Drivers of Economic Development: A Regression Analysis of Health and Education Impacts on GDP

A study on the factors influencing GDP per capita in emerging economies through health and education indicators

Artem Khomytskyi -20221686 | Davyd Azarov - 20221688 | Timofii Kuzmenko - 20221690

Introduction

This project examines the main drivers of economic development by exploring how health and education outcomes relate to GDP per capita. Focusing on four large developing countries — Brazil, China, India, and South Africa — we use macroeconomic data from 2010 to 2020 to build and test a series of econometric models. These countries represent different regions and socioeconomic contexts but share similar challenges in improving living conditions. By assessing the role of human capital, we aim to understand whether public investment in health and education contributes significantly to economic performance.

Data

We collected data from the World Bank's World Development Indicators (WDI) for the years 2010–2020. The selected countries are Brazil, China, India, and South Africa. The variables include GDP per capita, population, health expenditure per capita, and primary school enrollment. After downloading the data, we cleaned and prepared it in R, applying transformations such as logarithms and standardization.

Variables Used in the Analysis

log_GDP_per_capita: Natural logarithm of GDP per capita (current US$), used to measure economic development.

log_Health_Expenditure: Natural logarithm of current health expenditure per capita (US$), reflects investment in healthcare.

log_Primary_Enrollment: Natural logarithm of primary school enrollment rate (% of relevant age group), indicates access to basic education.

log_Population: Natural logarithm of total population, used to account for country size and scale.

GDP_growth: Annual growth rate of GDP per capita (log difference), captures economic change over time.

HealthExp_growth: Annual growth rate of health expenditure per capita (log difference), shows change in healthcare investment.

Enrollment_growth: Annual growth rate of primary enrollment (log difference), reflects changes in education access.

Country: Categorical variable indicating the country: Brazil, India, China, South Africa.

Theoretical Background

The relationship between human capital and economic development has been widely discussed in the economic literature. One influential line of research suggests that investments in health and education play a central role in boosting productivity and long-term growth. Studies by Barro (1996) and Bloom & Canning (2000), for example, highlight how better health outcomes and increased schooling are associated with higher GDP levels. However, some findings also show that the effect of education or health spending can vary significantly depending on institutional quality, demographic structure, and policy implementation. While statistical models often confirm a positive link between human capital and economic performance, challenges remain in capturing all relevant country-specific factors and separating causality from correlation.

MODEL SPECIFICATION AND DIAGNOSIS

CORRELATIONS

When analyzing the correlations between log_GDP_per_capita, the dependent variable, and the independent variables studied, we observed that log_Health_expenditure_per_capita had the strongest positive correlation with the dependent variable. The variables log_Primary_education_enrollment and log_Population_total also showed moderate correlations. On the other hand, the growth rate variables (such as growth in population or enrollment) showed weaker correlations, typically below 0.3, suggesting they may have a smaller direct influence on GDP per capita in the short term.

MODEL ANALYSIS

Our regression model aimed to explain the variation in GDP per capita across four developing countries (Brazil, India, China, and South Africa) over the period 2010–2020. Since the original values of GDP per capita and other variables were highly skewed and difficult to interpret, we applied a logarithmic transformation to all key continuous variables. This allowed us to estimate a log-log model, where the coefficients can be interpreted as elasticities — that is, the percentage change in GDP per capita for a 1% change in each explanatory variable.

To avoid issues of multicollinearity, we computed the Variance Inflation Factor (VIF) for all explanatory variables. All VIF values were below the commonly accepted threshold of 10, indicating no severe multicollinearity.

We performed individual significance tests for each variable. The variables log_Health_expenditure_per_capita, log_Primary_education_enrollment, and log_Population_total were statistically significant at the 1% level and retained in the final model. Variables that were not statistically significant (such as growth rates or untransformed indicators) were excluded step by step to improve model performance.

We also tested a quadratic term for log_Population_total, to check for non-linear effects, but it was not statistically significant. Similarly, interaction terms between predictors were explored but did not improve the model's explanatory power.

Finally, we conducted key model diagnostics, including:

RESET test to assess specification errors

Breusch-Pagan test for heteroskedasticity

Durbin-Watson test for autocorrelation in residuals

All tests confirmed that our final model meets standard OLS assumptions.

Final Estimated Model

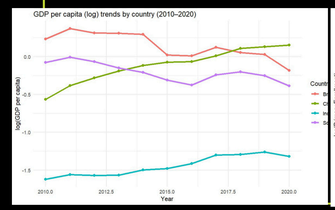
$$\log(\text{GDP\_per\_capita}) = 0.031 + 1.073 \cdot \log(\text{Health}) - 0.038 \cdot \log(\text{Enrollment}) + 0.118 \cdot \log(\text{Population})$$

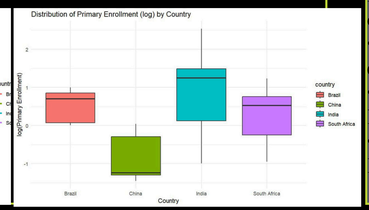
The model shows that higher public health expenditure is strongly associated with higher GDP per capita, while primary school enrollment has a small but statistically significant negative coefficient, possibly reflecting diminishing returns or data limitations. Population also has a positive effect, suggesting scale advantages in larger economies.

$R^2 = 0.9856$ — this means that 98.56% of the variance in log(GDP per capita) is explained by the variance in health, education, and population indicators across countries and over time.

Assumptions of the Final Model

1. Our final model is linear in parameters and assumes that the relationship between the dependent variable (log of GDP per capita) and the independent variables (log of health expenditure, log of school enrollment, and log of population) is correctly specified. The error term captures all other random influences not included in the model.
2. The data consist of panel observations from four countries over the period 2010–2020. Although not a strictly random sample in the statistical sense, we consider these observations representative for our comparative analysis of emerging economies.
3. We tested for multicollinearity by calculating the Variance Inflation Factors (VIFs) for all regressors. All values in model1 were well below the common threshold of 10 (log_Health: 2.32, log_Enroll: 1.40, log_Pop: 2.19), indicating that the explanatory variables are not perfectly correlated. Thus, the model satisfies the assumption of no perfect collinearity.
4. To evaluate whether the error term has zero conditional mean (E[u|X] = 0), we analyzed the residuals and confirmed that their sample average is extremely close to zero. Additionally, the RESET test for model1 did not reject the null hypothesis of correct functional form (p-value = 0.1911), providing further support for this assumption.
5. To test for heteroskedasticity, we applied both the Breusch-Pagan and White tests. The Breusch-Pagan test produced a p-value above 0.05, suggesting homoskedasticity, while the White special test was not significant (p = 0.7481), confirming that the variance of the errors does not systematically vary with the fitted values. Thus, model1 appears to satisfy the constant variance assumption. For robustness, we still report robust (HC1) standard errors.
6. We checked the normality of residuals using histograms and Q-Q plots. The histogram of residuals for model1, when plotted with probability density, closely matched the theoretical normal distribution curve. This suggests that the residuals are approximately normally distributed, supporting the validity of inference based on t- and F-statistics.
7. Lastly, we tested for autocorrelation using the Durbin-Watson test. The result for model1 (DW = 1.88, p = 0.35) indicates no significant autocorrelation, satisfying the sixth Gauss-Markov assumption in the context of time-ordered panel data.





All Tests

To assess potential misspecification in the regression models, the Ramsey RESET test was applied, examining whether adding nonlinear transformations of the fitted values improves the model.

model1 - RESET Statistic = 1.7291, df1 = 2, df2 = 38, p-value = 0.1911.
Conclusion - No evidence of misspecification

model2 - RESET Statistic = 1.4338, df1 = 2, df2 = 35, p-value = 0.2521.
Conclusion - No evidence of misspecification

model3 - RESET Statistic = 1.4338, df1 = 2, df2 = 35, p-value = 0.0195.
Conclusion - Evidence of misspecification, indicating omitted nonlinearities or incorrect functional form

Variance Inflation Factor (VIF)

To detect multicollinearity:

In Model 1, all VIF values were acceptable (below 5): log_Health (2.31), log_Enroll (1.40), and log_Pop (2.19).

In Model 2, VIFs were very low (around 1.00), showing no multicollinearity.

In Model 3, strong multicollinearity was detected: log_Pop had a VIF of 46.67, and the country variable had a GVIF of 6.19. Thus, care must be taken in interpreting coefficients in Model 3.

Breusch-Pagan Test

This test was used to check for heteroskedasticity:

Model 1: BP = 4.462, p-value = 0.2157 → no evidence of heteroskedasticity.

Model 2: BP = 0.627, p-value = 0.7309 → homoscedasticity assumed.

Model 3: BP = 14.138, p-value = 0.0281 → evidence of heteroskedasticity; error variance is not constant.

Durbin-Watson Test

This test checks for autocorrelation in residuals:

Model 1: DW = 1.879, p-value = 0.354 → no autocorrelation detected.

Model 2: DW = 1.369, p-value = 0.018 → positive autocorrelation likely.

Model 3: DW = 1.927, p-value = 0.407 → residuals appear to be independent.

Conclusion

This study allowed us to better understand how health and education indicators relate to GDP per capita in developing countries. Our final model confirms a strong link between health expenditure and economic performance, while the effects of education and population are less clear.

Still, not all relevant factors are captured — aspects like governance, institutions, or historical context remain outside the scope of this model. Therefore, while useful, the model cannot fully explain economic development on its own.

REFERENCES

- Barro, R. J. (1996). Determinants of Economic Growth: A Cross-Country Empirical Study (No. w5698). National Bureau of Economic Research. https://doi.org/10.3386/w5698
- Bloom, D. E., & Canning, D. (2000). The health and wealth of nations. Science, 287(5456), 1207–1209. https://doi.org/10.1126/science.287.5456.1207
- The World Bank. (2024). World Development Indicators. Retrieved from https://databank.worldbank.org/source/world-development-indicators
- Fox, J. (2023). An R Companion to Applied Regression (3rd ed.). SAGE Publications.
- Wooldridge, J. M. (2016). Introductory Econometrics: A Modern Approach (6th ed.). Cengage Learning.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67. https://doi.org/10.2307/1267351