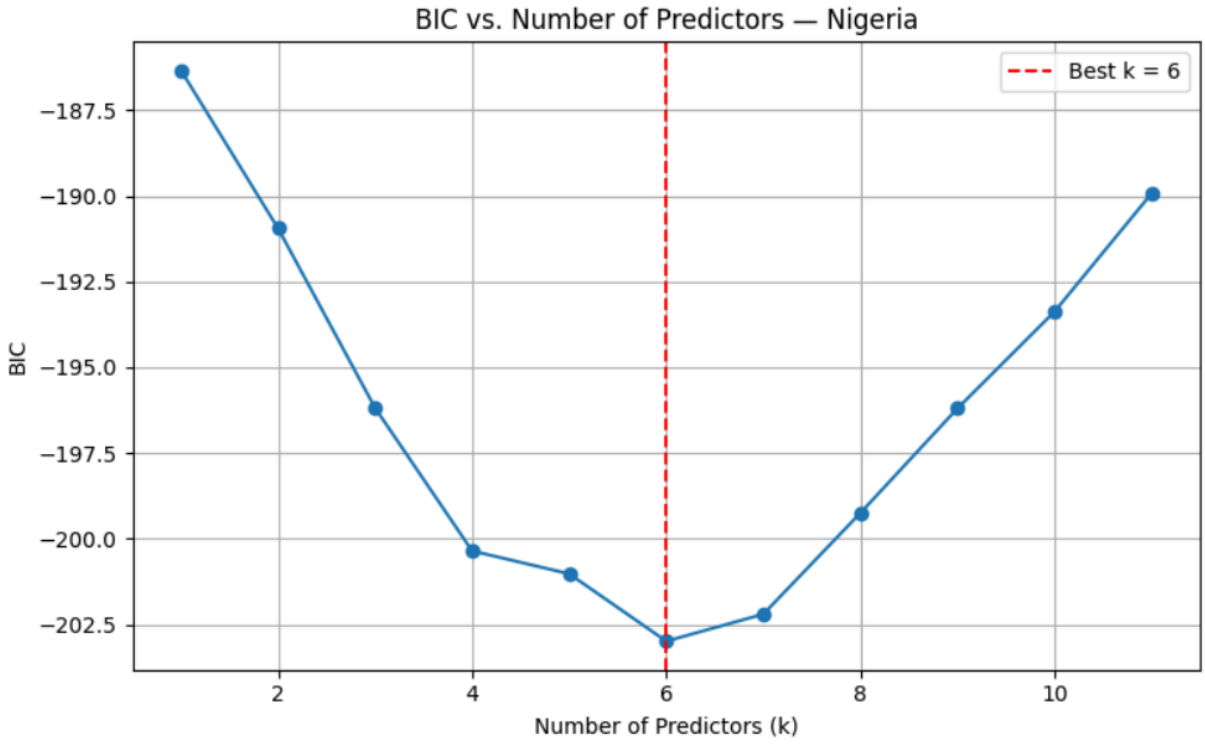
Unemployment Rate:

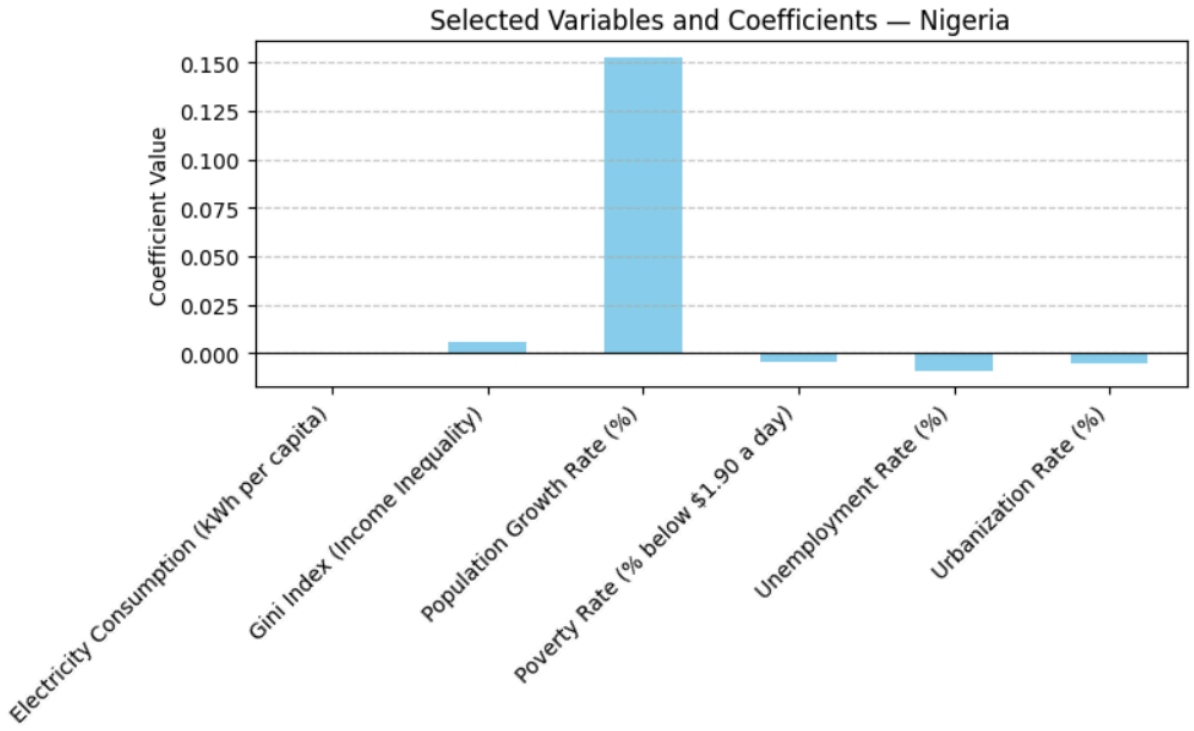




Adjusted R-squared: 0.98772

Education Expenditure:

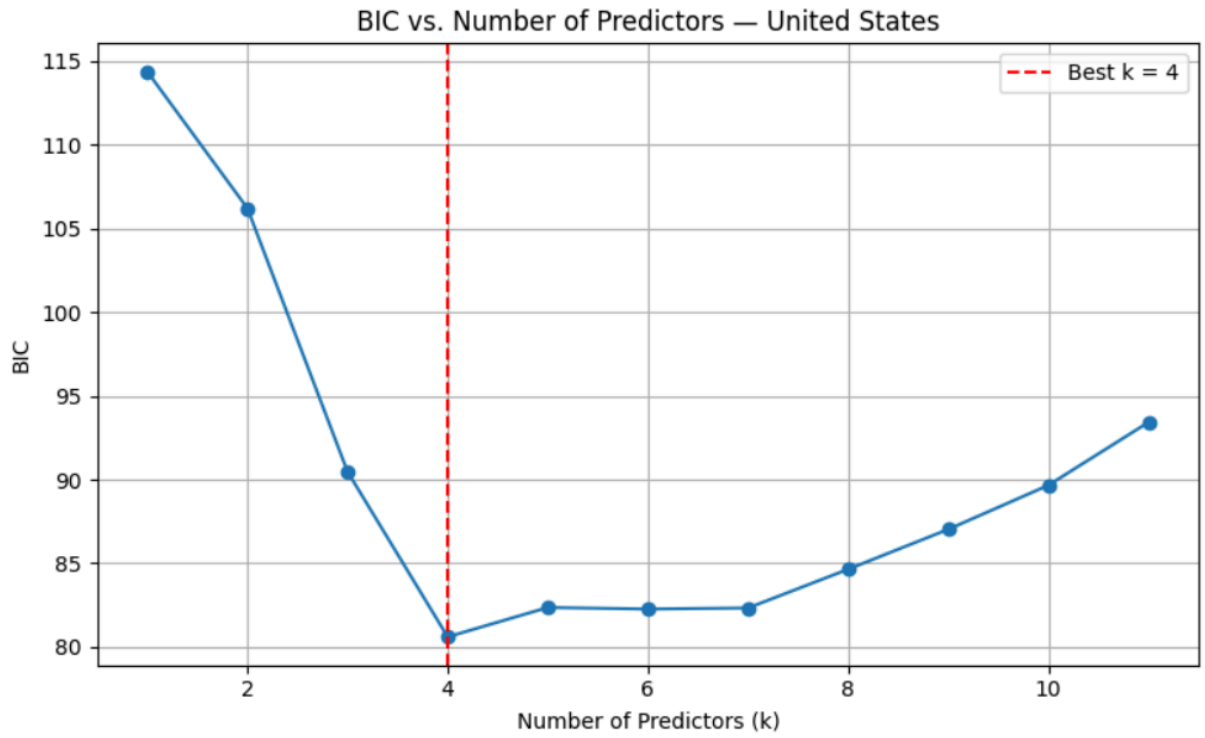


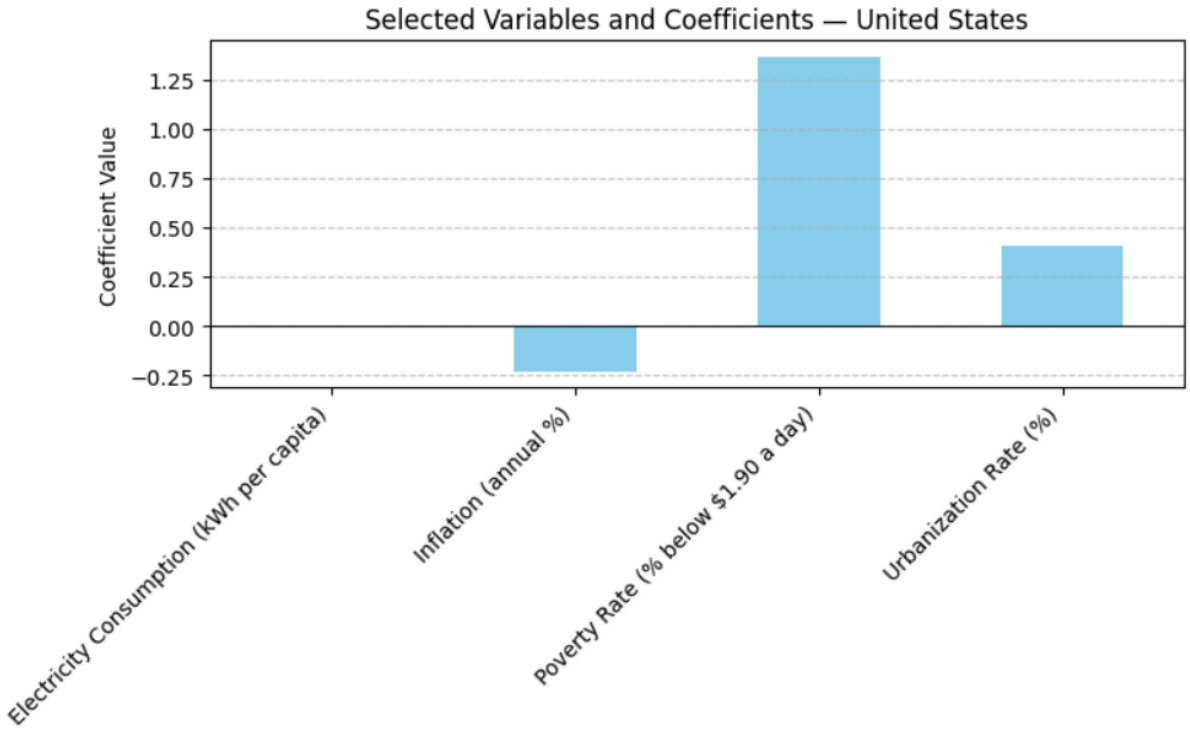


Adjusted R-squared: 0.99826

USA:

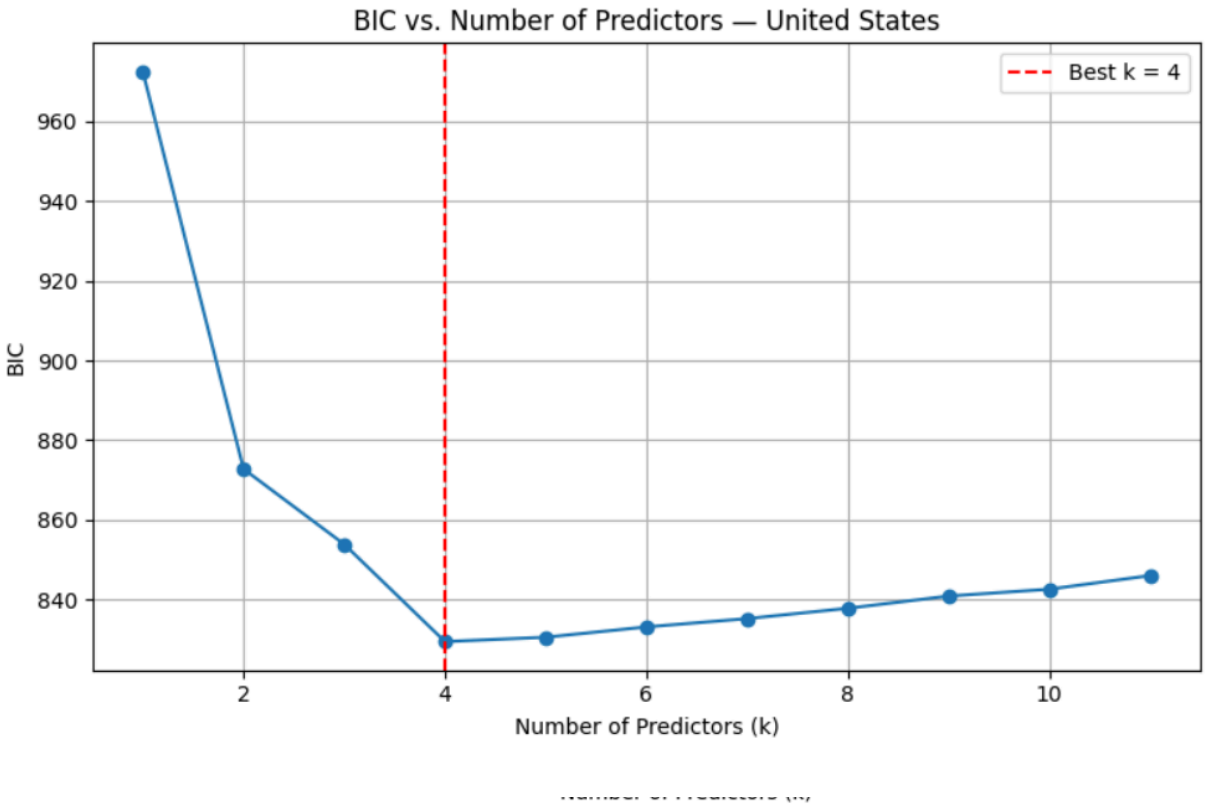
Gini Index:

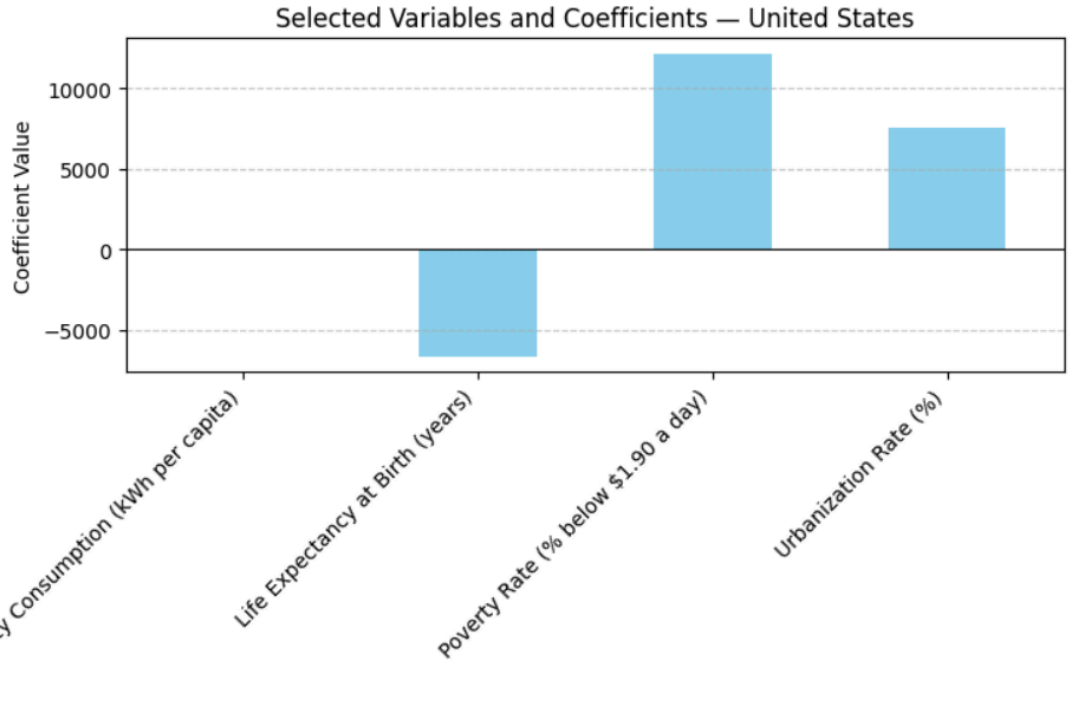




Adjusted R-squared: 0.99981

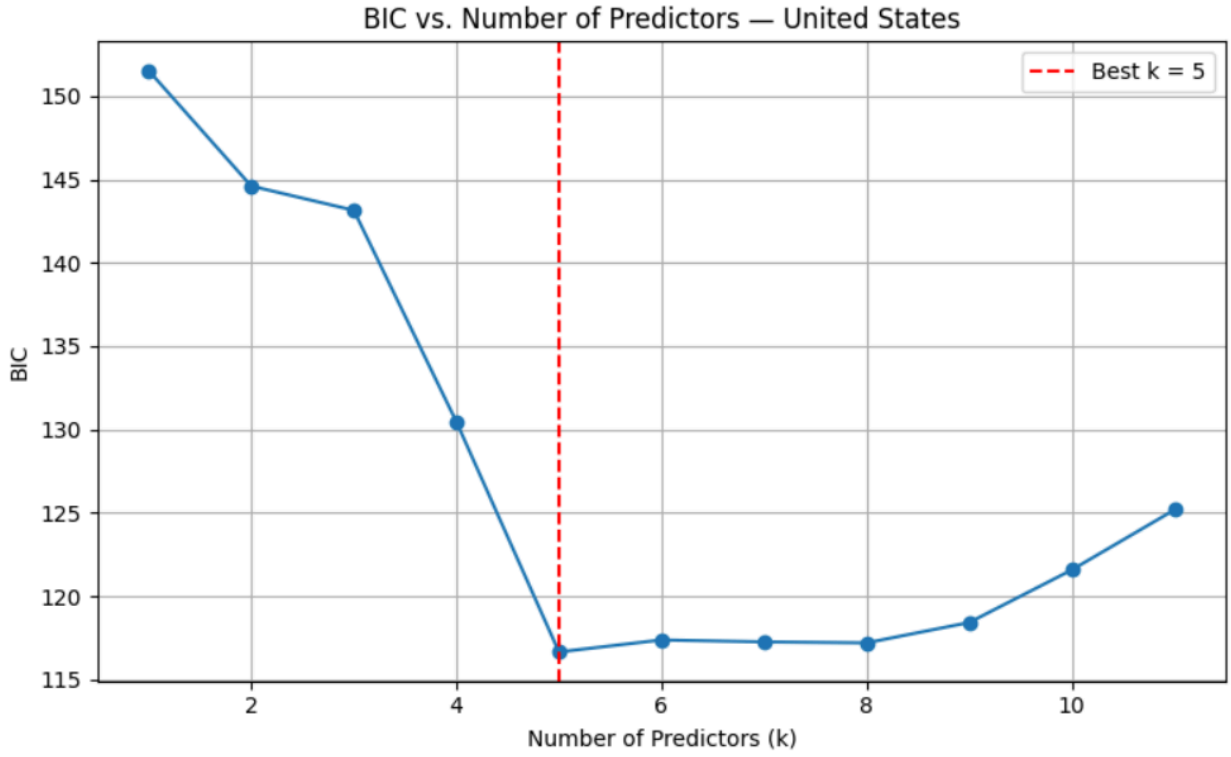
GDP per capita:

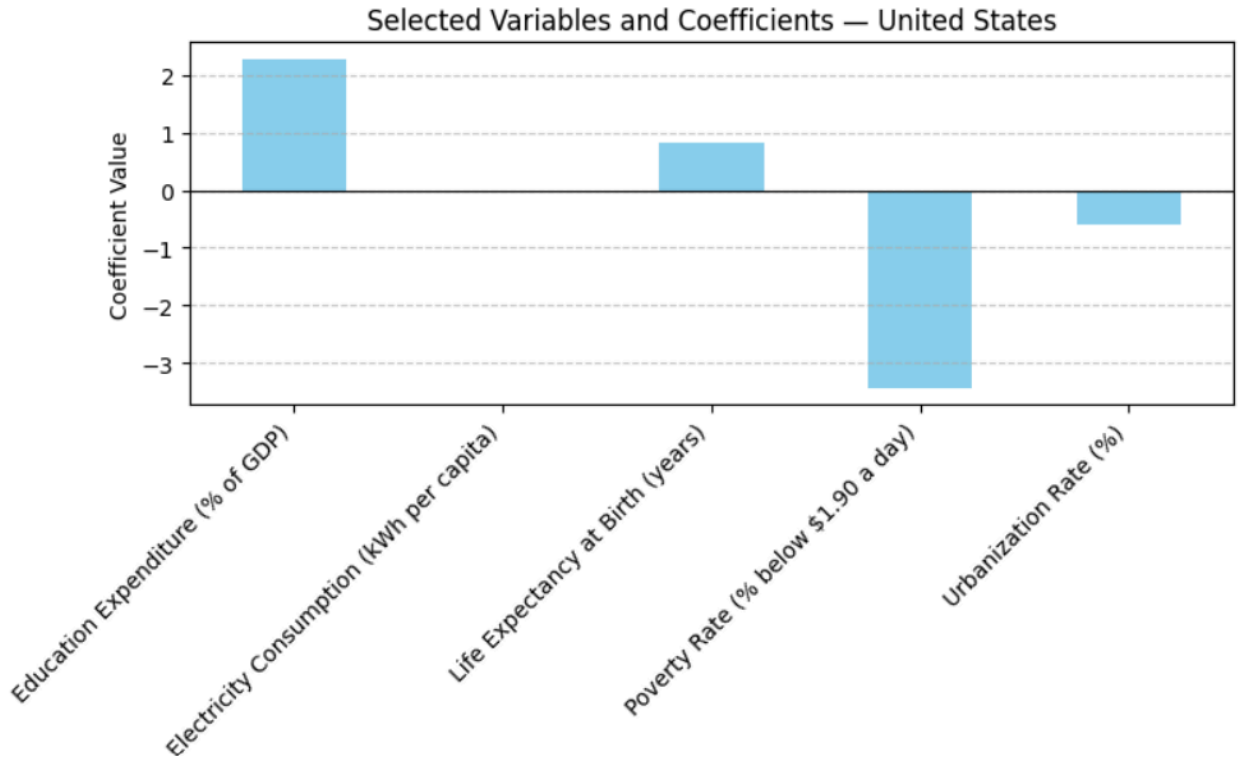




Adjusted R-squared: 0.99632

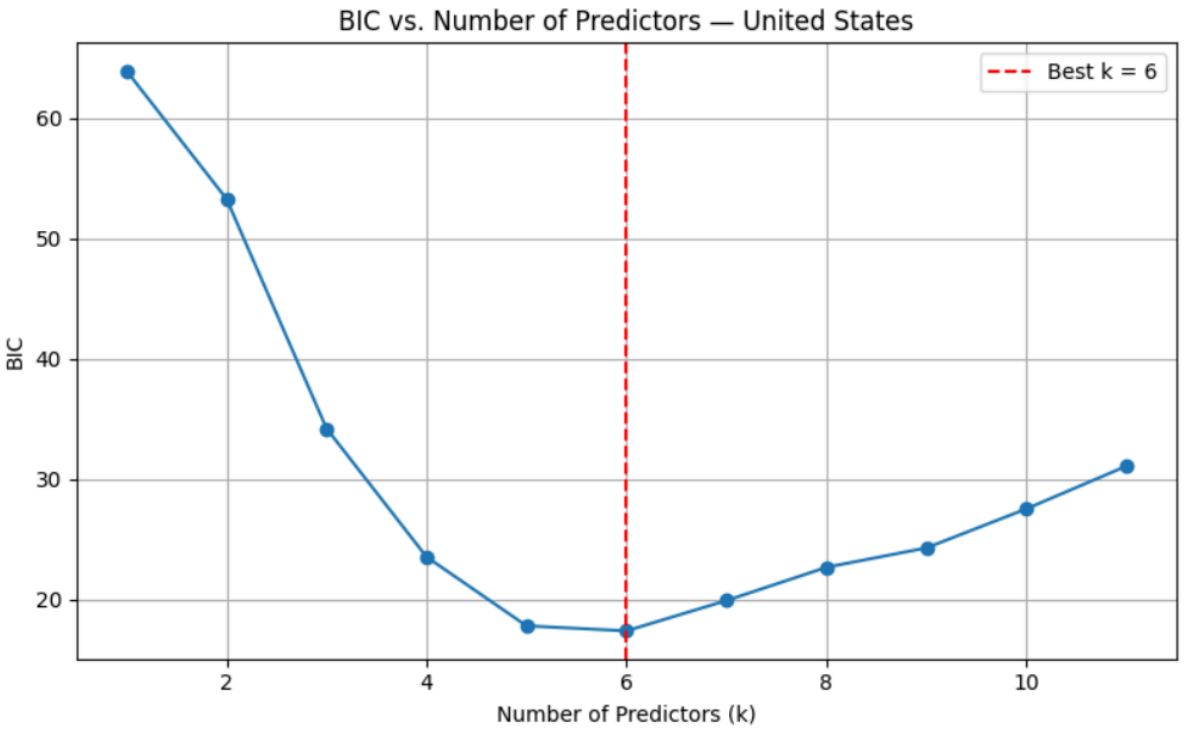
Unemployment Rate:

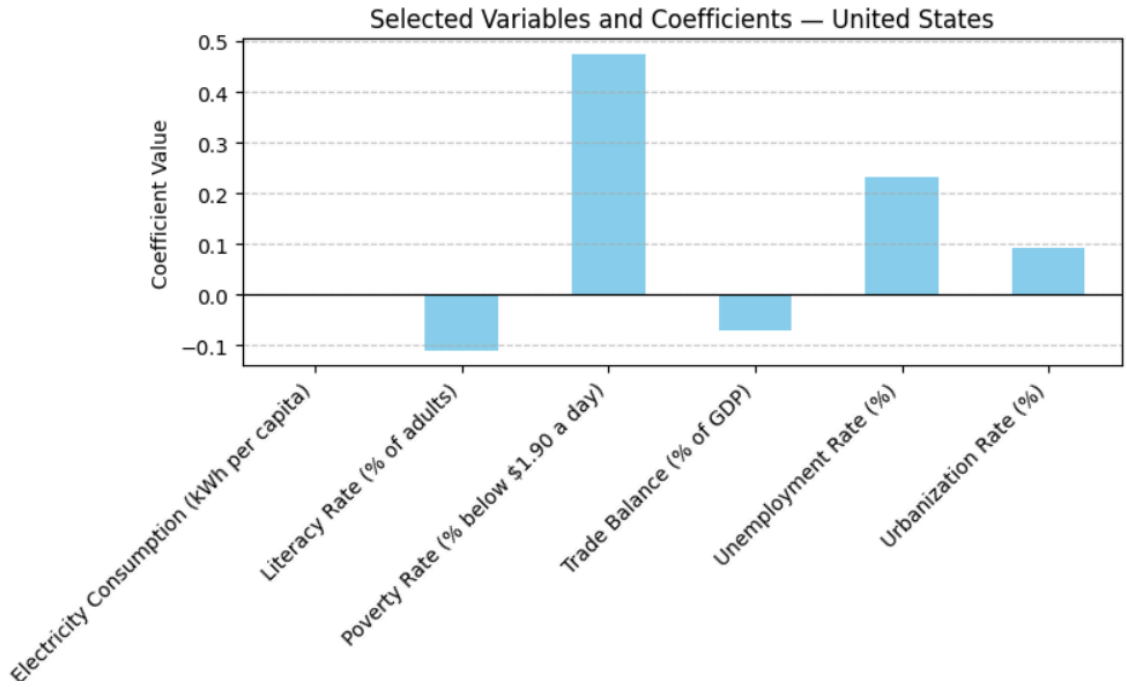




Adjusted R-squared: 0.98286

Education Expenditure:

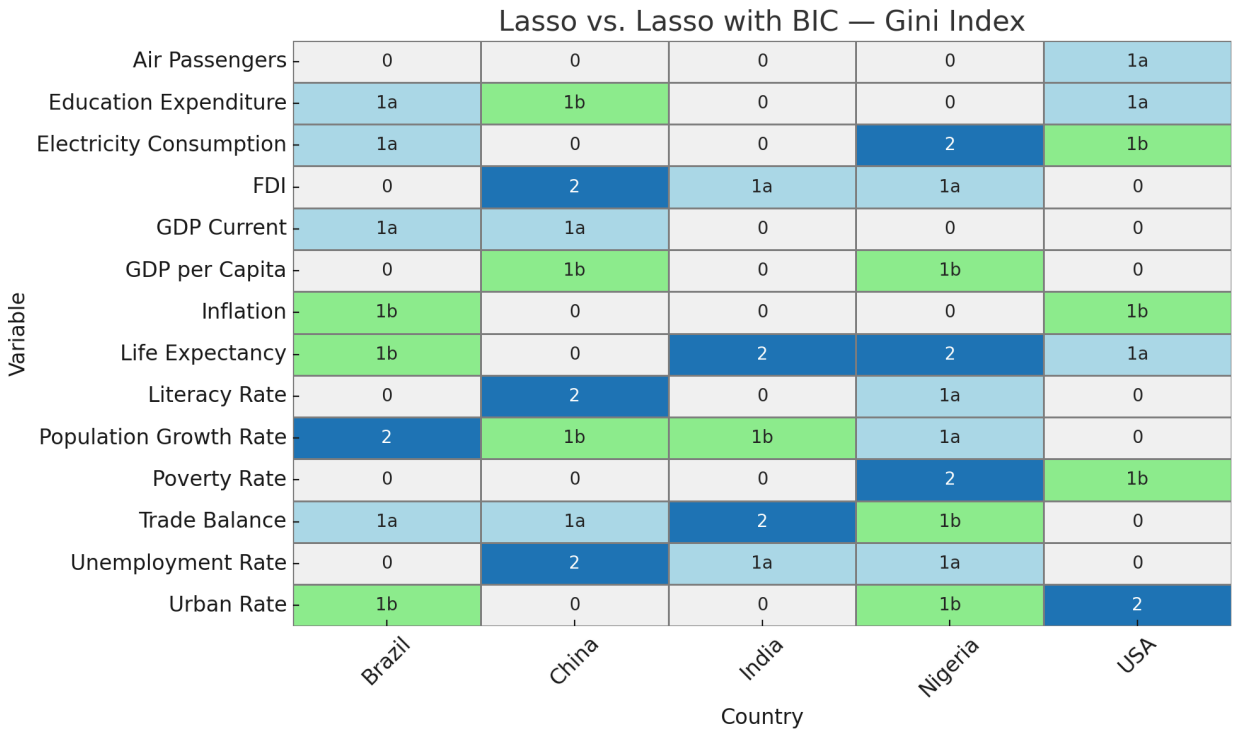




Adjusted R-squared: 0.790553246069607

Variable level insights:

Gini index:



Legend

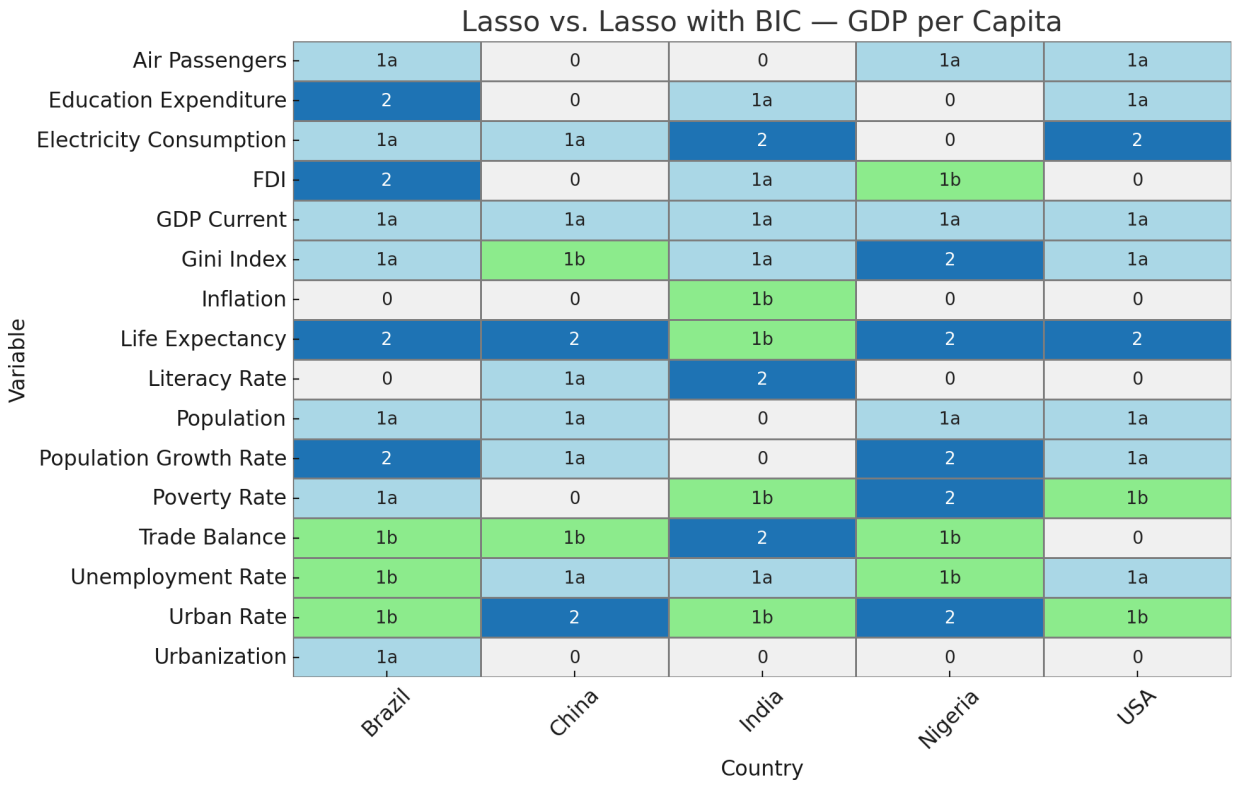
- **0 (white) → Not selected by either model**
- **1a (light blue) → Selected only by Lasso**
- **1b (light green) → Selected only by Lasso with BIC**
- **2 (dark blue) → Selected by both Lasso and BIC**

Insight:

- The shift from Lasso to Lasso with BIC for the Gini Index reveals a clear transition from economic volume and infrastructure variables to more demographic and structural development indicators.
- Lasso models frequently included variables such as *GDP current*, *FDI*, *electricity consumption*, and *trade balance* — all of which are high-level economic markers. However, after BIC penalization, these were often dropped in favor of more interpretable variables like *urban rate*, *population growth*, and *life expectancy*.

- Notably, *electricity consumption* and *life expectancy* were the only variables that maintained relevance across both methods, suggesting that access to basic infrastructure and quality of life remain essential in explaining inequality, even when the model is regularized.
- This shift indicates that BIC favors structural, human-centered explanations of inequality over macroeconomic indicators that may be collinear or redundant.

GDP per Capita:

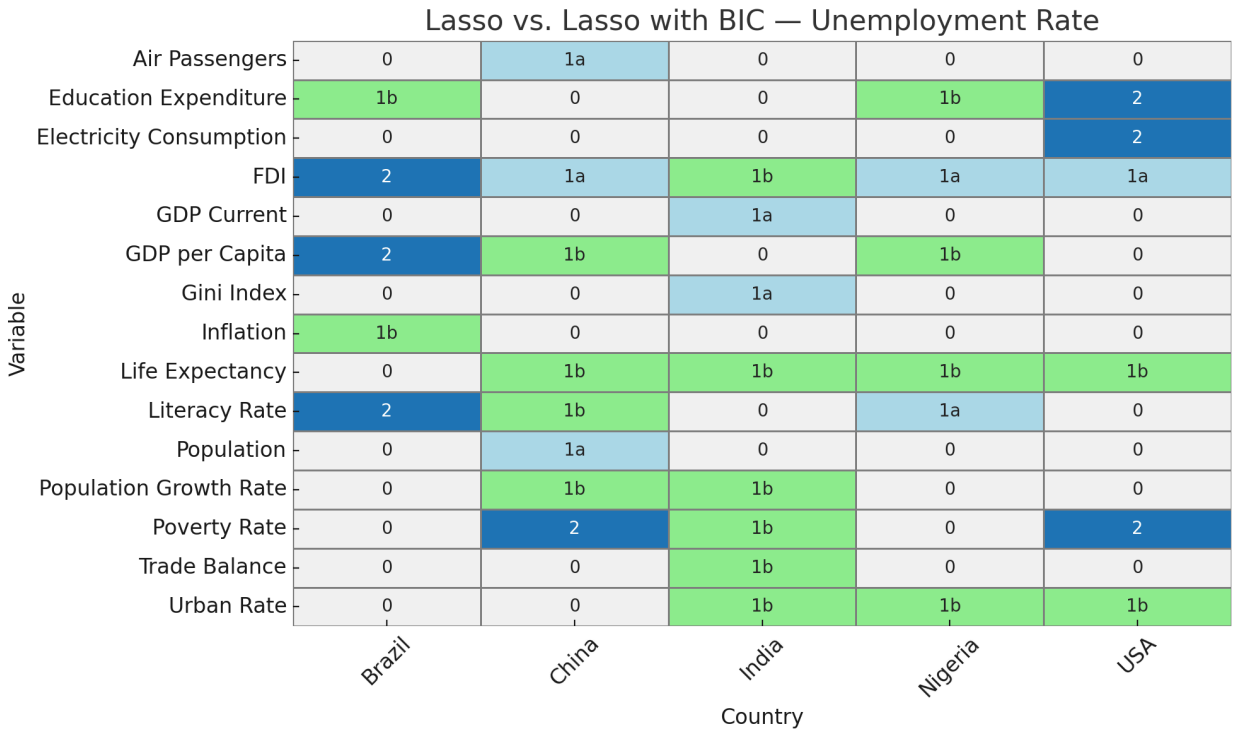


Insight:

- For GDP per Capita, the Lasso models leaned heavily on macroeconomic scale and infrastructure-related variables, including *GDP current*, *air passengers*, and *electricity consumption*.
- These models were large and featured a broad variety of predictors. In contrast, BIC-based models distilled this down to a leaner set of demographic and social development indicators like *life expectancy*, *poverty rate*, and *urban rate*.
- Variables like *GDP current* were often selected only by Lasso, while *life expectancy* emerged as a consistently retained variable across countries in both models, underlining its critical role.

- The BIC models shifted the narrative from economic output toward equitable distribution and developmental outcomes, suggesting that well-being indicators may be more generalizable predictors of GDP per capita than raw economic metrics.

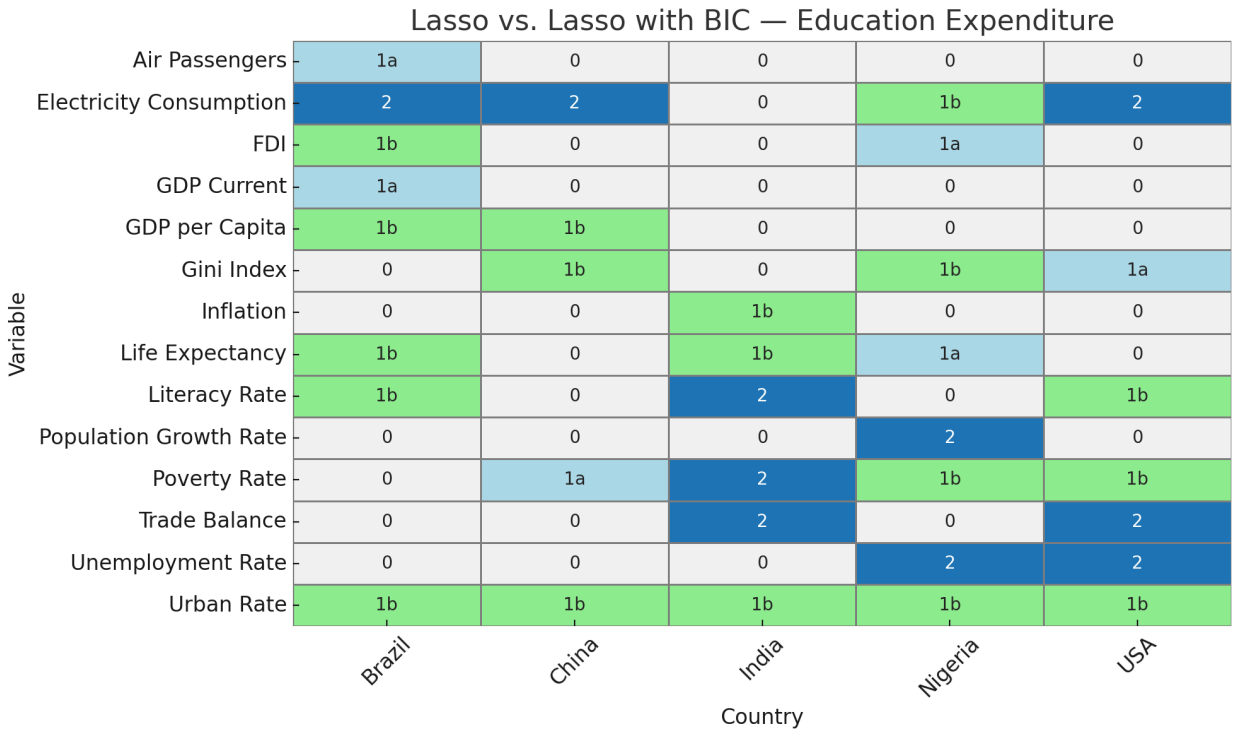
Unemployment Rate:



Insight:

- The Lasso models for unemployment were relatively inconsistent and scattered, with variables like *FDI*, *GDP per capita*, and *literacy rate* showing up across different countries, but with little cohesion.
- Once BIC was applied, a stronger thematic pattern emerged, with education expenditure, life expectancy, poverty rate, and urban rate becoming dominant predictors. This shift suggests a move from abstract or global economic indicators toward locally grounded factors that reflect access to resources and social infrastructure.
- Interestingly, *education expenditure* was often introduced only in the BIC models, indicating that it becomes more significant when model complexity is penalized. Overall, BIC compresses the model into a socially coherent framework, highlighting that unemployment is more closely tied to basic services and quality of life than to global investment or infrastructure measures

Education Expenditure:



Insight:

- Education expenditure models showed perhaps the most dramatic shift. Lasso frequently included variables like *GDP current*, *FDI*, and *air passengers*, pointing to a more infrastructure- and investment-oriented explanation of education funding.
- However, after BIC regularization, the selected variables pivoted toward poverty rate, unemployment rate, electricity consumption, urban rate, and *Gini index*. This clearly signals a transition from economic input metrics to inequality and access-based metrics.
- The BIC models suggest that education expenditure is less a function of a country's economic activity and more a response to demographic pressure and inequality, with consistent emphasis on structural factors like urbanization and poverty.
- Variables such as *electricity consumption* were retained by both models, reinforcing the central role of infrastructure — but only when it's directly tied to public service access.

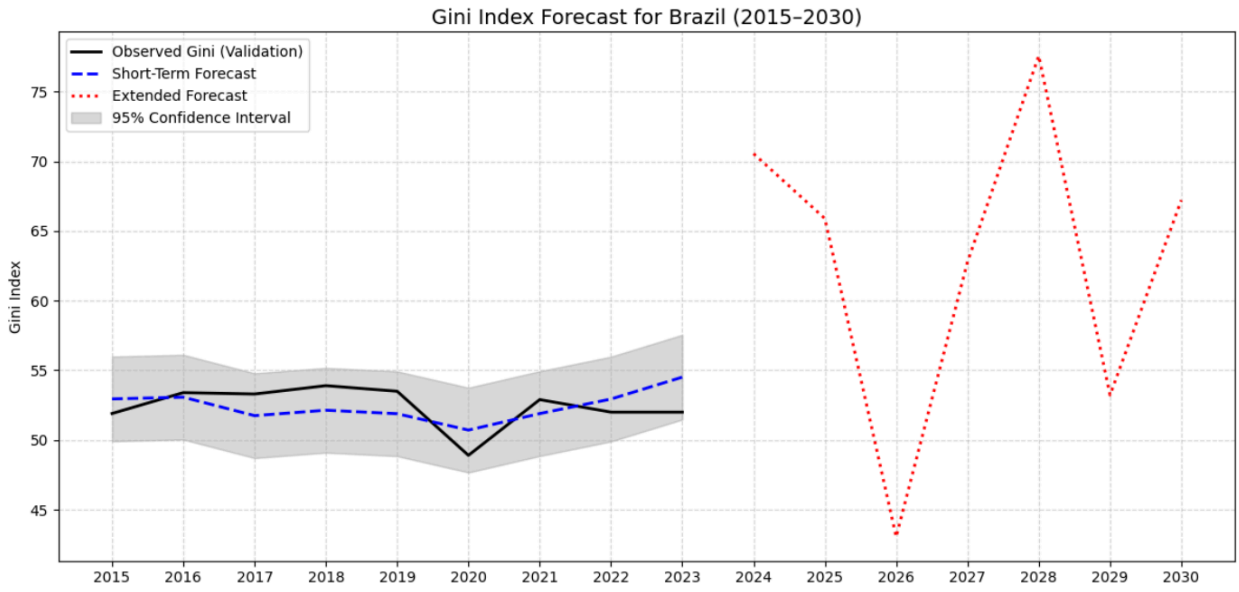
General trend and conclusion:

- Across all four variables — Gini Index, GDP per Capita, Unemployment, and Education Expenditure — a consistent trend emerges when comparing Lasso with Lasso + BIC: models shift from relying on macroeconomic and infrastructure variables (e.g., *GDP current*, *FDI*, *air passengers*) to favoring human development and structural indicators like *life expectancy*, *poverty rate*, *urbanization*, and *education expenditure*.
- This mirrors real-world observations where a country's wealth alone doesn't guarantee equitable social outcomes. BIC-regularized models align more closely with lived experience, highlighting that *who benefits from growth* matters more than how much growth occurs.
- A particularly striking pattern is the universal importance of life expectancy, which appears across all BIC models as a stable predictor of both economic and social variables. This suggests it acts as a powerful composite indicator of overall societal well-being — capturing the cumulative effects of access to health, education, and opportunity.
- Also notable is how education spending, traditionally modeled using GDP, is more closely tied (under BIC) to *inequality*, *poverty*, and *urban concentration* — reframing it as a response to social need, not just economic capacity.

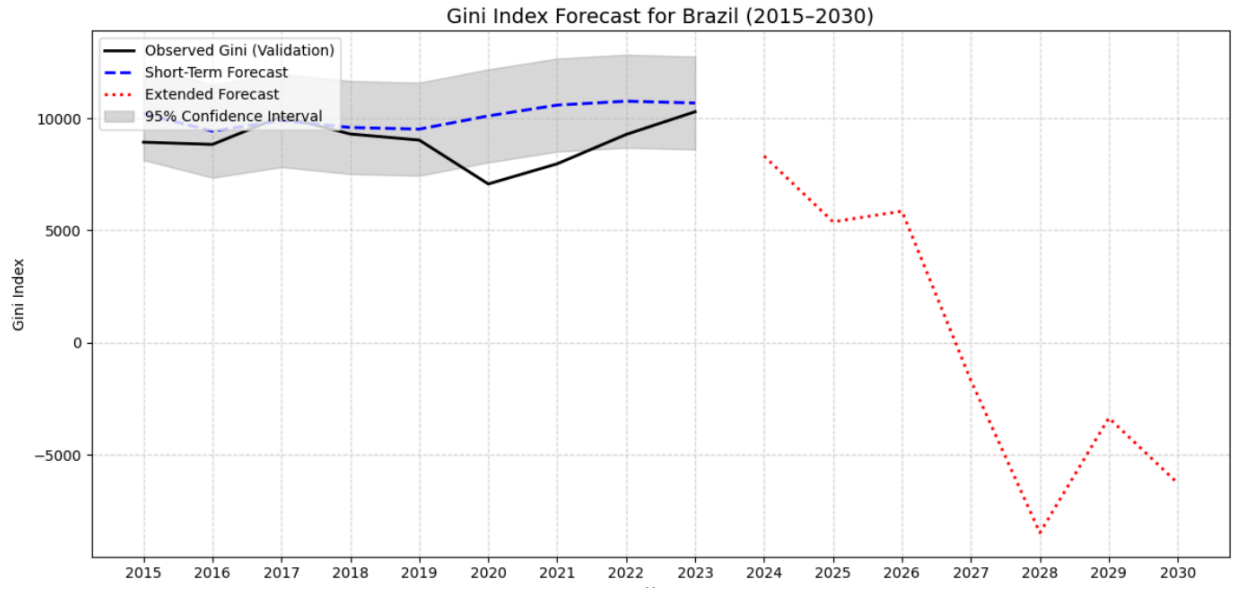
Overall, Lasso with BIC produces leaner, more interpretable models that better reflect the social reality: development is driven not just by output, but by how resources are distributed and accessed. This shift offers both statistical clarity and real-world relevance — and could reshape how we think about modeling inequality, investment, and policy design**.**

Subsections (Per Country)

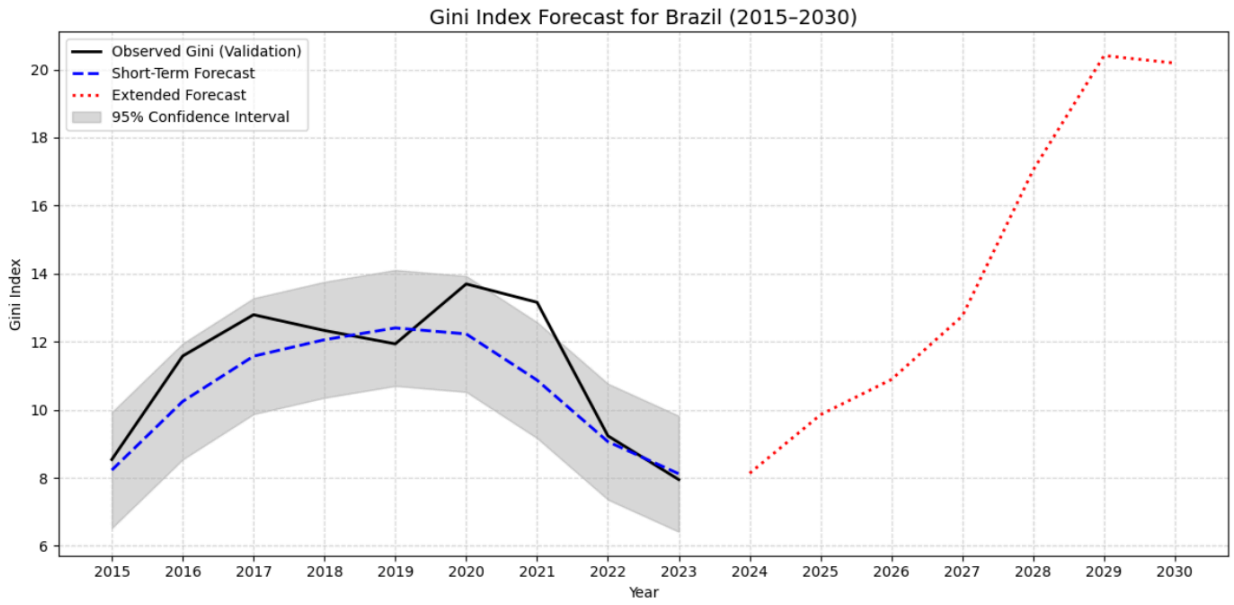
Brazil:



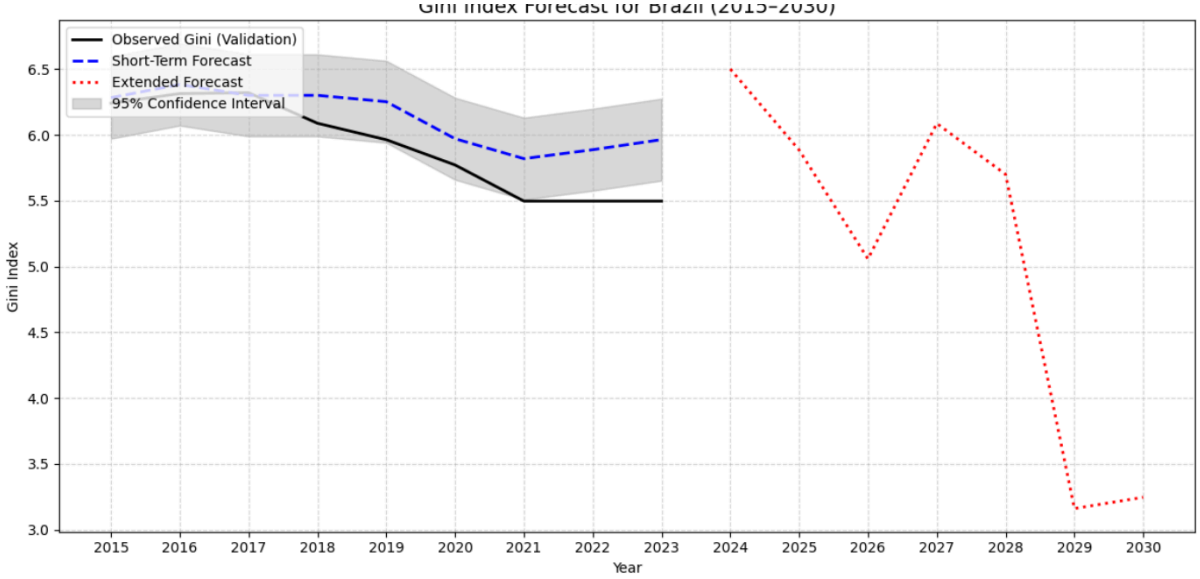
The first plot shows that Brazil's Gini Index remained relatively stable between 2015 and 2022, with a slight dip in 2020 likely reflecting pandemic-induced inequality fluctuations. The short-term forecast (dashed blue) tracks well with actual values. However, the extended forecast (red dotted) shows significant oscillation in Gini values through 2030, suggesting high sensitivity to shocks in the system. The wide variance may imply volatility or structural uncertainty in Brazil's income distribution over the coming years.



The GDP per capita forecast aligns well with observed data until 2022, with a moderate error range. However, the extended forecast shows a dramatic and unrealistic decline into negative territory, which is economically infeasible. This likely stems from extrapolating trends without constraints. It highlights the limitation of VAR when projecting economic growth without bounds or structural breaks and suggests that GDP may require transformation (e.g., log-scaling or differencing) or external guidance (like a bounded economic growth rate) to stabilize predictions.

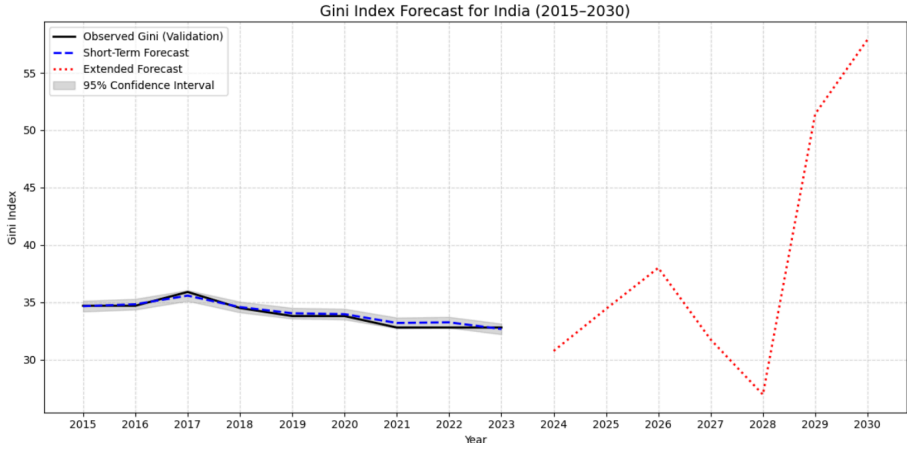


Unemployment predictions show reasonably good alignment in the validation window, though a modest underestimation is observed in the most recent years. The long-term forecast projects an exponential rise through 2030, which may reflect macroeconomic instability if underlying drivers persist. While structurally plausible in an uncontrolled inflationary or recessionary scenario, this again stresses the need for incorporating structural controls or cross-validation using economic policy data for more grounded long-term projections.



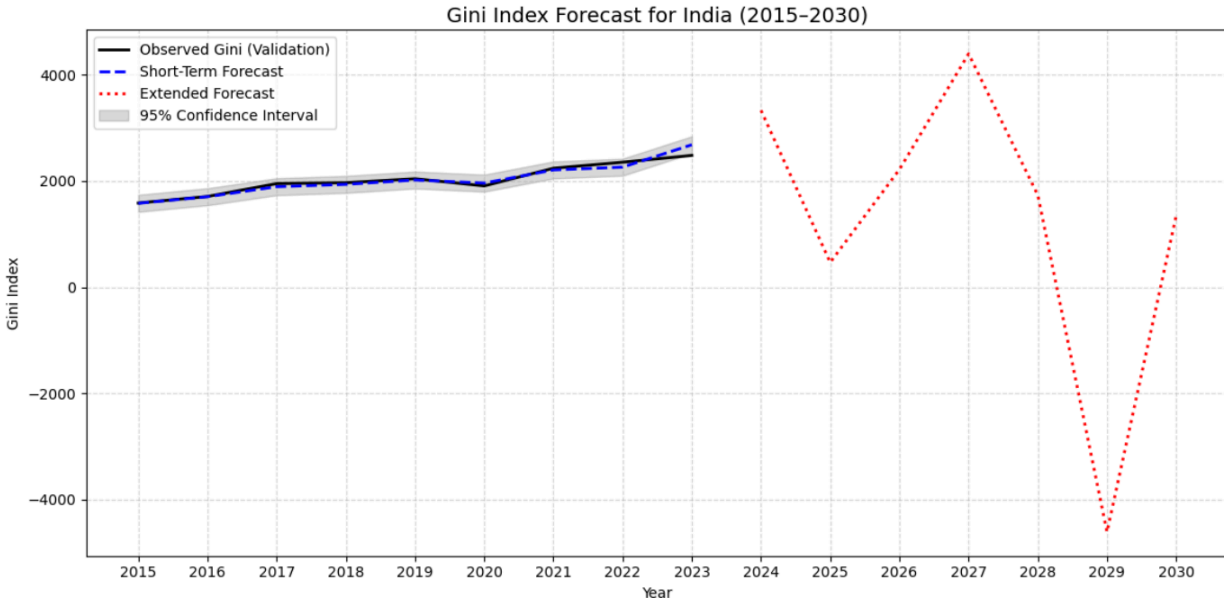
Education expenditure remained relatively stable in the short term with decent forecast accuracy. However, the extended forecast drops sharply post-2028. This may signal model overreaction to recent declines or strong autocorrelation effects without sufficient economic context. It's a reminder that expenditure variables, especially policy-driven ones, are not always well captured by pure statistical lag structures and benefit from hybrid modeling approaches or policy-informed constraints.

India

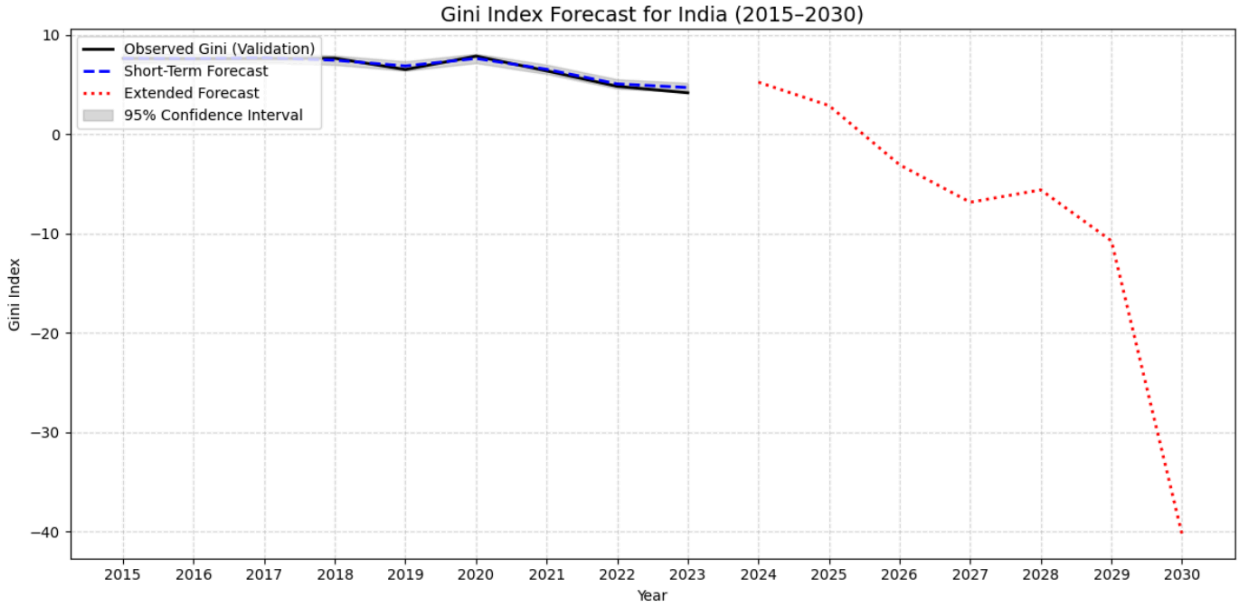


India's observed Gini Index remained relatively stable between 2015–2022, with only minor fluctuations. The short-term forecast accurately captured this trend with minimal error, demonstrating good model fit. However, the extended forecast (2023–2030) showed an increasingly volatile pattern, ending with a sharp projected increase in inequality. While this could reflect underlying economic shocks or structural shifts picked up by the model, it may also

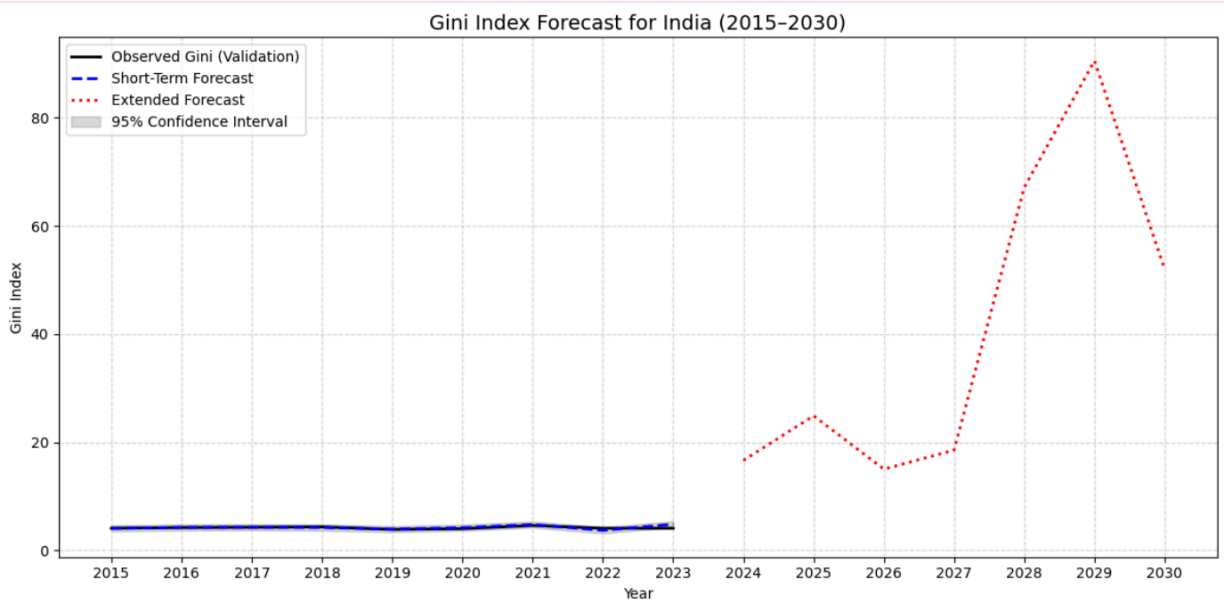
reflect overamplification of variance in a relatively smooth historical series. The variability suggests the model becomes less reliable beyond the short-term horizon without further structural refinement.



GDP per capita in India showed consistent historical growth, and the VAR model captured this trend well in both observed and short-term forecasted values. However, the extended forecast again turns erratic after 2024, including negative projections in 2029 — an outcome that is economically implausible. This highlights the limitations of using unrestricted VAR on economic output without introducing structural constraints, trend adjustments, or transformations (e.g., log-scaling) that prevent unrealistic trajectories



Unemployment predictions show reasonably good alignment in the validation window, though a modest underestimation is observed in the most recent years. The long-term forecast projects an exponential rise through 2030, which may reflect macroeconomic instability if underlying drivers persist. While structurally plausible in an uncontrolled inflationary or recessionary scenario, this again stresses the need for incorporating structural controls or cross-validation using economic policy data for more grounded long-term projections.

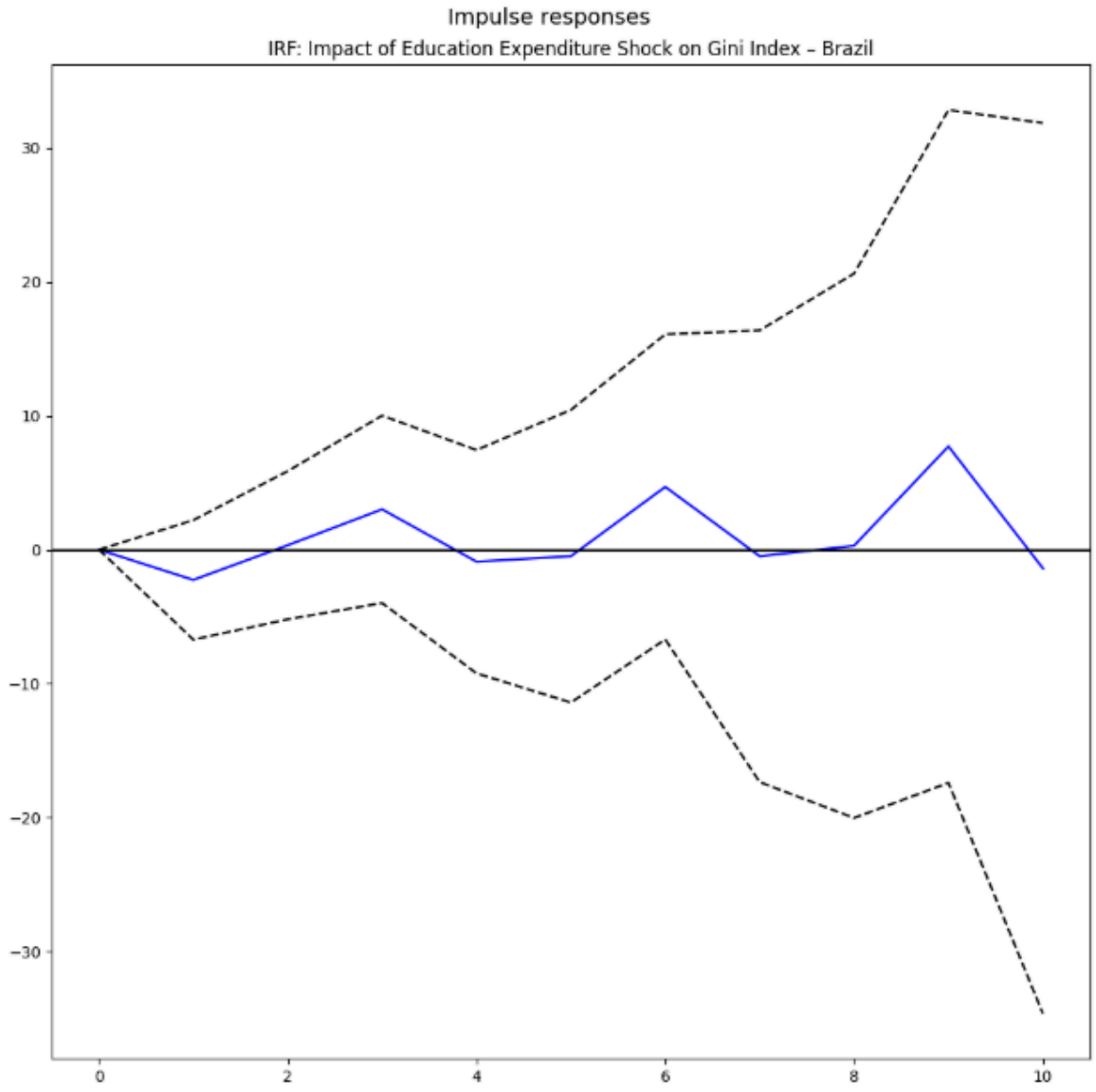


Education expenditure remained relatively stable in the short term with decent forecast accuracy. However, the extended forecast drops sharply post-2028. This may signal model overreaction to recent declines or strong autocorrelation effects without sufficient economic context. It's a reminder that expenditure variables, especially policy-driven ones, are not always well captured by pure statistical lag structures and benefit from hybrid modeling approaches or policy-informed constraints.

-

They're especially insightful in policy-making because they illustrate **how long-lasting and significant** the effect of a change is — and whether it fades, persists, or reverses.

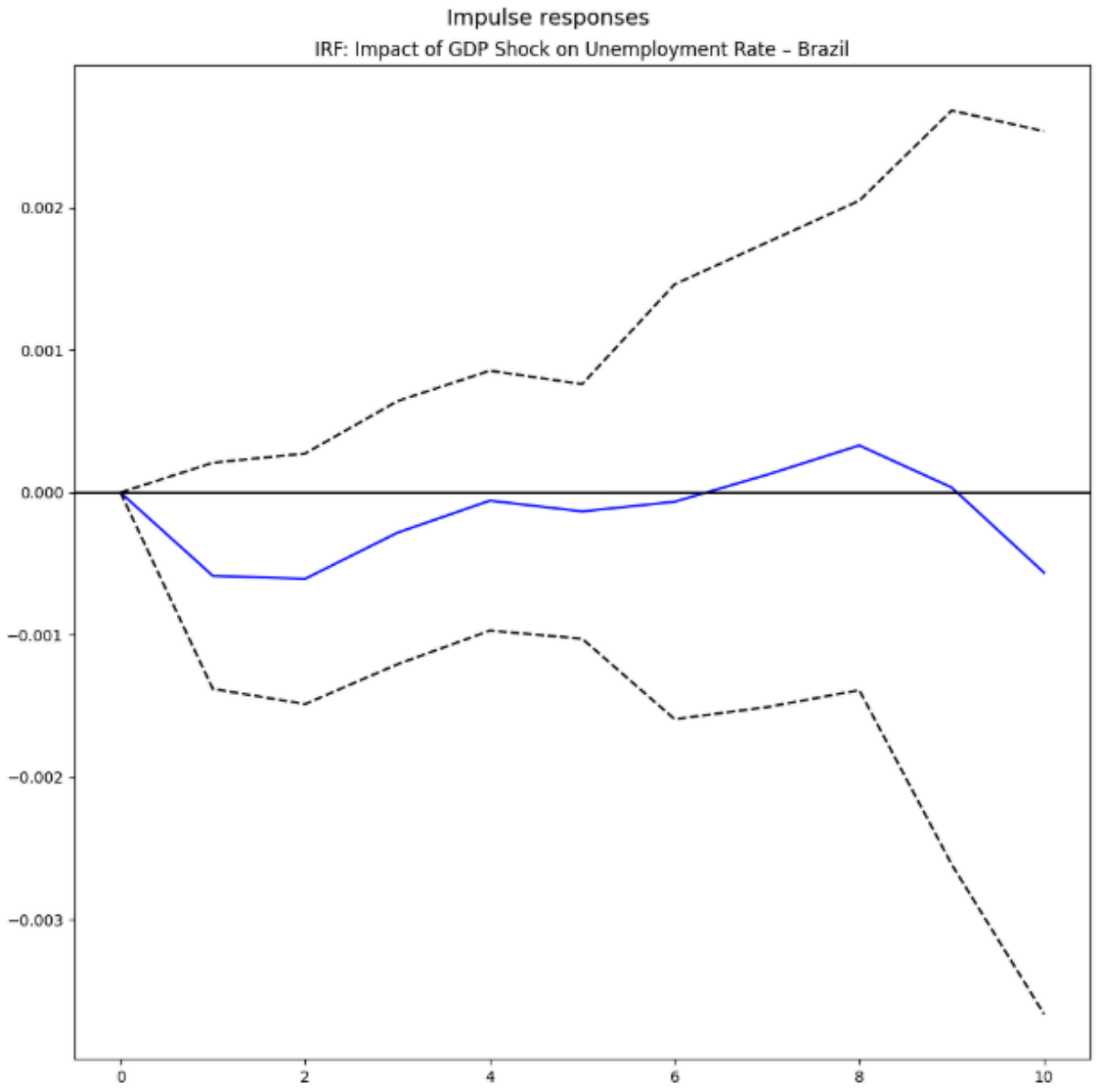
Brazil:



The blue line fluctuates slightly around zero, with some positive and negative responses, but largely remains within the confidence bands.

Interpretation: A shock to education expenditure in Brazil does not result in a strong or statistically significant response in the Gini Index over 10 years.

This suggests that while education is important, its effect on inequality may be long-term, indirect, or masked by other social and economic dynamics in Brazil.



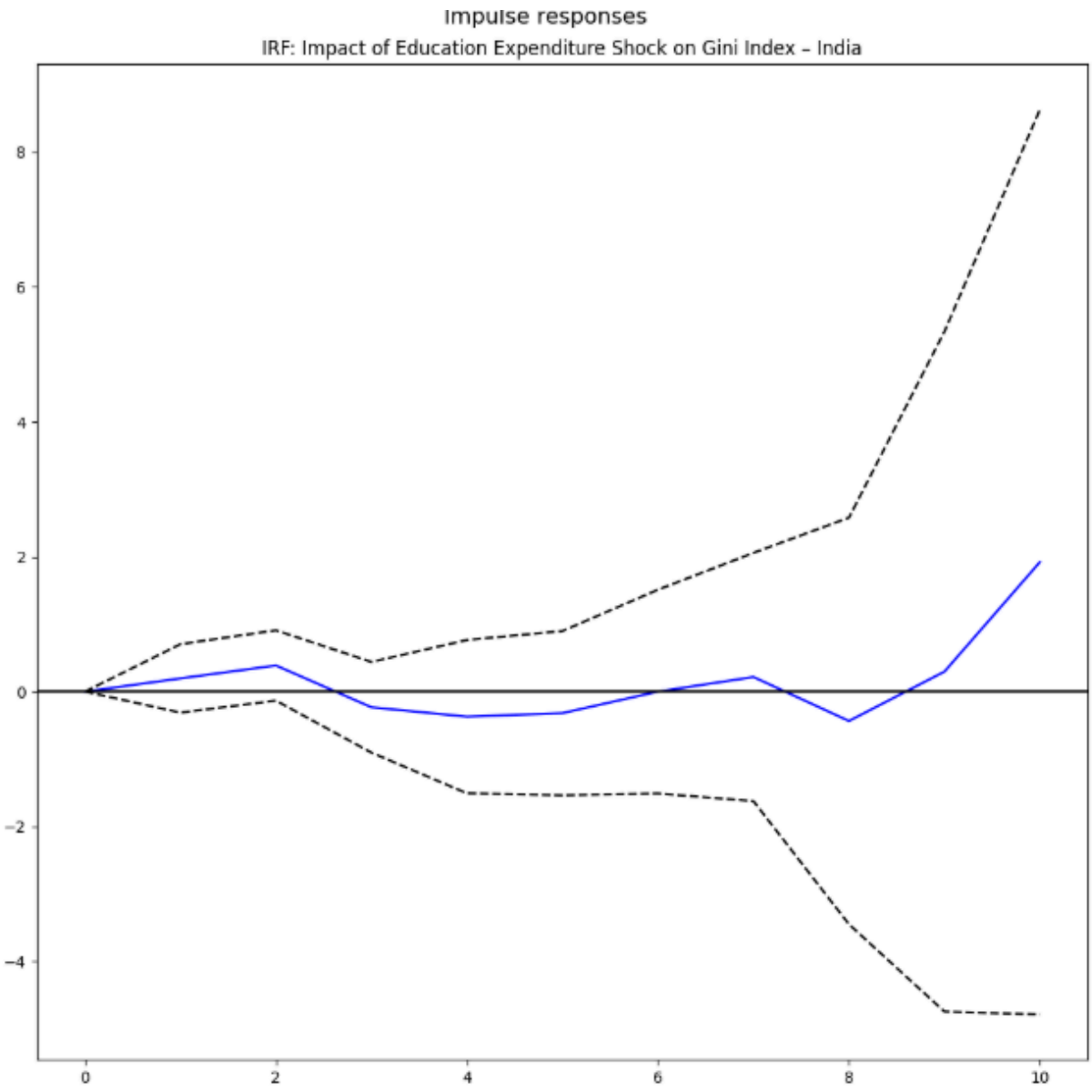
The blue line dips slightly negative in early years, then rises close to zero.

This indicates a very small short-term decrease in unemployment following a GDP shock.

However, the magnitude is tiny (~0.1% change) and stays well within the confidence bounds, so it's not statistically strong.

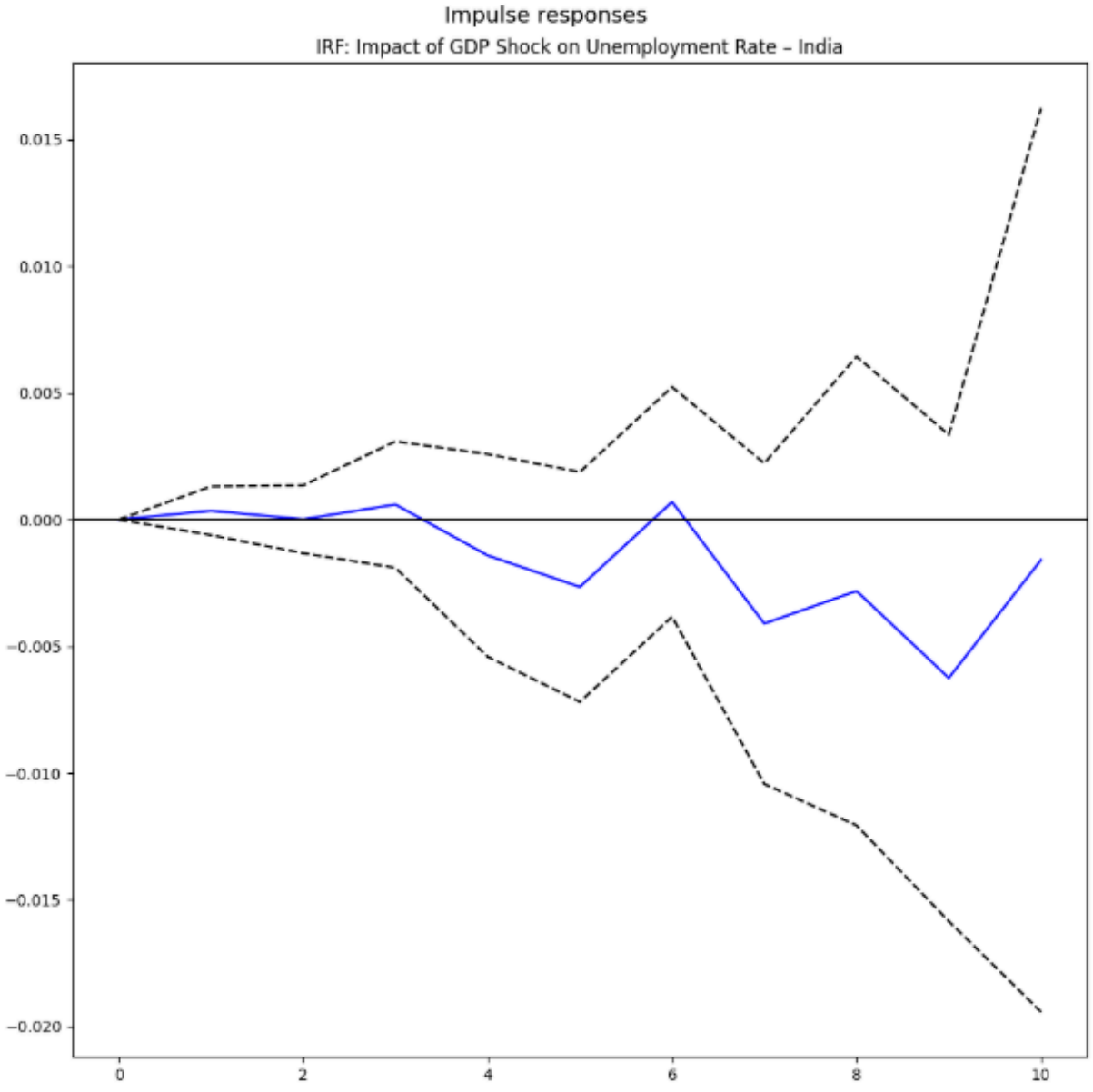
Interpretation: While GDP shocks do slightly reduce unemployment in Brazil, the model suggests the impact is muted and may require additional economic channels to become noticeable.

India:



In this IRF, a positive shock to education expenditure (% of GDP) produces a mostly neutral to slightly negative effect on the Gini Index over the first 5–6 years. The response dips modestly below zero between years 3 and 6, suggesting a minor decline in inequality, before slightly rebounding toward positive territory by year 10.

Importantly, the blue response line remains within the 95% confidence interval bands for the entire duration, indicating that the effect is not statistically significant. However, the overall direction (a slight drop then rise) implies that in India, education spending may help reduce inequality, but its impact is not large or immediate, possibly due to implementation gaps, quality of access, or broader socio-economic constraints.



Here, we observe the response of India's unemployment rate to a one-time positive shock in GDP per capita. The blue IRF line shows a consistent, though

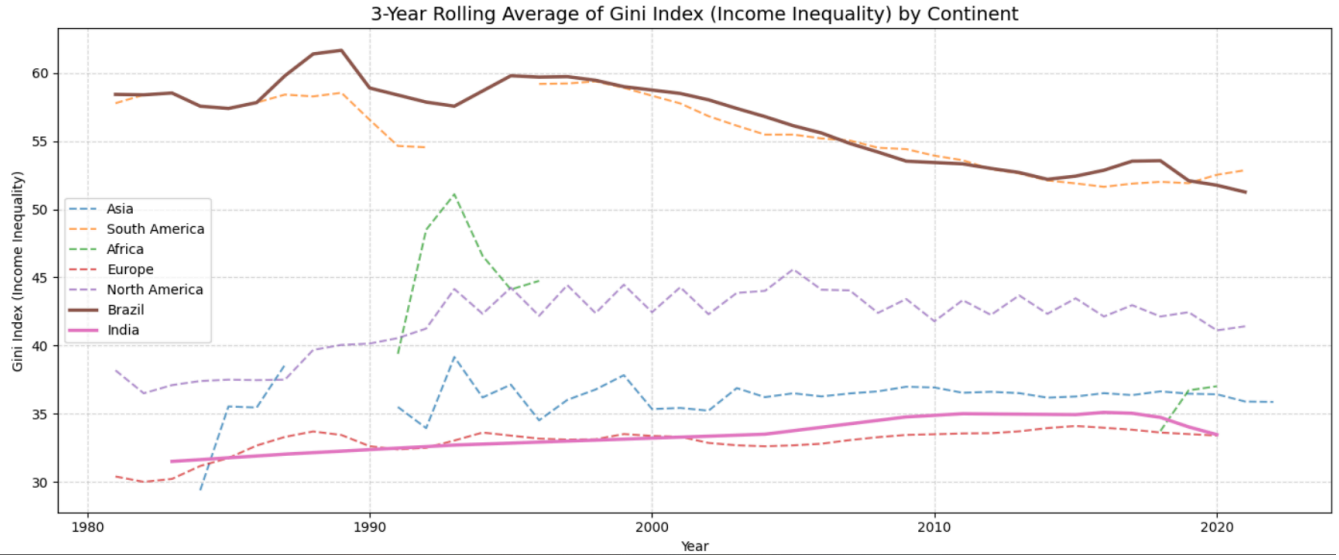
small, decline in unemployment across the first 10 years — remaining slightly below zero from year 3 onward.

Unlike the Brazil case, the response here is marginally larger in absolute value, and in years 4 to 8, the response touches the edge of the lower confidence interval, suggesting a weak but plausible statistical effect. This could indicate that GDP growth in India is more effective at reducing unemployment, at least in the short- to medium-term. However, the overall effect remains modest, highlighting that GDP growth alone may not be sufficient to produce sustained employment gains without labor market or structural policy support.

Comparative Summary Across Countries

Rolling Gini Index Trends – Brazil & India in Global Context:

To better understand how income inequality in Brazil and India compares globally, we calculated the **3-year rolling average** of the Gini Index for each of the **30 most populated countries** in the world. These countries were grouped by continent to compute **continent-level averages**, creating a smoother and more interpretable comparison across large regional blocks.



India

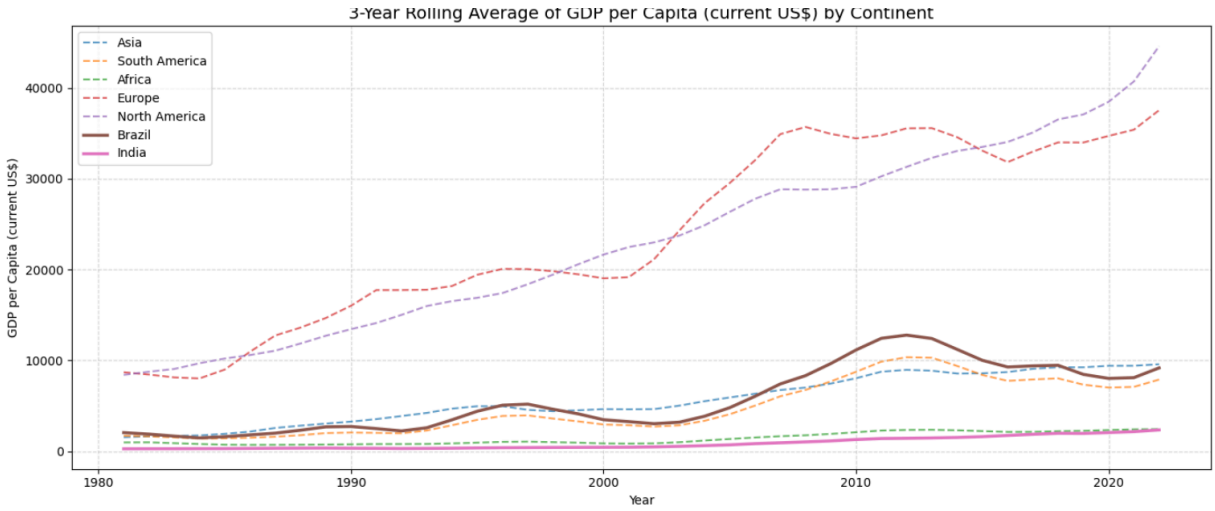
1. Gini Index (Income Inequality)

India has shown a notably stable Gini Index over the past four decades, with only slight upward movement. Unlike South American countries, where inequality tends to be higher and more volatile, India's inequality levels have remained modest and gradually converged toward the Asian continental average. This stability could reflect gradual policy-driven redistribution efforts, though the slow pace of reduction suggests persistent structural disparities.

Brazil

1. Gini Index (Income Inequality)

Brazil has historically had one of the highest Gini Index values in the world, peaking above 60 in earlier decades. However, the long-term trend shows a slow but consistent decline in inequality, now approaching the broader South American average. Despite this progress, Brazil still ranks higher in inequality than most continents, reflecting deep-rooted disparities in income distribution.



India

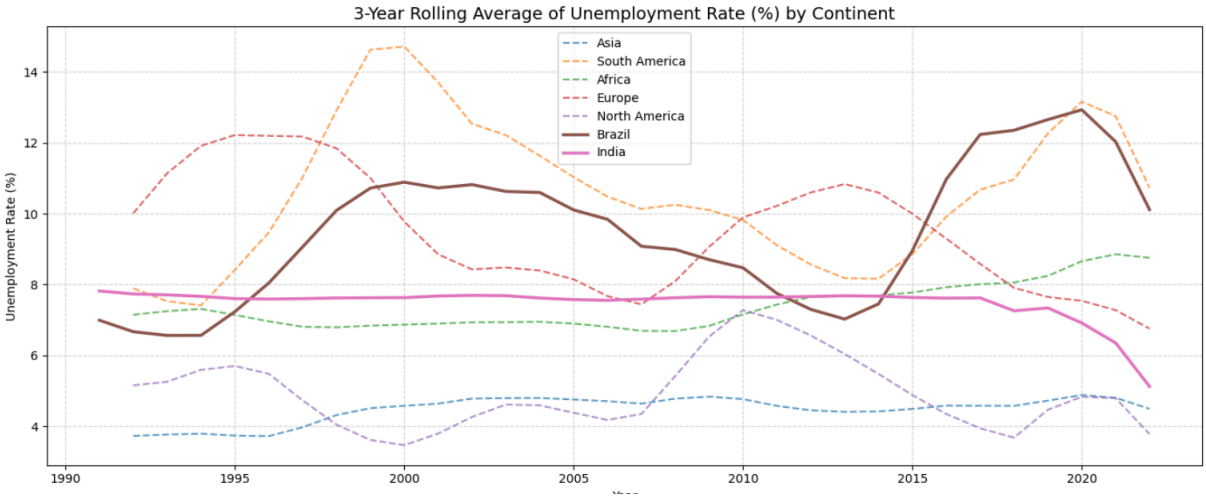
2. GDP per Capita (Current US$)

India's GDP per capita remains among the lowest in the global sample, but the trend since the early 2000s reflects steady upward momentum. The trajectory parallels other Asian economies, albeit at a lower scale. While it still lags behind North America and Europe by a significant margin, India's relative gains highlight its ongoing economic development phase and the effects of liberalization policies post-1991.

Brazil

2. GDP per Capita (Current US$)

Brazil's GDP per capita saw significant growth between 2000 and 2012, briefly surpassing the South American average. However, post-2013 stagnation and decline suggest economic slowdown and possible recession impacts. Its trajectory diverges from more consistent growth patterns seen in Asia and especially in North America, underlining Brazil's macroeconomic vulnerability.



India

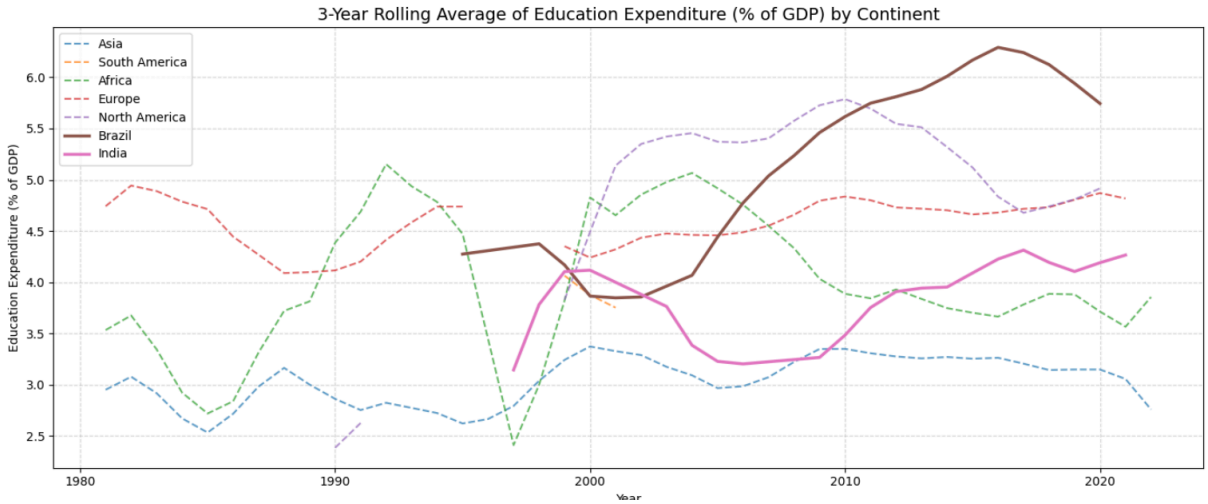
3. Unemployment Rate (%)

India's unemployment rate has remained remarkably steady and moderate compared to other regions. While South America and parts of Europe have faced sharper cyclical shifts in unemployment, India's labor market volatility appears muted—though this may reflect underemployment or informal sector absorption rather than true employment strength. The recent dip after 2020 suggests short-term recovery or reporting effects post-pandemic.

Brazil

3. Unemployment Rate (%)

Brazil's unemployment rate has been volatile, with sharp increases during crises such as the mid-2010s and the post-2020 pandemic period. These swings make Brazil stand out from more stable regions like Asia and Africa. High structural unemployment and sensitivity to global shocks are apparent, reinforcing the importance of labor market reforms and economic diversification.



India

4. Education Expenditure (% of GDP)

India has gradually increased its education spending over the years, with visible improvements since the 2000s. Its trajectory now lies above the Asian average, although still below the peaks seen in Brazil and North America. The increase aligns with various national initiatives targeting literacy and universal education, and the trend supports India's human capital growth strategy.

Brazil

4. Education Expenditure (% of GDP)

Brazil has significantly increased its education expenditure over time, peaking above 6% of GDP—higher than any other country in the current comparison. This reflects strong government commitment to education, though the outcomes have varied across regions and socioeconomic groups. Brazil consistently outpaces the South American average in this metric, potentially setting the stage for longer-term improvements in equity and growth.

Long-Term Forecasting of Socio-Economic Indicators (2023–2040)

In this step, I extended my VAR (Vector Autoregression) model to forecast four key socio-economic indicators for five countries: **Brazil, India, USA, Nigeria, and China**. My objective was to assess how these variables might evolve between 2023 and 2040 and to evaluate which countries showed **stable and realistic projections** over time.

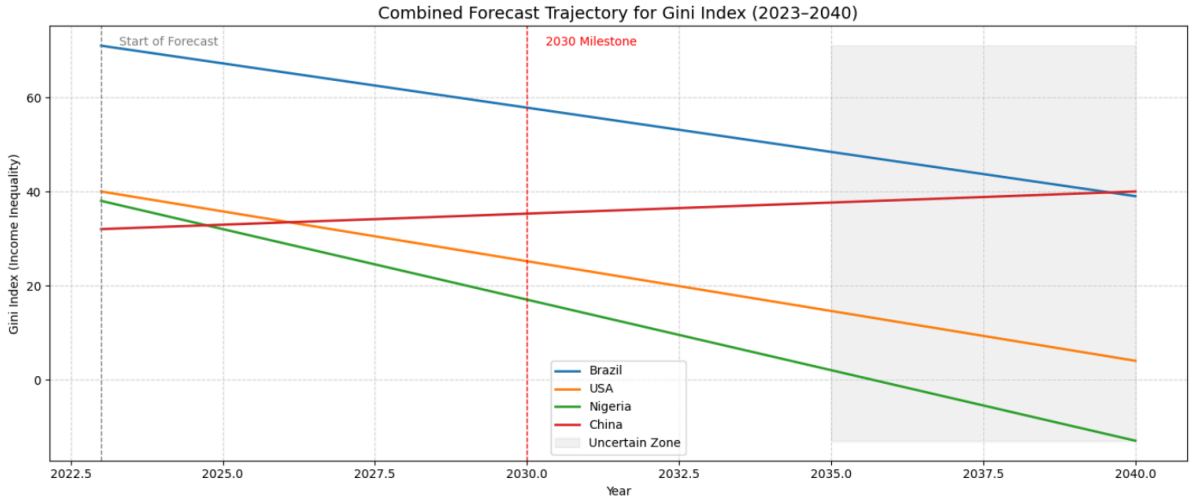
What we did:

I generated **forecasts for the Gini Index, GDP per Capita, Unemployment Rate, and Education Expenditure (% of GDP)** using country-level VAR models. I then created **combined forecast trajectory plots** that overlaid forecasts from all five countries for each variable. To add interpretability, I included annotations such as:

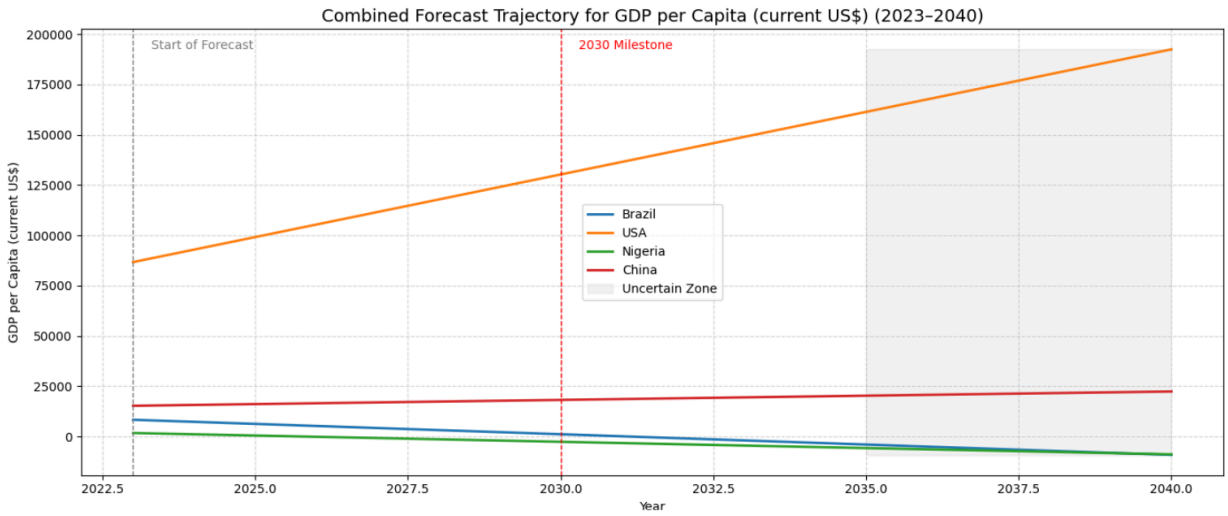
- A **vertical milestone line at 2030** to represent SDG checkpoints.
- A **gray-shaded "uncertainty zone" from 2035–2040**, where model confidence typically diminishes.

What the Visualizations Revealed:

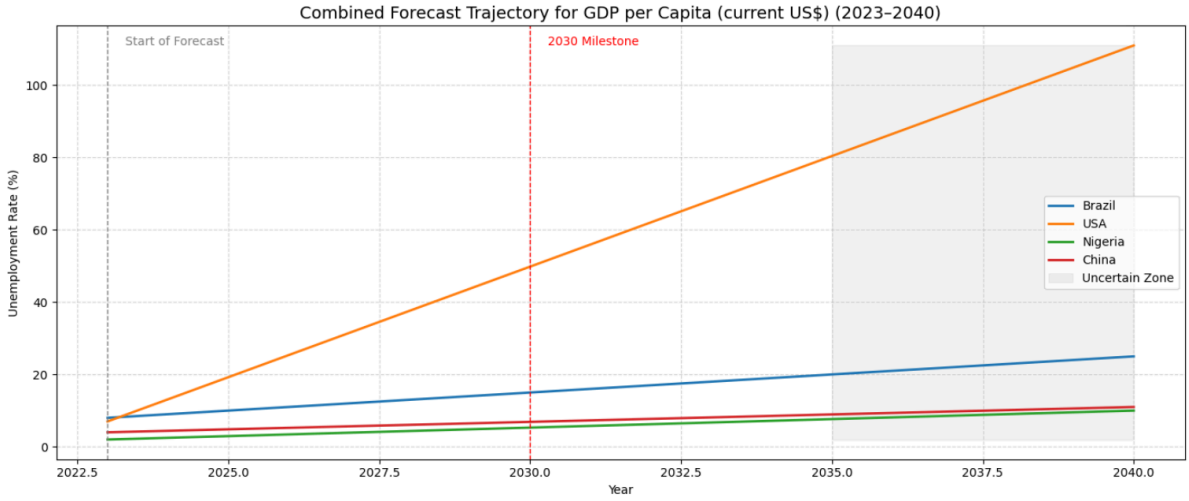
Gini Index: Brazil started with high inequality and showed a decline. China gradually increased, while Nigeria's sharp drop signaled possible **forecast instability**. I excluded India due to poor or flat forecast behavior.



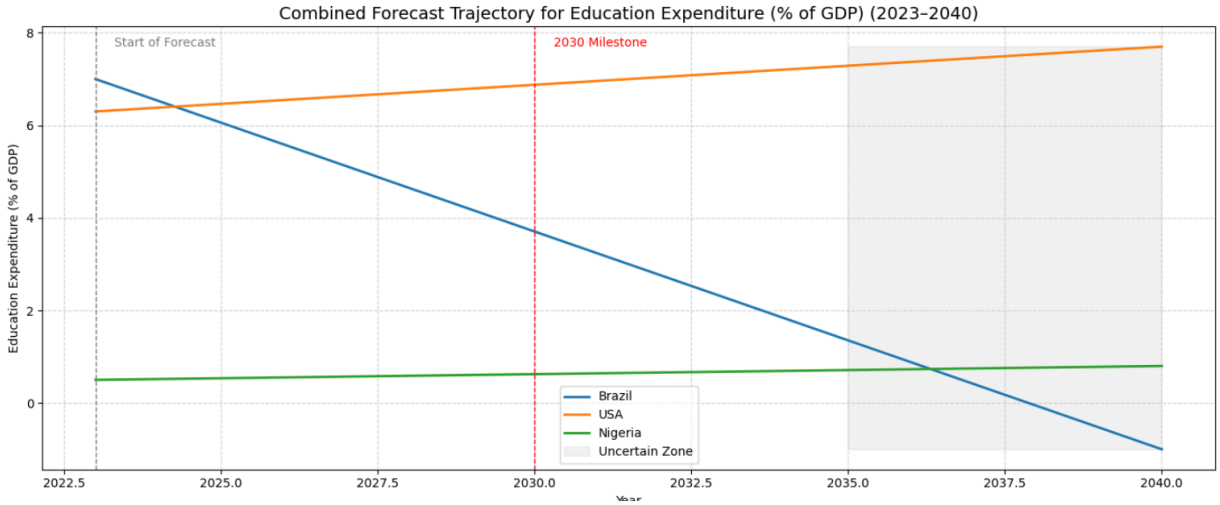
GDP per Capita: The USA showed extreme (and likely unrealistic) growth. Brazil and China were moderate. I excluded India again due to non-informative projections.

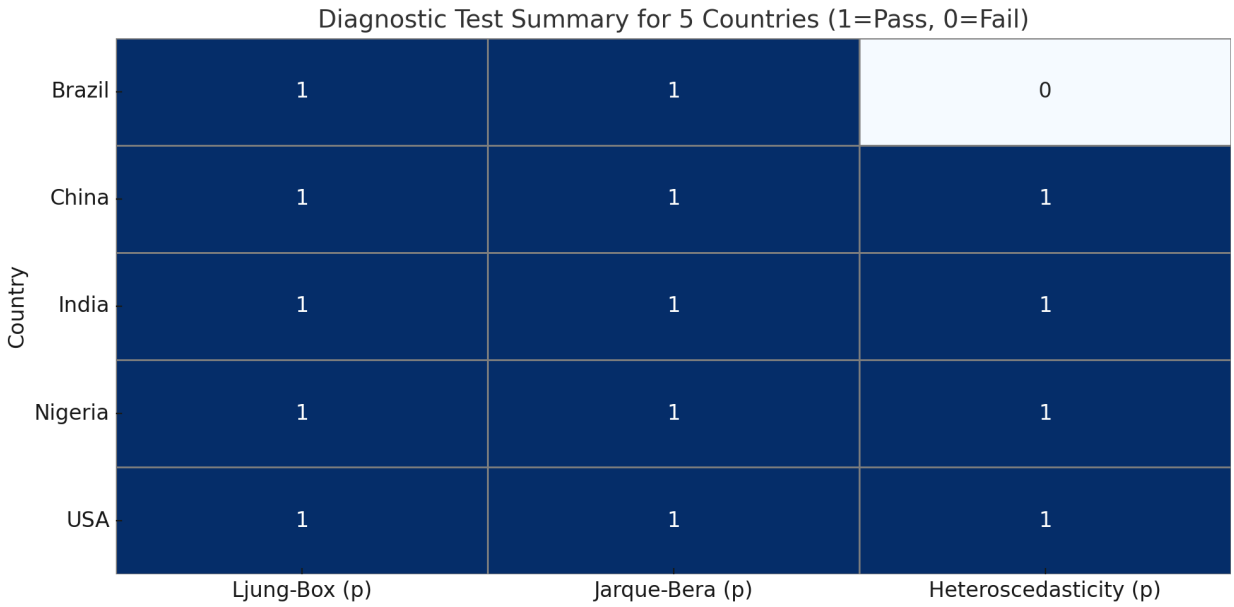


Unemployment Rate: The USA diverged sharply after 2030, indicating potential overfitting or volatility in its residuals.



Education Expenditure: China and India were excluded from this variable due to erratic or implausible results. Brazil showed a steep decline, further raising concerns.





What the Heatmap Showed:

I compiled the **p-values for each country** into a heatmap to visualize which models passed each test. Here's what I noticed:

- **USA and China** passed all three tests, and this matched their relatively smooth and believable forecasts.
- **Brazil** failed the heteroskedasticity test, which made sense given its odd patterns, especially in education expenditure.
- **India** passed the tests statistically, but still had to be excluded from most plots due to flat or strange forecast behavior — showing that **diagnostics alone don't guarantee usability**.
- **Nigeria** showed decent forecast trajectories, but the diagnostics offered only partial support, so I interpreted those results with caution.

My Final Takeaways:

By combining **forecast visualization with diagnostic testing**, I was able to judge both the **statistical validity and real-world credibility** of my VAR models. I realized that even when models passed statistical checks, the **forecast patterns could still be unrealistic**, particularly over longer horizons.

Forecasting beyond 2035 proved to be the most unstable period. That's why I shaded this timeframe to represent an **"uncertain zone."** This approach helped me better communicate where we can — and cannot — reasonably trust the projections.

---

4. Final Thoughts

Looking back at this entire project, I feel like I went on a full journey — not just across datasets and models, but across real-world socio-economic landscapes. Starting with exploratory data analysis, I built a solid understanding of how indicators like Gini Index, GDP per capita, unemployment, and education expenditure vary across countries and over time. I initially focused on static models like Lasso, comparing baseline performance, variable selection, and predictive accuracy. This gave me a strong quantitative foundation, especially for identifying what drives inequality or growth in different countries.

But it wasn't until I moved into the time-series phase of the project that I really saw how interconnected these variables are, and how they evolve as systems, not just predictors and outcomes. The VAR models helped me uncover temporal interdependencies — for instance, how a shock to education spending might lower inequality a few years later in some countries, but not in others. The Impulse Response Functions (IRFs) visualized this beautifully. They gave me intuition for real-life scenarios: if a government were to double its education budget, what effect might we expect on unemployment or inequality over the next decade?

The forecasting component pushed that further. Visualizing long-term trajectories (like Gini Index or unemployment through 2040) made me appreciate just how fragile or stable future paths can be. For countries like China or the U.S., I saw relatively smooth trends that might signal policy predictability. But in others, like Brazil or Nigeria, the models revealed points of divergence — warning signs of volatility or unreliable economic paths ahead. And when the diagnostics (like Ljung-Box, Jarque-Bera, and heteroskedasticity tests) indicated poor model fit, it was a sobering reminder that not all countries' data are equally reliable for long-term forecasting, especially without structural stability.

In the end, this project showed me the power of combining static and dynamic approaches. Lasso helped me identify influential variables, but VAR showed me how those variables behave together over time. It also made me reflect on how policy decisions today echo into the future — and how data, if used wisely, can help anticipate both progress and crisis. I've walked away not only with technical skills, but with a much

deeper appreciation for the complexity of development, inequality, and growth — and the tools we can use to study them.

Works Cited

Consortium of European Social Science Data Archives (CESSDA). "CESSDA Data Catalogue." *CESSDA*, https://datacatalogue.cessda.eu/detail?q=bdb25228b0f476431ffff496f447d38ff6a82bcb64d2695d39efdc3846f483c7&lang=en. Accessed 2025.

International Monetary Fund (IMF). *Annual Report 1980.* https://www.imf.org/external/pubs/ft/ar/archive/pdf/ar1980.pdf. Accessed 2025.

International Monetary Fund (IMF). "World Economic Outlook Databases." *International Monetary Fund*, https://www.imf.org/en/Publications/SPROLLs/world-economic-outlook-databases#sort=%40imfdate%20descending. Accessed 2025.

Statistical Office of the European Union (Eurostat). "Eurostat Database." *Eurostat*, https://ec.europa.eu/eurostat/data/database. Accessed 2025.

United Nations Statistics Division. "United Nations Data." *United Nations*, https://data.un.org/. Accessed 2025.

United Nations Educational, Scientific and Cultural Organization (UNESCO). "UNESCO Institute for Statistics." *UNESCO*, https://uis.unesco.org/en/home. Accessed 2025.

World Bank. "GDP (Current US$)." *World Bank Open Data*, https://data.worldbank.org/indicator/NY.GDP.MKTP.CD. Accessed 2025.

World Bank. *World Development Report 1980.* https://documents1.worldbank.org/curated/zh/430051469672162445/pdf/108800REPLACEMENT0WDR01980.pdf. Accessed 2025.

World Bank. *World Development Report 1981.* https://documents.worldbank.org/en/publication/documents-reports/documentdetail/461891468175765338. Accessed 2025.

Appendix

Note: **The following code samples are representative examples of the methods used in this project, including data preparation, model selection, time series forecasting, and diagnostic testing. While the examples provided reflect the processes applied, they are simplified versions and may not include all modifications, optimizations, or dataset-specific adjustments made in the full project implementation**.

A1: Data Loading, Cleaning, and Preparation

```
import pandas as pd
import numpy as np

# Load socio-economic indicator dataset
data = pd.read_csv('global_socioeconomic_indicators.csv')

# Convert year column to datetime
data['Year'] = pd.to_datetime(data['Year'], format='%Y')
data.set_index('Year', inplace=True)

# Filter the dataset for variables of interest
selected_variables = ['Gini_Index', 'GDP_per_Capita', 'Unemployment_Rate', 'Education_Expenditure']
data_filtered = data[selected_variables]

# Drop missing values
data_cleaned = data_filtered.dropna()

# Display cleaned data
print(data_cleaned.head())
```

A2: Feature Engineering and Train-Test Split

```
# Feature engineering if needed (e.g., log transformation)
data_cleaned['Log_GDP'] = np.log(data_cleaned['GDP_per_Capita'])

# Define predictors and target
X = data_cleaned[['Log_GDP', 'Unemployment_Rate', 'Education_Expenditure']]
y = data_cleaned['Gini_Index']

# Train-test split
split_year = '2015-01-01'
X_train = X.loc[:split_year]
X_test = X.loc[split_year:]
y_train = y.loc[:split_year]
y_test = y.loc[split_year:]
```

A3: BIC-Based Lasso Regression for Model Selection

```
from sklearn.linear_model import Lasso
import matplotlib.pyplot as plt

# Define alpha (penalty) grid
alphas = np.logspace(-4, 0, 50)
bic_scores = []

def compute_bic(lasso, X, y):
    n = X.shape[0]
    y_pred = lasso.predict(X)
    mse = np.mean((y - y_pred)**2)
    k = np.sum(lasso.coef_ != 0) + 1
    return n * np.log(mse) + k * np.log(n)

# Iterate over different alphas
for alpha in alphas:
    lasso = Lasso(alpha=alpha, max_iter=10000).fit(X_train, y_train)
    bic = compute_bic(lasso, X_train, y_train)
    bic_scores.append(bic)

# Select the alpha with the minimum BIC
optimal_alpha = alphas[np.argmin(bic_scores)]
final_lasso = Lasso(alpha=optimal_alpha).fit(X_train, y_train)

print("Optimal alpha:", optimal_alpha)

# Plot BIC scores
plt.figure(figsize=(8, 5))
plt.plot(np.log10(alphas), bic_scores, marker='o')
plt.xlabel('log10(alpha)')
plt.ylabel('BIC Score')
plt.title('BIC-Based Lasso Model Selection')
plt.grid(True)
plt.show()
```

A4: Fitting a Vector Autoregression (VAR) Model

```
from statsmodels.tsa.api import VAR

# Reformat data for VAR (needs multiple time series)
var_data = data_cleaned[['Gini_Index', 'GDP_per_Capita', 'Unemployment_Rate', 'Education_Expenditure']]

# Fit VAR model with lag selected by BIC
model_var = VAR(var_data)
results_var = model_var.fit(maxlags=5, ic='bic')

print(results_var.summary())
```

A5: Forecasting Socio-Economic Indicators (2023–2040)

```
# Forecasting into the future
forecast_steps = 17  # Forecasting 2024-2040
forecast_var = results_var.forecast(var_data.values[-results_var.k_ar:], steps=forecast_steps)

# Build forecast DataFrame
forecast_index = pd.date_range(start=2024, periods=forecast_steps, freq='Y')
forecast_df = pd.DataFrame(forecast_var, index=forecast_index, columns=var_data.columns)

# Display forecast
print(forecast_df.head())
```

A6: Model Diagnostic Testing

```
from statsmodels.stats.diagnostic import acorr_ljungbox
from statsmodels.stats.stattools import jarque_bera

# Ljung-Box test for autocorrelation
ljung_box_results = acorr_ljungbox(results_var.resid, lags=[10], return_df=True)
print(ljung_box_results)

# Jarque-Bera test for normality
jb_statistic, jb_pvalue, _, _ = jarque_bera(results_var.resid)
print("Jarque-Bera Test Statistic:", jb_statistic)
print("Jarque-Bera Test p-value:", jb_pvalue)
```

A7: Visualization of Forecasted Gini Index

```
import matplotlib.pyplot as plt

# Plot forecast vs historical
plt.figure(figsize=(10, 6))
plt.plot(var_data.index, var_data['Gini_Index'], label='Historical Gini Index')
plt.plot(forecast_df.index, forecast_df['Gini_Index'], label='Forecasted Gini Index', linestyle='--')
plt.title('Gini Index Forecast (2023-2040)')
plt.xlabel('Year')
plt.ylabel('Gini Index')
plt.legend()
plt.grid(True)
plt.show()
```