WILHELM MATHEUS HARTZE

RETURNS TO EDUCATION IN BRAZIL: A PSEUDO PANEL AND SAMPLE SELECTION BIAS APPROACH

Sorocaba-SP

2020

WILHELM MATHEUS HARTZE

RETURNS TO EDUCATION IN BRAZIL: A PSEUDO PANEL AND SAMPLE SELECTION BIAS APPROACH

Thesis presented to the Graduate Program in Economics in order to obtain the title of Master in Applied Economics.

FEDERAL UNIVERSITY OF SÃO CARLOS
SCHOOL OF MANAGEMENT AND TECHNOLOGY
GRADUATE PROGRAM IN APPLIED ECONOMICS

Advisor: Prof[a] Dr[a] Andrea Rodrigues Ferro

Sorocaba-SP
2020

HARTZE, WILHELM MATHEUS

RETURNS TO EDUCATION IN BRAZIL: A PSEUDO PANEL AND SAMPLE SELECTION BIAS APPROACH / WILHELM MATHEUS HARTZE. -- 2020.
79 f. : 30 cm.

Dissertação (mestrado)-Universidade Federal de São Carlos, campus Sorocaba, Sorocaba
Orientador: ANDREA RODRIGUES FERRO
Banca examinadora: MARIUSA MOMENTI PITELLI, ELAINE TOLDO PAZELLO
Bibliografia

1. Pseudo panel. 2. Sample selection bias. 3. Returns to education. I. Orientador. II. Universidade Federal de São Carlos. III. Título.

Ficha catalográfica elaborada pelo Programa de Geração Automática da Secretaria Geral de Informática (SIn).
DADOS FORNECIDOS PELO(A) AUTOR(A)
Bibliotecário(a) Responsável: Maria Aparecida de Lourdes Mariano – CRB/8 6979

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências em Gestão e Tecnologia
Programa de Pós-Graduação em Economia

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Wilhelm Matheus Hartze, realizada em 03/02/2020:

Profa. Dra. Andrea Rodrigues Ferro
UFSCar

Profa. Dra. Mariusa Momenti Pitelli
UFSCar

Profa. Dra. Elaine Toldo Pazello
FEARP - USP

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Elaine Toldo Pazello e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Profa. Dra. Andrea Rodrigues Ferro

Acknowledgements

I would like to thank everyone that was part of this amazing journey. My parents, Wilhelm and Adriana, who have always supported me in every possible way. My sister and her husband, Isabela and Vinícius, who have always encouraged me to keep going forward.

Everyone at the Department of Economics at UFSCar Sorocaba, specially the professors, whose passion for science and education helped me broaden my view of the world. CAPES for the financial support to develop this research.

The members of the examination board for the great contributions to this work. A special thanks to Dra. Andrea Ferro for being an amazing advisor, helping with the development and guidance of this work and teaching how to think deeply about social problems.

My classmates who shared the hardships and joy of this process with me. Our friendship made everything better.

Above everything and everyone, God. He has never forsaken me.

"For of him, and through him, and to him, are all things:
to whom be glory for ever.
Amen.
(Holy Bible, Romans 11.36)

Resumo

HARTZE, Wilhelm Matheus. *Returns to education in Brazil: A pseudo panel and sample selection bias approach.* 2020. 79 f. Dissertação (Mestrado em Economia Aplicada) - Centro de Ciências e Tecnologias para Sustentabilidade, Universidade Federal de São Carlos, Sorocaba, 2020.

O objetivo do presente trabalho é analisar os retornos à educação no salário-hora no Brasil, para o período de 2011 a 2015, os últimos anos da PNAD. Apesar de ser um tema que vêm sendo estudado há anos, a precisão dos resultados de modelos tradicionais, como o de Mínimos Quadrados, deve ser checada devido a presença de viés de seleção amostral e de variáveis não observadas. A hipótese testada é de que há efeitos significativos de ambos os fatores, levando a estimadores viesados. Através da abordagem de pseudo painel, possibilitado pelo uso de coortes, verificou-se que a heterogeneidade coletiva leva a uma subestimação dos retornos pelo modelo WLS. Apesar de haver diferença significativa entre os modelos tradicionais e os de viés de seleção não é possível afirmar com unanimidade de resultados a direção do viés. Os resultados obtidos pelos estimadores de dois estágios mostram retornos menores para homens e mulheres em relação ao modelo WLS, enquanto que o método de Máxima verossimilhança apresentou resultados divididos.

Palavras-chave: Pseudo painel. Viés de seleção amostral. Retornos à educação.

Abstract

The goal of this work is to analyze the returns to education in the hourly wages in Brazil for the period of 2011 to 2015, last five years of PNAD data. Even thought the relationship between years of education and wages has been long studied, the accuracy of the results from traditional models like the Ordinary Least Squares are to be questioned due to the necessity to check the presence of group heterogeneity and sample bias. The hypothesis tested in this work is that there is significant effect from non observable factors that leads to biased estimators if not taken into account. By estimating a pseudo panel, through the creation of cohorts, it is seen that the group heterogeneity leads to underestimation of the returns to education from the WLS model. The sample bias also affects the magnitude of the coefficients, however, the two different models used to capture and correct the estimators present different results. The two step model reports smaller returns in comparison to the WLS benchmark, while the Maximum Likelihood model had mixed results.

Keywords: Pseudo panel. Sample selection bias. Returns to education.

List of Figures

List of Tables

List of abbreviations and acronyms

PNAD	National Household Sample Survey
GDP	Gross Domestic Product
LSMS	Living Standard and Measurement Survey
OLS	Ordinary Least Squares
ML	Maximum Likelihood
US	United States
PME	Monthly Employment Research
IBGE	Brazilian Institute of Geography and Statistics
IPCA	Broad Consumer Price Index
WLS	Weighted Least Squares

Contents

1 Introduction

Understanding the relationship between a person's educational level and their income has been the subject of worldwide study for a long time, with first works on the matter being published as early as the 1950s, as seen in the works of Sargan (1957), Lebergott (1959) and Zwick (1957). Even with the existence of extensive literature, studies about the returns to education continue to be developed, either to consolidate discoveries already found with new econometric techniques (ZHANG, 2019) or to test new hypotheses (GUO; ZHANG; YE, 2019), the scientific effort directed to this field of study shows that there are still many questions to be answered.

The idea that people with better educational background have higher wages is scientifically verified by numerous researches. But, what is the effect of education on the salary of the Brazilian people in the recent period? Do men and women have different gains from more years of study? Is there a significant effect from group heterogeneity or sample selection bias? To answer these questions, the present work proposes to use the linear regression fixed-effects estimator from pseudo panel data by creating 352 different cohorts from the National Household Sample Survey, stratifying the groups by characteristics that have significant effects on the returns, like year of birth (age), skin color, sex and rural or urban workers, in order to obtain an overview of the returns to education in Brazil accounting for group heterogeneity. To account for the sample selection bias effect, the two step model from Heckman (1979) and the Maximun Likelihood model are estimated. These three results are then compared to the a Weighted Least Squares, used as benchmark.

There has been a continued wealth growth, in terms of GDP[1] and GDP per capita[2] over the years. But how can this growth be sustained over time? In the early models of economic growth such as the ones presented by Swan (1956) and Clark (1949) it was considered that the product of an economy was a function of two main factors, capital and labor. Thus, by accumulating capital (such as buying more inputs or more machines) and increasing working hours would lead to economic growth, allowing the achievement of new levels of wealth and consequently higher levels of quality of life.

However, Kuznets (1966) realizes that the constant economic growth that has been sustained for so long is not simply a result of buying more inputs or making people work more (or harder), because the benefit would be marginal, much smaller than what has been seen. So what is the unknown source that has been maintaining the level of economic growth in recent

[1] World GDP has grown approximately $100 billions in the last two decades. Starting at US$28,956 trillion in 1990 going to US$128,756 trillions in 2017 according to (The World Bank, 2017)

[2] Data from (The World Bank, 2017) shows a continue growth for GDP per capita in US$. From 1960 to 2017, the GDP per capita has grown more than US$10.000,00 going from US$452,36 to US$10.748,71.

years? By considering the theory behind the initial economic growth model, education would be able to positively affect both the marginal product of labor and capital as stated by Nelson e Phelps (1966), Becker, Murphy and Tamura (1990) and Barro (2001).

A well-educated person would be able to perform better at work and produce the same as a person with less education, in less time. In this way, it allows him to produce more, comparatively, in a workday of the same duration. In the same way, people with more education, would be able to understand better how a machine works and it could make improvements in such equipment making it to be more efficient, productive, producing more at a lower cost. Thus, technological expansion produced from time to time by investing in studying and increasing understanding of technology would also allow an increase in production. Education would then affect the two fundamental factors of economic growth and would not continue to be seen as just something to improve the individual himself, but also as a form of capital.

Not only Kuznets (1966) but also Schultz (1960) thought about the mystery behind the continued economic growth since both capital and labor are limited. By understanding the importance of education for the economy, the author proposes that education should be considered as a form of investment and its consequences a form of capital, so that the term equivalent to education, considering this economic aspect, would be human capital. The hypothesis behind this theory would be that important unexplained increases in national income, due to production factors of traditional economic development models, such as labor and physical capital, would be the result of an increase in the stock of human capital.

Nowadays, human capital is seen as a set of skills, which are characteristics of individuals, that can be used to obtain financial resources through paid work, as Josan (2012) points out. It is possible to improve these skills by investing in human capital (investing in some kind of education) so that the final result (work) is performed more efficiently and increasing the productivity, leading to products with better quality. This benefits the individual, who will be rewarded financially by the work, and the economy of the country as well, since it will increase the quantity and quality of its production. However, the definition of human capital today is much wider than when it was first introduced.

The formalization of the relationship between education and remuneration was initially carried out by Schultz (1961). The author notes that, although knowledge and skills are considered to be important tools for all, they were not yet seen as a form of investment and, like any investment, it would have positive effects on the economy. Increasing human capital would be responsible for improving the results for different activities, increasing labor productivity, which would lead to higher rates of economic growth as Schultz (1960) and Becker (1962) points out. The effects of this investment would also affect the income of each worker, because, due to the increase in productivity, they would receive higher salaries when compared to the less qualified workers.

From the human capital theory, modeling the returns to additional years of study began

to be developed. The function created by Mincer (1974) is composed of a unique equation model that aims to capture the relationship between income (in logarithm to capture the percentage effects) as a function of schooling and experience. The hypothesis tested was that the years of study and experience of the individual would have positive effects on the wages, since they would be more instructed with better working techniques, increasing their productivity and consequently their wages.

The author observed that there was a positive difference in the wage differential of white men from urban areas to different educational groups, according to age and experience, indicating that these factors affected wages. The level of schooling of a person had a positive effect on wages, although at a younger age the relationship was opposite, due to the high costs of studying. The returns responded positively to age and it was found out that there was a positive correlation between age and educational level. It was also possible to verify the positive relation between investment in post-school education and wage, empirically contributing to the theory of human capital.

According to Card (2001) and Duflo (2001), returns to education are higher in Global South countries compared to its Global North counterparts, with the caveat that much of the literature corroborating this assertion was produced in a period where statistical concepts such as bias selection and sample weight had not been developed yet. Initially, it seems logical to follow this line of reasoning that leads to such an assertion. Global North countries have greater accessibility of high-quality education, this way, comparatively, the skilled labor market tends to be in balance, or at least have a smaller shortage of labor supply for these jobs than in countries with lower educational levels.

To have a better understanding on where the returns to education are at, in different countries of the world, the work of Peet, Fink and Fauzi (2015) investigates 25 different developing countries from Africa, Asia, Eastern Europe and Latin America, through the LSMS survey, between 1985 and 2012, in order to verify if there is any significant difference from the results found and the returns to education in developed countries found by other works.

Using a pooled OLS, the results indicate that, on average, developing countries presented percentage returns of 7,6%, which is similar to countries like United States and Central Europe, that showed a rate of return of 6,6% . It was argued that, this is a evidence that the gap between returns to education in Global North and Global South countries are closing. There was also no declining in the returns over time, which, according to the authors, indicates that there are no unbalance between supply and demand of highly educated workers.

The average returns per continent showed the following results: Africa (9,6%), Latin America (8,6%), Eastern Europe (6,3%) and Asia (4,4%). It should be noted that there has been significant differences between countries in the same continent, as in the returns in Iraq (0,9%) and Pakistan (9,8%). The country with the highest return among all was Peru (13,6%). Interestingly enough, the results showed that the returns to one year of education were higher

to females (8,6%) than to males (7,1%), which is the opposite of to what was found here in Brazil by Davanzo and Ferro (2014) and that the difference between urban to rural residents was 1 percentage point (7,9% to 6,9% respectively).

The difference between the returns of different educational levels is shown by Adams (2017), who studied the case of Slovenia using longitudinal registration data of matches between employers and employees for the period of 1993 to 2007. The groups were divided into primary education (14,53% of the total population), secondary education (28,85%), secondary general education (29,13%) and tertiary education[3] (23,87%).

It was found that returns to education in all categories were increasing during the period from 1993 to 2000 when it reached its apex. For secondary education, the returns started at approximately 14% (1993) growing to 21% in 1998 when it began to decline. The returns remained practically constant, with a slight annual decrease until the year of 2007, where it reached 18%. For secondary general education, the returns started already higher than in comparison to the previous category, at a rate of 45% in the year 1993, which showed a growing trend until 1998, when it reached 58%. Starting in 2000, the downward trend started from 60% to just over 50% in 2007. Finally, for the highest level of education (tertiary education) returns were the highest of all. In the period from 1993 to 1998, returns to education for this group had the biggest increase, going from 100% in 1993 to 145% in 1998. However, its decrease in the second period was also significant. In 2000, the effect of having tertiary education was 160% but in 2007 was 15 percentage points less. These findings suggests that there has been a downward trend to returns in the latest years for all groups, regardless of their educational level.

Analyzing the returns for Brazil in the period from 1981 to 2004, Barbosa Filho and Pessoa (2008) used the Internal Rate of Return method to calculate the marginal effect of years of study on wage differential, following the hypotheses that the rate of return of education is independent of the individual being employed or not and that education raises productivity both at work and in other activities. In the estimation, it was considered levels of schooling from 1 to 15 years of study for individuals with 30, 40 and 50 years of work. The rates of returns across groups have remained virtually constant at almost all levels of education. The lowest returns to schooling were those who had only 1 year of study while the highest returns varied according to the number of years that the individual was at work. For the group of 30 years the highest return was 5 years of study with a 27% average return.

Considering the 40 years group, the highest return was for people with 11 years of study with a marginal effect of 28,8% in their income, while for 50 years was for individuals with 5 years of study, with 27,4%. In the analysis by level of education, the highest returns found was in the group of 8 to 11 years, which corresponds to the complete secondary education.

[3] Secondary general education refers to high school with teachings focus on the market while tertiary is any type of education beyond high school level.

Regarding the evolution of returns over the years, comparing to the first period (1981) and the last one (2004), all groups presented lower returns in the last year, except for the group 4 to 8 years of study, which increased 1,7 percentage points, corroborating with the hypothesis that, as years go by, the positive effects of year of study on wages are getting smaller.

Menezes Filho (2001) studying a PNAD data set from 1997, argues that, the expansion of education in Brazil is affecting directly employment and the returns to years of study. The increase in the number of people with better (more years of study) education leads to a higher supply of qualified workers, which is not necessarily followed by an increase in the demand of these kind of jobs, which means that there is an excess of labor supply. Consequently, people with this characteristics will earn less and will even be unemployed.

Davanzo and Ferro (2014) also studies the effects of the growth in number of workers with higher educational level. It is investigated if this expansion led to a reduction in productivity, due to loss in education quality and how different groups would react to it. The authors separates individuals into four categories (men and women from the rural and urban regions) and analyze the effects of education, experience, formal work, public sector, metropolitan region and Brazilian macro-regions. The Oaxaca-Blinder decomposition was used to see if there was differences in the productivity for each of the four groups and how much of the difference in wages was due to differences in characteristics. If there was a decrease in the productivity of the groups, the increase in the number of universities in Brazil was not accompanied by an increase in the quality of education. The results show that returns to education for the period from 2001 to 2012 fell for all four groups, but nothing could be said about the quality of the education itself.

As Wansuri and McNown (2010) argues, one of the biggest problems when analyzing returns to education, especially when it comes to standard methods like linear regression, is the endogeneity of years of study. It is most likely that the reason why an individual choose to study that amount of years is due to other non observable variables, like innate ability or parental influence.

An alternative to eliminate the individual heterogeneity bias that may rise when it comes to returns to education, is to estimate a model that takes such effect into consideration, like a fixed or random effects model from panel data. The issue is that, there is a lack of genuine panel data, especially for countries in development like Brazil, due to the high cost and low convenience of having to interview the same individual every time [4]. Not only that, the panel data sets that are available, like the continuous PNAD, have friction problems that greatly reduces the effectiveness and the statistical inference from the panel, due to inconsistency of the answers[5] as seen by Ribas and Soares (2008) which causes the panel to suffer from major loss of data and possibilities of miss specification for the individuals.

[4] To characterize as panel data, the same individual must be followed overtime

[5] A person may answer differently the same question in two different periods

A possible solution is the use of pseudo panel approach, which uses repeated cross section data, that is usually more common and easily accessible, even for Global South countries. This way, it is possible to mimic the advantages [6] of traditional panel, by stratifying the sample into groups that shares the same characteristics and allows the researcher to estimate a fixed effect like [7] model, in order to take into account individual heterogeneity and eliminate its bias in the estimators.

Another issue that may also rise when it comes to estimating unbiased returns to education is the sample selection bias. As Heckman (1979) shows, there are specific groups that can not get jobs due to the presence of certain characteristics that affects negatively the insertion of a certain individual in the labor market. This means that part of the sample is censored, and because of this, only individuals with similar characteristics will be part of the sample with jobs. Since the main source of income are wages, the sample used for the study of returns to education will not be random, but biased.

In order to take this effect into account it is necessary to insert an auxiliary equation to correct the parameters from the main equation. As the main problem for the returns analysis is the probability of the person being employed, a probit model to estimate the likelihood of a person having a job according to the presence of specific characteristics is estimated and the results used to correct the estimators from the main equation (returns).

This work contributes to the literature by investigating the three most important phases of the returns to education, according to Marcelo and Wyllie (2006), using the 5 latest years of PNAD, one of the biggest survey data sets in Brazil. Three different types of models are used in order to investigate the returns to education accounting for different factors that may affect the returns. First, by following a similar approach as to classical works like Mincer (1974) and Griliches (1977) a Weighted Least Squares is estimated to be used as benchmark, since it is one of the most common econometric methods in economics. Then, by accounting the existence of sample selection bias first studied by Heckman (1979), two models are estimated. The Maximum Likelihood and the Two step Heckman Sample Selection model, like suggested by Nawata (1994). Lastly, to account for the group heterogeneity effect, a cohort stratified analysis using pseudo panel is used to estimate the returns to education to 352 different groups created by determinant wage factors (WILLIS, 1986), which are: year of birth, sex, skin color and rural or urban worker.

After estimating all of the models, a comparison is carried one to see the differences between estimators from models that account for different issues. This way it will be possible to test the first hypothesis, that there are significant differences in the returns due to the

[6] Been able to take into account individual heterogeneity and consider both time and regional effects are the biggest advantages of panel data. Pseudo panel can perform similarly, as it will be shown in the methodology section

[7] Considering that both are corrected by the mean, but pseudo panel takes the mean of a group, namely cohorts, instead of individuals.

presence of collective heterogeneity and sample bias, making the traditional Least Squares approach biased and that the magnitude of the effects of years of education are different for men and women for every approach. The second hypothesis that both men and women respond differently to both biases and face distinct challenges when it comes to employment and wages.

Besides this introduction, chapter 2 talks about the concept of human capital and the importance of education to income, the pseudo panel approach and results found using this method and lastly the sample selection bias problem. Section 3 gives an overview of the Brazilian work force. Section 4 explains the methodology and data used. Section 5 reports the results for every model and compares them. Section 6 is the concluding remarks.

2 The relationship between education and earnings

2.1 Evolution of the Human Capital Theory

In one of the first works regarding the effects of education in economic growth, Schultz (1960) points out that education can be pure consumption or investment, but regardless of what exactly is, the costs of education are high, both for students[1] and for government [2]. In this way, the author proposes to estimate the amount of expenditure (time spend studying and its opportunity cost and monetary value for the maintenance of teacher's salary, material and other factors) in updated prices considering the period from 1900 to 1956 in order to obtain an estimation of what would be the investment in human capital and to establish a relationship with this value and the unexplained national income growth.

The results indicate that resources directed to elementary education in the United States have grown less than high school and higher education, but still has grown at a higher rate than the gross formation of physical capital. In 1900 the total cost of elementary education was equal to 5% of the gross formation of physical capital, whereas in 1956 it was 9%. The costs of college level education and high school initially rose from only 4% (combined) to about 25% in 1956. Resources for education in the United States in this period increased 3,5 times more than the income for consumption and the formation of physical capital.

These results give evidence to a trend that, as years goes by, both government and individuals tend to invest more in education, which is something that has already been verified by Jorgenson, Ho and Stiorh (2003) and Gloom and Ravikumar (1992) but it also makes the author question if there is a governmental preference for higher education over others, especially when compared to basic education, over the years is seen, most likely, in order to offset the high opportunity cost that it has on individuals, highlighting the importance to invest in universities and consequently promote technological development. Second, the importance of education in economic growth, considering that the physical formation of capital, which is widely seen today as a pivotal factor to economic growth by traditional economic growth theory such as the ones presented by Chow (1993) and Solow (1962), has shown smaller than

[1] It is important to remember that this work was made in the last century using more than 100 years old data, so education was not so accessible as it is today with the advances of the internet, technology and public policies across the word. Also important to keep in mind that, the country studied in this work was the US and both public and private universities have tuition fees. Although public universities tend to be less expensive than private ones. According to the Department of Education in the US data set, the total cost to attend university is the US depends both on the duration and if its public or private. The total tuition cost paid for a 4 year private university in the US in 2012 was $62,585 billions.

[2] According to the Depart of Education in United States the total expenses of 4 year private universities in 2012 was $159,295 billions

the expenditure in education.

One year after its original publication, Schultz (1961) expands the concept of human capital, going beyond just education and it starts to consider other factors as well. He divides them into 5 categories: 1) All expenditures that affect life expectancy, strength, stamina, vigor and vitality of people such as medical facilities and services; 2) Works focused on developing experience and providing internal training as internships; 3) Education of all levels (from primary to the highest specialization); 4) Study programs other than those mentioned (such as extension programs); 5) Migration of individuals and families to suit employment opportunities.

After expanding its concept and commenting on how human capital is improved by key factors that also enhances quality of life (in his work both terms are almost interchangeable), the author concludes that in the US, at the time, there were too many taxes that affected negatively and directly the improvement on human capital, despite of the fact that just like a stock of physical capital can depreciates, human capital (in an individual point of view) can also decrease (since people get sick, older and eventually dies) there should be laws to promote the growth of human capital too.

Schultz also concludes that unemployment affects negatively human capital as, if a person is not working, he/she is not using its skill, their set of abilities will get rusty and less efficient. These skill that were polished over the course of a person's life will only make a difference if they can be put in use. Its third point develops the idea that there are many social frictions in the US (but this can be generalized to any country) such as social, racial and religious prejudice that, unfortunately, keeps people from getting into their career they want and end up in jobs they do not want so much, reducing the human capital effectiveness on economic growth, as they are not working in its most optimal conditions.

In the early 60s Becker (1962), could already see a problem that has become so common in the US, the lack of incentives to education. According to his paper, both government aid through laws and public policy to reduce the cost to education and financial reforms in banks to increase the accessibility to student loans should have been considered a long time ago. The great disparity between rural and urban areas and the lack of internal mobility to allow people to go from a rural area to a urban place, where there are more job options and more access to better education. In a similar view with Oaxaca (1973) and Blinder (1973), Becker (1962) raises awareness on earnings's differences between different groups of people, especially Puerto Rican, Mexican, indigenous migrants and black people. By being paid less, these people would also have worse education and worse health[3], not only reducing human capital growth but increasing its depreciation, due to the lack a minimal working conditions.

Contrary to the works of Schultz (1960) and Schultz (1961), which focus on the positive effects of education in a country, Becker (1962), using a microeconomic approach, studies

[3] Even thought Becker (1962) and Schultz (1961) uses two different approaches to understand human capital, both authors agree that these two aspects are important parts of human capital

the effects of human capital on the individual's income. The author tries to understand how investment in their training would be beneficial to their future real income and how the disparity of investment in human capital between different social classes that have different amount of resources available have perpetuated the inequality of income in the world.

One of Becker's most important finds is that, the majority of the possibilities of investment in human capital affect earnings positively. But first it is important to account that, the returns differ depending on the age of the person. It was seen a positive correlation between age and returns to human capital investment. Older people would have higher returns because of the accumulation of investments in human capital. Younger ones surely also benefits from the investments in human capital, but, since they have, generally lower income, the costs of the investment are comparably higher to them, which makes seen, at first glance, that the investment in human capital would not be beneficial to the younger people.

The theory developed by Becker (1962) states that there is a direct relationship between investment in human capital and returns. Therefore, the more you invest in human capital higher the return, this way, people with more resources would be able to expand their investments in human capital and promote even greater grow to their income. This way people that can invest more would have greater benefits and it would lead to a unbalanced distributions of income, indicating that the inequality problem is a vicious cycle.

Not only monetary resources are important to returns but how effective is the result of such self-investment would be also depends in your innate ability, talent. However, it was seen that there is a positive correlation between ability and investment in human capital. That could be due to people that are interested in activities that enhance their abilities are most likely to be good at them and also because they have a clear edge over other less talented ones, leading them to be more successful and have higher income to invest even more.

Mincer (1958), also using a microeconomic approach and the model of permanent and transitory income components by Friedman (1957), studies the relationship of income, years of study, age, and inequality of American men in 1949. Considering the positive relationship between human capital and productivity increase in the model, there are differences of gains in the same occupations and such differences increase with age.

In an attempt to understand the distributions and observable structures of market remuneration, Mincer (1974) considers the cumulative distribution of human capital investments of workers to build a human capital remuneration function. When calculating the effect of the years of study on the individual's remuneration, the author uses one of Becker (1962) conclusions, taking to consideration that the gains that the person failed to earn during the period he was studying (opportunity cost) is equivalent to a reduction in the total duration of the gains. In this way the gains function in relation to the years of study would be:

$$lnY_s = lnY_0 + r_s \tag{2.1}$$

Where Y_s is the yearly earning of a person with s years of study, Y_0 the base earning with zero years of study and r_s is the discount rate of education. From this equation, the author came to the conclusion that the percentage increase in gains are strictly proportional to the absolute difference in the amount of time in the school. This means that, the log of gains is a linear function of the time that is spent studying. But, as previously noted, human capital is not only education at its pure form (sitting in a chair at school, going to classes) but a set of skills, so the author extended the initial model (1) to consider the experience (accumulation and enhancement of such skills) of the individual in the equation:

$$lnY_s = lnY_0 + r_s + \beta_1 X + \beta_2 X^2 \tag{2.2}$$

Equation (2) is almost the same to equation (1) with the addition of two new terms. X is the years of experience of the individual and β_1 is the angular coefficient of the parameter (in econometric terms), which is equivalently to the rate of return of years of experience. Something that not only Becker (1962) had seen previously with a different methodology approach, but also Mincer (1974) with a more traditional method for today's economists, observed that the investments made in the past does not fade away completely with time, but continue to add to value to a persons total human capital, even thought it also experiences depreciation, as Schultz (1961) states. This way, the greater the experience, the more accumulated knowledge and better developed the person's abilities is, this way, it is natural that the return of the experience is expected to increase, assuming a quadratic form. That's why the second angular coefficient β_2 is accompanied by X^2, to try to capture the effect of continue positive effect of experience over the course of the years. Griliches (1977), starting from the idea of that there could be other factors that also affects the returns of an individual, expands the original return model from Mincer (1974) and formalized it through a standard [4] econometric approach, generalizing the Mincer returns equation, creating the equation called income-generating function:

$$lnY_i = \alpha + \beta S_i + \Gamma X_i + u_i \tag{2.3}$$

Where y_i is a measure of income, S_i corresponds to the educational level, whether it is years of study or full grade (middle school, high school, university degree, master degree and so on...), X is the set of variables that affect income[5], depending on each author (since it is a general equation).Γ is the set of angular coefficients (magnitude of the effect) of each variable included in X and u_i is the error term, also known as residue[6], independently distributed[7] in relation to X and S. This model has become the basis of many linear regression estimation models to study the relationship between returns to education.

[4] Traditional linear regression models, like the Ordinary Least Squares, includes two important factors initially left out by the first works on this matter, which are a constant term and an error term.

[5] The early works of Becker (1962) and Mincer (1974), despite having an already expanded view of human capital available developed by Schultz (1961), he was not able to incorporate factors like sex and skin color, which later on with Oaxaca (1973) and Blinder (1973) made specifically papers on it.

[6] in a classical linear regression model the error residue or term is all information that affects the dependent variable y that was not included in the model

[7] This is one of the main assumptions from the linear regression model

Leal and Werlang (1989) studied the returns in education in Brazil for the period from 1976 to 1986 using PNAD data, considering as income the number of minimum wages for an 8-hour journey for people aged from 25 to 50 years old. Based on the Mincer equation, previously explained in this section, a model that considers the income of the individual according to the years of study was developed. This way there is a possibility that the person starts working immediately and receives Y with $s = 0$ or studies one year and receives $Ys + 1$. Additionally, it was considered the costs to continue studying, like an opportunity cost, that was standard due to Schultz (1961), Becker (1962) and Mincer (1974) earlier works . In this way, the years of study would be considered as a rate of return on investment to education, as shown in the following equation:

$$r_{s+1} = lnY_{s+1} - lnY_s \tag{2.4}$$

Where $r_(s + 1)$ is the rate of return for one additional year of study, $Y_(s + 1)$ is the income resulting from an additional year of study, with Y_s as the base income. In accordance with the equation above, the authors adopt a more financial approach considering education purely as an investment that can produce return to a certain cost, considering a more limited concept of human capital, which is more appropriate to the work that seeks to verify the returns to education purely , since other factors pointed out by Schultz (1961) are characterized in other categories as experience and well being factors that are difficult to measure[8].

The results show that the returns (average increase in the income per year of study) grew for every educational category. The effects of having basic education (from 1 to 4 years of study) on wages grew from one period to another, when compared to those with less than a full year of study (considered illiterates), the returns for year of study were 14,24% for the period 1976 to 1981 and 15,73% for 1982 to 1986. For higher categories of study such as higher education (more than 11 years of study) against high school level individuals (from 9 to 11 years of study), the additional years of study presented a rate of return of 10,16% and 14,82% in the first and second cuts. The highest increase and highest absolute value was from the high school category in comparison to middle school (from 5 to 8 years of study) with 11,87% and 16,36% of annual marginal return for each time interval, respectively.

Menezes Filho (2001) highlights the importance of considering the impacts of the evolution of education in Brazil on the labor market. Analyzing PNAD data, the author reports that there was a significant improvement in the educational training of people in Brazil. Comparing the relative numbers of people in different educational levels for the generations born in 1921 and 1971, the number of illiterates in the country fell from 44% to 8% and that of people completing high school grew from approximately from 3% to 20%. For cuts of 1977 and 1997 the constant educational evolution in Brazil becomes evident. The relative number of people with neither a full year of study dropped from 25% to 12% and people with 15 years of schooling (university degree level) reached approximately 5%. In a more recent data set, also

[8] Factors like migration mobility, in-work training, health, diet, general well being of the individual were even more difficult to include in a model at the time.

from PNAD, in 2015 there was 97% of young people between 6 and 17 years of age (school age in Brazil) are studying and more than 90% only study (without being bound to any other type of work, either formal or informal).

Regarding the returns to education the authors found that the wage differentials for years of study declined in all the study years. The income of individuals with high school level in 1977 was 7 times higher than that of an illiterate, whereas in 1997 it was 4 times. For university degree level, the reduction was from 13 to 9 times and in the postgraduate category, the greatest overall difference between wages was observed, from 17 to 12 times. The author suggests that the increase in education is leading the decrease in the returns for additional years of study.

There are two main hypotheses as for why reductions in returns to education occur in response to the expansion of education. Andrade and Menezes Filho (2005) emphasize the importance of market equilibrium, where an increase in the supply of skilled labor would reduce wage bonuses by educational level. Davanzo and Ferro (2014), using PNAD data from 2001 to 2012, studies the hypothesis that the expansion of education is not necessarily accompanied by investments that are necessary to maintain quality of education. This way, the quality of the education available is reduced and the individual's productivity falls. Considering that the productivity of the work affects greatly its wage, this decrease in productivity would affect the person's income, which is measured by using the Oaxaca-Blinder decomposition.

Using the PNAD data from 1981 to 1999, Andrade and Menezes Filho (2005) calculated the supply and demand of skilled, intermediate and unqualified labor in Brazil to verify if the trend of reduction of returns to education in the recent period is due to the increase in the supply of skilled labor, which would reduce the market imbalance characterized by a repressed demand of university-educated workers. It is important to highlight the possibility that the expansion of the supply of skilled labor will grow at a level lower than the increase in demand for such workers, in a scenario where educational expansion would be accompanied by economic growth and the installation of new labor-demanding firms.

The study found that the number of people with low qualification (up to 4 years of study) presented a decrease over the considered period, while the average qualification (from 5 to 11 years of study) and higher qualification (more than 11 years of study) both grew. Labor demand presented similar results to supply. A decrease in the demand for unskilled workers was seen and an increase for both high and medium skilled workers, which, according to the authors, is attributed to the commercial opening of the 1990s.

Davanzo and Ferro (2014), separated individuals into four categories urban and rural men, urban and rural women and analyses the effects of education, experience, formal work, public sector, metropolitan region and Brazilian state regions. The results show that one additional year of study for urban men and women represented an increase of approximately 11% in income in 2001, while in 2012, 9%. For rural men and women, the behavior was the

same, from 9,1% and 7,7% to 5,9% and 5,6%, respectively.

Experience was not significant for all groups and in the ones that were significant, presented an effect of less than 5% for an additional year. Performing formal work and residing in the metropolitan region were significant for all groups analyzed, increasing income by up to 68,9% and 31,4%, respectively, in the case of rural women. Working in the public sector affected income significantly in the years 2001 and 2012 for women's groups, while for all groups residing in the North and Northeast had a negative effect on income compared to the Southeast.

Using the Oaxaca-Blinder decomposition method, it was possible to separate the evolution of the real wages of the four groups of individuals into explained and unexplained factors. Comparing the years 2001 to 2012, productivity explained 50,18% of the wage differential of rural women, 41% of urban women, 58,19% of rural men and 34,6% of urban men, so that, it presented a determinant role in the income behavior of all the groups considered. However, the author stated that, through this method, it was not possible to verify the reduction in the quality of training of individuals that would affect productivity, since the real wage increased in all categories.

2.2 Schooling and returns using pseudo panel: A cohort stratified analysis

One of the biggest problems of estimating panel data models, especially in Brazil, is the lack of data. The Continuous PNAD, a rich survey designed for panel data started in 2012, therefore, is a very recent data set. But if someone intends to estimate a panel model using the Continuous PNAD, it must consider that, as for now, it does not have an official identifier for each individual, just for the family itself. Another issue to take into account, is the attrition of the panel itself, shown by Gonçalves and Menezes Filho (2015) in his work with the Continuous PNAD. The paper shows that it was not possible to see the same consecutive observations (the same individuals) for one full rotation of the panel.

In the Continuous PNAD, a household is interviewed once every two months, up to a year. After this period there is a new panel rotation and new houses are visited in order to collect data from different individuals and also to not make the people that are been interview annoyed or tired from so many visits in such a short period of time. Another problem that is very common with this kind of data is that, when the interviewer visits the same household in the second turn of interviews it may not be the same person that was interviewed in the first one. If this happens already in the second turn, even if in the third visit is the first person interviewed answering the questions again, the sequence of the panel is broken for this individual and there is going to be a time gap in between visits, consequently, the panel will not be balanced. [9]

[9] If the analysis is considering the same time period as the Continuous PNAD, the panel will have

An alternative to the Continuous PNAD for panel estimation is the PME also from the IBGE. The PME has been conducted by IBGE since 1980. It has a much smaller sample than the continuous PNAD in both number of people interviewed and regions considered. The research is done in six metropolitan areas of Brazil: Belo Horizonte, Porto Alegre, Recife, Rio de Janeiro, Salvador and São Paulo. The questionnaire contemplates questions mainly about employment, age and general characteristics of the individuals. Alongside with the Continuous PNAD, is the only IBGE survey that interviews the same household continually, making it possible (at least theoretically) to use to estimate panel data models.

The PME sample of March 2012 covered 33.809 households with 95.122 individuals. The panel rotation system for the PME works like this: Each household is interviewed eight times over 16 months (interviewed four consecutive months, absent for eight months and re-interviewed for another four months). This way it is possible to have both consecutively monthly data from two periods, while also allowing the household to have a time to "rest" and not get annoyed by the consecutive visits which may lead the household to stop participating willingly. However, the PME suffers from a similar problem from the Continuous PNAD, which is that it does not have an official individual identifier. This leads to leave up to the researcher itself to try to develop one on their own. However, since this data set is much older than the Continuous PNAD, it does exist unofficial individual identifiers developed. Ribas e Soares (2008) is one of the most notable identifiers available.

Even thought true panel data in Brazil is scarce, the PNAD (not the Continuous one) is one of the biggest surveys data in the country and it has a wide variety of information about the Brazilian population. Unfortunately, it falls into the same problem as most of the big surveys, which is that it is a cross sectional data set, therefore, the data does not follow the same individual overtime, which means it is not possible to estimate true panel. Fortunately, there is a way to mimic the traditional panel data analysis using repeated cross sectional data, through the pseudo panel approach.

Panel data for social studies is extremely scarce for almost every country, only a few of them have wide availability of true panel data. Most of them generally have only repeated cross section data sets that have important information to study social phenomena. It is much more common to see panel data in medical studies, as it is key to observe the same patient over time to see how he reacts to a certain medicine. Another rich field is the one of macroeconomic series as it is easier to keep track of something that is so important and is most likely will ceased to exist, a country. However, it is highly costly to create and maintain a large data set of the same persons, especially if the periodicity is high, that is why panel data surveys are not often seen.

gaps, but it is possible to consider the analysis for a different period. For example, if the analysis considers a 6 month period it would still be possible to develop an analysis with a selected sample as the individuals are interviewed every three months. But it is important to keep in mind that every other observations should behave the same way.

The solution that Deaton (1985) considers is that trough the use of cohorts constructed from these random individuals from repeated cross section data it would be possible to take the mean of this sample and use it like an individual from true panel data. According to the author, the individuals relationships seen in panel can be seen in a different way by using cohorts. By taking the average of these groups, the model would be able to take into account the average fixed effects for these cohorts as a parallel alternative to individual effects from panel.

By considering the use of pseudo panel as an alternative to true panel data, one may think that the data set constructed from the cohorts is inferior to panel data but that is not true. Panel data faces many problems, like, missing information, poor identification, attrition, continuous breaks, all of these problems are present in the panel data options for Brazil (PME and continuous PNAD). These problems do not happen with pseudo panels, since it is using the available data (cross section) stratified by groups. So, as long as it is possible to classify the individuals into any of these groups it will be possible to estimate the model.

Deaton (1985) Points out that one can see the use of cohorts as an possible solution to the previously mentioned problems with panel data, like an instrument. When trying to estimate an OLS model using regressors that are not correlated with the error, but are meaningful to the dependable variable, instrumental variable approach solves this issue. The same thing using a group instead of an individual to capture the variations that are related to the fixed effects. As cohorts take into themselves information every time a new survey comes out, it can contain a large amount of observations allowing the researcher to be more judicious with the creation of a cohort, increasing the number of characteristics used leading to higher stratification. If there are very big differences in cohort sizes it, especially if the sample is not large enough, that may lead to heterokedastic errors. In this case, each observation in the cohort should be corrected by some sort of sample weights, Dargay (2007) suggests to use the square root of the cohort size, the same idea as an Generalized Least Squares.

Browing, Deaton and Irish (1985) uses the Family Expenditure Survey to create cohorts, to study labor supply and commodity demands basing on the life cycle theory, as it was a traditional random sample cross sectional data. Since it was not possible to follow the same household over time, the authors considered using groups to distinguish the households and in this way they could follow their behavior over time, just like a panel data would with an specific individual. First the cohorts were constructed using year of birth because the authors realized that tracking a group of individuals of 30 years old in 1975 and a 31 year old in 1976, and so on, they would be tracking the same cohort. After defining the groups, the mean from the individuals within the group are taken, because that would represent the average behavior of those people that are inside that cohort.

The more specified the cohort, the better at representing the behavior of such individuals, using a more detailed approach, as its taken into account more determining factors mimicking the individual treatment that panel data has but without the attrition's problem. However,

it is important to keep in mind that the higher the number of cohorts, lower the number of observations in each cohort there will be. This way, if a limited data set that does not have a big sample is being used it may lead to biased estimators. The mean number of observations used to construct one cohort must be at least 100, according to Verbeek (2008).

It is important to note that the model has to be linear in the parameters, otherwise the mean of the cohort will not be able to reproduce the average behavior of the individual. However, there is no problem with non-linear data. So first the individuals are selected, then the data is obtained in its "final form"[10]. Then, one proceeds to take the means and store in the cohorts. Since the authors were working with a limited data set that did not have many observations to create the cohorts, they divided the sample into 1 year interval, as was initially toughed. However, by stratifying in this way, the division would lead to groups with too few observations. [11].

To overcome the problem of too many cohorts for a small sample, the authors divided the individuals into 8 different age intervals[12] and into manual and non manual workers. The year of birth division is the most common one in studies that use cohorts, but, it is especially important in economics because different generations go through different economic stages, both in their own country and in the world, experiencing drastic changes in their lives that goes beyond economic measurements, but that are very important to the formation of human capital.[13]. The second characteristic is also key to capture these individual effects, because the type of labor that a person did[14] was very important to determine his wage and standard of living according to Haskel (1996) and Theodossiou (1990).

It would be ideal if the cohorts that were created had the same number of individuals every year, or at least, grew in the same proportion of the population growth, but, unfortunately it is not always like this. There is migration and immigration factors that could drastically change a cohort size over the years. Consider a specific cohort that has most of its member in a location that suffers from natural disasters. It is expected that there would be a higher migration rate (exodus) from that region than others. With age stratification, it is most likely that a cohort that had already older individuals will suffer more losses from death than the younger groups. One should be specifically careful when selecting the characteristics to create the cohorts as if a very specific group is created, the number of observations used to create that cohort would be much smaller than the other ones and this must be taken into account at some point.

[10] Linear, log, squared. The data itself has no restrictions, just the parameters.

[11] It was respectively: 2557 observations in 1970 to 1971, 2865 in 1971 to 1972, 3050 in 1972 to 1973, 2992 in 1973 to 1974, 3046 in 1971 to 1972, 3156 in 1975 to 1976 and 3020 in 1976 to 1977

[12] Respectively:18-23, 24-28, 29-33, 34-38, 39-43, 44-48, 49-53, 54-58

[13] As Schultz (1961) shows that technological improvements can greatly improve human capital, by improving health, urban and rural mobility and many others.

[14] Nowadays it is not as common to see analysis using a cohort like Manual and Non manual work, but one may consider factors like formal work and college degree present similarities with the variable used by the authors as both have significant impact in their wages.

The method developed by Deaton (1985) is the standard approach to construct a pseudo panel, which consists in first, creating cohorts, taking its mean and performing the estimations. Very common in medical research, cohorts are a way to separate individuals into groups that share the same characteristics. They are usually divided by year of birth and gender, but there is relative flexibility to combine different characteristics in order to create a more complex analysis, the few rules of creating a cohorts will be discussed later in this section.

There are different approaches when dealing with pseudo panel, no just linear models, but also nonlinear and dynamic as shown by Moffitt (1993), Collado (1997) and Girma (2000), just like the traditional panel models. As Verbeek (2008) argues, diving the sample into cohorts is the same as using instrumental variables, estimation wise. Therefore, it should fulfill the same conditions as the latter.

Menezes Filho, Mendes and Almeida (2004) uses PNAD data from 1981 to 2001 (with breaks for the years of 1991 and 1994) to investigate the differences in wages for formal and informal work using the pseudo panel approach to account for individual heterogeneity. The sample used was composed of male from 24 to 57 years of age with jobs. The cohorts were constructed using age intervals and level of education: People born from 1924 to 1933, 1934 to 1943, 1944 to 1953, 1954 to 1963 and 1964 to 1973, with 7 years of study or less and 8 years of study or more.

The results are divided into 3: First, for the full sample. Second for the decade of 1980. Third for the decade of 1990. Four models were estimated, first using only the formal work dummy, second introducing the educational dummy, third the cohort dummies and lastly using the residuals for corrections.

For the full sample, the effects of having a formal work on the natural log of hourly wages were 0,66, 0,47, 0,47 and -2,76 for the four models. The educational dummy (which takes value of 1 for people with higher education) was 1,04, 1,07 and 1,56, respectively. The second sample reported lower effects of formal work on the wages with coefficients of 0,63, 0,457, 0,458 and -3,3 for the four models. The returns of better educational level were higher than the overall sample with scoring 1,14, 1,18 and 1,64 total. The last sample had slightly lower effects for the group with more years of study, 0,87, 0,89 and 1,3 for the last three models. The effects of formal work were the highest overall at a rate of 0,69, 0,51 0,50 and -1,6. The model corrected by the residuals showed that individuals with formal work had much smaller wages than those with informal work, while all the other models showed the opposite.

2.3 Sample Selection Bias in labor force and its effects on the returns

Heckman (1979) proposes a different approach to the bias problems that arise when using a sample with missing data due to specification error. One of the conditions to have efficient estimators is that the sample must be random, however each different observation (or group of observations) do not have the same "treatment" as others. In the sense that, when performing a wage analysis for male and female, for example, there may exist factors that affect the two groups differently, leading to the observations that have positive values of wages (that are part of the labor market) to have similar characteristics, leading to the exclusion of individuals with traits that affect negatively the insertion in the labor market. This leads to a sample of individuals with similar characteristics and therefore, non random. This way, when estimating a linear regression, the estimators will be biased because the sample that is being used is not random.

Sample selection models models are used when a dependable variable from the regression is only seen in one part of the data, which means that the sample is being censured somehow. Considering that the goal of this work is to estimate the returns to education in Brazil, it is expected that part of the sample used does not have any income because they do not have jobs. The data set used here (PNAD from 2011 to 2015) shows that about 50% of the total sample were employed, considering individuals that are not in school age (at least 18 years old) and out of the retirement range (65 years old or older). Since part of the sample does not work and consequently does not have any income, the OLS estimators will not be the most efficient ones because there is a part of the sample that is being censored and will not be representative, affecting negatively the statistical inference.

In order to have non biased and efficient estimators, first, it is needed to take into consideration the factors that affect the insertion of a person in the labor market. Curi and Menezes Filho and Scorzafave and Menezes Filho (2001) shows that the most important determinants of participation in the labor market in Brazil are age, years of study, being a partner living with the other. Both papers agree that men and women react differently regarding the insertion in the work force, the former uses sex and year of birth cohorts to stratify the analysis while the latter uses a sex dummy. In this work the estimations will be separated by sex and the determinants will also include marital status. Unfortunately, in this data set, there is not a variable that determinate if men are parents or not, so two proxies will be used in order to try to capture this effect. The first one is if the person is a male of 18 years or older and the type of family that he has is a couple with children under 18 years of age, the second one is if his family has 3 or more members.

Since the determinants of the participation in labor market and the returns from education are not the same, each problem should be offer a different solution. So, there will be

two different models, one for each purpose. The two step Heckman selection model and the Maximun Likelihood Heckman selection model are both capable to take into account these issues and return efficient non biased estimators. Despite having many similarities on how to approach the problem both of them are perform the estimation differently.

The two step Heckman selection model first estimates a probability model[15], in this case it will be the probit model, to captures the effects of the factors that are censoring the sample and them creates a new variable called mills ratio that enters the second regression as an independent variable, leading to efficient OLS estimators. The Maximun likelihood model estimates both models together, without the necessity to estimate first a probabilistic model and then include the reverse mills ratio into the regression as the maximun likelihood model already takes into account the bivariate distribution density according to the probability of the event happening (the person being in the labor market).

Heckman (1974) argues that the deciding factors for a certain person to enter the labor market is different for different groups of people. Married women have a unique set of factors that not only decides if she is getting a job (or is even looking for one) but that also affects her wage, total number of hours worked and level of education. Because there is such difference between this specific group and others, it is necessary to take into account the probability of married women entering the market and estimate the functions that determinate the number of hours she works, wage and the expected wage.

Using data from the National Longitudinal Survey for women from 30 to 44 years old, the author considered the number of children, net worth of the household (money that does not come from work), years of experience, years of education and wage of the husband to determine the wage that the women expects. The results show that the husband's wage and years of study increased the expected wage by 15% and 5% (for every additional dollar in his wage). The number of years of education and experience affected positively the number of hours worked per week (5%) and per month (4%) for the former and 4,5% for both frequencies the latter.

Brand and Xie (2010) using the National Longitudinal Survey of Youth of 1979 and Wisconsin Longitudinal Study from 1957 studies the effects of college degree on men and women in different age groups to see if there is significant difference. The hypothesis that is being tested is that people that have characteristics that positively affect the decision (possibility) to get into college are the one that get the most return out of it. In a way to deal with the problem of heterogeneity and sample bias the methods consisted in estimating separated regressions for different groups according to the probability of that group having a college degree.

The results finds that having a college degree positively affects the natural log of hourly

[15] In theory it could be either a logit or probit but the majority of the research using sample selection model chooses probit for its more simplistic method

wages in 18% for the group of ages from 29 to 32 years old, 30% for the group from 33 to 36 years old and 41% for the group of 37 to 40 years old. For women, the returns were, 27% for the youngest group, 18% for the group from 33 to 36 years old and 21% for the oldest group.

Grogger (2009) estimates the effects of experience in wages using data from reservation wage data in order to create a probabilistic model to capture the labor force participation, the person chooses to work if the wage that the person is offered is higher than the wage she expects. The author argues that the traditional OLS model underestimate the returns to education, especially for older ages. Using data from Florida's Family Transition Program between May of 1994 to February of 1995, the author estimates two equations, one for wage and another for the reservation wage. If the real wage exceeds the reservation wage, the individual will choose to be part of the labor force.

The results show that, on one hand, age had no effect on either the actual wage or the expected wage, skin color and children on the other hand significantly affect both estimations. Non-white and having children reduced both the wages and the reservation wages. The returns for experience are smaller for both traditional OLS and Heckman's two-step model, in comparison the the Maximun Likelihood sample selection model.

Silva, Carvalho and Neri (2006) uses data from 2003 PNAD to estimates the mincer equation correcting the selection bias using the sample selection model from Heckman (1979) and them decomposes the differences of wages using the oaxaca-blinder decomposition. Two estimations for wages were carried on using as explanatory variables years of study, years of study squared, experience, experience squared and urban area. The second equation considered the same variables but also added two dummies, one for sex (males taking the value of 1) and one for skin color (white individuals taking the value of 1).

First it was estimated two traditional OLS models. The results show that, both years of study and experience had similiar effects, 5% each, on log linear hourly wages for models 1 and 2. Squared experience had a very small(0,005%) negative value, while urban residents had an income 20% higher. The effects of sex and skin color were the highest in the models, account for 30% and 26% increment in the wages.

After the first model, the Heckman Selection model was estimated to account for selection bias. To calculate the probability of a person having a job or not was used a probit model, that was constructed using years of study, squared years of study, experience, squared experience, if the person is the family householder and if there were any children from 0 to 5 years of age. The probit results shows that the higher the experience, more likely is the person to have a job, while the negative coefficient for the squared experience indicates that the positive effect of experience decreases with time. It was also seen that, the more educated the person (the higher the number of years of education) higher the chances to participate in the labor market. Another important factor was if the person was a householder or son/daughter of the householder, the former had positive effects in the market participation while the latter,

negative.

After the estimation of probit, the inverse mills ratio was calculated and the linear regression was perfomed again, this time corrected by the lambda. The results show that the returns for years of study for men and women were at the 2% rate, while for white women was negative. The effects of experience was 0,003% for every group regardless of the skin color or sex and the urban dummy effect was also reduced to 0,002%. The results reported using the two step model were much lower than the OLS bechmark.

Using data from PNAD 2004, Sampaio (2007) studied the returns to education in Brazil and in Paraná (a Brazilian state) using the OLS method, instrumental variables to account for individual heterogeneity and two step heckman to eliminate the selection bias. For the three different approaches, the author estimated two different regressions. One for the wages and years of study and another one adding the experience and squared experience. The first and second regressions using the OLS method scored the exact same result, an effect of 8,9% of returns per additional year of study for Brazil and 7,6% for Parana. For the instrumental variables (it was used a two stage OLS) the results were significantly higher. The first regression reported 12,3% of returns for Brazil and 9,4% for Parana. The second estimation had even higher scores reaching 13,8% and 14,5% respectively. For the sample selection model the returns were the lowest, at 5% for both regressions for Brazil and 2,2% and 2,9% for Parana, the difference between the instrumental variable estimator and the selection bias model was of 9 percentage points for Brazil and 12 percentage points difference for Parana.

Pereira, Braga and Mendonça (2013) uses data from PNAD 2009 to estimate the returns to education in brazil for males from urban and rural regions accounting for the selection bias problem. The authors used the traditional returns equation from Mincer (1974) as base and added dummies for skin color, region, if the person that answer the questionnaire was the household or the son/daughter of the household, if the person was part of a syndicate and used age as proxy for experience. The probabilistic model estimated to determine the market participation of the person shows that both age and years of study affect positively the matter. The condition in the family was also significant, if the person was the son/daughter of the householder the chances of that person being in the labor force were smaller. Skin color did not have significant effect as all 4 different categories included (Black, yellow, brown and indigenous) were not statistically significant at an α of 1%. Out of the four regions considered in the analysis, only the south region was significant at 1%, having a positive effect in labor market participation.

The coefficients corrected by the inverse mills ratio from the probabilistic model shown an overall rate of return per year of study of 9,7%, while the age had the highest effect among the time-variable coefficients, 14,4%. Out of the 4 skin color categories only one was statistically significant at 1%. Having brown skin color had a negative effect on earnings of 13,5% compared to white. Being part of a syndicate led to a wage increment of 14% and the

two regions that were statistically significant were South and Northeast, both had negative effects compared to Southeast.

Coelho, Veszteg and Soares (2010) studies the returns to education for white and non white women in brazil using a quantilic regression accounting for selection bias. Two models are proposed to correct the parameters. The first model is the traditional two step selection model of Heckman (1979) were it is first estimated a probit model to determine the probability of a certain sample (sub sample in this case, for skin color groups for women only) to obtain the inverse mills ratio that will be used in the returns regression to corrected the parameters. The second model is a qualitic regression from Buchinsky (1998), a technique to divide the sample into different quantils of the conditional distribution. This way, the absolute value of the residuals are weighted according to the distribution divided into quantils, in order to have a more symmetrical distribution.

The data used is from PNAD 2007 for females only, ranging from 20 to 60 years old, in a total of 107.634 observations. Approximately 61% of the sample had paid jobs, which means that 39% of the sample was being censored. This is an evidence of potential significant selection bias. Both models show the same signs for all variables and with almost identical coefficients for the labor market equation. The results show that, the number of children that the family has, presence of other sources of income besides work, total income from the family and being white have negative effects in the insertion in the labor market. Years of education, being the household and age, all had positive effects.

The results of the returns regression, for the full sample, showed that the effects of education on log linear hourly wages, for the 0,1 quantil, was 10,8%. The magnitude of the returns diminished for every 0,1 quantil, until quantil 0,5. After that, the returns started increasing for every sample expansion, reaching 13,7% for the 0,9 quantil. Using the selection bias correction, the returns were smaller compared to the ones without correction until the 0,5 quantil, with a difference of 0,06 percentage points. From the 0,6 to the 0,9 quantil, the coefficients reported 11,5%, 12%, 13% and 14,5% respectively. Using the symmetrical correction from Buchinsky (1998), the results were almost identical, scoring for the last 4 quantils, 11,6%, 12,1%, 13,1% and 14,5%, respectively.

As for white females, the returns were higher than for the full sample. For the model without correction, the results were 13,3%, 15,25%, 16% and 16,1%, for the quantils 0,6 to 0,9. For the same interval the two step model reported 13%, 14,3%, 15,25% and 15,5% while the third model showed 12,7%, 14,3%, 15,1% and 15,4%, respectively.

3 Brazilian Labor Force: an overview of 2011-2015 PNAD dataset

PNAD data from the years of 2011 to 2015 was used to perform different estimations in this work,so it is important to have a overview of the main characteristics of the labor force in these years. This period comprises the first year after the 2010 census (there was no PNAD in 2010) and 2015 the last available year of the database in question. For the pseudo panel, the analysis took into account 24 different groups, namely the cohorts, which are basically formed from two fundamental characteristics: Age and sex. In the PNAD, interviewees respond if they were occupied (had any job, regardless if it was paid or not) in the reference week of the research. For the pseudo panel approach, only those who answered affirmatively to this question were considered. In this way it is possible to also include those who are not formal workers with a formal contract, which is important for a more complex analysis. The sample selection bias models used a slightly different sample, considering also those who were not working in order to correct the coefficients for the sample censoring.

The labor market in Brazil, in the years 2011 to 2015 was dominated by men. In all years, males were the majority in the number of workers, however, there was a small reduction, consecutive from year to year. In the first year of the analysis, men were 57,8% of the total workforce in Brazil, while in 2015 this number decreased to 56,2%, a reduction of 1,6 percentage points. It is important to note that the increase in the number of women in the labor market occurred mainly in the last two years of the sample, with the increase in the from 2011 to and 2012 to 2013 being practically negligible, less than 0,03 percentage points, combined.

The average age of the Brazilian worker has been increasing in recent years. In 2011 the total average was 37,7 years old, being 37,8 for men and 37,5 for women, while in the year 2015 the total average was 38,4. The age difference between men and women however remained the same, 0,3 years. Average experience, naturally, followed the same trend as the age increase in the national workforce, growing 1 year over this 5-year period from 22,8 years in 2011 to 23,8 years of experience in 2015. It is interesting to note that the difference between experience for men and women is decreasing over time. In 2013, the year where there was the biggest difference, men had on average 24 years of experience, while women had 2 years less, a difference of 9,1%. In only two years the difference decreased to 6,9%, with women reaching 22,8 years of average experience while men had an average of 24,4 years.

A very interesting variable, included in the PNAD questionnaires, is the age that the individual began to work. In addition to exposing one of the most striking Brazilian characteristics, which is the rapid insertion of children in the labor market, this information is also illuminating to better understand why the significant difference in experience between

men and women, even when the average age of both male and female workers are practically the same. The average age that the employed individuals started working in 2011 was 14,8 years, however, the highlight is the difference between men and women, which was 1.5 years. This difference was maintained throughout the analyzed period, with the exception of the year of 2012, where the difference was 2 years. However, both men and women started to enter the labor market at an older age as years went by. In 2015 the average age was 15,1 years, with men starting to work at 14,5 years and women at 16 years.

Since the main objective of this study is to analyze the returns of additional years of education in people's income, it is necessary to have a better understanding regarding the general panorama of education in Brazil. Fortunately, in recent years, the average years of study for workers has been growing for both men and women. In 2011, the overall average was 9,37 years, with 8,8 years for men and 10,1 years for women, a 14,8% difference in favor of females. Although in 2015 there was a reduction in this difference, the change was not constant over time, in fact, the difference grew reaching 15,4% in 2013 and in the following years it fell to 14,1% and 13,3% respectively. In the year of 2015, the general average was 10 years of study, with 9,47 years of study for men and 10,73 years for women.

In the recent years, not only the average number of years of study grew, but also, the number of people with higher education has been growing as well. In 2011, only 13,9% of the workforce had a college degree, whereas in 2015, 16,2% of the workers were graduates. There was no significant change in the difference between of men and women with higher education in Brazil. Females kept their dominance in universities in every year of the sample. In 2011, 55.4% of the graduates were female, while in 2015 the difference grew 0,4 percentage points.

Three other categories were also considered, those being high school (11 years of study), middle school (8 years of study), primary school (4 years of study) and the illiterates (those that stated that did not know how to read and write). The number of people with high school and primary school did not change during the whole period. As for the illiterates in the labor market, there was a very small reduction to the already low percentage of this group. In 2011, 5% of the sample workers were in this situation, while in 2015, only 4% a total variation of 1 percentage points. Regarding gender differences in this segment, there is an overwhelming male dominance, where 58,5% of the ones with no education were men.

It was also analyzed the evolution of real wages over the course of the 5 years of the sample for four different categories: White Men, Non-White Men, White Women and Non-White Women. It is important to keep in mind that the series was deflated for 2015 prices using the yearly IPCA.

As table 1 above shows, there has been a significant increase in the inflation for the 2015 period, so, it is expected that there will be negative effects in the wages considered in the sample. Besides 2015, there is no other year where the difference is bigger than 1 percentage point. One may think that the spike in the inflation in the last year should not have an impact

Table 1 – Yearly IPCA from 2011 to 2015

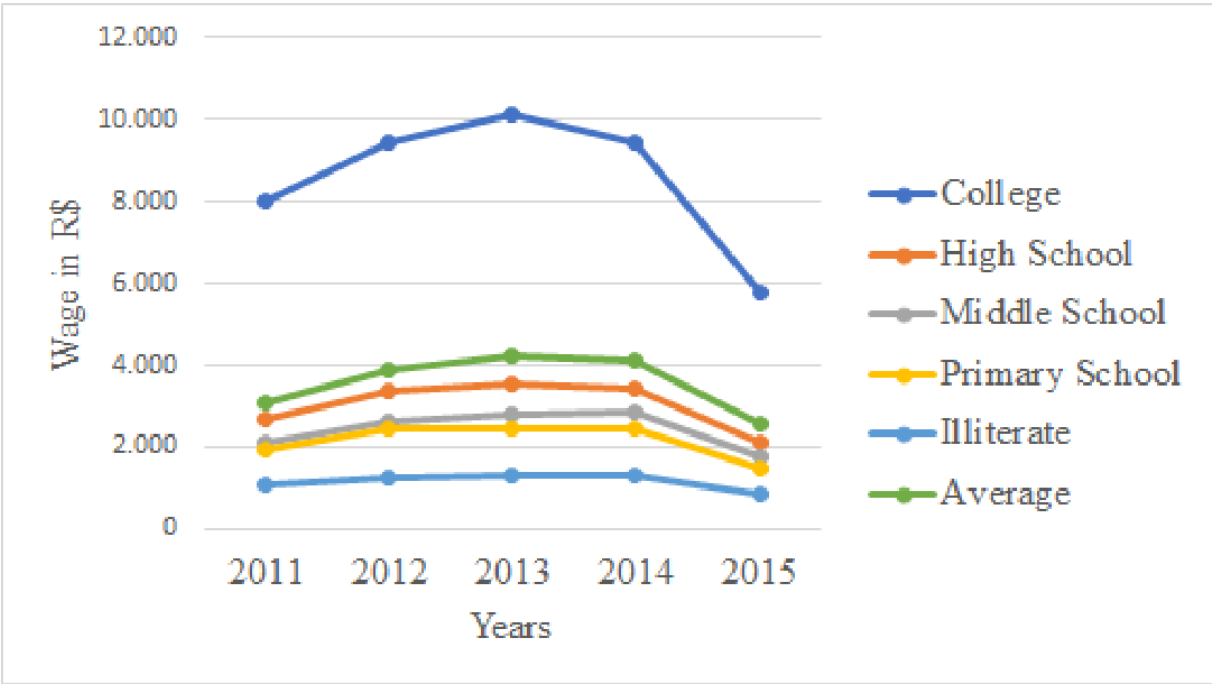
Year	Yearly IPCA (%)
2011	6,50
2012	5,84
2013	5,91
2014	6,41
2015	10,67

Source: Own elaboration using data from Ipeadata.

in the real wage since most wages also increase above the inflation, but it is important to keep in mind that it was considered people working in informal jobs that do not have wage correction and has wages lower than the legal minimun.

It was considered 5 different educational groups and, in green, an "Average" group, which is the overall wage for every individual in the sample, regardless of their educational level. The other four categories are:In dark blue there is the category "College", which is made of individuals with 15 or more years of study. In orange there is the category "High School", which is made of individuals with 11 years of study.In grey there is the category "Middle School", which is made of individuals with 8 years of study. In yellow there is the category "Primary School", which is made of individuals with 4 years of education. Lastly, in clear blue, there is the category "Illiterate", which is made of individuals who responded negatively to the question if they could read and write.

Figure 1 – Real wage and years of study of white men



Source: Own elaboration using PNAD data.

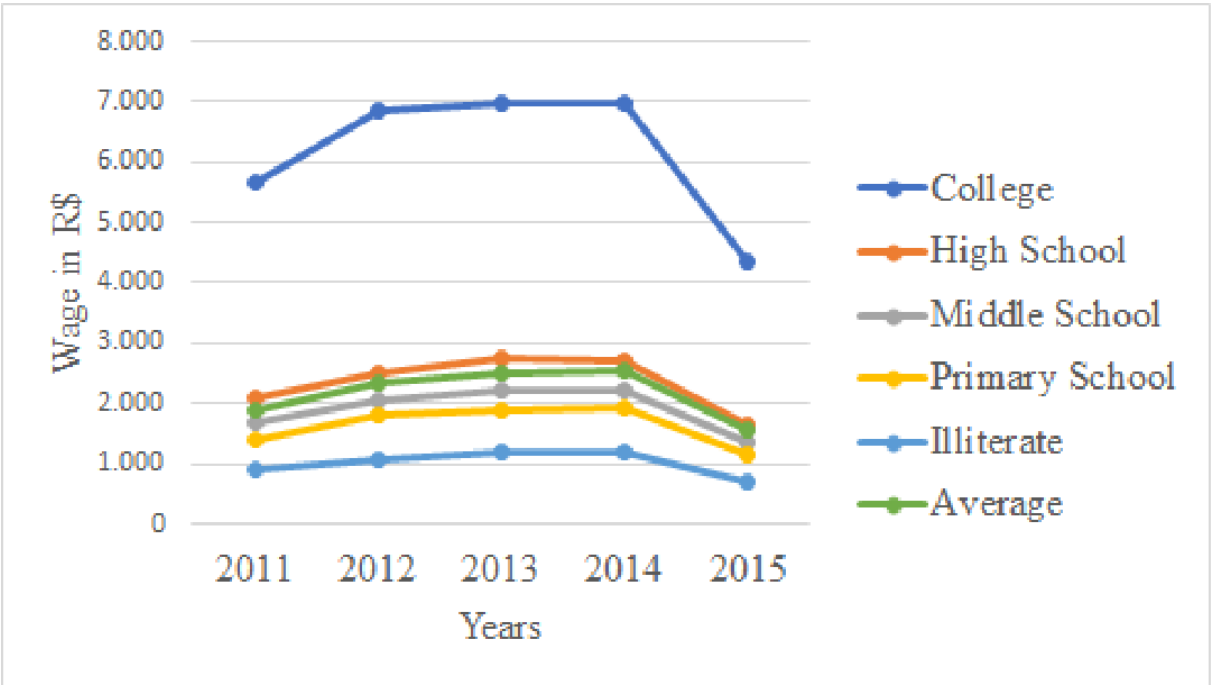
Starting with the group of white males, the figure 1 shows that there is an enormous

difference between the real wages of individuals with college degree and everyone else for every year of the period of 2011 to 2015 in Brazil. All other categories are much closer with each other but, show results that are expected from the literature. The highest earnings are from graduates, followed by, high school graduate, middle school, those who finished primary school and lastly illiterates.

Considering the behavior of all five educational categories, the one that has fallen the most, in the last two years are graduates. In 2011, the average wage for a college degree holder in Brazil was R$7.993,11, the closest to this group was high school graduates with a real wage of R$2.690,97, which is R$5.302,14 less. In comparative terms, a high school graduates' wage corresponded to 33,66% of a graduate one. However, the gap has reduced since then and in 2015 the difference was R$3.695,48 with average wages of R$5.789,40 and R$2.093,92. Obviously, both took a major hit from the high inflation in 2015 but the relative difference has also reduced and in the latest year of the sample, a high school graduate had a wage of 36,16% of the total earning from an individual of with higher education.

The other three categories showed similar pattern, showing slight increase every year until 2015, when there was a significant fall . Middle school graduates were the overall third in terms of earnings, with a total wage of R$2.251,54 in 2011 and R$1.744,57 in 2015. Both primary school and illiterate categories were in a growing trend until 2014 when it started to fall as well. The former went from R$1.934,43 in 2011 to R$1,492,48 in 2015, while the latter started at R$1.058,00 and finished at $873,19.

Figure 2 – Real wage and years of study of non white men



Source: Own elaboration using PNAD data.

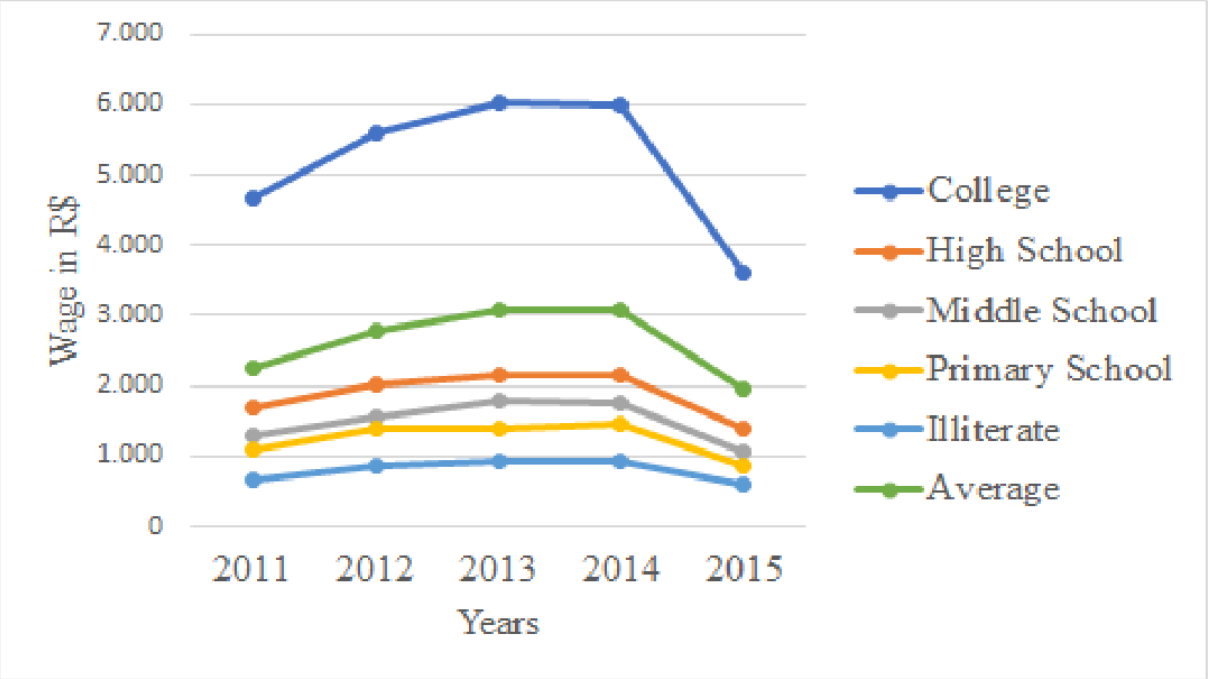
The figure 2 shows the behavior of the real wages for non-white men, which shows the

same trend as for white men. For college graduates, in the first four years, went down from R$5.650,39 to R$4.346,70. Every other category has shown the same pattern, with 2014 being a year where there was significant reduction in wages for this group.

Even thought it shows that the average wage is smaller than the the high school's, which is opposite of what the group of white men had, it still holds the expected relationship when it comes to educational groups income. The group with the second highest income is High School with an average wage of R$2.088,30 in 2011 and R$1.632,18 in 2015. Middle school comes at a rather smaller margin from the High School group at a R$414,91 difference in 2011 as their average income was R$1.673,39 at the time. In 2015, the difference was smaller, at R$274,79, considering its R$1.357,39 average wage in that year.

Both Primary school and Illiterate groups had an average income of R$1.406,27 and R$915,93 ,respectively, in 2011. In 2015, reached its lowest point, finishing at R$1.145,19 for the Basic Education group and R$696,57 for the Illiterates.

Figure 3 – Real wage and years of study of white women



Source: Own elaboration using PNAD data.

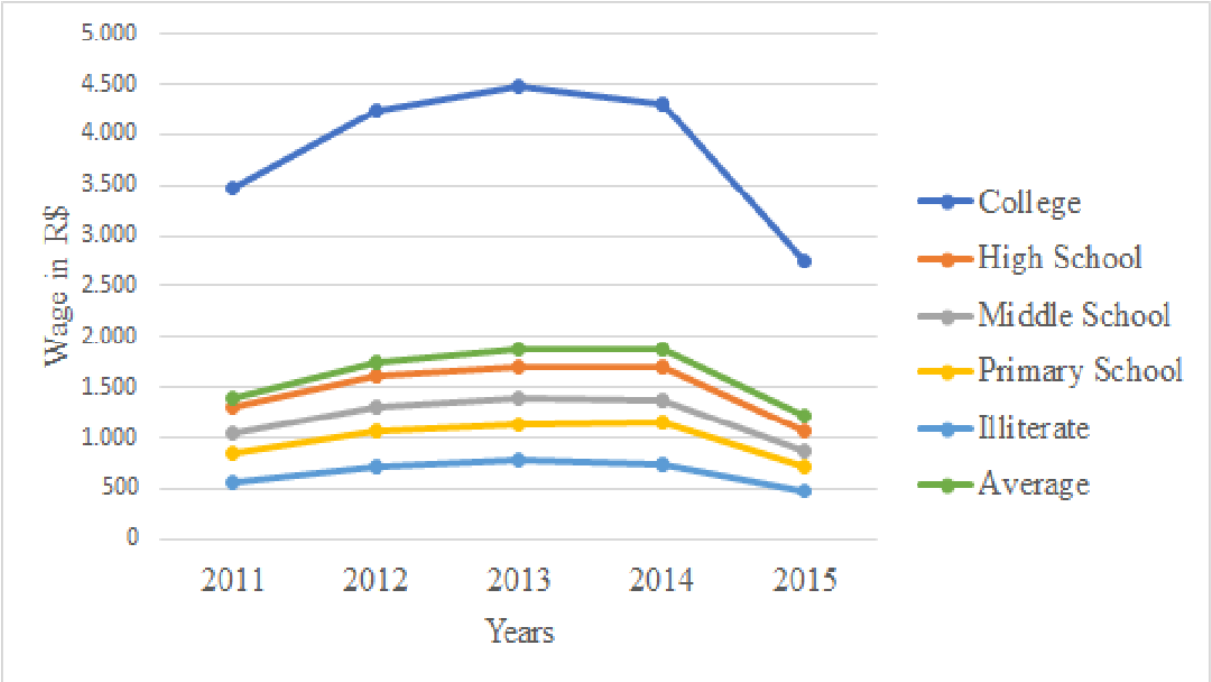
The real wage returns for white women with college degree seen in figure 3 has a very similar pattern as white men with college degree. Both show significant increase in the first years while having a steep downward spike towards 2015. Starting at R$4.676,07 in 2011, the average wages continued to increase up until 2013, where it reached its peak at R$6.039,43 and then proceeded to fall down until it settled at R$3.625,00 in 2015.

High school and Middle school both started and finished very close to each other. In 2011 the mean income for the former was R$1.673,08 while for the latter was R$1.282,44, a R$390,64 difference. The gap got significantly bigger in 2012, when the higher education

group had an average of R$2.011,28, while the lower education one, R$1.572,86. Finally, in 2015, both meet at much lower values as High School graduates had a average income of R$1.376,86 while the Junior High group was at R$1.053,76.

Both Primary School and Illiterate groups scored the lowest wages so far and showed the same pattern from as the other groups. In 2011, the wages were R$1.077,26 and R$675,22, respectively. The ones with the lowest education showed continued increase in their incomes up until 2014 and then proceed to fall reaching R$589,04 as average. The group with 4 years of study had a rising trend until the second last year as well, but it settled at R$876,42.

Figure 4 – Real wage and years of study of non white women



Source: Own elaboration using PNAD data.

Regarding to the real wages for Non-White Women in Brazil shown in figure 4, we see a similar pattern to the two other groups, White Women and White Men as Non-White Men showed a higher wage for the High school group compared the average. There was a continuous increase up until 2013, where it reached its peak, for every group, except for Primary school which had no significant difference between 2013 and 2014. All of the groups had the smallest wage in the series in 2015. The behavior of the wages of the college degree group showed the same pattern to the ones from other groups. Starting at R$3.479,16 and continuously grew reaching R$4.484,07 in 2013 and then finishing at R$2.752,09 in 2015.

The second and third highest wages are from High School and Middle School groups, as expected. In 2011 they had a real income of R$1.298,16 for the first group, and R$1.048,89, for the second group. The R$249,27 difference did get R$65,98 smaller for the last period considered, where the highest education group had an average wage of R$1.059,50 while the lower one was at R$876,21.

Lastly the two lowest groups in terms of years of study not only scored the lowest wages for Non-White Women but also was the lowest overall. In 2011, the 4-year study group had an average income of R$841.84 while the illiterates where at R$567,17. In 2015 the wages got even lower. The second less educated group overall had a R$234,08 lead over the less educated one. The former had an average wage of R$707,29, while the latter, R$473,21.

4 Methodology

4.1 Pseudo Panels

Starting from a simple linear regression equation, we have:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \tag{4.1}$$

Where Y_{it} denotes the explained variable, X_{it} the explanatory variable, β_1 its coefficient, $alpha_i$ an intercept and u_{it} an error term. At first glance, it looks like a traditional panel data regression, but, despite the subscript t still ranging from 1 to the last time frame, i does not range from the first individual to the last for each cross sectional data. That is because the individuals are not the same, so each year (t) will have different numbers for i. If the explanatory variables and the individual effects are not correlated, there is no violation in the presupposes of linear regression and, therefore, could be estimated via a pooled OLS. However, if there is correlation, it should be corrected.

In panel data, the individual effect is an unknown fixed parameters, giving each individual its own intercept. This is only possible for panel data, as the individuals are the same every time frame. One way to try to mimic the advantages for the estimation, of an unique individual, is by creating cohorts. When creating them, is important to consider characteristics that are not time variant. The characteristics selected to separate the sample into cohorts should be observable in every individual, in every time frame, should stay the same (always) and be designed in a way that a person must be part of one and only one cohort. By aggregating everyone into cohorts the linear regression equation can be written as:

$$\bar{Y_{ct}} = \beta_1 \bar{X_{ct}} + \bar{\alpha_{ct}} + \bar{u_{ct}} \tag{4.2}$$

The key to understand this approach lies in the subscripts c and t, which correspond to cohort and time respectively. Looking closely, one may realize that the way the pseudo panel mimics a true panel is by creating cohorts, which takes the place of the individual itself. By taking the mean, denoted by the bar over the variables, of the whole group, observations for each year are created making an artificial panel. Considering that $\bar{\alpha_{ct}}$ is unobserved and very likely to be correlated with X, especially because of the group separation made by the cohorts, treating as part of the error term, making it a compound error like in the random effects model will violate the one of the presupposes of the linear regression. This way, it is correct to treat as a fixed effect just like true panel, since there is no variation over time, therefore, β_1 is a within estimator.

4.2 Sample Selection models

Consider that, in the PNAD data set that will be used for the analysis has N observations[16], however, only n have positive wages due to unemployment, as $N>n$. So, the dependent variable in the returns regression (real wages) will only exist if the dependent variable in the probability regression is equal to one (has a job), this way the probability model of a person being in the work force will be the following:

$$Z_i = \gamma_1 + \gamma_2 W_i + u_i \tag{4.3}$$

Where Z_i denotes the participation of individual i in the work force, γ_1 is the constant term, γ_2 is the magnitude of the effect of the factor W_i in Z_i and u_i is the error term. The range of i goes from 1 to N (total sample size). Like it was said before, men and women will have different explanatory variables due to the lack of data. The men's and women's workforce participation equations will be, respectively:

$$\begin{aligned} Z_{im} = \gamma_1 + Years_{im} + Age_{im} + Partner_{im} + Children_{im} \\ + Metropolitan_{im} + Region_{im} + White_{im} + Urban_{im} + u_{im} \end{aligned} \tag{4.4}$$

$$\begin{aligned} Z_{iw} = \gamma_1 + Years_{iw} + Age_{iw} + Partner_{iw} + ChildrenUnder14_{iw} + ChildrenOver14_{iw} \\ + Metropolitan_{iw} + Region_{iw} + Urban_{iw} + White_{iw} + u_{im} \end{aligned} \tag{4.5}$$

Where Z_{im} denotes the participation of man i and Z_{iw} woman i in the work force, γ_1 is the constant term, Age is the age of the person, $Years$ years of study, $Partner$ if lives with the partner, $Children$ if the person has any children, $Urban$ if the person lives in an urban area, $White$ if the person has white skin, $Region$ taking value 1 depending on the region that the person lives (North, Northeast, South and Midwest) and u_i is the residual. After estimating each probability model, the coefficients from every explanatory variable will be taken to calculate the mills ratio, which is given by the following equation:

$$\lambda_1 = \frac{\psi(\tilde{\gamma_1} + \tilde{\gamma_2} W_i)}{\Psi((\tilde{\gamma_1} + \tilde{\gamma_2} W_i))} \tag{4.6}$$

One can realize that this equation is for the general case of W_i determinant factors. λ_1 is the inverted mills ratio, $\tilde{\gamma_1}$ is the estimated constant from the probability model, $\tilde{\gamma_2}$ is the estimated angular coefficient from the explanatory variable, ψ is the standard normal probability density function and Ψ is the cumulative distribution function for a standard normal random variable.

Now that the mills ratio has been calculated, it is added to the returns regression as shown by the equation below:

$$Y_i = \beta_0 + \beta_1 Years + \beta_2 Exp_i + \beta_3 Exp2_i + \beta_4 Urban + \beta_5 White + \beta_\lambda \tilde{\lambda_i} + \epsilon_i \tag{4.7}$$

[16] Since this method will be carried on all the years considered, both separately (individual regressions for each year) and all at once (a regression using every year) the notations used here will be general and will not use a specific year as base.

Where Y_i is the real wage for the person i, β_0 is the constant of the regression, β_1 the angular coefficient of Years (years of education), β_2 the angular coefficient of Exp (years of experience), β_3 the angular coefficient of Exp2 (years of experience squared), β_4 the effect of living in an urban area, β_5 the effect of having white skin and β_λ is the angular coefficient of the inverted mills ratio[17].

The Maximun Likelihood Heckman Selection Model has a key difference from the Two-step Heckman Selection Model, which is that it estimates a joint regression using the probability model and the returns to education model:

$$Z_i = \gamma_1 + \gamma_2 W_i + u_i \tag{4.8}$$

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i \tag{4.9}$$

Then, the estimation equation is given by:

$$L = \prod_{i=1}^{n} [(Pr(Z_i) \leq 0)^{1-Z_i} f(Y_i|Z_i) > 0 Pr(Z_i > 0)]^{Z_i} \tag{4.10}$$

The log-likelihood equation is compound from the results when Z_i does not happen (the person is unemployed, Z_i is zero) and the second, when it does happen (the person has a job and consequently income, Z_i is 1). The $f(Y_i|Z_i)$ is the bivariate density.

4.3 Weighted Least Squares

As the main focus of the work is to see the effects of sample selection bias and omitted variables it is necessary to also compute the returns for a approach that does not cover any of these issues, which will be the traditional least squares model corrected for heterokedasticy. Starting with a general equation for linear regression:

$$Y_i = \alpha_i + \beta_1 X_i + u_i \tag{4.11}$$

Where Y_i denotes the explained variable, α_i an intercept, X_i the explanatory variable, β_1 its coefficient and $u_i t$ an error term. For the model to be robust it will need a weight that cancels out the heterokedasticy. This is possible by estimating the uncorrected equation by traditional OLS and storing the residuals. Then compute the variance of the residuals to transform the observations of the traditional model, in a way that the actual equation to be estimated is:

$$\frac{Y_i}{\tilde{\sigma_i}} = \frac{\alpha_i}{\tilde{\sigma_i}} + \beta_1 \frac{X_i}{\tilde{\sigma_i}} + \frac{u_i}{\tilde{\sigma_i}} \tag{4.12}$$

Then, we estimate the equation by OLS applying the σ_i as weights, eliminating the heterokedasticy problem, thus, producing robust estimators.

[17] For those who are interested who men and women compared when considering a model with the same variables (using the same measure for children), the estimations are in the appendix A, as model 2.

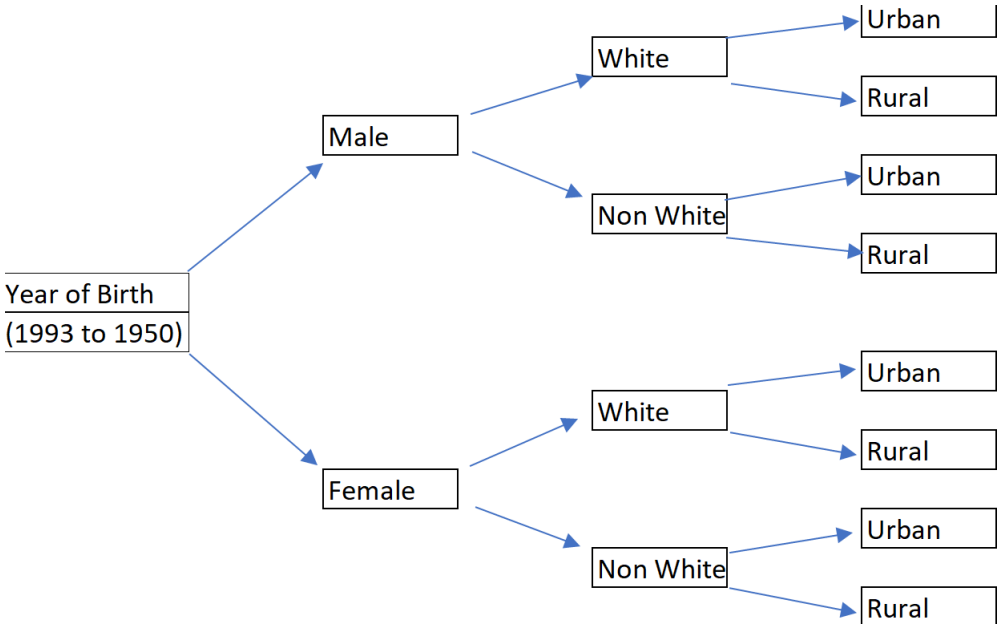
4.4 Data

The data set used in this work is the PNAD from the period of 2011 to 2015. This data frame was chosen because 2011 is the first year after the last demographic census and 2015 is the last year of data available. The PNAD is an yearly survey applied to every state of Brazil, including the Federal District, which aims to gather information about a wide range of subjects such as sex, age, skin colour, area that the person lives and work (urban and rural), wage, work, education, and many other demographic and socioeconomic characteristics.

The core of estimating a pseudo panel model is to divide the sample into different groups (cohorts) using characteristics that are important to the analysis. First of all, the sample was divided into year of birth as it is not only traditional in cohort analysis but it is also an important factor to understand the determinants of the income of a person. It was chosen the same age interval for both pseudo panel and sample selection model, as the estimates will be compared to each other.

Since cohorts are made of characteristics that are constant over time (it must not vary every year like age or experience) it was created a group for each year of birth from 1993 (these persons were 18 years old in 2011) to 1950 (these persons were 65 years old in 2015), this way, the sample was made of people from at least 18 years old to a maximum of 65 years old. In order to perform the sex stratified analysis that will be carried on for the four models, every year of birth cohort had its male and female version. Skin color and rural workers were also characteristics considered to construct the cohorts as enriches the data set with more diverse observations allowing the pseudo panel to capture these important effects. The construction of the cohorts is seen in the diagram below below:

Figure 5 – Cohorts construction



Source:Own elaboration.

Each year has 8 possible cohorts.

1. Year of birth 1 , male, white, rural.
2. Year of birth 1, male, white, urban.
3. Year of birth 1 , male, non white, rural.
4. Year of birth 1 , male, non white, urban.
5. Year of birth 1 , female, white, rural.
6. Year of birth 1 , female, white, urban.
7. Year of birth 1 , female, non white, rural.
8. Year of birth 1, female, non white, urban.

Considering that the year of birth interval has 44 years, there is a total of 352 cohorts per year. The asymptotics of pseudo panel developed by Verbeek (2008) shows that, the longer the panel (more years) and the higher the number of cohorts, more consistent are the estimators. However, there are trade off in both cases. As for longer panels, age cohorts start to comprehend individuals that are too old, reducing drastically the number of observations used to create the cohort. If the length of the panel is 15 (for yearly frequency), and the oldest cohort in the first year is 60 years old, in the last year it will be 75 years old and there won't be many people in this age working, so the number of observations used to create that cohort will be much smaller then the others.

Increasing the number of cohorts also has its disadvantages as, the higher the number of groups, the smaller the number of individuals within each group. That is why it is important to work with large data sets like PNAD, to construct as many number of cohorts possible in the range of 100 observations mean per cohort. Since the number of observations used for each cohort differs from one to another it was used the weight of the person provided by the own data set to have adequately weighted cohorts without having to correct the estimators later on.

In order to create the weekly wage, it was first consider total monthly wage from its main job divided by the number of hours worked by that person in that month. This method was used because there are people in the sample that worked both more or less than what is expected from a full time job, 40 hours per week. So it is safe to assume that people are working at part time jobs (which would make them earn less than a full time job) and people working extra time (which would make them earn more than a full time job), since the wage is from the main work. If a person did not have a job at the week of the interview, she was not included in the sample, only those who were part of the Brazilian workforce were considered. To account for inflation, the wages were deflated using the yearly IPCA, with the base year as 2015.

Table 2 – Variables used on the Pseudo Panel model

Variable	Description
Returns equation	
Real Wage	It was considered the hourly wage of a person as the income from its main job divided by the total number of hours worked.
Experience	Generally, age is used as a proxy to experience. In this work, however, the experience was calculated by subtracting the age of a person from the age he/she started working.
Squared Experience	The experience value to the power of 2.
Cohorts construction	
Skin Colour	In PNAD questionnaires there are 4 different options. White, black, yellow and indigenous. In order to not stratify the sample too much it was separated in two different groups: White and non-white.
Years of Study	The number of years of study completed by the individual.
Sex	Sex of the person with two possible categories: Male or female.
Year of Birth	Year that the person is born. This variable was used, instead of age, because it is better suited to create cohorts, as explained in the methodology section.
Rural Worker	If the person is a rural worker.

Source: Own elaboration.

Experience is an important factor in wages, it is expected that people with more experience will earn more as it is assumed that they would be able to handle the work better as they have been acumulating human capital longer than others, leading to an increase in productivity, this way, individuals with more experience would be rewarded with higher salaries as shown by Topel (1991).

The squared experience is included to capture the effects of human capital depreciation, as stated by Schultz (1961) and Becker (1962). In theory, the positive effects of experience in wages should not be always increasing as people with high levels of experience are also advanced in age, which tends to lead to a decrease in vitality and strenght, which are also key factors of human capital and would directly reduced productivity as was shown by Skirbekk (2008).

Skin color and sex are factors that directly affect a person's wage. Oaxaca (1973) and Blinder (1973) showed that, there was a significant difference between wages for men and women and white and non-white people in the US. The authors also investigated the reason why there was such a large difference between groups, using the Oaxaca-Blinder decomposition, which performs a decomposition of the difference between the estimated values in two different regressions (by groups, sex and race in this case) and sees if the difference comes from the coefficients, the explanatory variables or unknown factors. In Brazil, Davanzo and Ferro (2014) performed the same methodology for sex and rural/urban workers. It was shown that there was a significant difference in both wage and returns for both factors. This way, it is important to consider individuals from these groups as well.

For the sample selection model there are additional variables to be considered to

estimate the probability model.

Table 3 – Variables used on the Sample selection model

Variable	Description
Returns equation	
Real Wage	Income from its main job divided by the total number of hours worked.
Experience	The age of a person minus the age that he/she started working.
Squared Experience	The experience value to the power of 2.
Urban	Dummy equals to 1 if the person lives in a urban area.
White	Dummy equals to 1 if the person is white.
Labor market participation equation	
Partner	Dummy equals to 1 if the person lives with the partner.
Children	Dummy equals to 1 if the person has children.
Children Over 14*	Dummy equals to 1 if person has children over 14.
Children Under 14*	Dummy equals to 1 if person has children under 14.
Age	The person's age in years.
Urban	Dummy equals to 1 if the person lives in a urban area.
White	Dummy equals to 1 if the person is white.
Metropolitan	Dummy equals to 1 if the person lives in a metropolitan area.
North	Dummy equals to 1 if the person lives in the North region.
Northeast	Dummy equals to 1 if the person lives in the Northeast region.
South	Dummy equals to 1 if the person lives in the South region.
Midwest	Dummy equals to 1 if the person lives in the Midwest region.

Source: Own elaboration.
* Variables only available to females.

The marital status of a person has direct effect on its labor participation, income, education and other key variables to understand earnings and schooling. Being married affects men and women differently as seen by Hill (1979). Married men, regardless of skin color, has better wages, works more hours and has a higher chance of being employed than those who are single. For white women, the highest earnings, years of study and labor participation were for those who didn't have partner due to divorce or death (widows).

Having children or not it is also very important. Waldfogel (1997) shows that women without children have more experience (1 year more than those with children), earns more (US$0,30, more than women with children) and are the majority of the female workforce (71% were working at full time jobs, 10 percentage points more than women with children). Considering that younger children are more dependant of the parents, it was also included in the model, two other variables indicating the age of the child. It is expected that children under 14 will have a higher impact in the participation of women in the workforce than those who are over 14. This variable is only available to women due to the limitations of the original data set (PNAD).

The age and years of study are also expected to affect the labor participation as it was shown in the chapter "General Overview of the Brazilian Workforce" that the mean age for

men and women worker were 37,8 and 37,5 years old respectively and the average years of study was 10 years overall.

As seen in Chapter 3, overview of the Brazilian workforce, the skin colour has an important effect in the persons wages. For the 2011-2015 data set, for both men and women, people of white skin color had higher income than those who were not white. So, this variable was also tested to see if affected the labor participation as well.

Where the person lives can also have significant impact in the probability of a person working or not. Even thought the income of a person living in a urban area is higher, people living in the urban area have a higher probability of having a job.

5 Results

5.1 Pseudo Panel

In order to be able to perform a comparative analysis between different methods, the estimations were divided into different samples. For the pseudo panel, there is the total sample, which refers to every individual in the sample, with no restrictions. The female sample refers to individuals of the female sex while the male sample refers to individuals of only male sex. The full sample considers both men and women. By dividing the sample, it is going to be possible to see how the returns of different groups react to different models. In this section, first it is going to be analyzed the descriptive statistics of the pseudo panel, then the regression results by sample. After finishing the pseudo panel analysis, the descriptive statistics of the sample used in the sample bias selection models will be presented, then the results of each model for males and females will be investigated. Lastly, the returns of education for every model (including the WLS as benchmark) will be compared to one another.

Table 4 – Descriptive Statistics - Pseudo Panel - Males

Variable	Mean	Standard Deviation		
	Overall	Overall	Between	Within
Year	2013	1,414615	0	1,414615
Year of Birth	1971,5	12,7056500	12,7346500	0
Skin Colour	0,5	0,5002843	0,5014265	0
Rural Worker	0,5	0,5002843	0,5014265	0
Years of Study	6,8949180	2,4969980	2,4798040	0,3369761
Experience	28,0149200	13,6893800	13,6415300	1,4678030
Squared Experience	984,0961000	782,493000	778,7682000	92,6019900
Real Wage	11,8275400	7,2072490	6,0883240	3,8788430
Total Number of Cohorts	880			
Number of cohorts	176			
Length of the Pseudo Panel in years	5			

Source: Own elaboration.

Table 5 – Descriptive Statistics - Pseudo Panel - Females

Variable	Mean	Standard Deviation		
	Overall	Overall	Between	Within
Year	2013	1,414615	0	1,414615
Year of Birth	1971,5	12,70565	12,73465	0
Skin Colour	0,5	0,5002843	0,5014265	0
Rural Worker	0,5	0,5002843	0,5014265	0
Years of Study	7,68277	2,792094	2,706329	0,7105642
Experience	26,9641920	13,7989200	13,7355300	1,6136840
Squared Experience	938,0676000	770,5808000	766,0203000	98,3770700
Real Wage	10,1618800	18,8684900	10,1410500	15,9263000
Total Number of Cohorts	880			
Number of cohorts	176			
Length of the Pseudo Panel in years	5			

Source: Own elaboration.

The table 4 and 5 shows the overall mean (considering the total sample) of each variable and the standard deviation overall (total sample), between (between cohorts) and within (within cohorts over the years) for both samples. The variable year does not change between cohorts as the pseudo panel is strongly balanced[18]. All of the dummies used to create the cohorts do not change within the cohort itself, as if a person was a certain type of worker (rural or urban) or had a specific skin color, it did not change with the course of the years. Skin colour and rural worker have a overall mean of 0,5. This result shows that the cohorts were divided equally for each of this categories. Even thought year of birth is also a qualitative variables, it does not take value of 1 or 0 like the other dummies. Each year takes its own value, that is why the overall mean is not 0,5 but 1971,5 which is the exact half of the period considered.[19]

For the male sample, shown in table 4, the average years of study for the female sample was 6,89 years with 28,01 years of experience and a hourly wage of R$10,16. The average years of study for the female sample shown in table 5 was 7,68 years with 26,96 years of experience and a hourly wage of R$10,16. The number of cohorts per year was 176, in a total of 880 cohorts for the 2011-2015 period (176 per year for 5 years). The total number of cohorts for the full sample is 1760 (880 from the male sample +880 from the female sample) with a number of cohorts per year of 352. The average number of observations used for each cohort is 440, which is 340 more than what is needed to have a consistent estimation, according to Verbeek (2008).

[18] Strongly balanced means that every observation is present for every year.

[19] The period considered ranges from 1950 to 1993, which is equal to a 43 year interval. 43 divided by 2 is 21,5. 21,5 plus 1950 is 1971,5 or 1993 minus 21,5 is also equal to 1971,5. This means that there are no repeated years of birth for the same cohort in the same time period.

Table 6 – Results - Pseudo Panel

Full Sample		
Variable	Coefficient	Standard Error
Years of Study	0,1476237*	0,0140719
Experience	0,0988162*	0,0115452
Squared Experience	-0,0004462*	0,0001852
Constant	-1,1816810*	0,1876597
Females		
Years of Study	0,1472002*	0,0184380
Experience	0,0795175*	0,0182847
Squared Experience	-0,0002279	0,0002994
Constant	-0,9985667*	0,2952120
Males		
Years of Study	0,1353641*	0,0267687
Experience	0,1265675*	0,0136730
Squared Experience	-0,000756*	0,0002084
Constant	-1,433938*	0,2227393

Source: Own elaboration.
* Statistically significant at 5%.

The estimation results for all samples are reported in table 6, which shows that all coefficients were significant for an α of 5%, except the squared experience for females. All three coefficients had the expected signs in every sample, with years of education and experience affecting positively the wages, while the squared experience had a small negative effect.

There was a small difference between the effect of years of study in the natural log of the hourly wage. For the full sample, the returns for one year of study was of 14,76% which is only 0,04 percentage points more than for the female sample, that had an effect of 14,72%. The lowest return of education for the three samples were for men. Each year of study will lead to an increment of 13,53% in its hourly wage. A 1,23 and 1,19 percentage points less than the total sample and the female sample, respectively.

The marginal effects of experience in the log of hourly wages presented significant differences across the three samples. The group of males, that had the lowest return per year of education, had the highest return for experience, an increment of 12,5% a 4,6 percentage point difference to the female sample, that had an average return of 7,9%. The magnitude of the marginal effects of experience for the total sample was in between the sex groups, with an effect of 9,8% per additional year of experience. As expected, the squared experience had a very small impact in wages, reducing the income of the individuals from the all three samples. However, the value itself is not what should be focused, as this variable is included to see if the positive effects of experience gets bigger or smaller over the years. The negative sign shows that experience has diminishing returns. The constant was statistically significant at 5% for

every sample.

5.2 Sample Selection Models

In this section the results for the sample selection model will be analyzed. First it is going to be reported the descriptive statistics for every variable used in the model to have a better understanding of the characteristics of the sample.

Table 7 – Descriptive statistics - Sample selection model

	Females		Males	
Variable	Mean	Standard Deviation	Mean	Standard Deviation
Age	38,9085000	13,1584400	38,2463600	13,1359000
Partner	0,6021720	0,4894500	0,6250591	0,4841081
Hours	36,1199600	13,5922100	42,1768800	11,6460300
Age that started working	16,0231900	4,9233090	14,5346300	3,8614540
Children	0,5875842	0,4922697	0,5060618	0,4999637
Years of study	8,7819880	4,3927890	8,1920580	4,3849740
Job	0,5769963	0,4940364	0,82285432	0,3818089
Experience	22,1048400	13,2795000	23,7081600	13,6832800
Squared Experience	664,9685000	665,8759000	749,3087000	723,1111000
Urban	0,8769537	0,3284906	0,854492	0,3526126
White	0,4404369	0,4964400	0,4177587	0,4931904
Real Wage	9,5477740	58,4249800	16,4487300	85,8822800
Metropolitan	0,3911546	0,4880093	0,372809	0,4835524
North	0,1474912	0,3545952	0,1546895	0,3616088
Northeast	0,2871908	0,4524518	0,2787878	0,4484033
South	0,1057151	0,3074728	0,1061466	0,3080254
Midwest	0,1062593	0,3081695	0,1079865	0,3103636
Children Under 14*	0,2440657	0,4295323	-	-
Children Over 14*	0,2470742	0,4313106	-	-

Source: Own elaboration.
* Variables only for women.

For the descriptive statistics shown in table 7 of all of the dummy variables, the results that are reported in the mean category reflects the percentage relative to the total population that has that characteristic. The only restriction of the sample was the age interval, ranging from 18 to 65 years. The results indicate that, comparatively, there were more males living with the partner than the females, a 2,3 percentage point difference. For the number of people with jobs, 82,28% of the males in the sample had jobs, while only 57,69% of the females. The difference between them is so big that even in total number of people, man outnumber women, even though there are slightly more women than men in the sample. The number of women living with their children corresponded to 58,75% of the total sample, while men, 50,06%. The PNAD data set also shows if the mother lives with the child, without the father, separating for age groups. For this sample it was considered children younger and older than 14 years old.

The former represented, 24,44% of the women in the sample, the latter 24,70. The number of females living in a urban area was higher than males, by a very small margin, 87,67% in total.

For the non dummy variables it is possible to compare all the variables as they are either used in the returns regression (years, exp, exp2 and wage) or to create a variable that was used (hours, age that started working, age), which means that they are available for men and women, as the returns equation is the same for both of them. The mean age of the sample used was of 38,9 years for women and 38,2 years for men about 8,4 months of difference[20]. As expected, women have, in general, more years of study than men as the female group had 8,78 years while men had 8,19 years of study. According to the data set, men have 1,6 years of experience more than women, that is because, even thought men and women have almost the same mean age, men start to work earlier, at 14,5 years of age, while women at 16. Men also have higher hourly wages, with a R$6,90 difference over women. As for the number of hours worked per week, women are reported to work 36 hours and men 42.

Next is the analysis of the results from the labor market participation estimation and the return regressions. The estimations are separated by sex because it is expected that the coefficients will behave differently for each group (HILL, 1979) and also because in the female probability equation was used two different variables for children, as explained in the last paragraph. This way, the female equation has the same variables as the male one except for the variable $children$, that was broken into different age groups for females. However, for those who are interested in how the results would vary for both models, using the same variables for men and women and for the full sample, there are additional estimations in the appendix A.

Considering the Maximum Likelihood results for females presented in table 8, it is seen that all of the coefficients that influence the insertion in the labor market are statistically significant at 5%. At first, it is seen that women with younger children have less chances of being employed as for those with older children. Living with the partner and having white skin colour also effected negatively the probability of women working. As expected, those with higher education were more likely to have jobs. The place were they lived also had an impact on labor market participation, as regions like North and Northeast affected negatively as living in the South region or in a Metropolitan area had positive effects. Considering that the Rho is different than zero, there is correlation between the errors of the two equations and, therefore, there is need to use the sample selection bias. The LR test p-value rejects the null hypothesis that the returns equation is independent from the labor market participation. The error covariance, lambda[21] of both equations, since it is significant it also provides evidence that there is dependence between the two equations

[20] It is important to keep in mind that these descriptive statistics are related to the total sample with only the age restriction. As one may think that men should be older than women because in the General overview of the Brazilian Labor Force the average wage was higher for men. That holds true only for those with jobs, not for the full sample.

[21] Lambda is Rho x Sigma. Sigma is the standard error of the residuals from the first equation.

Table 8 – Maximum Likelihood Results - Females

Female		
Variable	Coefficient	Standard Error
Returns		
Years of Study	0,0892405*	0,0004821
Experience	0,0254142*	0,0003713
Squared Experience	-0,0002935*	0,00000076
Urban	0,2372658*	0,0055159
White	0,208858*	0,0028705
Constant	0,9534562*	0,0107101
Labor Market Participation		
Children Under 14	-0,0301642*	0,0043836
Children Over 14	0,067293*	0,0044087
Age	-0,0019332*	0,0001546
Years of Study	0,07300955*	0,0004369
Partner	-0,0558455*	0,0036397
Urban	0,1313082*	0,0061034
White	-0,0574388*	0,0037595
Metropolitan	0,0249391*	0,0036989
North	-0,1946496*	0,0054266
Northeast	-0,2189895*	0,0044514
South	0,1401852*	0,0060317
Midwest	0,0289422*	0,0058926
Constant	-0,5761138*	0,0106074
Rho	-0,3615879*	0,007326
Sigma	0,797183*	0,0016966
Lambda	-0,2882517*	0,0063431
LR Test P-value	0,000	

Source: Own elaboration.
* Statistically significant at 5%.

All of the variables that explained the log linear hourly wage were statistically significant at 5%. The results report that the years of study for women have a positive effect of 8,92%, while experience only 2,5%. Living in the urban area and having white skin colour also effects positively the wages of women, 23,75% and 20,88%, respectively. The squared experience was also statistically significant even thought its coefficient has a very low number, indicating that the returns per year of experience decreases over time.

As the table 9, the Maximum Likelihood results for males reports that all of the coefficients that determines the insertion in the labor market were statistically significant at 5%. Having children affected positively the chances of being employed. Most of the variables share the same signs as for the females group. The ones that were different are living with the partner, that had a positive effect, the variables metropolitan and urban area showed negative effects, in the labor market, for men. Considering that the Rho is different than zero, there is correlation between the errors of the two equations and, therefore, there is need to use the

Table 9 – Maximum Likelihood Results - Males

Male		
Variable	Coefficient	Standard Error
Returns		
Years of Study	0,0906189*	0,0003157
Experience	0,0316451*	0,0003235
Squared Experience	-0,0003499*	0,00000061
Urban	0,3799218*	0,0036581
White	0,2100286*	0,0024833
Constant	0,8434751*	0,0055709
Labor Market Participation		
Age	-0,0107985*	0,0001758
Years of Study	0,0342749*	0,0005167
Children	0,3596426*	0,0057236
Partner	0,5536478*	0,0056479
Urban	-0,3410256*	0,0069362
White	-0,0465879*	0,0046093
Metropolitan	-0,0215015*	0,0044891
North	-0,0602363*	0,0065858
Northeast	-0,2218731*	0,0053734
South	0,0194717*	0,0073605
Midwest	0,1336992*	0,0073908
Constant	0,6912437*	0,0113565
Rho	-0,3555412*	0,004652
Sigma	0,7899831*	0,0010025
Lambda	-0,2808716*	0,0038746
LR Test P-value	0,000	

Source: Own elaboration.
* Statistically significant at 5%.

sample selection bias. Lambda was also different than zero, indicating that the probabilistic model showed that the variables included in the labor market equation affected significantly the coefficients of the returns estimation. The LR test p-value rejects the null hypothesis that both equations are independent.

As for the effects of education on wages for males, the results show that one additional year of study would lead to an increment of 9% in the income, while experience 3,1%, one percentage point more than for females. Having white skin colour and living in the urban area had a significant increase of 38% and 21% respectively. The squared experience had the expected negative sign showing the decrease of the returns for experience over time. The constant was statistically significant at 5% as well.

As for the results in table 10, which shows the two step model for females, the probit model reported that all of the variables considered were statistically significant at 5%, except for the Midwest area. Years of study, living with older children and living in the urban area had

Table 10 – Two Step Results - Females

Female		
Variable	Coefficient	Standard Error
Returns		
Years of Study	0,0689756*	0,0010693
Experience	0,0264568*	0,0003756
Squared Experience	-0,0003062*	0,0000076
Urban	0,1808977*	0,0065089
White	0,2019419*	0,0032064
Constant	1,506855*	0,0278318
Labor Market Participation		
Children Under 14	-0,0299302*	0,0044853
Children Over 14	0,0803773*	0,00449
Age	-0,0027141*	0,0001539
Years of Study	0,0742234*	0,0004413
Partner	-0,0735396*	0,0036918
Urban	0,1539123*	0,0060893
White	-0,0469739*	0,0037701
Metropolitan	-0,0160734*	0,0036726
North	-0,1950049*	0,0055461
Northeast	-0,1646937*	0,0044584
South	0,1509094*	0,0061656
Midwest	0,0028327	0,0060302
Constant	-0,5004749*	0,0105188
Rho	-0,78502	-
Sigma	0,95376747	-
Lambda	-0,7487264*	0,0219034

Source: Own elaboration.
* Statistically significant at 5%.

also positive effects on labor market insertion for females, just like the Maximum Likelihood model. Living with the partner and with younger children had a negative effect as expected. The constant was also statistically significant and negative. Lambda was statistically different than zero[22], showing that the inverse mills ratio affects the coefficients from the returns equation significantly.

The results for the two step model showed the smallest returns to education so far, at a return rate of 6,8%. Living in the urban area e being white was also beneficial for the income with an effect of 18% and 20%, respectively. The experience was also statistically significant at 5% with a slightly higher effect than the Maximum Likelihood model, of 2,6%. The squared experience presented similar results as the maximum likelihood model, negative with a very small coefficient. The constant was positive and statistically significant.

[22] The two step model does not show the standard error of Rho and Sigma as they are estimated directly. In the ML estimation, Rho and Sigma are estimated separately, that is why there are standard errors.

Table 11 – Two Step Results - Males

Male		
Variable	Coefficient	Standard Error
Returns		
Years of Study	0,0877295 *	0,0003395
Experience	0,0288402*	0,0003305
Squared Experience	-0,0002965*	0,00000062
Urban	0,4085718*	0,0039029
White	0,2093495*	0,0064973
Constant	0,9478572*	0,0055066
Labor Market Participation		
Age	-0,0115115*	0,0001755
Years of Study	0,0351332*	0,000528
Children	0,3655825*	0,0058354
Partner	0,5466241*	0,0057655
Urban	-0,3090695*	0,0069344
White	-0,0313736*	0,0046143
Metropolitan	-0,0645545*	0,0045233
North	-0,0397972*	0,0066928
Northeast	-0,1452209*	0,0053998
South	0,0312903*	0,0075184
Midwest	0,0974045*	0,0075665
Constant	0,7670319*	0,0114216
Rho	-0,64225	-
Sigma	0,83414856	-
Lambda	-0,535728*	0,0086073

Source: Own elaboration.
* Statistically significant at 5%.

Lastly, table 11 shows the results for the two step model for males. Again, all of the variables were statistically significant at 5% for the labor market equation. Having children and living with the partner had a positive effect on the insertion in the labor market. Years of study and living in the South region had a positive cofficient for all 4 models. Interestingly, living in the urban area, being white and living in the metropolitan region had negative effects on the matter. The constant was statistically significant and positive. The statistically significant lambda shows that there is negative error covariance between the two estimations, therefore, there is need for selection bias correction.

As for the returns estimation, the results show that one additional year of study would lead to an increment of 8,7% in the income, while experience 2,8%, a smaller return per experience comparing with the Maximum Likelihood model. The squared experience had the expected negative sign, showing the decrease of the returns for experience over time. Again, living in the urban area and having white skin colour had important effects on the income, 40% for the former and 20% for the latter. The constant was statistically significant at 5%, affecting the wages positively.

5.3 Model comparison

Table 12 – Returns to education comparison - All models

Female		
Model	Coefficient	Standard Error
Pseudo Panel	0,1472002*	0,018438
ML	0,0892405*	0,0004821
Two Step	0,0689756*	0,0010693
WLS	0,0870374*	0,0004419
Male		
Pseudo Panel	0,1353641*	0,0267687
ML	0,0906189*	0,0003157
Two Step	0,0877295*	0,0003395
WLS	0,0919131*	0,0003049

Source: Own elaboration.
* Statistically significant at 5%.

Table 12 groups all the results for the four models. In everyone of them, the returns were statistically significant at 5%, which means it is possible to see the differences from every approach. Comparing the four models for the female sample, the results show that the pseudo panel method leads to the highest returns per education with 5,8 percentage points more than the second highest score, from the Maximum Likelihood model. The smallest returns came from the two step model at 1,2 percentage points less than the WLS benchmark, that had the second lowest rate. In comparison to the traditional model, accounting for group heterogeneity increased the returns of education in 6 percentage points, while accounting for the sample bias (ML model) led to higher coefficients as well, compared to the two step model, with an increase of 1,81 percentage points.

As for the differences between models for the male sample, the gaps were not so wide as for the females. The highest return was reported by the pseudo panel approach, followed by the WLS, ML and finally the two step model. The ML and WLS reported very close returns, as the difference between then is only 0,1 percentage point. Accounting for group heterogeneity led to a coefficient of 4,8 percentage points higher than the two step model and 4,4 for the WLS benchmark. For both groups the Pseudo Panel had the highest returns, while the two step model, the lowest. The ML and WLS reported very similar returns, getting as close as 0,1 percentage point difference for the male sample.

5.4 Discussion

The results found in this work suggests that there is a significant difference between the returns of education when considering group heterogeneity and sample selection bias. The

effect of an additional year of study in the wages of a person was 41% higher (6 percentage points) in the pseudo panel model, in comparison to the WLS for females and 33% higher (4,4 percentage points) for males. These results are in concordance with the literature as shown by Mendiratta and Gupt (2013) whom found that, for the regular least squares estimator, the returns to education were of 11,6% per additional year of study, while the pseudo-panel estimator had 15,04% which represents a 3,37 percentage point difference. Sampaio (2007) also sees similar results, with IV estimations being 28% bigger than the traditional OLS and 43% smaller for the sample selection model.

The differences between estimators for different genders is also seen in the work of Warunsiri and McNown (2010) , as the results shows higher returns to pseudo panel estimators for both sexes. For men, the traditional model had returns of 10,7% , while accounting for group heterogeneity showed an average return of 12,6%, total difference of 1,9 percentage points. As for women, the returns were of 12,9% for the OLS model and 17,8% for the cohort estimation.

As for the effect of selection bias, the results are mixed. The two step and maximum likelihood had different results, as the former had significantly more different results than the WLS while the latter was much closer. The literature shows that it can happen to have little difference between the results found by sample selection model and OLS. This is shown by Himaz and Aturupane (2016) whose findings see no difference (less than 0,1 percentage point) between the coefficients found using OLS or using the Heckman approach for additional years of study, years of experience and living in the rural area. The only difference between estimators was found for one variable for the female group, been non muslim had a 0,4 percentage higher effect on the OLS estimator.

Even though, in this work there was only one case where the ML estimation showed positive bias, when compared to the WLS estimation, it is something that can happen. Ribeiro and Bastos (2014) shows that for people with 8 to 10 years of study for differents states of Brazil, the number of statistically significant positive and negative selection bias (compared to OLS), at 5% significance level, was the same, with a difference of 4 percentage points in total (all of positive bias minus the negative), which means that in some cases the returns found by a sample selection model is higher than the OLS and sometimes lower.

The probability model estimated in the sample selection approach showed how different characteristics affected the participation of men and women in the labor market. As expected, education affects positively the probability of the person being employed, a statement that has already been seen in previous papers such as the ones from Curi and Menezes Filho (2004) and Barbosa Filho and Pessoa (2008).

The effects of living with the partner were the opposite for both genders. While males had a positive effect for living with their females partners, females saw a decrease in labor participation when found in this situation. Stolzenberg and Waite (1984) found similar evidence

to this finds, except, the authors were using marital status instead of effectively living with the parents as used in this work.

As for the presence of children, the estimations also show different results for the two groups. There are papers which studies the effects of the presence of children in the household, regarding labor market participation like Eissa and Hoynes (2004) and Guimarães and Santos (2010). But, in this work, there was an additional desegregation of this variable in children under and over 14 years of age, for females. The presence of younger children in the household affected negatively the chances of women being employed while having older children had the opposite effect. For males, having a child in the house, had positive effects on the matter. This results are in accordance with Pagani and Marenzi (2008) and Waldfogel (1997).

Skin color and the area of residence was also an important factors when it came to employment. Being white had a negative impact in the probability of getting a job but had a positive effect on wages, which corroborates with the findings of Coelho, Veszteg and Soares (2010). Living in the metropolitan area also had a negative impact on the matters, evidence found by Scorzafave and Menezes Filho (2001) as well.

6 Concluding Remarks

In order to test the hypothesis that there are effects which are not observable in the returns to education in Brazil, and how each of them would affect the magnitude of the coefficients, this work estimates the mincer equation (MINCER, 1974) for four different models. The first model is the pseudo panel, which takes into account non observable group characteristics that affect wages, like the skill and talent of an individual that could lead to biased estimators for a model that does not consider this differences (heterogeneity). By stratifying people into groups of characteristics that do not change over time, it is possible to analyze the changes in the effects of years of study and experience in wages.

The second and third models are based on the work of Heckman (1979), which considers the possibility of selection bias in the labor market. As different groups of people faces different challenges to get a job, the sample of working people may consist of individuals with a predominant set of characteristics, which would lead to a sample bias in traditional estimation methods such as the OLS. This way, it is estimated, both the probability of the person being employed and the returns for education corrected by the probability model. The two-step approach estimates them separately while the Maximum likelihood model has the advantage estimating them jointly. Finally, a WLS model was used as a benchmark, considering that it was a variation of one of the first models to be used in this type of analysis as seen in the work of Griliches (1977).

The cohort stratified analysis shows supporting evidence that there is significant effects of group heterogeneity in the returns to education in Brazil for the years of 2011 to 2015, corroborating to the first hypothesis tested in this work. For all three samples (full sample, only men, only women) the effect of years of study in the wages were higher for the pseudo panel than those estimated by the WLS method. The biggest difference seen was in the female sample, 6 percentage point gap between these two estimators.

As for labor market participation, the findings show that men and women do face different circumstances when trying to get a job, as the same characteristics showed different effects for each gender, providing supportive evidence to the second hypothesis tested in this work. Having children had different effects depending on their age, as younger children affected negatively the chances of females getting jobs. As for females with older children and males with children of any age the effects were the opposite. Another factor that did have the opposite effect comparing males and females was the variable partner. It was shown that living with the partner influenced positively the probability of being part of the work force for man and negatively for women.

The results for the sample selection bias, show that not only men and women behave

differently in entering the labor market, but these differences also reach the returns per year of education, especially. One additional year of study led to a 6,8% increase in the wage for females in the two step model, while not accounting for this bias the coefficient was of 8,7%. The difference was closer for males as the two step was only 0,6 percentage point less than the WLS.

As for the effects of experience in the hourly wages, the two step model estimated smaller coefficients than the Maximun Likelihood and WLS. The difference between the ML and WLS models were of 0,5 percentage point for females and 0,6 for males, indicating that the selection bias does not have as great as of impact in experience as the individual heterogeneity does, but still significant, as these differences represent 17,85% of the total value of the return. Experience was statistically significant for both men and women for all for models. Men were the ones who benefit the most out of additional years of experience as they had the highest returns for both sample selection and the benchmark model.

The comparison of returns per year of study among models showed that the WLS tends to underestimate the effect of higher education when it does not take into consideration the group heterogeneity but overestimates when it does not consider the selection bias, therefore, there is bias from both ends in the WLS model. Determining which group had the highest return per additional year of education would depend on the model chosen. The pseudo panel reported that females have the highest return, while the WLS, Maximun Likelihood and Two step model, men.

This work, therefore, contributes to the field by providing supporting evidence that the group effects and sample bias affect directly the rate of return per additional year of education in the hourly wage in Brazil for males and females. The WLS estimator is downward biased in the first case and upward biased in the second, with the biggest gap been for the female sample, especially between the pseudo panel and the two step model. There is also significant difference on how males and females are affected by each bias as the biggest change in the coefficients was seen in the female sample.

As for future contributions to this field, it will be interesting to see how a longer pseudo panel would compare to the one used in this analysis, as the bigger the sample, more information is considered, leading to more robust models. In addition to the estimation of the pseudo panel and the selection models, one could consider adding the model from Garen (1984) of selective bias and see how it compares for the last years of PNAD available.

The addition to the new data set provided by IBGE, the PNAD continua would also be an important complement to the data set used, as there are some similar variables in both of them, allowing to estimate similar models to the ones used in this work. One possible problem that the researcher may face is that more complex models are not possible to be constructed using the PNAD continua data, as the data set is much smaller and less detailed. The biggest advantage of using it, is that it has a periodicity of 4 months which increases the robustness of

the pseudo panel.

As to regard of the utilization of true panel data, this possibility would definitely be a great contribution to the research, especially if compared to a pseudo panel model to see if both approaches are similar in the results and to see how the differences over time behave between the different estimators proposed. At the moment of writing this work, the only true panel data available is the PME, but it does not have an official individual identifier. The author would have to rely on the unofficial ones or create one of his own. Even then, the models would have to be compatible as the PNAD data set is completely different from the PME, so, depending on the subjects that the researcher wants to study it may even be impossible to have a model that can provide a fair comparison between estimators.

As for policy recommendation, the pseudo panel estimations showed that the returns to education, when considering the group heterogeneity bias are even higher, which means that, it is even more beneficial for the individual to study when considering the non observable variables that also influence the wage of a person. This way, investments in education by the government to increase the supply of higher levels of study, such as universities, would lead to even higher gains to society than previously thought. The sample selection bias showed the reduction in returns especially for women, this means that, due to the different challenges that females face when looking for work influence negatively their wages. This evidence suggests that public policies that can help women enter the workforce is necessary to reduce the inequality of income.

More aggregations are still possible with the sample available. Adding the estimation of the returns equation for different professions, formal and informal work (all stratified by sex, if possible) are also viable possibilities to see how these four different models perform, allowing more comparison between groups and methods and try to find what is the exact cause of each bias. Developing a model that can take into account both biases would be able to produce even more robust results, as the pseudo panel and sample selection bias approaches are relevant in the studies to returns to education in Brazil.

References

ADAMS, B. L.; KING, J.; PENNER, A. M.; BANDELJ, N.; KANJUO-MRČELA, A. The returns to education and labor market sorting in slovenia, 1993–2007. *Research in Social Stratification and Mobility*, Elsevier, v. 47, p. 55–65, 2017.

ANDRADE, A. A. S. d.; MENEZES FILHO, N. A. O papel da oferta de trabalho no comportamento dos retornos à educação no brasil. Instituto de Pesquisa Econômica Aplicada (Ipea), 2005.

BARBOSA FILHO, F. d. H.; PESSÔA, S. Retorno da educação no brasil. Instituto de Pesquisa Econômica Aplicada (Ipea), 2008.

BARRO, R. J. Human capital and growth. *American economic review*, v. 91, n. 2, p. 12–17, 2001.

BECKER, G. S. Investment in human capital: A theoretical analysis. *Journal of political economy*, The University of Chicago Press, v. 70, n. 5, Part 2, p. 9–49, 1962.

BECKER, G. S.; MURPHY, K. M.; TAMURA, R. Human capital, fertility, and economic growth. *Journal of political economy*, The University of Chicago Press, v. 98, n. 5, Part 2, p. S12–S37, 1990.

BLINDER, A. S. Wage discrimination: reduced form and structural estimates. *Journal of Human resources*, JSTOR, p. 436–455, 1973.

BRAND, J. E.; XIE, Y. Who benefits most from college? evidence for negative selection in heterogeneous economic returns to higher education. *American sociological review*, Sage Publications Sage CA: Los Angeles, CA, v. 75, n. 2, p. 273–302, 2010.

BROWNING, M.; DEATON, A.; IRISH, M. A profitable approach to labor supply and commodity demands over the life-cycle. *Econometrica: journal of the econometric society*, JSTOR, p. 503–543, 1985.

BUCHINSKY, M. Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of human resources*, JSTOR, p. 88–126, 1998.

CARD, D. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, Wiley Online Library, v. 69, n. 5, p. 1127–1160, 2001.

CHOW, G. C. Capital formation and economic growth in china. *The Quarterly Journal of Economics*, MIT Press, v. 108, n. 3, p. 809–842, 1993.

CLARK, C. Theory of economic growth. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 112–116, 1949.

COELHO, D.; VESZTEG, R.; SOARES, F. V. *Regressão quantílica com correção para a seletividade amostral: estimativa dos retornos educacionais e diferenciais raciais na distribuição de salários das mulheres no Brasil*. [S.l.], 2010.

COLLADO, M. D. Estimating dynamic models from time series of independent cross-sections. *Journal of Econometrics*, Elsevier, v. 82, n. 1, p. 37–62, 1997.

CURI, A. Z.; MENEZES FILHO, N. A. Os determinantes das transições ocupacionais no mercado de trabalho brasileiro. *Anais do XXXII Encontro Nacional da Anpec*, 2004.

DARGAY, J. The effect of prices and income on car travel in the uk. *Transportation Research Part A: Policy and Practice*, Elsevier, v. 41, n. 10, p. 949–960, 2007.

DAVANZO, E. S.; FERRO, A. R. Retornos à educação: uma análise da redução do diferencial salarial por anos de estudo no brasil no período de 2001 a 2012. *Anais do Encontro da Associação Nacional de Pós-graduação em Economia–ANPEC*, 2014.

DEATON, A. Panel data from time series of cross-sections. *Journal of econometrics*, Elsevier, v. 30, n. 1-2, p. 109–126, 1985.

DUFLO, E. Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American economic review*, v. 91, n. 4, p. 795–813, 2001.

EISSA, N.; HOYNES, H. W. Taxes and the labor market participation of married couples: the earned income tax credit. *Journal of public Economics*, Elsevier, v. 88, n. 9-10, p. 1931–1958, 2004.

FRIEDMAN, M. The permanent income hypothesis. In: *A theory of the consumption function*. [S.l.]: Princeton University Press, 1957. p. 20–37.

GAREN, J. The returns to schooling: A selectivity bias approach with a continuous choice variable. *Econometrica (pre-1986)*, Blackwell Publishing Ltd., v. 52, n. 5, p. 1199, 1984.

GIRMA, S. A quasi-differencing approach to dynamic modelling from a time series of independent cross-sections. *Journal of Econometrics*, Elsevier, v. 98, n. 2, p. 365–383, 2000.

GLOMM, G.; RAVIKUMAR, B. Public versus private investment in human capital: endogenous growth and income inequality. *Journal of political economy*, The University of Chicago Press, v. 100, n. 4, p. 818–834, 1992.

GONÇALVES, S. L.; MENEZES FILHO, N. A. *O salário mínimo e a oferta de trabalho das famílias pobres: uma abordagem coletiva com os dados da PNAD Contínua (2012-2015)*. [S.l.], 2015.

GRILICHES, Z. Estimating the returns to schooling: Some econometric problems. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 1–22, 1977.

GROGGER, J. Welfare reform, returns to experience, and wages: using reservation wages to account for sample selection bias. *The Review of Economics and Statistics*, MIT Press, v. 91, n. 3, p. 490–502, 2009.

GUIMARÃES, P. W.; SANTOS, C. M. dos. Determinantes da ocupação no mercado de trabalho de maridos e esposas. *Revista Brasileira de Gestão e Desenvolvimento Regional*, v. 6, n. 2, 2010.

GUO, M.; ZHANG, Y.; YE, J. Does a foreign degree pay? the return to foreign education in china. *Review of Development Economics*, Wiley Online Library, v. 23, n. 1, p. 415–434, 2019.

HASKEL, J. *The decline in unskilled employment in UK manufacturing*. [S.l.], 1996.

HECKMAN, J. Shadow prices, market wages, and labor supply. *Econometrica*, v. 42, n. 4, p. 679–694, 1974.

HECKMAN, J. J. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, JSTOR, p. 153–161, 1979.

HILL, M. S. The wage effects of marital status and children. *Journal of Human Resources*, JSTOR, p. 579–594, 1979.

HIMAZ, R.; ATURUPANE, H. Returns to education in sri lanka: a pseudo-panel approach. *Education Economics*, Taylor & Francis, v. 24, n. 3, p. 300–311, 2016.

JORGENSON, D. W.; HO, M. S.; STIROH, K. J. Growth of us industries and investments in information technology and higher education. *Economic Systems Research*, Taylor & Francis, v. 15, n. 3, p. 279–325, 2003.

JOSAN, I.-J. Human capital investment. *Manager*, Editura Universităţii din Bucureşti, n. 15, p. 104–112, 2012.

KUZNETS, S. *Modern economic growth: findings and reflections*. [S.l.]: Nobel foundation, 1966.

LEAL, C. I. S.; WERLANG, S. R. d. C. Retornos em educação no brasil: 1976/1986. Escola de Pós-Graduação em Economia da FGV, 1989.

LEBERGOTT, S. The shape of the income distribution. *The American Economic Review*, JSTOR, p. 328–347, 1959.

MARCELO, R.; WYLLIE, R. Retornos para educação no brasil: evidências empíricas adicionais. *Economia Aplicada*, SciELO Brasil, v. 10, n. 3, p. 349–365, 2006.

MENDIRATTA, P.; GUPT, Y. Private returns to education in india by gender and location: A pseudo-panel approach. *Arthaniti-Journal of Economic Theory and Practice*, SAGE Publications Sage India: New Delhi, India, v. 12, n. 1-2, p. 48–69, 2013.

MENEZES FILHO, N. A. A evolucão da educacão no brasil e seu impacto no mercado de trabalho. *Instituto Futuro Brasil*, v. 43, 2001.

MENEZES FILHO, N. A.; MENDES, M.; ALMEIDA, E. S. d. O diferencial de salários formal-informal no brasil: segmentação ou viés de seleção? *Revista Brasileira de Economia*, SciELO Brasil, v. 58, n. 2, p. 235–248, 2004.

MINCER, J. Investment in human capital and personal income distribution. *Journal of political economy*, The University of Chicago Press, v. 66, n. 4, p. 281–302, 1958.

MINCER, J. Schooling, experience, and earnings. human behavior & social institutions no. 2. ERIC, 1974.

MOFFITT, R. Identification and estimation of dynamic models with a time series of repeated cross-sections. *Journal of Econometrics*, Elsevier, v. 59, n. 1-2, p. 99–123, 1993.

NAWATA, K. Estimation of sample selection bias models by the maximum likelihood estimator and heckman's two-step estimator. *Economics Letters*, Elsevier, v. 45, n. 1, p. 33–40, 1994.

NELSON, R. R.; PHELPS, E. S. Investment in humans, technological diffusion, and economic growth. *The American economic review*, JSTOR, v. 56, n. 1/2, p. 69–75, 1966.

OAXACA, R. Male-female wage differentials in urban labor markets. *International economic review*, JSTOR, p. 693–709, 1973.

PAGANI, L.; MARENZI, A. The labor market participation of sandwich generation italian women. *Journal of Family and Economic Issues*, Springer, v. 29, n. 3, p. 427–444, 2008.

PEET, E. D.; FINK, G.; FAWZI, W. Returns to education in developing countries: Evidence from the living standards and measurement study surveys. *Economics of Education Review*, Elsevier, v. 49, p. 69–90, 2015.

PEREIRA, V. d. F.; BRAGA, M.; MENDONÇA, T. de. Avaliação dos retornos à escolaridade para trabalhadores do sexo masculino no brasil. *Embrapa Semiárido-Artigo em periódico indexado (ALICE)*, Revista de Economia Contemporânea, Rio de Janeiro, v. 17, n. 1, p. 153-176 . . . , 2013.

RIBAS, R. P.; SOARES, S. S. D. *Sobre o painel da Pesquisa Mensal de Emprego (PME) do IBGE*. [S.l.], 2008.

RIBEIRO, E. P.; BASTOS, V. M. Viés de seleção, retornos à educação e migração no brasil. *ENCONTRO BRASILEIRO DE ECONOMETRIA*, v. 26, p. 1–19, 2004.

SAMPAIO, A. V. Retorno de escolaridade no brasil e no paraná em 2004. *V Ecopar*, 2007.

SARGAN, J. D. The distribution of wealth. *Econometrica, Journal of the Econometric Society*, JSTOR, p. 568–590, 1957.

SCHULTZ, T. W. Capital formation by education. *Journal of political economy*, The University of Chicago Press, v. 68, n. 6, p. 571–583, 1960.

SCHULTZ, T. W. Investment in human capital. *The American economic review*, JSTOR, p. 1–17, 1961.

SCORZAFAVE, L. G.; MENEZES FILHO, N. A. Participação feminina no mercado de trabalho brasileiro: evolução e determinantes. Instituto de Pesquisa Econômica Aplicada (Ipea), 2001.

SILVA, D. B.; CARVALHO, A.; NERI, M. C. Diferenciais de salários por raça e gênero: aplicação dos procedimentos de oaxaca e heckman em pesquisas amostrais complexas. Escola de Pós-Graduação em Economia da FGV, 2006.

SKIRBEKK, V. Age and productivity capacity: descriptions, causes and policy options. *Ageing horizons*, Oxford Institute of Population Ageing, v. 8, p. 4–12, 2008.

SOLOW, R. M. Technical progress, capital formation, and economic growth. *The American Economic Review*, JSTOR, v. 52, n. 2, p. 76–86, 1962.

STOLZENBERG, R. M.; WAITE, L. J. Local labor markets, children and labor force participation of wives. *Demography*, Springer, v. 21, n. 2, p. 157–170, 1984.

SWAN, T. W. Economic growth and capital accumulation. *Economic record*, Wiley Online Library, v. 32, n. 2, p. 334–361, 1956.

The World Bank. *GDP per capita*. 2017. Data retrieved from World Development Indicator <https://data.worldbank.org/indicator/ny.gdp.pcap.cd>,.

THEODOSSIOU, I. A quarterly earnings series for manual and non-manual workers in great britain. *International Journal of Manpower*, MCB UP Ltd, v. 11, n. 6, p. 23–26, 1990.

TOPEL, R. Specific capital, mobility, and wages: Wages rise with job seniority. *Journal of political Economy*, The University of Chicago Press, v. 99, n. 1, p. 145–176, 1991.

VERBEEK, M. Pseudo-panels and repeated cross-sections. In: *The econometrics of panel data*. [S.l.]: Springer, 2008. p. 369–383.

WALDFOGEL, J. The effect of children on women's wages. *American sociological review*, JSTOR, p. 209–217, 1997.

WARUNSIRI, S.; MCNOWN, R. The returns to education in thailand: A pseudo-panel approach. *World Development*, Elsevier, v. 38, n. 11, p. 1616–1625, 2010.

WILLIS, R. J. Wage determinants: A survey and reinterpretation of human capital earnings functions. *Handbook of labor economics*, Elsevier, v. 1, p. 525–602, 1986.

ZHANG, J. Estimates of the returns to schooling in taiwan: Evidence from a regression discontinuity design. 2019.

ZWICK, C. Demographic variation: Its impact on consumer behavior. *The Review of Economics and Statistics*, JSTOR, p. 451–456, 1957.

A Complete Estimations

Table 13 – Complete Results - Pseudo Panel

Full Sample		
Variable	Coefficient	Standard Error
Years of Study	0,1476237*	0,0140719
Experience	0,0988162*	0,0115452
Squared Experience	-0,0004462*	0,0001852
Constant	-1,181681*	0,1876597
R-squared within	0,2194	
R-squared between	0,2550	
R-squared overall	0,2185	
P-value F test	0,000	
Number of Obs	1760	
Number of groups	352	
Females		
Years of Study	0,1472002*	0,018438
Experience	0,0795175*	0,0182847
Squared Experience	-0,0002279	0,0002994
Constant	-0,9985667*	0,295212
R-squared within	0,1702	
R-squared between	0,2301	
R-squared overall	0,1851	
P-value F test	0,000	
Number of Obs	880	
Number of groups	176	
Males		
Years of Study	0,1353641*	0,0267687
Experience	0,1265675*	0,013673
Squared Experience	-0,000756*	0,0002084
Constant	-1,433938*	0,2227393
R-squared within	0,3292	
R-squared between	0,2947	
R-squared overall	0,2675	
P-value F test	0,000	
Number of Obs	880	
Number of groups	176	

Source: Own elaboration.

* Statistically significant at 5%.

Table 14 – Maximum Likelihood Results - Females

Female		
Variable	Coefficient	Standard Error
Returns		
Years of Study	0,0892405*	0,0004821
Experience	0,0254142*	0,0003713
Squared Experience	-0,0002935*	0,00000076
Urban	0,2372658*	0,0055159
White	0,208858*	0,0028705
Constant	0,9534562*	0,0107101
Labor Market Participation		
Children Under 14	-0,0301642*	0,0043836
Children Over 14	0,067293*	0,0044087
Age	-0,0019332*	0,0001546
Years of Study	0,07300955*	0,0004369
Partner	-0,0558455*	0,0036397
Urban	0,1313082*	0,0061034
White	-0,0574388*	0,0037595
Metropolitan	0,0249391*	0,0036989
North	-0,1946496*	0,0054266
Northeast	-0,2189895*	0,0044514
South	0,1401852*	0,0060317
Midwest	0,0289422*	0,0058926
Constant	-0,5761138*	0,0106074
Rho	-0,3615879*	0,007326
Sigma	0,797183*	0,0016966
Lambda	-0,2882517*	0,0063431
LR Test P-value	0,000	
P-value Chi2	0,000	
Censored Obs	253.363	
Uncensored Obs	309.969	
Log likelihood	-721097,4	

Source: Own elaboration.

* Statistically significant at 5%.

Table 15 – Maximum Likelihood Results - Males

Male		
Variable	Coefficient	Standard Error
Returns		
Years of Study	0,0906189*	0,0003157
Experience	0,0316451*	0,0003235
Squared Experience	-0,0003499*	0,00000061
Urban	0,3799218*	0,0036581
White	0,2100286*	0,0024833
Constant	0,8434751*	0,0055709
Labor Market Participation		
Age	-0,0107985*	0,0001758
Years of Study	0,0342749*	0,0005167
Children	0,3596426*	0,0057236
Partner	0,5536478*	0,0056479
Urban	-0,3410256*	0,0069362
White	-0,0465879*	0,0046093
Metropolitan	-0,0215015*	0,0044891
North	-0,0602363*	0,0065858
Northeast	-0,2218731*	0,0053734
South	0,0194717*	0,0073605
Midwest	0,1336992*	0,0073908
Constant	0,6912437*	0,0113565
Rho	-0,3555412*	0,004652
Sigma	0,7899831*	0,0010025
Lambda	-0,2808716*	0,0038746
LR Test P-value	0,000	
P-value Chi2	0,000	
Censored Obs	98.207	
Uncensored Obs	428.195	
Log likelihood	-728872,6	

Source: Own elaboration.

* Statistically significant at 5%.

Table 16 – Two Step Results - Females

Female		
Variable	Coefficient	Standard Error
Returns		
Years of Study	0,0689756*	0,0010693
Experience	0,0264568*	0,0003756
Squared Experience	-0,0003062*	0,0000076
Urban	0,1808977*	0,0065089
White	0,2019419*	0,0032064
Constant	1,506855*	0,0278318
Labor Market Participation		
Children Under 14	-0,0299302*	0,0044853
Children Over 14	0,0803773*	0,00449
Age	-0,0027141*	0,0001539
Years of Study	0,0742234*	0,0004413
Partner	-0,0735396*	0,0036918
Urban	0,1539123*	0,0060893
White	-0,0469739*	0,0037701
Metropolitan	-0,0160734*	0,0036726
North	-0,1950049*	0,0055461
Northeast	-0,1646937*	0,0044584
South	0,1509094*	0,0061656
Midwest	0,0028327	0,0060302
Constant	-0,5004749*	0,0105188
Rho	-0,78502	-
Sigma	0,95376747	-
Lambda	-0,7487264*	0,0219034
P-value Chi2	0,000	
Censored Obs	253,363	
Uncensored Obs	309,969	

Source: Own elaboration.

* Statistically significant at 5%.

Table 17 – Two Step Results - Males

Male		
Variable	Coefficient	Standard Error
Returns		
Years of Study	0,0877295 *	0,0003395
Experience	0,0288402*	0,0003305
Squared Experience	-0,0002965*	0,00000062
Urban	0,4085718*	0,0039029
White	0,2093495*	0,0064973
Constant	0,9478572*	0,0055066
Labor Market Participation		
Age	-0,0115115*	0,0001755
Years of Study	0,0351332*	0,000528
Children	0,3655825*	0,0058354
Partner	0,5466241*	0,0057655
Urban	-0,3090695*	0,0069344
White	-0,0313736*	0,0046143
Metropolitan	-0,0645545*	0,0045233
North	-0,0397972*	0,0066928
Northeast	-0,1452209*	0,0053998
South	0,0312903*	0,0075184
Midwest	0,0974045*	0,0075665
Constant	0,7670319*	0,0114216
Rho	-0,64225	-
Sigma	0,83414856	-
Lambda	-0,535728*	0,0086073
P-value Chi2	0,000	
Censored Obs	98.207	
Uncensored Obs	428.195	

Source: Own Elaboration.

* Statistically significant at 5%.

Table 18 – Maximum Likelihood Results Model 2 - Females

Female		
Variable	Coefficient	Standard Error
Returns		
Years of Study	0,0898256*	0,0004818
Experience	0,0245716*	0,0003716
Squared Experience	-0,0002765*	0,0000076
Urban	0,2385185*	0,0055108
White	0,2087528*	0,0028665
Constant	0,9472136*	0,0108601
Labor Market Participation		
Age	-0,0019264*	0,0001418
Years of Study	0,074182*	0,000438
Children	0,1853045*	0,0039641
Partner	-0,1401438*	0,0038813
Urban	0,1302943*	0,0061112
White	-0,0503862*	0,0037659
Metropolitan	0,0261227*	0,0037102
North	-0,1961845*	0,0054388
Northeast	-0,217537*	0,0044681
South	0,1398338*	0,0060468
Midwest	0,0269579*	0,0059097
Constant	-0,6654592*	0,0103826
Rho	-0,3508777*	0,0074796
Sigma	0,83414856*	0,0016703
Lambda	-0,2789993*	0,0064241
LR Test P-value	0,000	
P-value Chi2	0,000	
Censored Obs	253.363	
Uncensored Obs	309.969	
Log likelihood	-721224	

Source: Own Elaboration.

* Statistically significant at 5%.

Table 19 – Two Step Results Model 2 - Females

Female		
Variable	Coefficient	Standard Error
Returns		
Years of Study	0,0792575*	0,0008652
Experience	0,0243904*	0,0003734
Squared Experience	-0,0002677*	0,0000078
Urban	0,2085049*	0,0060348
White	0,2047579*	0,0030131
Constant	1,246165*	0,0226712
Labor Market Participation		
Age	-0,0025939*	0,0001409
Years of Study	0,0754323*	0,0004429
Children	0,1960117*	0,0040325
Partner	-0,1627267*	0,0039086
Urban	0,1523961*	0,0060959
White	-0,0396200*	0,0037751
Metropolitan	0,0134869*	0,0036762
North	-0,1959962*	0,0055517
Northeast	-0,1643571*	0,0044621
South	0,1496864*	0,0061732
Midwest	0,0017537*	0,0060380
Constant	-0,5981102*	0,013106
Rho	-0,60833	-
Sigma	0,86331764	-
Lambda	-0,5251835*	0,0173555
P-value Chi2	0,000	
Censored Obs	253.363	
Uncensored Obs	309.969	

Source: Own Elaboration.

* Statistically significant at 5%.

Table 20 – Maximum Likelihood Results Model 2 - Full Sample

Female		
Variable	Coefficient	Standard Error
Returns		
Years of Study	0,084705*	0,0002575
Experience	0,0284301*	0,0002454
Squared Experience	-0,0003069*	0,0000048
Urban	0,3103723*	0,0030826
White	0,2068139*	0,0019244
Constant	0,9924404*	0,0049224
Labor Market Participation		
Age	-0,0037263*	0,000104
Years of Study	0,0459954*	0,0003155
Children	0,1333609*	0,0030385
Partner	0,1881458*	0,0030082
Urban	-0,126206*	0,0041822
White	-0,0576564*	0,0027809
Metropolitan	0,0117474*	0,0026984
North	-0,1236771*	0,0039244
Northeast	-0,2255374*	0,0032287
South	0,0874931*	0,004451
Midwest	0,0752237*	0,0043397
Constant	0,5981102*	0,013106
Rho	-0,4222159*	0,0036347
Sigma	0,8180443*	0,0010000
Lambda	-0,3453913*	0,0032981
LR Test P-value	0,000	
P-value Chi2	0,000	
Censored Obs	351.570	
Uncensored Obs	738.164	
Log likelihood	-1526347	

Source: Own Elaboration.

* Statistically significant at 5%.

Table 21 – Two Step Results Model 2 - Full Sample

Female		
Variable	Coefficient	Standard Error
Returns		
Years of Study	0,062223 *	0,000541
Experience	0,0241846*	0,0002976
Squared Experience	-0,0002123*	0,0000057
Urban	0,3579338*	0,0045222
White	0,1998064*	0,0028003
Constant	1,681742*	0,013507
Labor Market Participation		
Age	-0,0043958*	0,0001041
Years of Study	0,0471357*	0,0003207
Children	0,1513873*	0,0031205
Partner	0,1560757*	0,0030900
Urban	-0,0956985*	0,0041793
White	-0,0443294*	0,0027907
Metropolitan	0,0358813*	0,0027423
North	-0,1141728*	0,0040467
Northeast	-0,151805*	0,0032883
South	0,1040765*	0,0045948
Midwest	0,0369211*	0,0044927
Constant	0,1379128*	0,0071399
Rho	-1	-
Sigma	1,3076051	-
Lambda	-1,307605*	0,0170688
P-value Chi2	0,000	
Censored Obs	351.570	
Uncensored Obs	738.164	

Source: Own Elaboration.

* Statistically significant at 5%.

Table 22 – Complete WLS - Females

Variable	Coefficients	Std. Error	P-value
Years of Study	0,0870374	0,0004419	0,000
Experience	0,0218568	0,0004177	0,000
Squared Experience	-0,0003716	0,00000849	0,000
Urban	0,5419405	0,0054703	0,000
White	0,1926766	0,003323	0,000
Constant	0,3630838	0,007482	0,000
Number of Obs	309.969		
Adj. R-squared	0,1944		

Own elaboration.

Table 23 – Complete WLS - Males

Variable	Coefficients	Std. Error	P-value
Years of Study	0,0919131	0,0003049	0,000
Experience	0,0342111	0,0003162	0,000
Squared Experience	-0,0004091	0.0000064	0,000
Urban	0,3593112	0,0034917	0,000
White	0,2086084	0,002418	0,000
Constant	0,8118351	0,0050806	0,000
Number of Obs	428.195		
Adj. R-squared	0,2753		

Source: Own elaboration.