POLI 381 Data Project – Data Analysis

Adam Cheng

1 Introduction

In this project, the goal is to address how economic performance (denoted X) influences public approval of national governments (denoted Y) over time. This research topic is important because findings can offer valuable insights to governments about the impact economic factors have on their survival, helping to develop more effective strategies that can generate mutually beneficial effects where governments prolong their lifespan (assuming that is their goal) and citizens enjoy stable governments.

My model expects that the relationship between X and Y is positive (or negative for countercyclical variables like unemployment rate) with more complex characteristics than a simple linear relationship, such as:

1. Non-linearity: Like the law of marginal diminishing returns, the relationship between X and Y should be initially positive until reaching a plateau point, where further economic growth produces diminishing returns on government approval as individuals substitute to other concerns responsible by the government.
2. The strength of correlation between X and Y (denoted by Spearman's ρ, discussed in detail later) should be dependent on countries':
    - GDP per capita level: The wealthier the country, the weaker the ρ because fluctuations in economic performance would be less impactful to them.

- Economic stability through time: The more stable the economy is across time, the weaker the ρ because there are minimal fluctuations (i.e. essentially constant), meaning changes in Y are likely influenced by other factors.

Although there are certainly more confounders to control, only two will be discussed due to the project's scope and limited word count.

2 Data and Measurement

2.1 Variables and Theoretical Conjecture

The coverage of the dataset is extensive, with 110 countries between 1990-2023. Therefore, to reduce noise, the variables will be carefully structured to examine the correlation between X and Y as follows:

Independent variables (measures X):

1. `gdp_pc_growth`: % change of `gdp_pc` from the previous year.
2. `unemployment_rate`: Unemployment rate (% of labor force).
    - Source: International Monetary Fund (2025).
3. `cpi_growth`: % change of Consumer Price Index from the previous year.
    - Source: International Monetary Fund (2025).

…, where implementing numerous independent variables can better capture multiple dimensions of X and determine the dimension most impactful on Y.

Dependent variable (measures Y):

1. `approval_smoothed`: Approval rating of national government smoothed via exponential smoothing (% of survey respondents).
    - Source: Carlin et al. (2023).

All variables are measured annually and nationally, and transformations of these variables only occur in the visualizations and tables of this paper, which will be discussed in detail when presented.

Control variables (compares differences in ρ):

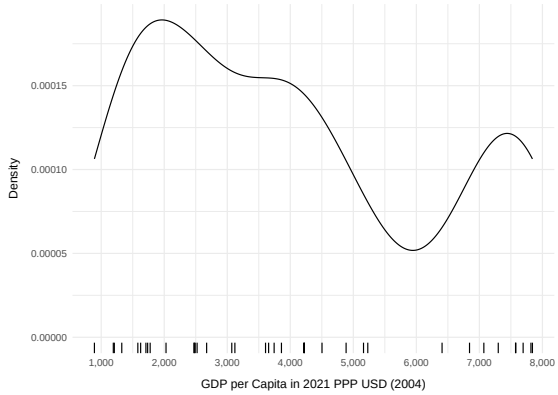
1. `year`: Calendar year, controlled by reducing the range to 2004-2013, setting 2008 as a cutoff point, and producing two strata (2004-2007 & 2008-2013).
2. `gdp_pc`: GDP per capita in 2021 PPP USD, controlled by classifying countries based on tertiles of its 2004 values (i.e. bottom, middle, and top 33% of the data).
    - Source: The World Bank (2025).

ρ will be calculated and visualized separately by `year` and `gdp_pc` strata to compare values between times of relative economic stability (2004-2007) and economic instability (2008-2013: global financial crisis) and three ordinal income levels of countries at a fixed year respectively.

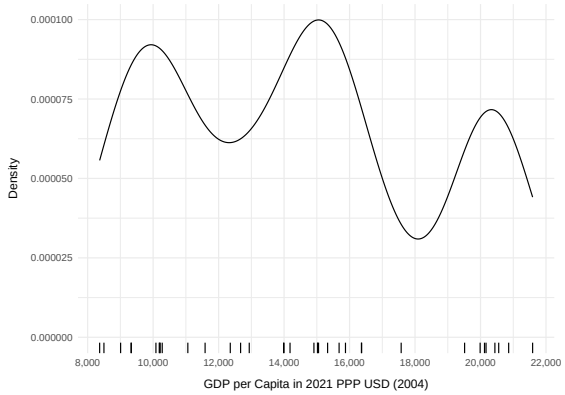
2.2 Examining Variation

Before assessing correlations, it is vital to verify if control variables `gdp_pc` and `year` exhibit the necessary variation for a meaningful analysis.

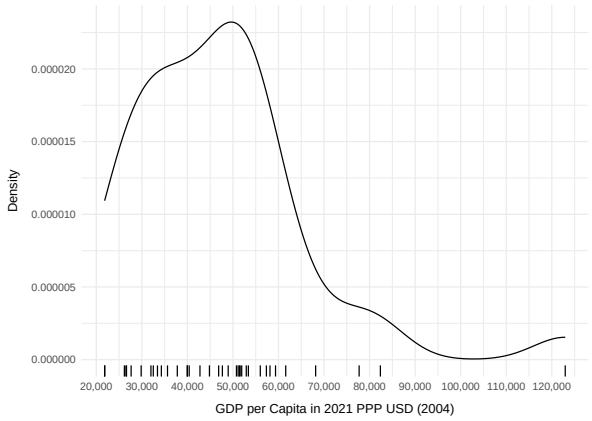
Figure 1 uses kernel density estimation and rug plots to examine the distribution of 2004 `gdp_pc` values stratified by their tertiles. The x-axis and its tick marks indicate 2004 `gdp_pc` values and the y-axis indicates the proportion of total data shared.



(a) KDE plot for the bottom 33% of `gdp_pc` values in 2004 (bandwidth = 650)



(b) KDE plot for the middle 33% of `gdp_pc` values in 2004 (bandwidth = 1100)



(c) KDE plot for the top 33% of `gdp_pc` values in 2004 (bandwidth = 7000)

Figure 1: Kernel density estimation (KDE) plots for `gdp_pc` values in 2004 separated by their tertiles

Each subplot of Figure 1 indicates each stratum sufficiently isolated a reasonable range of `gdp_pc` between levels while providing a balanced number of observations, adequate variation, and no redundancy in values between each stratum. These characteristics suggest `gdp_pc` and its 2004 base year can meaningfully contribute to the analysis.

Figure 2 utilizes line plots to illustrate the variation of variables across time and `gdp_pc` levels, highlighting differences in economic stability before and after the 2008 cutoff point and trends that may suggest potential correlations. Since variables are measured

by country and year, and the `gdp_pc` strata contain multiple countries, each variable is transformed to its mean value by its corresponding strata each year, plotting one simple but intuitive line within each facet.

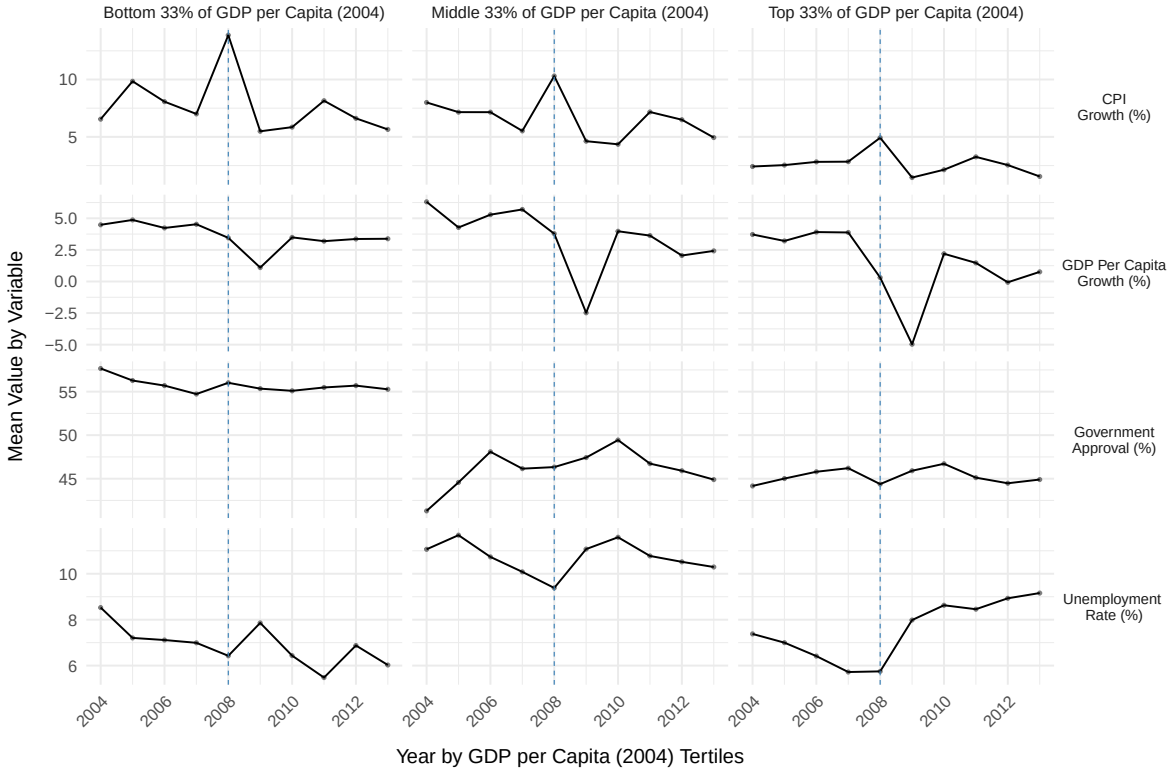


Figure 2: Line plots for mean values of variables (y-axis) for countries between 2004-2013 (x-axis) grouped by variable (row) and tertiles of 2004 `gdp_pc` values (column). The vertical blue line indicates the 2008 financial crisis cutoff point

By comparing fluctuations of independent variables before and after the cutoff point for each subplot, it is evident that 2008 is a strong separator that can control for periods of economic stability and instability. However, it is difficult to ignore the lack of fluctuations in `approval_smoothed` before or after 2008, suggesting the correlation between X and Y may be weak. Overall, the plot indicates trends between `approval_smoothed` and `unemployment_rate` appear the most correlated. Furthermore, it is vital to note that variables are transformed and do not visualize the true correlation until scatterplots or

ρ values are produced. The goal of these figures is to show that both control variables can meaningfully contribute to the analysis.

3 Descriptive Analysis

To first address my research question, Spearman's rank correlation coefficients (ρ) are computed in Table 1 separately for each control and independent variable to compare changes in correlation (denoted $\Delta\rho$) and determine which variable has the strongest correlation with `approval_smoothed` respectively. Given the context of the research question, ρ is chosen as the correlation metric because it is non-parametric and rank-based, meaning ρ does not assume variables are normally distributed (beneficial to my right-skewed economic variables) and is more suited to capture the assumed non-linear monotonic relationship between X and Y.

Table 1: Differences in Spearman's rank correlation coefficeint (ρ) separated by independent variables, cutoff point (2008 financial crisis), and tertiles of 2004 `gdp_pc` values. ρ ranges from -1 to 1, where values near 1 or -1 indicate a strong positive or negative relationship, and values near 0 suggest little to no association

(a) ρ for gdp_pc_growth and approval_smoothed

GDP Per Capita Tertile (2004)	ρ (2004-2007)	ρ (2008-2013)	$\Delta\rho$
Bottom 33%	-0.35	0.08	0.43
Middle 33%	-0.13	0.10	0.23
Top 33%	0.16	0.18	0.02
Total (Not grouped by tertiles)	-0.08	0.22	0.30

(b) ρ for unemployment_rate and approval_smoothed

GDP Per Capita Tertile (2004)	ρ (2004-2007)	ρ (2008-2013)	$\Delta\rho$
Bottom 33%	0.22	0.00	-0.22
Middle 33%	0.06	-0.38	-0.44
Top 33%	-0.24	-0.39	-0.15
Total (Not grouped by tertiles)	-0.03	-0.34	-0.31

(c) ρ for cpi_growth and approval_smoothed

GDP Per Capita Tertile (2004)	ρ (2004-2007)	ρ (2008-2013)	$\Delta\rho$
Bottom 33%	0.01	-0.18	-0.19
Middle 33%	-0.08	0.04	0.12
Top 33%	0.06	-0.02	-0.08
Total (Not grouped by tertiles)	0.10	0.10	0.00

By examining the effects of independent variables, it is evident that `unemployment_rate` has the strongest correlation with `approval_smoothed`. However, all ρ values suggest a weak to moderate relationship between X and Y, meaning findings do not strongly support my theoretical conjecture. Furthermore, ρ values with `cpi_growth` as the inde-

pendent variable are too weak to generate conclusive findings. However, differences in ρ between control variables offer some crucial insights.

When comparing differences in ρ before and after 2008, the relationship between X and Y overall strengthened (i.e., $\Delta\rho$ noticeably increased/decreased), suggesting ρ is likely dependent on economic stability. However, differences in ρ between `gdp_pc` levels are not as conclusive because they vary between all independent and control variables with no evident but still, non-random patterns, suggesting further conditioning before concluding ρ is independent of `gdp_pc` levels.

Since `unemployment_rate` exhibits the strongest monotonic correlation with `approval_smoothed` post-2008, a LOESS regression plot is produced given those conditions to visualize the correlation.

Figure 3: LOESS regression for `unemployment_rate` and `approval_smoothed` values between 2008-2013 separated by tertiles of 2004 `gdp_pc` values (bandwidth = 0.7)

As shown by Figure 3, the LOESS regression captures the expected moderate and negative relationship from Table 1b, with the top 33% `gdp_pc` level displaying the most consistent pattern. Although the plot shows signs of a plateau point at `unemployment_rate` values lower than ~8%, the data is too noisy for the evidence to be conclusive.

In conclusion, Table 1 and Figure 3 suggest current data is still noisy and lacks strong evidence to support the argument and assumptions, demanding further data conditioning.

4 Interpretation and Conclusion

From all the findings, it is evident that the model requires improvement but still produced insights that were consistent with my expectations, such as `unemployment_rate` exhibiting the strongest correlation with `approval_smoothed` amongst the other two independent variables, which is reasonable considering unemployment effects are likely the most damaging and immediate to individuals' livelihood. Though the correlation strength between X and Y is likely dependent on countries' economic stability across time, it remains inconclusive for `gdp_pc` levels. Lastly, the relationship between X and Y seems non-linear and exhibits plateaus, but with more complexity as results vary often depending on the control and independent variables used.

Ultimately, the unavoidable caveats to this analysis reside in the size of the dataset, meaning there are countless variables to consider and condition for. Therefore, for future reference, improvements to the model can be made in numerous ways:

- Reducing scope: Concentrating on countries similar in many categories to reduce variation from unobserved variables.
- Improving precision: Shortening the time gap between observations can increase precision as annual gaps may be too large to fully capture public opinion, considering they are often very elastic and suffer from recency bias.

- More variables: Adding more relevant control, independent, and dependent variables can improve consistency in results.

References

Carlin, R. E., Hartlyn, J., Hellwig, T., Love, G. J., Martínez-Gallardo, C., Singer, M. M., … Sert, H. (2023). Executive Approval Database 3.0. Retrieved from https://executiveapproval.org/

International Monetary Fund. (2025). International Financial Statistics (IFS). Retrieved from https://data.imf.org/?sk=4c514d48-b6ba-49ed-8ab9-52b0c1a0179b

The World Bank. (2025). World Development Indicators. Retrieved from https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators