GLOBAL POVERTY ESTIMATION

THEORETICAL AND EMPIRICAL VALIDITY OF PARAMETRIC LORENZ CURVE ESTIMATES AND REVISITING NONPARAMETRIC TECHNIQUES

Joao Pedro Azevedo and Shabana Mitra[1]

January, 2014

. The parametric method requires that the estimated Lorenz curve satisfy four conditions to be theoretically valid. Our analysis shows that for 30% of the cases at least one of the four conditions is violated. Further the cases where the Lorenz conditions fail have on average higher inequality than that for the region. Poverty estimates based on invalid Lorenz curves are more inaccurate compared to household estimates than if the Lorenz curve was valid, especially for poverty lines of 2.00; 4.00 and 10.00 USD-PPP. Given the high rates of failure we use non-parametric methods as an alternative. In general, poverty estimates from the empirical CDF are more accurate than parametric estimates. This paper concludes that microdata estimates are preferable, and compromises relying on group data come at cost, especially so at higher poverty lines. Our findings reinforce the need for data producers to make anonymized versions of their microdata available, or provide analysts with non-parametric distributions of their welfare aggregate with at least 400 bins. In this paper we use data from 19 countries in from the Latin American and Caribbean region from 1995 to 2010 to assess the performance of the parametric method to estimation of poverty and inequality as is used for the World Bank's $1.25 per day estimates.

Keywords:

JEL Codes:

[1] Joao Pedro Azevedo: Senior Economist, World Bank, jazevedo@worldbank.org. Shabana Mitra: Visiting Scholar, ESOP, University of Oslo and Consultant, World Bank shabanasingh83@gmail.com. The findings, interpretations, and conclusions expressed in this paper are entirely those of the author. They do not necessarily represent the views of the International Bank for Reconstruction and Development / World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent. This work was produced as part of the World Bank Global Poverty Monitoring Working Group validation effort for Latin America and Caribbean conducted in December, 2012 and as a background exercise to the Latin America Middle Class Flagship (Ferreria et al, 2013). The author's would like to thank for the discussions all members of the Global Poverty Working Group, in particular Shoau Chen, Prem Sangrula, Nobuo Yoshida and Johan A. Mistien. The usual disclaimer applies.

1. Introduction

Over the past few decades several institutions and independent researchers have used grouped data from household surveys to compute and monitor poverty and inequality. There have been several reasons for this including the limited computational power to process microdata, particularly during the 1970s and 80s; limited access to microdata as only recently countries have begun to make their data publically available; loss of data during institutional transitions and natural disasters, that has often resulted in data only being available at an aggregate level in the form of distributional tables and charts from printed publications; and, comparability of the recent microdata estimates with the early grouped data estimates. These reasons continue to remain valid even today though more microdata is now available in the public domain than was available to researchers at the beginning of the 1990s.

An important contribution in this area is the World Bank international poverty monitoring exercise that uses the $1.25 per person per day to provide internationally comparable estimates for over 100 countries (Chen and Ravallion 2010)[2]. To account for differences in purchasing power of the dollar across countries the World Bank uses the private consumption PPP conversion factors from the 2005 round of the International Comparison Program, which surveyed prices in 146 countries. The international poverty line of $1.25 per person per day is an average of national poverty lines of the approximately 80 countries in the World conditional on their level of per capita expenditure (PCE) (all PPP adjusted, 2005 prices) with a larger weight to the 15 of the poorest countries[3], whose PCE per capita from national accounts is less than $60 per day (Ravallion, Chen and Sangrula, 2008). The method used to estimate inequality and poverty from grouped data first estimates the parameterized Lorenz curve and uses the fitted values to estimate the inequality and poverty (Villasenor and Arnold 1989; Kakwani 1980a).

The World Bank's global poverty numbers are the largest single source of comparable poverty numbers spanning the longest time horizon available to researchers. These poverty numbers have been used widely by both academics and policy makers. In fact, the global poverty estimates are the metric which is used to measure success of the first Millennium development goal (or MDG1) which aims to reduce world poverty by half by the year 2015 based on the World Bank's estimate of global poverty with the base year of 1990. These numbers and the approach are also of great import for the monitoring of one of the World Bank's goals, namely to reduce the proportion of the world population living on less than 1.25 USD-PPP a day to three percent by 2030 (World Bank, 2013). However since their introduction the methodological choices pertaining to these estimates have been keenly debated.

The debate surrounding these global poverty numbers can be broadly classified into four categories. The one that has received the most attention has been the actual choice of the poverty line at $1.25 per day per capita (Reddy 2009). The second is concerned with the use of PPP figures to bring all the poverty lines to comparable international dollars (Ackland, Dowrick and Freyens 2012; Reddy 2009). The third set of criticisms pertain to the choice of data- there are two alternative sources of data that could be used for the average consumption or income distribution coming from either the National Accounts or the household surveys (Deaton 2003; Dhonge and Minoiu, 2011, Karshenas 2003). Povcalnet uses data from household surveys to estimate poverty. The fourth set of concerns pertains to the methodology that is used to

[2] In the latest release of Povcalnet the methodology has been changed significantly for the Latin American Countries. For the Latin American countries Povcalnet has shifted to the use of household data rather than grouped data for the estimation of poverty and inequality.

[3] The reference countries include Malawi, Mali, Ethiopia, Sierra Leone, Niger, Uganda, Gambia, Rwanda, Guinea-Bissau, Tanzania, Tajikistan, Mozambique, Chad, Nepal and Ghana.

estimate the poverty number. Methodological critiques range from the issue of using grouped data versus the household level data to the specific parametric and non-parametric forms used to arrive at the estimates (Minoiu and Reddy 2009, 2012, Dhongde and Minoiu 2011).

In general the World Bank's method fares well and provides estimates which are very close to those obtained from micro-level data. There have been several studies conducted which have evaluated the performance of these methods (Miniou and Reddy, 2008; Dhonge and Minoiu, 2011). Most studies have shown that the results using the parametric methods are consistent, with the biases being small. The comparison of different non-parametric and parametric methods reveals that the results are similar across these methods. Further on comparison with estimates from household level data using a select set of countries the results show that when the underlying distribution is uni-modal then the estimates from parametric methods are more accurate than for multimodal distributions. However these studies usually use a small number of countries (typically no more than five) or do not have access to household data to make the comparison with estimates from a typical poverty analysis. More importantly these papers did not pay any attention to the tails of the distribution, which becomes salient in countries with low poverty levels at $1.25 per day, such as the middle and upper middle income Latin American countries, nor at higher values of poverty lines, such as the recently announced $ 10.00 per day World Bank Middle Class line for Latin America (Ferreria et al 2013).

In this paper we use data from 19 LAC countries from over 15 years with approximately 69 surveys. Therefore this study uses a larger number of countries and more years than has been done previously in similar exercises, and goes beyond the 1.25 USD-PPP poverty line, and explores the results at higher poverty lines. Using this large database we revisit the question of the validity of the estimated parametric Lorenz curve. Theoretical validity or internal validity of the parametric estimation requires checking four conditions that need to be met by the parametric estimate of the Lorenz curve. These consistency checks correspond to properties of the Lorenz curve. The estimated parametric Lorenz curve should pass through the origin and the point (1,1). There are two other conditions that ensure that the Lorenz curve is monotonically increasing and convex.[4] Though this issue has been mentioned in passing in the Povcalnet estimates and also in the studies done to evaluate the global poverty estimates, it has not received much attention. Minoiu and Reddy (2009) find no correspondence between the validity of Lorenz curve estimating and the quality of the poverty and inequality estimates produced. Further Kakwani 1980 defends the use of parametric Lorenz curves even if they are theoretically invalid due to their better overall performance at fitting the income distributions.

We find that in over 30% of the estimations at least one of the four conditions of the Lorenz curve is violated. Further we see that the Lorenz curve is more likely to fail for the following countries-Bolivia, Brazil, Chile, Ecuador, Mexico, Nicaragua and Paraguay. These countries in general have higher inequality than the average for the region. We also find that the differences between the estimates using these methods and the household data directly are on average about 3 percentage points different for the poverty headcount. A further point to be noted is that the difference increases when the parametric Lorenz was actually not valid.

Given the fact that in a significant number of cases the Lorenz curve is invalid and in these cases the bias of the estimates is larger than when it is valid we next revisit non-parametric techniques to estimate poverty and inequality. To estimate the poverty rates and the Gini coefficient it is sufficient to estimate the nonparametric cumulative distribution from the grouped data. This is advantageous since there is no need to impose any conditions on the behavior of the curve (such as passing through any specific points), the non-parametric estimate of the CDF is by construction well behaved. We use the empirical CDF to estimate poverty and inequality from grouped data for the 19 LAC countries. The difference in estimates from household data versus grouped data using non-parametric forms is less than one percentage point.

[4] More on this later.

The average inequality (measured by Gini coefficient) in the LAC region ranges from 0.525 in 1995 to 0.51 in 2010. In general this is a region of high inequality and using grouped data has the concern of neglecting important distributional issues that may arise in such economies. Therefore we always compare the results from grouped data with those from household surveys. The debate on the use of household data versus grouped data is a long standing one. Use of household level data for the estimation of poverty and inequality is preferred to estimations from grouped data (Davies and Shorrocks 1989; Atkinson and Brandolini 2001). However often due to the availability and access to micro-data we are restricted to the use of grouped data for the estimation. Our results show that on average there is about a 3 point difference in the poverty headcount estimated using household data versus the parametric forms in 1995 whereas the difference was less than one percent for the non-parametric estimates. In general, the non-parametric estimates have a lower bias than parametric estimates.

The results discussed so far have been for the $1.25 per day poverty line. However other poverty lines have also been used in studies. Therefore as a robustness exercise we use four different poverty lines to check our results. We use the poverty line of $2.0, $4.0, $10.0 and $50.0 and find similar results. Since we are using data from LAC countries which primarily use income data we find that there is a high incidence of zero incomes which cannot occur with consumption expenditure data. Therefore we conduct a second set of robustness exercises dropping all zero income, to check whether the invalidity of Lorenz is caused by the concentration of zero incomes. We find that even when we drop all zero incomes the Lorenz curve is still invalid in about 30% of the cases, therefore the high rate of invalidity is not due to the presence of zero incomes.

It is important to notice that this paper only attempts to document the impact of a methodological choice, such as the use of group data to estimate poverty numbers.[5]

The rest of the paper is organized as follows: Section 2 provides a brief description of the methodology used to estimate global poverty numbers by the World Bank. Section 3 provides some intuition as to the pattern of cases where the parametric method fail the internal validation. Next we use the empirical CDF as an alternative to the parametric Lorenz to provide estimates of poverty and inequality using grouped data and present a few robustness checks. Section 4 concludes with a summary of the findings.

2. The global poverty estimation methodology

The approach used to estimate poverty rates at the international poverty line is different from the typical approach used to estimate poverty from household survey data. In the latter, poor households are first identified by comparing their household expenditure per capita with poverty lines, and then the percentage of the poor in the population is estimated to give the poverty headcount ratio. On the other hand, poverty rates at an international poverty line are estimated by first fitting a parametric Lorenz curve to the grouped data and then using the functional relationship between the slope of the Lorenz curve and the headcount rate of poverty. The following relation when evaluated at the point representing the proportion of the poor in the population, the slope is equal to the ratio of the international poverty line to mean household expenditure (or income) per capita (see equation 1).

$$L'(p) = z/\mu \text{ at } p = H \quad \ldots\ldots\ldots\ldots (1)$$

[5] The simulations and exercises presented in this paper are based on algorithms written by the author's and computed in the statistical package Stata. All group data results presented in this paper are based on 20 bins (sensitivity exercises were also conducted with 40, 60 and 80 bins and results were not qualitatively different). Both the algorithms and group data reported are available upon request from the author's.

where H refers to headcount rate, L' is the first derivative of the Lorenz curve, p is a cumulative proportion of population, z is a poverty line, and μ is the mean household expenditure (or income) per capita.

To obtain poverty headcount estimates from this equation, we need three key inputs: (i) the estimate of the Lorenz curve (and its slope), (ii) the international poverty line in local currency for a survey year, and (iii) mean nominal household expenditure (or income) per capita. Grouped distribution data provide the mean household expenditure per capita and are also used to estimate the Lorenz curve. The international poverty line is converted to the country's local currency using the latest private consumption PPP conversion factors. If a survey year is different from the reference year of PPP (currently 2005), the poverty line is further adjusted for inflation using CPI data. [6]

To calculate the slope of the Lorenz curve, the Lorenz curve is estimated using one of the following two functional forms – the Beta Lorenz curve and the General Quadratic (GQ) Lorenz curve. For example, if the Beta Lorenz Curve $L(p) = p - \theta p^{\gamma}(1-p)^{\delta}$ were used, three parameters $\theta, \gamma,$ and δ need to be estimated. There are four conditions which need to be satisfied by the estimated parameters for the Lorenz curve to be theoretically valid. There conditions are:

1. $L(0) = 0$
2. $L(1) = 1$
3. $L'(0^{+}) \geq 0$
4. $L''(p) \geq 0, p \in (0,1)$

The first two conditions, which may be called boundary conditions, imply that 0 and 100 percent of the population account for 0 and 100 percent of the total income or expenditure, respectively. The third and fourth conditions ensure that the Lorenz curve is monotonically increasing and convex. There is no guarantee that the estimated parameters of the Lorenz curve will satisfy all these conditions.[7]

If the Beta Lorenz curve is adopted, equation (1) becomes:

$$1 - \theta H^{\gamma}(1-H)^{\delta}\left[\frac{\gamma}{H} - \frac{\delta}{(1-H)}\right] = \frac{z}{\mu} \quad \ldots\ldots\ldots(2)$$

Equation (2) clearly indicates that if we have the three parameters of the Lorenz curve, the poverty line and the mean household expenditure (or income), we can solve this equation to get the estimate of the poverty headcount rate (H). Poverty gaps, severity of poverty, and Gini coefficients can also be calculated from specific equations derived from the Lorenz curves (see also Datt 1998).

3. Findings from 19 LAC countries from 1995-2010

3.1 Theoretical validity of the parametric Lorenz curve estimation

The two most popular parametric Lorenz curves are the Generalized Quadratic and the Beta Lorenz curve. Among the 69 surveys over 15 years for the 19 countries we see that in 38% cases the GQ fails and in 93% cases the Beta fails and in 35% both fail. Table 1 shows the list of countries for which the Lorenz curves fail at a specific point in time. There are some countries for which the Lorenz curve fails in most years reported in Table 1. Further the average inequality in countries over the whole period when the Lorenz

[6] As far as possible, the methodology uses the same CPI series as those used in the country. A requirement is that the series be available for the whole period starting form 1981. However in some cases when there can be a convincing argument made for using an alternative price index series, that may also be done (though not very often).

[7] The functional form of the beta Lorenz guarantees conditions (1) and (2), although conditions (3) and (4) need still to be tested. Regarding the GQ all four conditions need to be tested.

fails is 0.550 compared to 0.509 when at least one of the Lorenz is internally valid.[8] For the Beta Lorenz curve estimation the Gini coefficient on average is lower by almost 0.05 points for the cases where the curve is internally valid than the average inequality as measured by any of the other three methods. This coupled with the additional fact that the Beta Lorenz in invalid in 93% of the data points in this study is a cause for concern. For the GQ Lorenz, the performance is better for the estimation of Gini and for the validity of the estimation. However as can be seen in Table 2 the average inequality is higher of the cases when the GQ Lorenz is invalid than for when all the four conditions are met.

Table 1 List of countries where the Lorenz curve fails at least one condition

Circa 1995		Circa 2000		Circa 2005		Circa 2010	
GQ	Beta	GQ	Beta	GQ	Beta	GQ	Beta
Bolivia*	Bolivia	Bolivia*	Bolivia	Colombia	Bolivia	Bolivia*	Bolivia
Brazil*	Brazil	Brazil*	Brazil	Mexico*	Brazil	Brazil*	Brazil
Ecuador*	Chile	Ecuador*	Chile	Ecuador*	Chile	Mexico*	Chile
Mexico*	Ecuador	Nicaragua*	Ecuador	Nicaragua*	Ecuador	Nicaragua*	Ecuador
Nicaragua*	Mexico	Paraguay*	Guatemala	Paraguay*	Mexico	Paraguay*	Mexico
Paraguay*	Nicaragua	Chile*	Nicaragua	Chile*	Nicaragua	Guatemala*	Nicaragua
	Paraguay	Guatemala*	Paraguay		Paraguay		Paraguay
	Argentina	Peru	Argentina		Argentina		Argentina
	Costa Rica		Costa Rica		Costa Rica		Costa Rica
	El Salvador		Panama		El Salvador		Guatemala
	Panama		Venezuela		Panama		Panama
	Venezuela		Dominican Republic		Venezuela		Venezuela
	Dominican Republic		Uruguay		Dominican Republic		Dominican Republic
	Uruguay		Peru		Peru		Peru
			El Salvador		Colombia		Colombia
							Uruguay

Sources and Notes: The * indicates that for that year for the specific country, neither of the two parametric Lorenz curves are valid.

Table 2 Average inequality in cases where the Lorenz conditions fail

	Beta						GQ						Micro		
	when alteast one condition fails			when all conditions are met			when alteast one condition fails			when all conditions are met					
	Av.	Std. Dev	No. of obs	Av.	Std. Dev	No. of obs	Av.	Std. Dev	No. of obs	Av.	Std. Dev	No. of obs	Av.	Std. Dev	No. of obs
1995	0.49	0.06	15	0.59	-	1	0.55	0.05	6	0.50	0.05	10	0.52	0.05	16
2000	0.50	0.06	16	0.60	0.04	2	0.55	0.07	8	0.52	0.06	10	0.54	0.05	18
2005	0.48	0.06	15	0.57	0.06	3	0.51	0.03	7	0.53	0.06	11	0.53	0.04	18
2010	0.46	0.07	16	0.58	-	1	0.52	0.05	6	0.48	0.04	11	0.51	0.04	17

When only grouped data is available then there is a proportion of inequality on which information is not available. This is the inequality within each group. If the overall inequality is high it is more likely that there will be high levels of "within-group" inequality that is left out by the use of grouped data. A

[8] These are the average Gini Coefficients computed from micro-data.

measure of the extent of inequality that is missing from the grouped data is the ratio of within group inequality over the overall inequality in the distribution or the percentage of inequality attributable to the within component.

To capture this measure, we use the Theil inequality measure. This measure has the property that the overall observed inequality can be divided into two components- the within group inequality and the across group inequality.[9] Therefore using this measure we are able to capture the notion of what is the extent of information we lose when we use grouped data.

The Theil measure of inequality is given by :[10]

$$L = \frac{1}{N}\sum_{i=1}^{N} \ln\left(\frac{\bar{y}}{y_i}\right)$$

There the extent to which we are missing information on inequality is given by the decomposition of the Theil index into two components: the within group inequality and the across group inequality. For the decomposition assume that the total income in given by $Y = N\bar{y}$ and the income of subgroup j is given by Y_j and there are N_j members of this group and the mean income of the group is $\bar{y}_j = N_j Y_j$. The decomposition is given by:

$$L = \sum_j \frac{N_j}{N} \mathrm{L_j} + \sum_j \frac{N_j}{N} \ln\left(\frac{\bar{y}}{y_j}\right)$$

where L_j is the Theil index for the subgroup j. The first component of the above equation is the within group inequality and the second component is the across group inequality. Therefore the share of inequality that is not included when using grouped data is the second component divided by the total inequality.

Table 3 Within group proportion of total inequality

	Beta						GQ					
	when at least one condition fails			when all conditions are met			when at least one condition fails			when all conditions are met		
	Av.	Std. Dev	No. of obs	Av.	Std. Dev	No. of obs	Av.	Std. Dev	No. of obs	Av.	Std. Dev	No. of obs
1995	7%	6%	15	6%	-	1	9%	8%	7	6%	4%	10
2000	12%	10%	16	38%	37%	2	15%	10%	8	16 %	19%	10
2005	14%	8%	15	7%	2%	3	14%	5%	7	11 %	10%	11
2010	13%	13%	16	21%	-	1	21%	16%	6	9%	8%	11

Source: based on author's calculations.

[9] We cannot use the Gini for this exercise since the Gini measure is not decomposable into within group and across group inequality.

[10] When there are zero income observations in the data we have two options: one is to drop all the zero income values or two add a very small positive number all the zero value observations. Though both these have been used in the literature previously the second approach is captures the full extent of inequality.

Table 3 shows that the average values of the within group inequality in cases where all Lorenz conditions are met versus when at least one condition fails. We see that the average within group inequality is higher in cases where the Lorenz conditions fail for the GQ estimates. For the beta Lorenz since most of the cases the Lorenz fails we don't have enough observations for the case when the Beta Lorenz satisfies all conditions.

The internal validation of the Lorenz curve methodology applied to the LAC countries has yielded the following results: The Beta Lorenz Curve fails at least one of the four conditions for almost all the counties and for all four years. The GQ Lorenz curve is valid for most countries. However for the following cases neither of the two Lorenz curves is valid-Bolivia (1995, 2000, 2010) Brazil (1995, 2000, 2010), Ecuador (1995, 2000, 2005), Mexico (1995, 2005, 2010), Nicaragua (1995, 2000, 2005, 2010), Paraguay (1995, 2000, 2005, 2010), Chile (2000, 2005) and Guatemala (2000, 2010). Further the average inequality by the GQ Lorenz for the countries where GQ Lorenz is valid is lower than where there it is invalid. For the Beta Lorenz estimation in cases when the Lorenz curve is invalid we also see that the estimates of inequality are lower than those obtained from the micro-data. To see if the grouping of the data into bins for high inequality countries which by itself hides some distributional features was at least in part the reason of the invalidity we next explored if the within group inequality was systematically different between the cases where the Lorenz curve was valid versus where it failed at least one of the four conditions. We find that within group inequality as a proportion of total inequality is higher in cases where the Lorenz failed versus when it was valid indicating that it may be the case that the analytical solution is not able to account for the entire within group inequality.

The fact that the parametric Lorenz curve estimation is invalid in some cases and these cases tend to be the ones where inequality is high and the grouping of income into bins leaves a high proportion of the inequality out of the estimation leads us to explore non-parametric techniques to the estimation of poverty and inequality from grouped data. We compare how the non-parametric method compares to inequality and poverty calculations from micro data.

3.2 Comparison of parametric estimates with Microdata estimates.

We have seen that the parametric Lorenz curve fails at least one of the conditions for internal validity. However, next we compare the estimated poverty and inequality from these two techniques with the estimates from micro level data. Table 4 shows that recent years have less difference from the microdata and earlier years. Secondly, Beta Lorenz estimates for poverty have less difference from the microdata estimates than the GQ estimates. However for inequality the GQ is more accurate than Beta estimates.

Table 4 Average Difference from micro level estimates

	Poverty headcount		Poverty gap ratio		Poverty squared gap		Inequality (Gini)	
Year	GQ	Beta	GQ	Beta	GQ	Beta	GQ	Beta
1995	3.88	3.09	2.23	1.99	1.81	1.67	0.01	0.08
2000	3.95	2.67	2.7	1.77	2.13	1.37	0.01	0.08
2005	3.74	2.23	2.01	1.28	1.39	0.96	0.01	0.05
2010	1.84	1.14	0.91	0.58	0.7	0.4	0.01	0.03

Table 5 has the point estimates for 2010 for poverty and inequality. In general GQ estimates are lower than Beta estimates. However there is no general trend to if they are higher or lower than the microdata estimates.

Table 5 Poverty and inequality estimates for 2010

	Poverty headcount			Poverty gap			Poverty squared gap			Inequality (Gini)		
	Beta	GQ	Microdata	Beta	GQ	Microdata	Beta	GQ	Microdata	Beta	GQ	Microdata
Argentina	0.7	-	1.8	0.3	-	0.9	0.3	-	0.7	0.4	0.5	0.4
Bolivia	17.8	17.3	15.6	7.9	6.3	8.6	-	3.0	6.3	0.6	0.6	0.6
Brazil	5.3	4.8	6.1	3.2	1.1	3.6	-	0.4	2.9	0.5	0.5	0.5
Chile	0.8	-	1.3	1.3	-	0.7	4.5	-	0.5	0.3	0.5	0.5
Colombia	10.9	10.4	8.2	4.6	3.8	3.8	-	1.9	2.6	0.5	0.5	0.6
Costa Rica	1.4	-	3.1	0.9	-	1.8	-	-	1.5	0.4	0.5	0.5
Dominican Republic	0.1	-	2.2	0.0	-	0.5	0.0	-	0.2	0.3	0.4	0.5
Ecuador	2.5	0.6	4.6	1.3	0.0	2.1	-	0.0	1.5	0.4	0.5	0.5
Guatemala	7.1	8.7	13.5	3.1	3.2	4.7	-	1.6	2.4	0.5	0.5	0.6
Honduras	23.0	21.5	17.8	12.4	10.6	9.3	-	6.8	6.8	0.6	0.6	0.6
Mexico	1.7	-	4.0	0.7	-	1.8	-	-	1.3	0.5	0.5	0.5
Nicaragua	7.8	8.7	15.8	2.5	1.8	5.5	-	0.5	2.8	0.4	0.5	0.5
Panama	1.3	-	3.0	0.3	-	0.8	-	-	0.3	0.5	0.5	0.5
Peru	2.5	2.8	4.9	0.8	0.5	1.3	0.4	0.1	0.5	0.4	0.5	0.5
Paraguay	6.0	5.7	7.2	2.7	1.1	3.0	-	0.3	1.9	0.6	0.5	0.5
Uruguay	0.7	-	0.2	-	-	0.1	-	-	0.0	0.5	0.4	0.5
Venezuela	5.7	6.0	6.8	3.1	1.8	3.8	3.2	0.8	3.1	0.4	0.4	0.4

Source: based on author's calculations.

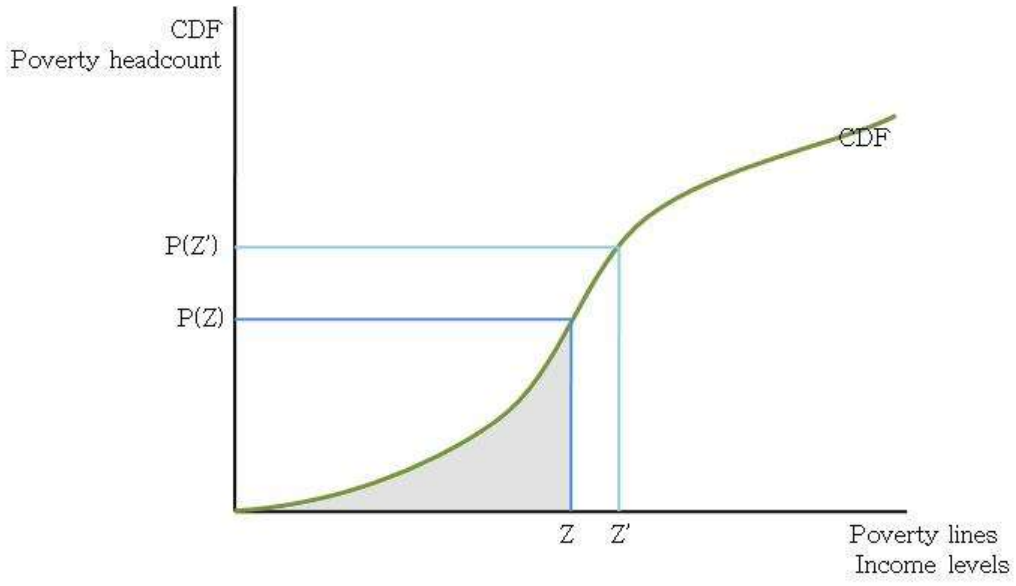
3.3 Poverty and inequality from grouped data using Empirical CDFs

In this section we use the grouped data to estimate the non-parametric cumulative distribution function which is in turn can be used to estimate both the poverty and inequality.[11] The non-parametric estimation of the CDF corresponds to an exercise of using the available data to estimate a smooth cumulative distribution function. Given the empirical CDF the method to estimate poverty is straight forward with the poverty headcount being the height of the CDF at the poverty line and the poverty gap is the area under the CDF up to the poverty line. In Figure 1 for the poverty line z the headcount is given by P(z) and the shaded area is the poverty gap estimate. Similarly we can also estimate the poverty squared gap ratio. The CDF can be easily converted into the Lorenz curve that can be used to estimate the inequality in the distribution. The empirical CDF is also known as the poverty incidence curve by Ravallion (1994).

[11] The analysis presented here is done using grouped data with 400 bins. Other bin sizes were also tested with similar results.

The use of the CDF is beneficial since it is a completely empirical technique with no parametric requirements or constraints. This allows us to estimate the CDF irrespective of the properties of the underlying distribution such as high inequality countries. [12]

Figure 1 Estimating the poverty headcount and gap measures from the CDF



Using this method to estimate poverty and inequality in 19 LAC countries we find that the poverty numbers are closer to the poverty numbers from micro-data and the inequality is also accurately estimated.

Table 6 Non-parametric and micro-data estimation results for 2010.

Country	Poverty Headcount ratio		Poverty gap		Poverty squared gap		Gini Coefficient	
	Micro-data	Nonparametric	Micro-data	Nonparametric	Micro-data	Nonparametric	Micro-data	Nonparametric
Argentina	1.836	1.751	0.906	0.888	0.696	0.671	0.445	0.445
Bolivia	15.614	15.502	8.644	8.641	6.307	6.303	0.563	0.565
Brazil	6.143	6.250	3.619	3.619	2.877	2.875	0.547	0.546
Chile	1.347	1.250	0.687	0.684	0.489	0.486	0.521	0.518
Colombia	8.161	8.000	3.783	3.785	2.618	2.620	0.559	0.555
Costa Rica	3.123	3.001	1.792	1.787	1.454	1.446	0.507	0.507
Dominican Republic	2.244	2.260	0.519	0.501	0.206	0.180	0.472	0.473
Ecuador	4.609	4.751	2.104	2.101	1.454	1.447	0.493	0.498
Guatemala	13.533	13.499	4.724	4.723	2.396	2.393	0.559	0.558
Honduras	17.822	17.751	9.299	9.299	6.833	6.833	0.569	0.570
Mexico	3.998	4.001	1.786	1.787	1.254	1.255	0.478	0.477
Nicaragua	15.804	15.751	5.481	5.479	2.830	2.827	0.524	0.524

[12] However this technique, since it is completely non-parametric does not stop the CDF from being non-monotonic, though empirically this is hard to come across.

Panama	2.994	3.001	0.801	0.802	0.326	0.319	0.520	0.521
Peru	4.908	5.001	1.308	1.311	0.541	0.543	0.482	0.480
Paraguay	7.160	6.999	3.022	3.019	1.917	1.917	0.524	0.526
Uruguay	0.199	0.251	0.068	0.057	0.042	0.013	0.453	0.453
Venezuela	6.840	6.750	3.819	3.818	3.080	3.078	0.448	0.447

Source: based on author's calculations.

Comparing the estimation results using the empirical CDF we see that the estimates are close to the numbers from micro data (see Table 6). For the poverty headcount the estimates are always within one percentage point of the micro-data estimates. For the Gini coefficient as well the estimates are accurate up to two decimal places with some difference being seen at the third decimal place. This is also true for the other three years (see Table 7).

Table 7 Average difference of estimates of Nonparametric estimates from micro-data estimates.

Year	Headcount	Poverty gap	Poverty squared gap	Gini
1995	0.13	0.01	0	0
2000	0.12	0	0	0
2005	0.12	0.01	0.01	0
2010	0.09	0.02	0.02	0

Source: based on author's calculations.

The performance of the empirical CDF is better than the performance of the parametric Lorenz curves.

Table 8 gives the poverty headcount ratios for 2010 using the four different methods. Two points are evident from here. The Non-parametric estimates are closest to those obtained from micro data. Secondly, the difference is on average larger for the GQ estimates that are more likely to be theoretically valid.

Table 8 Poverty headcount ratio for 2010 using all four methods.

Country	Micro-data	Nonparametric	Beta Lorenz	GQ Lorenz
Argentina	1.836	1.751	0.700	-
Bolivia	15.614	15.502	17.770	17.255
Brazil	6.143	6.250	5.313	4.794
Chile	1.347	1.250	0.820	-
Colombia	8.161	8.000	10.925	10.352
Costa Rica	3.123	3.001	1.394	-
Dominican Republic	2.244	2.260	0.060	-
Ecuador	4.609	4.751	2.471	0.621
Guatemala	13.533	13.499	7.078	8.663
Honduras	17.822	17.751	23.044	21.537
Mexico	3.998	4.001	1.699	-
Nicaragua	15.804	15.751	7.768	8.706
Panama	2.994	3.001	1.291	-
Peru	4.908	5.001	2.455	2.800
Paraguay	7.160	6.999	5.972	5.698
Uruguay	0.199	0.251	0.660	-
Venezuela	6.840	6.750	5.659	6.021

Source: based on author's calculations.

3.4 Robustness exercises

We have conducted two sets of robustness exercises. The above analysis was done for the $1.25 a day per person poverty line. However there are other poverty lines that are also popular in the literature. We repeated the exercises above for four other poverty lines- $ 2.0, $4, $10 and $50. In general, the findings discussed above hold with the average difference being larger for the GQ and Beta Lorenz than for the non-parametric estimation compared to the estimates from the micro-data (See Table 9).

Table 9 Average Bias in estimation using various poverty lines.

	GQ Lorenz Curve					Beta Lorenz Curve					Nonparametric estimation				
Poverty Line / Year	$1.25	$2	$4	$10	$50	$1.25	$2	$4	$10	$50	$1.25	$2	$4	$10	$50
1995	3.88	5.35	6.41	5.88	0.88	3.09	3.76	5.45	7.91	1.45	0.13	0.07	0.05	0.06	0.09
2000	3.95	5.16	8.23	6.05	0.9	2.67	5.11	8.15	7.05	1.2	0.12	0	0.04	0.04	0.06
2005	3.74	3.68	5.26	2.82	0.33	2.23	3.29	4.98	4.07	0.7	0.12	0.11	0.04	0.04	0.1
2010	1.84	3.86	3.32	2.32	0.03	1.14	2.54	3.09	3.2	0.64	0.09	0.16	0.08	0.07	0.02

Source: based on author's calculations.

The second set of robustness exercises related to treatment of the income variable used in this paper. Most countries in the LAC region use income as the primary welfare for poverty measurement and therefore in

this paper we are working with income data. Use of income rather than consumption brings two features into the data which do not exist for consumption data[13]:

1. Zero income is actually possible and is present in a number of cases. This is different from consumption since zero consumption is not a physical possibility.
2. There are many observations with the same income values (happens especially at the minimum wage).

The second of the peculiarities of income data, which is clustering at certain values, is not much of a concern if we are working with grouped data. The clustering at particular values is hidden behind the envelope of the grouping of the data and therefore cannot directly effect the parametric estimation from grouped data.

On the first point, to see whether the zeros introduce an error into the measurement the analysis is done without taking the zero values. Of course this grossly underestimates the actual distribution of incomes in the respective countries since we are leaving out the bottom end of the distribution. However even when we repeat the parametric Lorenz analysis without the zero the failure rates are very similar to those with the zeros with the Beta Lorenz failing in 90% (one less case of failure) of the cases and the GQ Lorenz failing in 36% (one extra case of failure) of the cases Error! Reference source not found.. **Table 10 shows that in general the estimates of Gini are gain lower than those obtained from Micro data. However there is a less clear pattern when comparing inequality levels when conditions are met versus when the Lorenz is invalid. Further**

[13] The author's have also replicated this validation exericse using consumption data from 28 Europe and Central Asia countries. The results of the internal validity of the estimates are better but external validation issues remained. Results are available upon request.

Table 11 shows that the estimates poverty headcount are most accurate using the non-parametric approach even when the zeros are removed from the sample.

Table 10 Average inequality in cases where the Lorenz conditions fail (without the zero incomes)

	Beta Lorenz		GQ Lorenz		Micro data
	At least one condition is violated	All conditions are met	At least one condition is violated	All conditions are met	
1995	0.479	0.588	0.468	0.498	0.518
2000	0.491	0.588	0.549	0.522	0.532
2005	0.471	0.564	0.433	0.533	0.522
2010	0.455	0.568	0.520	0.482	0.505
Overall	0.474	0.575	0.495	0.509	0.519

Source and Notes: based on author's calculations. The cells give the average value of the Gini coefficient.

Table 11 Poverty headcount rates (at $1.25 per day) without the zeros.

	Beta Lorenz	GQ Lorenz	Non Parametric	Micro-data
Argentina	0.33	0	1.25	1.35
Bolivia	16.3	15.39	14.25	14.24
Brazil	3.25	1.543	4.25	4.17
Chile	0.69	0	1.00	1.11
Colombia	9.91	9.13	7.00	7.15
Costa Rica	0.6	0	2.00	2.10
Dominican Republic	0.05	0	2.25	2.22
Ecuador	1.72	0	4.00	3.91
Guatemala	6.99	8.56	13.50	13.46
Honduras	20.53	18.20	14.75	14.69
Mexico	0.89	0	3.25	3.31
Nicaragua	7.52	8.41	15.76	15.64
Panama	1.24	0	3.00	2.95
Peru	2.44	2.79	5.00	4.91
Paraguay	5.36	4.82	6.49	6.68
Uruguay	0.7	0	0.25	0.18
Venezuela	2.98	2.51	4.75	4.64

Source: based on author's calculations.

Table 10-

Table 11 show that use of data which has clustering around zero is not the reason why we see the failure in the parametric Lorenz curves or the reason for poorer estimates from parametric Lorenz curves than from the non-parametric Lorenz curves.

4. Summary and Discussion

In this paper we use data from 19 countries in the Latin American and Caribbean region at four points in time from 1995-2010 to do a sensitivity analysis of poverty estimations based on grouped data. The decision to use grouped data is in most cases justified. Particular when countries do not provide access to anonymized records of the microdata, or when records have been lost, and the only available information is the one reported in the distributional tables from printed publications, or for methodological comparability of the recent microdata estimates with the early grouped data estimates. However, it is also important to better understand and document the some of the limitation of this approach, specially as we focus the policy debate on the poorest of the poor (which often are at the very far lower tail of the distribution) or those belonging to the middle class, which will required to analyze higher moments of the distribution.

We find that in approximately 30% of the cases the estimated parametric Lorenz fails at least one of the four conditions necessary for its theoretically validity. Further the countries and years when the Lorenz fails are in general high inequality countries. A second observation is that when the Lorenz curve has failed, the estimates of poverty and inequality have larger biases than when it has not failed.

Though shifting to use of microdata for estimation of poverty and inequality may be desirable, data access issues still restrict us to the use of grouped data. In the background of the parametric Lorenz failing its conditions in about one-third of the cases we re-estimate poverty and inequality using non-parametric techniques.

Since we also have access to the household data for these 19 countries we are able to compare the Lorenz curve estimates of poverty and inequality with those from household data. We find that on average there was a 3 percentage point difference in poverty headcount rates for the year 1995 and over 1 percentage point for 2010, more important this difference is not systematic across countries, choice of poverty line, nor poverty indicator (i.e. FGT0, FGT1, and FGT2). However the non-parametric estimates have biases that are lower than one percentage point in all cases. Therefore, the non-parametric estimation gives us more accurate estimates and does not require any restriction on the shape of the income distribution. This is an important factor in cases when the inherent inequality is high in the country.

The availability of better technology for non-parametric estimation coupled with increased access to better information on income and consumption expenditure alludes to the need to rethink the methodology used for global poverty estimation. This paper merely alludes to the shortfalls of parametric estimation and the need to revisit non-parametric estimation as a potential alternative. As the developing world grows and in most cases with greater inequality the restrictions imposed by parametric methods may become a larger hindrance to the accurate estimation of poverty and inequality and therefore there is a need to explore technologies which can capture the changing distribution of income and consumption expenditure.

Statistical agencies also have an important role to play in this debate, strengthening institutional capacity could would enable statistical agencies to make available household surveys properly anonimized and with clear terms of use. Initiatives such as the Accelerated Data Program (ADP) and the Statistical Disclosure Control (SDCMicro), led by Paris 21 and the World Bank, are pioneers in this agenda (see link for more details).

The result of this paper have implications to the Post-MDG discussions also. We illustrate how the choice of indicators and the underlying data used in its computation can be crucial for our ability to understand both levels and trends of indices. This once again reflects the shared responsibility between governments and the international community to create this essential public good.

One of the greatest risks currently facing policy makers and governments is inability to deliver results and to successfully assure that prosperity is shared within and across countries. However, our ability to deliver results is directly linked to the capacity to demonstrate results, and the later is a consequence of the measurement choices, data availability, and data quality. This paper shows that all three matter.

References

Ackland, R., Dowrick, S., & Freyens, B. (2012). Measuring global poverty: Why PPPs matter. *Review of Economics and Statistics*. http://www.mitpressjournals.org/doi/abs/10.1162/REST_a_00294

Atkinson, A.B. and Brandolini, A. (2001). "Promise and Pitfalls in the Use of "Secondary" Data-Sets: Income Inequality in OECD Countries as a Case Study", *Journal of Economic Literature*, Vol. 39, No. 3, pp. 771-799.

Chen, S. and Ravallion, M., 2010, "The developing world is poorer than we thought, but no less successful in the fight against poverty," *Quarterly Journal of Economics*, Vol. 125(4), pp. 1577- 1625.

Datt, G. (1998) "Computational Tools for Poverty Measurement and Analysis", IFPRI Food Consumption and Nutrition Division Discussion Paper No. 50 (Washington: International Food Policy Research Institute).

Davies and Shorrocks 1989

Deaton, A., 2003, "Household Surveys, Consumption, and the Measurement of Poverty," *Economic Systems Research*, Vol. 15, pp. 135–159.

Dhinge, S. and Miniou, C. 2011. "Global Poverty Estimates: A Sensitivity Analysis". *World Development*, Volume 44, April 2013, Pages 1–13.

Ferreira, Francisco H.G., Julian Messina, Jamele Rigolini, Luis-Felipe López-Calva, Maria Ana Lugo, and Renos Vakis, 2013. "Economic Mobility and the Rise of the Latin American Middle Class," World Bank Publications, The World Bank, number 11858.

Kakwani, N. C. (1980a) "On A Class of Poverty Measures", *Econometrica*, Vol. 48, Issue 2, pp 437−446.

Kakwani, N. C. (1980b) "Functional Forms for Estimating the Lorenz Curve: A Reply", *Econometrica*, Vol. 48, Issue 4, pp 1063−1064.

Karshenas, M (2003). "Global Poverty: National Accounts Based versus Survey Based Estimates", *Development and Change*, Vol. 34. Issue 4, pp 683-712.

Minoiu, C. and Reddy, S., 2008, Kernel Density Estimation Based on Grouped Data: The Case of Poverty Assessment," IMF Working Papers No. 183 (Washington: International Monetary Fund).

Minoiu, C. and Reddy, S., 2009, "Estimating poverty and inequality from grouped data: How well do parametric methods perform?" *Journal of Income Distribution*, Vol. 18(2), pp. 160–178.

PovcalNet. An Online Poverty Analysis Tool (Washington: The World Bank Group). Available on: http://go.worldbank.o rg/NT2A1XUWP0

Ravallion, Martin, Shaohua Chen, Prem Sangraula (2008) Dollar a Day Revisited Policy Research Working Paper 4620. World Bank: Washington, DC.

Reddy, S.G. 2009. "The Emperor's New Suit: Global Poverty Estimates Reappraised", DESA Working Paper No. 79. http://www.un.org/esa/desa/papers/2009/wp79_2009.pdf

United Nations, 2000, Millennium summit goals (New York: United Nations)

Villasenor, J. and Arnold, B., 1989, "Elliptical Lorenz curves," *Journal of Econometrics*, Vol. 40, pp. 327–338.

World Bank Group. 2013. World Bank Group Strategy. Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/16095 License: Attribution-NonCommercial-NoDerivs 3.0 Unported.