Estimation of Opportunity Inequality in Brazil using Nonparametric Local Logistic Regression[*]

Erik Alencar de Figueiredo
Department of Economics
Universidade Federal do Rio Grande do Norte, Brazil
Phone: +55 84 3642-1675
E-mail: eafigueiredo@gmail.com

Flávio Augusto Ziegelmann
Department of Statistics
Graduate Program in Economics
Universidade Federal do Rio Grande do Sul, Brazil

[*] The authors thank professors Fernando A. Veloso and Sérgio Ferreira. We take full responsibility for any errors or omissions.

Estimation of Opportunity Inequality in Brazil using Nonparametric Local Logistic Regression

Abstract: This paper measured opportunity inequality in Brazil by combining a series of theoretical and empirical tools. The database was built using a two-sample instrumental variable (TSIV), developed by Angrist & Krueger (1992). After that, the axiomatic approach put forward by O'Neill et al. (2001) was used, in which the estimation of children's income distribution function is conditional on their parents' wages. The inference process was based on nonparametric local logistic regression. The results indicate that Brazil has a high level of opportunity inequality. In other words, in the context of intergenerational mobility, those whose parents belong to lower income strata have to expend greater effort in order to attain a certain income level.

Keywords: Intergenerational mobility; opportunity inequality; Nonparametric local logistic regression.

1. Introduction

Income inequality is often regarded as undesirable. Theoretical approaches, such as the one used by Atkinson (1970), posit that, under some assumptions, a lower inequality level produces a higher social optimum. Philosophical theories of justice usually preach about the social benefits of a more equal income distribution through the maximization of a poorer individual's utility (see, among others, Ralws (1971)). In this regard, outlooks on equality rely on greater political and popular appeal, which call for the adoption of income redistribution policies.

However, the relation between inequality and welfare might not be so strong in some societies. Recently, Alesina et al. (2001) have argued about the impact of inequality on the welfare of European and U.S. citizens. Their conclusions point out that income inequality has a negative effect on the happiness of European people, but not on the happiness of the U.S. population, since Americans associate poverty with inefficiency, whereas Europeans reckon it as "bad luck." Such heterogeneity shows that income distribution policies may or may not be socially desirable.

The latter argument, if taken to the extreme, raises a question: is income inequality a social problem? The answer necessarily encompasses a conceptual expansion, leading to the notion of opportunity inequality. Conventional approaches are usually based on the inequality of results, which is not totally satisfactory given that income differences may originate from the different opportunity levels people find during their lifetime.[2] Therefore, factors related to individual characteristics, such as social background, genetic makeup, race, sex, place of origin, among others, end up influencing inequality.[3]

Based on these facts, this study aims to measure opportunity inequality in Brazil. The choice of Brazil as subject of study in this case is due to at least two aspects. First, because it has one of the worst individual income distributions of the world.[4] Second, due to the idea that Brazilian society has a high social debt, which has been the basis of income redistribution and compensation policies.[5] Nevertheless, the formulation of measures to fight opportunity inequality runs into a major difficulty, since these indicators are not calculated by taking into

[2] See Fleurbaey (1995) and Roemer (1996).

[3] The literature on opportunity inequality has grown in the latest years, including the following: Van der gaer (1993), Ok & Kranich (1998), Bossert et al. (1999), O'Neill et al. (2001), among others. Roughly speaking, the studies conclude that income inequalities will only become a social problem if they result from opportunity inequality.

[4] See United Nations Development Program (2006).

[5] From the second half of the 1990s, federal governments have adopted a series of income redistribution policies, classified as "affirmative policies." Among them, one may cite the system for preferred admission for racial minorities (blacks and indians) at public universities, food grants and school attendance programs.

account earnings inequality only, but also the inequality conditional on the opportunity set for every individual.[6]

In order to circumvent this problem, the study uses the axiomatic approach put forward by O’Neill et al. (2001). This framework is based on the concept of intergenerational mobility, allowing for the construction of the opportunity set based on an individual’s parental income. This enables calculating children’s income indicators conditional on parental income. To do that, it is used a nonparametric statistical method based on local logistic estimation of the income cumulative distribution function of individuals conditional on parental income.

The remainder of the paper is organized as follows. Section 2 deals with the concept of opportunity equality, establishing the axioms for the construction of the opportunity set. Section 3 describes the statistical methodology used. The major empirical results are shown in Section 4. To conclude, Section 5 presents the final remarks.

2. Opportunity Equality

The concept of opportunity set can be easily captured from the observation of two situations. First, suppose two individuals, **A** and **B**, had identical opportunities, went to the same school, had a similar family environment, and belong to the same race and sex. After finishing their studies, they pursued the same profession. However, individual **A** strives in his job, whereas **B** would rather set leisure as a priority. These choices lead to a situation in which **A** has a higher income level than **B**. Then, income inequality between them is not a social problem, as it results from a rational choice.

Following the same line of thought, suppose new agents: **C** and **D**. However, postulate that **C** has a wider set of opportunities, produced by a more favorable social situation. Therefore, the opportunities for **D** constitute a subset of the opportunities available for **C**, such that even if he shows similar effort, **C** will probably be richer than **D**. In this case, inequality between them does not arise from aspects that are related to choice, but rather from factors that are beyond an individual’s control.

In brief, some people, due to primitive characteristics, have a poorer opportunity set than others. These primitive characteristics can stem from the social base of racial lineage, of sexual orientation, among others. Henceforth they will be referred to as non-responsibility characteristics. The remaining ones will be designated as responsibility characteristics.

As previously put, the construction of the opportunity set is not an easy task. The key issue is how to define it. Thereafter, resorting to theories of equality, one may ask the following: after detecting opportunity inequality between individuals, should the underprivileged ones be compensated for such inequality? This second topic deserves special attention, but it is not within the scope of the present paper. The debates related to compensation are summarized in Fleurbaey & Maniquet (2005).

Given that compensatory actions are not considered, one should define the opportunity set. Following O’Neill et al. (2001), it is considered that the opportunity set for an individual, S_x, is determined by his non-responsibility characteristics, x. The fact is that the results, or the income level of an individual, will depend on his effort, or responsibility characteristics, conditional on the opportunity set. These results may be summarized by $z = y[e, x]$, where z is the utility or the income throughout life and e is a variable that represents effort.

Assume that the distribution function of e is continuous. Some axioms should also be postulated.

[6] Bossert et al. (1999) demonstrate that this procedure is not an easy task.

SINC (Strictly Increasing): $y[e,x]$ is strictly increasing in e.

The interpretation of this axiom is straightforward: a higher effort will give rise to a higher level of utility. By defining $F_z^*\left(z\,|\,x\right)$ and $F_e^*\left(e\,|\,x\right)$ as accumulated distribution functions of z and e conditional on x, respectively, then **SINC** can be written as:

$$F_z^*\left(y[e,x]\,|\,x\right)=F_e^*\left(e\,|\,x\right). \qquad [2.1]$$

i.e., the level of effort of an individual conditional on a class of the opportunity set will be lower than an α-th percentile of the distribution of this effort if and only if the result is smaller than an α-th percentile of the distribution of this result conditional on the same class of the opportunity set.

IND (Independence): $F_e^*\left(e\,|\,x\right)$ is independent from x.

This means that no differences are assumed in the effort distribution functions between different types of individuals, i.e., between individuals with different non-responsibility characteristics. Admitting that such assumption is not valid would be the same as stating that an individual would exhibit different levels of effort depending on his (non-responsibility) characteristics.

Finally, (2.1) and IND provide the following axiom:

RIA (Roemer's Identification Axiom): $F_z^*\left(y[e',x_1]\,|\,x_1\right)=F_z^*\left(y[e'',x_2]\,|\,x_2\right)\Rightarrow e'=e''$.

This axiom informs that two people with different opportunity levels, but with the same distribution percentile within their type, have the same level of effort.

That being said, define $\pi_i=F_z^*\left(z\,|\,x\right)$ as the accumulated distribution of result (income level) z conditional on non-responsibility characteristics, x. Assume that this function is strictly increasing in z. Thus, $F_z^{*-1}\left(\pi\,|\,x\right)$ expresses the income level obtained by an individual of type x who belongs to percentile 100*π. On account of RIA, observing the value of $F_z^{*-1}\left(\pi\,|\,x\right)$ will be the same as observing the value of $y[e,x]$.

Therefore, $F_z^{*-1}\left(\pi\,|\,x\right)$ will provide information on responsibility characteristics and non-responsibility characteristics of individuals. O'Neill et al. (2001) assert that it allows us to draw income as a function of $\pi\in[0,1]$ for different values of *x*. The opportunity set for a particular type *x* is given by the outcomes someone of type *x* can obtain by varying his responsibility characteristic *e* or π. This way, the opportunity set of individuals of type *x* will be

$$S_x=\left\{(z,\pi)\in(R^+\times[0,1])\,|\,z=F_z^{*-1}\left(\pi\,|\,x\right)\right\}, \qquad [2.2]$$

where R^+ is the set of non-negative real numbers. The visual inspection of set (2.2) will show the level of opportunity inequality of different types of individuals and also to what extent different options or levels of effort yield different results (income).

Finally, one should determine the elements of x. A possibility is to consider the context of intergenerational mobility, as dealt with in Van de gaer et al. (1998) and O’Neill et al. (2001). Thus, z will represent the children’s income and x the parental income. The estimation of children’s income conditional on parental income will be made using a nonparametric method, namely, the nonparametric local logistic regression, which will be described in the subsequent section.

3. Estimation of the Conditional Distribution Function using Local Logistic Regression

Nonparametric methods are quite flexible for the estimation of unknown curves. In this context, kernel smoothing tools provide estimators that are both flexible and intuitive, besides having good asymptotic properties. Silverman (1986), Wand & Jones (1995) and Fan & Gijbels (1996) comprehensively describe these techniques, while Simonoff (1996) and Bowman & Azzalini (1997) introduce a more intuitive and applied approach.

In this paper, as well as in a wide variety of statistical problems, the objective curve to be estimated is the cumulative conditional distribution function. Consider, for instance, the estimation of the quantile function of Z given X, using a random sample with pairs $\{(X_1, Z_1),...,(X_n, Z_n)\}$. Yu & Jones (1997) suggest using the "double kernel" approach for the local linear regression. O’Neill et al. (2001) use a multiple-step procedure, whereby they first estimate the joint and marginal densities and then the joint, marginal and conditional distributions.

A drawback of such methods is that they do not guarantee that the estimated functions are non-decreasing and restricted to the interval [0,1]. In this regard, Hall et al. (1999) propose an estimator that satisfies these conditions. This estimator is known as local logistic estimator of the conditional distribution function. The idea is to create an indicator response variable which assumes only values 0 or 1, and locally regressing it via a logistic function against the covariate of interest. Note that this type of logistic estimator is a particular case of a broader class, which we may designate as functional class. Ziegelmann (2002) uses this class of estimators to guarantee a non-negative estimator for conditional variance.

Let $\{(X_i, Z_i)\}$ be a random two-dimensional sample of size n. Our aim is to estimate the conditional distribution function $\pi(z\,|\,x) \equiv P(Z_i \le z\,|\,X_i = x)$. Note that, if we write $W_i = I(Z_i \le z)$ where $I(\square)$ is the indicator function, then

$$E(W_i\,|\,X_i = x) = \pi(z\,|\,x).$$

Now assume that, for a fixed z, $\pi(z\,|\,x)$ has $r-1$ continuous derivatives. Therefore, we will choose the following logistic form to approximate $\pi(z\,|\,x)$ locally (i.e., in the neighborhood of x)

$$L(x,\theta) = \frac{\exp\{p(u-x,\theta)\}}{1+\exp\{p(u-x,\theta)\}} \quad , \qquad [3.1]$$

where $p(u-x,\theta) = \theta_0 + \theta_1(u-x) + ... + \theta_{r-1}(u-x)^{r-1}$ for u in the neighborhood of x.

This way, adjusting this model locally to the data of indicator function W_i will give us the estimator $\hat{\pi}(z\,|\,x) \equiv \exp\{p(0,\hat{\theta})\}/\{1+\exp\{p(0,\hat{\theta})\}\} = \exp\{\hat{\theta}_0\}/\{1/\exp\{\hat{\theta}_0\}\}$, where $\hat{\theta}_0$ denotes the value of θ_0 which minimizes

$$\sum_{1}^{n}\left\{W_i - \frac{\exp\{p(X_i - x,\theta)\}}{1+\exp\{p(X_i - x,\theta)\}}\right\}^2 K_h(X_i - x) \quad , \qquad [3.2]$$

where $K_h(x) = (1/h)K(x/h)$, in which $K(x)$ is a kernel function (traditionally a symmetric probability density function around 0) and h is the smoothing parameter, or bandwidth, which controls the level of complexity of the estimated curve. Hall et al. (1999) derive the asymptotic properties of this estimator in the broader context of time series, where the case of an i.i.d. sample is a particular case.

In this paper $K(\cdot)$ will be the probability density function of a standard normal distribution, while the smoothing parameter h will be chosen by cross-validation, i.e., so as to minimize something like an out-of-sample "forecast" error (see Fan & Gijbels, 1996, for details).

4. Results

This section presents the major results of this study. First, the nature and manipulation of data are discussed, and later, the accumulated nonparametric conditional distribution functions described in Section 3 are estimated.

4.1. Data

The data used in this study were obtained from the Brazilian National Household Survey (PNAD). This survey has been conducted by the Brazilian Institute of Geography and Statistics (IBGE) since the late 1960s and consists of a basic questionnaire that includes questions about household and personal characteristics, such as family size, household income, educational background, number of working hours, personal income, among others. In some years, some special characteristics are investigated and then summarized in supplemental issues. These special characteristics include, for instance, health, food safety, child labor and social mobility.

The latest supplement contains data on individuals and their parents, and was applied in years 1973, 1976, 1982, 1988 and 1996. More specifically, it gives information on education, schooling and parent's and children's occupation. However, the major problem with this survey is the lack of a "father's wage" variable. That is, in a study of intergenerational mobility, it is not possible to regress children's income on parental income.

Due to the lack of this information, it is necessary to use a method that allows building an approximation for parental income. This is provided by the two-sample instrumental variable (TSIV), designed by Angrist & Krueger (1992). In summary, one seeks to estimate parental income (predicted wage) based on characteristics such as educational and professional background.

To do that, we considered two samples, called respectively parent's sample and children's sample. Both contain information on male household heads aged 25 to 65 years, with a workload of 40 hours or more per week in all jobs, who live in the urban zone.

The parent's sample was constructed using the 1976 PNAD data. These data provide information about (hourly) wage,[7] years of schooling and occupation. The first step was to create dummy variables for occupation and schooling. The 927 occupations described in the survey were classified into six categories, according to Pastore & Silva (1999). Education was

[7] Earnings from all jobs divided by the number of working hours.

divided into seven categories: no education (less than one year of schooling); incomplete lower elementary education (from 1 to 3 years of schooling); complete lower elementary education (4 years of schooling); incomplete upper elementary education (from 5 to 7 years of schooling); complete upper elementary education (8 years of schooling); incomplete or complete high school education (9 to 11 years of schooling) and incomplete or complete college education (over 11 years of schooling).

The ultimate goal of this stage is to perform the regression of the logarithm of wage on dummy variables for education and occupation. Given that some studies, including those by Menezes Filho et al. (2003), Ferreira (2003) and Ferreira and Veloso (2006), highlight the change in the school premium in several cohorts throughout time in Brazil, it is important to include dummies for year of birth and for the relationship between occupation and schooling. Finally, the synthetic profile of parents was built using a wage equation.

The second stage consisted in collecting information on male household heads aged 25 to 65 years, who worked 40 hours or more a week in all jobs, and lived in the urban zone, by using the 1996 PNAD supplement. This database is referred to as “children’s sample.” This sample includes information on children’s income and on parent’s education and occupation. The wages for the “synthetic” fathers were based on the coefficients estimated in the first stage (in the parent’s sample). Thus, we obtained the two variables of interest of this study: children’s wage and parent’s predicted wage.

4.2. Results

As previously seen, the opportunity set was estimated using the income level estimated for the parents. That is, the context of intergenerational mobility was considered. In a recent study, Ferreira and Veloso (2006) showed that Brazil has low intergenerational income mobility. In other words, parental income is a determining factor for children’s income. In brief, elasticities ranged between 0.58 and 0.73, depending on the tools used in the regressions.

Actually, these results indicate high opportunity inequality in Brazil. However, to confirm this assumption, it is necessary to take into account the “children’s effort” variable. Table 1 shows the first sign of opportunity inequality for Brazil. This table summarizes the results for the accumulated distribution conditional on children’s income vis-à-vis the parental income. The values were calculated using the tools described in Section 3. The analysis is simple; consider the first column of results, and in this case, parents belong to the fifth percentile, i.e., the poorest 5%. Therefore, the estimated probability for children belonging to the poorest 25%, for example, amounts to approximately 0.43. In the same column, note that the estimated probability for the children finding themselves below the 95th percentile amounts to nearly 0.99. The differences between these two probabilities become more evident when we compare them with the results shown in the last column, where parents belong to the 95th percentile. Note that, under this circumstance, the estimated probability for children being among the poorest 5% is virtually equal to zero. In turn, the estimated probability for them to be on the top of the distribution (the richest 5%) is 0.242 (1-0.758).

Insert Table 1 here

In brief, there is some strong evidence that confirms the major conclusion drawn by Ferreira & Veloso (2006): parental income level seems to be a crucial factor for the determination of children’s income. Nevertheless, how much of this behavior is due to non-responsibility characteristics? The answer to this question is provided in Figure 1, where we have the estimates for the Brazilian opportunity set. This figure shows the results for the estimated probabilities of children having the same income as or a lower income than a given relative income, since their parents belong to the 5th, 50th and 95th percentiles. In this case, we set the parent's percentile, say, at 5%, and observe the accumulated probabilities of their children in relation to their relative income. One of the results from Section 2, RIA (Roemer's Identification Axiom), indicates that two people with different levels of opportunity, but belonging to the same distribution percentile within their type, exhibit the same effort level. This axiom allows us to establish an income level where individuals with different opportunity sets can be assessed. The rationale behind it is simple: what is the level of average effort each individual should have in order to get to this distribution percentile.

Insert Figure 1 here

Figure 1 indicates that the higher the parental income level, the lower the effort level children have to make to obtain a relative income equal to one. Therefore, if a parent belongs to the 95th percentile, his child must have an effort level of approximately 0.19 to reach the average income. On the other hand, if the parent is poor (5th percentile), his child should have an effort level of approximately 0.70 to obtain a relative income equal to one. In other words, between the extremes of the opportunity set, effort has to be more than three times greater in order to obtain the same income level.

In terms of the significance of the differences between cumulative conditional distribution functions, we used the Kolmogorov-Smirnov test for two independent samples. In this regard, we then provide a justification for its use. Although the samples that originated the estimates were actually the same for all curves (a priori not characterizing independence), only one part of the sample is used for the estimation of each one of the curves (local estimate). This means that, depending on the parent's percentile, only parental income values close to the percentile at issue are used for the estimation conditional on that percentile. Considering the smoothing parameters obtained by cross-validation, there is no overlap of observations taken into account in the three curves of Figure 1. This way, we have the equivalent to independent samples, with approximately 400 observations used for each estimated curve. Thus, using the traditional Kolmogorov-Smirnov test for the two pairs of curves (5th percentile against the 50th percentile and the 50th percentile against the 95th percentile), we may conclude that the curves should be different at a 5% significance level.

In summary, the results show the high level of opportunity inequality in Brazil. When compared with international results, more specifically, with the U.S. case, investigated by O'Neill et al. (2001), there is a remarkable difference between indicators. In the case of the USA, the difference in effort between individuals belonging to the 25th and 75th percentiles amounts to approximately 56%. Brazilian data, shown in Table 1, reveal a 112% difference ([0.679/0.321]-1).

5. Final Remarks

This paper measured opportunity inequality in Brazil. An approach based on intergenerational mobility was used as the main factor for the construction of the opportunity set. The database could only be built due to the use of the two-sample instrumental variable (TSIV), developed by Angrist & Krueger (1992). After the construction of the vectors using the children's and parental wages, a nonparametric local logistic regression was used to estimate the children's income distribution conditional on parental wage.

This information, along with the axiomatic approach described in O'Neill et al. (2001), allowed inferring on the level of effort put in by individuals conditional on the level of opportunity. Then, it was detected that Brazil seems to have a high level of opportunity inequality. In other words, those individuals whose parents belong to lower income strata have to expend greater effort in order to attain a certain income level. However, the detection of a high level of opportunity inequality does not put an end to the debate. Therefore, the next step, left as a suggestion for further research, would be to discuss the necessity to adopt compensatory policies.

References

Alesina, A., Di Tella, R. & MacCulloch, R. (2001). *Inequality and happiness*: are Europeans and Americans different? NBER: Working Paper W8198.

Angrist, J. & Krueger, A. (1992). The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of American Statistical Association,* 87:238-336.

Atkinson, A. (1970). On the measurement of inequality. *Journal of Economic Theory*, 2: 244-263.

Bossert, W., Fleurbaey, M. & Van de gaer, D. (1999). Responsibility, talent and compensation: a second best analysis. *Review of Economic Design*, 4: 35-55.

Bowman, A. & Azzalini, A. (1997). *Applied smoothing techniques for data analysis*. New York: Oxford University Press.

Fan, J. & Gijbels, I. (1996). *Local polynomial modeling and its applications*. London: Chapman and Hall.

Ferreira, S. (2003). *Skinning the cat: education distribution, changes in the school premium and earnings inequality*. Anais do Encontro Nacional de Economia da ANPEC.

Ferreira, S. & Veloso, F. (2006). Intergenerational mobility of wages in Brazil. *Brazilian Review of Econometrics*, 26:181-211.

Fleurbaey, M. (1995). Equal opportunity or equal social outcome? *Economics and Philosophy*, 11: 25-55.

Fleurbaey, M. & Maniquet, F. (2005). Compensation and Responsibility. Texto não publicado.

Hall, P., Holff, R. & Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94: 154-163.

Menezes Filho, N., Fernandes, R. & Picchetti, P. (2006). Rising human capital but constant inequality: the education composition effect in Brazil. *Revista Brasileira de Economia*, 60:407-424.

Ok, E. & Kranich, L. (1998). The measurement of opportunity inequality: a cardinal-based approach. *Social Choice and Welfare*, 15: 263-287.

O'Neill, B., Sweetman, D. & Van de gaer, D. (2001). Equality of opportunity and kernel density estimation: an application to intergenerational mobility. In: Fomby, T. & Hill, C.

(eds.). *Applying kernel and nonparametric estimation to economic topics. Advances in Econometrics*, vol. 14, Stanford, Conn: JAI Press.

Pastore, J. & Silva, N. (1999). *Mobilidade social no Brasil*. São Paulo: Makron Books.

Rawls, J. (1971). *A theory of justice*. Oxford: Oxford University Press.

Roemer, J. (1996). *Theories of distributive justice*. Harvard University Press, Cambridge.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Simonoff, J. (1996). *Smoothing methods in statistics*. New York: Springer-Verlag.

United Nations Development Program. *Human development report*, New York, 2006.

Van de gaer, D. (1993). *Investment in human capital and equality of opportunity*. Doctoral dissertation, K. U. Leuven.

Van de gaer, D., Schokkaert, E. & Martinez, M. (1998). *Measuring intergenerational mobility and equality of opportunity*. Economics Department Working Papers Series N78/05/98, NUI Maynooth.

Wand, M. & Jones, M. (1995). *Kernel smoothing*. London: Chapman and Hall.

Yu, K. & Jones, M. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, 93: 228-237.

Ziegelmann, F. (2002). Nonparametric estimation of volatility functions: the local exponential estimator. *Econometric Theory*, 18: 985-991.

Appendix A

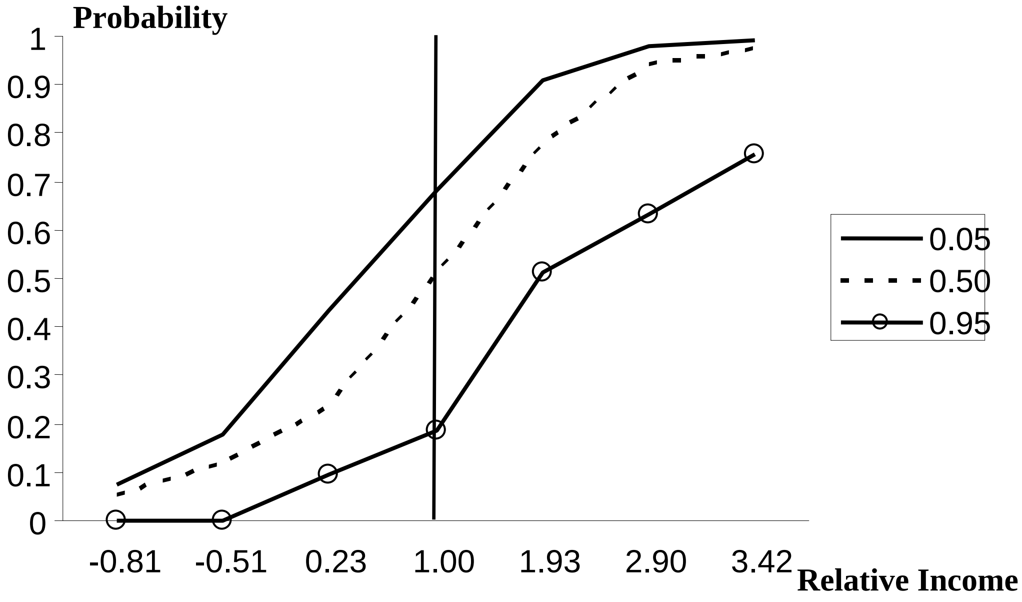

Figure 1: Estimate for the Opportunity Set in Brazil.
Note: The horizontal axis corresponds to the relative income of children (z), whereas the vertical axis represents the accumulated probability, P(Z<=z|X=x), for x in different percentiles for the parents.

Table 1: Accumulated Conditional Probabilities – Parents and Children

Children's percentiles	Parent's percentiles				
	5th	25th	50th	75th	95th
5th	0.076	0.108	0.054	0.005	0.000
25th	0.433	0.375	0.237	0.107	0.093
50th	0.683	0.679	0.506	0.321	0.185
75th	0.909	0.889	0.775	0.632	0.511
95th	0.993	0.987	0.977	0.947	0.758

Source: Research data.