ERF 31ST Annual Conference

YOUTH

DEMOGRAPHIC DIVIDEND & MIGRATION IN MENA:
CHALLENGES & OPPORTUNITIES IN UNCERTAIN TIMES

reshaping the future

2025

Poverty Estimation in Data-Poor Countries:

Parametric Approximations Based on Limited Statistics

Hassan Hamie,
Jinane Jouni
and Vladimir Hlasny

ECONOMIC
RESEARCH
FORUM

منتدى
البحوث
الاقتصادية

Poverty estimation in data-poor countries:

Parametric approximations based on limited statistics

Abstract

Technical issues with income datasets and heterogeneous statistical approaches for addressing them give rise to discrepancies in poverty estimates across different studies. We assess how alternative parametric modeling approaches perform under various modes of data granularity. We use a large worldwide set of household income surveys – including notably conflict-affected and high-income countries in the MENA region – on which we artificially impose one of four alternative degrees of data granularity: individual-level microdata, random extraction from the microdata, grouped data, and a pair of basic distributional statistics. We then correct for the data limitations, and estimate poverty headcount ratio and poverty gap, using several parametric distribution functions advanced in prior studies. We find that, when only basic distributional statistics are available, lognormal and Fisk distributions demonstrate a similar, moderate degree of estimation accuracy, compared to other functional forms. With grouped income data, three- and four-parameter models, particularly the beta and generalized beta functions, perform well. With microdata, three- and four-parameter models again outperform two-parameter models. These findings underscore the accurate fit of four-parameter models in various data environments, particularly compared to two-parameter alternatives. The findings also highlight the challenges in modeling the bottom of income distributions when only basic distributional statistics are available.

JEL Codes: D31, I32, N35

Authors: Hassan Hamie, Jinane Jouni, Vladimir Hlasny

1. Motivation

Technical inconsistencies in income datasets and statistical approaches to addressing them often give rise to discrepancies in poverty estimation across different studies. This includes the use of different data sources or measurement instruments, different granularity of data, choice of welfare aggregate such as income or consumption, choice of limiting assumptions about the shape of the distribution, and the application of different poverty lines.

These considerations are particularly relevant in developing and conflict-affected countries such as those across the MENA region, where data collection and dissemination are highly politicized and data quality may be compromised. For instance, data may not be nationally representative, or some geographic or demographic groups may be omitted entirely. Some statistical agencies also provide only limited subsamples such as 25–75% of the original survey microdata, or report only grouped-data information or basic distributional statistics. Agencies in other developing countries provide top or bottom coded data.

To what extent do the discrepancies in poverty estimates stem from the alternative estimation methods deployed in response to limited data availability? This study evaluates poverty estimates obtained under different modes of data availability, using microdata from 800+ income and consumption surveys from 60+ countries worldwide, including 28 surveys from 7 MENA countries. This database encompasses diverse country-income groups, geographic regions, and years from the 1950s to the 2020s. These surveys are obtained from the Luxembourg Income Study (LIS) database. Most are nationally-representative and harmonized, ensuring consistent definitions of incomes or expenditures (henceforth for simplicity: incomes) and sampling parameters such as population weights.

We consider three scenarios for the counterfactual restrictions on data availability. The first scenario consists of the unrestricted raw microdata themselves, comprising detailed household income values for all surveyed households. As a modality on the first scenario, microdata are randomly subsampled, to only feed 25–75% of observations into the analysis.

In the second scenario, grouped data for income distributions are utilized. This may involve information on income and population shares (a.k.a., Lorenz coordinates) of specific population strata, such as income quantiles. The third scenario restricts the data availability to just two summary statistics such as the mean and the Gini coefficient of inequality.

These data-availability scenarios are quite realistic given that data repositories such as the World Income Inequality Database, the World Bank’s Poverty and Inequality Platform (PIP), and the World Inequality Database also report data in these formats.

For each country–year observation (including some cases where 2 alternative surveys exist in the LIS database for the same country–year), the three scenarios of data availability are constructed and used in estimating poverty parametrically, one scenario at a time. While numerous parametric functional forms have been proposed in extant literature for poverty and inequality studies, no prior study has investigated systematically whether data-availability constrains the accuracy of poverty estimation and to what extent. This analysis sheds light on the singular impact of the modalities of data availability on parametric estimation of poverty.

The rest of the study is organized as follows. Section 2 first summarizes the data format for each scenario and reviews the literature on the most applicable parametric functional models for estimating poverty under that scenario. Next, we highlight the performance of each functional form within each data-availability scenario, and undertake a theoretical assessment of the alternative approaches. Section 3 proposes an application of all the proposed parametric functional forms across all three data-availability regimes using the LIS database. Finally, section 4 presents the sensitivity of poverty estimates across various functional forms within each data-availability regime and, critically, between all three regimes of data availability. Section 5 concludes.

2. Models of Income Distribution: Literature Review

The study of income poverty or inequality often requires researchers to model income distributions with functional forms that reflect real-world income dynamics. However, data limitations—such as the common availability of only distributional statistics, grouped or tabulated income data, rather than complete microdata—have long challenged these efforts. Consequently, a variety of mathematical models have been developed to work with data in various forms, enabling the analysis of income distributions even in the absence of complete data.

To estimate income distributions and assess poverty accurately, researchers have employed both parametric and non-parametric techniques. Non-parametric techniques, including simple linear or quadratic interpolation and kernel density estimation, are often used to construct empirical Lorenz curves from available data. Parametric methods on the other hand estimate the income density function whether based on unrestricted ('raw') microdata containing detailed household income/expenditure values; limited 'grouped data' (information on income and population shares – i.e Lorenz curve); or even just restricted grouped data using at least two summary statistics (the mean and the Gini coefficient of inequality).

2.1 Non-parametric: Interpolation

Linear or quadratic interpolation of income shares can be used to connect available Lorenz curve points based on income shares. While straightforward, interpolation methods can yield inaccurate estimates, particularly at specific points along the distribution. This approach, widely used in studies on global distribution of income inequality (Bourguignon and Morrison, 2002; Lakner and Milanovic, 2016), has often been noted to underestimate inequality, making relative inequality measures lower-bound approximations (Kakwani, 1980).

2.2 Non-parametric: Kernel density

Kernel density estimators, such as those applied by Sala-i-Martin (2006) for global poverty estimation and approximation of national income distributions, provide a more flexible non-parametric alternative for grouped data. However, a comparison between kernel density estimation and parametric estimation of the Lorenz curve—both applied to grouped data—suggests that the latter performs better and should be the preferred approach (Minoiu and Reddy, 2014).

2.3 Parametric

Parametric models, meanwhile, offer a structured approach that typically outperforms non-parametric techniques in poverty estimation from grouped data (Dhongde and Minoiu, 2013; Bresson, 2009; Jorda et al., 2018). These methods require an ex-ante assumption about the underlying income distribution shape, and depending on data availability, may be applied to raw microdata, grouped data, or even summary statistics like mean income and inequality metrics.

Parametric: Microdata

When observed income data are available, the parameters of parametric income distributions can be estimated through methods like Maximum Likelihood Estimation, quantile matching, or multinomial likelihood functions. These techniques, however, face challenges with heavy-tailed distributions allowing the presence of extreme values. For example, Cowell and Flachaire (2007) contend that standard bootstrapping techniques often perform poorly for income distributions with a heavy-tailed shape, such as those in the GB2 family, resulting in inaccurate estimates of inequality. However, this issue can be mitigated by using a non-standard bootstrap; the semiparametric bootstrap has demonstrated superior performance compared to other methods, providing accurate inference in finite samples.

Parametric: grouped data

Given that microdata are frequently unavailable, researchers rely heavily on grouped data and summary statistics to estimate income distributions. Parametrized Lorenz curves such as the Generalized Quadratic and Beta Lorenz Curves are common parametric methods that have been adopted to estimate a complete Lorenz curve from limited data (Villasenor and Arnold, 1989; Kakwani, 1980; Minoiu and Reddy, 2009).

Parametric: Restricted data

When only restricted data are available, researchers can derive model parameters by solving a set of linearly independent equations, although this becomes increasingly complex as more parameters are added (Bourguignon, 2003). For instance, a two-parameter model can be easily derived using summary statistics such as mean and the Gini index, but a three-parameter model – such as the Singh–Maddala distribution – requires an additional statistic, like the Theil index. Due to the increasing complexity, it is advised to adopt an approach that focuses on estimating the Lorenz curve associated with each distribution (Bresson, 2009).

GB2 family functions

Among the various parametric families, the generalized beta distribution of the second kind (GB2) has gained prominence for modeling income distributions. The GB2 and its variants, widely used in national and global studies, have proven to fit income data across different countries and periods effectively (Feng et al., 2006; Hajargasht et al., 2012; Jorda and Niño-Zarazua, 2019). McDonald (2008) demonstrated that, given any arbitrary estimation criterion, higher branches (with more parameters) typically yield improved fit as they retain more parameter flexibility.

Historically, income distribution modeling began with the Pareto principle and evolved through the introduction of other strong candidate distributions such as the lognormal (Aitchison and Brown, 1957, Lopez et al., 2006). More recently, studies have shown that alternative functional forms, including the Gamma and Weibull (Pinkovskiy and SalaiMartin ,2009) and generalized Pareto distributions (Bourguignon et al., 2016) may provide a better fit than the lognormal (Bandourian et al., 2002). Other notable parametric forms, like those introduced by Maddala and Singh (1976) and Dagum (1977), as well as the generalized beta distribution, have been shown to capture complex income dynamics (Jenkins, 2009; Hajargasht et al., 2013).

The estimation of parameters for these various functional forms heavily relies on data availability. Although this family of distributions offers flexible modeling options, it can present estimation challenges, particularly in country-specific studies where limited data may be available (Burkhauser et al., 2012; Jenkins et al., 2011).

Our study contributes to the extant literature by examining how data limitations affect poverty estimation. We assess alternative parametric estimation approaches, systematically comparing the outcomes across data forms: summary statistics, grouped data, and detailed microdata. By comparing parametric methods across various data types—from aggregated summary statistics to microdata—our study highlights the effects of data constraints on poverty estimation accuracy and provides insights into potential discrepancies in poverty metrics.

3. Methodology

In this section, we outline the methodology employed for fitting a range of parametric distribution functions to income or consumption data under varying data availability scenarios. Our approach is structured to provide accurate parameter estimation for income distribution models, enabling consistent poverty measurements across different poverty line thresholds.

3.1 Parameter estimation using grouped data (data regime I):

In the first layer of analysis, we rely on aggregated summary statistics, specifically the average income (v) and the Gini coefficient (G). Based on these, we apply three parametric distributions – lognormal, Fisk, and an adjusted Pareto distribution referred to as New Pareto. The following paragraphs describe the parameter estimation techniques for each distribution.

Table (1) summarizes the linear relationships between distribution parameters and the available aggregate statistics. The lognormal distribution function's two parameters μ and σ are estimated using equations 1 and 2. Similarly, equations 3 and 4 estimate the two parameters a and b for the Fisk distribution.

The lognormal distribution is characterized by two parameters, μ and σ, which can be directly estimated from the mean and Gini coefficient as in equations (1) and (2). For the Fisk distribution, two parameters, a and b , are estimated using equations (3) and (4).

The New Pareto distribution, parameterized by α and β, involves a more complex estimation procedure. Due to the lack of a closed-form solution for the integral in equation (5), we numerically estimate α based on the Gini coefficient, following the methodology proposed by Bourguignon (2016). Once α is obtained, we determine β by optimizing it according to the first moment equation (i.e., $E(Y^{r=1}) = v$) as specified in equation (6). This iterative approach ensures that both parameters align with the given summary statistics, allowing for a robust representation of income data.

Table 1 – Formal definition of parametric distribution functions

	Parameters	
Lognormal	$\mu = \log(v) - \frac{\sigma^2}{2})$	(1)
	$\sigma = \sqrt{2}\,\Phi^{-1}\left(\frac{1+G}{2}\right)$	(2)
Fisk	$a = \frac{1}{G}$	(3)
	$b = \frac{\mu}{\Gamma(1+\frac{1}{a})\Gamma(1-\frac{1}{a})}$	(4)
New Pareto	$G = 1 - 2\int_0^1 L(p)dp = 1 - \frac{2}{\frac{\alpha}{\alpha-1}\,{}_2F_1\left(1,-\frac{1}{\alpha};2-\frac{1}{\alpha};-1\right)}\int_0^1\int_0^p\left(\frac{1+t}{1-t}\right)^{\frac{1}{\alpha}}dtdp$	(5)
	$E(Y^r) = 2*\alpha*\beta^\alpha\int_\beta^\infty\frac{y^{r+\alpha-1}}{(y^\alpha+\beta^\alpha)^2}dy = \frac{2*\alpha*\beta^r\,{}_2F_1\left(2,2-\frac{r+\alpha}{\alpha};3-\frac{r+\alpha}{\alpha};-1\right)}{\alpha-r}$ $o < r < \alpha, \beta^\alpha \neq 0$	(6)

${}_2F_1\left(1,-\frac{1}{a};2-\frac{1}{a};-1\right)$ is a hypergeometric function

3.2 Parameter estimation using grouped data (data regime II)

In the second layer of analysis, we incorporate additional data—specifically, grouped income data in conjunction with the mean income $(\bar{Y})$ and the Gini index $(G)$to estimate distribution parameters for models with a larger parameter space. The grouped data consists of cumulative population proportions, denoted as$c' = (c_1, \dots, c_{n-1}, 1)$ which are treated as fixed values, and the corresponding cumulative income shares $y' = (y_1, \dots, y_{n-1}, 1)$ considered as random variables.

This study focuses on the four-parameter generalized beta distribution of the second kind (GB2) and ITS related models, including the three-parameter distributions (such as Singh–Maddala (SM), Dagum, and Beta 2) and two-parameter distributions (such as Lognormal and Fisk). Each distribution has a cumulative distribution function (CDF), denoted as $F(y;\, \theta)$, and a Lorenz curve $L(c;\, \theta)$. Table 2 outlines the functional forms of each Lorenz curve for the selected models.[1]

In addition to the maximum likelihood estimation using a multinomial likelihood function proposed by McDonald (2008), two econometric strategies are introduced to estimate these parametric distributions using the nonlinear least squares method: the equally weighted minimum distance (EWMD) estimator and the optimally weighted minimum distance (OMD) estimator. Both methods involve estimating the parameters by minimizing the distance between the observed cumulative income share (y_n) and the theoretical Lorenz curve $L(c;\, \theta)$ based on the assumed income distribution. The main difference between the two estimators lies in the weight matrix: EWMD applies equal weights, while OMD employs variable weights (as discussed in Jorda et al., 2018). The EWMD estimates serve as starting values for the two-step OMD method.

The EWMD shape estimator (denoted by $\theta^h \subset \theta$) is computed by minimizing the following objective function

$$\hat{\theta} = \arg\min_{\theta} M(\theta)' M(\theta), \tag{7}$$

where $M(\theta) = L(c;\, \theta) - y$, represents the difference between the theoretical and observed moments, treated equally without any additional weighting matrix.

Numerical optimization using the Levenberg-Marquardt Algorithm is implemented, which requires a starting value. For two-parameter distributions (with one shape parameter), the starting value is obtained – as in the first layer – by solving the Gini index expression $G(\theta^h) = g$. For three-parameter distributions involving two shape parameters (θ_1^h, θ_2^h) are needed, a grid search method is used to avoid local convergence of the equation above. A range of $\theta_1^h$is defined, for each value, the second θ_2^h is computed by solving for $G(\theta_1^h, \theta_2^h\,) = g$. Equation (7) is then estimated over the grid ranges, retaining the parameters that yield the minimum residual sum of square.

For the GB2 distribution, which includes three shape parameters, initial estimates are derived from restricted models (SM, B2, and Dagum) by setting one shape parameter to “1”. A similar approach as used for two-shape parameters is applied to each restricted model, resulting in a larger number of initial value

[1] Parameter estimation is performed using the *fitgroup* function in the *GB2group* package within RStudio.

combinations that are used to generate different sets of estimates. The estimates with the smallest residual sum of squares are selected.

Since the Lorenz curve is scale-independent (its shape is independent from the income unit of measurement), scale parameter (θ^s) could not be computed from above equation. Instead, it is determined by equating the observed mean income with the first moment expression of the specified distribution $E(Y;\theta\,) = \bar{Y}$, using the estimated shape parameters. Standard errors of parameter estimates are calculated by Monte Carlo simulation.

The OMD estimation introduces a weight matrix and is carried out as

$$\hat{\theta} = \arg\min_{\theta} M(\theta)'\,\Omega^{-1} M(\theta), \tag{8}$$

where Ω is the variance-covariance matrix of the moment conditions (detailed in the appendix). This matrix depends on partial first and second moments, in addition to income class bounds which are not available. Consequently, consistent EWMD estimates are used to obtain an initial consistent estimate $\hat{\Omega}$. Given the large sample sizes, bias is not a concern. Substituting $\hat{\Omega}$ in Equation (8) yields the second-step OMD estimates. The sample size must be known to compute standard errors.

Table 2 - Lorenz curve under each distributional form

Distribution	$\boldsymbol{F(x;\ \theta)}$	$\boldsymbol{E(X;\ \theta)}$	$\boldsymbol{L(p;\ \theta)}$
GB2	$B\left(\dfrac{\left(\frac{x}{\beta}\right)^{\alpha}}{1+\left(\frac{x}{\beta}\right)^{\alpha}};\ \boldsymbol{p},\boldsymbol{q}\right)$	$\dfrac{b\,B\left(p+\frac{1}{\alpha},q-\frac{1}{\alpha}\right)}{B(p,q)}$, $q > \dfrac{1}{\alpha}$	$L_{GB2}(c;\alpha,p,q) = B\left(B^{-1}(c;p,q);\ p+\dfrac{1}{\alpha},q-\dfrac{1}{\alpha}\right)$
Beta-2	$B\left(\dfrac{x/\beta}{1+x/\beta};\ \boldsymbol{p},\boldsymbol{q}\right)$	$\dfrac{b\,B(p+1,q-1)}{B(p,q)}$, $q > 1$	$L_{B2}(c;\alpha,q) = B(B^{-1}(c;p,q);\ p+1,q-1)$
Singh–Maddala	$1-\left(1+\left(\frac{x}{\beta}\right)^{\alpha}\right)^{-\boldsymbol{q}}$	$\dfrac{b\,\Gamma\left(1+\frac{1}{\alpha}\right)\,\Gamma\left(q-\frac{1}{\alpha}\right)}{\Gamma(q)}$, $q > 1/a$	$L_{SM}(c;\alpha,q) = B\left(1-(1-c)^{\frac{1}{q}};1+\dfrac{1}{\alpha};q-\dfrac{1}{\alpha}\right)$
Dagum	$\left(1+\left(\frac{x}{\beta}\right)^{-\alpha}\right)^{-\boldsymbol{p}}$	$\dfrac{b\,\Gamma\left(p+\frac{1}{\alpha}\right)\,\Gamma\left(1-\frac{1}{\alpha}\right)}{\Gamma(p)}$, $a > 1$	$L_{Da}(c;\alpha,p) = B\left(c^{1/p};p+\dfrac{1}{\alpha};1-\dfrac{1}{\alpha}\right)$
Lognormal	$\Phi\left(\dfrac{\log x-\ \mu}{\boldsymbol{\sigma}}\right)$	$\exp(\mu+\ \sigma^2/2)$	$L_{LN}(c;\ \sigma) =\ \Phi(\Phi^{-1}\,(c)-\ \sigma)$
Fisk	$1-\left(1+\left(\frac{x}{\beta}\right)^{\boldsymbol{\alpha}}\right)^{-1}$	$\dfrac{b\ (\pi/\alpha)}{\sin(\pi/\alpha)}$, $\alpha > 1$	$L_F(c;a) = B\left(c;1+\dfrac{1}{\alpha};1-\dfrac{1}{\alpha}\right)$

Shape parameters are in bold

3.3 Parameter estimation using microdata (data regime III)

In the third layer of analysis, parameters of two-parameter functional forms are estimated directly from microdata. While microdata provides rich detail, handling large samples requires careful consideration of several factors that can impact the accuracy and reliability of estimates. Outliers and data errors (such as zero, negative or extreme income values) may skew results, making it essential to apply data transformations to ensure meaningful estimates. Additionally, the choice of estimation method—whether Maximum Likelihood Estimation (MLE) or Generalized Method of Moments (GMM)—can significantly influence the stability and precision of parameter estimates. Adequate degrees of freedom are also necessary to balance model complexity with data size, preventing issues of over- or under-fitting. By addressing these factors through robust modeling and validation practices, reliable estimates are achievable.[2]

Under the assumption that income data is independently and identically distributed, distribution parameters are estimated by maximizing the likelihood function:

$$L(\theta) = \prod_{i=1}^{N} f(y_i;\ \theta),$$

where y_i represents the N income observations. For distributions with more than one parameter, numerical optimization via the Nelder–Mead algorithm is applied to maximize the log-likelihood function. Classical distributions are predefined and do not require initial values, except for the custom defined New Pareto distribution. Standard errors are derived from the Hessian matrix at the optimal solution, and bootstrap techniques can reduce uncertainty in parameter estimates.

Although full microdata ideally offers detailed parameter estimates, aggregated data (such as mean, Gini index, and Lorenz curve points) can sometimes be more efficient, robust to quality issues, and suitable for specific analyses, especially with simple distribution models. Additionally, aggregated data can reduce computational demands and clarify relationships by smoothing out noise, making it useful when microdata is complex or error prone. For specific applications, like estimating the poverty headcount in our case, aggregated data may be sufficient, offering adequate insights without the need for full microdata.

3.4 Poverty estimation

Once the distribution parameters are estimated, poverty headcount and other metrics are derived using the Foster–Greer–Thorbecke (FGT) class of poverty measures, which approximates the poverty level through the probability density function $f(y)$:

$$P_{\varphi} = \int_{0}^{z} \left[\frac{z-y}{z}\right]^{\varphi} f(y)\ dy. \qquad (9)$$

Where φ is a positive integer in $\varphi \in \{0,1,2\}$, each corresponding to a specific poverty measure: poverty headcount, poverty gap, and squared poverty gap respectively. In this study, we focus on estimating the poverty headcount under several poverty lines, denoted as denoted as P_0^z.

[2] In R, the *fitdistr* function from the MASS package is used to estimate parameters using multiple estimation techniques, including MLE, moment matching (MME), and quantile matching (QME).

This estimation can be approached by solving the integral of the PDF in equation (9). Alternatively, the headcount P_0^z can be derived more directly using the cumulative density function (CDF) as we are dealing with normal distributions. Table 3 and Annex I provide detailed equations for the PDF, CDF, and quantile functions for each parametric model.

Observed poverty headcount, which is taken as the anchor for comparison, is computed simply by counting income or consumption values below a specified threshold.

Table 3 – Poverty estimation for all functional form (2,3, and 4 parameters)

Model	*Parametric Functional Form*	*Notes*
Lognormal	$F(z;\mu,\sigma) = \Phi\left(\frac{\log\left(\frac{z}{v}\right)}{\sigma} + \frac{\sigma}{2}\right)$	$\Phi(.)$is the standard normal cumulative density function
Fisk	$F(z;\alpha,\beta) = \left(1 - \left(1 + \left(\frac{z}{\beta}\right)^{\alpha}\right)^{-1}\right)$	
New Pareto	$F(z;\alpha,\beta) = \left(1 - \frac{2\beta^{a}}{z^{\alpha} + \beta^{a}}\right)$	
Singh–Maddala	$F(z;\varphi) = 1 - \left(1 + \left(\frac{z}{b}\right)^{a}\right)^{-q}$	φ is the is the set of distribution-specific parameters (a, b, q)
Dagum	$F(z;\varphi) = \left(1 + \left(\frac{z}{b}\right)^{-a}\right)^{-p}$	φ is the is the set of distribution-specific parameters (a, b, p)
Beta2	$F(z;\varphi) = B\left(\frac{\frac{z}{b}}{1 + \frac{z}{b}}; p, q\right)$	$B(v;p,q) = \int_0^v \frac{y^{p-1}(1-y)^{q-1}}{B(p,q)}$ Denotes the incomplete beta function ratio
	$F(z;\varphi) = B\left(\frac{\left(\frac{z}{b}\right)^{a}}{1 + \left(\frac{z}{b}\right)^{a}}; p, q\right)$	

4. Results and Discussion

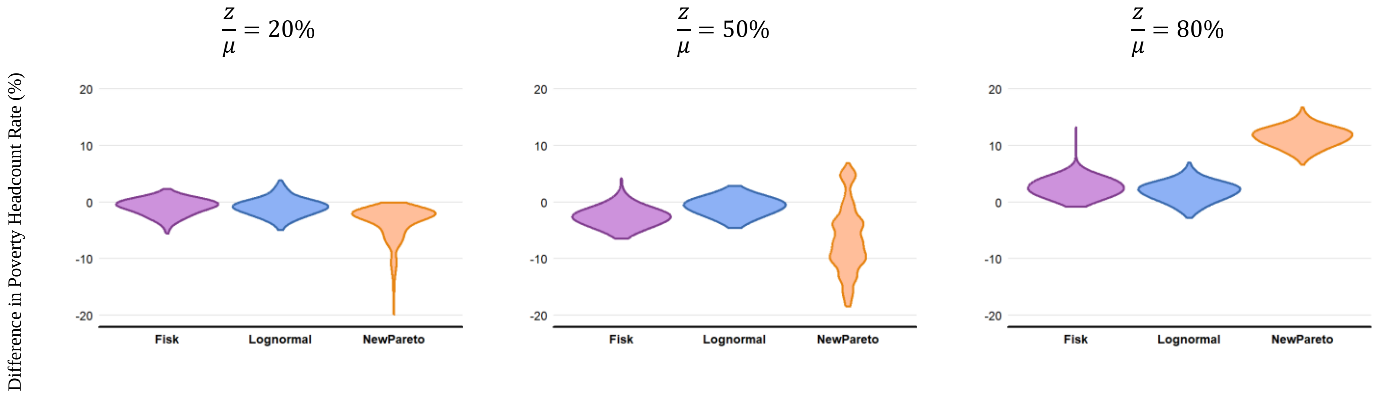
4.1 Data regime 1: Basic distributional statistics

In assessing the accuracy of different functional models for predicting poverty levels, we compare the predicted and observed poverty headcount across various country–year datasets. Figure 1 below depicts the headcount difference across various poverty lines for 2-parameter distributions ($\widehat{P_0^z} - P_0^z$) using limited grouped data (data regime 1).[3]

In evaluating poverty estimates, the lognormal distribution consistently yields the most accurate results. However, in cases where data aligns closely with both lognormal and Fisk distributions, the latter typically provides a more precise poverty estimate (as observed in non-outlier cases with data regime 3).

Overall, poverty estimates derived from grouped data exhibit more stability and consistency compared to those derived from microdata (to be seen later), especially for the lognormal and Fisk distributions. This improvement is evident in the absence of outliers in the boxplots for the Fisk distribution in grouped data and the concentration around "zero difference" as shown in Figures 1. The New Pareto distribution, however, does not show this same stability.

Figure 1 - Comparison of poverty headcount estimates derived from observed data versus those obtained through imputation (data regime 1), across different two-parameter functional forms and various poverty lines



$\frac{z}{\mu} = 20\%$

$\frac{z}{\mu} = 50\%$

$\frac{z}{\mu} = 80\%$

While aggregate statistics, such as the mean and Gini coefficient, may be limited in scope, they nonetheless provide a wealth of information and effectively represent income and consumption data in its entirety. Simple parametric forms, including lognormal and Fisk distributions, succeed in capturing the data's core characteristics when parameters are estimated using grouped data. Additionally, aggregate data helps mitigate noise that frequently affects microdata, leading to more reliable estimates. This empirical robustness is validated through a broad and diverse dataset.

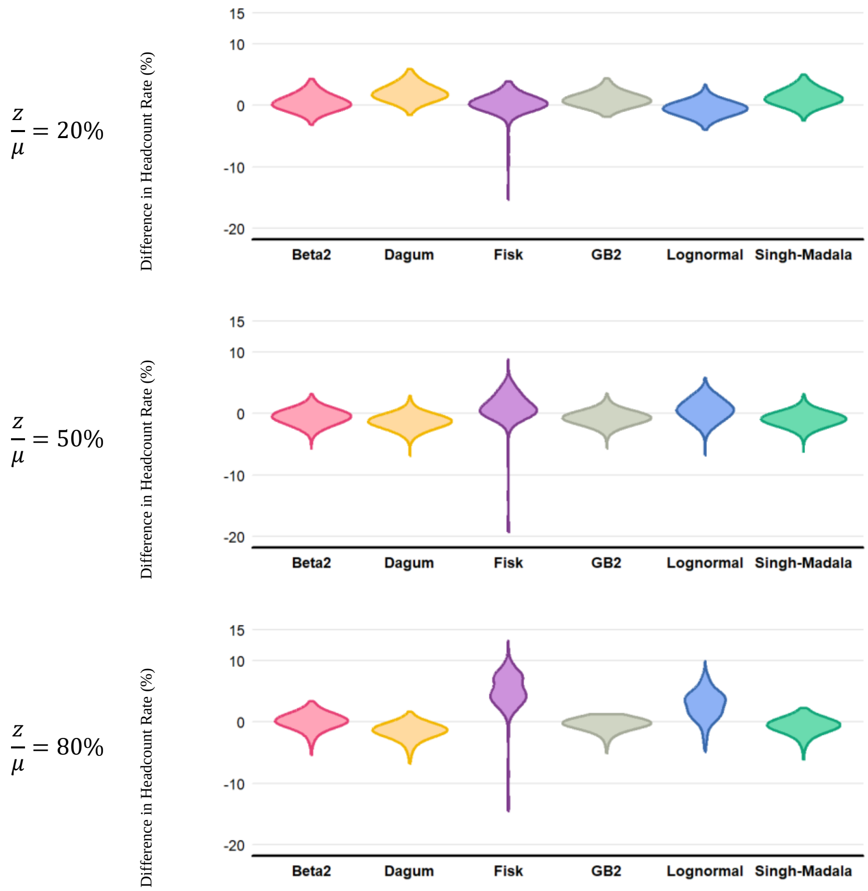
[3] Detailed results are available in Tables 3 and 4 in Annex II, which reveals how limitations in data availability influence the accuracy of poverty estimates across different functional forms.

4.2 Data regime 2: Added Lorenz coordinates

Empirical analysis shows that using only a limited set of income distribution statistics – notably the mean and Gini index – can produce poverty estimates closely aligned with observed values. However, the inclusion of additional information, such as income and population shares (e.g., deciles or quintiles) via Lorenz coordinates (data regime 2), enhances parameter estimation and accuracy. With this richer dataset, three- and four-parameter models become feasible, yielding more precise poverty estimates.

Figure 2 illustrates these improvements but for the Fisk distribution which seems to be worsened by the additional information (poverty underestimation). Poverty headcount differences show reduced variability across poverty lines, indicating a more reliable estimation. Shorter whiskers and larger belly at level zero suggest narrower error margins and more accurate estimations.

Figure 2 - Comparison of poverty headcount estimates derived from observed data versus those obtained through imputation (data regime 2) across various poverty lines



4.3 Data regime 3: Microdata

Income microdata for multiple countries employ parametric functions with 2, 3, and 4 parameters. Literature indicates that models with more parameters can better capture income distribution patterns due to greater flexibility. However, this added complexity can lead to overfitting and the marginal gain may not be worth it. Therefore, evaluating whether additional parameters lead to meaningful improvements is essential.

Figure 3 compares the headcount difference across various distributions and poverty lines. Detailed results are available in Tables 3 and 4 in Annex II, which reveals how limitations in data availability influence the accuracy of poverty estimates across different functional forms.

GB2 model is clearly effective in capturing income distribution intricacies across different poverty thresholds. The GB2 model's four-parameter structure offers substantial flexibility, enabling it to more accurately model complex income distributions than simpler models. On the opposite side, the three-parameter Singh–Maddala exhibits some extreme outliers in cases where fit is inappropriate.

For lower income thresholds (where the poverty line is set at 20% of the mean income), the lognormal distribution generally provides a robust fit, with only a few outliers. This alignment at lower thresholds arises because the lognormal distribution places most data points near its central region, governed by the location parameter. However, as the threshold increases (moving from $\frac{z}{\mu} = 20\%$ to 50% and 80%), the fit weakens in the distribution's tail, where the lognormal model is less precise and other models become more accurate.

Conversely, the Fisk distribution generally provides a stronger fit across most country–year observations for higher ratio, yet it displays notable outliers for low ratios. In fact, even slight inaccuracies in parameter estimation (specifically the shape parameter α) influences the cumulative distribution's tail behavior and subsequently alters the poverty headcount substantially. Figure 4 highlights that outliers arise when the estimated parameter $\hat{\alpha}$ significantly diverges from the observed α_0 (i.e., $\frac{\alpha_0}{\hat{\alpha}} > 1$), leading to a marked increase in the headcount difference. This inconsistency is often due to the empirical income distribution in certain country–years not conforming closely to the Fisk model.

The New Pareto distribution performs reasonably well but does not achieve the same level of accuracy as the lognormal and Fisk distributions. While effective in many cases, this model shows notable discrepancies in poverty estimates for some country–year datasets, suggesting it is more sensitive to variations in data structure or underlying distribution characteristics.

Despite its simplicity, the two-parameter lognormal model outperforms the more flexible Singh–Maddala model at lower poverty thresholds. However, as thresholds rise and extend into the distribution's tail, the lognormal model's fit diminishes. In contrast, the Fisk and New Pareto models exhibit improved performance at higher poverty lines, although the Singh–Maddala model's performance at these levels remains inconsistent due to outliers that hinder its accuracy. Similar issues persist for the Fisk and New Pareto models, underscoring the challenges in modeling income distributions when only limited distributional statistics are provided.

Figure 3 - Comparison of poverty headcount estimates derived from observed data versus those obtained through imputation (data regime 3) across different functional forms and various poverty lines

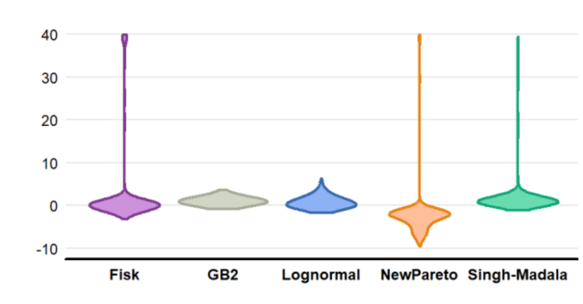
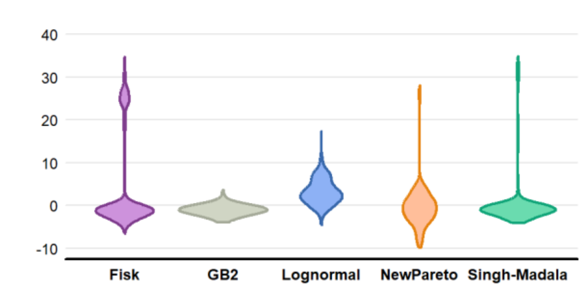
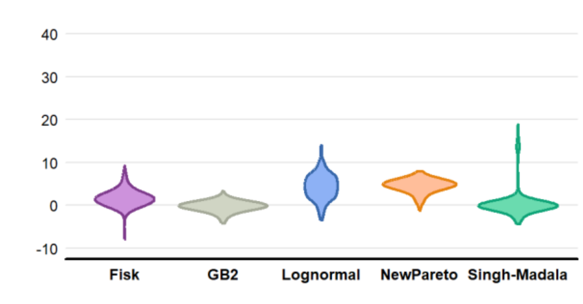
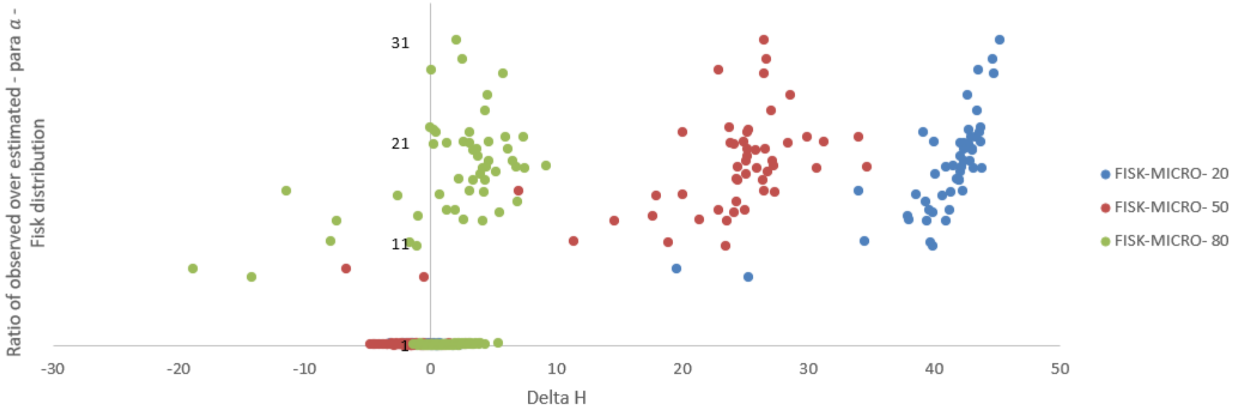
$\frac{z}{\mu} = 20\%$	Difference in Headcount Rate (%)	
$\frac{z}{\mu} = 50\%$	Difference in Headcount Rate (%)	
$\frac{z}{\mu} = 80\%$	Difference in Headcount Rate (%)	

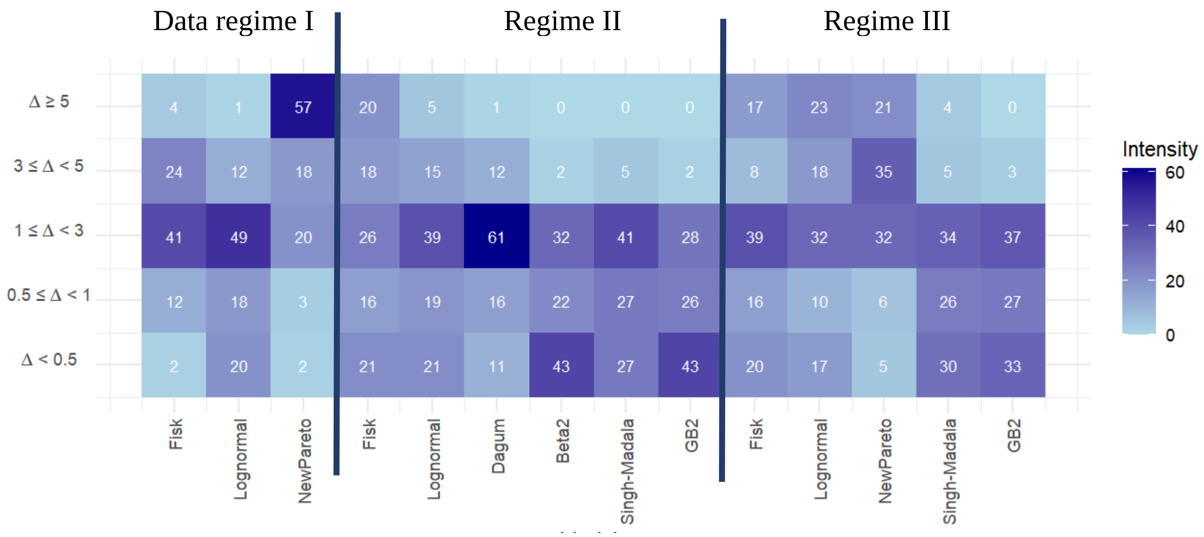
Figure 4 – Ratio of observed over estimated Fisk distribution parameter α



4.4 Final assessment

To assess the sensitivity of parameter estimates across different functional forms and data availability levels, poverty headcount differences were categorized into five ranges, from 0.5 percentage points to approximately 5 percentage points. For each country–year observation, data availability regime, and functional form, the delta headcount was assigned a category corresponding to one of these ranges. These results are presented systematically in Table 4.

Table 4 - Poverty headcount difference (%) across data regimes and parametric functional forms



	Data regime I			Regime II						Regime III				
	Fisk	Lognormal	NewPareto	Fisk	Lognormal	Dagum	Beta2	Singh-Madala	GB2	Fisk	Lognormal	NewPareto	Singh-Madala	GB2
Δ ≥ 5	4	1	57	20	5	1	0	0	0	17	23	21	4	0
3 ≤ Δ < 5	24	12	18	18	15	12	2	5	2	8	18	35	5	3
1 ≤ Δ < 3	41	49	20	26	39	61	32	41	28	39	32	32	34	37
0.5 ≤ Δ < 1	12	18	3	16	19	16	22	27	26	16	10	6	26	27
Δ < 0.5	2	20	2	21	21	11	43	27	43	20	17	5	30	33

Intensity
60
40
20
0

4.5 Within-regime comparison

- Data regime I: both the Lognormal and Fisk distributions demonstrate a similar level of accuracy in estimating poverty headcounts, whereas the New Pareto distribution falls short of this standard, showing larger discrepancies in alignment with observed values.

- Data regime II: except for the Dagum distribution, all three- and four-parameter functional forms generally perform well. Notably, the Beta2 and GB2 distributions achieve strong accuracy, with approximately two-thirds of all poverty estimates differing by less than one percentage point from the observed values, highlighting their robustness when additional grouped data (such as Lorenz coordinates) is available.

- Data regime III: The three- and four-parameter distributions in data regime III also yield accurate poverty estimates, with over 50% of headcount estimates deviating by less than one percentage point from observed values. For two-parameter models, poverty estimates are still acceptable except for the New Pareto where most of the estimates deviate between one to five percentage points.

- As supported in the literature, and given an arbitrary estimation criterion within a specified data regime (Regime I, II, or III), a distribution higher on a branch will generally perform better according to the same criterion. Table 1 empirically supports this pattern, demonstrating that the

GB2 distribution (a higher-branch distribution) consistently outperforms the Singh–Maddala (SM), which in turn performs better than the Fisk distribution.

4.6 Between-regimes comparison

- Distributions in higher branches generally perform better because they offer greater flexibility and can more effectively capture tail behavior and income inequality than distributions with a limited number of parameters. However, this increased complexity—due to the greater number of parameters—can make estimation more challenging and require more extensive data. Importantly, our results indicate that greater data availability does not necessarily lead to better estimation, even when using the same functional form from a higher branch. Therefore, the method of estimation and its sensitivity to data availability become crucial factors. For example, the GB2 estimation results under grouped data or microdata illustrate this point. Although microdata have significantly more data available, the complexity of the model and the estimation method have led to less accurate outcomes compared to the estimation method used when only grouped data are provided. Issues such as overparameterization and the misfitting of certain parameters may contribute to this discrepancy. In essence, this is primarily due to the estimation of parameters for the various functional forms based on data usage and the parameter estimation methodology.

4.7 Comparison across data regimes

- Surprisingly, the lognormal distribution performs best when the least amount of data are available, where only aggregate statistics like the mean and Gini index are used. This finding suggests that even with minimal data, the lognormal distribution effectively captures income distributions for poverty estimation, possibly due to its simplicity and fewer parameters, which reduce the risk of overfitting.

- Working with grouped data using specific estimation techniques (such as EWMD or OMD estimators) can yield results that are both accurate and stable. This implies that grouped data, despite its relative simplicity, provides a wealth of information, while the applied methods maintain robustness in parameter estimation. These methods effectively balance the limitations of data volume with parameter estimation accuracy, resulting in reliable poverty measures across varied distributions.

- Distributions with more parameters, such as the GB2, generally perform better across data regimes as they provide increased flexibility, particularly in capturing tail behavior and income inequality. However, this added complexity often requires extensive data for reliable estimation, making estimation more sensitive to issues like overfitting or parameter instability.

- Our findings indicate that more data does not always enhance estimation accuracy. For instance, the GB2 distribution, when estimated on grouped data, yielded more accurate results than on microdata despite the latter’s larger data volume. This discrepancy likely arises from challenges associated with overparameterization in the presence of noisy or sparse data, underscoring the importance of carefully matching estimation methods to data availability and distributional complexity.

4.8 Goodness of fit comparison

Data regimes 1 and 2

Our analysis is based on the key assumption that a more accurate approximation of the Lorenz curve will yield more reliable poverty estimates. This premise holds true for the applications derived from data regimes I and II, where the objective is to minimize prediction errors associated with the Lorenz curve. The traditional goodness-of-fit approach can be expressed mathematically as follows:

$SSR =$	$\sum_{t=1}^{N}\left(L(p_t)-\widehat{L(p_t)}\right)^2$	$SAE =$	$\sum_{t=1}^{N}\left\lvert L(p_t)-\widehat{L(p_t)}\right\rvert$
$AIC =$	$e^{\frac{2K}{N}} * \frac{\sum_{t=1}^{N}\left(L(p_t)-\widehat{L(p_t)}\right)^2}{N}$	$BIC =$	$N^{\frac{K}{N}} * \frac{\sum_{t=1}^{N}\left(L(p_t)-\widehat{L(p_t)}\right)^2}{N}$
$wssr = \frac{\sum_{t=1}^{N}\left(L(p_t)-\widehat{L(p_t)}\right)^2 * \left(1-\lvert p_t-\widehat{P_0}\rvert\right)^2}{\sum_{i=1}^{N}\left(1-\lvert p_t-\widehat{P_0}\rvert\right)^2}$			

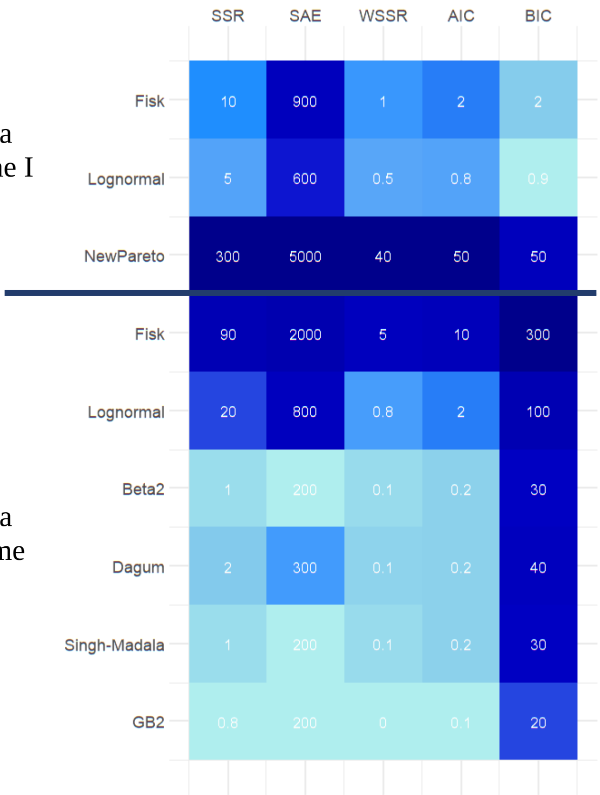
Where $L(p_t)$ denotes the observed cumulative income share at the cumulative at the cumulative share of people p_t, $\widehat{L(p_t)}$ represents the estimated cumulative income share, and N is the total number of data points available.

If one wants to concentrate the analysis on the poor subgroup of the population, we can employ the $wssr$ proposed measure. The weights decrease as the distance from the estimated headcount index $\widehat{P_0}$ increases, thereby allowing for a more nuanced evaluation of fit for low-income populations.

Furthermore, to facilitate the comparison of non-nested models while penalizing for the inclusion of additional parameters, we utilize AIC and BIC. These criteria allow us to assess model performance while accounting for model complexity, helping us identify the functional form that best balances fit and parsimony.

Table 5 presents the average of the different g-o-f measures, with lowest values (lighter color) signal better fit. Note that values have been multiplied by thousands for ease of comparison for numbers. In summary, similar conclusions regardless of the specific measure can be drawn: the superiority of the GB2 distribution in accurately modeling the empirical income distribution and improving poverty estimation. This is followed by favoring higher branch distributions. Concerning data regime 1, the lognormal distributions shows another time its successful fit for income data.

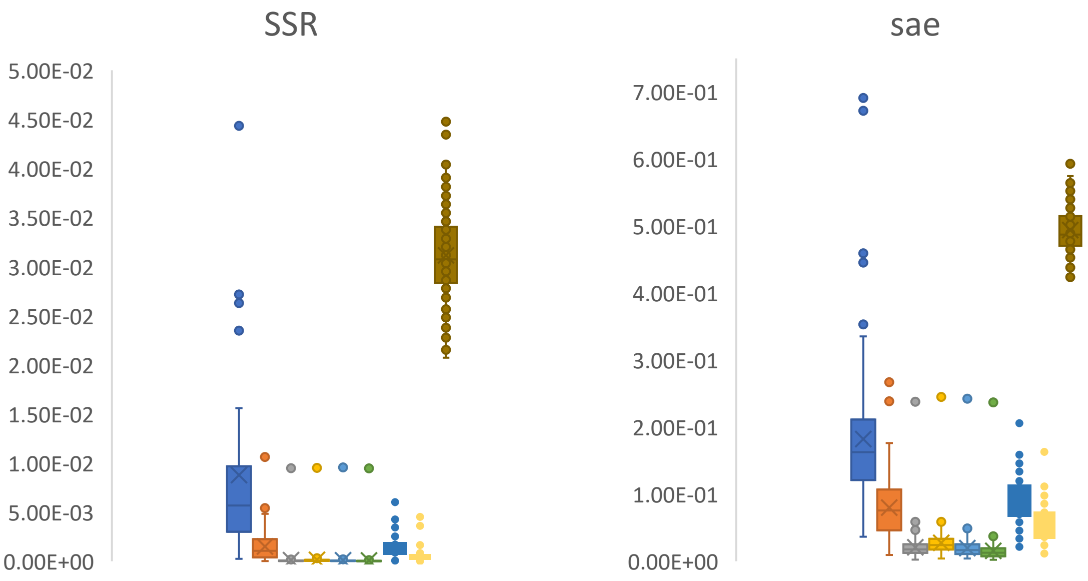
Table 5 - Goodness-of-fit, mean values (in thousands)

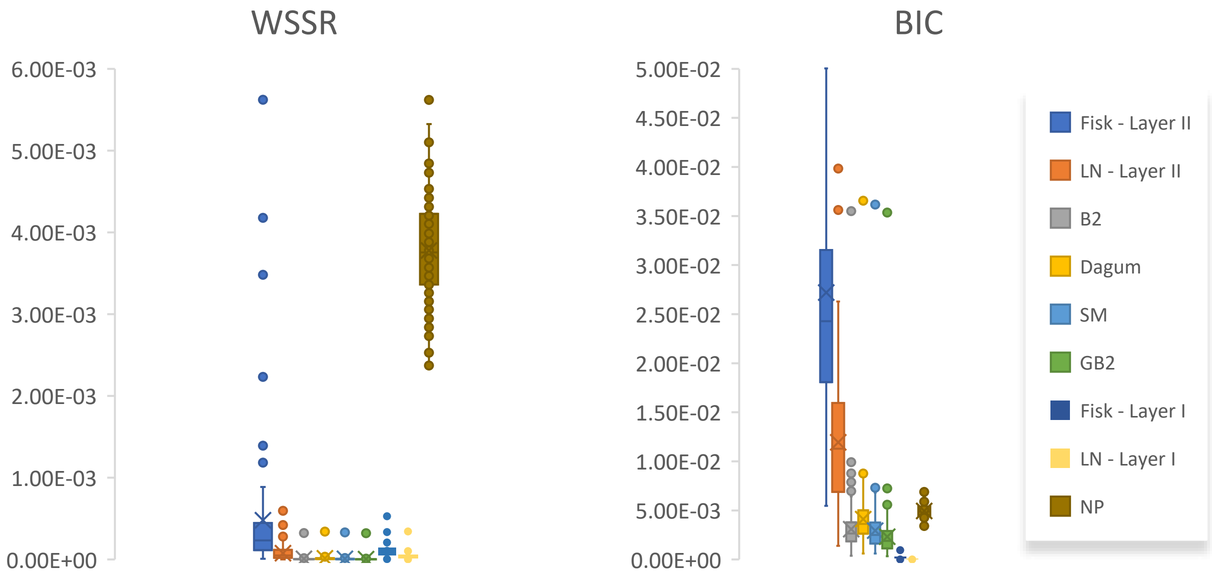


		SSR	SAE	WSSR	AIC	BIC
Data regime I	Fisk	10	900	1	2	2
	Lognormal	5	600	0.5	0.8	0.9
	NewPareto	300	5000	40	50	50
Data regime II	Fisk	90	2000	5	10	300
	Lognormal	20	800	0.8	2	100
	Beta2	1	200	0.1	0.2	30
	Dagum	2	300	0.1	0.2	40
	Singh-Madala	1	200	0.1	0.2	30
	GB2	0.8	200	0	0.1	20

Note: The tabulated results for WSSR measure are estimated using the ratio of Poverty line to mean equal to 0.5. As measured by 50% of the mean

Figure 5 - Goodness-of-fit boxplot




WSSR
6.00E-03
5.00E-03
4.00E-03
3.00E-03
2.00E-03
1.00E-03
0.00E+00
BIC
5.00E-02
4.50E-02
4.00E-02
3.50E-02
3.00E-02
2.50E-02
2.00E-02
1.50E-02
1.00E-02
5.00E-03
0.00E+00
Fisk - Layer II
LN - Layer II
B2
Dagum
SM
GB2
Fisk - Layer I
LN - Layer I
NP

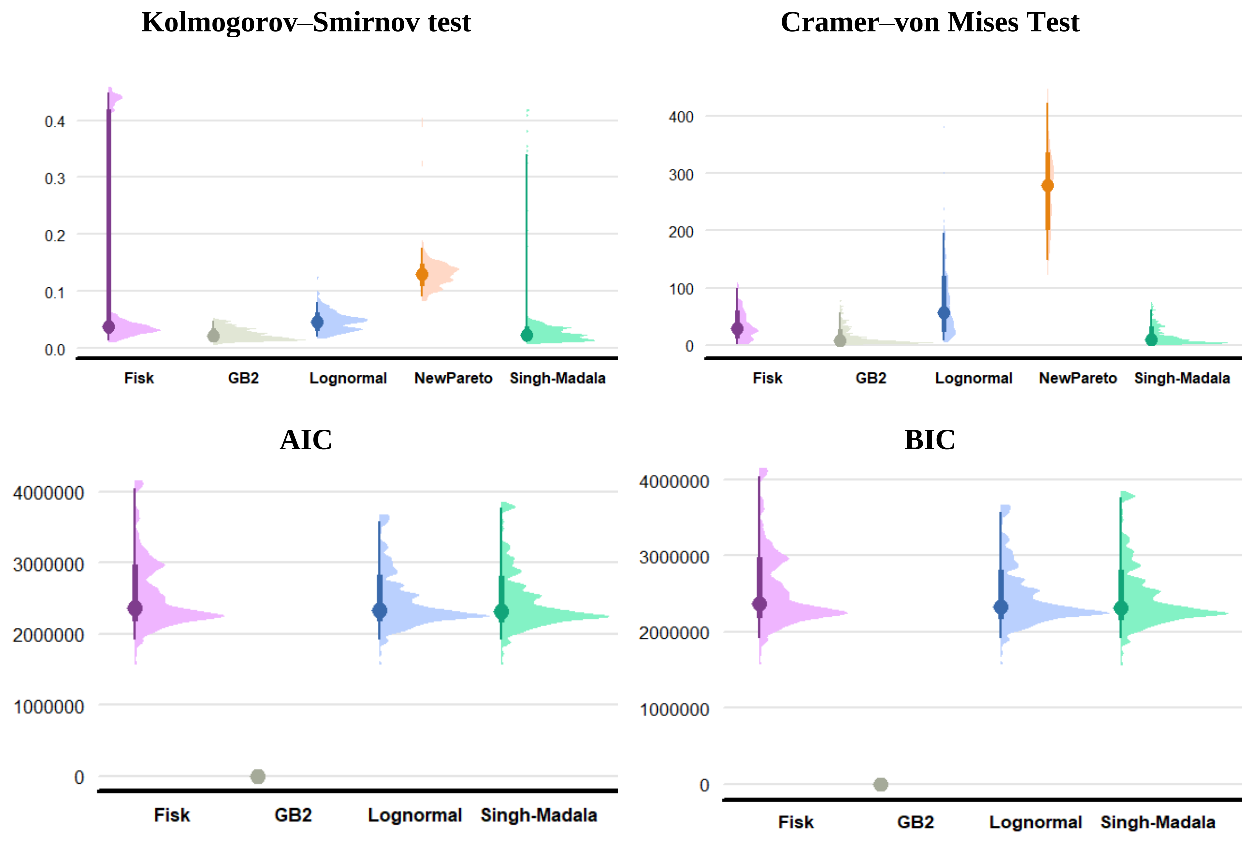
Data regime 3

In this section, we assess the goodness-of-fit statistics for estimation on microdata. Our analysis focuses on quantifying the alignment between the fitted parametric distributions and the empirical distribution of income. To accomplish this, we utilize two classical goodness-of-fit tests: the Cramér–von Mises test and the Kolmogorov–Smirnov test. These tests are widely recognized in the literature for evaluating how closely the fitted distributions match the observed data.

In addition to these traditional tests, we acknowledge the complexity inherent in our modeling framework, which incorporates more than two-parameter functions. To account for this complexity, we incorporate classical penalized criteria based on the log-likelihood, specifically the AIC and BIC. The results demonstrate that the GB2 consistently outperforms its competitors.

In general, the KS and CVM tests are more informative when comparing the different distributional models. As expected, the GB2 exhibits the lowest values in all tests signaling its superiority over other models. The Singh–Maddala competes with the Lognormal although differences in number of parameters. The New Pareto seems to have a weak fit (AIC and BIC measures were omitted). The Fisk distribution seems to be struggling by being extremist: sometimes it hits the fit while sometimes it goes too far.

Figure 6 - Goodness-of-fit measures on microdata



Kolmogorov–Smirnov test

Cramer–von Mises Test

AIC

BIC

5. Policy Implications

This study has contributed novel insights regarding the effect of data constraints on parametric estimation of poverty statistics. We have presented poverty estimates across income surveys worldwide, under various common data-availability regimes and various modeling specifications advanced in prior studies. The analysis has notably included the conflict-affected and high-income countries in the MENA region.

By examining global evidence and comparing outcomes across varying regimes of data availability, the study sheds light on some sources of discrepancies in poverty estimates, namely those linked to the constraints on estimation approach as dictated by limited data availability.

Understanding the relative performance of various estimation methods, in various data availability regimes, can inform practitioners and policymakers about the pitfalls of relying on any of the three evaluated data-availability options for poverty assessments, and understanding the extent to which limited data may affect or even bias the analysis – compared to having access to complete 100% microdata. Understanding the relationship between the poverty estimates derived in alternative ways under alternative data settings can also enhance general confidence in the reliability of poverty measures, which is a challenge in many parts of the world not least in the MENA region.

References

1. Aitchison, J. and Brown, J. A. C. (1957). The Lognormal Distribution. Cambridge University Press. For a one-page summary, see www.komkon.org/~tacik/science/lognorm.pdf For pages 12 and 11-113, see www.StatLit.org/pdf/1957-Aitchison-Brown-Excerpts.pdf
2. Lopez, J. Humberto, and Luis Servén. *A normal relationship?: poverty, growth, and inequality*. Vol. 3814. World Bank Publications, 2006.
3. Bandourian, Ripsy, James McDonald, and Robert S. Turley. "A comparison of parametric models of income distribution across countries and over time." (2002).
4. Pinkovskiy, M. and X. Sala-i-Martin, 2009, "Parametric estimations of the world distribution of income," NBER Working Paper No. 15433 (Cambridge: The National Bureau for Economic Research).
5. Pinkovskiy, Maxim, and Xavier Sala-i-Martin. "Africa is on time." *Journal of Economic Growth* 19 (2014): 311-338.
6. Bourguignon, Marcelo, Helton Saulo, and Rodrigo Nobre Fernandez. "A new Pareto-type distribution with applications in reliability and income data." Physica A: Statistical Mechanics and its Applications 457 (2016): 166-175.
7. Singh, Surendra K., and Gary S. Maddala. "A function for size distribution of incomes." Modeling income distributions and Lorenz curves. New York, NY: Springer New York, 2008. 27-35.
8. Dagum, Camilo. "A new model of personal income distribution: specification and estimation." *Modeling income distributions and Lorenz curves*. New York, NY: Springer New York, 2008. 3-25.
9. Jenkins, Stephen P. "Distributionally-sensitive inequality indices and the GB2 income distribution." *Review of Income and Wealth* 55.2 (2009): 392-398.

10. Hajargasht, Gholamreza, et al. "Inference for income distributions using grouped data." *Journal of Business & Economic Statistics* 30.4 (2012): 563-575.
11. Chotikapanich, Duangkamon, et al. "Calculating poverty measures from the generalised beta income distribution." Economic Record 89 (2013): 48-66.
12. McDonald, J. B. (2008). Some generalized functions for the size distribution of income. In *Modeling income distributions and Lorenz curves* (pp. 37-55). New York, NY: Springer New York.
13. Bourguignon, François. "The growth elasticity of poverty reduction: explaining heterogeneity across countries and time periods." (2003).
14. Bresson, Florent. "On the estimation of growth and inequality elasticities of poverty with grouped data." Review of Income and Wealth 55.2 (2009): 266-302.
15. Bourguignon, François, and Christian Morrisson. "Inequality among world citizens: 1820–1992." American economic review 92.4 (2002): 727-744.
16. Lakner, Christoph, and Branko Milanovic. "Global income distribution: from the fall of the Berlin Wall to the Great Recession." The World Bank Economic Review 30.2 (2016): 203-232.
17. Kakwani, Nanak. "On a class of poverty measures." Econometrica: Journal of the Econometric Society (1980): 437-446.
18. Dhongde, Shatakshee, and Camelia Minoiu. "Global poverty estimates: A sensitivity analysis." World Development 44 (2013): 1-13.
19. Jorda, Vanesa, Jose Maria Sarabia, and Markus Jäntti. "Estimation of income inequality from grouped data." arXiv preprint arXiv:1808.09831 (2018).
20. Chotikapanich, Duangkamon, DS Prasada Rao, and Kam Ki Tang. "Estimating income inequality in China using grouped data and the generalized beta distribution." Review of Income and Wealth 53.1 (2007): 127-147.
21. Jordá, Vanesa, José María Sarabia, and Faustino Prieto. "On the estimation of the global income distribution using a parsimonious approach." Economic Well-Being and Inequality: Papers from the Fifth ECINEQ Meeting. Vol. 22. Emerald Group Publishing Limited, 2014.
22. Feng, Shuaizhang, Richard V. Burkhauser, and John S. Butler. "Levels and long-term trends in earnings inequality: overcoming current population survey censoring problems using the GB2 distribution." *Journal of Business & Economic Statistics* 24.1 (2006): 57-62.
23. Hajargasht, Gholamreza, et al. "Inference for income distributions using grouped data." Journal of Business & Economic Statistics 30.4 (2012): 563-575.
24. Jordá, Vanesa, and Miguel Niño-Zarazúa. "Global inequality: How large is the effect of top incomes?." *World Development* 123 (2019): 104593.
25. Burkhauser, Richard V., et al. "Recent trends in top income shares in the United States: reconciling estimates from March CPS and IRS tax return data." Review of Economics and Statistics 94.2 (2012): 371-388.
26. Jenkins, Stephen P., et al. "Measuring inequality using censored data: a multiple-imputation approach to estimation and inference." Journal of the Royal Statistical Society Series A: Statistics in Society 174.1 (2011): 63-81.
27. Villasenor, JoséA, and Barry C. Arnold. "Elliptical lorenz curves." Journal of econometrics 40.2 (1989): 327-338.
28. Minoiu, Camelia, and Sanjay G. Reddy. "Estimating poverty and inequality from grouped data: How well do parametric methods perform?." *Journal of Income Distribution* 18.2 (2009).
29. Sala-i-Martin, Xavier. "The world distribution of income: falling poverty and… convergence, period." The quarterly journal of economics 121.2 (2006): 351-397.
30. Minoiu, Camelia, and Sanjay G. Reddy. "Kernel density estimation on grouped data: the case of poverty assessment." The Journal of Economic Inequality 12 (2014): 163-189.

31. Cowell, Frank A., and Emmanuel Flachaire. "Income distribution and inequality measurement: The problem of extreme values." Journal of Econometrics 141.2 (2007): 1044-1072.