Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Policy Economics

Proximity to Prosperity: Estimating Returns to Higher Education using Distance to Institutions in Brazil

Milena M. Nakashima (521731)

Abstract

This paper explores the causal effect of higher education (HE) attainment on income in São Paulo, Brazil. The methodology and theoretical framework draws from Becker's Human Capital Model, Mincer's Equation and Cards' Instrumental Variables approach to estimate returns to HE, using proximity to a HE institution as an instrument. A new data set was created for this study, combining 2000 Brazilian Institute of Geography and Statistics (IBGE) Census data with Statewise System for Data Analysis Foundation (SEADE) coordinates and collected information on the opening dates of public university campuses. Individual-level findings using ordinary least squares (OLS) indicate a 1.0%-1.16% increase in income and a 4.3%-5% increase using instrumental variables (IV). These findings align with previous studies. To address possible sample selection bias, a regression discontinuity design (RDD) at the cohort level is conducted, exploiting a sudden decrease in campus distance, post-1955. The RDD analysis estimates a 6.7% return for the entire sample and 2.0% for the mixed-race population. While the results support the theory of positive returns to higher education, robustness checks raise concerns about accuracy and external validity.

Keywords: higher education, returns to education, ordinary least squares, instrumental variables, regression discontinuity design, Brazil

Supervisor:	Olivier Marie
Second assessor:	Dinand Webbink
Date final version:	October 9, 2023

Contents

List of Figures

List of Tables

1 Introduction

"(...) the return to education is not a single parameter in the population, but rather a random variable that may vary with other characteristics of individuals(...)"
- Card (1999)

Tertiary education attainment has rapidly increased in Brazil. Between 1995 and 2017, undergraduate enrollment increased from 1.7 million to an estimated 8 million (OECD, 2018). As a growing number of Brazilians pursue higher education, it is essential to explore its effect on later life outcomes. Notably, returns to education have been extensively addressed in economic literature and explored across various contexts. The general consensus in the field is that education wields a significant influence on both individuals and society. Specifically, educational attainment has been linked to labor market outcomes and economic growth. Undoubtedly, studying the returns to education can provide insightful policy implications. This importance is amplified in Brazil; a country characterized by great levels of inequality among socio-economic and ethnic groups (Arretche, 2018). Notably, the country's turbulent history with democratic institutions hindered the development of higher education in the country. Further details on the context during which higher education institutions (HEIs) developed in the country is elaborated on. This study focuses on the state of São Paulo (SP), the most populous state in the country, with 44.4 million inhabitants accounting for 21.8% of the Brazilian population (*2022 IBGE Census*, 2023).

In the on-going discussion about returns to higher education, two main perspectives stand out. On one hand, it is argued that higher education attainment is beneficial for individual later life outcomes. Based on the seminal work by economists such as Becker (1962), Mincer (1974) and Card (1993), theoretical and empirical studies have supported this perspective. A number of papers suggest that an additional year of education and higher education attainment have a positive effect on consequent earnings. Moreover, previous research highlights the heterogeneous returns to education by gender, ethnicity and socio-economic background. Given Brazil's diverse population, this is particularly relevant. Studies focused on Brazil suggest that higher education attainment has been suggested to significantly reduce social inequalities by increasing social mobility (Menezes Filho & Kirschbaum, 2019). On the other hand, there is skepticism about the importance of higher education. According to Hanushek (2016), studies often examine the impact of higher education in developed countries, prompting concerns about the applicability of their conclusions to developing countries. Particularly in countries where the provision of basic education is limited, expanding access to higher education may have lower social benefits. Furthermore, the mere establishment of higher education institutions does not address the main limitations (financial, social and geographic) which hinder higher education attainment.

A number of studies have investigated the returns to education in the context of Brazil. However, a literature review indicates that few studies focus on estimating the returns to higher education in particular. In many studies, the returns to an additional year of education, regardless of the level, are estimated. Various methods of estimating returns to education have been introduced in previous literature, however, studies focused on Brazil seem to usually employ a simple ordinary least squares (OLS) approach. This study aims to contribute to the understanding of returns to higher education in the context of Brazil, by employing a variety of estimation methods. Notably, this paper applies the instrumental variable (IV) approach introduced in Card (1993) and draws a causal inference of higher education on income, using proximity to higher education institutions as an instrument. To the best of my knowledge, this study is the first to exploit variations in the distance to HEIs in the context of Brazil.

To pursue this methodology, a new data set is constructed for this study. The study combines individual observations reported in the 2000 Population Census data from the Brazilian Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatistica, IBGE) and coordinates from The Statewide System for Data Analysis Foundation (Fundação Sistema Estadual de Análise de Dados, SEADE). Additionally, new data was gathered on the opening dates of university campuses. Leveraging this new data set, this paper employs the OLS, IV, and a Sharp Regression Discontinuity Design (RDD) approaches to estimate returns. As aforementioned, the instrument used in the IV approach is an individual's distance to a HEI. Moreover, the RDD exploits a discontinuity in the average distance to a HEI in 1955.

The rest of the paper is structured as follows. Section 2 provides context to the Brazilian higher education system, particularly focusing on the development of HEIs. Next, Section 3 provides a literature review and Section 4 details the theoretical framework of this study. The data and identification strategy are presented in Sections 5 and 6, respectively. Results are presented in Section 7 with robustness checks conducted in Section 8. The limitations of the study are detailed in Section 9 followed by a discussion in Section 10. Section 11 concludes.

2 The Brazilian Higher Education System

The Brazilian education system can be divided into three main phases: fundamental (primary and secondary), high school and higher education. Both private and public institutions offer education for these phases.

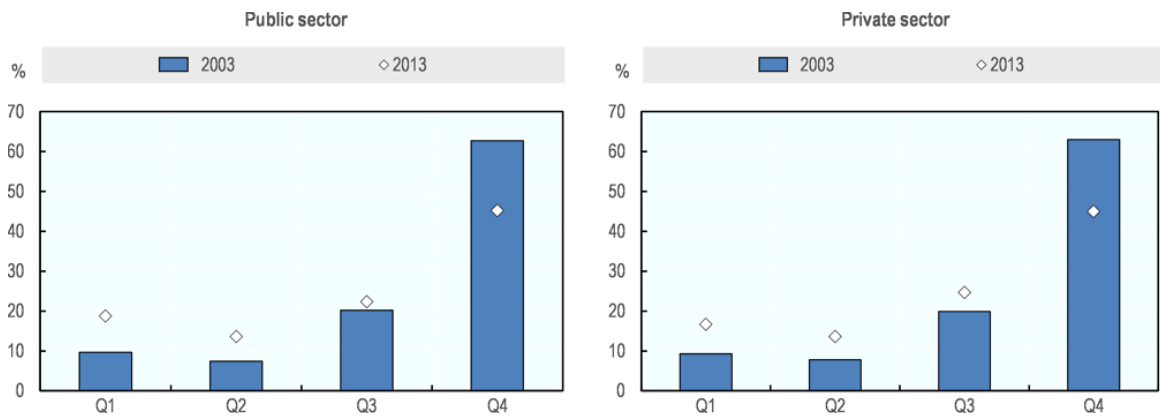
Brazil's higher education system has undergone significant reforms and expansion over the past few decades leading to an increase in the number of public and private institutions. Contrary to the common belief that private universities outperform their public counterparts, this is not the case in Brazil. In fact, public universities in the country provide the best quality of higher education and outperform private institutions in global rankings (Mello, 2022). In order to obtain the highest quality education, high-income students often begin their academic journey in the private sector and later enroll in public universities. Students attending public schools for fundamental education and high school often received a lower level of education than their private school counterparts (McCowan, 2007).

2.1 Admissions & Enrollment

Before 2010, students were required to take admission exams for particular higher education institutions, referred to as '*vestibular*'. Currently, admission to a public university is most dependent on the student's performance on the national entrance exam known as the Exame Nacional do Ensino Médio (ENEM). The ENEM assesses students' performance in various subjects related to their field of interest. A common practice for students who want to score highly on the ENEM is to attend preparatory classes called '*cursinho*'. Notably, these courses are typically offered by private institutions at a substantial tuition fee. Consequentially, the current education system limits low-income students from obtaining higher education due to financial restrictions.

The number of individuals participating in the ENEM has varied throughout the years. In 2016, an estimated 5 million registrations were recorded. This figure declined in 2021, to an estimated 2 million people and increased to 3.9 million registrations in 2023. Notably, a substantial number of students apply for higher education compared to the available spots. For example, Universidade de São Paulo (USP), one of the most renowned university in Brazil, only offered a total of 11.147 positions for new students in 2023. As a result, there is fierce competition among students who want to attend public institutions. Given the systemic inequalities faced by low-income students from public high schools, a larger proportion of students from private high schools attend higher education. Consequentially, fewer low-income individuals participate in higher education, as illustrated in Figure 1.

Figure 1: Share of enrollment in higher education institutions, by income quartile and sector



Note: Q1 refers to the 25% poorest individuals from the total population, whereas Q4 refers to the 25% richest. *Source*: Nascimento and Verhine, 2017[19]

Given these barriers, a variety of quotas were introduced as part of efforts to promote inclusivity across ethnic and socio-economic groups in higher education (Mello, 2022). An example of this is the Quota Law (Lei de Cotas), introduced in 2012. This affirmative action policy set quotas on federal universities for the enrollment of low-income students with a public school background. Despite these measures, inequality in higher education persists, as depicted in Figure 1.

2.2 Political Context

The development of higher education institutions in Brazil has been heavily influenced by the country's complex relationship with democracy (dos Santos & Custodio, 2022). A simplified timeline of the main political developments in Brazilian history is as follows:

- **1824 : Promulgation of Brazil's First Constitution** after the proclamation of independence in 1822.
- **1888 : The "Lei Áurea" (Golden Law)** is signed, officially abolishing slavery.
- **1889 - 1930 : The First Brazilian Republic (The Old Republic).** Characterized by political instability, military interventions, and the dominance by coffee oligarchs.
- **1930 - 1945 : The Vargas Era.** Getúlio Vargas assumes power through a coup and establishes an authoritarian regime.
- **1945 - 1964 : The Second Brazilian Republic.** Brazil returns to democracy, adopting a new constitution.
- **1964 - 1985 : The Military Dictatorship.** Following a military coup, a repressive authoritarian military dictatorship takes power.
- **1985 - 1988 : Transition to Democracy.** Civilian rule is restored and democratic elections are held.
- **1988 : Promulgation of Brazil's Seventh Constitution.**

Evidently, democratic institutions in Brazil have faced significant challenges throughout the country's history. These developments influenced HEIs as they impacted educational funding, public policy and choices in governance. Between 1930 and 1985, the country was ruled under an authoritarian regime with a brief democratic period during The Second Brazilian Republic. This hindered the mitigation of social and economic inequalities.

Nevertheless, inequality in higher education attainment existed before authoritarian governments (Klein & Schwartzman, 1993). Historically, education was exclusively available to the elite, perpetuating long-lasting socioeconomic and racial inequalities (dos Santos & Custodio, 2022). While the turbulent political context exacerbated these inequalities, their deep-rooted nature in Brazilian history cannot be overlooked.

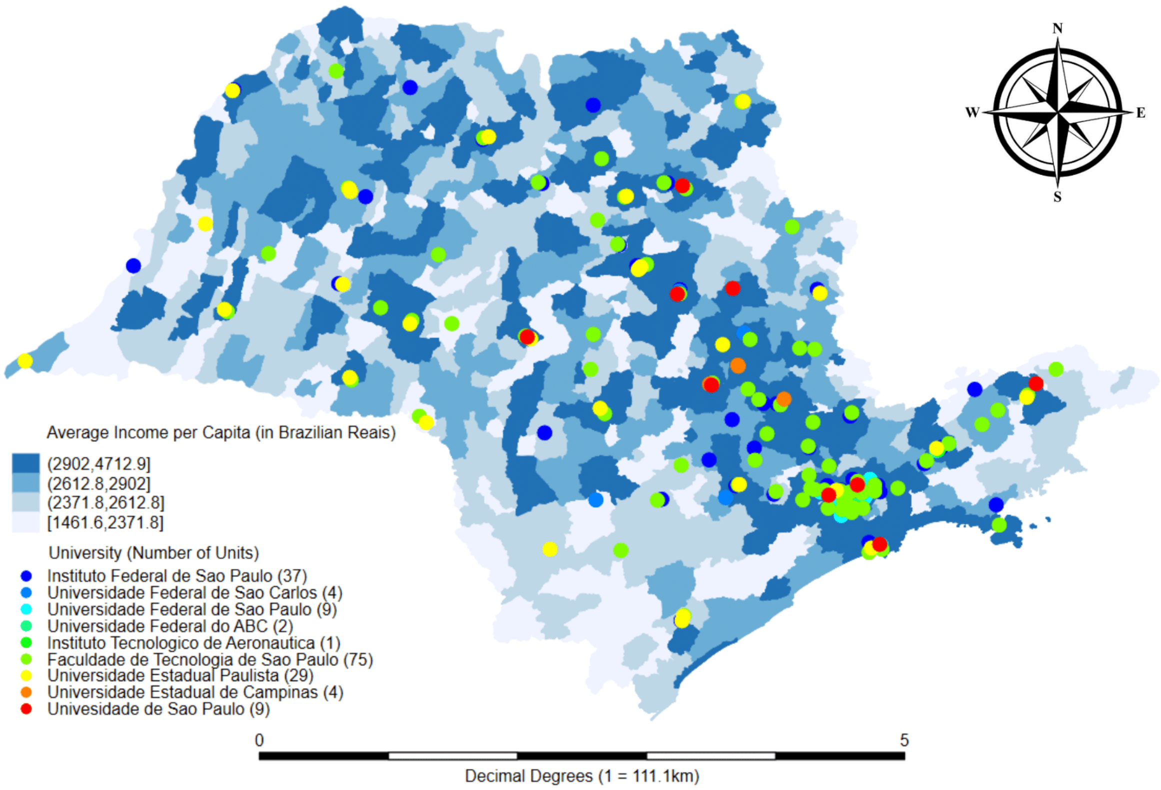
2.3 Development of Higher Education Institutions (HEIs)

Brazil was colonized by Portugal between 1500-1822. During this time, the creation of HEIs in the country was prohibited by the colonizers. Consequentially, the development of HEIs in Brazil began relatively late compared to other South American countries. The first institutions to train professionals in the country were created in the 19th century. At the start of the 20th century, the government attempted to establish the first university called *Universidade de Amazonas* (De Mello E Souza, 1991).

Since then, a large number universities have been founded across the country. The expansion of public HEIs in the country is often linked to the cycles of rapid economic growth (Schwartzman, 2012). According to the 2021 Higher Education Census, 2574 HEIs were active, of which 2261 were private and 313 public. In 2021, the public sector was composed of 119 federal, 134 state and 60 municipal institutions. Brazil's position as the fifth largest country in the world explains the large number of HEIs in the country. In order to simplify the analysis, this paper focuses on public HEIs in the state of SP, depicted in Figure 2.

Figure 2 illustrates the state of São Paulo, which consists of 645 municipalities. A total of 9 public HEIs are active in the state and they have various campuses scattered around the area. Table 10 in the Appendix presents the names of the 9 HEIs. The number of campuses registered per institution in 2021 varied from 1 to 75. In this paper, university campuses will also be referred to as units or institutions. Visually, the map presents fewer units in municipalities with a lower GDP per capita. The number of units are most densely packed in the metropolitan megacity SP (seen South East in Figure 2) This raises an interest in the relationship between economic outcomes and HEIs.

Figure 2: Map of the state of São Paulo illustrating the distribution of state and federal university units



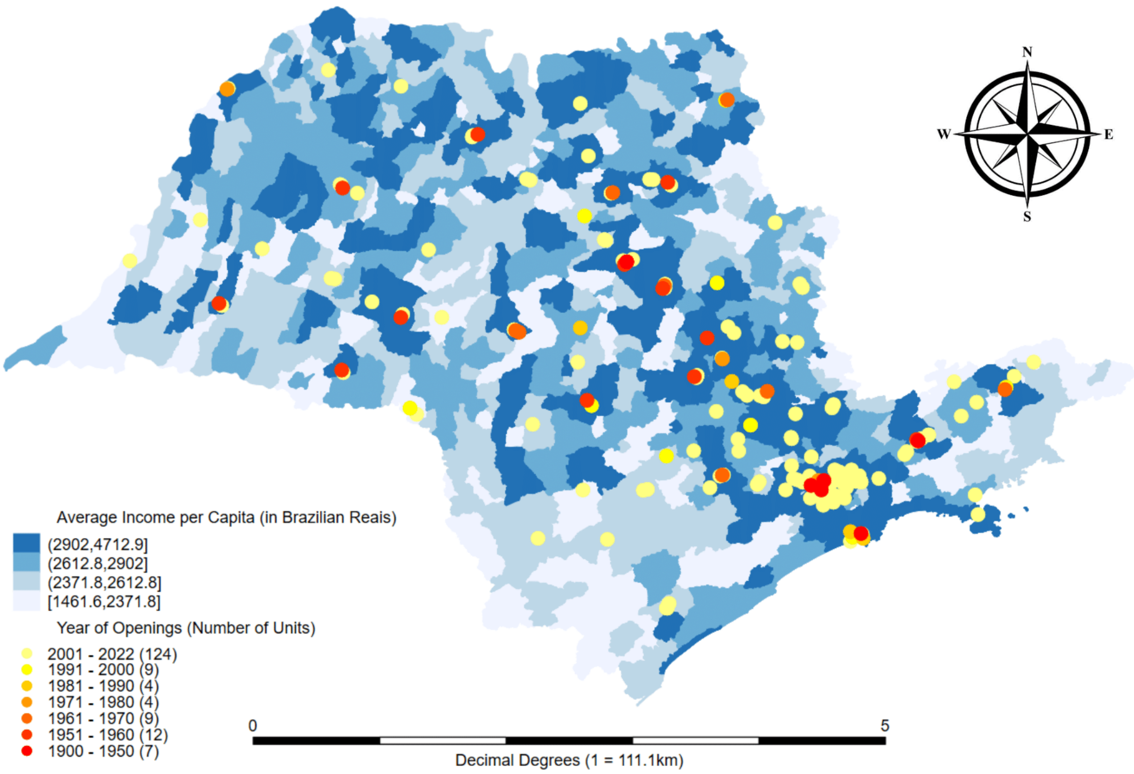
Note: The term university units is synonymous to university campuses

Source: Opening dates collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

Similar to national trends, HEIs have greatly expanded in SP. This development is illustrated in Figure 3. The map highlights two major factors in the history of HEIs in the state. First, the largest proportion of units were established after 2000. As depicted in the map, 124 units were established between 2001-2022. This increase is considerably higher than the earlier phases of the expansion of HEIs. For example, between 1900 and 1950, a total of 7 units were established. Second, earlier established units brought HEIs closer to the population in remote areas. As illustrated in Figure 3, the established units in the 1950s greatly reduced the proximity between universities and municipalities compared to the openings in the 21st century.

Around 1955, four university campuses opened in the state of SP which greatly reduced the average distance to a HEI: São Paulo State University(Universidade Estadual Paulista, UNESP) Aracatuba (1954), UNESP São Jose dos Campos (1954) UNESP Assis (1956), and Universidade Estadual de Campinas in Piracicaba (1955). Notably, the cities of Piracicaba and São Jose dos Campos are located near the city of SP. The cities of Aracatuba and Assis are located further away, in the west of the state.

Figure 3: Map of the state of São Paulo illustrating the opening years of state and federal university units

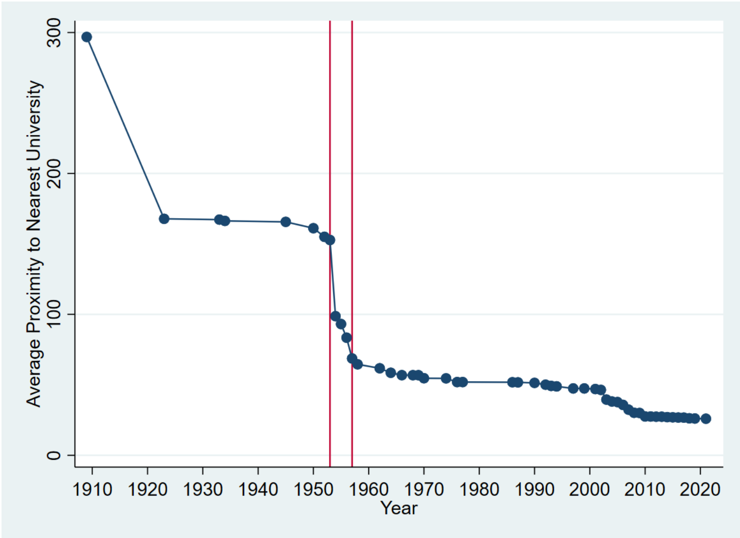


Note: The term university units is synonymous to university campuses

Source: Opening dates collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

Figure 4 further highlights the significant reduction in the average university proximity in the 1950s. While a considerable number of campuses opened in the early 21st century, the average proximity to a HEI was not greatly impacted due to the pre-existing closeness between them.

Figure 4: Average proximity (in Km) to nearest university unit in the state of São Paulo



Note: The term university units is synonymous to university campuses. Observations included in the figure reflect years during which at least one new campus opened.

Source: Opening dates of university units collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university coordinates.

Despite the turbulent political environment in Brazil, HEIs have persistently expanded. This development raises interest in understanding the effects of higher education in the country.

3 Literature Review

The literature review conducted for this study focused on classic papers that have shaped our understanding of returns to higher education. Sections 3.1 and 3.2 outline general fundamental concepts while Section 3.3 focuses on studies in the context of Brazil.

3.1 The Human Capital Model

Human capital refers to the skills, abilities and knowledge that an individual possesses (Becker, 1962). Introduced by Becker (1962), the Human Capital Model (HCM) suggests that individuals can enhance their human capital by acquiring education and training. These inputs are also referred to as schooling and post-school investments, respectively. As a result of these investments, individuals can improve their productivity, obtain higher earnings and stimulate economic growth. Therefore, investing in the development of human capital through education can benefit individuals and society as a whole (Becker, 1964).

According to the HCM, individuals make rational decisions about investing in their human capital based on a cost-benefit analysis. The costs associated with obtaining education, such as tuition fees and opportunity costs, are then weighed against the potential future benefits, such as improved labor market outcomes and higher earnings. An individual chooses to invest in their human capital when the potential benefits are greater than the costs. This concept was further developed by Ben-Porath (1967) who argued that individuals choose to invest in their human capital as long as the marginal cost is smaller than the marginal benefit.

Both Becker (1962) and Ben-Porath (1967) provide a rational explanation why some individuals seemingly underinvest in their human capital. The authors argue that individuals differ in factors which determine the costs and benefits of investing in human capital. In that case, individuals with higher marginal benefits or lower marginal costs are expected to invest more in their human capital. For example, individuals with more economic stability could face lower marginal costs from attaining more education. Given that individuals make educational choices under different conditions with varied rates of return, numerous studies explore heterogeneous factors which influence human capital investment choices. Further elaboration on this topic is outlined in Section 3.2.2.

3.2 Returns to Higher Education

Drawing from Becker's (1962) HCM, Mincer (1974) pioneered an empirical strategy to estimating monetary returns to education. Following his model, a number of papers have attempted to estimate the monetary returns to education (Card, 2001). Mincer's original specification (Equation 1) posited a simple relationship between schooling (S) and consequent earnings (Y_s).

$$ln(Y_{si}) = ln(Y_0) + r * S_i \tag{1}$$

In the model, all individuals i have the same original earning capacity of Y_0. At an average rate of return of r, individuals can attain more years of schooling S_i which enhances their human capital. As a result, individuals can increase their earnings to Y_{si}. While the simple model does not account for post-schooling investments such as work experience, other models outlined in Mincer (1974) address this relationship. Broadly, equations that estimate the relationship between schooling and earnings are commonly referred to as Mincerian equations.

Alternatively, studies have argued that the monetary returns to education are not attributed to the improvement of human capital. Derived from Akerlof's signaling theory (Akerlof, 1978), Spence (1978) argued that individuals use education to signal their innate ability level to the labor market. The author suggests that the primary function of education is to serve as a credential, presenting qualifications to potential employers, rather than a human capital investment. Earlier empirical studies on signalling theory find limited evidence supporting the proposed mechanism (Altonji & Pierret, 1998). However, later studies such as Castagnetti, Chelli, Rosti, et al. (2005) find supporting results. While the signalling theory and the HCM support different mechanisms through which education increases earnings, both provide evidence of a positive causal effect of education on earnings (Weiss, 1995).

Another factor in exploring the relationship between education and earnings focuses on whether the returns display linearity or non-linearity. Linear returns in education would suggest that each additional year of education results in the same monetary returns. In terms of economic returns, this would imply that all levels of education are equally important. However, earlier papers such as Becker (1964) suggest decreasing marginal returns to education. In the case of decreasing marginal returns, greater economic benefits could be gained from investing in basic education. Park (1994) finds non linearity in the returns to education, particularly around the level of higher education. At the level of higher education, the results present a sudden increase in returns. The author suggests that the sudden increase in returns are due to the signalling power of higher education. Despite the common assumption of linear returns to education, studies have shown the importance of differentiating returns at different levels of education. Section 3.2.1 presents general and higher education specific estimates to provide an overview of the existing literature.

3.2.1 Estimation Methods

Empirical studies on the returns to higher education have explored a range of estimation techniques. A common method for estimating the returns to higher education in previous literature is the Ordinary Least-Squares approach (OLS). An overview of numerous international studies using OLS are presented in Psacharopoulos (1985). On average, an additional year of schooling increases earnings by 5-15% according to OLS estimates. The identifying assumption of OLS is that the choice to attain more education is randomly assigned. In that case, the causal effect of education on earnings could be estimated.

However, estimating returns to education using OLS raises endogeneity concerns. Previous academic literature has prominently focused on the omitted variable bias (OVB) derived from ability levels. Assuming that high-ability individuals attain more education and have higher earnings capacity regardless of education level, this would lead to an upward bias in the estimate(Griliches, 1977). However, Griliches (1977) and Becker (1964) argue that critics overestimate the ability bias.

Nevertheless, studies have employed various methods to avoid potential OVB. The use of instrumental variables (IV) stands out as a particularly convincing approach. In their seminal paper, Angrist and Krueger (1991) used a change to the compulsory schooling law in the US as an IV to estimate returns to education. The study focused on early years of education and concludes that an additional year of schooling increases earnings by 7-10%. Also applying an IV method, Card (1993) estimates the returns to higher education by exploiting the exogenous variation of college proximity. In the study, a sample of 5,525 men, ages 14-24, living in the United States were observed between 1966-1981. Individuals living closer to a college are assumed to be more likely to attend higher education. It is reasoned that proximity to a college allows students to live with their parents, potentially reducing relocation expenses. The study presented three main findings. First, college proximity had a strong, statistically significant effect on schooling.

As a result, the paper set the precedence of using college proximity as an IV. Second, according to the IV estimate, attaining a college degree increased earnings by 10-14%. Thirdly, the IV estimate obtained was 25-60 % above the OLS estimate.

Assuming that the IV estimate is closer to the true returns to education, Card (1993) highlights that the result contradicts the theoretical framework that OLS estimates are overestimated due to ability OVB. The author provides two main explanations. First, as individuals self-declare the level of schooling they obtained, responses may have suffered from measurement errors. Second, the author notes that proximity had the greatest effect on men from low-income families with the lowest propensities to continue studying. Assuming decreasing marginal returns to education (Becker, 1964), compliers would thus lead to a greater positive estimate.

The findings from Card (1993) set the precedence for further research into the returns to education using IVs. However, the author outlines a number of limitations to his study. In the context of this study, the most relevant criticism are the “imperfect indicators of residence”. Simply put, Card (1993) did not use the location of birth to calculate an individual’s proximity to college. Instead, the author used the location of the individual when the first interviews were held in 1966. This may introduce a bias to the estimate as individuals who wanted to attend college could have moved closer to the institution later in life. The possible reverse causality between proximity to university and university attendance would weaken the IV’s independence assumption. In that case, the overestimation of the first-stage estimates could bias the IV estimates.

3.2.2 Heterogeneous Returns

The earliest empirical works on returns to education found heterogeneous returns across gender, ethnicity and socioeconomic background (Card, 1999). Recent empirical studies support these findings (Oreopoulos & Petronijevic, 2013).

Studies found that females often experience a greater rate of return from higher education relative to males. Notably, this does not imply that educated females earn more than their male counterparts. Numerous studies find that females earn lower salaries compared to males (Weichselbaumer & Winter-Ebmer, 2005). Card, Cardoso, and Kline (2016) suggests that females with higher education degrees possess greater bargaining power. As a result, they can bargain for higher wages that match those of their male counterparts. Furthermore, females are less likely to attain higher education compared to males (Filmer, 2000). This may result in a scarcity of qualified female workers in the labor market which enhances their bargaining power. Psacharopoulos and Patrinos (2018) review 1,120 empirical studies on the returns to education across 139 countries between 1950-2014. The authors find that the rate of return from an additional year of education for females is on average 2% higher than for males. Therefore, empirical studies commonly control for gender to account for this heterogeneity.

Heterogeneous returns to higher education may also stem from employer-based discrimination (Becker, 2010). This includes gender-based and ethnicity-based discrimination. Individuals facing discrimination may have limited opportunities, which reduce their potential returns on their educational investment (Neumark, 2018). Thus, acquiring higher education credentials could enhance their job prospects. As a result, these individuals could increase their chances at receiving equitable returns. Nevertheless, despite acquiring higher education credentials, individuals facing discrimination are prone to experience lower returns. Ethnicity-based differences in returns may also stem from racial minorities often belonging to lower socioeconomic brackets. This results in additional constraints to racial minorities. Section 3.3 provides empirical evidence of ethnicity-based heterogeneous returns to education.

Individuals with disadvantaged socioeconomic background tend to achieve low education outcomes. This could be due to their limited access to educational institutions and financial resources. Those least likely to continue their studies are often referred to as marginal students. Card (1999) suggests that marginal students have higher marginal returns to education. On one hand, high marginal returns imply that marginal students should invest in more education. On the other hand, these individuals often face other limitations and opportunity costs. Consequentially, pursuing higher education may not yield sufficient returns for them to investment.

The purpose of this subsection was to elaborate on three predominant factors that impact rates of return. A comprehensive analysis into drivers of heterogeneous returns to higher education goes beyond the scope of this study. However, acknowledging a few drivers of heterogeneity allows for a deeper understanding of individual educational choices, group differences in earnings and provides a basis for control variables.

3.3 Returns to Higher Education in Brazil

The earliest empirical studies on the returns to education in Brazil were Castro (1970) and Langoni (1974). According to Barbosa Filho and Pessoa (2008), both indicated that marginal costs were smaller than marginal benefits using data from the 1960s. This results indicated that it was profitable for Brazil to invest in education. Due to access restrictions to the papers, the exact estimates and methods applied were not available.

On behalf of The World Bank, Jallade (1977) estimated the effects of basic education on income using an OLS model. Using 1972 Household Survey and 1970 Population Census data from the Brazilian Institute of Geography and Statistics (*Instituto Brasileiro de Geografia e Estatistica (IBGE)*, the study found that an additional year of primary schooling increased earnings by 13.9%-23.5% and secondary schooling increased earnings by 10%-13.1%. These findings support the theory of decreasing marginal returns to education. Furthermore, the study found that rates of return varied across socioeconomic group due to unequal access to education. Given the focus of the study on basic education, these estimate do not reflect returns to higher education. Nevertheless, the results provided early evidence of heterogeneous returns to education and similar estimates to international papers.

Griffin and Edwards (1993) estimated returns using OLS and 1989 IBGE National Household Sample Survey (*Pesquisa Nacional por Amostra de Domicílios (PNAD)*) data. The study found that on average, an additional year of education increased earnings by 12.8%-15.1%. However, the author also found non-linear returns to education. According to the results, a person with 16 years of education or more, had higher returns to an additional year of education relative to their less educated counterparts. This finding suggest that higher education has a greater positive effect on earnings. Similarly, Tannen (1991) estimated returns to education using OLS based on 1980 IBGE Census microdata. The study provided estimates for the South-East of the country, which includes the state of São Paulo. According to the results, completing high school increased earnings by 14.1%-14.6%, while completing college increased earnings by 18.8%-20.7%. Evidently, the results contradict the theory of diminishing marginal returns. However, the identifying strategy may suffer from OVB, given that the treatment is not randomly assigned.

Following earlier studies on the returns to education in Brazil, interest in the topic has persevered. Sachsida, Loureiro, and Mendonça (2004) use OLS to estimate the returns to education for men ages 24-56, using PNAD 1992-1999 data. The results suggest that an additional year of education increases earnings by and average 14%. Stefani et al. (2006) and Stefani and Biderman (2009) use PNAD data from different years to estimate the returns to education in Brazil using an IV approach. The studies use PNAD 1988 and PNAD 1996 data, respectively. Both studies apply an IV Quantile Regression technique which uses an individual's initial endowments as the instrument. Despite the use of different data sets, the studies provide similar results. For clarity, the estimates derived from both papers are presented in Table 1.

Table 1: Monetary returns from an additional year of education estimated by Stefani et al. (2006) and Stefani and Biderman (2009)

	PNAD 1988	PNAD 1996
White Females	23%	15.2%
White Males	19%	14.1%
Black Females	14.3%	6.3%
Black Males	11.4%	10.5%

Note: The table presents the percentage increase in earnings from attaining an additional year of education. Estimates are calculated for four groups based on gender and ethnicity. PNAD 1988 and 1996 data are derived from Stefani et al. (2006) and Stefani and Biderman (2009), respectively.

As presented in Table 1, estimates from both studies show that the rate of return from an additional year of education are greatest for white females followed by white males. Compared to their white counterparts, black individuals presented lower returns. Based on PNAD 1988 data, the estimates for black females were greater than the estimates for black males. The inverse is observed for estimates derived from PNAD 1996 data. Evidently, both studies provide empirical evidence of heterogeneous returns to education across ethnicity and gender.

Teixeira and Menezes-Filho (2012) also use the IV approach to estimate returns to education. Their study exploits the systemic change introduced by Law 5692 in 1971 which enforced the compulsory 8 years of primary schooling. Using 1997-2007 PNAD data, the authors find that the returns of an additional year of education are 5.5% for the IV estimate 11.6% for the OLS estimate. These results coincide with the ability OVB theory. Similar to Jallade (1977), the results are based on primary schooling and may not reflect returns to higher education. However, the use of IV to estimate returns to education proves to be growing in popularity in Brazil.

A number of key aspects stood out when reviewing literature related to the returns to higher education in Brazil (*see Table 2*). First, previous studies predominantly use an OLS approach. Although the IV method is favored, its application within the context of Brazil remains relatively limited. Second, the majority of papers investigate the general effects of an additional year or for primary or secondary education. In comparison, fewer studies focus on the returns to higher education. Lastly, despite the dynamic history of the Brazilian education system, studies have not exploited the potential insights derived from the changes in the number of HEIs.

Table 2: Overview of estimates from the literature review on returns to education

Paper	Method	Scope	Returns from	Estimates
Psacharopoulos (1985)*	OLS	All levels of education	Additional year	5-15%
Angrist and Krueger (1991)	IV	Early years of education	Additional year	7-10%
Card (1993)	OLS	Higher education	Graduating college	7.3%
	IV	Higher education	Graduating college	10-14%
Studies focusing on Brazil				
Jallade (1977)	OLS	(A) Primary schooling	Additional year	(A) 13.9-23.5%
	OLS	(B) Secondary schooling	Additional year	(B) 10-13.1%
Griffin and Edwards (1993)	OLS	All levels of education	Additional year	12.8-18.1%
Tannen (1991)	OLS	(C) High school	Completing HS	(C) 14.1-14.6%
	OLS	(D) Higher education	Completing HE	(D)18.8-20.7%
Sachsida et al. (2004)	OLS	All levels of education	Additional year	14%
Stefani et al. (2006)	IV	All levels	Additional year	*See Table 1*
Stefani and Biderman (2009)	IV	All levels	Additional year	*See Table 1*
Teixeira and Menezes-Filho (2012)	OLS	Primary schooling	Additional year	11.6%
	IV	Primary schooling	Additional year	5.5%

Note: This table presents the estimates from the studies discussed in the main text. OLS: Ordinary Least Squares Approach. IV: Instrumental Variables Approach. *This study is a collection of various international studies with different scopes.

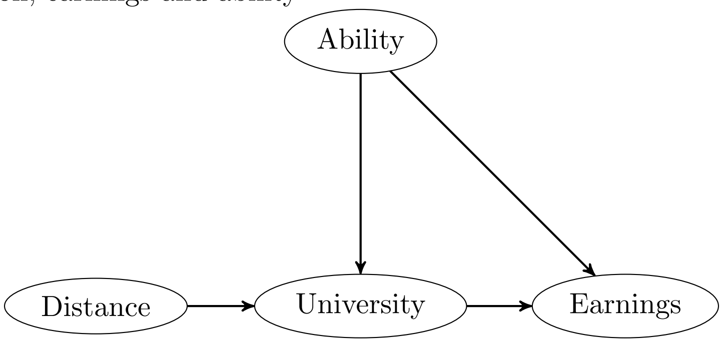
This study aims to contribute to the existing literature on the returns to higher education in Brazil. By applying an IV approach to estimate the causal effect of higher education on returns, this paper builds on the methodology's growing popularity. From the best of my knowledge, no previous studies in the context of Brazil have followed the IV approach introduced by Card (1993); estimating returns to higher education using proximity to university as an instrument.

Therefore, this research aims to provide new estimates exploiting the geographic variation of HEIs in the country. Moreover, the data set compiled for this analysis may serve as a tool for future research.

4 Theoretical Framework

Derived from previous literature, the following section outlines the theoretical framework of this study, as illustrated in Figure 5. Furthermore, this section establishes the main hypotheses.

Figure 5: Directed Acyclic Graph illustrating the relationship between distance to university, university graduation, earnings and ability



Source: Created by the author, based on Card (1999)

Using Mincerian equations, this study aims to empirically estimate the returns to higher education using Brazilian data. In other words, the focal estimation pertains to the causal effect of university education attainment on subsequent earnings. To provide a basis for the estimate, an OLS identification strategy will be applied. However, as outlined by previous literature, confounders such as ability are likely to lead to OVB. This relationship is illustrated in Figure 5. In order to avoid potential OVB, this study will apply an IV identification strategy, using proximity to university as an instrument. For the IV estimates to be reliable and valid, four fundamental assumptions need to hold: relevance, independence, monotonicity and exclusion.

The relevance assumption holds in the case that the first-stage of the IV strategy is meaningful. In this case, this would require that the proximity to university affects higher education attainment. Given that this study follows the same identification strategy as Card (1993), the arguments raised in the original study provide a clear explanation of the relationship. Furthermore, the study finds empirical evidence of this relationship which strengthens the relevance assumption. Intuitively, living closer to a HEIs stimulates higher education attainment as the costs associated with attendance, such as travel, are smaller. Proximity to university can also inspire individuals through exposure to role models and potential career paths, thereby motivating them to pursue higher education.

For the independence assumption to hold, an individual's proximity to a HEI has to be as good as random. Two main arguments suggest this to be true. First, people cannot choose where they are born. Therefore, their proximity to a HEI is essentially random. Second, the exact opening dates of university units are usually unknown to the population. While campuses may open near densely populated areas to meet demand, the precise timing of these openings is rarely precisely predetermined. Therefore, the opening of a campus which reduces the distance to a HEI is as good as random. In the context of this study, another argument can be made. This study exploits the sudden reduction in distance to HEIs in the state of SP. To the best of my knowledge, no legislation is tied to the abrupt change. Consequently, there is no reason to believe that people relocated prior to this development. These factors support the randomness of proximity to HEIs and justifies its selection as the instrumental variable for this study.

The third assumption relates to monotonicity. This require that people living closer to a university to always be more likely to attend HEIs. In other words, the IV approach requires that there are no defiers of the instrument and only compliers. The earliest work of Becker (1962) supports this assumption. People who live closer to HEIs face fewer financial and geographic restrictions to higher education attainment compared to those living further away. Ceteris paribus, the cost-benefit analysis for individuals in the proximity of a HEI has lower costs compared to their distanced counterpart. Therefore, the monotonicity assumption in this model is suggested to hold.

Lastly, the IV approach requires the exclusion restriction to hold. This would suggest that proximity to university affects subsequent earnings solely through its effect on higher education attainment. Economic intuition supports this assumption as simply living near a HEI does not impact someone's future earnings in the labor market. For example, moving 10km closer to a HEI does not directly increase an individual's earning potential. The inverse holds true. Intuitively, an individual moving 10km away from a HEI would not directly change their earnings.

The theoretical framework used in this study suggests that the IV method is appropriate for the estimation of returns to higher education. As presented in the literature review, an extensive number of studies find that education has a positive causal effect on earnings. Studies focused on the effects of higher education coincide with these results. Furthermore, this research bears the closest resemblance to the study conducted by Card (1993). In the earlier paper, it is observed that the IV estimates are greater than the OLS estimates when determining the returns to higher education. Therefore, the following hypotheses are tested in this paper:

Hypothesis 1 (H1): *Higher education attainment has a positive causal effect on earnings*

Hypothesis 2 (H2): *Higher education attainment has a greater positive effect on earnings when calculated using and IV approach, compared to an OLS approach*

5 Data

For the purpose of this study, a new data set was constructed using secondary data. This section presents the main data sources used and outlines the data collection process. Next, the intuition behind the sample selection is provided. To offer a preliminary understanding of the composition of the data set, descriptive statistics are included at the end of this section.

5.1 Data sources and collection

This paper combines various data sources for the empirical analysis using the statistical software Stata/MP 17 (64-bit), henceforth also referred to as Stata. The main data acquisition process was divided into three focuses: individual data, distances between municipalities and campuses and the opening dates of campuses.

Collected by the IBGE, the 2000 Population Census provides the individual level observations used in this study. Purchased from the IBGE web shop, the observations were originally in Text Data format. Next, a dictionary file in SAS format provided by the IBGE was manually converted into a Stata dictionary. This dictionary allowed for the 2000 Population Census to be opened on Stata. The census registered a total of 4,038,219 observations in July in 2000. Individual characteristics such as age, gender and ethnicity were reported. However, the survey did not collect the municipality of birth of the individual. Instead, the municipality where the individual was living at the time of the survey was reported, in addition to a dummy variable that indicated if they were living in the same municipality where they were born. Based on these two variables, the municipality of birth for a specific groups could be identified. Specifically, individuals who have not moved from their municipality of birth. Henceforth referred to as non-movers. Furthermore, the census did not report the exact address of an individual. Therefore, we use the geographic center of their municipality as a proxy.

The main data source used to determine the distance between municipalities and universities was the national reference center for the state of São Paulo called The Statewide System for Data Analysis Foundation (Fundação Sistema Estadual de Análise de Dados), commonly referred to as Fundação SEADE or SEADE. SEADE's database compiles various demographic data sets related to the regions and municipalities in the state of São Paulo. The coordinates for the center of each municipality was provided in a single data set. However, the coordinates for the university units were provided by a number of SEADE data sets from 2020 and 2021. This study included 9 public universities and a total of 170 campuses across the state. For the names and abbreviations of the aforementioned universities, see Table 10 in the Appendix. To combine the data sets, official IBGE municipal codes were consulted. Based on the municipal coordinates and the campus coordinates, the instrumental variable for the empirical analysis was calculated; proximity to university.

In addition to the individualized data and the distances between municipalities and universities, the opening dates of the campuses were required to accurately determine the distance faced by individuals within the sample. Most universities (as listed in Table 10, Appendix) were contacted via email and telephone to obtain the most accurate opening years for each unit. Following my request, only São Paulo State Technological Colleges (FATEC) and Federal Institute of São Paulo (IFSP) provided the opening dates as registered internally. Due to the smaller number of units, Instituto Tecnológico de Aeronáutica (ITA) and Federal University of ABC (UFABC) data were easily available on their main website. The remaining universities did not respond. Therefore, the opening dates published on the website for each campus were consulted. The websites often listed three dates: when it was declared that the unit would exist, when administrative work began and when the academic year began. Intuitively, students could only enroll to a new campus once the administrative process was in place. Therefore, this study uses the start of administrative work as the reference date to determine when a unit became active. However, not all units reported the starting date of administrative work. In those cases, the start of the academic year is used as an estimate.

5.2 Sample selection

The IBGE 2000 Population Census provides a sample of 4,038,219 individuals living in the state of São Paulo. The population of interest in this study are individuals who completed an undergraduate degree. Therefore, a sample of the dataset is selected to target this population. Individuals below 18 years of age are typically still in high school, haven't attended university and have a low/no income. Hence, including this demographic in the sample would introduce an upward bias as the effect of university education on wages would be overestimated. Consequently, 1,319,103 individuals below the age of 18 were dropped from the sample.

On average, the length of a bachelor's degree in Brazil takes up to 5 years to complete. Assuming a starting age of 18, the student would graduate at the age of 23. Therefore, the lower bound of the sample age is 23. Given that the working population in Brazil includes individuals ages 15-64 (OECD, 2019), an upper bound can be set at the age of 64. Individuals between the ages of $23 - 64$ could be categorized into three groups: young $(23 - 30)$, prime $(31 - 50)$ and mature $(51 - 64)$. The distribution of individuals within these three categories is illustrated in Figure 7 in the Appendix. Evidently, the sample size per age category decreases by age. In other words, there are a greater number of young people relative to the people in their prime or mature individuals.

The target demographic for studies on the returns to higher education have varied in age range. In the original study focused on returns to higher education using the IV approach, Card (1993) observed the later life outcomes of men between the ages of 29-39. Mincer (1974) suggests that individual returns to education are best measured eight years post-schooling. Given the lower bound of 23, this would indicate that 31-year-old individuals are the best age group to observe. Nevertheless, the wide age range of individuals in this study offers comprehensive insights into returns to education.

Furthermore, the study is limited to individuals who were born within the state of SP. Intuitively, people born within the state are most affected by the changes in HEIs. Given the empirical strategy of this study, the following section outlines the sample selection for the individual level and cohort level analyses.

The first section of this study exploits an individual's proximity to a HEI. As seen in the Figure 7 in the Appendix, the largest proportion of the sample are individuals in their prime. Moreover, individuals in the prime age group seem to have the most consistent earnings (see Figure 8, Appendix). As young individuals are at the start of their career, their earnings present a steady increase. Furthermore, mature individuals who are at the end of their careers may entire retirement which decreases their earnings. Given the sample size and the earnings stability of individuals in their prime, the individual level analyses focuses on this group.

Additionally, this study focused on creating a representative sample of the Brazilian population. As a result, individuals from the megacity of SP were excluded. The city of São Paulo, is the largest and most densely populated urban center in Brazil. The city exhibits extreme economic disparities and socio-economic trends that are not representative of the broader Brazilian population. Notably, university campuses are densely located in the city. Therefore, individuals from the city of SP are excluded from the sample. Furthermore, Table 3 shows that out of 645 cities, the population from the city of SP account for 26% of the sample. This suggests a disproportionate representation of individuals from the megacity in the sample and further justifies their exclusion.

The second part of this study exploits changes to the proximity of HEIs at a cohort level. Specifically, the sudden decrease in average distance around 1955, illustrated in Figure 4. Given the year of the change, the sample used for the cohort level analysis includes mature individuals.

5.3 Data description

In order to gain insights into the individuals observed in the data set, the table below presents descriptive statistics of our sample. Table 3 provides an overview of the entire sample and Table 11 (in the Appendix) takes a closer look at the characteristics of the three age categories: young, prime and mature. The purpose of describing the data set at both levels is to better understand how limiting the empirical analysis to one age category could impact estimates.

Table 3: Descriptive statistics for adult sample (above 18 years old) in 2000

Variable	Description	Mean	SD	Min	Max	N
Gender	=1 if male, =0 if female	.4820	.4997	0	1	2,672,772
Age	Age registered in complete years	39.24	15.76	18	100	2,672,772
Racial Minority	=0 if white, 1 otherwise	.2831	.4505	0	1	2,656,632
University	=1 if completed university, 0 otherwise	.0727	.2596	0	1	2,672,772
Years of Education	Years of education across all levels	7.11	4.41	0	20	2,672,772
Literacy	=1 if able to read & write, 0 otherwise	.9285	.2576	0	1	2,672,772
Employed	=1 if formally employed, 0 otherwise	.5542	.4971	0	1	2,672,772
Income	Per month, Registered in Brazilian Reais	830.74	2364.76	1	900000	1,516,397
Hours worked	Number of formal hours worked	45.23	13.64	1	140	1,540,156
Retired	=1 if retired, 0 otherwise	.1228	.3282	0	1	2,672,772
Distance to University	Distance from municipality in Km	31.22	42.79	0.30	597.08	2,672,772
From SP City	=1 if from SP city, 0 otherwise	.2642	.4409	0	1	2,672,772

Source: Based on 2000 Brazilian Institute of Geography and Statistics (IBGE) Census Data, created by the author

Furthermore, descriptive statistics by age category can provide a baseline reference to interpret the impact of findings. The most relevant insights from the descriptive statistics relate to education, income and proximity to a HEI. According to the average in 2000, individuals in their prime had a greater literacy rate compared to the mature group, however, lower rates compared to the young demographic. This indicates a natural development of basic education in the state. Notably, individuals in their prime and at a mature phase are a main focus of this study. A larger portion of the group in their prime (10.2%) had completed a higher education degree compared to the mature group (7.4%). On average, individuals in their prime attained around 2 more years of education relative to mature individuals. These averages suggests a trend of improved higher education attainment across generations.

Moreover, the average income of groups differed. The prime group reported an average income of 985 Brazilian Reais (Reais) per month, while the mature individuals reported an average of 1095 Reais. To contextualize, the minimum wage in 2000 in Brazil was 266 Reais (Saboia, 2007). This highlights the extreme income disparity in the country given that the standard deviations for the prime and the mature group are 2,398 Reais and 1,095 Reais, respectively. Descriptive statistics further support the relevance of the instrument used in this study. The older the group, the greater the average distance to a HEI when they were 18. For the young, prime and mature group, this coincided with 23km, 25km and 36km, respectively.

6 Identification Strategy

The aim of this study is to explore the causal effect of higher education attainment on earnings. For that purpose, this section outlines the identification strategies applied.

6.1 Individual Level

As explained in the previous section, the individual level analysis focuses on non-movers.

6.1.1 Ordinary Least Squares Approach

To begin, a simple OLS estimation (Equation 2) is conducted to obtain initial insights and most importantly, to set a baseline for the study.

$$Log(Income_{imcu}) = \beta_{2,0} + \beta_{2,1} * Uni_i + \alpha_2 * X_i + \pi_{2,m} + \lambda_{2,c} + \upsilon_{2,u} + \epsilon_{2,imcu} \quad (2)$$

In Equation 2, for individuals i in municipality m, cohort c, attending a university u, earnings (*Log(Income)*) are regressed on graduating university (*Uni*). *Log(Income)* is a continuous variable measured in Brazilian Reais. *Uni* is a binary variable that takes the value of 1 if the individual completed a higher education degree and 0 otherwise. X denotes a matrix of individual controls consisting of characteristics such as gender and ethnicity. Furthermore, municipal, cohort and university fixed effects are denoted as π, λ and υ, respectively. The notation includes the constant, $\beta_{2,0}$, and the error term ϵ. The identifying assumption of this approach is that individuals who graduated university are identical to those who did not. In other words, it should be expected that graduates and non-graduates would have the same earnings if both did not attend higher education. However, previous studies argue against this assumption. As aforementioned, confounders such as ability are expected to influence both the treatment (graduating university) and outcome (earnings). This endogeneity concern prompts the need for the IV approach.

6.1.2 Instrumental Variable Approach

The IV approach leverages variations in the proximity to HEIs for individuals i. Given the development of HEIs (see Figure 3), individuals from the same municipality could face varied distances at different points in time. In addition to the outcome variable (*Log(Income)*) and treatment variable (*Uni*) introduced in the OLS approach, *Log(Distance)* denotes the continuous instrument. The distance to a HEI is measured in kilometers. Furthermore, the same controls and similar fixed effects are included. The following notations present the first-stage and reduced form of the IV approach, respectively:

$$Uni_{imc} = \gamma_0 + \gamma_1 * Log(Distance_{mc}) + \alpha_3 * X_i + \lambda_{3,c} + \upsilon_{3,u} + \varepsilon_{imc} \quad (3)$$

$$Log(Income_{imcu}) = \theta_0 + \theta_1 * Log(Distance_{mc}) + \alpha_4 * X_i + \lambda_{4,c} + \upsilon_{4,u} + \sigma_{imcu} \quad (4)$$

The causal estimates are then derived using the following equation:

$$Log(Income_{imcu}) = \beta_{5,0} + \beta_{5,1} * \hat{Uni}_{mc} + \alpha_5 * X_i + \lambda_{5,c} + \upsilon_{5,u} + e_{imcu} \tag{5}$$

Notably, the IV estimate derived from Equation 5 is the ratio between the reduced form estimate and the first-stage estimate: $\beta_1 = \theta/\gamma$. Moreover, it is important to highlight that this approach estimates the Local Average Treatment Effect (LATE). This means that the causal effect of university graduation on earnings is estimated for individuals who were affected by the instrument (proximity to university).

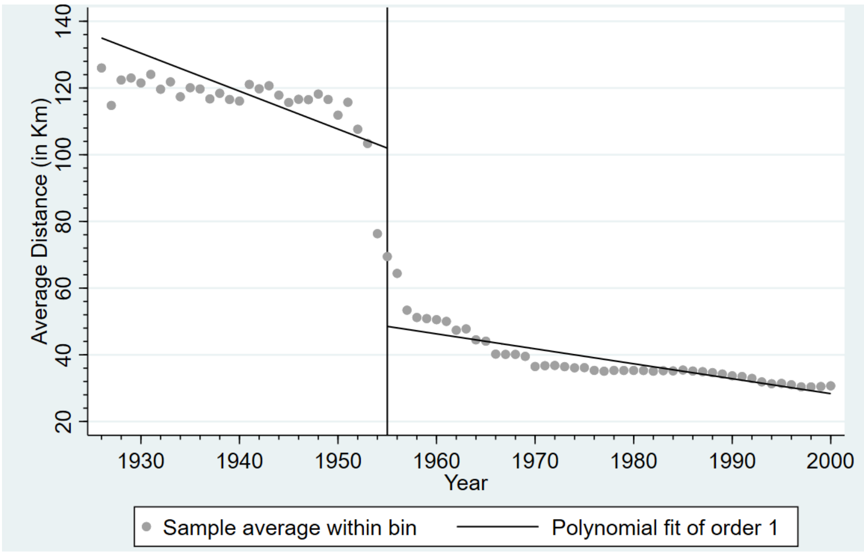
6.2 Samples Selection Bias Concern

A main concern in the individual level analysis is sample selection bias. Given the data limitations, only non-movers are included in the sample. Intuitively, attaining a university degree is likely to influence an individual's propensity to reallocate. For example, due to an improvement in labor market opportunities in a different city. Given that the correlation between graduating university and moving is likely to bias the individual level results, this study explores this concern. Moreover, to address this concern, an alternative identification strategy is implemented.

6.3 Cohort Level

Examining the development of HEIs in the state of SP revealed a sudden shift in the average distance to a university campus. As aforementioned, this change is linked to the opening new locations around 1955. Depicted in Figure 6, this study leverages the sudden change around 1955 by applying a Sharp Regression Discontinuity Design (RDD). First introduced by Thistlethwaite and Campbell (1960), the RDD approach allows for a treatment allocation which is as good as random.

Figure 6: Discontinuity of average distance to a university unit in the state of São Paulo



Note: For calculating the average distance to a university campus within the state of São Paulo, the data excluded observations in the city of São Paulo due to a potential over-representations.

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

6.3.1 Regression Discontinuity Design Approach

Similar to the notation provided for the individual level IV approach, the RDD estimates the effect of graduating university (*Uni*) on earnings (*Log(Income)*), using proximity to HEI to assign treatment status. However, in this case, the treatment allocation depends on the threshold at which the average distance decreased, 1955. The control group consists of individuals who were 18 before the sudden change in average distance. Individuals in the cohort after the shift are in the treated group. The following equations coincide with the first-stage and reduced-form of the RDD:

$$Uni_{ct} = \phi_0 + \phi_1 * Decrease_c + \tau_t + \zeta_{ct} \quad (6)$$

$$Log(Income_{imctu}) = \pi_0 + \pi_1 * Decrease_c + \alpha_7 * X_i + \tau_{7,t} + \upsilon_{7,u} + \rho_{imctu} \quad (7)$$

Causal inference is then derived from the following equation:

$$Log(Income_{imctu}) = \beta_{8,0} + \beta_{8,1} * \hat{Uni}_{ct} + \alpha_8 * X_i + \tau_{8,t} + \upsilon_{8,u} + \mu_{imctu} \quad (8)$$

Furthermore, the RDD approach controls for the same individuals controls used in the individual analysis: gender and ethnicity. Additionally, year (τ) and university (υ) fixed effects are included. The results of the RDD approach also provide LATE estimates. Essentially, this method estimates the causal effect of university graduation on earnings for individuals for whom the sudden decrease in average distance affected university graduation.

To begin the cohort level analysis, the optimal bandwidth of the RDD framework around the cutoff point was determined based on the minimized mean square error (MSE). The optimal bandwidth was identified as 1945 - 1965. Interestingly, this time frame coincides with "The Second Brazilian Republic". Figure 10 in the Appendix shows that the threshold at 1955 is highlighted when the bandwidth is restricted to 1945 - 1965.

After determining the optimal bandwidth for the RDD, a visual inspection of relevant graphs need to occur to determine if the approach is applicable. In other words, a discontinuity at the threshold needs to be present for the trends of university graduation and earnings. Figure 12 and 13 in the Appendix illustrate the trends for university graduation and earnings, respectively. A clear discontinuity for the relevant variables are not visible when analyzing the complete samples. However, when the cohorts are divided into gender and ethnic categories, discontinuities are found. Given the limited number of observations for Indigenous and Asian individuals, inferences cannot be drawn from the graphs. A clear jump in university graduation rates at the threshold are visible for individuals who identify as male or mixed-race, known in Brazil as pardos. Notably, no clear discontinuity is visible for the white population. However, given the large number of white individuals in the sample, it is interesting to investigate this demographic. Moreover, these three groups exhibit a clear decrease in earnings at the threshold. Thus, it suggests that the effects of the sudden change had a greater impact on these demographic groups. This explains the focus on these three demographic groups for the cohort level analysis.

7 Results

In this section, the empirical findings derived from the identification strategies are presented.

7.1 Individual Level

As aforementioned, the individual level analysis consists of two approaches: OLS and IV. Table 4 presents the OLS estimates for the effect of graduating university on income. This approach coincides with Equation 2. A preliminary Breusch-Pagan and Cook Weisberg test indicates heteroskedasticity and prompts the use of robust standard errors. Columns 1 and 2 controls for individual characteristics. This is not the case for Columns 3 and 4. Furthermore, only estimates presented in Columns 1 and 3 account for the fixed effects presented in Equation 2.

Table 4: The effect of university graduation on income, Ordinary Least Squares estimates

Dependent Variable: Log(Income)	(1) OLS	(2) OLS	(3) OLS	(4) OLS
University	1.063***	1.158***	1.031***	1.128***
	(0.016)	(0.016)	(0.017)	(0.016)
Male	0.608***	0.594***	-	-
	(0.012)	(0.012)		
Minority	-0.304***	-0.311***	-	-
	(0.015)	(0.015)		
Fixed Effects	Yes	No	Yes	No
Observations	18,742	18,742	18,836	18,836
R-squared	0.382	0.290	0.282	0.189

Note: Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

The OLS estimates across all four columns show that completing university increases income by more than 1 percent. Notably, the estimates do not greatly vary despite differences in controls and fixed effects. Furthermore, all the estimates are statistically significant at a 0.01 significance level and have similar robust standard errors. These results support Hypothesis 1; attaining a university degree increases consequent earnings. However, previous literature suggest that these OLS estimates suffer a bias.

Tables 12 (*in Appendix*) and 5 present the first-stage and second-stage estimates of the individual level IV identification strategy, respectively. Across both tables, Column 1 estimates account for all individual controls and fixed effects outlined in Equation 3, while Columns 2-5 present various combination of controls and fixed effects to assess estimate robustness. The first-stage estimates presented in Tables 12 indicate that a decrease in the distance to a university campus by 1% would increase an individuals probability of graduating higher education by 3.3-3.5%. This coincides with the belief that individuals are more likely to attain higher education when they live near a HEI. These estimates are statistically significant at the 0.01 level and the F-value suggest that the instrument is not weak.

Following the first-stage results, Table 5 presents the IV estimates. Notably, estimates displayed in Columns 1-5 of both tables coincide with each other. Column 1 includes individual controls and fixed effects, while Columns 2-5 accounts for a combination of these variables. The statistically significant findings suggest that attaining a university degree increases income by 4.3%-5.0%. These results support Hypothesis 1 that university increases consequent earnings.

Furthermore, Hypothesis 2 is also supported given that the IV estimates are greater than the OLS estimates. Nevertheless, an investigation into the potential sample selection bias is necessary.

Table 5: The effect of university graduation on income, Instrumental Variables (2SLS) estimates

Outcome Variable:	(1)	(2)	(3)	(4)	(5)
Log(Income)	2SLS	2SLS	2SLS	2SLS	2SLS
University	4.551***	4.523***	4.348***	4.320***	5.034***
	(0.332)	(0.321)	(0.323)	(0.312)	(0.343)
Male	0.836***	0.838***	-	-	0.866***
	(0.033)	(0.033)			(0.035)
Ethnicity	Yes	No	Yes	No	Yes
White	-0.112	-	-0.135	-	-0.191
	(0.142)		(0.134)	-	(0.153)
Black	0.018	-	0.009		0.009
	(0.147)		(0.139)	-	(0.159)
Asian	-0.628***	-	-0.620	-	-0.800***
	(0.228)		(0.336)		(0.246)
Mixed	0.071	-	0.043	-	0.069
	(0.142)		(0.133)		(0.154)
Indigenous	-0.013	-	-0.019	-	0.013
	(0.206)		(0.218)		(0.229)
Fixed Effects	Yes	Yes	Yes	Yes	No
Observations	18,836	18,836	18,836	18,836	18,836
F-value	184.75	338.55	121.28	191.59	137.54

Note: Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

7.2 Sample Selection Bias Concern

As expected, a correlation between graduating university and moving is found (*see Table 6*).

Table 6: The correlation between university graduation, moving trends and earnings

Variables	Mean	SD	N	1. University	2. Moved	3. Earnings
1. University	0.09	0.28	818,695	1.0000	-	-
2. Moved	0.05	0.22	548,900	0.0554	1.0000	-
3. Earnings	858.83	2252.58	543,526	0.2015	0.0118	1.0000

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

In this sample, the correlation between graduating university and moving is 0.06. Intuitively, this coincides with the fact that university graduates may reallocate due to an improvement in labor market opportunities. This relationship is further analyzed by estimating a person's probability of moving, given they graduate university (*see Table 13, Appendix*).

Probability models indicate that university graduates are more likely to move compared to individuals who have not completed a higher education degree. Table 13 (in the Appendix) presents this difference for the whole sample (all), by sex (male and female) and for three ethnic groups (white, black and mixed). For the whole sample, graduates were more than twice as likely (+107%) to move, relative to non-graduates. The group with the greatest difference in relocation between graduates and non-graduates were females (+111%). Mixed individuals had the least relocation difference post-graduation among the groups (+79%). Despite the variation across the groups, the findings consistently show that graduates have a different moving behaviour compared to non-graduates. Thus, excluding movers results in a sample selection bias when estimating returns to higher education. For example, if all high earning graduates move after their studies, then the sample would underestimate the returns. Moreover, graduates with innately lower abilities may be restricted to local labor market opportunities. In that case, these individuals would be over-represented in the sample. Consequently, the returns to higher education would be further underestimated. These findings support the need for the cohort level analysis to estimate returns to higher education.

7.3 Cohort Level

The results of the cohort level RDD (Equations 6 - 8) are presented in three main tables. Estimates for the first-stage, reduced form and second-stage coincide with Table 14, 15 and 7, respectively. Notably, Table 14 and 15 are in the Appendix. As presented in Table 14, the results find that the sudden decrease in average distance had a positive effect on university graduation. Across the three tables, Column 1 presents estimates which account for year fixed effects and individual controls. In that case, the sudden decrease in average distance post-1955 increased an individual's likelihood of attaining a university degree by 5.3%. Given that the effect is statistically significant and that the F-value is greater than 10, the sudden decrease in average distance is suggested to be a strong instrument.

In Table 14 (Appendix), Columns 2-6 show estimates which account for different combinations of ethnic and gender criteria. These estimates also show a statistically significant increase in university graduation after the change. However, only estimates for mixed/pardo individuals of all genders (seen in Column 2) present F-values greater than 10. One way to interpret this is through the socio-economic environment at the time; white individuals prominently tended to have greater access to educational opportunities compared to mixed people/pardos due to higher socio-economic backgrounds. Furthermore, historical inequalities and discrimation faced by mixed people/pardos would have limited their access to educational opportunities. These factors may explain why the decrease in average distance to HEIs had a greater impact on mixed/pardo individuals compared to white individuals. While the allocation costs would be reduced for everyone in the cohort, it is possible that this limitation was a greater barrier for mixed/pardo individuals.

Table 15 in the Appendix presents estimates of the sudden decrease on income. Across all columns, results show that the decrease in average distance had a positive effect on income. This estimate varies between 24.5%-40.5%. Given that the findings are statistically significant and given that exclusion holds, these results indicate that the sudden change is a good instrument.

The relevant estimates for causal inference are presented in Table 7 and suggest that there is a positive effect of university graduation on income. These results are statistically significant for all columns excluding Column 5. Given that the sudden decrease only had a strong impact on the complete sample (Columns 1) and the mixed people/pardos (Columns 2), the causal estimates of these groups are of most interest. Column 1 indicates that university graduation increases an individuals income by 6.75%. For mixed/pardo people only, university graduation increases earnings by 2.02%. The lower effect of having a university degree for mixed/pardo individuals may reflect the groups limited bargaining power in the labor market. Given the ethnic inequalities present at the time, this is a possible explanation for the lower returns.

Table 7: The effect of university graduation on income, regression discontinuity design estimates

Dependent: Log(Income)	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) 2SLS	(5) 2SLS	(6) 2SLS
University	6.749***	2.016*	5.360***	7.368***	0.919	7.467***
	(0.419)	(1.079)	(0.389)	(0.654)	(1.180)	(0.722)
F-statistic	259.01	3.49	190.25	126.97	0.61	107.01
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	No	No	No	No	No
Ethnic Group	All	Mixed	White	All	Mixed	White
Gender	All	All	All	Male	Male	Male
Observations	79,008	8,208	64,391	55,646	5,629	45,881

Note: Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

Compared to earlier estimates on the return to higher education, these findings suggest lower returns. Nevertheless, the results consistently indicate that higher education attainment has a positive effect on consequent earnings, supporting Hypothesis 1. To recall, the average income for the mature group was 1,095 Brazilian Reais per month. A 6.749% increase in average income would coincide with an increase of 73.87 Reais per month. However, descriptive statistics indicated that income greatly varied across the sample. Thus, the percentage increase in income due to higher education could have different effects. Therefore, the percentage estimate should not be interpreted as a uniform benefit for all. Instead, it supports the existence of a positive effect of higher education on income, in general.

8 Robustness Checks

8.1 Continuity Checks

The identifying assumption of the RDD requires other variables related to the outcome variable to be continuous at the cutoff. To test this assumption, it is possible to visually analyze the distribution of control variables used in the empirical approach. Figure 14 and 15 in the Appendix illustrate the distribution of gender and ethnicity, respectively. Both suggest no clear discontinuity of these variables at the threshold. Furthermore, a t-test analysis is conducted for both variables and suggest that the differences in distribution before and after the cutoff are not statistically significant.

Furthermore, the relocation trends of the population at the threshold is important to analyze. If a large number of individuals decide to move around the cutoff to be located closer to a HEI, then the average distance to a campus would decrease. Thus, the sudden decrease in the average distance would not be random due to the openings of the new locations. Instead, the observed decrease would be driven by individuals moving closer to these openings. Figure 16 in the Appendix illustrates the moving trends in the sample. The graphs show no clear discontinuity at the threshold and t-test results suggest statistically insignificant differences before and after the cutoff. This suggests that the decrease in average distance was not driven by an increase in internal relocation.

8.2 Density Checks

It is important that the assignment of individuals into the treatment or control group is random. Therefore, individuals should have no control over their cohort. In other words, the random treatment assignment would require a continuous number of students in cohorts around the cutoff. Notably, it is possible, however unlikely, that the sudden decrease in average distance to HEIs influenced a parents decision to have a child. The original approach to test manipulation around the cutoff was the McCrary density test, introduced by McCrary (2008). Similar methods have since been developed to plot observations around the cutoff and identify sorting. In this study, the Manipulation Test introduced by Cattaneo, Jansson, and Ma (2020) is used. The null-hypothesis of this test is that there is no difference in the density of observations before and after the cutoff. For the running variable, Years, the test calculated a p-value of 0.000. This means that the test rejects the null-hypothesis and suggests that there is a difference between the number of observations before and after the cutoff. However, it is possible that the result does not reveal manipulation. Instead, given the natural decline of an older population, the difference in the number of observations may be driven by the sample availability of mature individuals. Additional analysis would be required to support manipulation at the threshold.

8.3 Endogeneity of Opening Locations

It is reasonable to assume that universities strategically choose the location of new campuses. For example, university's might target areas where there is a high demand for higher education institutions and insufficient supply. This provides an explanation to why new campuses in the state of SP were opened in such remote areas. In that case, the estimates would suffer from sample selection bias as campus placements would have targeted a certain group. To evaluate this concern, Table 8 explores the characteristics of individuals living in the municipalities where a new campus was opened around the sudden decrease in average distance. The statistics describe the characteristics of individuals who turned 18 before the change. Specifically, the 1950 cohort. For comparison, the characteristics of individuals across municipalities are included in the table.

Table 8: Descriptive statistics of 1950 cohort from municipalities where university campuses were later opened around 1955

Municipality	**Education**	**Graduate**	**Literacy**	**Employed**	**Income**	N
Aracatuba	4.9	0.10	0.14	0.07	637.75	59
Assis	4.6	0.11	0.18	0.18	222.86	38
Piracicaba	5.1	0.08	0.12	0.11	653	105
Sao Jose dos Campos	4.19	0.04	0.12	0.11	367.88	74
All	4.49	0.05	0.17	0.15	1,053.88	10,732

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

Note: Education refers to years of education. Graduate refers to university graduations (1 if yes, 0 otherwise). Literacy refers to literacy rate (1 if individual can read and write, 0 otherwise). Employed refers to employment status at the time of surveying (1 if employed, 0 otherwise). Income declared in Brazilian Reais based on monthly rates.

As expected due to the large number of municipalities in the state, very few observations are available for the municipalities of interest. Nevertheless, it is possible to draw potential insights from the limited results. Table 8 shows that São Jose dos Campos is the only municipality with the number of years of education and university graduation rate lower than the average across all municipalities. Furthermore, literacy rates in all municipalities are below the average rate, excluding Assis. The combination of high university graduation rates, low number of years of education and low literacy rate suggests education disparity. Furthermore, it may suggest that individuals who pursued more education followed higher education to completion. Moreover, the average income for all municipalities of interest are lower than the average.

It is possible that the government opened campuses in these municipalities to drive an increase in education attainment. Assuming that individuals from these municipalities were more likely to pursue higher education relative to other locations, this strategic placement could greatly benefit the target population. Notably, the opening of HEI near these low-income municipalities could stimulate attendance by lowering costs. Based on the high graduation rates, one could speculate that the individuals from these municipalities have a greater innate motivation or interest in higher education. In the case that these individuals are more motivated, their expected income regardless of graduating would be greater. This could suggest an overestimation of the returns to higher education. However, this is a superficial analysis with a small sample size for each municipality. Thus, the implications outlined remain speculative.

8.4 Regression Discontinuity Design Analysis for Non-Movers

The results of the main RDD analysis suggest that the decrease in distance acts as a strong instrument to identify the positive causal effect of higher education attainment on income. An additional test to check the robustness of this findings is to conduct the same analysis with the group of non-movers. In theory, the sudden decrease should have a similar effect on everyone regardless of their moving behaviour. Thus, similar estimates should be attained when limiting the analysis to non-movers.

Table 9: Regression discontinuity design returns to university graduation using the 1955 cohort-level decrease in average distance to a university campus, for non-movers

	(1)	(2)	(3)	(4)	(5)	(6)
	University	Log(Income)	Log(Income)	University	Log(Income)	Log(Income)
Decrease in distance	0.03*	0.28***	-	0.03***	0.17	-
	(0.02)	(0.07)		(0.01)	(0.17)	
University	-	-	5.48***	-	-	2.52**
			(0.49)			(1.12)
F-statistic	5.90	105.30	364.19	4.38	10.38	69.83
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
Ethnic Group	All	All	All	Mixed	Mixed	Mixed
Gender	All	All	All	All	All	All
Observations	34,671	34,671	34,671	3,159	3,159	3,159

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

Note: Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Columns 1, 2 and 3 refer to the same estimation regression. Columns 4, 5 and 6 refer to the same estimation regression.

Interestingly, while the original RDD analysis indicated that the instrument was strong for two groups (all, mixed), the results based on non-movers contradict this finding. The F-values presented in Columns 1 and 4 (Table 9) indicate that the instrument is weak. Notably, the IV estimates presented in Columns 3 and 4 of the table, present similar results as the original RDD. Nevertheless, the potential IV bias from the weak instrument indicate that the estimates of interest are imprecise or unreliable. The results of this robustness check were expected to coincide with the results of the main analysis. However, it is shown that for non-movers, the sudden decrease in average distance is a weak instrument when estimating the returns to higher education. This suggests that the opening of new university campuses had a greater impact on movers and points to an area of further research.

9 Limitations

Given that the limitations of the individual level analysis have been discussed in previous sections, this segment focuses on the limitations related to the cohort level analysis.

9.1 Data

A main data constraint encountered in this research was the imprecise measures of individual distances to HEIs. This is due to the fact that the available individual observations did not report exact home addresses. Given the lack of precise individual addresses, this study opted to using the midpoints of the respective municipalities as an estimate. While this approach provided a practical alternative, it also introduced imprecision across observations. For example, two individuals living at different distances from the center of the municipality would be assigned the same distance to a HEI. Essentially, the IV for the individual level analysis would be assumed to be the same across individuals from the same municipality. While this limitation restricts the reliability and causal inference of our estimates, insights from the available data provide a preliminary understanding into the relationship under investigation. Ideally, individual addresses from where they lived before turning 18 years-old should be used as the instrument.

Furthermore, the available data does not provide exact birthdays for the individuals in the sample. As a result, an individuals cohort was calculated based on their reported age in 2000. This means that individuals of the same age are all expected to have made the decision to attain higher education in the same year. Intuitively, it is possible that people born in the same year made higher education attainment choices at different ages. Thus, the precision of the treatment assignment based on the threshold may be compromised. Given this limitation, this study was also unable to control for age due to collinearity. This is due to the fact that the cohort and year fixed effects coincided with the age. Therefore, due to collinearity, age could not be a control variable. In addition, the data lacks information regarding an individuals years of experience. As highlighted by previous literature, post-schooling experiences also greatly impact earnings. As the analysis was unable to estimate the effects of experience on earnings, it is possible that the findings suffer from bias.

Another data limitation in this study pertains to the lack of controls for background characteristics. For example, the available data set did not report an individuals family characteristics. Factors such as parental income or parental education are therefore not elaborated on. As previous studies such as Card (1999) have suggested, characteristics of an individuals family background have an impact on their later life outcomes. Thus, the effects of unobserved background characteristics could bias the estimates. In an ideal analysis, this study would have access to additional background characteristics to control for.

Descriptive statistics point out a potential concern for the generalizability of the findings. Table 16 shows that the vast majority of the sample are white. According to the 2000 Census (Loveman, Muniz, & Bailey, 2013), 45% of the population identified as mixed or black and indicates an underrepresentation in the sample. This underrepresentation may be driven by the fact that more white individuals were willing to participate in the survey. The larger sample size of white individuals could explain the greater estimation power of this demographic. Meanwhile, the estimation of returns for mixed/pardos might have benefited from a larger sample size.

9.2 Identification Approach

As outlined, the identification approach of this study would have greatly benefited from additional controls. Moreover, this section raises three issues related to the identification process.

Firstly, the accuracy of the measured distance to a HEI for individuals could further lack accuracy. The method used to measure the distances relied on the reported opening dates on the available websites, reports and contact people. In most cases, there were multiple dates to consider. This included the date when the campus became a legal entity and the begin of operations. This study used the date when administrative work in the campus began as the official opening date. However, it is likely that the effect of a new campus may suffer a lag given that the news of a new HEI may slowly gain salience. Notably, this potential lag is not accounted for in this study. At the individual level analysis this would suggest an inaccurate instrument and at the cohort level analysis it may indicate that the cutoff may lack precision.

To estimate the causal effect of higher education attainment on earnings, it is important to account for other factors which influence outcomes. A schooling alternative not investigated in this study are private HEIs. Using time fixed effects the estimation method attempts to control for trends in private HEIs. However, as it is not thoroughly explored, changes in higher education attainment could be affected by developments in the network of private HEIs. Moreover, controlling for school quality could also reduce bias. As previous studies have highlighted, not all schooling is the same. The quality of schooling plays an important role in understanding its impact on individuals. This study uses university fixed effects to control for the time-invariant quality level of campuses. However, it is possible that campuses vary in educational quality. In that case, controlling for educational quality could strengthen the identification strategy.

Another concern is the use of a continuous instrumental variable in the individual level analysis. The estimation of the LATE depends on the assumption of monotonicity. As outlined by Kennedy, Lorch, and Small (2019), causal inference is usually drawn using a binary instrument. In the individual level analysis this is not the case. Kennedy et al. (2019) outlines that continuous instruments result in weighted estimates. Individuals who are most responsive to the instrument are given greater weights. This would suggest that the estimate would be skewed towards individuals who are most responsive to a change in proximity to a HEI. For example, individuals from low-income families may be more responsive to a change in distance compared to their wealthier counterpart. In that case, the estimated effect of higher education on earnings would be skewed towards the former. This is less of a concern for the cohort level analysis given the binary instrument. Nevertheless, accounting for the features of a continuous instrument is essential to interpretation of results.

10 Discussion

This section briefly outlines policy implications that can be drawn from the findings of this study and potential extensions. Notably, this study recognizes that the increase in earnings could be driven by the skills developed in higher education or the credentials attained by completing the degree. This 'sheep skin effect' is a general concern in the study of returns to education but is not further elaborated in this section.

10.1 Policy implications

The approach of this study highlights a few potential policy implications related to the instrument. As the study suggests, a decrease in the distance to HEIs seem to increase an individual's likelihood of attaining higher education. However, if the goal of the government is to decrease the distance to HEIs for individuals, opening new campuses may be less effective in the current context. As aforementioned, the opening of 120 new campuses after 2000 did not reduce average distance to a HEI to the extent that the opening of 4 campuses did around 1955. It seems that the state of SP has become saturated by the number of public HEIs and opening of new campuses may be less effective at reducing average distance. Notably, the opening of new HEIs in the city may be driven by population growth which results in an increase in demand. Nevertheless, policies related to the opening of new HEIs need to account for the evolving landscape of higher education accessibility.

The results of this study also highlight the importance of determining who benefits the most from the opening of new HEIs. As shown in earlier maps, HEIs are concentrated in the city of SP. Compared to other municipalities, the city of SP has historically experienced more economic growth due to its metropolitan nature. Consequently, individuals from the city have a greater average income relative to other municipalities. This raises the question if opening more campuses in this city is an efficient way for the state government to stimulate higher education attainment. In theory, it might be a more efficient use of resources for the government to target individuals who would otherwise stop their educational journey. As public HEIs do not charge a tuition fee, providing low-income individuals with more access could further their educational attainment.

10.2 Extensions

Building on this study, a number of extensions could be explored. This section presents three potential directions for exploration which were not examined due to time limitations.

In the RDD analysis, this study established the cutoff year as 1955. However, Figure 9 in the Appendix shows a progressive decrease between 1954 and 1956. This transition period introduces some ambiguity in the treatment allocation. To address this concern, a simple one-year 'donut-hole' RDD could be conducted. This is illustrated in Figure 11, in the Appendix.

Applying this method, the observations one year before and after the cutoff period would be excluded from the analysis. By removing one observation before and after 1955, the cutoff seems to be sharper. Thus, it is reasonable to assume that we would see a more pronounced effect of the instrument.

Furthermore, it would also be valuable to explore a cost-benefit analysis of higher education attainment based on the findings. This study briefly estimates the potential returns to higher education in absolute monetary value. However, this estimate is obtained using sample averages which might not reflect the profile of everyone in the population. Thus, to interpret the implications of these findings, future research could explore the heterogeneous returns to higher education in absolute terms. Furthermore, this study did not address the additional costs associated with schooling. Tuition was accounted for as the only cost, however, opportunity costs, study material costs and reallocation costs are all factors which influence an individual's educational choices. By evaluating a more complete set of costs and benefits, future research could explore for whom higher education attainment is beneficial.

In addition, the analysis conducted in this paper could be replicated using a more complete data set. As outlined in the previous section, a number of additional controls could add to the estimation power of the analysis. In the case of the individual analysis, a data set with complete observations including an individual's precise address at the age of 18 would allow for a more accurate instrument. Moreover, exploring a more recent sample could also provide further insights. The cohort level analysis focuses on the mature population. These individuals made educational decisions during a time when women and non-white individuals had a limited access to opportunities. This factor may have impacted the educational attainment of individuals in the sample. While inequalities are still observed in Brazil, the current context is more equitable compared to older generations. Therefore, basing future research on more recent data sets could result in an improved causal inference. For example, the IBGE population census of 2010 or the most recent 2022 edition.

11 Conclusion

This paper aimed to contribute to the available literature on the returns to higher education. Aforementioned studies outline a variety of identification approaches. Based on these papers, this study applied three approaches to estimate returns. Namely, the OLS, IV and RDD approaches. For the IV and RDD approaches, variations in proximity to HEIs were used to attain estimates. The analysis in this study were divided into two parts: individual and cohort level. The former used the individual's distance to a HEI as the instrument, while the latter used a sharp discontinuity of the average distance to HEIs in 1955. OLS estimates suggest that university graduation increases income by 1.03% - 1.16%. IV estimates suggest an increase in income by 4.3% - 5.0%. Moreover, RDD estimates for the complete sample find an increase in income of 6.7% and an estimate of 2.0% for the mixed/pardo group. Notably, all estimation approaches indicate positive returns to higher education, which supports Hypothesis 1 of this study. As the IV estimates are greater than the OLS estimates, findings also support Hypothesis 2. However, robustness checks conducted on non-movers raises doubts on the external validity of the estimates. As outlined in the previous section, data limitations could be hindering causal inference. Future research could replicate the RDD approach introduced in this paper in combination with a more complete data set on individual characteristics and addresses. To the best of my knowledge, no previous study in the context of Brazil used distance to HEI as an instrument. Thus, this paper highlights an interesting approach to be further explored.

References

2022 ibge census. (2023). Instituto Brasileiro de Geografia e Estatística, IBGE. Retrieved from `https://censo2022.ibge.gov.br/en/about/learning-about-brazil.html`

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics* (pp. 235–251). Elsevier.

Altonji, J. G., & Pierret, C. R. (1998). Employer learning and the signalling value of education. In *internal labour markets, incentives and employment* (pp. 159–195). Springer.

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014.

Arretche, M. (2018). *Paths of inequality in brazil: a half-century of changes.* Springer.

Barbosa Filho, F. d. H., & Pessoa, S. d. A. (2008). Retorno da educação no brasil.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of political economy*, *70*(5, Part 2), 9–49.

Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis.* University of Chicago Press, With Special Reference to Education.

Becker, G. S. (2010). *The economics of discrimination.* University of Chicago press.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of political economy*, *75*(4, Part 1), 352–365.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. *National Bureau of Economic Research Cambridge, Mass., USA*.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics*, *3*, 1801–1863.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, *69*(5), 1127–1160.

Card, D., Cardoso, A. R., & Kline, P. (2016). Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *The Quarterly journal of economics*, *131*(2), 633–686.

Castagnetti, C., Chelli, F., Rosti, L., et al. (2005). Educational performance as signalling device: Evidence from italy. *Economics Bulletin*, *9*(4), 1–7.

Castro, C. d. M. (1970). *Investment in education in brazil: a study of two industrial communities* (Unpublished doctoral dissertation). Vanderbilt University.

Cattaneo, M. D., Jansson, M., & Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, *115*(531), 1449–1455.

De Mello E Souza, A. (1991). Higher education in brazil: Recent evolution and current issues. *Higher Education*, *21*(2), 223–233.

dos Santos, Y., & Custodio, J. (2022). *Racismo brasileiro: Uma história da formação do país.* Todavia. Retrieved from `https://books.google.nl/books?id=DaBvEAAAQBAJ`

Filmer, D. (2000). The structure of social disparities in education: Gender and wealth. *Available at SSRN 629118*.

Griffin, P., & Edwards, A. C. (1993). Rates of return to education in brazil: Do labor market conditions matter? *Economics of Education Review*, *12*(3), 245-256. Retrieved from `https://www.sciencedirect.com/science/article/pii/027277579390007 4` doi:

https://doi.org/10.1016/0272-7757(93)90007-4

Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. *Econometrica: Journal of the Econometric Society*, 1–22.

Hanushek, E. A. (2016). Will more higher education improve economic growth? *Oxford Review of Economic Policy*, *32*(4), 538–552.

Jallade, J.-P. (1977). Basic education and income inequality in brazil: The long-term view. world bank staff working paper no. 268.

Kennedy, E. H., Lorch, S., & Small, D. S. (2019). Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society Series b: Statistical Methodology*, *81*(1), 121–143.

Klein, L., & Schwartzman, S. (1993). Higher education policies in brazil: 1970–90. *Higher Education*, *25*(1), 21–34.

Langoni, C. G. (1974). As causas do crescimento econômico do brasil.

Loveman, M., Muniz, J. O., & Bailey, S. R. (2013). Brazil in black and white? race categories, the census, and the study of inequality. In *Accounting for ethnic and racial diversity* (pp. 109–126). Routledge.

McCowan, T. (2007). Expansion without equity: An analysis of current policy on access to higher education in brazil. *Higher education*, *53*, 579–598.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698-714. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0304407607001133` (The regression discontinuity design: Theory and applications) doi: https://doi.org/10.1016/j.jeconom.2007.05.005

Mello, U. (2022, August). Centralized admissions, affirmative action, and access of low-income students to higher education. *American Economic Journal: Economic Policy*, *14*(3), 166-97. Retrieved from `https://www.aeaweb.org/articles?id=10.1257/pol.20190639` doi: 10.1257/pol.20190639

Menezes Filho, N., & Kirschbaum, C. (2019). Education and inequality in brazil. *Paths of inequality in Brazil: a half-century of changes*, 69–88.

Mincer, J. (1974). *Schooling, experience, and earnings. human behavior & social institutions no. 2.* National Bureau of Economic Research, Inc.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature*, *56*(3), 799–866.

OECD. (2018). *Education at a glance 2018.* Retrieved from `https://www.oecd-ilibrary.org/content/publication/eag-2018-en` doi: https://doi.org/https://doi.org/10.1787/eag-2018-en

Oreopoulos, P., & Petronijevic, U. (2013). *Making college worth it: A review of research on the returns to higher education.* National Bureau of Economic Research.

Park, J. H. (1994). *Returns to schooling: a peculiar deviation from linearity* (Tech. Rep.). Princeton University, Department of Economics, Industrial Relations Section.

Psacharopoulos, G. (1985). Returns to education: A further international update and implications. *Journal of Human resources*, 583–604.

Psacharopoulos, G., & Patrinos, H. A. (2018). Returns to investment in education: A decennial review of the global literature. *Education Economics*, *26*(5), 445–458.

Saboia, J. (2007). Efeitos do salário mínimo sobre a distribuição de renda no brasil no período 1995/2005–resultados de simulações. *Revista Econômica*, *9*(2).

Sachsida, A., Loureiro, P. R. A., & Mendonça, M. J. C. d. (2004). Um estudo sobre retorno em escolaridade no brasil. *Revista Brasileira de Economia*, *58*, 249–265.

Schwartzman, S. (2012). Economic growth and higher education policies in brazil. *International Higher Education*(67).

Spence, M. (1978). Job market signaling. In *Uncertainty in economics* (pp. 281–306). Elsevier.

Stefani, P. C., & Biderman, C. (2009). The evolution of the returns to education and wage differentials in brazil: a quantile approach. *Applied Economics*, *41*(11), 1453–1460.

Stefani, P. C., Biderman, C., et al. (2006). Returns to education and wage differentials in brazil: a quantile approach. *Economics Bulletin*, *9*(1), 1–6.

Tannen, M. B. (1991). New estimates of the returns to schooling in brazil. *Economics of Education Review*, *10*(2), 123–135.

Teixeira, W. M., & Menezes-Filho, N. A. (2012). Estimando o retorno à educação do brasil considerando a legislação educacional brasileira como um instrumento. *Brazilian Journal of Political Economy*, *32*, 479–496.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, *51*(6), 309.

Weichselbaumer, D., & Winter-Ebmer, R. (2005). A meta-analysis of the international gender wage gap. *Journal of economic surveys*, *19*(3), 479–511.

Weiss, A. (1995). Human capital vs. signalling explanations of wages. *Journal of Economic perspectives*, *9*(4), 133–154.

A Supplementary Tables

Table 10: University names and abbreviations

University Name	**Abbreviation**
Instituto Federal de São Paulo	IFSP
Universidade Federal de São Carlos	UFSC
Universidade Federal de São Paulo	UFSP
Universidade Federal do ABC	UFABC
Instituto Tecnologico de Aeronautica	ITA
Faculdade de Tecnologia de São Paulo	FTSP
Universidade Estadual Paulista	UEP
Universidade Estadual de Campinas	UEC
Universidade de São Paulo	USP

Source: 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus names

Table 11: Descriptive statistics for adult sample, by age category (23-30, 31-50, 51-64)

Variable	**Mean**	**SD**	**Min**	**Max**	**N**
Age category 1 (23 - 30)					
Gender	.4944	0.5000	0	1	554,619
Age	26.43	2.30	23	30	554,619
Racial Minority	.3112	.4630	0	1	550,841
University	.0736	.2611	0	1	554,619
Years of Education	8.28	4.00	0	20	554,619
Literacy	.9725	.1634	0	1	554,619
Employed	.6604	.4736	0	1	554,619
Income	650.44	993.35	1	120000	373,027
Hours worked	45.25	12.72	1	140	377,581
Retired	.0081	.0897	0	1	554,619
Distance to University	23.19	27.26	0.30	200.77	554,619
From SP City	.2739	0.4460	0	1	554,619
Age category 2 (31 - 50)					
Gender	.4840	.4997	0	1	1,112,739
Age	39.68	5.62	31	50	1,112,739
Racial Minority	.2901	.4538	0	1	1,106,579
University	.1020	.3026	0	1	1,112,739
Years of Education	7.34	4.47	0	30	1,112,739
Literacy	.9488	.2204	0	1	1,112,739
Employed	.6512	.4766	0	1	1,112,739
Income	985.37	2398.68	1	600000	742,758
Hours worked	45.83	14.01	1	140	750,991
Retired	.0420	.2005	0	1	1,112,739
Distance to University	25.26	28.64	.30	222.37	
From SP City	.2643	.44093	0	1	1,112,739
Age category 3 (51 - 64)					
Gender	.4732	.4993	0	1	385,034
Age	56.73	4.01	51	64	385,034
Racial Minority	.2553	.4360	0	1	382,925
University	.0743	.2623	0	1	385,034
Years of Education	5.23	4.59	0	30	385,034
Literacy	.8641	.3427	0	1	385,034
Employed	.4020	.4903	0	1	385,034
Income	1095.31	4531.99	1	900000	160,426
Hours worked	44.88	15.17	1	140	163,778
Retired	.3120	.4633	0	1	385,034
Distance to University	36.32	38.41	0.30	286.01	385,034
From SP City	.2572	.4371	0	1	385,034

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

Table 12: The effect of campus distance on university graduation, Instrumental Variables first-stage estimates

Endogenous Variable:	(1)	(2)	(3)	(4)	(5)
University	IV FS	IV FS	IV FS	IV FS	IV FS
Log(Distance)	-0.033***	-0.035***	-0.034***	-0.035***	-0.035***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Male	-0.069***	-0.072***	-	-	-0.069***
	(0.006)	(0.006)			(0.006)
Ethnicity	Yes	No	Yes	No	Yes
White	0.090***	-	0.094***	-	0.088***
	(0.033)		(0.033)		(0.033)
Black	-0.053	-	-0.052	-	-0.055
	(0.034)		(0.034)		(0.034)
Asian	0.339***	-	0.0	-	0.336***
	(0.046)		(0.046)		(0.046)
Mixed	-0.062*	-	-0.060***	-	-0.068**
	(0.033)		(0.033)		(0.033)
Indigenous	-0.066	-	-0.067	-	-0.063
	(0.051)		(0.052)		(0.052)
Fixed Effects	X	X	X	X	
Observations	18,836	18,836	18,836	18,836	18,836
F-value	146.17	150.47	150.11	154.72	168.33

Note: Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

Table 13: The estimated probability of moving based on university graduation status

		University		
	N	0	1	Difference
All	2,676,085	0.045	0.093	+107%
Male	1,289,892	0.045	0.089	+98%
Female	1,386,193	0.046	0.097	+111%
White	1,906,605	0.053	0.098	+85%
Black	129,110	0.032	0.061	+91%
Mixed	587,715	0.028	0.050	+79%

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

Table 14: The effect of the 1955 cohort-level decrease in average distance to a university campus on university graduation

Treatment:	(1)	(2)	(3)	(4)	(5)	(6)
University	FS	FS	FS	FS	FS	FS
Decrease in distance	0.053***	0.024***	0.053***	0.040**	0.022***	0.037***
	(0.010)	(0.005)	(0.012)	(0.012)	(0.006)	(0.014)
F-statistic	11.83	10.47	8.96	5.61	7.57	4.70
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	No	No	No	No	No
Ethnic Group	All	Mixed	White	All	Mixed	White
Gender	All	All	All	Male	Male	Male
Observations	79,008	8,208	64,391	55,646	5,629	45,881

Note: Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

Table 15: The effect of the 1955 cohort-level decrease in average distance to a university campus on income, reduced form estimates

Dependent: Log(Income)	(1) RF	(2) RF	(3) RF	(4) RF	(5) RF	(6) RF
Decrease in distance	0.405***	0.245**	0.345***	0.380***	0.370**	0.378***
	(0.041)	(0.103)	(0.047)	(0.049)	(0.112)	(0.053)
F-statistic	397.40	6.33	23.58	31.53	7.23	28.29
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	No	No	No	No	No
Ethnic Group	All	Mixed	White	All	Mixed	White
Gender	All	All	All	Male	Male	Male
Observations	79,008	8,208	64,391	55,646	5,629	45,881

Note: Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and 2000 Brazilian Institute of Geography and Statistics Census Data.

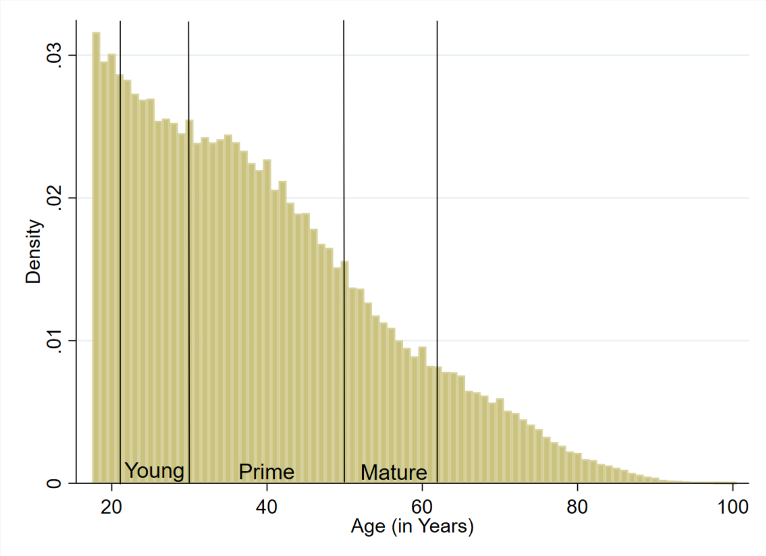
Table 16: Descriptive statistics of adult population by ethnicity: White, Black and 'Pardo'

Variable	**Mean**	**SD**	**Min**	**Max**	**N**
Ethnicity: White					
Gender	.4715	.4992	0	1	1,906,605
Age	39.95	16.14	18	100	1,906,605
University	.0904	.2866	0	1	1,906,605
Years of Education	7.56	4.54	0	30	1,906,605
Literacy	.9409	.2359	0	1	1,906,605
Employed	.5495	.4975	0	1	1,906,605
Income	939.32	2638.23	1	900000	1,073,234
Hours worked	45.01	13.70	1	140	1,091,696
Retired	.1350	.3417	0	1	1,906,605
Distance to University	32.66	43.65	0.30	597.08	1,906,605
From SP City	.2491	.4325	0	1	1,906,605
Ethnicity: Black					
Gender	.5147	.4998	0	1	129,110
Age	39.39	15.48	18	100	129,110
University	.0241	.1535	0	1	129,110
Years of Education	5.93	4.21	0	30	129,110
Literacy	.8785	.3267	0	1	129,110
Employed	.5738	.4945	0	1	129,110
Income	513.19	823.27	1	150000	75,755
Hours worked	45.33	13.52	1	140	76,437
Retired	.1213	.3264	0	1	129,110
Distance to University	27.60	40.49	0.30	556.05	129,110
From SP City	.3074	.4614	0	1	129,110
Ethnicity: 'Pardo'					
Gender	.5085	0.5000	0	1	587,715
Age	36.83	14.32	18	100	587,715
University	.0156	.1239	0	1	587,715
Years of Education	5.89	3.98	0	30	587,715
Literacy	.8969	.3040	0	1	587,715
Employed	.5654	.4957	0	1	587,715
Income	506.05	733.83	1	96000	339,477
Hours worked	45.90	13.40	1	140	343,270
Retired	.0839	.2772	0	1	587,715
Distance to University	27.77	40.17	0.30	597.08	587,715
From SP City	.2920	.4547	0	1	587,715

Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and Individual observations from the 2000 Brazilian Institute of Geography and Statistics Census Data.

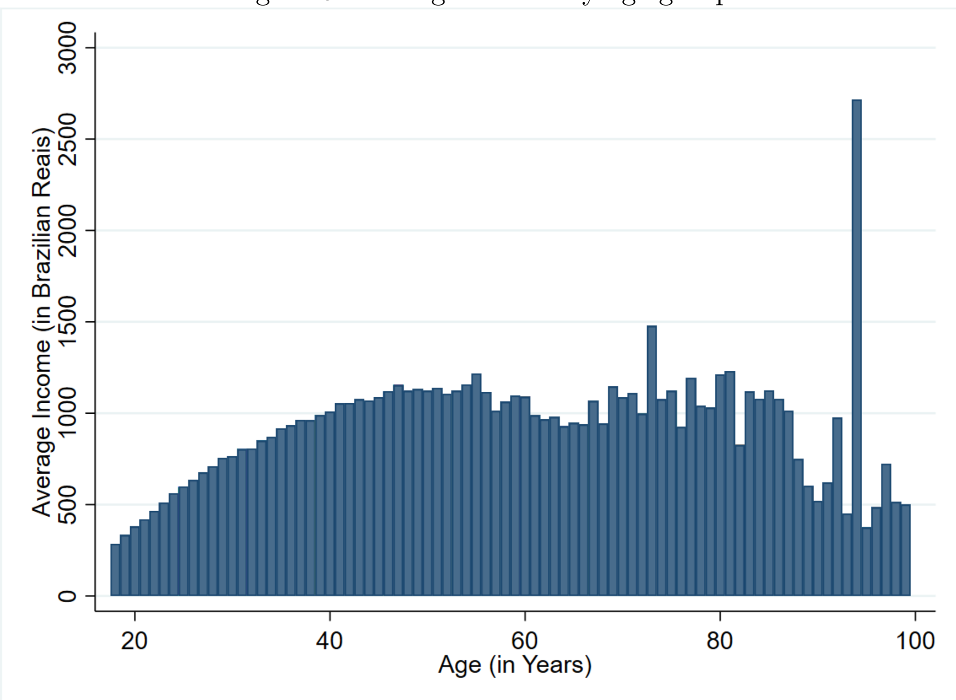
B Supplementary Figures

Figure 7: Age distribution across adult sample (above 18 years old)



Source: Based on 2000 Brazilian Institute of Geography and Statistics Census Data, created by the author

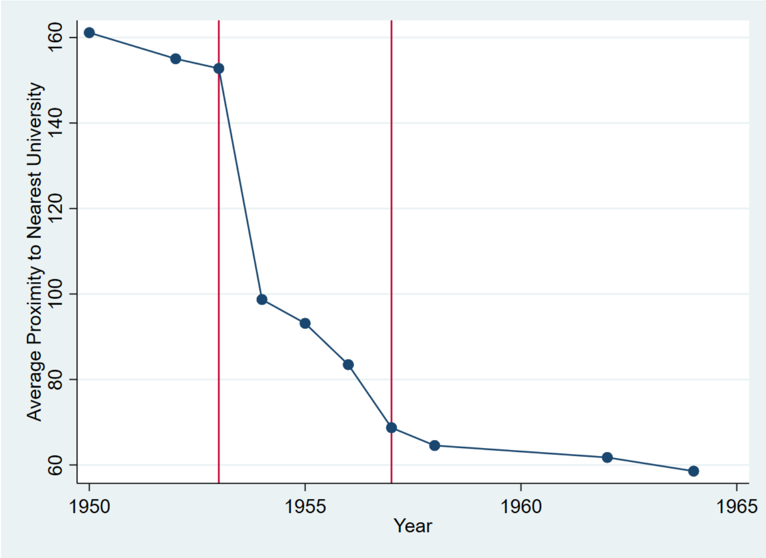
Figure 8: Average income by age group



Note: This figure depicts the average income of individuals by age groups, by years. Average income is measured in Brazilian Reais and is not adjusted to inflation.

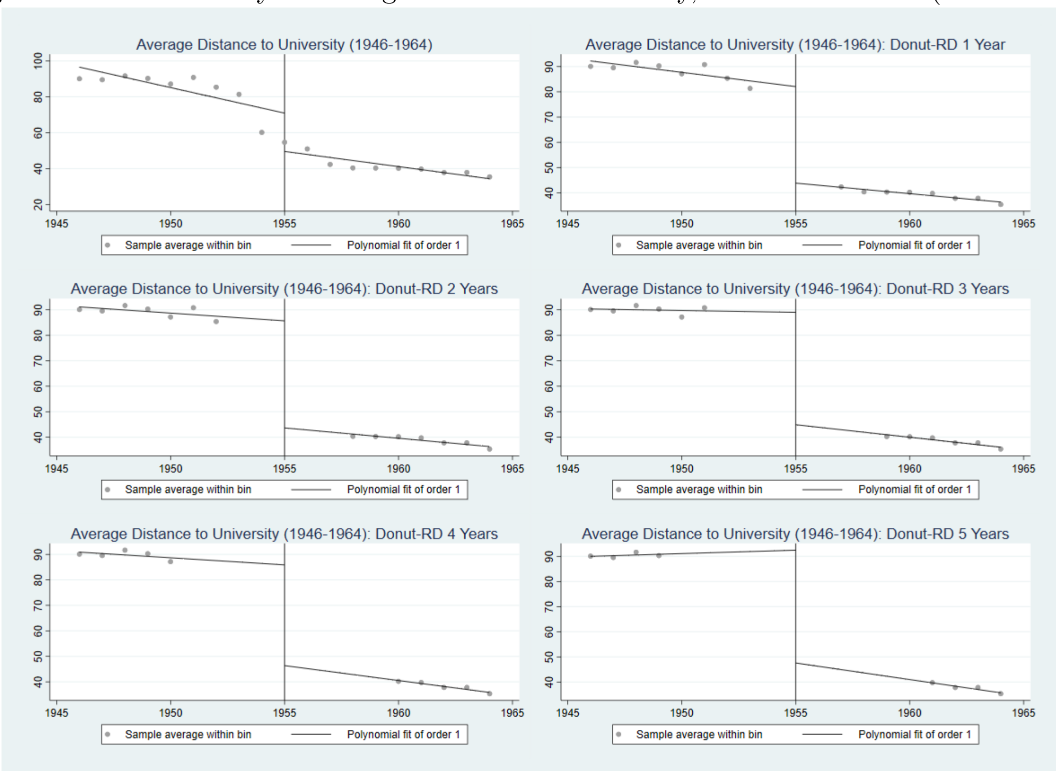
Source: Based on 2000 Brazilian Institute of Geography and Statistics Census Data, created by the author

Figure 9: Average proximity to nearest university unit per year (detailed jump)



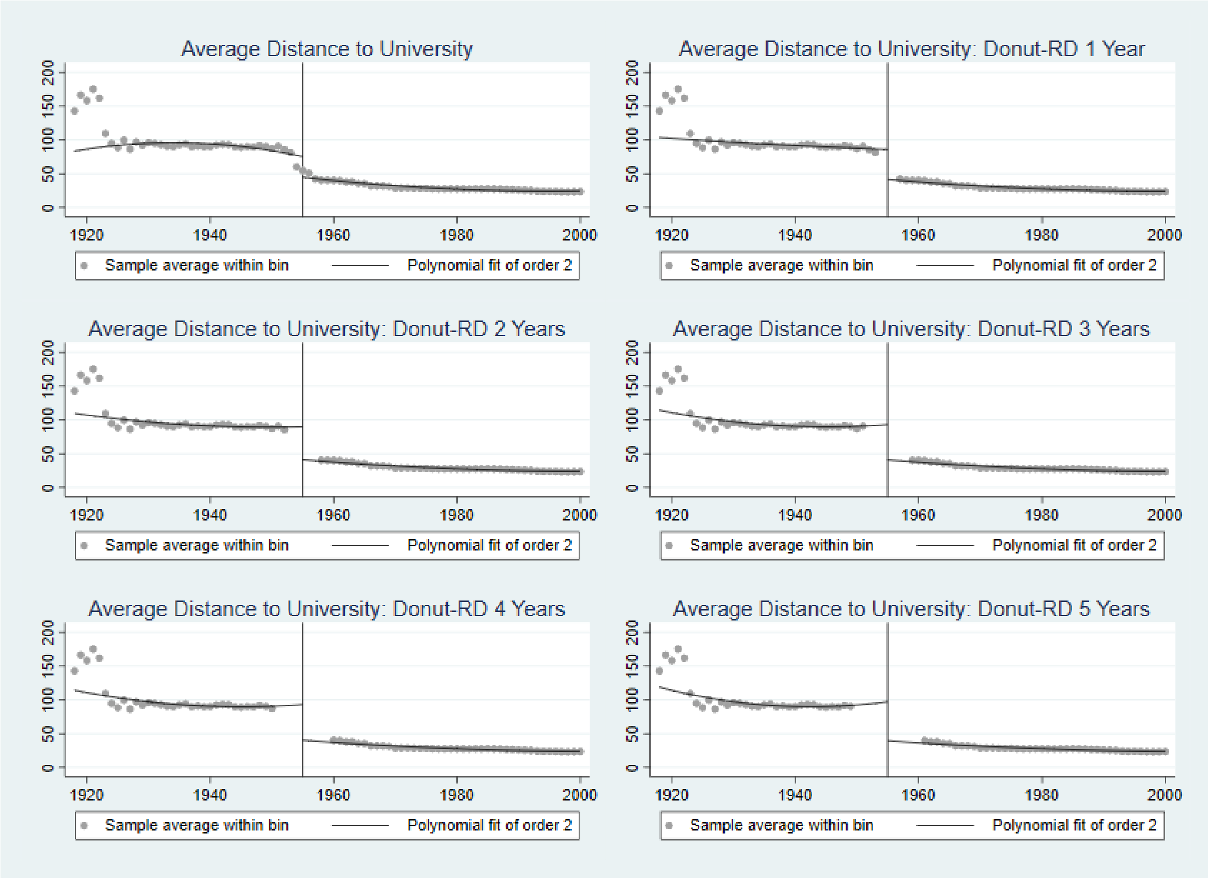
Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates.

Figure 10: Discontinuity of average distance to university, with donut-holes (1945 - 1965)



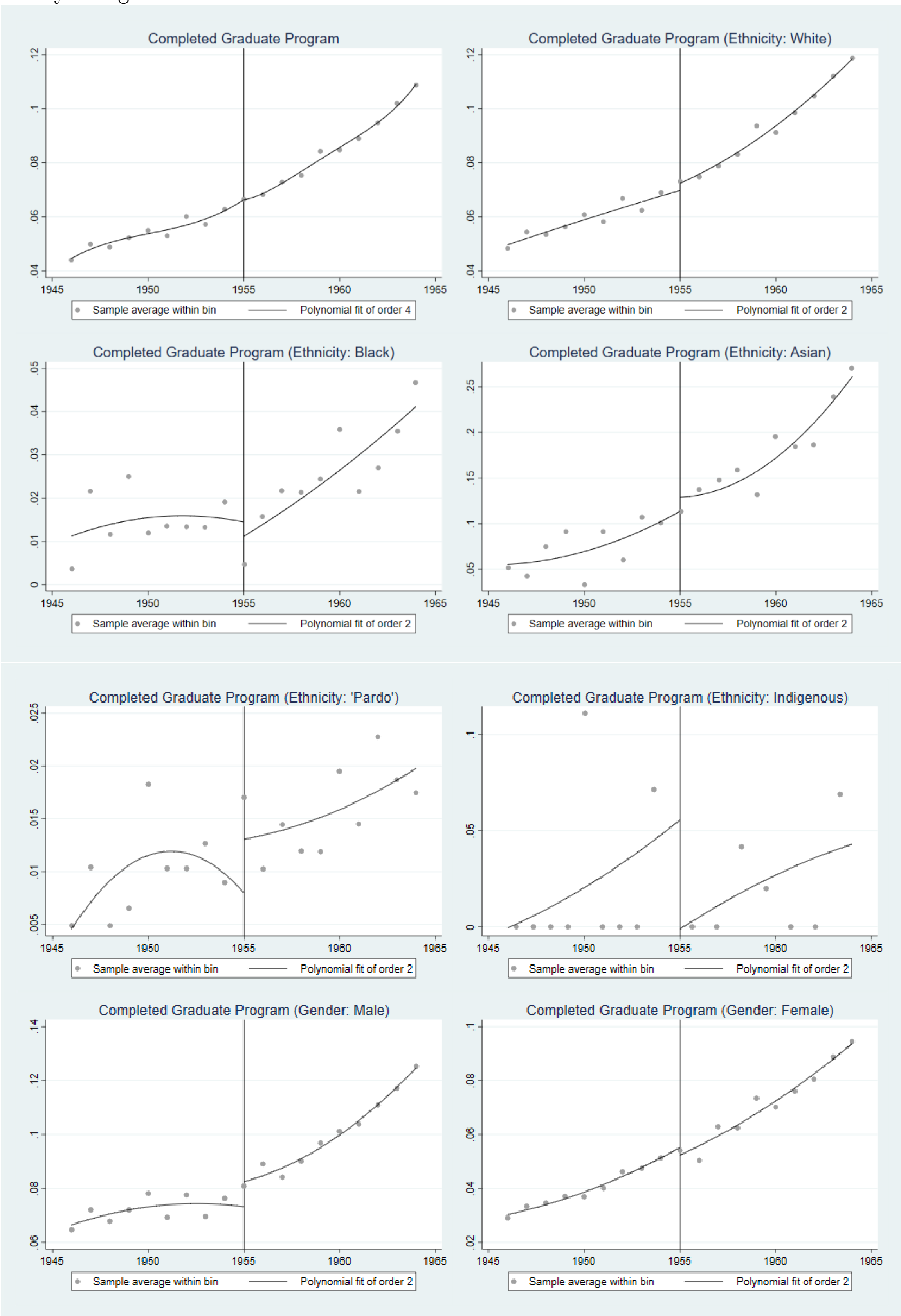
Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and Individual observations from the 2000 Brazilian Institute of Geography and Statistics Census Data.

Figure 11: Discontinuity of average distance to university, with donut-holes



Source: Opening dates of university campuses collected by author, combined with 2020/2021 Statewise System for Data Analysis Foundation municipal and university campus coordinates, and Individual observations from the 2000 Brazilian Institute of Geography and Statistics Census Data.

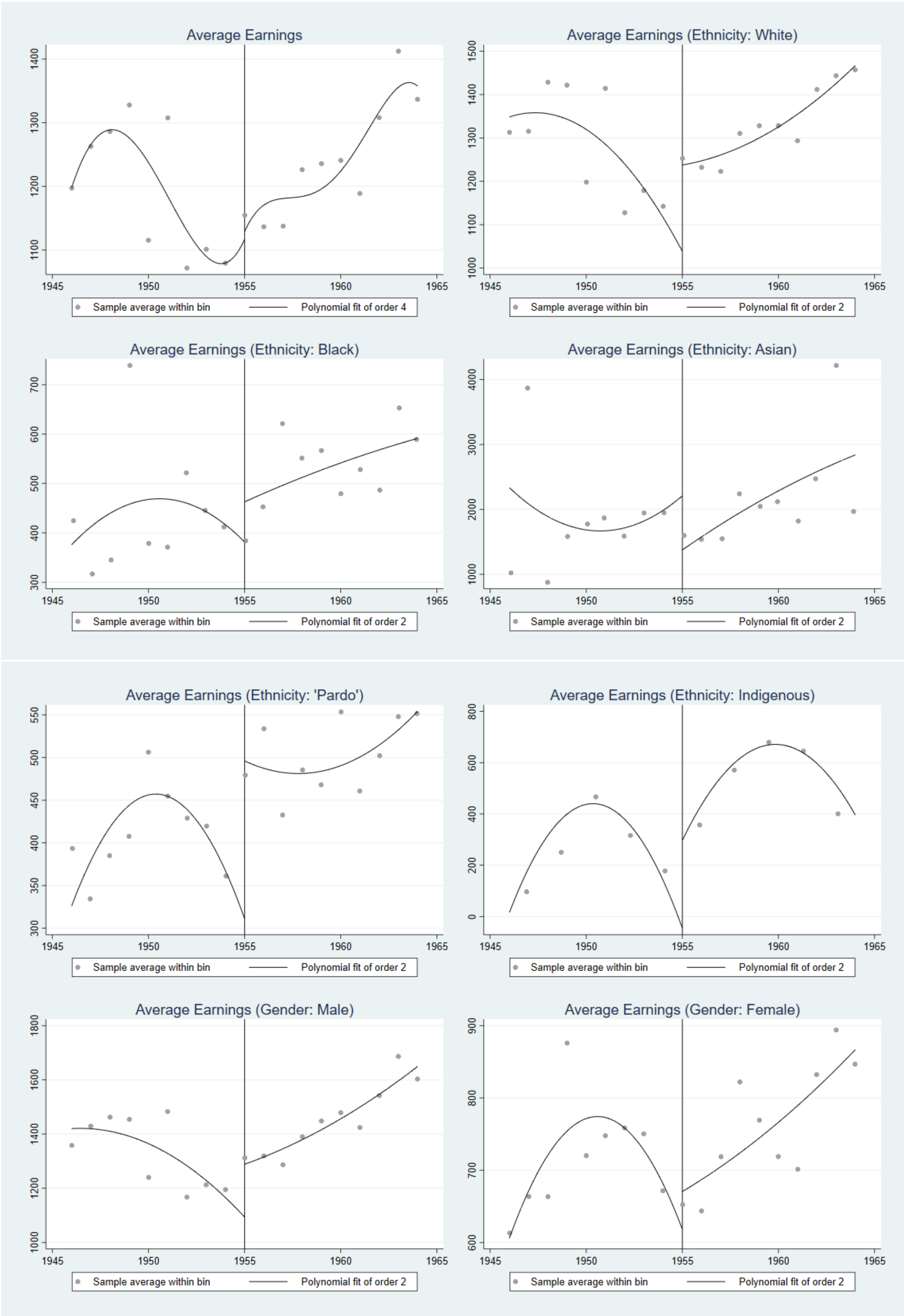
Figure 12: Discontinuity of average university completion after 1955 for complete sample, by ethnicity and gender



Note: Line at 1955 illustrates cutoff point used in the study

Source: Individual observations from the 2000 Brazilian Institute of Geography and Statistics Census Data.

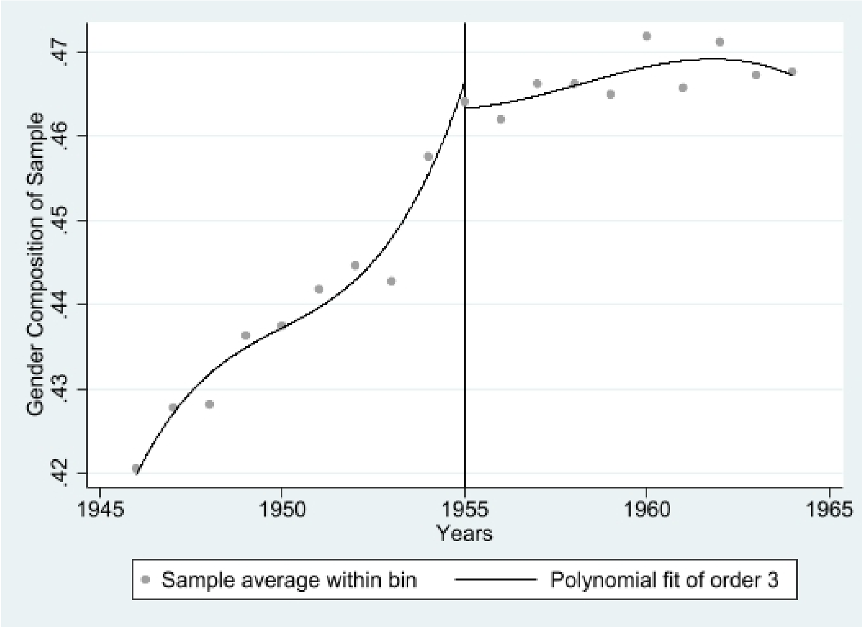
Figure 13: Discontinuity of average earnings after 1955 for complete sample, by ethnicity and gender



Note: Line at 1955 illustrates cutoff point used in the study

Source: Individual observations from the 2000 Brazilian Institute of Geography and Statistics Census Data.

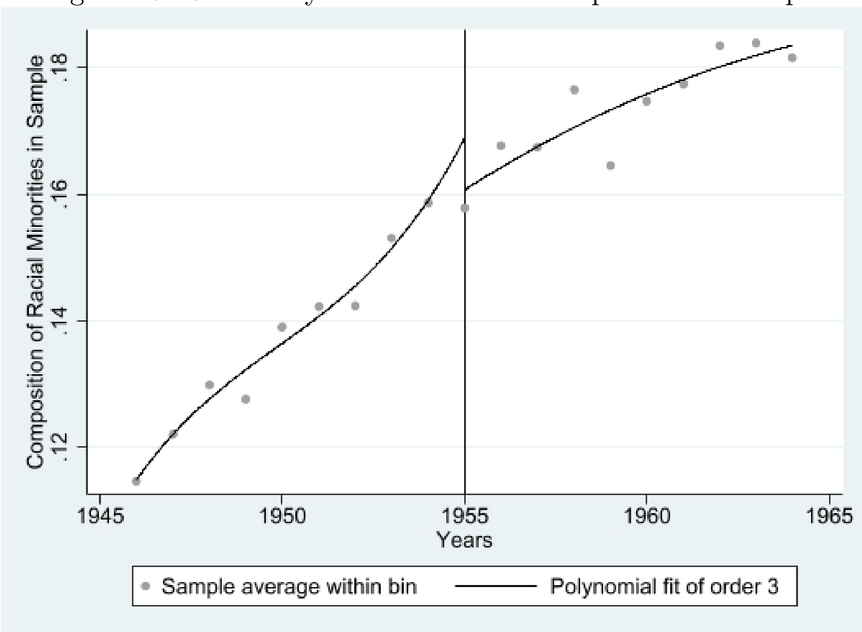
Figure 14: Continuity check for gender composition of sample



Note: Line at 1955 illustrates cutoff point used in the study

Source: Individual observations from the 2000 Brazilian Institute of Geography and Statistics Census Data.

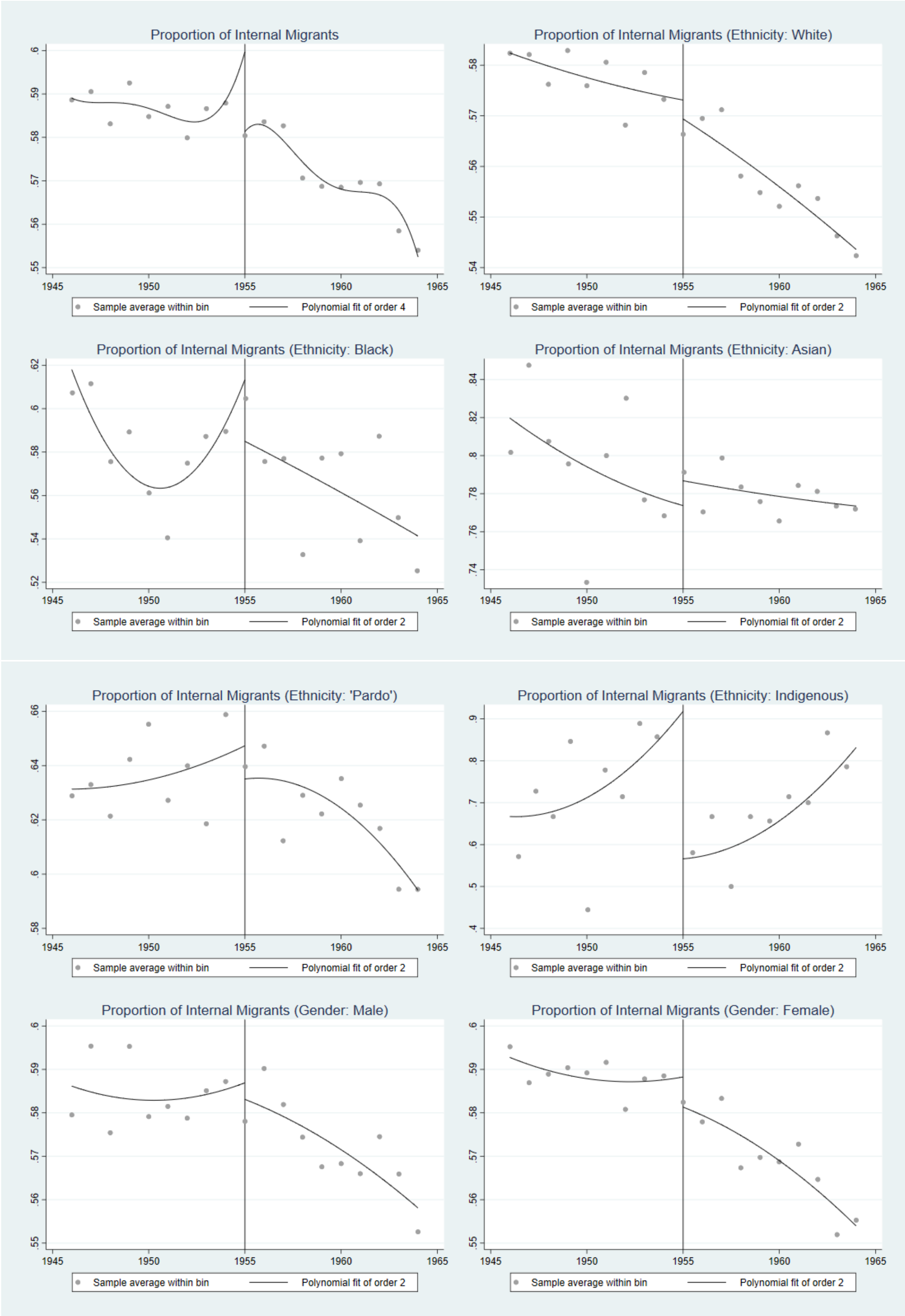
Figure 15: Continuity check for ethnic composition of sample



Note: Line at 1955 illustrates cutoff point used in the study

Source: Individual observations from the 2000 Brazilian Institute of Geography and Statistics Census Data.

Figure 16: Discontinuity of average internal migration (moving rate) for complete sample, by ethnicity and gender



Note: Line at 1955 illustrates cutoff point used in the study

Source: Individual observations from the 2000 Brazilian Institute of Geography and Statistics Census Data.