Inequality in the Brazilian Labor Market: A Heckman's Procedure Analysis

Gilnei Costa Santos[1]
Rosa M. O. Fontes[2]
Patrícia M. A. Bastos[3]
João E. de Lima[4]

ABSTRACT

This paper analyzes the inequality in the Brazilian labor market generating an earning equation for Brazil in 2006 with the PNAD database and identifying the factors that have the highest impact on income and on gender inequality. The adopted model was the Heckman sample selection procedure to avoid selectivity bias. Additionally, it was verified that if PNAD was not considered as a complex sample some bias could be generated in the results. The results of the participation of individuals in the labor market showed that if the estimates did not consider the complex sample plan, all the variances of the variables would be underestimated, and the same result would be valid for the earning equation. With respect to gender inequality, although women have presented higher qualification than men, the male earnings had been 30% higher than the female earnings. In addition, men tend to have a higher probability to be in the labor market. These results suggest a major problem in the Brazilian labor market, which is the fact that there are some worker's differences caused by characteristics that are not related to the individual's productivity. With respect to race, being black, mulatto, yellow or indigenous increased the probability of being employed compared to the whites. This may suggest that the blacks need to develop more activities or have a lower reservation wage than the whites. Individuals in urban areas were less likely to be in the labor market. Finally, each additional year invested in education increased the likelihood of participating in the labor market by 4.41 p.p., showing that investment in human capital is extremely important for entrance into the labor market. Furthermore, a parabolic pathway was detected for the experience proxy, that is, an inverted U shaped relationship for the variables age and age squared. The results for the income determinants showed that individuals with 10 or more years of study increased income by 18% for each additional year of schooling.

Keywords: Gender inequality, work income, Heckman's procedure, PNAD.

[1] Economist, Graduate Student in Economics - Universidade Federal de Viçosa (MG), Departament of Economics. E-mail: gilnei.santos@ufv.br
[2] Professor at the Departament of Economics - Universidade Federal de Viçosa. E-mail: rfontes@ufv.br
[3] Economist. Graduate Student in Economics - Universidade Federal de Viçosa (MG), Departament of Economics. E-mail: patiabrita@yahoo.com.br
[4] Professor at the Departament of Agricultural Economics - Universidade Federal de Viçosa (MG). E-mail: jelima@ufv.br

1. Introduction

Income inequality in Latin America is among the highest in the world, justifying the existence of an extensive literature on its determinants, especially studies by Morley (2001) and Ribeiro (2006), and specifically for Brazil, by Barros and Mendonça (1993, 1995 a and b), Hoffmann (1989, 1992, 1996, 1998, 2000), Teixeira (2006) and Barros *et al* (2007). Brazil stands out from other Latin American countries as the country with the greatest income inequality. The 1999 Human Development Report of the United Nations Program for Development (PNUD, 1999) showed that only South Africa and Malawi have a greater degree of inequality than Brazil. The literature points out that among the determinants of this disparity, besides historical factors, education is an important variable and it is poorly distributed in Brazil. Another explanatory factor for Brazilian inequality is the influence of the labor market, because of its characteristics of discrimination and segmentation, and the differentials in regional remuneration.

Ramos and Vieira (2001) reported the high explanatory power of schooling on income in several Latin American countries, except for Chile and Argentina. However, the contribution of education is even greater in Brazil, both in relative and absolute terms. According to the study by the IPEA-CAIXA (2007), education greatly influences the inequality of the Brazilian wage for two reasons: the high educational inequality among workers and the high sensitivity of wages in relation to the educational level. This same thought is shared by Morley (2001), who emphasized that in Brazil the returns for a high educational level are greater than in other countries with the same degree of development, such as the Asian countries. According to the same author, in Brazil there is a pro-inequality educational policy, because it prioritizes investments in university education.

On the other hand, as Gandra (2002) stated, attributes such as race, gender, age, ethnicity and formal credentials are also determinant variables in the process of worker selection and wage determination and they can be also used to determine the fragmentation of the labor market. For example, Hoffmann and Leone (2004) reported that "the less valued and traditionally female occupations in the labor market continue to reproduce, implying the persistence of occupational cluster, for example, domestic employment". Another form of discrimination refers to race or color, and the main representative is the black race that, similarly to the employed female population, tends to occupy low-paying positions.

Thus the focus of the present paper is to analyze some factors that determine the entry of the individual to the labor market and to generate an earning equationfor Brazil in 2006, in order to identify the factors that most impact on income. Additionally, some variables were incorporated into the equation regarding segmentation and discrimination on the labor market, such as color or race, gender, geographic region etc. Furthermore, it was verified that if the PNAD was not considered as a complex sample, bias would be generated on the results. According to Resende and Wyllie (2006), a good part "of the econometric studies about returns for education ignore the research sample design", such as PNAD. Ratifying these authors, Silva and Pessoa (2002) stated that regression analyses with PNAD are often performed by analysts who work outside the agency that produced the data and frequently using statistical packages for modeling that are based on hypotheses that are valid only when the data are obtained by simple random samples with replacement.

The model adopted here is a sample selection model using the Heckman's procedure to avoid selectivity bias. The earning equation is a hybrid of the Jacob Mincer theory, and it takes into consideration the variables used by Scampini (1996) and Hoffmann and Simão (2005). The purpose is to better portray the effects of the education variable estimating the threshold effect as proposed by Ney and Hoffmann (2003).

In addition to the introduction and conclusions, the study is organized in three sections. The first presents the theoretical model, the second develops the methodology, dealing with questions such as the particularities of the National Household Survey (PNAD - Pesquisa Nacional por Amostra de Domicílios), and the estimating methods with this type of sample, the sample selection model and finally, the variables used. The main results are discussed in the fourth section.

2. Theoretical Model

2.1 - Labor Market Participation Decision

This section reports the theoretical foundations that corroborate the participation of the individual in the labor market. According to Berndt (1996) *apud* Scorzafave and Menezes Filho (2001), the decision to offer work of a given individual is determined by a utility function described in the following manner:

$$\text{Max } U(G,L) \qquad (1)$$

Subject to

$$P_G G = P_L(T - L) + V \qquad (2)$$

That is, the objective of the agent is to maximize the utility (U), and the arguments of this function (U) are the quantity of goods (G) and hours of leisure (L). Maximization is subject to a budget constraint, determined by the income not derived from employment (V), number of hours available (T), the price of the goods (P_G) and the price of leisure (P_L). The total number of hours worked (H) is defined as H = T – L. The total expenditure on goods should equal the earnings from work, and the income not derived from work. Solving the conditioned maximized problem, the following is obtained[5]:

$$\frac{UMg_L}{UMg_G} = TMS_{GL} = \frac{P_L}{P_G} \qquad (3)$$

That is, the marginal utility of leisure per monetary unit spent on leisure should be equal to the marginal utility generated by the goods consumed over the price of these goods. In other words, the marginal substitution rate (MSR) of quantities of goods (G) by hours of leisure (L) is equal to the relative prices. According to Scorzafave and Menezes Filho (2001), "graphically, this condition implies the condition of tangency between the indifference curve and the budget constraint. At this point of tangency, the number of hours worked and the quantity of goods to be consumed are determined. However, it is observed that this condition is only satisfied in the case of interior solutions, where $L < T$ and $H > 0$".

The individual decision to enter the labor market comes, therefore, from a corner solution where L = T and H = 0, that is, the agent is willing to offer zero hours of work. It is assumed, therefore, that the satisfaction of one more hour of leisure is greater than the relative price of leisure and, thus, the agent does not offer work, that is, he does not participate in the Economic Active Population. It can be stated therefore that in the case of the corner solution, the reservation wage of the individual is greater than that offered on the market, and the reservation wage is $TMS_{G/L}$. The rule to decide to take part in the market will be: the agent participates in the labor force if the wage offered on the market is greater than his reservation

[5] Considering that the second order condition is satisfied.

wage. Thus, as stated by Scorzafave and Menezes Filho (2001), the reservation wage plays a fundamental role in determining the entry to the labor market.

2.2 - Determinants of Labor Income

The human capital theory was used to analyze the income determinants based on the approach by Mincer (1974). This view starts from the assumption that individual earnings in any period correspond to the return on the level of their skills (human capital stock incorporated and accumulated by the individual over time). In terms of econometric estimates, the functional form proposed by Jacob Mincer to estimate the rate of return on instruction and experience, according to Chiswick and Mincer (2003) and Moretto (2000), can be specified as:

$$\ln Y_i = \ln Y_0 + \beta_1 S_i + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 S_i X_i + u_i \qquad (4)$$

$$i = 1, 2, ...$$

where :

ln Yo = natural logarithms of the earnings of the individual without instruction;

ln Yi = natural logarithms of the earnings of the individual i;

Si = years of schooling or education or formal instruction of the individual i;

Xi = years of experience on the labor market of the individual i;

ui = term of random error.

This approach was extended in the present paper in what can be defined as a hybrid model of the human capital theory. This new model was based on studies by Hoffmann and Scampini (1996) and Hoffmann and Simão (2005), with the details presented in section 3.2.

3. Methodology

3.1 - Heckman's procedure

As reported by Hoffmann and Kassouf (2005), the Heckman's procedure has become very popular recently. Heckman developed a relatively simple method to correct the possible

problem of sample selection[6], the procedure consists of estimating two equations. The first determines the decision of the individual to participate or not in the labor market, following a probit model where the Mills inverse ratio is obtained. This equation is known as the selection equation and is defined as follows:

Considering L_i^* as the difference in wages offered on the market and the reservation wage of the individual i and noting that L_i^* is a latent variable, that is, not observable, we have:

$$L_i^* = \alpha' Z_i + \mu_i \ , \qquad (1)$$

where Z is the vector of exogenous variables that affect the decision to participate or not in the labor force, and μ is the random error. Even when L_i^* is not observable, it can be verified whether a determined agent works or not, thus the decision is taken as follows:

$$\begin{aligned} L_i &= 1 \quad se \quad L_i^* > 0 \\ L_i &= 0 \quad se \quad L_i^* \leq 0 \end{aligned} \qquad (2)$$

If the individual works (L_i=1), the reservation wage of this agent is less than that offered on the market. On the other hand, if the individual does not work, his reservation wage is greater than that of the market.

Considering W now as being the wage logarithm, we have:

$$W_i = \beta' X_i + v_i \ , \qquad (3)$$

where X represents the vector of explanatory variables that affect the level of earnings and v is the random error.

According to Hoffmann and Kassouf (2005), considering that u_i and v_i have a normal bivariate distribution with average zero and standard deviation σ_u and σ_v and correlation coefficient ρ and that W_i is observed only when L_i is greater than zero, the expected value will be defined as:

[6] See Heckman (1979), Hoffmann and Kassouf (2005), Resende and Wyllie (2006) for details on sample selection problems.

$$E(W_i \mid L_i^* > 0) = \beta' X_i + \rho\sigma_v \lambda_i \ , \qquad (4)$$

where λ is the Mills inverse ratio, given by:

$$\lambda_i = \frac{\phi\left(\frac{\alpha' Z_i}{\sigma_u}\right)}{\Phi\left(\frac{\alpha' Z_i}{\sigma_u}\right)} \ , \qquad (5)$$

and ϕ and Φ are, respectively, the standard normal density function and standard normal distribution function.

It is clear that direct estimation of the earning equation(3) with only the economically active individuals would generate specification bias, known as selectivity bias. Including the Mills inverse ratio in the equation (4) explicitly takes into consideration the decision of the individual to take part or not in the labor market, so that this variable eliminates the selectivity bias and consequently gives consistent estimates of the parameters of the earning equation.

Due to the sampling characteristic of the database adopted (PNAD)[7], the sample selection model was estimated by maximum likelihood instead of the two-stage method. Estimation by maximum likelihood can generate inconsistent estimates if the variance of the random error is heteroscedastic, according to Greene (2003). Thus the test was carried out for each explanatory variable of the model.

It was further observed in the sample selection model that the participation equation is indeed a probit model that aims to analyze the factors that influence the probability of an individual being on the labor market. It was also emphasized that to avoid problems of multicolinearity, the explanatory variables considered in the earning equation were a subset of those considered in the selection equation.

3.2 - Description of the Equations and Selected Variables

The participation equation was defined as:

[7] The PNADs sample plan is not a simple random sample, it is a complex sample as it will be discussed in a further section.

$$Z_i = \alpha_0 + \alpha_1 income\_fa + \alpha_2 fam\_comp + \alpha_3 D_{child} + \alpha_4 D_{gender} + \alpha_5 age_i + \alpha_6 age_i^2 + \alpha_7 D_{color_i} + \alpha_8 EDU_i + \alpha_9 ELEDU_i + \alpha_{10} D_{urban} + \alpha_{11} D_{region_i} + u_i$$

(6)

where,

Z_i = *dummy* variable concerning the condition of activity of the individual i, that is, economically active (1) or not economically active (0);

$income\_fa$ = Represents the monthly family earnings of the individual i;

fam_comp = A number of components in the family of individual i;

D_{child} = *dummy* variable that presents value 1 if there is a child 5 to 17 years old in the family, 0 if not;

D_{gender} = *dummy* variable for gender, where the control category is the female gender;

age_i = age of the individual i in tens of years (following the methodology proposed by Hoffmann and Simão, 2005);

age_i^2 = age squared in tens of years;

EDU_i = years of education of the individual i;

$ELEDU_i$ = threshold effect of the years of study of the individual i;

$D_{regions_i}$ = set of *dummy* variables to describe the geographic regions of Brazil, divided into: Central West, Southeast, North and South, and the Northeast region is the control;

D_{color_i} = set of *dummy* variables to describe color or race of the individuals, divided into: black and indigenous, mulatto and yellow. The white color was used as control. Color or race was clustered because of the low representiveness of the yellow and indigenous individuals in the sample;

D_{urban} = *dummy* variable for the place of residence of the individual, divided into rural and urban, and the first was the control variable;

ui = random error.

The earning equation was specified in the following form:

$$\ln\_income_i = \beta_0 + \beta_1 D_{gender} + \beta_2 age_i + \beta_3 age_i^2 + \beta_4 D_{color_i} + \beta_5 EDU_i + \beta_6 ELEDU_i + \beta_7 D_{urban} + \beta_8 D_{region_i} + \beta_9 \lambda_i + v_i$$

(7)

where,

$\ln\_income$ = natural logarithm of the income of all the jobs of the individual i;

λ_i = Mills inverse ratio;

v_i = random error.

The other variables follow the same specifications mentioned above.

Some aspects of the model and the variables selected should be emphasized:

i) The sample expansion factors used in all the econometric analysis was available with the PNAD;

ii) The use of the activity condition variable can generate questioning regarding this choice as a proxy for work supply, thus some arguments should be presented. The objective of considering the activity condition variable and not the condition in occupation aims to be the most coherent as possible with the theory. That is, when analyzing the economically active and the economically inactive individuals, the option of the individual to offer work was taken into account even if this individual was not employed at the time of the interview. In spite of being unemployed, it should be remembered that this condition is temporary and that the agent has already exercised his choice of offering hours of work and took effective search measures (IBGE[8], 2007). On the other hand, if the variable condition of occupation was considered as proxy for the supply of work, one would be omitting part of the sample that decides to offer work but does not find employment;

iii) The functional log-normal form for the earning equation was chosen due to the expectation that the income does not vary linearly with the variables related to productivity and experience. Authors including Mincer (1974 and 1993), Chiswick (2003), Hoffmann (1996 and 1998), Ney and Hoffmann (2003) and Teixeira (2006) support the use of this functional form, even when discriminatory variables are incorporated and the functional form adopted would not present serious problems because the sample used is considerably high, as will be verified;

iv) To analyze the earnings, the individuals who did not declare income were excluded. In PNAD the non-declaration of income is computed as R$ 999.999.999.999[9], that is, the inclusion of these values would bring considerable biases to the analysis;

v) The variable age of the person was measured in tens of years, and also the square of this variable, bearing in mind that Y does not vary linearly with age (Ney and

[8] Brazilian Geographical and Statistical Institute (IBGE).
[9] IBGE (2007).

Hoffmann, 2003 and Hoffmann and Simão, 2005). Age is measured in tens of years to prevent the coefficients from being very small. These two variables were included in the model to measure the experience of the individual and the impact on the entrance into the labor market and the earnings. The results show the mean age when the worker tends to receive the maximum return in the life cycle;

vi) Regarding the years of study, the observations with value equal to 17 (undetermined or undeclared) were not included for reasons similar to item (ii); the years ranged from 1 to 16 years of study (PNAD, 2006). A restrictive factor of the education variable, for this database, was the impossibility of determining the teaching quality. As stated by Behrman and Birdsall (1983) *apud* Resende and Wyllie (2006), failure to consider the quality factor would lead to an overestimation of the return for education. Furthermore, it is expected *a priori* that there are differences in terms of public and private teaching (in this case, depending whether it is elementary, high school or university) and in regional terms;

vii) Individuals were excluded from the analysis when they did not declare color and when they did not fit as members of the family.

3.3 - The Threshold Effect of Education

The education threshold effect can be considered as the growing impact of the years of study on income starting at a determined value, that is, "the value of schooling starting at which the rate of schooling becomes greater" (Hoffmann and Simão, 2005). The education threshold can be described as follows:

$$S^* = D(S - L) \qquad (8)$$

where:

S^* = education threshold effect;

S = years of study of the individual;

L = education threshold that, according to Ney and Hoffmann (2003), should be around 10 years of study;

D = a *dummy* variable that assumes zero value for $S \leq L$ and assumes value 1 for $S > L$.

Considering that $S = EDU_i$ and $S^* = ELEDU_i$, and K the other variables and coefficients of the earning equation, the value of the income logarithms of all employment will be:

$$\ln\_renda_i = K + \beta_5 EDU_i + \beta_6 ELEDU_i \quad (9)$$

When $S \leq L$, D = 0 and the equation is reduced to:

$$\ln Y_i = K + \beta_5 EDU_i \quad (10)$$

that is, if the education threshold is not considered, each additional year of study is associated to an increase of [exp(β_5)-1]100 in the income of the individuals.

When $S > L$, D = 1, expression (5) becomes:

$$\ln Y_i = K - \beta_6 L + (\beta_5 + \beta_6) S \quad (11)$$

Thus, after the threshold level, each further year of schooling will cause a return on earnings of $[\exp(\beta_5 + \beta_6) - 1]100$.

3.4 - Considerations on the Database

The database used in this paper consisted of the micro data of the National Household Survey (PNAD) supplied annually by the Brazilian Institute of Geography and Statistics, IBGE. The PNAD micro data are individual data of the main social and economic characteristics of individuals and families, some permanent in character, such as the general characteristics of the population, education, work, income and housing, and others with variable periodicity, such as the characteristics on migration, fertility, marriage, health, nutrition and other subjects that are included in the system according to the Brazilian information needs (IBGE, 2007). September has been the reference period of the PNAD since 2000.

The National Household Survey (PNAD) is carried out by a probabilistic sample of households obtained in three selections stages: primary units - municipalities; secondary units – censual sectors; and tertiary units - household units (private households and collective housing units), according to PNAD (2006). Thus this database is characterized by being a complex sample and the incorrect treatment of the sample plan[10] would generate biased results, such as standard errors, variance, *quantis*, *percentis* and regression analysis (Silva and Pessoa, 2002). Kish (1965) developed a method to assess the impact of incorporating the effects of the sample plan known as DEFF (design effect). The method is defined as the ratio of the variance obtained considering the sample plan and the variance obtained by ignoring the sample plan (that is, the variance estimated as if the sample was random sample with replacement). DEFF values further away from 1 indicate that ignoring the sample plan in

[10] That is, to consider the sample process as being a simple random sample with replacement.

estimating the variance leads to biased and incorrect estimates. High DEFF values (>1) indicate that the "naïve" estimates of the variance obtained by ignoring the complex sample plan lead to underestimation of the true variance of the estimation, according to Silva and Pessoa (2002). A further alternative is the MEFF[11] statistic, which compares the variance estimate of the parameter obtained considering the sample plan with another from the same model but not considering weight, conglomerate and stratification[12], Skinner *et. al* .(1989).

The PNAD presents some limitations that should be considered. The questionnaire aims to capture both the earnings in cash and kind, but does not consider the value of the production for self consumption, which is an important component of the real income of small farmers according to Graziano da Silva and Del Grossi (2001) and Del Grossi and Graziano (2002). Furthermore, until 2003 the data research did not include the rural area of the Northern Region and did not allow the inclusion of an agricultural area that, although relatively small in terms of activity, is not negligible. Data collection is based on a specific reference month, September, and does not "capture the variety of agricultural activities in Brazil over the year" (Corrêa, 1998). These facts tend to underestimate the results for the individuals employed in the agricultural sector and affect the regression results. Another constraint is the declaration of less than real income, especially the highest. Hoffmann and Simão (2005) estimated that in the state of Minas Gerais for the 2000 census the declared incomes were underestimated by about 31% of their real value. In spite of these factors that cause underestimation of the results, they do not invalidate the analysis of income data from the PNAD questionnaires[13].

4. Empirical Analysis

4.1 - Descriptive Statistics

The economically active persons in Brazil, for the year 2006, represented about 62% of the considered population (156,300,000), that is, 96,906,000, compared to 59,394,000 outside the labor market. Some important characteristics can be observed in Table 1. For example,

[11] *misspecification effect*
[12] The MEFF analysis is identical to the DEFF analysis.
[13] For further details on PNAD constraints, see Hoffmann (1998), Del Grossi and Graziano (2002) and Rocha (2002).

the individuals in the labor market had an average income, in 2006, of around R$ 785.75 and a standard deviation of R$ 1,500.00, indicating high income concentration.

Family income was R$ 1,605.39, the families had on average four members and about 24% had children between five and 17 years of age. There was a predominance of the individuals in the Southeast and South Regions, which was expected, given the level of economic activity of these regions. The predominant color or race was white with 50% of the population, followed by the mulattos or yellow individuals with 43% and black or indigenous individuals with 7%.

Table 1. Descriptive Statistics of Major Variables, Brazil 2006*.

Variables	Obs	Population	Mean	St. Dev.	Min.	Max.
Southeast	118,598	79,754,039	0.43	0.49	0	1
South	61,121	27,369,544	0.15	0.35	0	1
Central West	44,771	13,314,996	0.07	0.26	0	1
North	56,244	15,080,183	0.08	0.27	0	1
Activity condition (Z)	340,914	156,300,000	0.62	0.48	0	1
Log of income of all jobs	170,468	78,395,586	6.24	1.00	1.10	11.70
Income of all jobs	190,907	88,077,045	785.75	1,499.53	0	120,000
Family earnings	401,526	182,900,000	1,605.39	2,545.63	0	138,800
Number of people in the family	408,978	186,700,000	3.83	1.61	1	16
Children from 5 to 17 years	99,795	187,200,000	0.24	0.43	0	1
Gender (masculine)	199,689	91,199,656	0.49	0.50	0	1
Age (in tens of years)	410,254	187,200,000	3.07	2.04	0	11.90
Age^2 (in tens of years)	410,254	187,200,000	13.57	15.48	0	141.61
Black or indigenous	29,939	12,908,501	0.07	0.25	0	1
Mulatto or yellow	192,416	80,706,837	0.43	0.50	0	1
Years of education	408,754	186,600,000	6.74	4.62	1	16
Threshold effect of education	408,754	186,600,000	0.78	1.56	0	6
Location (urban)	343,753	155,937,868	0.83	0.37	0	1

Source: Authors elaboration using micro data from PNAD 2006.
* values presented for mean, standard deviation, minimum and maximum values are related to the population and not to the sample.

In the sample considered, less than half were men (49%) and out of 91.2 million men, 54.9 million were in the labor market. About 42.6 million out of 96 million women in this sample were part of the labor force. These preliminary results had indicated that there is discrimination in the Brazilian labor market. This is explained by the fact that although the women represent more that half of the analysed sample, the total women that participate in the qork force is considerably less than the men, with a differential of approximately 12,3 millions. These results are partially influenced by the wage reservation differential between

men and women. Women with small children and/or with many children tend to have a higher reservation wage than men in the same situation. Consequently, it is expected that the number of economically active men be greater than the women.

With respect to the variables that represent the human capital, it was verified that the average age was around 31 years, although the data had showed a reduced level of qualification, with the years of schooling being on average 6,7 years. The low values for these variables can confirm the low average earning for all labor (R$ 785,00).

The women on average have a higher level of qualification in terms of years of studying (7) than men (6,5), according to PNAD (2006). Therefore, the criteria related to the individual's capability, in terms of years of studying, are a good proxy for the income level determination, then it is expected that the earnings for the female gender would be higher or close to the ones obtained for the male gender, in the absence of discrimination.

Finally, it is worthwhile mentioning that the number of children in the sample has considerable effect on human capital variables and Table 2 confirms it.

Table 2. Characteristics of Children Activities in Brazil, 2006.

	Activity Condition				
Age (years)	Non Eco. Active	Eco. Active	Not Considered*	Total	%
0 a 9	0	0	30,944,181	30,944,181	16.9
10	3,350,822	201,035	0	3,551,857	1.94
11	3,328,192	259,425	0	3,587,617	1.94
12	3,277,503	335,415	0	3,612,918	1.96
13	3,040,456	451,753	0	3,492,209	1.9
14	2,796,097	661,466	0	3,457,563	1.89
15	2,516,777	928,524	0	3,445,301	1.86
16	2,024,524	1,405,078	0	3,429,602	1.84
17	1,781,405	1,768,447	0	3,549,852	1.91

Source: Authors elaboration using micro data from PNAD 2006.
* Individuals from 0 to 9 years are not considered for the analysis of activity condition.

It was verified that about 31 million of the considered individuals were children aged 0 to 9, therefore they were not considered in terms of activity condition. However, of the total of more than 59 million children, about 10% were economically active.

When the participation of the individual in the labor market is considered, it can be stated that if the estimates do not consider the complex sample plan, the variance would be underestimated in all the variables, especially in the Location, DEFF statistic (Table 3). Thus

the sample plan had to be considered to obtain unbiased estimates. The analysis of the heteroscedasticity for each variable showed that out of 15 variables considered in the model, only two presented this problem, namely the Southern and Northern Regions[14]. Regarding the statistical significance, the variable related to the Central Western geographic region was not statistically significant in determining the entry of the individual into the labor market [15]. The other variables were highly significant, all at the level of 1%. The global significance statistic (F) of the model corroborated the previous results and was statistically significant at 1%, showing that this model properly described the determinants of the individual probability of entry into the labor market.

Some variables presented a negative sign regarding the probability of the individual being in the labor market. The existence of children in the family reduced the possibility of the individual being in the market. An possible explanation for this fact maybe the high participation of women in the sample, with a greater reservation wage than men due to the physical and psychological characteristics of preserving descendents; the number of components in the family also reduced the possibility of the individual being economically active. The results regarding the geographic regions showed direct similarities to the study by Hoffmann and Kassouf (2005), mainly regarding the signs. That is, the Southern Region was the only location where a positive relationship was found regarding the probability of the individual being employed. Furthermore, individuals in urban areas were less likely to be economically active than those in the rural areas. This can be explained, at least in part, by the greater possibility of the rural agent being busy in activities for his own consumption and for his own use.

On the other hand, some variables presented positive relation with the probability of the individual being economically active, especially family income. The fact that the family income, in the present paper, presented a positive relationship with the probability of the individual being economically active may, at least in part, be linked to the family social network, that is, richer families tend to relate to others with the same standard of living so that access to the labor market can be facilitated. In the literature on determinants of the entrance into the labor market for the female gender, for example in Hoffmann and Kassouf (2005) or Scorzafave and Menezes Filho (2001), the relationship of these variables was negative.

[14] Table A.1 in the Appendix presents these results.
[15] At 10% significance level.

Table 3 - Equation for the labor market participation using the Heckman's procedure, Brazil 2006.

Variables	Coefficients	St. Dev.	t	p-value	DEFF	MEFF	Marg. Effect
Constant	-2.002806	0.0403	-49.73	0.00	2.02	2.85	-
Family earnings	0.0001	0.0000	17.11	0.00	2.16	16.33	0.0000917
Number of people in the family	-0.0575	0.0022	-25.86	0.00	1.49	1.80	-0.0574129
Children from 5 to17 years	-1.0580	0.0171	-61.85	0.00	1.40	2.37	-1.060496
Gender (masculine)	0.8102	0.0097	83.42	0.00	1.82	2.76	0.8097712
Age (in tens of years)	1.2804	0.0157	81.31	0.00	1.50	2.65	1.276071
Age^2 (in tens of years)	-0.1669	0.0019	-90.13	0.00	1.46	2.78	-0.1663745
Black or indigenous	0.1701	0.0129	13.22	0.00	1.40	1.41	0.1701751
Mulatto or yellow	0.1139	0.0075	15.12	0.00	1.41	1.46	0.1142218
Years of Education	0.0442	0.0015	28.89	0.00	1.60	1.77	0.0441344
Threshold effect of education	-0.0293	0.0044	-6.70	0.00	1.51	2.20	-
Location (urban)	-0.3143	0.0172	-18.22	0.00	3.91	4.11	-0.3139364
Southeast	-0.0337	0.0135	-2.50	0.01	2.90	3.35	-0.0332672
South	0.0768	0.0159	4.84	0.00	2.49	2.96	0.0770383
Central West	-0.0258	0.0161	-1.60	0.11	1.79	2.73	-0.0256047
North	-0.0461	0.0187	-2.46	0.01	3.06	4.72	-0.0460761
Num. Strata = 652	Number obs = 293,460						
Num. PSU* = 5,628	Considered Pop. = 134,264,412						
F(12, 4965) = 1,340.26	Prob > F =		0.0000				

* Primary Sample Unit.
Source: Authors elaboration using micro data from PNAD 2006.

For the practical treatment of the selection model, the marginal effects of the variables that are not constant must be considered in the estimation of the probit model [16] and therefore the coefficients cannot be interpreted directly, according to Gujarati (2000) and Long and Freese (2006).

Beginning the analysis of the impact of the explanatory variables on the probability of the individual being in the labor market, Table 3 shows that an increase of R$1,000.00 in family income increases the probability of the individual being economically active by 9.17 p.p.. For each additional member of the family, the probability of the individual being on the labor market is reduced by 5.74 p.p.

Age and age squared showed signs in line with those found in the literature, for example, in the study by Resende and Wyllie (2006). That is, the inverted U shape in the interaction among age and age squared suggest, that the return on experience presented a maximum point after which the return in income decreased. The performance of these

[16] As it is done in the literature, the marginal effects were calculated for the sample average point.

variables reflected the depreciation of human capital growth over life cycle. In the present paper, the maximum point was around 38 years of age, with values very close to those reported by Hoffmann and Kassouf (2005) considering only the case of women. Finally, each additional year invested in education raised the probability of the individual participating in the labor market by 4.41 p.p., showing that investment in human capital is an extremely important factor for entrance into the labor market.

The analysis of the *dummy* variables in the model showed that the individual being of the male gender raised his probability of being in the labor market by approximately 81p.p. This result tends to demonstrate some segmentation in the labor market, as reported by Gandra (2002), and confirming the results obtained by the descriptive statistics. Regarding the color of the agent, the fact that the individuals were black, mulatto, indigenous or yellow raised the likelihood of this person being employed. This occurred mainly for the black or indigenous individuals, that have 17 p.p. more probability of being in the labor market than the white inidividuals[17]. This fact may demonstrate that the individuals of this color or race need to develop more activities or have a smaller reservation wage than the white individuals [18]. That will be discussed later with the earning equation.

The analysis of the impacts of the explanatory variables on the income logarithm, Table 4, showed first that the exclusion of the sample plan structure would generate biased results of variance, because the DEFF statistic[19] presented positive values for all variables, especially for the censual situation that was close to 7.00. Regarding the statistical significance, all the variables of the earning equation presented significance at 1% as seem with the F statistic. That is, the global fit of the model followed that detected for the probit model in Table 3[20]. The inverse Mills ratio (Lambda) was also statistically significant at 1% showing that its inclusion was necessary to prevent biased selectivity.

The individuals of the male gender tended to present 30% higher income than women. Although women have higher qualification than men, they presented considerably lower earnings than men, suggesting a substantial discrimination level. Confirming this analysis, Santos *et al.* (2008) reported that there is a tendency to perpetuate this differential in the years 2000. This result showed that discrimination on the labor market by gender is extremely high,

[17] Although this result seems to be good for individuals of this color or race, they do not determine if the individual is in a formal activity or in an informal one, and additionally it is not related to earnings. For more details see the methodological notes from PNAD (2006).
[18] It may again indicate market segmentation.
[19] *Design Effect.*
[20] This occurs because the selection equation and earnings equation are jointly estimated by Heckman's procedure.

and the authors further concluded that the average income of the men in 2006 was approximately 1.5 times greater than that of the women.

Ratifying the previous analyses on color and race, it was clear, as shown in Table 4, that the reservation wage of black or indigenous or mulatto or yellow individuals was less than that of the white individuals. This is justified because black or indigenous individuals tended to earn 17% less, and the yellow or mulatto individuals earned about 20% less. However, it should be considered that, in terms of average years of study, the white individuals have approximately 7.51 years compared to 6.44 and 5.59 years of study for the black or indigenous, mulatto or yellow individuals, respectively. Although the individuals resident in urban areas were less likely to be in the labor market than those in the rural areas, the average income of the first was 37% higher than the latter. Regarding analysis of the geographic regions, individuals resident in the Central Western Region had the highest income, compared to the Northeastern Region, followed by the Southeast, South and North, with values of around 55%, 48%, 45% and 39%, respectively.

Table 4 – Earnings equation using Heckman's procedure, Brazil 2006[*].

Variables	Coefficients	St. Dev.	t	p-value	DEFF	MEFF	Marg. Effect(%)
Constant	4.5294	0.0463	97.89	0.00	1.75	4.48	-
Gender (masculine)	0.2616	0.0074	35.54	0.00	1.72	2.77	29.90
Age (in tens of years)	0.2395	0.0185	12.94	0.00	1.60	4.10	27.06
Age^2 (in tens of years)	-0.0090	0.0023	-3.97	0.00	1.56	4.15	-0.90
Black or indigenous color	-0.1873	0.0100	-18.81	0.00	1.95	1.82	-17.08
Yellow or mulatto color	-0.1707	0.0056	-30.21	0.00	1.65	1.66	-15.69
Education in years of schooling	0.0500	0.0013	38.69	0.00	1.89	2.27	5.13
Threshold effect of education	0.1221	0.0028	43.21	0.00	2.16	2.27	18.79
Census situation (urban)	0.3136	0.0169	18.50	0.00	6.85	7.98	36.83
Southeast Region	0.3981	0.0124	32.19	0.00	5.71	6.04	48.90
South Region	0.3728	0.0141	26.42	0.00	4.57	5.18	45.18
Central West Region	0.4446	0.0159	28.00	0.00	4.06	5.75	55.99
North Region	0.3327	0.0163	20.46	0.00	4.65	7.17	39.47
Lambda	-0.6446	0.0114	-56.41	0.00	-	-	-
Num. strata = 652	Number obs = 293,460						
Num. PSU[**] = 5,628	Considered Pop.= 134,264,412						
F(12, 4965) = 1,340.26	Prob > F = 0.0000						

[*] The dependent variable is the logarithm of earnings of all jobs.
[**] Primary Sample Unit.
Source: Authors elaboration using micro data from PNAD 2006.

Considering the variable years of education and the education threshold, it was verified that up to 10 years of schooling, an additional year represented an average increase of 5.13% in income. Nevertheless, taking into consideration the effects of the education threshold, after 10 years of study, the increase was close to 19%. Santos *et. al.* (2008) stated that "years of study and the education threshold decreased continuously between 2002 and 2006…it was perceived that education, although extremely important to explain the income level, has become less important". Experience presented the desired sign for age and age squared, showing an inverted U shape relationship, representing a parabolic pathway for experience, suggesting that human capital depreciates over time.

5. Conclusions

The results had showed that considering the PNAD data as a complex sample plan was extremely important to obtain robust and unbiased statistics of variance of the regressors. Furthermore, the importance of the Mills inverse ratio made it possible to avoid selectivity bias when estimating the earning equation.

Regarding the determination of participation in the labor market, only the variable concerning the Central Western Region was not significant for the model. All the variables were significant at 1% for the earning equation.

In the present paper the data indicated traces of segmentation in the market by race or gender. For the first, it was verified that black and mulatto individuals presented greater probability of being on the labor market than the white individuals. However, in terms of income the black and mulatto individuals presented on average considerably lower income than that of white individuals. On the other hand, the white's years of studying are relatively higher than the other groups. In terms of gender, although the women had presented a higher level of qualification, the male gender had 30% higher earnings. Additionally, men had a higher probability of being in the labor market in comparison to women. These results show a major problem in the Brazilian labor market, which is the worker's distinction based on characteristics that are not related to the individual's productivity.

Finally, from the analysis of the investments in human capital, it was found that the proxy of experience performed as expected, with a parabolic pathway indicating that the returns of experience increase to a maximum point and then decrease, corroborating the theory that human capital depreciates over time. Regarding education, it was confirmed that the level of schooling has greater impact on the individual earnings at 10 years of study or

more, showing that investments in education are one of the main sources of income generation.

6. References

BARROS, R. P., MENDONÇA, R. A evolução do bem-estar e da desigualdade no Brasil desde 1960. In: Teixeira, E. C. Desenvolvimento agrícola na década de 90 e no século XXI. Viçosa: 1993.

BARROS, R. P., MENDONÇA, R. Os determinantes da desigualdade no Brasil. Rio de Janeiro: IPEA, 1995(a).

BARROS, R. P., MENDONÇA, R. Bem-estar, pobreza e desigualdade de renda: uma avaliação da evolução histórica e das disparidades regionais. Rio de Janeiro: IPEA, 1995(b).

BARROS, R. P. de, CARVALHO M. de, FRANCO, S., MENDONÇA, R. A queda recente da desigualdade de renda no Brasil. In: HENRIQUES, (org). Rio de Janeiro: IPEA, 2007.

CHISWICK, B. R., MINCER, J. Experience and the distribution of earnings. University of Illinois at Chicago and IZA Bonn, 2003. Available in: <http://ssrn.com/abstract=435260>

CORRÊA, A. M. C. J. Distribuição de renda e pobreza na agricultura brasileira (1981-1990). Piracicaba: Editora UNIMEP, 1998. 260 p.

DEL GROSSI, M. E., GRAZIANO, J. S. O uso das PNADs para áreas rurais. Rio de Janeiro: IPEA, **Texto para Discussão** 874, Abril de 2002.

GANDRA, R. M. O debate sobre a desigualdade de renda no Brasil: da controvérsia dos anos 70 ao pensamento hegemônico nos anos 90. Dissertação de Mestrado. Niterói (RJ): UFF, 2002.

GRAZIANO DA SILVA, J., DEL GROSSI, E. O novo rural brasileiro: uma atualização para 1992-98. IE/Unicamp.**Texto para discussão**. 2001.

GREENE, W. H. Econometric analysis. New York: Pearson, 2003. 1026p.

GUJARATI, D. N. Econometria básica. Terceira Edição. São Paulo: Macro Books, 2000.

HECKMAN, J. J. Sample selection bias as a specification error. **Econometrica**, Vol. 47, No. 1, Jan. 1979, pp. 153-161.

HOFFMANN, R. Evolução da distribuição de renda no Brasil, entre pessoas e entre famílias, 1979 e 1986. In: SEDLACEK,G. L., BARROS, R. P. (eds). Mercado de trabalho e distribuição de renda: uma coletânea. Rio de Janeiro: IPEA/INPES, 1989.

_____. Vinte anos de desigualdade e pobreza na agricultura brasileira. **Revista de Economia e Sociologia Rural,** v. 30, n.2, p. 97-113, abr./jun. 1992.

HOFFMANN, R., SCAMPINI, P. J. Desigualdade e pobreza na agricultura do estado de Minas Gerais. **Nova Economia**. Belo Horizonte, v.6, n.2, p.67-84, nov. 1996.

HOFFMANN, R. Distribuição de renda: medidas de desigualdade e pobreza. São Paulo: Edusp, 1998.

_____.Mensuração da desigualdade e da pobreza no Brasil. In: HENRIQUES, R. (eds) Desigualdade e pobreza no Brasil. Rio de Janeiro: IPEA, 2000.

HOFFMANN, R.; LEONE E. T. Participação da mulher no mercado de trabalho e desigualdade da renda domiciliar per capita no Brasil: 1981-2002. **Nova Economia**. Belo Horizonte.14 (2) 35-58. maio-agosto de 2004.

HOFFMANN. R., SIMÃO, R. C. S. Determinantes do rendimento das pessoas ocupadas em Minas Gerais em 2000: o limiar no efeito da escolaridade e as diferenças entre mesorregiões. **Nova Economia,** v. 15, n. 2, p. 35-62, maio/ago. 2005.

HOFFMANN, R., KASSOUF, A.L. Deriving conditional and unconditional marginal effects in log earnings equations estimated by Heckman's procedure. **Applied Economics,** Londres, v. 37, n. 11, p. 1303-1311, June 2005.

IBGE, Instituto Brasileiro de Geografia e Estatística. 2007. disponível em: ,<http://www.sidra.ibge.gov.br/bda/tabela/protabl.asp?z=p&o=16&i=P>.

IPEA-CAIXA. As desigualdades nos retornos do ensino superior no Brasil. Rio de Janeiro: IPEA, 2007.

KISH, L. Survey Sampling. New York: Wiley, 1965.

LONG, J. S., FREESE, J. Regression models for categorical dependent variables using stata. Second Edition. College Station Texas, 2006.

MINCER, J. Schooling, experience and earnings, New York: National Bureau of Economic Research, 1974.

MINCER, J. Studies in human capital. Aldershot; Vermont: Edward Elgar, The Collected essays of Jacob Mincer: v. 1 (Economists of the twentieth century), 1993.

MORLEY, S. The income distribution problem in Latin America and the Caribbean. ECLAC, 2001. Disponível em: <http://www.eclac.cl/cgibin>.

MORETTO, C. F. Função minceriana de determinação dos rendimentos individuais: uma aplicação do método de variáveis instrumentais. **Teoria e Evidência Econômica**, v. 8, n. 15, p. 47-65, 2000.

NEY, M. G., HOFFMANN, R. Desigualdade de renda na agricultura: o efeito da posse da terra. **Economia,** v. 4, n. 1, p. 113-152. NPEC, Niterói, jan./jun. 2003.

PESQUISA NACIONAL POR AMOSTRA DE DOMICÍLIOS 2006. Brasil. Rio de Janeiro: IBGE, v. 27, 2007.

PNUD. Relatório do Desenvolvimento Humano 1999. Disponível em: <http://www.pnud.org.br/rdh/rdh99/index.php>.

RIBEIRO, C. A. C. Um panorama das desigualdades na América Latina. Rio de Janeiro, IUPERJ/UCAM, 2006. Disponível em: < http://observatorio.iuperj.br/>.

RAMOS, L., VIEIRA, M. L. Desigualdade de rendimentos no Brasil nas décadas de 80 e 90: evolução e principais determinantes. Rio de Janeiro: Ipea, **Texto para Discussão** 803, 2001.

RESENDE, M.; WYLLIE, R. Retornos para educação no Brasil: evidências empíricas adicionais. **Economia Aplicada**, Ribeirão Preto, v. 10, n. 3, 2006 . Disponível em:<http://www.scielo.br/scielo.php?script=sci_arttext&pid=S141380502006000300003&lng=pt&nrm=iso>. Acesso em: 10 Jul 2008. doi: 10.1590/S1413-80502006000300003.

ROCHA, S. A investigação do rendimento na PNAD – comentários e sugestões à pesquisa nos anos 2000. Rio de Janeiro: IPEA, **Texto para Discussão** 899. Agosto de 2002.

SANTOS, G. C. , BASTOS, P. M. A., ROCHA, L. E. V. Determinantes da renda do trabalho no Brasil no período de 2002 a 2006. In: **XLVI Congresso da SOBER**, 2008, Rio Branco. XLVI Congresso da SOBER, 2008.

SCORZAFAVE, L. G., MENEZES FILHO, N. A. Participação feminina no mercado de trabalho brasileiro: evolução e determinantes. **Pesquisa e Planejamento Econômico**, Rio de Janeiro, v.31, n.3, p. 441-477, 2001.

SILVA, P. L. do N., PESSOA, D. G. C., LILA, M. F. Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. **Ciênc. Saúde Coletiva**, 2002, vol.7, no.4, p.659-670. ISSN 1413-8123.

SKINNER, C., Holt, D. & Smith, T. (1989), Analysis of Complex Surveys, John Wiley & Sons.

TEIXEIRA, W. M. Equações de rendimentos e a utilização de instrumentos para o problema de endogeneidade da educação. Faculdade de Economia, Administração e Contabilidade (FEA). Disponível em <http://www.teses.usp.br/teses>, 2006.

7. Appendix

Table A.1 – Heteroscedasticity Test

Variables	Chi2(1)	Prob > chi2
Activity condition	0,00	0,9989
ln(earnings from all jobs)	0,00	1,0000
Family earnings	0,00	1,0000
Number of people in the family	0,00	0,9896
Children from 5 to 17 years	0,00	0,9950
Gender (masculine)	0,00	0,9998
Age (in tens of years)	0,00	1,0000
Age^2 (in tens of years)	0,00	0,9975
Black or indigenous color	0,00	0,9698
Yellow or mulatto color	0,00	0,9994
Education in years of schooling	0,00	1,0000
Threshold effect of education	0,00	0,9806
Census situation (urban)	0,00	0,9853
Southeast Region	0,00	1,0000
South Region	0,00	-
Central West Region	0,00	1,0000
North Region	8,90E+08	0,0000

Source: Authors elaboration using micro data from PNAD 2006.