Method Section

Heyang Liu

March 2025

This research's benchmark method is a cross-country least squares regression of our panel data. The specification of this regression will be as follows:

$$y_i = \alpha_1 \left(\frac{rem}{GDP}\right)_i + \alpha_2 \left(\frac{rem}{GDP}\right)_i^2 + \beta_1 school\_13_i + \beta_2 school\_58_i + \beta_3 school\_58_i^2 + \epsilon_i.$$

In the equation above, y_i is our income inequality indicator, represented by two measures: the standardized Gini coefficient (gini_std), and the difference between the income share of the top 20 percent earners and the bottom 20 percent earners (top20-bottom20). Our variable of interest, which is the share of international remittance in GDP (Remittance_as_percent), is denoted as $\left(\frac{rem}{GDP}\right)_i$. According to previous theoretical and empirical literature on remittances and inequality, an inverted U-shaped relationship is expected. Therefore, we include both the linear and quadratic terms of remittances in the regression model. Following this assumption, we expect a positive sign for the linear coefficient α_1, and a negative sign for the quadratic coefficient α_2.

The main control variables are education levels, The control variable in this regression is education level, which is composed of two different variables. Specifically, $school\_13_i$ corresponds to school expectancy for ISCED levels 1 to 3, while $school\_58_i$ represents school expectancy for ISCED levels 5 to 8 (for details on ISCED levels, see the data section). We use these two variables to capture both the general level of educational attainment in a country and the proportion of the population receiving higher education and to examine how they influence the relationship we are studying.

As for the expected signs of the two education-related control variables, we expect a negative sign for $school\_13_i$ because it measures access to basic education, and greater educational inclusiveness should lead to lower income inequality. For $school\_58_i$, we anticipate a non-linear relationship with inequality. At lower levels of higher education attainment, access is often limited to wealthier or more privileged groups, potentially widening inequality. However, as higher education becomes more broadly accessible across different segments of society, its negative effect on inequality should decrease. So we expect a non-linear relationship between the access to high-level education and the income inequality in the country.

Since our data is based on a country-year panel, endogeneity issues are inevitable in the regression analysis. A conventional approach to mitigate this problem is to include country and year-fixed effects. However, due to the incompleteness of the dataset, many countries or years have only one or very few observations, which makes the estimation of fixed effects less reliable and potentially unstable.

To address this issue, grouped regressions are performed. Sourcing from the data obtained from the International Monetary Fund, 12 different economic indicators are selected to form the basis for classification. These indicators include real GDP, GDP per capita, total investment, gross national savings, inflation, unemployment rate, general government revenue, general government total expenditure, general government net lending/borrowing, general government net debt, and current account balance. All these indicators are either presented as indices, expressed in purchasing power parity (PPP) terms, or shown as a percentage of GDP, making them comparable across countries and over time.

The grouping is conducted using the k-means clustering algorithm, an unsupervised learning method that partitions observations into k clusters based on similarity, aiming to minimize within-cluster variance. This allows us to group countries with similar macroeconomic profiles, as measured by the selected indicators. By running regressions within these clusters, we can partially mitigate endogeneity concerns, as comparisons are made among countries with comparable structural characteristics. During the analysis, we gradually increase the number of clusters to identify an appropriate cutoff point. This is determined by observing when additional clusters lead to only marginal improvements in model fit, known as the 'elbow point.' The chosen number of clusters strikes a balance between capturing sufficient heterogeneity across groups and maintaining enough observations within each cluster to ensure model stability and interpretability of the results.