Global Economy and Development at BROOKINGS

How much do we really know about inequality within countries around the world? Adjusting Gini coefficients for missing top incomes

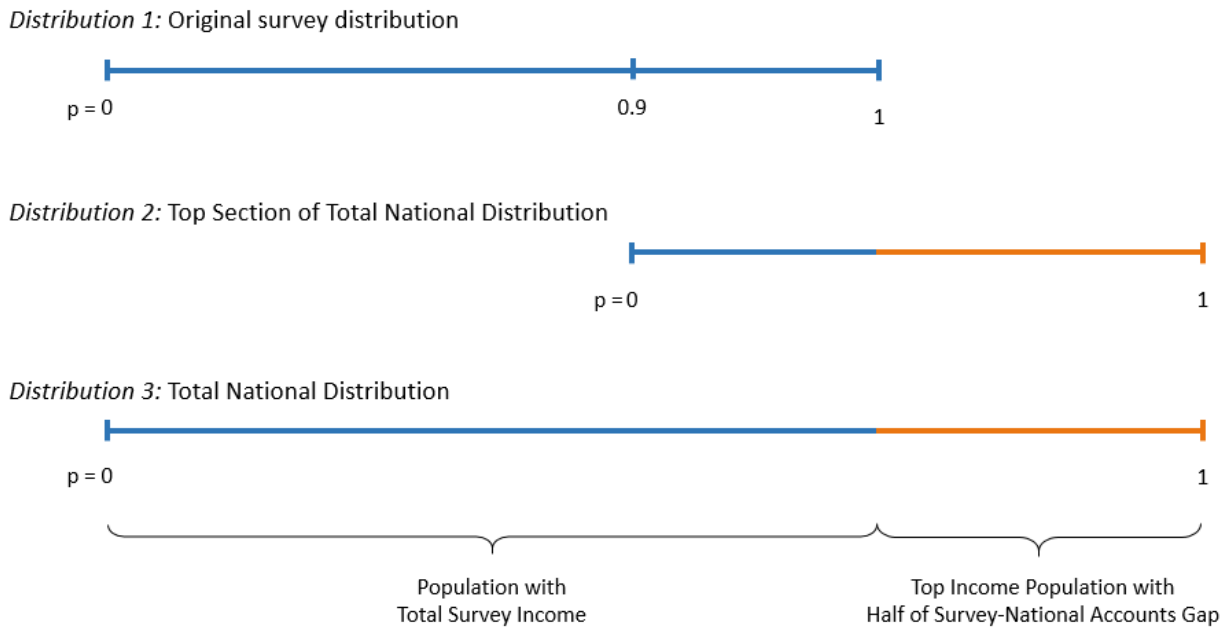
By Laurence Chandy and Brina Seidel

Appendix: Allocating Income from Survey-National Accounts Gap to Top Incomes According to Pareto Distribution

Overview

This appendix describes the method we use to adjust the income distribution captured by a household survey to account for missing top incomes. We attribute half of the survey-national accounts gap to missing top incomes, and we assume that the incomes of those at the top of the adjusted distribution follow a Pareto distribution. The stylized relationship between the population in the original survey distribution and the population in our adjusted national distribution is illustrated in Figure A1.

Figure A1: Relationship between population captured by survey and total population including missing top incomes



We use the top decile of the population captured by the survey to determine the value of the Pareto parameter α that characterizes the top section of the total adjusted national distribution. In other words, we estimate α based on the total income earned by the population in the blue section of Distribution 2, and use the α value to calculate distributional values for the orange section. In our final adjusted national distribution, or Distribution 3, we obtain values for the blue section by rescaling values from Distribution 1 and we obtain values for the orange section by rescaling our calculated values from the orange section of Distribution 2.

This means that our method preserves the shape of the original survey data for all individuals except the missing top incomes, and appends Pareto-imputed values for the missing top incomes. Our specific method for determining the value of α also enables us to determine the size of the population that makes up the missing top incomes endogenously, as a function of the size of the survey-national accounts gap.

Description of Methodology

We begin with the formula for $L(y)$, where L is the percent of total national income owned by individuals with incomes at or below y, k is the minimum income, and α is the Pareto parameter.

$$L(y) = 1 - \left(\frac{k}{y}\right)^{\alpha-1}$$

This comes from substituting the Pareto CDF formula into the Pareto Lorenz curve formula, which are respectively:

$p(y) = 1 - \left(\frac{k}{y}\right)^{\alpha}$ = percent of the population with an income at or below y

$L(p) = 1 - (1-p)^{1-\frac{1}{\alpha}}$ = percent of total national income owned by the bottom p percent of the population.

Solving $L(y)$ for α yields the following equation:

$$\alpha = 1 + \frac{\log(1-L)}{\log\left(\frac{k}{y}\right)}$$

This formula can be used to calculate the value of α that characterizes the top of the national income distribution, or Distribution 2 in Appendix Figure 1. We define Distribution 2 as the top decile captured by the survey plus the top income individuals that are missing from the survey, to whom we attribute half of the national accounts income that is missing from the survey. The minimum income k in this distribution is the mean income of individuals at the 90th percentile of the survey (that is, the bottom of the top decile).

Let y^* denote the maximum income recorded in the survey, and let L^* denote the income in the top decile of the survey as a share of the total in Distribution 2. We calculate L^* in terms of the mean survey income $\bar{y}_{svy}$, the mean national accounts income $\bar{y}_{na}$, and the share of income in Distribution 1 (survey income) attributable to the top decile of the survey population, $l_{0.1}$.

$$L^* = \frac{l_{0.1}\bar{y}_{svy}}{l_{0.1}\,\bar{y}_{svy} + \frac{1}{2}(\bar{y}_{na} - \bar{y}_{svy})}$$

Because we attribute the shortfall in survey income to missing top incomes, $L(y^*) = L^*$ in Distribution 2. We use this relationship to solve for α^*, the α value that characterizes the entire top section of the income distribution.

$$\alpha^* = 1 + \frac{\log(1-L^*)}{\log\left(\frac{k}{y^*}\right)}$$

We use the calculated value of α^* to calculate the percent of the Distribution 2 population that was captured by the survey, p^*.

$$p^* = p(y^*) = 1 - \left(\frac{k}{y^*}\right)^{\alpha^*}$$

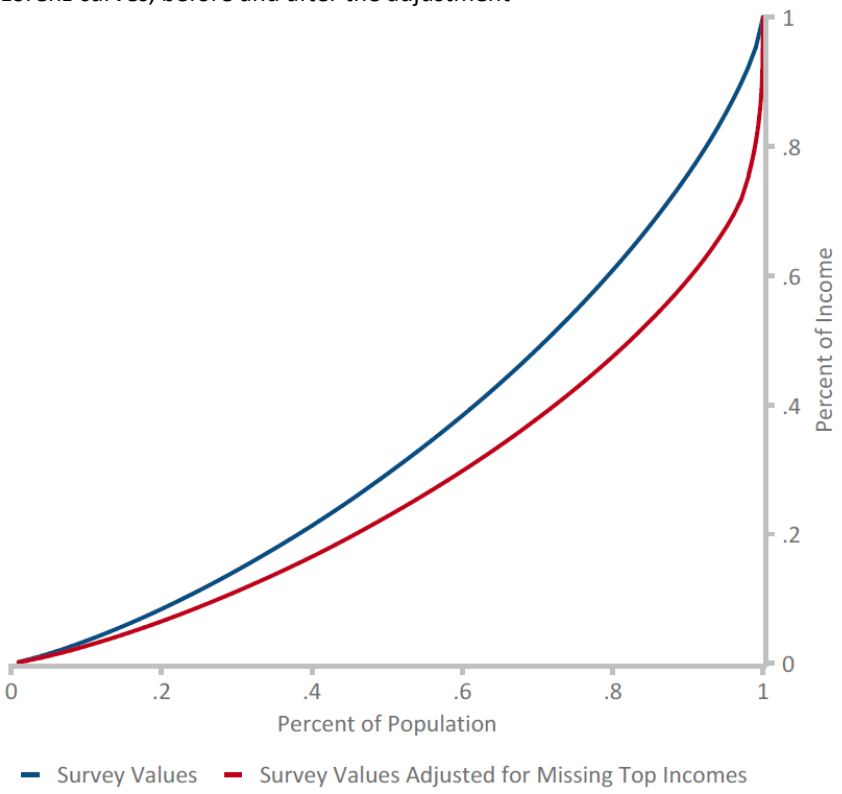
For the remaining population $p^* < p \leq 1$ that was not captured by the survey – that is, the missing top incomes – we use the formula for the Lorenz curve $L(p) = 1 - (1-p)^{1-\frac{1}{\alpha}}$ to calculate the share of income belonging to each centile of the population according to a Pareto distribution.

We then transform the p and L values values from the original survey, Distribution 1, and the p and L for the missing top income population from the top section of the distribution, Distribution 2, to create a single set of adjusted p and L values for the total national distribution, Distribution 3. We re-scale the original survey p values by a factor of $\frac{1}{1+\frac{1}{10}(1-p^*)}$, and we rescale the L values by a factor of $\frac{2\bar{y}_{svy}}{\bar{y}_{svy}+\bar{y}_{na}}$. We transform the top income p values by $\frac{1-\frac{1}{10}p^*}{1+\frac{1}{10}(1-p^*)} + p(\frac{\frac{1}{10}}{1+\frac{1}{10}(1-p^*)})$, and we transform the L values by $\frac{2\bar{y}_{svy}(1-l_{10})}{\bar{y}_{svy}+\bar{y}_{na}} + L\left(1-\frac{2\bar{y}_{svy}(1-l_{10})}{\bar{y}_{svy}+\bar{y}_{na}}\right)$.

Methodology in Practice

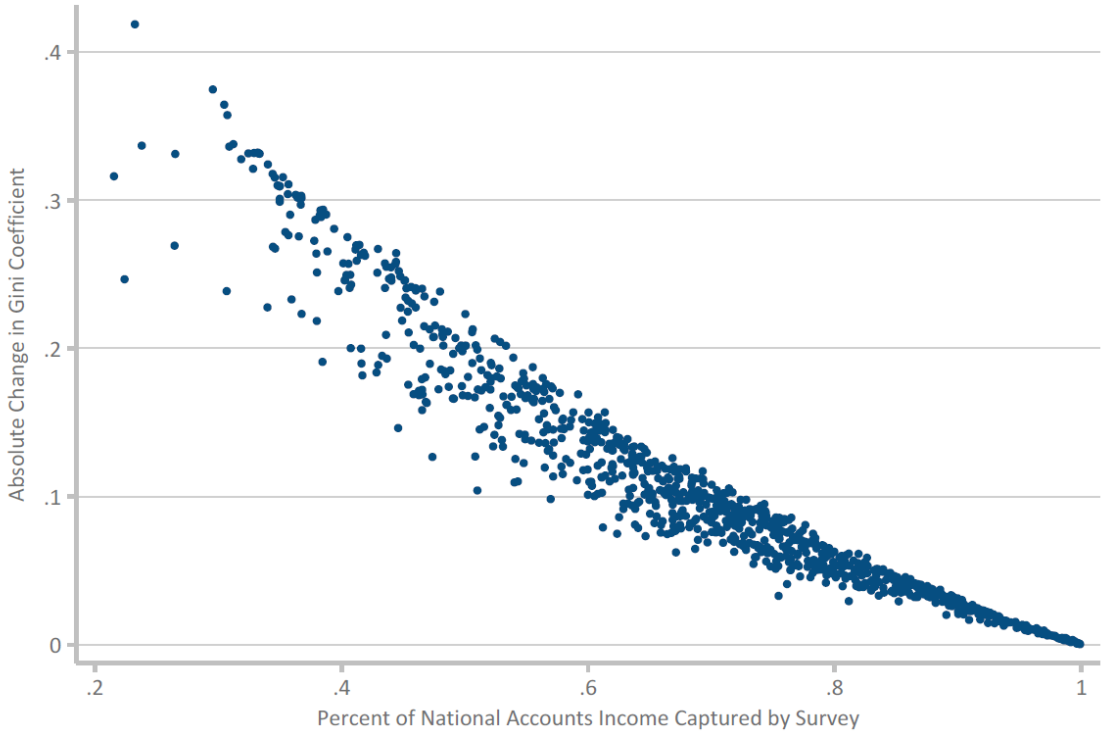
An example of how a country's Lorenz curve changes after this adjustment can be seen in Figure A2 below.

Figure A2: Sample Lorenz curves, before and after the adjustment



As expected, this method leads us to adjust the Gini coefficient by a larger amount, on average, in countries where the gap between surveys and national accounts is greater. Figure A3 below shows the relationship between the size of the gap and the magnitude of our adjustment.

Figure A3: Absolute change in Gini coefficient after adjustment



Similarly, our estimate of the percent of the population that is missing from the survey – that is, the missing top incomes – increases with the size of the gap, as shown in Figure A4.

Figure A4: Percent of adjusted population missing from survey

The full procedure for adjusting raw p and L values can also be found in the accompanying code. The procedure described in this document is implemented in the do-file "Calculate Pareto-Adjusted P's and L's."