APPLICATION OF THE RECORD LINKAGE OF DATA BASES AND OF THE REGRESSION DISCONTINUITY TO EVALUATE THE IMPACT OF THE "BOLSA FAMILIA" PROGRAM ON THE EDUCATION IN BRAZIL

Julio Racchumi Romero[+]
Diana Oya Sawyer[*]

ABSTRACT

This work explores the sole possibilities generated with the relationship of data bases originated from different information sources, to apply the Regression Discontinuity (RD) in the analysis of the impact of the Bolsa Familia Program (BFP) on the education of children from age 7 to age 14. To answer this objective, it is first utilized the relationship of the data bases from the field research of the Avaliação de Impacto do Programa Bolsa Família and from the administrative records of the CadÚnico. With the information obtained in this relationship, specifically with the income of the CadÚnico families are estimated with RD the effects of the BFP on the education. Two important results were observed: the number of families that it was possible to find was considered satisfactory to estimate the effect with the RD, and the results this estimation indicate the presence of few significant impacts for the indicators of the education, confirming that the program, still by itself, dos not impact all of the educational aspects.

[+] Cedeplar/UFMG
[*] Ministério do Desenvolvimento Social e Combate à FOME

APPLICATION OF THE RECORD LINKAGE OF DATA BASES AND OF THE REGRESSION DISCONTINUITY TO EVALUATE THE IMPACT OF THE "BOLSA FAMILIA" PROGRAM ON THE EDUCATION IN BRAZIL

Julio Racchumi Romero[+]
Diana Oya Sawyer[*]

INTRODUCTION

This work as objective to explore the sole possibilities generated with the relationship of the data bases originated from different sources of information, to apply the Regression Discontinuity (RD) in the analysis of the impact of the "Bolsa Família" Program (BFP) on the education of children from age 7 to age 14.

The BFP is one of the main social programs coordinated and inspected by the Federal Government of Brazil, which envisage alleviating or fighting poverty. The BFP has as objective to reduce the poverty and inequality of today, furnishing cash transferences to the poor families and to reduce the poverty and inequality of tomorrow, providing incentives to the investment in human capital of the benefited families, making possible that these families can come out of poverty. The Program establishes the condition that these families keep the children and adolescents at school age attending school and that the basic care with health are accomplished (Brasil, 2007). Considering the objectives foreseen and the importance of the Program for Brazil, it became necessary to measure the differentials achieved by the program in the benefited groups. Thus, the Pesquisa de Avaliação de Impacto do Programa Bolsa Família (AIBF) carried out in 2005 had as objective to evaluate the impact of the Bolsa Familia Program on the dimensions derived from the budget restrictions and from the operation of behavioral aspects connected with the conditions of the program, having been analyzed the following aspects: Expenditures Relative Structure, Anthropometry, Health, Education, Work performed by the children and by the mother. This research as gained importance due to the coverage that the Bolsa Fmilia Program has achieved in the Brazilian population (Oliveira, etc al. 2007).

Among the results of the social indicators that have achieved more relevance during the past years are the educational ones, because several authors state that to invest in education is to invest in human capital, which constitutes a fundamental component in the process of economical growth, associated to the improvements in the standard of living, more social cohesion and better equality of opportunities; moreover, for the more lacking parcels of the population (Barros, 2001; Willms, 1997). Although this information seems to be reasonable, one shall be careful about the conclusions to be realized with the results found when the Bolsa Familia Program is evaluated, mainly because the program was created from the migration and integration of various previous programs, making impossible the definition of a "before" moment to carry out the experiment and because there are non observable factors that may influence the results.

As in the implementation and the Evaluation of the Impact of the Program was not possible to effect an experimental design, one has opted for the application of method for non-experimental design to evaluate the impact of the program. In the AIBF it was used the methodology Propensity Score Matching (PSM). For our case it is proposed to use the design

[+] Cedeplar/UFMG
[*] Ministério do Desenvolvimento Social e Combate à FOME

Regression Discontinuity (RD); this is a method used when the data are originated from a non-experimental design. It is characterized for considering that the probability of receiving the benefits of the program (be part of the treatment group) is a discontinued function of one or more fundamental variables for the eligibility of the program (Thistlethwaite & Campbell, 1960). In the past years the RD has become the bases of the standard evaluation to solve casual themes with non-experimental data. An intrinsic characteristic of this method is that the treatment group is given to individuals if and only if a co-variable observed intercepts a known threshold. The first to ones to relate design of the RD for the literature of assessment of programs were Hahn, Todd and van der Klauuw (2001), together with Porter (2003), who have formally established at least the conditions for the identification of the model.

The design of the RD that will be applied in this work supposes that, in principle, there is a continuous and "agreeable" relationship between the income of the families, preferably from the administrative record of the "Cadastro Único" for Social Programs (CadÚnico)[2] and the variable of impact to evaluate the differentials of the BFP on the education of children from age 7 to age 14. However, as it is required that to be part of the Program the families needed to have a monthly income of up to R$100,00 (one hundred reais) per individual duly registered in the CadÚnico, this income classifies the families that receive the benefit from the BFP and those who do not receive them. In this sense, there is a "defined" point that separates these two types of families and that can be considered the monthly income of up to R$100,00. Based on this idea, it is expected that that the "agreeable" relationship of the variable monthly income of the family presents a discontinued at the cut or separation point (R$100,00). This discontinued will be explained by the fact that the families that have received benefits form the Program would have better indicators of education, supposing that the benefits of the program had the expected impact. This, for example, a positive impact of the BFP upon the school leave would intuitively show to be as a displacement until below the line that indicates the relationship between both variables, precisely before the cut point that separates the families as benefited or not from the BFP. In the analysis of the discontinued of this method, one shall consider only the families that are in the vicinity of the threshold or cut point. For this work, the size of the neighborhood was defined in such a way that it is obtained a sufficient sample to have statistical power in the estimates.

Before the necessary information for an adequate application of the Regression Discontinuity (RD), it is necessary to using the family income of the beneficiaries in the administrative records of the CadÚnico. In this sense, as the data available for assessment of the impact of the BFP on the education come from a field research of the AIBF and the information to use the RD shall be obtained from the CadÚnico, it became necessary to use the relationship of the data basess originated from different sources. As the relationship of data bases should identify the same entity: which in this case corresponds to each one of the families and their respective members, some problems of conciliation shall be considered. Thus, the process of relationship of data used in this work is defined as a comparison of two or more records of the bases, which contain identification information to determine if these records refer to the same entity (Howe, 1988). At this point, it is important to consider that for the works that make se of data bases, when there is some identifying number of the records the problem is facilitated; but, otherwise, when attempting to relate the data one shall consider other variables, such as name, sex, birth date, municipality code among others (Camargo & Coeli, 2002). These

[2] Basis formed of information from the members of the potential family that has enrolled itself to recfeive some benefit of the programs of transfer of income from the Federal Government

characteristics make important to assess the data bases available in the social area, because frequently these bases do not contain information of codes or undoubted identifiers of the individual or events are not present, requiring a strategy where it is considered more than one variable identifying the entity or individual that is being related. Such alternative strategy is called probabilistic record linkage, which is based on the statistical theory developed by Fellegi and Sunter (19670) and is adequate when the data bases to be related do not contain at least one only identifier, common at the bases to be related, with main applications in joining two heterogeneous data bases. The probability record linkage, like many fields of human endeavour, has progressed as a highly fruitful interplay between theory and experiment, axioms and pragmatism. One viewpoint would see record linkage as primarily a highly practical enterprise based on common-sense and close attention to the empirical characteristics of the data sets involved in any linkage. Another would emphasize the rigorous grounding of record linkage practice in statistical theory and the theory of probability (Fellegi and Sunter, 1969). However, different authors recognizes the value of both perspectives.

With the information obtained from the relationship it is possible to apply the Regression Discontinuity (RD) to assess the impact of the Bolsa Familia Program upon the indicators of education of children from age 7 to age 14. Such indicators are: attendance to school, evasion from school, progression at school, school retaining and allocation between labor and study.

Relevance and results expected from the work

The relevance of the work is disclosed in the methodologies used to assess social programs that in their implementation it was not possible to carry out an experimental design. In this sense, the relationship of the data bases supplies possibilities of application of non-experimental methods in the assessment of the impact of the social programs, such as the Regression Discontinuity (RD). Both the record linkage and the RD enable an integrated view on the information available in various sources of information and enable an alternative analysis to the application effected in the research of Evaluation of the Impact of the BFP. In this work, specifically the record linkage of data bases has permitted to know the additional information of the families interviewed in the field research of the AIBF and to apply the RD with information from the CadÚnico, an exercise that would not be feasible using only one source of information.

With regard to the results of the RD, it was expected to verify the presence of significant impacts for the indicators of education; however, one shall consider that there is a possibility that the results be influenced by the configuration of the model, which considers the families that are around a reduced neighborhood of the threshold or discontinuity cut point. That is, the families those are in the edges or with an income distant from the cur points will not be clearly represented, families that in their majority are in a situation of extreme poverty and for which it is supposed that the benefits of the Bolsa Familia Program reach them in a better measurement. Nevertheless the assessment of the impact results is direct, as the effects will indicate that the results can be associated to the Bolsa Familia Program or to the improvements in the existing programs to achieve the objectives of the social policy.

2. DATA BASE LINKING: AIBF AND CADÚNICO

2.1. The research "Avaliação de Impacto do Programa Bolsa Familia" (AIBF).

The assessment of the impact of the Pesquisa de Avaliação de Impacto do Programa Bolsa Família (AIBF) is a research that was carried out in 2005 with the purpose of assessing the impact of the "Bolsa Familia" program in the dimensions deriving from budget restrictions and from the operation of behavioral aspects connected to the program conditions: Relative Expenditures Structure, Anthropometry, Health, Education, infantile work and Mother's work. This research gained much importance for the coverage that the "Bolsa Familia" program has been achieving in the Brazilian production (OLIVEIRA et al, 2007).

It was used a non-experimental design, as the program was created from the migration and integration of various previous programs, without the possibility of defining a "before" moment. Besides, once the program has as target the universalization between the population below the misery line and the poverty line, the establishment of an random control group would create an ethical problem as to the denial of a benefit to a certain number of wanting families. Taking into consideration the non-experimental assessment, AIBF has opted for the elaboration of a research of household base line, of observational character. The sample was defined to obtain representativeness for the Northeast region (NE), the Southeast and South regions (SE-S), jointly, the Northern and Center-West regions (N-CO), also jointly. The sample was distributed at households identified as: program benefited households (cases); households with families cadastre in the Sole Cadastre, but still not beneficiary of the program (control 1); and households without beneficiary of cadastre families (control 2); providing a different probability to each group, in the following proportions: 30% (cases), 60% (control 1), 10% (control 2). The collection resulted in a total of 15.426 completed questionnaires. For the SE-S stratum, this total was of 5.887. The NE and N-CO strata presented totals of 5.106 and 4.433, respectively.

2.2. CadÚnico Administrative Record

The Sole Cadastre administrative record for Social Programs (CadÚnico) is a fundamental instrument to identify the poor families in the country, to know their vulnerabilities and potentialities and subsidies the elaboration and implementation of public policies destined to these families. The CadÚnico allows for granting benefits from the BFP, guides the design and implementation of public policies under responsibility of different areas in the government, directed to the low income families, whenever possible. Enables to better characterize the various dimensions of poverty and vulnerability to beyond the monetary income. The CadÚnico also allows to identify, by means of multidimensional variables, the more vulnerable families, with priority for family follow up and those that may, as per their characteristics, be included in supplemental programs of the BFP (BARROS et al, 2008; RAMOS e SANTANA, 2002).

Taking into consideration the information contained in the CadÚnico, it is important to disclose that this can be associated to a Household Field Research; this is due to the fact that the data survey covers a group of individual and family information, besides surveying data about life conditions. That is, not only useful information for a type of program or programs is surveyed, but also it contemplates broader information, which is useful to asses social problems (BARROS et al, 2008).

2.3. The record linkage of Data Bases

One of the data base used for the record linkage was derived from the AIBF field research and the other from the administrative records of CadÚnico. When one works with these types of data bases and/or with national information systems such as the one of CadÚnico, frequently the single individual's codes or identifiers are not present and the possibilities information recuperation of a same individual from different data bases is often a hard task. In view of this, as we have records with single identification information of the units to list, it was used the deterministic record linkage strategy; but when this identifier was not present, special techniques that considered more than one variable identifier of the individual to list, different data bases were necessary, such as the strategy known as probabilistic record linkage.

(a) Preparation of the Bases

This is a previous stage to the record linkage process, in which are carried out field editions and standardizations (variables). Besides, duplicated records are excluded; that is, those that belong to the same entity, within the same data file. In the exploration of the data from the CadÚnico administrative record, are duplicated records, being found 18% of duplicated records of the CadÚnico throughout Brazil (The bases of the AIBF research has not presented duplicated cases). For the standardization of the fields, it were applied the same codifications of sex, residence municipality, formats for the Social Identification Number (SIN) , Name and surname, date of birth, aiming at making possible to be associated the fields of the different data banks. Besides, it was applied the phonetic code of Soundex[1] (Newcombe et al., 1988), useful for blocks.

(b) Deterministic record linkage

It was used the "Social Identification Number (SIN)". The SIN is a number that proves the inscription in the Federal Government social programs (such as Bolsa Escola, Bolsa Alimentação, Auxílio Gás or Bolsa Família), designated to the person who inscription to receive the benefit. After carrying out an automatic review of the paired records, the results indicate that 73,8% of the individuals in the AIBF research were found in the CadÚnico, who represented 35,24% of the total of families interviewed in the AIBF field research.

(c) Probabilistic record linkage

Although the Social Identification Number (SIN) is a single and non-transferable identifier, in some cases it presented problems in the statements by the interviewed families (SIN with less than 11 digits or inexistent) and in the records collected from the CadÚnico administrative records (SIN with zero value and duplicated). Due to this reason, it was necessary to carry out the probabilistic record linkage, which is based on the statistical theory developed by Fellegi and Sunter (1969), and enables to classify the paired results from total consonance (exact) to the entire non-consonance and with various levels of consonance among them (CHRISTEN e CHURCHES, 2006)[2].

[1] Soundex is a phonetic algorithm for indexing names by sound. The goal is for homophones to be encoded to the same representation so that they can be matched despite minor differences in spelling (Camargo & Coeli, 2000).

[2] The software utilized is Reclink II eveloped by Camargo & Coeli (2002b), which implements several routines for processing files, based on the technique of record linkage probabilistic.

The final results of the percentage of paired results that were considered as true pairs, indicate that the percentage of records found in the application of this relationship is around 73% for the entire Brazil, a percentage hat can be considered significant, as only 27% did not succeed to recuperate the administrative records. Finding a technical explanation to justify the lack of records not found may be hard; however, the study does not escape from some errors as the variables used for blockage and relationship are poorly filled out, or even unfilled, resulting to be impossible to identify a true pair.

(d) Consequences and utilization of results from the data bases relationship

Taking into consideration that the record linkage enables to find families interviewed in the AIBF field research in the CadÚnico record, the results indicated that from the total of 15.426 families interviewed in the AIBF field research, 5.437 families were found with the deterministic record linkage (35% of the total), and 4.550 families that represent 30% were found with the probabilistic record linkage. Jointly, it was found 65% (9.987) of the families in the AIBF field research in the CadÚnico record for the entire Brazil. In spite of the exclusion of records with errors derived from the filling out or duplicity of information, the number of families that it was possible to find is considered very good; this is because not all of the families interviewed in the AIBF field research are actually in the CadÚnico, because there are families cadastre, but that are not yet beneficiary of the "Bolsa Familia" program (BFP) (households may be beneficiary of other income transfer programs, but not of this program); and families not found or beneficiary (Oliveira et al, 2007)

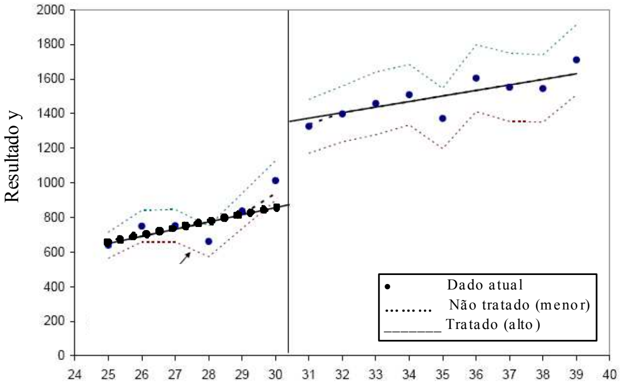
In special the record linkage results enable to identify in a special manner the groups potentially beneficiary and non-beneficiary of the social programs through of family income, to assess the impact upon the education indicators applying the Regression discontinuity (DR). The application of this technique is only possible when it is used a continuous variable such as the "family income" of the CadÚnico administrative record, presupposing that this variable is a pre-treatment and is not influenced by the income received by the beneficiaries, but that would influence the results of the "Bolsa Familia" Program (BFP) impact and in the participation of the families beneficiary of this program.

3. IMPACT EVALUATION METHODOLOGY: Regression Discontinuity

In the past years, the Regression Discontinuity (RD) has become the basis for the standard assessment to solve causal themes with non-experimental data. This method considers that the probability of receiving the benefits from the program (be part of the treatment group) is a discontinued function of one or more fundamental variables for the eligibility of the program (Buddelmeyer and Skoufias, 2004). One intrinsic characteristic of this method is that the treatment group is given to individuals if and solely if a co-variable observed intercepts a known threshold. Under these conditions the probability of receiving the benefits from the program near the variable threshold has a casual behavior. This is the only design that enables to identify the causal effect of the program without imposing arbitrary exclusive restrictions, suppositions about the selection process, functional form or the presupposed of error distribution (BLACK, GALDO and SMITH, 2005).

In the Regression Discontinuity the literature distinguishes general scenarios of the design, the regression discontinuity design of Sharp and Fuzzy (SRD and FRD respectively) (TROCHIM, 2001). In the Sharp (SRD) design the treatment is known and depends on a deterministic form of some noticeable variables, while in the Fuzzy (FRD) design the variable “treatment” is an random variable, given the observables variables, but the conditional probability known in the discontinued point in which the observable variable takes the threshold value; an example shown in Van der Klaauw (2002). To make operational the regression discontinuity design, it is necessary additional information to the selection rule; that is, to know the designation mechanisms to the treatment, which depend on the values of a observable continued variable, relative to the given threshold, or to its cutting score, in such a way that the corresponding probability of obtainment of the treated be a discontinued function of this variable in the cutting score (see Figure 1).

Figure 1 – Example of a design de regression discontinuity**.**

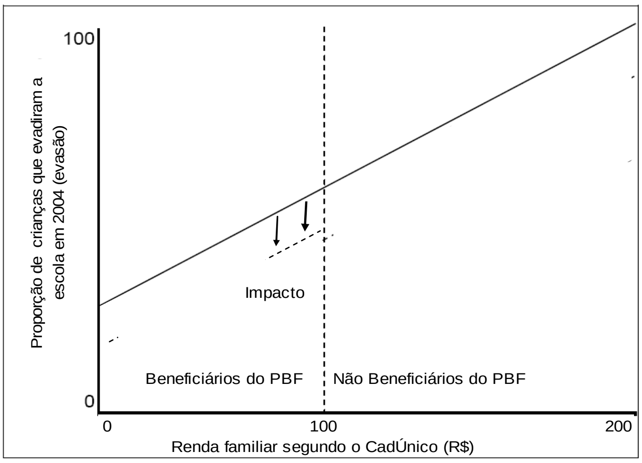


3.2. Implementation of the RD to evaluation impact of the BFP in the education

In this paper it is supposed in principle that there is a continuous or “smooth” relation between the income of the families from the CadÚnico record and the impact variable; that is, impact indicators to evaluation the BFP differentials in the education of children between 7 and 14 years of age. Nevertheless, we are using the characteristic of “minimum income” from the BFP; that is, the families needed to have a monthly income of up to R$50,00 and up to R$100,00 per individual duly inscription in the CadÚnico; thus, this income classifies the families who receive the BFP benefit and those who do not receive it. In this sense, there is a “defined” point that separates these two types of families and that may be considered a minimum monthly income. Based on this idea, it is expected that the “smooth” relation of the per capita income variable presents a discontinued at the cutting or separation point. This discontinuity will be explained by the fact that the families who received the benefits from the BFP would have better education indicators, supposing that the program benefits had the expected impact (example: seer Figure 2)

In the carrying out of the discontinued analysis it must be considered the families who are in the neighboring of the threshold or cutting point, arising the problem of how to define this neighborhood. When the neighborhood is defined as a very ample way – for example, taking into consideration practically all families who are considered in the study - then the estimates have a gain in terms of statistical power, but lose sense that the groups in each side have more heterogeneous families and consequently more difficult to be compared among themselves. When the neighborhood is defined in a narrow way, then exactly the contrary occurs. For this article, the size of the neighborhood has been defined in such a way that one can obtain a sufficient sample to have the statistical power in the estimates. However, with the purpose of verifying if the results are sensitive to the size of the selected neighborhood, one more neighborhood was defined, but solely as a test. Besides, for analyses issues, the estimates of the RD model will be carried out for two cutting points: family income up to R$50,00 and up to R$100,00.

Figure 2 Schedule of discontinuity family income of CadÚnico on the impact of the proportion of children who not attend the school in 2005



3.3. Variables for Impact Evaluation.

The data used to measure the impact correspond to the families who were found both in the data base of the AIBF field research and in the CadÚnico administrative records. Thus, the sample for impact evaluation using sharp discontinuity Regression is formed by 9.987 families.

Taking into consideration that in the educational component of the BFP, there is a condition that the children between 6 and 15 years old attend school regularly, it is expected that the program beneficiaries present positive effects upon the education indicators. Many studies have analyzed the importance of the family antecedents in determining the education results of children and adolescents. Behrman, Duryea and Székely (1999) analyze the influence of the family background in a direct way upon education gains of children and adolescents. About the human capital family production, Becker (1993) was one of the first to disclose that the household goods are produced by a combination of goods and domestic work. As for the case of families with children in school age, the differentials of the BFP may be measured by the variables of development in school which, as follows, are presented in Table 1.

Table 1 Dependent variables: Indicators to evaluation the differentials of the BFP in the education of children between 7 and 14 years old

Variables	Description
Did not miss school during the last month (or the supplement of it)	Proportion of girls and boys in the household that did not miss school during the last month
Evasion or abandonment	Proportion of girls and boys in the household who evaded from the teaching system between 2004 and 2005
Progression	Proportion of girls and boys in the household who were approved between 2004 and 2005
Allocation between work and study	Proportion of girls and boys in the household who have stated they are only studying currently, in face of those who have stated they only work, they work and study and who neither work nor study
Retention	Proportion of girls and boys who were unapproved between 2004 and 2005

The independent variables are specified to apply the RD, variables formed by a series of individual and household characteristics which control some aspect of the family eligibility to participate in the program and the education variables. The independent variables are in Table 2.

Table 2 Independent variables: variables used in the specification of regression discontinuity models to evaluate the differential PBF in education

Atributos do chefe de família:	
Raça do chefe de família	Branca Não Branca
Sexo do chefe de família	Masculino Feminino
Escolaridade do chefe de família	Até 3 anos de estudos* Até 4 anos de estudos* Até 7 anos de estudos*
Idade do chefe de família	Menor e igual há 50 anos Mais que 50 anos
Altura em metros do chefe de família	Medida em metros (mts)
Escolaridade da mãe do chefe de família	Mãe alfabetizada Mãe não alfabetizada
Tempo de permanência do chefe de família no município	Menos de 10 anos* Menos de 5 anos*
Tempo de permanência do chefe de família na área rural.	Viveu até os 14 anos Não viveu até os 14 anos
Características da família:	
Número de membros da família	Número de membros no domicilio
Crianças entre 0 a 3 anos de idade	Proporção de crianças de 0 a 3 anos
Crianças entre 0 a 6 anos de idade	Proporção de crianças de 0 a 6 anos
Crianças mulheres 7a14/ criança 0 a 14 anos	Proporção crianças mulheres 7 a 14/ crianças 0 a 14
Casal com filhos até 14 anos	O Casal tem filhos até 14 anos O Casal não tem filhos até 14 anos
Presença de pessoas de 60 anos ou mais	Há pessoa de 60 anos e mais no domicílio Há pessoa menor de 60 anos no domicílio.
Características do domicilio:	
Qualidade de domicilio[1]	Qualidade inferior* Qualidade media*
Área de residência do domicilio	Urbana Rural
Região de residência do domicílio	Nordeste* Norte – Centro Oeste*

* For each of these categories was constructed a dummy variable.

[1] This variable was generated as a Grade of Membership (GOM), with three categories for the quality of the conditions of households, classified as: very good, regular and bad.

4. RESULTS

The sample eligible to measure the impact of the BFP upon education is 12.514 households, this sample of households was selected considering only those presented among its members children from 7 to 14 years old, resulting, as per the AIBF research, in 3.248 households beneficiary of the BFP and 2.568 non-beneficiaries of any program. Same way, the distribution of the groups in accordance with the allocation of the relationship with CadÚnico, it was of 3.988 beneficiaries of the BFP and 2.955 are not beneficiaries or are not cadastre.

Analyzing the impact variables, in Table 3, it is observed that in Brazil 88,27% of the children between 7 and 14 years old did not miss school or nursery in October, 2005. Taking into consideration the distribution of the comparison group, the results are higher for the group of beneficiaries, both the allocated as per the AIBF field research and the allocated with the CadÚnico; thus, approximately 89% of the children from 7 to 14 years old that belong to families beneficiary of the BFP did not miss school in October, 2005, as compared to the group of non-beneficiaries (86,01% and 87,7%). Besides, the differences that exist between both groups are statistically significant for the two allocation procedures used.

Table 3: Indicators to Evaluate the differential of the BFP in the education of children from 7 to 14 years, by group of comparison, Brazil and Regions, 2005 (in%)

Variables de Impact	Group AIBF		P-value	Group CadÚnico		P-value	Total
	Beneficiaries BF	Not Beneficiaries		Beneficiaries BF	Not Beneficiaries		
Did not miss school during the last month							
Brazil	89,73	86,01	<0,01	89,52	87,70	<0,01	88,27
Man	89,14	89,07	NS	90,24	90,56	NS	88,78
Woman	90,38	83,12	<0,01	88,74	85,11	<0,01	87,70
Evasion or abandonment							
Brazil	1,05	2,12	<0,01	1,22	2,35	<0,01	1,59
Man	0,84	2,48	<0,01	0,94	2,51	<0,01	1,35
Woman	1,27	1,79	NS	1,53	2,22	<0,10	1,85
Progression							
Brazil	82,81	87,33	<0,01	83,58	86,59	<0,01	86,46
Man	80,00	86,59	<0,01	80,59	84,90	<0,01	85,16
Woman	85,90	87,98	<0,10	86,77	88,07	NS	87,88
Allocation between work and study							
Brazil	91,87	95,06	<0,01	92,37	94,23	<0,01	94,15
Man	90,71	93,75	<0,01	91,53	92,38	NS	93,44
Woman	93,14	96,30	<0,01	93,29	95,93	<0,01	94,94
Retention							
Brazil	16,01	11,19	<0,01	15,01	12,14	<0,01	12,22
Man	19,16	12,50	<0,01	18,41	14,10	<0,01	13,93
Woman	12,54	10,05	NS	11,39	10,43	NS	10,37

Source: AIBF e CadÚnico, 2005.

Note: Total column refers to values for all households with children 7 to 14 years.

P-value: the probability of observing a result as or more extreme than the sample, assuming that the null hypothesis is true.

Not significant

With regard to evasion from school or abandonment between 2004 and 2005, it is observed that approximately 2% of the children from 7 to 14 years old, abandoned school in 2005. Among the comparison groups, the group of beneficiaries of the BFP has presented a lower percentage of abandonment; 1,05% for those allocated as per the AIBF field research and 1,22% for those allocated with the CadÚnico.

The progression indicator indicated that 86,46% of the students from 7 to 14 years old were approved in 2005 in the entire Brazil. Taking into consideration the group of beneficiaries of the BFP and the non-beneficiaries, 83% of the students have progressed in accordance with the groups allocated with the AIBF field research and 84% as per the groups allocated with relationship with the CadÚnico, percentages being approximately 5 percent points lower than those observed in the group of non-beneficiaries, in both of the allocation procedures used. Regarding the allocation between work and study, the percentage of children from 7 to 14 years old who were only studying stands above 90%, while children that only worked form a small parcel.

Among the groups beneficiary of the BFP and the non-beneficiary, it is observed that the percentage of children from 7 to 14 years old in the allocation group that only studied in the beneficiaries of the BFP was of approximately 91,87% and 95,06% for the non-beneficiary, as it was obtained from the allocation as per the AIBF field research. In the data obtained from the allocation groups in accordance with the relationship with the CadÚnico, the children who only studied in the group of beneficiaries of the BFP was of 92,37% and 94,23% among the non-beneficiaries. In such case the differences found were statistically significant for the two allocation procedures used. Analyzing the allocation between work and study, of the children in accordance with their sex, it was found that 94,15% of the boys and 93,44% of the girls were only dedicated to study. Checking the results of the group of beneficiaries of the BFP and the non-beneficiaries, we find that the masculine children that only studied in the group of beneficiaries of the BFP is of 90,71%, a percentage lower that that of the group of non-beneficiaries, with 93,75%, in the allocation groups as per the field research, as related to allocation groups in accordance with the relationship with the CadÚnico.

The last impact indicator refers to the school repeaters, in which it is observed that only 12,22% of the students repeated the school year in Brazil. Taking into consideration the comparison groups, the results indicate that the percentage of repeaters is higher in the group of beneficiaries of the BFP, both for those allocated as per the AIBF field research (16,01%) and as per the relationship with the CadÚnico (15,01%), higher percentages as related to the group of non-beneficiaries.

In Table 5 are presented the results of estimates in the RD model for the differentials in the education of children from 7 to 14 years old in the BFP, only reporting the coefficients of the variable that indicates a statistically significant differential.

Table 4

Estimation of the regression discontinuity of indicators for evaluates differentials of BFP in the education of children 7 to 14 years old. Brazil and Regions, 2005

Variables/Regions	cutoff point					
	Up toR$100.00			Up to R$50.00		
	Total	Man	Woman	Total	Man	Woman
Did not miss school during the last month						
Brazil			-0,015**			-0,017*
Northeast	-0,026*					
Northern and Center-West				-0,023*		
b) children that approve a school 2004 to 2005						
Brazil						
Northeast					0,283*	
Northern and Center-West						
c) Children that repeated school among 2004 and 2005						
Brazil						
Northeast	-0,097*				-0,290*	
Northern and Center-West						
d) Children that only studied in 2005						
Brazil				-0,218***		
Northeast			-0,134***			
Northern and Center-West						

Source: AIBF e CadÚnico, 2005.

* Valor significant 10%; ** Valor significant 5%; *** Valor significant 1%.

Note: The group of Treatment is constituent for the households that currently receive the benefit of the BFP. The group of not-beneficiaries is composed of households who reported ever having received any benefit, independently be registered in a public program. The south\ southeast did not have any significant result.

Taking into consideration the discontinued in the cutting point up to R$100,00, we find that for the feminine children in the entire Brazil and the total of children in the Northeast region that evaded from school in 2004, they have significant differentials and are favorable to the families with income below R$100,00, as they are negative. Same way, taking into consideration the discontinued in R$50,00, the evasion of feminine children in the entire Brazil and the total of children in the North/Center-West region who evaded from school in 2004, this presents significant differentials and favorable to the families with income below R$50,00 reais, because the differentials are negative. In view of these results, it is possible to suppose that there is a difference favorable to the beneficiaries of the BFP as related to the children belonging to households that do not participate in any program, a result that favors the objectives of the program in these regions and groups of children.

With regard to the results in terms of proportion of those approved between 2004 and 2005, significant differentials of the BFP have a positive difference for the discontinued in R$50,00 and for the masculine children in the northeast region this result indicates a higher approval of the families with income below R$50,00 reais. From this, it is supposed that as the families receive the benefit of the BFP are those below the income cut, then it is suggested a potential positive effect for the program beneficiaries, as related to the group of non-beneficiaries.

For the variable of repeaters in school between 2004 and 2005, one finds significant and negative differentials for the discontinued cut of R$100,00 among the total of children in the Northeast and for the discontinued threshold of R$50,00 among the masculine children in the Northeast region. These results could be interpreted as favorable to the families with income below these specified income cuts, families that possibly receive benefits from the BFP and therefore it is supposed that there is a favorable difference to the beneficiaries of the BFP as related to the children in households that do not participate in any program.

Taking into consideration that the proportion of children that work in face of those that only study or do not work neither study, significant and negative differentials are found for the discontinuous of R$100,00 and among boys in the |Northeast region and for the discontinuous of R$50,00 among children in the entire Brazil. These results indicate a higher participation in the work force among children with family income below the income cuts and considered regions, as compared to the group of non-beneficiary families. Different results could be expected with this indicator because it is supposed that families below these income cuts receive the benefit from the BFP, but taking into consideration that still there may be a higher participation in the work force independently of the attendance to school by the children. Which may be the reflex of the conciliation between work and study which has not yet has not succeeded to be diminished or eliminated; but for future measurements different results are expected.

Finally, it shall be disclosed that the variable of not missing school during the last month was not shown on the table, because no differential was significant. Besides, the presence of few significant differentials for all education differentials and regions studied may be interpreted as a result of the configuration of the RD model. The RD model considers that the families in the reduced neighborhood surrounding at the cutting point is discontinued as related to a variable exogenous to the impact potential results which, for out case, is the family income in the Cadúnico administrative records (income cut of R$100,00 and R$50,00). This way, the families that are at the extremes or with an income far from the cutting points will not the explicitly represented; families that in their majority are in extreme poverty and who, it is supposed, the benefits of the BFP reach them in a broader measurement.

5. FINAL CONSIDERATIONS

To assess the effects of the BFP in the education indicators of children from 7toa 14 years old, as per the results obtained from variables considered as pre-treatment, starting from the data bases relationship, it was adopted the methodology of regression discontinuity (RD). This technique consists of comparing the families that are at the Program's eligibility threshold, considering the characteristic of minimum income. Given the data restrictions, the utilization of this technique seem to be the more indicated methodologies, being that the method of RD is an application as direct result of the data base relationship which in principle suppose that there is a continuous or "smooth" relationship among families' income in the CadÚnico and the impact variable; that is, impact indicators to assess the differentials of the BFP in the education of children from 7 to 14 years old.

In accordance with the results obtained, firstly it is possible to emphasize that the methodology of record linkage of databases is of fundamental importance for the application of the non-experimental technique, useful to assess the impact results of the social programs. This is pertinent because various areas applied the record linkage database as a tool to improve the quantity and quality of the information required for a research (GILL, 2001).

Secondly, making use of a special way of identifying groups potentially beneficiary and non-beneficiary of the BFP with a per capita income cutting point of R$50,00 and R$100,00, through the application of the Sharp RD method, it is verified the presence of some significant results for the education indicators and regions studied. It is possible that the non-significant results have been influenced by the configuration of the model, which considers only the families that are in the surroundings of a reduced neighborhood at the threshold or at the point of discontinuity cut. That is, families that are at the extremes or with an income far from the cutting points will not be explicitly represented, families that in their majority are in a condition of extreme poverty and for who, it is supposed, are reached in a better measurement by the benefits of the BFP. However, the expressive results that were found with the discontinued Regression confirmed some results found in the AIBF research project.

This paper aims at first contributing to explore the sole possibilities that are opened by the record linkage databases of two different sources, to analyze the impact of the social programs for income transfer; and secondly, apply data from the product of the record linkage databases to apply non-experimental methods, such as the method of regressions discontinuity, in which the families with per capita income near the poverty line proposed by the federal government are compared. In such case, it is eliminated the selection aspect attributed to individuals who opt to participate in the programs and compare the eligible with similar characteristics. This provides a higher strength to the results.

REFERENCES

BARROS, R.; CARVALHO, M.; MENDONÇA, R. S**obre as utilidades do Cadastro Único**. Niterói: Universidade Federal Fluminense, 2008. (Texto para Discussão, 244).

BECKER, G. S. **Human capital: a theorical and empirical analisis, with special reference to education**. London: The University of Chicago Press, 1993.

BEHRMAN, J. R.; DURYEA, S.; SZÉKELY.; M. S**chooling investments and aggregate conditions: a household-survey-based approach for Latin America and the Caribbean.** Washington, DC: Inter-American Development Bank, 1999.

BEHRMAN, J.; SENGUPTA, P.; TODD, P. **Progressing through PROGRESA: an impact assessment of a school subsidy experiment.** Washington, D.C: IFPRI, 2001.

BLACK, D.; GALDO, J.; SMITH, J. **Evaluating the regression discontinuity design using experimental data.** University of Michigan, 2005.

BRASIL. Ministério de Desenvolvimento Social e Combate à Fome. **Benefício de prestação continuada de assistência social (BPC)** Brasília, DF, [2006]

BRASIL. Ministério de Desenvolvimento Social e Combate à Fome. **O Programa Bolsa Família.** Brasília, DF, [2007].

CAMARGO JR., K. R; COELI, C. M. Reclink: aplicativo para o relacionamento de banco de dados implementando o método probabilistic record linkage. **Cadernos de Saúde Pública,** Rio de Janeiro: v. 16, n. 2, p. 439-47. abr./jun.. 2000.

CAMARGO JR., K. R.; COELI, C. M. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. **Revista Brasileira de Epidemiologia**, v. 5, n. 2, 2002a.

CAMARGO JR., K. R.; COELI, C. M. **Reclink II: guia do usuário**. Rio de Janeiro, 2002b.

CHENG, M.Y., FAN, J., MARRON, J.S. On automatic boundary corrections. **Annals of Statistics,** XXV, 1997.

CHRISTEN, P.; CHURCHES, T. **Secure health data linkage and geocoding: current approaches and research directions**. In: NATIONAL E-HEALTH PRIVACY AND SECURITY SYMPOSIUM, Brisbane, 2006. Proceedings... [2006].

FAN, J., GIJBELS, I., HU, T., HUANG, L. A study of variable bandwidth selection for local polynomial regression. **Statistica Sinica 6,** 1996.

FELLEGI, I. P.; SUNTER, A. A theory of record linkage. **Journal of the American Statistical Association**,v. 64, n. 328, p. 1183-1210, 1969.

HAHN, J., TODD, P., VAN DER KLAUUW, W. Evaluating the Effect of an Antidiscrimination Law Using a Regression-Discontinuity Design. **NBER Working Papers** 7131, 1999.

HOWE, G. R Use of computerized record linkage in cohort studies. **Epidemiologic Reviews**, v. 20, n. 1, p. 112-21, 1998.

JARO, M. A. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. **Journal of the American Statistical Association**, v. 84, n. 406, p. 414-420, 1989.

MCCRARY, J. **Manipulation of the running variable in the regression discontinuity design**. Working Paper, University of Michigan, 2005.

OLIVEIRA, A. et al. Primeiros resultados da análise da linha de base da pesquisa de avaliação de impacto do programa bolsa família. In: VAITSMAN, J.; SOUSA, R. P. **Avaliação de políticas e programas do MDS –Resultados: Bolsa Família e Assistência Social**. Brasília, DF: Ministério do Desenvolvimento e Combate a Fome, Secretaria de Avaliação e Gestão da Informação, 2007. v.2

RAMOS, C. E.; SANTANA, R. **Os pobres que levantem a mão (mas será que são mesmo pobres?). Uma tentativa de validar o cadastro único**. Brasília: Universidade de Brasília, 2002.

SCHUTT, R.I **Investigating the social world: the process and practice of research**. Thousand Oaks: Pine Forge Press, 2001.

THISTLETHWAITE, D.; CAMPBELL, D. Regression-discontinuity analysis: An alternative to the ex post facto experiment. **Journal of Educational Psychology**, 51: 309-317, 1960.

TROCHIM, W. Regression Discontinuity Designs. In: SMELSER, N.J., BALTES, P.B. (eds.) **International Encyclopedia of the Social and Behavioral Sciences**. Pergamon, Oxford, 2001.

VAN DER KLAAUW, W. Estimating the Effect of Financial Aid Offers to College Enrolment: A Regression-Discontinuity Approach. **International Economic Review,** 43(4), 1249-1287, 2002.

WILLMS, J. D. Literacy skills and social class. **Options Politiques**, v.18, n.6, p.22-26, 1997.