

Bentley University

Scholars @ Bentley

2020

Dissertations and Theses

2020

The Analytics of Vulnerable Populations in Brazil

Fernanda M. Araujo Maciel

Follow this and additional works at: https://scholars.bentley.edu/etd_2020



Part of the [Behavioral Economics Commons](#), [Econometrics Commons](#), [Health Economics Commons](#), and the [International Economics Commons](#)

©Copyright 2020
FERNANDA M. ARAUJO MACIEL



BENTLEY
UNIVERSITY

Bentley University: PhD Final Defense Form

This is to certify that we have examined this copy of a doctoral dissertation by

Fernanda M. Araujo Maciel

and have found that it is complete and satisfactory and that any and all revisions required by the final examining committee have been made

Committee Chair: Dr. Dominique Haughton

Signature

Committee Member: Dr. Dhaval Dave

Signature

External Reviewer: Dr. Jennifer Priestley

Signature

Date: April 13th, 2020

THE ANALYTICS OF VULNERABLE POPULATIONS IN BRAZIL

FERNANDA M. ARAUJO MACIEL

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy in Business

2020

Program Authorized to Offer Degree:
Department of Mathematical Sciences

ProQuest Number:27959669

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent on the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 27959669

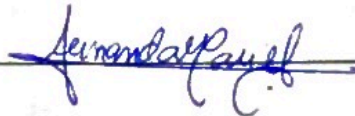
Published by ProQuest LLC (2020). Copyright of the Dissertation is held by the Author.

All Rights Reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at Bentley University, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to ProQuest Information and Learning Author Relations Team at (800) 521-0600 ext. 7020, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and /or (b) printed copies of the manuscript made from microform."

Signature 

Date April 13th, 2020

Abstract

THE ANALYTICS OF VULNERABLE POPULATIONS IN BRAZIL

FERNANDA M. ARAUJO MACIEL

Chair of the Supervisory Committee:

Professor Dominique Haughton

Department of Mathematical Sciences

Conditional Cash Transfer (CCT) programs became a popular measure to alleviate poverty in Latin American countries. The Brazilian CCT program, called *Bolsa Família*, is the largest social welfare program in the country, covering a quarter of all Brazilian households. The objective of the program is to reduce poverty and malnutrition, while providing low-income families access to public services, such as health, education, and social assistance. Since the program is based on conditions of maintaining health and schooling for children, my dissertation comprises two studies that examine the impact of *Bolsa Família* on educational outcomes and unhealthy behaviors of its participants. The third chapter investigates direct and indirect associations between negative body image perception, depression, and risk behaviors among adolescents in Brazil.

In Chapter 1, I evaluate the impact of Brazil's *Bolsa Família* program on the probability of dropping out of school, grade progression, and grade repetition. Prior literature has explored the impact of this program on educational outcomes, but these studies use methods that do not address endogeneity. To properly examine the effect of this program, this paper presents the Instrumental Variables method that controls for reverse causality and omitted variables. Consistent with the literature, my results

show that the rate of dropping out of school decreases if the household participates in the program, but when addressing endogeneity, this effect is not significant. Surprisingly, in contrast with what is documented by the literature, I find that the rate of grade progression decreases among participants. Finally, I estimate that program participants increase the chance of grade repetition, which was not previously studied. These findings are important to understand the effectiveness of the program in maintaining children in school, contributing to the literature of public policy and causality estimation.

In Chapter 2, I investigate if participants of *Bolsa Família* are increasing their expenses with ultra-processed foods, alcohol, and smoking products. For these analyses, I use the Propensity Score Matching method. Different from the existing methodology, I incorporate a machine learning approach for better predictability of the propensity score. I present a comparison between Random Forest, Gradient Boosting, Support Vector Machines, Neural Networks, and Logistic Regression in propensity score estimation. My results show that program participants purchase more food and increase expenses with snacks, such as cookies and out-of-home pastries, but they are not purchasing more unhealthy products than non-participants. This study also contributes to the literature on machine learning models for econometrics estimation.

In Chapter 3, I evaluate direct and indirect associations between negative body image perception, depression, and risk behaviors among adolescents in Brazil. Literature has shown associations between these factors among adolescents, however, few studies analyze the Brazilian population. In this paper, I estimate the effects using Directed Acyclic Graphs (DAG), a model that is based on a network of conditional independent nodes (Causal Markovian Condition), instead of theory. My results show similarities to the studies in the literature, validating DAG as a method of identifying directed links between variables. This model also finds associations not yet studied in the literature, shedding light on the vulnerability of Brazilian adolescents and their propensity to risk behaviors.

Table of Contents

Abstract	v
Table of Contents	vii
List of Tables	x
List of Figures	xii
List of Abbreviations	xiii
Introduction	1
1 Unintended Consequences of Welfare Programs in Schooling Outcomes: Evidence from Brazil	7
1.1 Introduction	7
1.2 Schooling Outcomes	8
1.3 Data	9
1.3.1 Study Sample	10
1.3.2 Descriptive Statistics	11
1.4 Methodology	12
1.5 Results	16
1.5.1 OLS Regression	16
1.5.2 Grade Progression	18
1.5.3 Grade Repetition	20
1.5.4 Dropout	20

1.5.5	Validity of the Instrument	23
1.6	Supplementary Models	26
1.7	Further Analysis	26
1.7.1	Grade Progression by Age Group	30
1.7.2	Grade Repetition by Age Group	31
1.7.3	Dropout by Age Group	31
1.7.4	Grade Progression by Gender	32
1.7.5	Grade Repetition by Gender	32
1.7.6	Dropout by Gender	32
1.8	Discussion and Conclusion	33
1.9	Appendix	36
2	The Impact of Cash Transfer Participation on Unhealthy Consumption in Brazil	50
2.1	Introduction	50
2.2	Food consumption among <i>Bolsa Família</i> participants	52
2.3	Ultra-processed Foods	52
2.4	Data	54
2.4.1	Study Sample	54
2.4.2	Variables	55
2.5	Descriptive Statistics	56
2.6	Methodology	57
2.6.1	Propensity Score Matching Method	58
2.7	Logistic Regression	60
2.8	Machine Learning Models	61
2.8.1	Random Forests	62
2.8.2	Gradient Boosting Machine	63
2.8.3	Support Vector Machines	64
2.8.4	Neural Networks	66

2.9	Propensity Score Matching Results	68
2.9.1	Model selection	68
2.9.2	Propensity Score Balance	71
2.9.3	Propensity Score Matching	73
2.9.4	Robustness Check	75
2.10	Conclusion	75

3 Direct and Indirect Associations between Body Image Perception, Depression, and Risk Behavior among Brazilian Adolescents 77

3.1	Introduction	77
3.2	Methodology	79
3.2.1	Exploratory Factor Analysis	79
3.2.2	Confirmatory Factor Analysis	79
3.2.3	Directed Acyclic Graphs	80
3.2.4	Structural Equations Modeling	81
3.3	Data	81
3.3.1	Variables	82
3.3.2	Descriptive Statistics	83
3.4	Results	86
3.4.1	Exploratory Factor Analysis	86
3.4.2	Confirmatory Factor Analysis	87
3.4.3	Directed Acyclic Graphs	89
3.4.4	Structural Equations Modeling	90
3.5	Discussion	92
3.6	Conclusion	93

References 95

List of Tables

1.1	Descriptive Statistics	13
1.2	OLS Estimates of Grade Progression, Grade Repetition and Dropout	17
1.3	Instrumental Variables Result for Grade Progression	19
1.4	Instrumental Variables Result for Grade Repetition	21
1.5	Instrumental Variables Result for Dropout	22
1.6	Instrumental Variables Result for Dropout for Ages 16 and 17	24
1.7	Instrumental Variables Result for Grade Progression: Low Education	27
1.8	Instrumental Variables Result for Grade Repetition: Low Education .	28
1.9	Instrumental Variables Result for Dropout: Low Education	29
1.10	OLS Estimates of Grade Progression, Grade Repetition and Dropout among Eligible Households	30
1.11	Instrumental Variables Result for Grade Progression by Age Group 6-12	36
1.12	Instrumental Variables Result for Grade Progression by Age Group 13-15	37
1.13	Instrumental Variables Result for Grade Repetition by Age Group 6-12	38
1.14	Instrumental Variables Result for Grade Repetition by Age Group 13-15	39
1.15	Instrumental Variables Result for Dropout by Age Group 6-12	40
1.16	Instrumental Variables Result for Dropout by Age Group 13-15 . . .	41
1.17	Instrumental Variables Result for Grade Progression by Gender: Male	42
1.18	Instrumental Variables Result for Grade Progression by Gender: Female	43
1.19	Instrumental Variables Result for Grade Repetition by Gender: Male	44
1.20	Instrumental Variables Result for Grade Repetition by Gender: Female	45
1.21	Instrumental Variables Result for Dropout: Male	46

1.22	Instrumental Variables Result for Dropout: Female	47
1.23	Instrumental Variables Result for Dropout for Age 16-17: Male	48
1.24	Instrumental Variables Result for Dropout for Age 16-17: Female . .	49
2.1	Correlation Matrix	55
2.2	Average Expenditure by Household	57
2.3	Descriptive Statistics of the Study Sample	58
2.4	Machine Learning methods comparison.	67
2.5	Model Comparison.	70
2.6	Marginal Effects on food, alcohol, and smoking spending	74
3.1	Descriptive Statistics of the Adolescents	84
3.2	Descriptive Statistics of the Variables	85
3.3	Factors for the boys sample.	88
3.4	Factors for the girls sample.	89
3.5	Structural Equations Modeling Estimates.	91

List of Figures

1.1	Study Sample: number of total observations in each variable	11
1.2	Instrumental Variable Framework	15
1.3	Instrumental Variable Framework and the main two validity checks: “Strong IV” and “Exclusion Restriction”	25
2.1	Random Forests Structure. x represents the test sample input, k_i is the prediction from each tree, and k the random forests prediction that is the average of all trees’ predictions (Nguyen et al., 2013).	62
2.2	Gradient Boosting Structure. Subsequent trees are built using the resid- ual r_i from the previous tree. Adapted from Kawerk (2020).	64
2.3	Linear separating hyperplanes for the separable case. The support vec- tors are circled. Adapted from Burges (1998).	65
2.4	A Neural Network with one hidden layer.	66
2.5	Lift chart - model comparison.	71
2.6	Density plots for the propensity score.	72
2.7	Box plots for the propensity score.	72
2.8	Standardized percentage bias across covariates.	73
3.1	DAG: Girls	90
3.2	DAG: Boys	90

List of Abbreviations

ANN	Artificial Neural Networks
ATE	Average Treatment Effect
AUC	Area Under the Curve
BMI	Body Mass Index
CART	Classification and Regression Trees
CCT	Conditional Cash-Transfer
CFA	Confirmatory Factor Analysis
DAG	Directed Acyclic Graphs
EFA	Exploratory Factor Analysis
IBGE	<i>Instituto Brasileiro de Geografia de Estatística</i>
IHS	Inverse Hyperbolic Transformation
IPCA	<i>Índice Nacional de Preços ao Consumidor Amplo</i>
IV	Instrumental Variables
MART	Multiple Additive Regression Trees
OLS	Ordinary Least Squares
PeNSE	<i>Pesquisa Nacional de Saúde do Escolar</i>
POF	<i>Pesquisa de Orçamentos Familiares</i>
PSM	Propensity Score Matching
SEM	Structural Equation Modeling
SNAP	Supplemental Nutrition Assistance Program
SVM	Support Vector Machines
UPF	Ultra-Processed Foods
WHO	World Health Organization

Introduction

Conditional Cash Transfer (CCT) programs have become a popular approach to alleviating poverty in Latin American countries. In these programs, the government provides funds to households in poverty, conditional on the participant's compliance with health and education requirements (Shei et al., 2014). The Brazilian CCT program, implemented in 2003, is called *Bolsa Família* and is the largest social welfare program in the country, covering about 14 million households – a quarter of all Brazilian households. Up to now, *Bolsa Família* is the largest CCT program in the world. The objective of the program is to reduce hunger, malnutrition, poverty, and familial deprivation, while providing low-income families access to public services, such as health, education, and social assistance.

In order to be eligible to participate in this program, families should be living in extreme poverty – those with monthly income *per capita* below R\$89 (approximately US\$23¹). For these families, the benefit is a base value of R\$89 and an additional R\$41 for each child, pregnant or nursing mother, up to five. Families are also eligible if they are living under the poverty line, with monthly income *per capita* below R\$178 (approximately US\$45) if they have children under 16 years old, pregnant or nursing mothers². For these families, the benefit includes only the amount of R\$41 for each child, pregnant or nursing mother³.

In terms of magnitude, *Bolsa Família* has a great impact on the income of these families. For example, a family of two parents with three children that has a monthly

¹Exchange rate in 2019: <https://data.oecd.org/conversion/exchange-rates.htm>

²<http://dab.saude.gov.br/portaldab/ape.bfa.php>

³Up-to-date benefits a family receives in 2020: <http://www.desenvolvimentosocial.gov.br/servicos/bolsa-familia/>

income of R\$400 (a monthly income *per capita* below R\$89) would receive a benefit of R\$212, leading to an increase of more than 50% of their original monthly income.

In addition, the program is based on health and schooling conditions. The health requirement is that children under seven years old and pregnant women comply with the immunization schedule and make monitoring visits to the doctor twice a year. The schooling requirement is that children between 6 and 17 years old be enrolled in school, maintaining a minimum daily school attendance of 85% (6 to 15 years old) or 75% (16 and 17 years old).

To participate in this program, the family should first be “eligible”, complying with the requirements aforementioned. If eligible, the household should be enlisted in a national registry, namely *Cadastro Único*. The registry contains self-reported information on household demographic characteristics, household income, and prior participation in transfer programs. Although all households are free to register in *Cadastro Único*, each municipality has quotas allocated by the federal government according to a poverty assessment based on poverty maps (Brollo et al., 2017). Since officials of each municipality are responsible for the registration process, there is substantial heterogeneity across municipalities in targeting for registration (De Brauw et al., 2015).

The school attendance data is collected on a daily basis by teachers and compiled by school directors. Monthly school attendance data are sent to the Ministry of Education that makes this information available to conditionality managers in each municipality. These managers are responsible for monitoring attendance information.

Beneficiary families that do not comply with the requirements receive warnings. In the first noncompliance, the family only receives a notification. The second warning comes with a penalty of having the benefits blocked for 30 days. In the third and fourth warnings, the benefits are blocked for 60 days, and in the fifth time, the benefits are canceled and they are suspended from the program (Brollo et al., 2017).

Since one of the conditions for low-income families to participate in *Bolsa Família*

and receive the cash-transfer benefit is to comply with a schooling attendance condition, my motivation is to evaluate whether the beneficiaries' children are performing better in school. In Chapter 1, I investigate the impact of participating in the program on three schooling outcomes. In particular, I evaluate the effects on grade progression, grade repetition, and school dropout of these children who need to maintain an attendance requirement.

Studying this effect is not new in the literature. However, studies in the literature use methods that do not control for endogeneity. In this problem, in particular, reverse causality should be controlled for, since analyzing the impact of schooling outcomes on beneficiaries' children is correlated with the fact that these children are in school to comply with the requirements to be a beneficiary in the first place. Thus, my main contribution to the literature is to analyze this impact controlling for endogeneity and to discuss selection bias by comparing my results to the results in the literature that do not account for it.

This chapter is organized as follows. Sections 1.1 and 1.2 introduce the topic and present the literature on the impact of *Bolsa Família* program on schooling outcomes. Section 1.3 describes the data, the study sample, and descriptive statistics. Section 1.4 describes the Instrumental Variables methodology that is used to control for reverse causality and selection bias. Section 1.5 introduces the main results, the change in probabilities of grade progression, grade repetition, and dropping out, controlling for endogeneity, and evaluates selection bias by comparing these results to the standard OLS regression. Section 1.6 presents supplementary models and robustness checks. Section 1.7 presents further analysis, comparing the results across gender and age groups. Section 1.8 offers a discussion on the findings and conclusion.

One of the objectives of the *Bolsa Família* program is to alleviate hunger and malnutrition. The majority of the recipient families use the benefit primarily to purchase food. However, there is a national trend of increasing consumption of ultra-processed foods, which motivates me to understand if participants of *Bolsa Família* program

are also increasing their unhealthy food consumption.

To address this question, in Chapter 2, I evaluate the impact of the *Bolsa Família* program on unhealthy habits and behaviors, which is measured by the purchase of ultra-processed foods, alcohol and smoking products.

For these analyses, I use the Propensity Score Matching method that consists of two stages. In the first stage, a propensity score is estimated, which is the probability of participating in *Bolsa Família* is estimated conditional to covariates that can predict being a participant. In the second stage, the effect is estimated by matching two similar propensity scores, one from a household that participates in the program, and the other from a household that does not participate.

In the literature, propensity scores are predominantly estimated by logistic regression, which has advantages such as interpretability and accessibility. However, using this method for propensity score estimation has some drawbacks. For instance, this method relies on the linearity assumption, and it lacks accuracy when estimating high dimensional models. In this chapter, I incorporate a machine learning approach for a better predictive power of the propensity score estimation.

Recent studies are implementing machine learning techniques to develop or improve econometric models (Athey et al., 2019), and some studies have used machine learning methods to estimate propensity scores and have compared them to logistic regression. My contribution is the inclusion of these methods, namely, Random Forests, Gradient Boosting, Support Vector Machines, and Neural Networks, for propensity score estimation, and compare their predictive power in the context of my analysis. Moreover, I improve the estimation of three effects (extensive margin, intensive margin, and the overall effect) to better understand the impact of *Bolsa Família* on unhealthy purchases, contributing to the literature of public policy and the impact of machine learning on economics.

This chapter is organized as follows. Section 2.1 introduces the topic and section 2.2 presents literature on the food consumption among *Bolsa Família* participants.

Section 2.3 presents literature on the rise of ultra-processed foods and unhealthy consumption in Brazil. Section 2.4 describes the data, the study sample, and the variables used in this study. Section 2.5 presents descriptive statistics of the sample data. Section 2.6 presents the methodology for these analyses, and sections 2.7 and 2.8 the models used for propensity score estimation. Section 2.9 presents the results and model validation. Section 2.10 concludes.

In Brazil, adolescents are increasing their chances of engaging in risk behaviors, such as unsafe sex, domestic violence, and gunfights involvement. Literature finds associations among negative body image perception, depression, and risk behaviors, but such studies were not found in the Brazilian adolescent context.

In Chapter 3, I use Directed Acyclic Graphs and Structural Equations Modeling to identify the direct and indirect effects of negative body image perception, depression, and risk behaviors among adolescents in Brazil. Using Exploratory and Confirmatory Factor Analysis, constructs found for risk behaviors are aggression, illegal substance use, and sexual behavior.

Directed Acyclic Graphs (DAG) are useful in cases where there is a lack of theory. This approach is based on conditional independence relations that satisfied a Causal Markovian Condition, which is an assumption made in Bayesian probability theory. DAG estimates directional links among variables or constructs, followed by the estimation of a Structural Equation Model based on directional links indicated by the DAG.

This paper contributes to the literature by shedding light on the vulnerability of Brazilian adolescents and their propensity to risk behaviors. Also, I find associations that were barely or even not studied in the literature, suggesting future research to investigate these relationships further. Furthermore, the analytical approach contributes to the literature since I present Directed Acyclic Graphs as a reliable model. My findings are consistent with those in the literature that are based on theory.

This chapter is organized as follows. Section 3.1 introduces the topic and literature

on associations among body dissatisfaction, depression, and risk behaviors. Section 3.2 presents the methodology for the proposed analyses. Section 3.3 describes the data and section 3.3.1 the variables and descriptive statistics. Section 3.4 presents the results of Exploratory and Confirmatory Factor Analyses, Directed Acyclic Graphs, and Structural Equations Modeling. Section 3.5 discusses the findings and section 3.6 concludes.

Chapter 1

Unintended Consequences of Welfare Programs in Schooling Outcomes: Evidence from Brazil

1.1 Introduction

The *Bolsa Família* program is a conditional cash-transfer program of the Brazilian Government and the largest in the world, with more than 14 million beneficiary households. This program was introduced in 2003 with the objective of reducing hunger and poverty while promoting the emancipation of families in situations of greater poverty in the country.¹

As a conditional program, in order to participate and receive benefits, families must comply with some requirements. One of the requirements to participate in this program is that children should be in school. Given this requirement, my motivation for this study is to understand how receiving government aid impacts the schooling outcomes of these children. Therefore, in this chapter, I evaluate the impact of Brazil's *Bolsa Família* program on schooling outcomes, in particular, dropout rates, grade progression, and grade repetition. The main question of investigation is what happens to the probability of dropping out of school, repeating a grade, or progressing in school when the family participates in the *Bolsa Família* cash-transfer program.

While the existing literature has formally addressed the impacts of dropping out of

¹<http://bolsafamilia.datasus.gov.br>

school, grade progression and other educational outcomes, these studies have predominantly used more traditional statistical modeling techniques, such as Linear Regression (e.g., Simões and Sabates, 2014) or Propensity Score Matching (e.g., Schaffland, 2011). In the current study, I address these limitations through the application of the Instrumental Variables method to reconsider the question.

1.2 Schooling Outcomes

One of the main problems in Latin America is inequality. According to the theory of inequality and intergenerational mobility of Becker and Tomes (1979), human capital is an important factor to reduce inequality because families maximize their utility function by investing in the human capital of their children.

In a study of intergenerational mobility in Latin America and the United States, Behrman et al. (2001) show that not only is mobility much higher in the United States, but economic growth does not improve Latin American countries' mobility. Their results show that, in these countries, the differences are associated with an investment in schooling. Thus, it is important to evaluate how the *Bolsa Família* program in Brazil has impacted children's schooling outcomes since these are some of the most significant predictors of later-life economic success.

Several researchers have studied the impact of *Bolsa Família* on educational and schooling outcomes. However, there is an analytical gap in the literature, since these studies use methods that do not address endogeneity. For instance, using data from 2004 and 2006, Schaffland (2011) uses propensity score matching and finds that the probability of school enrollment rises by around 4% for recipients' children. Her results also show a positive impact on school attendance among recipients' children, although this impact only happens in a short-term period.

Simões and Sabates (2014) use OLS regressions and find that the program contributes to improvements in school performance in test scores, pass-grades, and reduces the dropout rates for 4th-grade children, using data from 2005 to 2007. Glewwe

and Kassouf (2012) use census data from 1998 to 2005 (accounting for information on *Bolsa Escola*, a predecessor program to *Bolsa Família*). They find that the program increased school enrollment by 5.5-6.5%, reduced dropout rates by 0.4-0.5% and increased grade promotion rates by 0.3-0.9%. De Brauw et al. (2015) use the propensity score method to assess the impact of this program on schooling outcomes of children between 6 and 17 years old, using data from the years 2005 and 2009. They find that the program increases school participation by 8% and grade progression by 10% among girls.

The literature on the impact of *Bolsa Família* on schooling outcomes shows consistent results that are favorable to the beneficiaries of the program, for similar time-periods (at or around the year 2005). However, most of the methods applied are OLS regression or propensity score matching that do not control for selection bias. Here I propose analyzing the impact of this program on schooling outcomes through the Instrumental Variables method, which controls for unobserved selection and reverse causality (Wooldridge, 2010).

The objective of this study is to understand the real impact of *Bolsa Família* on schooling outcomes, addressing endogeneity. More specifically, I investigate what is the probability of students dropping out of school, progressing, or repeating a grade when they participate in the cash transfer program in Brazil.

1.3 Data

Most studies in the literature use data around the year 2005. In order to have comparable results and identify endogeneity, the data selected for the present study are from 2005. The data arise from a survey conducted by the *Centro de Desenvolvimento e Planejamento Regional (Cedeplar)*, commissioned by the Federal Ministry of Social Development in Brazil. The survey includes household-level questions on demographics, living conditions, assets, income, consumption, anthropometry, health, education, and participation in cash transfer and subsidy programs. A total of 15,426 households

distributed in 24 federal units were surveyed between October and December of 2005².

1.3.1 Study Sample

In order to select the appropriate sample for this study, I used the individual level version of the household data, which contains 68,392 observations. From this dataset, I selected only children from ages 6 to 17, obtaining a total of 22,457 observations. After data cleaning, my final dataset consists of 14,803 observations, each a child from 6 to 17 years old.

Variables

In order to answer the proposed question, I analyze three schooling indicators for children from 6 to 15 years old³, namely, dropout, grade progression, and grade repetition.

These three variables are dummy variables created based on a conditional measure on a child attending school in the year of the study (and which grade he or she attends) and in the previous year (and which grade he or she attended then). The child is considered a dropout if he or she attended school in the previous year but is not attending now. Grade progression means that the child attended a certain grade in the previous year and is attending the next grade in the current year. Grade repetition means that the child attended a certain grade in the previous year and is attending the same grade in the current year (failed in school). These three variables are mutually exclusive, i.e., a student can be a dropout, progressed in school, or repeated a grade.

The sample includes variables indicating whether the household participates in the *Bolsa Família* program (“Recipients”) and whether the household is registered in the Brazilian registry *Cadastro Único* (“Registered”).

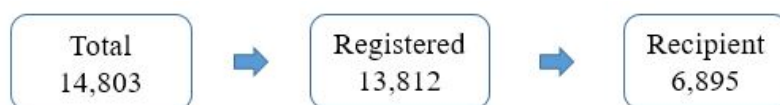
²<http://www.cedeplar.ufmg.br/pesquisas/projetos-concluidos/136-projeto-bolsa-familia>

³In 2005, to be eligible, children from 6 to 15 years old had to be enrolled in school. In 2009, this rule changed to include children from 6 to 17 years old. Source: www.planalto.gov.br/ccivil_03/_Ato2007-2010/2009/Decreto/D6917.htm.

Families that are eligible to participate in the program and comply with the requirements should register in a National Registry called *Cadastro Único* in order to be selected to participate in the *Bolsa Família* program. However, not all registered households are selected because each municipality has a quota. Since there are differences in quotas, it is likely that between two similar households in different municipalities, one is a recipient of the cash-transfer and the other is not.

Figure 1.1 shows the total number of children in the total sample, the number of children that are in registered households, and the number of children that are in a recipient household.

Figure 1.1: Study Sample: number of total observations in each variable



The sample also contains individual variables such as age, gender, race, geographic region, and whether they live in urban or rural areas.

1.3.2 Descriptive Statistics

A summary of the descriptive statistics of the variables used in the analysis is shown in Table 1.1. The number of children from 6 to 17 years old in the sample is 14,803. From the total sample (Total Mean column), 74.1% of students progressed (out of a total of 12,208), 11.9% repeated the grade (out of a total of 12,208), and 3.1% dropped out of school (out of a total of 14,796)⁴. The proportion of girls is 47%, and the most prevalent self-reported races are white (46%) and multiracial (45%). The most populated region is Southeast (44%), followed by North (30%) and Northeast (26%), with 83% living in an urban area.

⁴In the sample I have more observations for students who dropped out since to create this variable I could also use the variables “attend school this year” and “attend school last year” if there were missing cases for the grade attendance.

Table 1.1 also contains statistics among those who are registered (N=13,802) and those who are recipient (N=6,895). In the registered group, 70.7% of students progressed, 15.1% repeated the grade, and 3.3% dropped out of school. The rates of repeating and dropping out are higher among the registered group than the total sample, while the rate of grade progression is lower. The most prevalent races are multiracial (51%) and white (38%). Comparing to the total sample, there are fewer white people and more black and multiracial who are eligible and registered in *Cadastro Único*. There are also more registered living in the Northeast (35%) region, which consists of the poorer states in the country, while in the Southeast, the region with the richest region responsible for about 60% of the country's GDP, the rate of registered people is 33% versus 44% of the total sample. There are more registered people living in rural areas (24% versus 17% in the total sample).

In the recipient group, the statistics are very similar to the registered group. The rates of grade progression, grade repetition, and dropout are 71.6%, 16.4%, and 3.2% respectively. The percentage of multiracial (52%) and black (10%) are also higher compared to the total sample, while white is lower (36% versus 46%). They also live mostly in the Northeast region (39%), and also they live more in rural areas (26%) compared to the total sample.

1.4 Methodology

Ideally, to measure the effect of a program on the schooling outcomes, a randomized controlled experiment should be conducted, in which a sample of households would be divided into two groups: treatment and control. The household in the treatment group would participate in the cash-transfer program, and we could estimate the causal effect of this program on schooling outcomes of the children in these families. Since this particular examination is not feasible, we rely on a natural experiment.

The method chosen to analyze this causal effect is the Instrumental Variables (IV) method. Simulating a randomized experiment, IV accounts for all selection bias,

Table 1.1: Descriptive Statistics

Variables	Total Mean	Registered Mean	Recipient Mean
Grade progression	.741	.707	.716
Grade repetition	.119	.151	.164
Dropout	.031	.033	.032
Age			
6	.09	.10	.10
7	.09	.09	.09
8	.10	.10	.10
9	.09	.09	.09
10	.08	.10	.10
11	.07	.08	.10
12	.09	.09	.10
13	.06	.07	.07
14	.08	.07	.07
15	.07	.07	.07
16	.08	.06	.06
17	.09	.06	.05
Female	.47	.48	.48
Race			
White	.46	.38	.36
Black	.08	.10	.10
Multiracial	.45	.51	.52
Asian	.01	.01	.01
Indigenous	.00	.00	.00
Undeclared	.01	.01	.00
Region			
North	.30	.32	.30
Northeast	.26	.35	.39
Southeast	.44	.33	.32
South	.15	.17	.17
Center	.07	.07	.06
Urban	.83	.76	.74

including reverse causality (Angrist and Pischke, 2008). This method estimates via two-stage least squares (2SLS). The first-stage model estimates the probability of the household to be a recipient of *Bolsa Família* program, then the probability of the

schooling outcome (dropping out, grade repetition, and grade progression) is predicted based on the first stage.

$$Y_{it} = \hat{\beta}_0 + \hat{\beta}_1 T_{it} + \hat{\beta}_2 X_{it} + \varepsilon_{it} \quad (1.1)$$

Equation 1.1 shows the structural model. The parameter $\hat{\beta}_1$ captures the effect of the treatment variable Recipient T_{it} on the schooling outcomes represented by Y_{it} , adjusting for covariates, X_{it} . This is what I want to analyze, however, this model is susceptible to endogeneity. Therefore, this model cannot be used, and I use a two-stage model that relies on an instrumental variable.

In order to understand the causal effect of being a recipient of *Bolsa Família* (explanatory variable) on schooling outcomes (dependent variables), it is necessary to have a valid instrument variable that affects the explanatory variable but not the dependent variable. The instrument selected is “Registered” - those who are eligible and registered in *Cadastro Único*. Families registered in *Cadastro Único* fulfill the requirements to apply for the program, but not all of these families get to participate in the program to receive the cash-transfer. Also, since registration in *Cadastro Único* is voluntary, a self-selection bias into the program is possible. This bias can be controlled by the Instrumental Variables methodology.

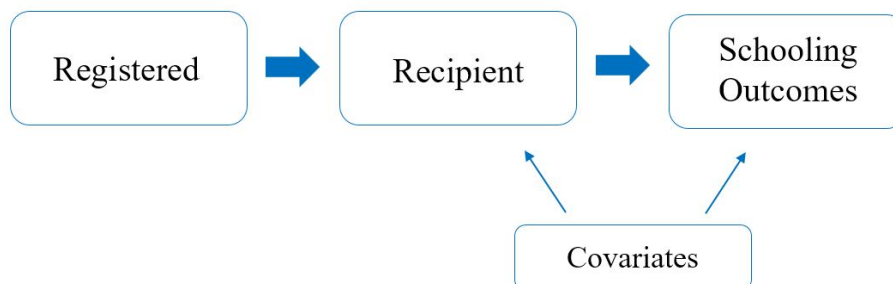
Equation 1.2 shows the first-stage model of the IV method. The parameter $\hat{\alpha}_1$ captures the first-stage effect of the instrument Registered Z_{it} on the predicted value of Recipient \hat{T}_{it} , adjusting for covariates, X_{it} . Equation 1.3 shows the second-stage model of the IV method. The parameter $\hat{\beta}_1$ captures the effect of the predicted value of Recipient \hat{T}_{it} on the schooling outcomes Y_{it} , adjusting for covariates, X_{it} .

$$\hat{T}_{it} = \hat{\alpha}_0 + \hat{\alpha}_1 Z_{it} + \hat{\alpha}_2 X_{it} + \mu_{it} \quad (1.2)$$

$$Y_{it} = \hat{\beta}_0 + \hat{\beta}_1 \hat{T}_{it} + \hat{\beta}_2 X_{it} + \gamma_{it} \quad (1.3)$$

Figure 1.2 shows a framework of the Instrumental Variable method. A valid instrument should be correlated to the endogenous variable (“Recipient” – those receiving the benefit). Since in order to participate in the program, the family should be first registered in *Cadastro Único*, it is clear that these variables are correlated. Additionally, the instrument should not be directly correlated with the schooling outcomes. Since any family can apply to *Cadastro Único*, given that they comply with the conditions mentioned previously, but they are not benefiting from the program yet, there should not be a direct correlation between being listed in a registry and the children’s schooling outcomes.

Figure 1.2: Instrumental Variable Framework



The IV model should control for covariates that could possibly predict the explanatory variable. Here, the covariates that could predict being a recipient are children’s age (from 6 to 15 years old), gender, race (white, black, multiracial, Asian, indigenous, or undeclared), region (North, Northeast, Southeast, South, and Center), and whether the child lives in an urban or a rural area.

$$Y_{it} = \hat{\Pi}_0 + \hat{\Pi}_1 Z_{it} + \hat{\Pi}_2 X_{it} + \omega_{it} \quad (1.4)$$

Furthermore, I also analyze the reduced-form model as a check for the IV method. Equation 1.4 shows that the parameter $\hat{\Pi}_1$ captures the effect of the instrument Reg-

istered Z_{it} on the schooling outcomes represented by Y_{it} , adjusting for covariates, X_{it} . The parameter $\hat{\Pi}_1$ should have the same direction and significance as the parameters of the IV model.

1.5 Results

In this section, I present the results of Instrumental Variable method for the schooling outcomes grade progression, grade repetition, and dropout. To estimate the regressions, the instrument “Registered” was used, the regressions were clustered at the household level to avoid the spillover effect onto siblings (an autocorrelation problem), and the sample weight was used as control (not shown)⁵.

1.5.1 OLS Regression

To have an idea of the association between being a *Bolsa Família* participant and the outcomes, first I present the OLS regression estimates (Table 1.2). The probability of grade progression decreases by 0.5 percentage points (pp) and it is not significant, the probability of grade repetition increases by 1.8 pp, and the probability of dropping out decreases by 0.6 pp among the program participants who are between 6 and 15 years old, controlling by age, gender, race, region, and area.

The probability of grade progression increases by 6 pp among girls, decreases among black and multiracial groups when comparing to the white (reference) group. It increases in the Southeast region by 4 pp comparing to the North (reference) and increases by 7 pp among those in an urban area.

The probability of grade repetition decreases by 5 pp among girls, increases among black and multiracial groups when comparing to the white (reference) group. It decreases in the Southeast region and increases in the South and Center compared to the North (reference) and decreases by 4 pp among those in an urban area.

⁵Weighted estimates are similar to not weighted. It was included as a control to avoid possible bias, since the design weight was calculated to compensate for over- and under-sampling of geographical regions, and region is a variable already being used as a control in the models.

The probability of dropping out does not change by gender, race, or area type. It decreases in the Southeast and Center regions comparing to the North (reference).

Table 1.2: OLS Estimates of Grade Progression, Grade Repetition and Dropout

	Grade Prog. (SE)	Grade Rep. (SE)	Dropout (SE)
Recipient	-.005(.009)	.018*(.008)	-.006*(.003)
Age			
6	Ref.	Ref.	Ref.
7	.45* (.02)	.05* (.02)	.00 (.00)
8	.63* (.02)	.06* (.02)	.00 (.00)
9	.69* (.02)	.05* (.02)	-.01* (.00)
10	.69* (.02)	.03 (.02)	.00 (.00)
11	.71* (.02)	.02 (.02)	.00 (.00)
12	.70* (.02)	.03 (.02)	.00 (.00)
13	.69* (.02)	.03 (.02)	.01 (.01)
14	.67* (.02)	.03 (.02)	.02* (.01)
15	.60* (.02)	.04* (.02)	.05* (.01)
Female	.06* (.01)	-.05* (.01)	.00 (.00)
Race			
White	Ref.	Ref.	Ref.
Black	-.07* (.02)	.05* (.02)	.00 (.00)
Multiracial	-.03* (.01)	.02* (.01)	.00 (.00)
Asian	-.06 (.07)	.05 (.05)	.01 (.02)
Indigenous	-.04 (.09)	-.01 (.06)	.01 (.03)
Undeclared	-.22* (.06)	.19* (.05)	.01 (.02)
Region			
North	Ref.	Ref.	Ref.
Northeast	.01 (.01)	.02 (.01)	.00 (.00)
Southeast	.04* (.01)	-.02* (.01)	-.01* (.00)
South	-.04 (.02)	.05* (.02)	-.01* (.01)
Center	-.01 (.02)	.03* (.02)	-.02* (.00)
Urban	.07* (.01)	-.04* (.01)	.00 (.00)

* denotes that the estimate is statistically significant at .05.

However, these estimates are likely to be biased because it does not account for reverse causality and unobserved variables. I hypothesize that there is selection bias

and it is positive. The reasoning is that it is likely that the unobserved variables that increase the probability of a family to be a participant of *Bolsa Família* also increase the probability for their children to repeat a grade or to drop out of school.

Another source of selection bias is self-selection into the program. Once a family is eligible to participate, they have to be aware of the program and voluntarily go to *Cadastro Único* registration center in their municipality. This motivates further analysis and understanding of the causal relationship and the direction of possible selection bias.

1.5.2 Grade Progression

Using Instrumental Variable method, the first finding is that the probability of grade progression decreases by 10.5 pp among students who belong to a family that participates in the *Bolsa Família* program (Table 1.3, Second Stage). Also, the probability of participating in the program given that the household is registered increases by 51.6 percentage points (pp) (Table 1.3, First Stage).

Analyzing the covariate variables (standard errors in parentheses), the probability of grade progression is about 60–71 pp higher for all ages comparing to age 6 (reference group). This result makes sense since a child entering school for the first time is not progressing in school, therefore most of the children at this age are considered as “not progressing”. Comparing to white students (reference group), black students have a lower probability of progressing by 6 pp, and multiracial students by 2 pp. The probability of grade progression among girls is 6 pp higher than boys. Comparing to the North region, students in the Southeast have a higher probability of progressing by 4 pp, and those in rural areas have a probability of 6 pp lower comparing to those in urban areas.

We can observe that the reduced form model is consistent with the IV model; the effect of the instrument is significant and has the same direction as the outcome variable in the IV model.

Table 1.3: Instrumental Variables Result for Grade Progression

Dep Var	First Stage (Recipients)	Second Stage (Grade Prog)	Reduced Form (Grade Prog)
Registered Recipient	.516*(.012)	-	-.054*(.019)
Age			
6	Ref.	Ref.	Ref.
7	-.03(.03)	.45*(.02)	.45*(.02)
8	-.03(.02)	.63*(.02)	.63*(.02)
9	-.01(.02)	.68*(.02)	.69*(.02)
10	.00(.02)	.69*(.02)	.69*(.02)
11	.01(.02)	.71*(.02)	.71*(.02)
12	.00(.02)	.70*(.02)	.70*(.02)
13	.01(.02)	.69*(.02)	.69*(.02)
14	-.03(.02)	.67*(.02)	.67*(.02)
15	-.02(.03)	.60*(.02)	.60*(.02)
Female	.00(.01)	.06*(.01)	.06*(.01)
Race			
White	Ref.	Ref.	Ref.
Black	.04(.03)	-.06*(.02)	-.07*(.02)
Multiracial	.03*(.01)	-.02*(.01)	-.03*(.01)
Asian	.10(.07)	-.05(.06)	-.06(.07)
Indigenous	-.07(.15)	-.04(.10)	-.04(.09)
Undeclared	-.13(.08)	-.23*(.06)	-.22*(.06)
Region			
North	Ref.	Ref.	Ref.
Northeast	.06*(.02)	.02(.01)	.01(.01)
Southeast	.02(.02)	.04*(.01)	.04*(.01)
South	.04(.03)	-.03(.02)	-.04(.02)
Center	.04(.03)	-.01(.02)	-.01(.02)
Urban	-.09*(.02)	.06*(.01)	-.07*(.01)

* denotes that the estimate is statistically significant at .05.

To evaluate the selection bias, I compare the effects from the IV model (-10.5 pp), the “true” causal effect, with OLS estimation (-0.5 pp), the estimated value. Equation 1.5 shows the estimation of the selection bias direction.

The parameter β refers to the causal effect and $E(\hat{\beta})$ the estimated OLS value. For

grade progression, the selection bias is positive. It means that unobserved variables that make the household more likely to participate in the program are causing the children to progress more.

$$E(\hat{\beta}) = \beta + \text{selection bias} \quad (1.5)$$

1.5.3 Grade Repetition

The IV model results show that the probability of repeating a grade increases by 8.4 pp among students who belong to a family that participates in the *Bolsa Família* program (Table 1.4, Second Stage).

Analyzing the covariate variables (standard errors in parentheses), the probability of grade repetition is about 4–6 pp higher for all ages comparing to age of 6 (reference group). Students in the first years of schooling are more likely to repeat the grade comparing to the very first year at the age 6. Comparing to white students (reference group), black students have a higher probability of repeating a grade by 5 pp, and multiracial students by 2 pp. The probability of grade repetition among girls is 5 pp lower than boys. Comparing to the North region, students in the Southeast have a lower probability of repeating by 2 pp, and those in rural areas have a probability of 4 pp higher comparing to those in urban areas.

Evaluating the selection bias using equation 1.5, I compare the effects from the IV model (8.4 pp), the “true” causal effect, with the OLS estimation (1.8 pp), the estimated value. The selection bias is negative, meaning that unobserved variables that make the household more likely to participate in the program are causing the children to repeat less.

1.5.4 Dropout

Performing the same analysis to evaluate the impact of the program on dropping out of school, this probability decreases by 0.3 percentage points (pp) among students who

Table 1.4: Instrumental Variables Result for Grade Repetition

Dep Var	First Stage (Recipients)	Second Stage (Grade Rep)	Reduced Form (Grade Rep)
Registered Recipient	.516*(.012)	-	.043*(.015)
Age		.084*(.029)	-
6	Ref.	Ref.	Ref.
7	-.03(.03)	.05*(.02)	.04*(.02)
8	-.03(.02)	.06*(.02)	.06*(.02)
9	-.01(.02)	.05*(.02)	.05*(.02)
10	.00(.02)	.03(.02)	.03(.02)
11	.01(.02)	.02(.02)	.02(.02)
12	.00(.02)	.03(.02)	.03(.02)
13	.01(.02)	.03(.02)	.03(.02)
14	-.03(.02)	.03(.02)	.03(.02)
15	-.02(.03)	.04*(.02)	.04*(.02)
Female	.00(.01)	-.05*(.01)	-.05*(.01)
Race			
White	Ref.	Ref.	Ref.
Black	.04(.03)	.05*(.02)	.05*(.02)
Multiracial	.03*(.01)	.02*(.01)	.02*(.01)
Asian	.10(.07)	.04(.05)	.05(.05)
Indigenous	-.07(.15)	-.01(.07)	-.01(.06)
Undeclared	-.13(.08)	.20*(.05)	.19*(.05)
Region			
North	Ref.	Ref.	Ref.
Northeast	.06*(.02)	.02(.01)	.02(.01)
Southeast	.02(.02)	-.02*(.01)	-.02*(.01)
South	.04(.03)	.04(.02)	.05*(.02)
Center	.04(.03)	.03(.02)	.03*(.02)
Urban	-.09*(.02)	-.04*(.01)	-.04*(.01)

* denotes that the estimate is statistically significant at .05.

belong to a family that participates in the *Bolsa Família* program and are between 6 and 15 years old (Table 1.5, Second Stage). However, this result is not significant.

Analyzing the covariate variables (standard errors in parentheses), the probability of dropping out of school is significant only for the age of 15, which is 5 pp higher

Table 1.5: Instrumental Variables Result for Dropout

Dep Var	First Stage (Recipients)	Second Stage (Dropout)	Reduced Form (Dropout)
Registered Recipient	.51*(.011)	-	-.001(.007)
Age			
6	Ref.	Ref.	Ref.
7	.02(.02)	.00(.00)	.00(.00)
8	-.02(.02)	.00(.00)	.00(.00)
9	.00(.02)	-.01(.00)	-.01(.00)
10	.02(.02)	.00(.00)	.00(.00)
11	.02(.02)	.00(.00)	.00(.00)
12	.00(.02)	.00(.00)	.00(.00)
13	.03(.02)	.01*(.01)	.01*(.00)
14	-.02(.02)	.02*(.01)	.02*(.01)
15	.00(.02)	.05*(.01)	.05*(.01)
Female	.00(.01)	.00(.00)	.00(.00)
Race			
White	Ref.	Ref.	Ref.
Black	.03(.02)	.00(.00)	.00(.00)
Multiracial	.02(.01)	.004*(.00)	.004*(.00)
Asian	.09(.07)	.01(.02)	.01(.02)
Indigenous	.06(.14)	.01(.03)	.01(.03)
Undeclared	-.12(.08)	.01(.02)	.01(.02)
Region			
North	Ref.	Ref.	Ref.
Northeast	.05*(.02)	.00(.00)	.00(.00)
Southeast	.01(.02)	-.01(.00)	-.01(.00)
South	.05(.03)	-.01(.01)	-.01(.01)
Center	.03(.02)	-.02(.00)	-.02(.00)
Urban	-.09*(.02)	.00(.00)	.00(.00)

* denotes that the estimate is statistically significant at .05.

comparing to the age of 6. This might happen since they are at their last age of compliance with the requirements of the program, soon to be out of the range, so they just drop out instead of finishing school. There is no region or gender difference in the probability of dropping out of school, and the only significant difference for race

is among multiracials – the probability of dropping out increases by .4 pp comparing to white students (reference group).

Evaluating the selection bias using equation 1.5, I compare the effects from the IV model (-.3 pp), the “true” causal effect, with OLS estimation (-.6 pp). For dropout, the selection bias is also negative, meaning that the unobserved variables that make the household more likely to participate in the program are causing the children to drop out of school less.

Dropout among older students

I analyzed the probability of dropping out among students from beneficiary families who are 16 and 17 years old and therefore were out of the range of compliance in 2005. Among these students, the probability of dropping out increases by 9.3 pp, although it is also not significant (Table 1.6). The probability of dropping out is higher for those 17 years old by 4 pp. There is no significant difference among regions or gender. For the race, the probability of dropping out among indigenous decrease by 13 pp and among undeclared race decrease by 12 pp comparing to the reference group (white).

Although the probability of dropping out is not significant, there is a considerable difference between 16-17 years old compared to those 6-15. One reason is that these students did not need to be in school to comply with the conditions of the program. Another explanation is that these are older and stronger children that bring immediate benefits to their families being part of the workforce rather than staying at school.

1.5.5 Validity of the Instrument

There are two requirements to evaluate the validity of an instrument. First, it is necessary to check its exclusion restriction - the instrument variable Registered should not be directly correlated with the schooling outcomes⁶. Since there is no reason for a child whose family is registered in *Cadastro Único* and for this child to progress or

⁶For the validity analysis, I am using the schooling outcomes grade progression and grade repetition, since dropout is not significant.

Table 1.6: Instrumental Variables Result for Dropout for Ages 16 and 17

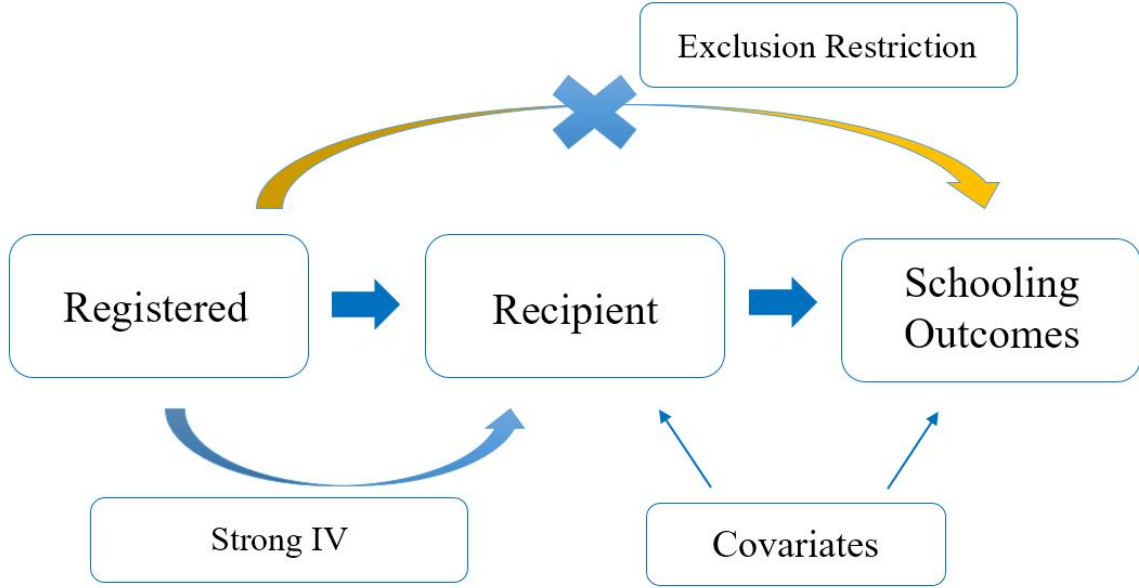
Dep Var	First Stage (Recipients)	Second Stage (Dropout)	Reduced Form (Dropout)
Registered	.467*(.016)	-	.043(.026)
Recipient	-	.093(.055)	-
Age			
16	Ref.	Ref.	Ref.
17	-.03(.02)	.04*(.01)	.03*(.01)
Female	.02(.02)	-.03(.01)	-.02(.01)
Race			
White	Ref.	Ref.	Ref.
Black	.03(.04)	-.02(.03)	-.02(.03)
Multiracial	.03(.03)	.01(.02)	.01(.02)
Asian	-.08(.11)	.13(.11)	.13(.11)
Indigenous	.29(.23)	-.13*(.04)	-.10*(.02)
Undeclared	.04(.12)	-.12*(.02)	-.11*(.02)
Region			
North	Ref.	Ref.	Ref.
Northeast	.06(.03)	-.01(.02)	-.01(.02)
Southeast	.00(.03)	-.01(.02)	-.01(.02)
South	.08(.06)	-.01(.04)	.00(.04)
Center	-.01(.04)	-.03(.03)	-.03(.03)
Urban	-.09*(.03)	.01(.02)	.00(.02)

* denotes that the estimate is statistically significant at .05.

fail in school, it can be assumed that there is no direct relationship between these variables, and therefore the exclusion restriction requirement can be verified.

The other requirement for an instrument is that it should be powerful (“strong IV”). Being registered at *Cadastro Único* is a powerful predictor of being a recipient, since in order to receive the cash transfer one needs to be registered in the first place. Therefore, the IV (Registered) should be strongly correlated with the endogenous variable (Recipient), which is measured by the F-test=1926, exceeding the conventional minimum standard of power of F-test=50 in the first stage. Figure 1.3 shows the framework of these two requirements.

Figure 1.3: Instrumental Variable Framework and the main two validity checks: “Strong IV” and “Exclusion Restriction”



Another test to check if the instrument is strong is to analyze the standard errors between OLS and IV models. For grade progression, the standard error of the treatment variable Recipient in the IV model is 0.037 (Table 1.3, Second Stage) and in the OLS model is 0.009 (Table 1.2). For grade repetition, the standard error of the treatment variable Recipient in the IV model is 0.029 (Table 1.4, Second Stage) and in the OLS model is 0.008 (Table 1.2). The standard error in both IV regressions is about 4 times higher than OLSs', which is under the benchmark for the magnitude of 10 times higher, when a concern of “weak IV” arises.

Additionally, it is necessary to test for endogeneity, which is measured by the robust regression F-test=7.42 ($p=0.01$) for grade progression and F-test=5.24 ($p=0.02$) for grade repetition. Both tests are significant at the 5% level, meaning that the null hypothesis is rejected, therefore, the treatment variable is endogenous. Thus, the recipient variable is endogenous and IV modeling is necessary.

Finally, the reduced-form models are estimated based on the IV models, which

are also significant (Tables 1.3 and 1.4, Reduced Form). The coefficients and their directions are consistent with the second-stage results from the IV model.

1.6 Supplementary Models

As a robustness check, I perform the analysis limiting the sample to families with lower education. I define a low educated family as a family in which the head of the household studied up to high school. In this sample, 82% of the households fit this category.

Tables 1.7, 1.8, and 1.9 present the result of these analyses. The probability of grade progression among lower education families decreases by 7.9 pp which is consistent with the decrease of 10.5 pp in the full sample. The probability of grade repetition increases by 7.5 pp, consistent with the increase of 8.4 pp in the full sample. The probability of dropping out among lower education families decreases by 1.1 pp, although it is not significant. This result is also consistent with the non-significant decrease by .3 pp in the full sample.

Another robustness check is the estimation of the schooling outcomes among registered households via OLS regression. Table 1.10 shows that these results are consistent to the IV results in relation to the direction, but not significance (Tables 1.3, 1.4, 1.5). This shows that being a recipient is significant to the effects of progression, repetition, and dropout relative to just being registered. Therefore, “recipient” is a truly exogenous variable within the registered group.

1.7 Further Analysis

In this section, I show the results of further analyses to explore if they are heterogeneous across different margins. I analyze differences in grade progression, grade repetition, and dropping out by age groups and gender.

Table 1.7: Instrumental Variables Result for Grade Progression: Low Education

Dep Var	First Stage (Recipients)	Second Stage (Grade Prog)	Reduced Form (Grade Prog)
Registered Recipient	.517*(.013)	-	-.041(.023)
Age	-	-.079(.044)	-
6	Ref.	Ref.	Ref.
7	-.01(.03)	.47*(.03)	.47*(.02)
8	-.03(.03)	.64*(.02)	.64*(.02)
9	-.01(.03)	.70*(.02)	.70*(.02)
10	.01(.03)	.71*(.02)	.71*(.02)
11	.00(.03)	.71*(.02)	.72*(.02)
12	.00(.03)	.70*(.02)	.70*(.02)
13	.01(.03)	.69*(.02)	.69*(.02)
14	-.04(.03)	.66(.02)	.67*(.02)
15	-.02(.03)	.60*(.02)	.60*(.02)
Female	.00(.01)	.05*(.01)	.05*(.01)
Race			
White	Ref.	Ref.	Ref.
Black	.04(.03)	-.05*(.02)	-.06*(.02)
Multiracial	.03(.02)	-.03*(.01)	-.03*(.01)
Asian	.10(.08)	-.06(.07)	-.07(.07)
Indigenous	-.17(.18)	-.12(.12)	-.10(.09)
Undeclared	-.09(.09)	-.18*(.06)	-.18*(.06)
Region			
North	Ref.	Ref.	Ref.
Northeast	.06*(.02)	.01(.02)	.01(.02)
Southeast	.02(.02)	.04*(.01)	.04*(.01)
South	.04(.04)	-.04(.03)	.04(.02)
Center	.07*(.03)	-.01(.02)	-.01(.02)
Urban	-.07*(.02)	.05(.02)	.05*(.01)

* denotes that the estimate is statistically significant at .05.

Table 1.8: Instrumental Variables Result for Grade Repetition: Low Education

Dep Var	First Stage (Recipients)	Second Stage (Grade Rep)	Reduced Form (Grade Rep)
Registered Recipient	.517*(.013)	-	.039*(.018)
Age	-	.075*(.034)	-
6	Ref.	Ref.	Ref.
7	-.01(.03)	.04*(.02)	.04*(.02)
8	-.03(.03)	.06*(.02)	.06*(.02)
9	-.01(.03)	.05*(.02)	.05*(.02)
10	.01(.03)	.03(.02)	.03(.02)
11	.00(.03)	.02(.02)	.02(.02)
12	.00(.03)	.04*(.02)	.04*(.02)
13	.01(.03)	.04*(.02)	.04*(.02)
14	-.04(.03)	.04(.02)	.03(.02)
15	-.02(.03)	.04*(.02)	.04*(.02)
Female	.00(.01)	-.04*(.01)	-.05*(.01)
Race			
White	Ref.	Ref.	Ref.
Black	.04(.03)	.04*(.02)	.04*(.02)
Multiracial	.03(.02)	.03*(.01)	.03*(.01)
Asian	.10(.08)	.03(.05)	.04(.05)
Indigenous	-.17(.18)	.06(.10)	.04(.09)
Undeclared	-.09(.09)	.21*(.06)	.20*(.06)
Region			
North	Ref.	Ref.	Ref.
Northeast	.06*(.02)	.02(.01)	.02(.01)
Southeast	.02(.02)	-.03(.01)	-.03*(.01)
South	.04(.04)	.04(.02)	.04*(.02)
Center	.07*(.03)	.03(.02)	.03(.02)
Urban	-.07*(.02)	-.03*(.01)	-.03*(.01)

* denotes that the estimate is statistically significant at .05.

Table 1.9: Instrumental Variables Result for Dropout: Low Education

Dep Var	First Stage (Recipients)	Second Stage (Dropout)	Reduced Form (Dropout)
Registered Recipient	.513*(.012)	-	-.006(.010)
Age	-	-.011(.019)	-
6	Ref.	Ref.	Ref.
7	.02(.02)	.00(.01)	.00(.01)
8	-.02(.02)	.00(.01)	.00(.01)
9	.00(.02)	-.01(.00)	-.01(.00)
10	.03(.02)	.00(.01)	.00(.01)
11	.02(.02)	.00(.01)	.00(.01)
12	.01(.02)	.00(.01)	.00(.01)
13	.03(.02)	.01(.01)	.01(.01)
14	-.02(.02)	.02*(.01)	.02*(.01)
15	.00(.02)	.07*(.01)	.07*(.01)
Female	.00(.01)	.00(.00)	.00(.00)
Race			
White	Ref.	Ref.	Ref.
Black	.04(.03)	.01(.01)	.00(.01)
Multiracial	.02(.01)	.00(.00)	.00(.00)
Asian	.12(.08)	.02(.04)	.02(.01)
Indigenous	-.08(.15)	.02(.04)	.02(.04)
Undeclared	-.08(.09)	.01(.03)	.01(.03)
Region			
North	Ref.	Ref.	Ref.
Northeast	.05*(.02)	.00(.01)	.00(.01)
Southeast	.02(.02)	-.02*(.01)	-.02*(.01)
South	.06(.03)	-.02*(.01)	-.02*(.01)
Center	.06*(.03)	-.02*(.01)	-.02*(.01)
Urban	-.08*(.02)	.00(.00)	.00(.00)

* denotes that the estimate is statistically significant at .05.

Table 1.10: OLS Estimates of Grade Progression, Grade Repetition and Dropout among Eligible Households

	Grade Prog. (SE)	Grade Rep. (SE)	Dropout (SE)
Recipient	-.001(.009)	.015(.008)	-.006*(.003)
Age			
6	Ref.	Ref.	Ref.
7	.45* (.02)	.05* (.02)	.00 (.00)
8	.63* (.02)	.06* (.02)	-.01 (.00)
9	.69* (.02)	.05* (.02)	-.01 (.00)
10	.69* (.02)	.03* (.02)	.00 (.00)
11	.71* (.02)	.02 (.02)	.00 (.00)
12	.71* (.02)	.03 (.02)	.00 (.00)
13	.69* (.02)	.03* (.02)	.01 (.01)
14	.67* (.02)	.03 (.02)	.02* (.01)
15	.61* (.02)	.04* (.02)	.05* (.01)
Female	.06* (.01)	-.05* (.01)	.00 (.00)
Race			
White	Ref.	Ref.	Ref.
Black	-.06* (.02)	.05* (.02)	.00 (.00)
Multiracial	-.03* (.01)	.02* (.01)	.00 (.00)
Asian	-.08 (.07)	.06 (.06)	.01 (.02)
Indigenous	-.04 (.09)	-.01 (.06)	.01 (.03)
Undeclared	-.22* (.06)	.19* (.05)	.02 (.02)
Region			
North	Ref.	Ref.	Ref.
Northeast	.02 (.01)	.02 (.01)	.00 (.00)
Southeast	.04* (.01)	-.02 (.01)	-.01* (.00)
South	-.03 (.02)	.05* (.01)	-.01* (.01)
Center	-.01 (.02)	.04* (.01)	-.01* (.01)
Urban	.07* (.01)	-.04* (.01)	.00 (.00)

* denotes that the estimate is statistically significant at .05.

1.7.1 Grade Progression by Age Group

The main IV result is that the probability of grade progression among students from 6 to 15 years old decreases by 10.5 pp when they participate in the *Bolsa Família* program. To better understand the differences between age groups, I perform the same analysis dividing into primary school children (from 6 to 12 years old) and

adolescents (from 13 to 15 years old).

Table 1.11 (in Appendix) presents that the probability of grade progression from ages 6 to 12 is negative and significant, decreasing by 10.9 pp. On the other side, the probability of progression among students from ages 13 to 15 decreases by 9.1 pp, but it is not significant (Table 1.12, appendix).

1.7.2 Grade Repetition by Age Group

The main IV result shows that the probability of repeating a grade among students from 6 to 15 years old increases by 8.4 pp when they participate in the *Bolsa Família* program. To better understand if younger or older students are repeating more, I perform the same analysis dividing into the two age groups.

Table 1.13 (in Appendix) presents that the probability of repeating the grade from ages 6 to 12 is positive and significant, increasing by 8.7 pp. On the other side, the probability of repetition among students from ages 13 to 15 increases by 7.5 pp, but it is not significant (Table 1.14, appendix). These results show that younger children are failing more in school.

1.7.3 Dropout by Age Group

The same analysis is performed to analyze the probability of dropping out. The main IV results show that the probability of dropping out among students from 6 to 15 years old decreases by .3 pp when they participate in the *Bolsa Família* program, however, it is not significant. Therefore, by dividing into age groups we can analyze if there is any significant difference between young and older students.

Table 1.15 (in Appendix) presents the probability of dropping out among students from ages 6 to 12, which decreases by .6 pp, and table 1.16 (in Appendix) the probability of dropping out among students from ages 13 to 15, which increases by .5 pp. Both results are not significant, but it can be noticed that older students have an inclination to drop out of school, while young students drop out less when participating

in the program.

1.7.4 Grade Progression by Gender

I also analyze if there is any heterogeneity across different genders. The main IV results show that the probability of grade progression decreases by 10.5 pp among program participants. Table 1.17 (in Appendix) presents that the probability of grade progression among boys decreases by 17.3 pp. The probability of progressing among girls decreases by 2 pp and it is not significant (Table 1.18, appendix). Boys are progressing in school about 8 times less than girls.

1.7.5 Grade Repetition by Gender

The main IV results show that the probability of repeating a grade increases by 8.4 pp among program participants and that the probability of repetition among girls is lower than boys.

Table 1.19 (in Appendix) presents that the probability of repeating the grade among boys is positive and significant, increasing by 12.9 pp. The probability of repetition among girls increases by 2.9 pp and it is not significant (Table 1.20, appendix). These results show that the probability of grade repetition is driven by the high and significant probability among boys, which is more than 4 times higher than the probability among girls.

1.7.6 Dropout by Gender

The same analysis is performed to analyze the probability of dropping out among boys and girls. The main IV results show that the probability of dropping out decreases by .3 pp among program participants, however, it is not significant. Table 1.21 (in Appendix) shows that the probability of dropping out among boys increases by .9 pp, and table 1.22 (in Appendix) shows that the probability of dropping out among girls decreases by 1.6 pp. Although both results are not significant, the rate of dropping

out among girls is negative while boys have a positive rate.

Dropout by gender among older students

In the previous section, I analyzed the probability of dropping out among students out of the range of compliance, i.e., those who are 16 and 17 years old. The results show that among these students, the probability increases by 9.3 pp. In this section, I analyze if there is any difference by gender among this age group.

Table 1.23 (in Appendix) shows that the probability of dropping out among boys increases by 21.5 pp. Table 1.24 (in Appendix) shows that the probability of dropping out among girls decreases by 2.2 pp and it is not significant. These results show that the probability of dropping out among 16 and 17 years old students is driven by the high and significant probability among boys.

1.8 Discussion and Conclusion

The objective of this study is to understand how much recipients of *Bolsa Família*'s children are progressing, failing, or dropping out of school. Consistent with the results obtained in the literature, the probability of dropping out of school at ages 6 to 15 decreases among the beneficiaries of the *Bolsa Família* program in 2005. However, when controlling for reverse causality and unobserved variables, this result is not statistically significant. Participating in the program does not significantly impact dropping out of school.

However, this can be seen as a positive result attributed to the condition of participating in the program, since students out of the compliance age range (16 and 17 years old) have the probability of dropping out increase by 9.3 pp.

Furthermore, my analyses show that teenagers (13 to 15 years old) are dropping out more if they participate in the program, while younger kids are dropping out less. A similar result arises by analyzing gender differences – among program participants, boys drop out more while girls drop out less. Although these results are also not

significant, the magnitude of these coefficients suggests a difference between these groups: younger students and boys are more prone to drop out.

Another interesting result is the analysis among those who are out of the compliance age range, divided by gender. With both groups combined, the result shows that there is an increase by 9.3 pp, however, it is clear that this result is driven by the high significant probability of dropping out among boys, which is 21.5 pp, while the probability of dropping out among girls decreases by 2.2 pp.

Analyzing grade progression, my results are not consistent with those found in the literature. Studies show that program participants progress more in school than non-participants. However, when controlling for endogeneity, I find that recipients of *Bolsa Família* are not progressing in school as much as non-participants. Furthermore, I find a significant gender difference. While the probability of progressing decreases by 2 pp and it is not significant among girls, the probability of progressing decreases by a significant 17 pp among boys.

The other outcome analyzed, grade repetition, was not studied in the presented literature. My results show that participating in the program increases the chance of repeating the grade by 8.4 pp. Since this result is controlled for reverse causality, one explanation is that the parents are keeping their children at school, regardless if they are studying or not, just to comply with the requirements of the program. These are probably marginal children, who otherwise might have dropped out of school, but are now staying in school because of the program, and then repeating the grade.

My analyses also show that younger students (6 to 12 years old) are failing more in school compared to teenagers (13 to 15 years old). Looking at differences in gender, boys are significantly failing more than girls (12.9 pp vs. 2.9 pp).

Analyzing the direction of the selection bias for the schooling outcomes, I found that it is negative for grade repetition and dropout, and positive for grade progression. Unobserved variables that increase the probability of participating in the program, decrease the probabilities of grade repetition and dropout, and increase the probability

of grade progression.

This study contributes to the literature by using the Instrumental Variables method that controls for endogeneity. Since I am analyzing the impact of a program on schooling outcomes, and being in school is a condition to participate in this program, this control is essential to properly estimate the effect of the Bolsa Família program. Literature shows studies in this theme using similar data, but methods that do not control for reverse causality or unobserved variables. By using data from 2005, I can compare my results with many of those in the literature, and show that the results vary when addressing endogeneity for two schooling outcomes, dropping out and grade progression. I also estimate that program participants increase the chance of grade repetition, an outcome not previously studied in the literature. Finally, in the present study, I estimate gender and age differences.

There are some limitations to this study. First, measures of schooling outcomes are self-reported. Ideally, this data should be collected from the schools, since they help to monitor the compliance, but there are no data available from schools. Another limitation is that the data do not contain information on when the household started participating in the program. There might be a difference from children belonging to a household that has received the cash-transfer from the beginning of the program in 2003 and those who started receiving just before the survey was performed. For the latter, we would not expect any impact on their schooling outcomes.

1.9 Appendix

Table 1.11: Instrumental Variables Result for Grade Progression by Age Group 6-12

Dep Var	First Stage (Recipients)	Second Stage (Grade Prog)	Reduced Form (Grade Prog)
Registered Recipient	.519*(.013)	-	-.056*(.022)
Female	.01(.01)	-.06*(.01)	-.06*(.01)
Race			
White	Ref.	Ref.	Ref.
Black	.05(.03)	-.06*(.02)	-.07*(.02)
Multiracial	.03(.02)	-.03*(.01)	-.03*(.01)
Asian	.07(.09)	-.05(.07)	-.05(.07)
Indigenous	-.16(.13)	-.04(.10)	-.03(.10)
Undeclared	-.12(.09)	-.24*(.07)	-.23*(.07)
Region			
North	Ref.	Ref.	Ref.
Northeast	.05*(.02)	.03(.02)	.03(.02)
Southeast	.03(.02)	.04*(.02)	.04*(.01)
South	.03(.04)	-.04(.03)	-.05(.03)
Center	.03(.03)	-.01(.02)	-.02(.02)
Urban	-.09*(.02)	.06*(.02)	.06*(.02)

* denotes that the estimate is statistically significant at .05.

Table 1.12: Instrumental Variables Result for Grade Progression by Age Group 13-15

Dep Var	First Stage (Recipients)	Second Stage (Grade Prog)	Reduced Form (Grade Prog)
Registered Recipient	.511*(.017)	-	-.046(.033)
Female	-.02(.02)	.05*(.01)	.05*(.01)
Race			
White	Ref.	Ref.	Ref.
Black	.03(.04)	-.05(.01)	-.06(.03)
Multiracial	.02(.02)	-.03(.01)	-.02(.02)
Asian	.17(.09)	-.07(.10)	-.09(.10)
Indigenous	.30(.24)	-.05(.22)	-.08(.20)
Undeclared	-.14(.10)	-.21*(.11)	-.20(.11)
Region			
North	Ref.	Ref.	Ref.
Northeast	.09*(.03)	-.01(.03)	-.02(.02)
Southeast	.01(.03)	.04(.02)	.04(.02)
South	.05(.05)	-.01(.04)	-.02(.04)
Center	.06(.04)	.02(.03)	.01(.03)
Urban	-.09*(.03)	.08*(.03)	.09*(.02)

* denotes that the estimate is statistically significant at .05.

Table 1.13: Instrumental Variables Result for Grade Repetition by Age Group 6-12

Dep Var	First Stage (Recipients)	Second Stage (Grade Rep)	Reduced Form (Grade Rep)
Registered Recipient	.519*(.013)	-	.045*(.017)
Female	.01(.01)	-.04*(.01)	-.04*(.01)
Race			
White	Ref.	Ref.	Ref.
Black	.05(.03)	.06*(.02)	.06*(.02)
Multiracial	.03(.02)	.03*(.01)	.03*(.01)
Asian	.07(.09)	.02(.05)	.03(.06)
Indigenous	-.16(.13)	-.04(.06)	-.06(.06)
Undeclared	-.12(.09)	.20*(.06)	.19*(.06)
Region			
North	Ref.	Ref.	Ref.
Northeast	.05*(.02)	.01(.01)	.02(.01)
Southeast	.03(.02)	-.03*(.01)	-.03*(.01)
South	.03(.04)	.06*(.02)	.06*(.02)
Center	.03(.03)	-.03(.02)	.04*(.02)
Urban	-.09*(.02)	-.03(.01)	-.03*(.01)

* denotes that the estimate is statistically significant at .05.

Table 1.14: Instrumental Variables Result for Grade Repetition by Age Group 13-15

Dep Var	First Stage (Recipients)	Second Stage (Grade Rep)	Reduced Form (Grade Rep)
Registered Recipient	.511*(.017)	-	.039(.027)
Female	-.02(.02)	-.05*(.01)	-.05*(.01)
Race			
White	Ref.	Ref.	Ref.
Black	.03(.04)	.03(.03)	.03(.03)
Multiracial	.02(.02)	.01(.02)	.01(.02)
Asian	.17(.09)	.10(.09)	.11(.09)
Indigenous	.30(.24)	.13(.23)	.16(.21)
Undeclared	-.14(.10)	.22*(.11)	.21(.11)
Region			
North	Ref.	Ref.	Ref.
Northeast	.09*(.03)	.03(.02)	.03(.02)
Southeast	.01(.03)	-.01(.02)	-.01(.02)
South	.05(.05)	.01(.03)	-.01(.03)
Center	.06(.04)	.02(.03)	.02(.03)
Urban	-.09*(.03)	-.06*(.02)	-.07*(.02)

* denotes that the estimate is statistically significant at .05.

Table 1.15: Instrumental Variables Result for Dropout by Age Group 6-12

Dep Var	First Stage (Recipients)	Second Stage (Dropout)	Reduced Form (Dropout)
Registered Recipient	.513*(.011)	-	-.003(.007)
Female	.01(.01)	.00(.00)	.00(.00)
Race			
White	Ref.	Ref.	Ref.
Black	.04(.03)	.01(.00)	.01(.00)
Multiracial	.02(.01)	.00(.00)	.00(.00)
Asian	.06(.08)	.02(.03)	.02(.03)
Indigenous	.06(.14)	.02(.04)	.02(.04)
Undeclared	-.13(.09)	.02(.03)	.02(.03)
Region			
North	Ref.	Ref.	Ref.
Northeast	.04*(.02)	.00(.00)	.00(.01)
Southeast	.02(.02)	-.01*(.00)	-.01*(.00)
South	.05(.03)	-.02*(.00)	-.02*(.00)
Center	.02(.03)	-.01*(.01)	-.01*(.01)
Urban	-.09*(.02)	.00(.00)	.00(.00)

* denotes that the estimate is statistically significant at .05.

Table 1.16: Instrumental Variables Result for Dropout by Age Group 13-15

Dep Var	First Stage (Recipients)	Second Stage (Dropout)	Reduced Form (Dropout)
Registered	.511*(.016)	-	.002(.02)
Recipient	-	.005(.037)	-
Female	-.03(.02)	.01(.01)	.01(.01)
Race			
White	Ref.	Ref.	Ref.
Black	.02(.04)	.00(.01)	.00(.01)
Multiracial	.02(.02)	.01(.01)	.01(.01)
Asian	.15(.09)	-.04*(.01)	-.04*(.01)
Indigenous	.05(.18)	-.04*(.01)	-.04*(.01)
Undeclared	-.10(.09)	-.01(.04)	.00(.04)
Region			
North	Ref.	Ref.	Ref.
Northeast	.08*(.03)	.01(.01)	.01(.01)
Southeast	.01(.03)	-.01(.01)	-.01(.01)
South	.05(.05)	-.01(.02)	-.01(.02)
Center	.04(.04)	-.03*(.01)	-.03*(.01)
Urban	-.09*(.03)	.00(.01)	.00(.01)

* denotes that the estimate is statistically significant at .05.

Table 1.17: Instrumental Variables Result for Grade Progression by Gender: Male

Dep Var	First Stage (Recipients)	Second Stage (Grade Prog)	Reduced Form (Grade Prog)
Registered	.517*(.014)	-	-.089*(.026)
Recipient	-	-.173*(.051)	-
Age			
6	Ref.	Ref.	Ref.
7	-.05(.04)	.43*(.03)	.44*(.03)
8	.00(.03)	.61*(.03)	.61*(.03)
9	.01(.03)	.66*(.03)	.66*(.03)
10	.01(.03)	.66*(.03)	.66*(.03)
11	.04(.03)	.69*(.03)	.68*(.03)
12	.01(.03)	.69*(.03)	.68*(.02)
13	.05(.03)	.69*(.03)	.68*(.03)
14	.00(.03)	.65*(.03)	.65*(.03)
15	.00(.04)	.56*(.03)	.56*(.03)
Race			
White	Ref.	Ref.	Ref.
Black	.02(.03)	-.06*(.03)	-.06*(.02)
Multiracial	.01(.02)	-.02(.01)	-.03(.01)
Asian	.18(.10)	.08(.07)	.04(.07)
Indigenous	-.29(.21)	-.07(.20)	-.02(.18)
Undeclared	-.20*(.09)	-.26*(.08)	-.23*(.07)
Region			
North	Ref.	Ref.	Ref.
Northeast	.10*(.03)	.04(.02)	.02(.02)
Southeast	.04(.03)	.05*(.02)	.05*(.02)
South	.06(.04)	-.01(.03)	-.02(.03)
Center	.07*(.03)	.00(.03)	-.01(.02)
Urban	-.06*(.02)	.05*(.02)	-.07*(.02)

* denotes that the estimate is statistically significant at .05.

Table 1.18: Instrumental Variables Result for Grade Progression by Gender: Female

Dep Var	First Stage (Recipients)	Second Stage (Grade Prog)	Reduced Form (Grade Prog)
Registered	.515*(.015)	-	-.010(.026)
Recipient	-	-.020(.050)	-
Age			
6	Ref.	Ref.	Ref.
7	-.01(.04)	.47*(.03)	.47*(.03)
8	-.06(.03)	.65*(.03)	.65*(.03)
9	-.04(.03)	.72*(.03)	.72*(.03)
10	-.01(.03)	.73*(.02)	.73*(.02)
11	-.03(.04)	.74*(.03)	.74*(.03)
12	-.02(.03)	.72*(.03)	.72*(.03)
13	-.03(.04)	.69*(.03)	.69*(.03)
14	-.06(.04)	.69*(.03)	.69*(.03)
15	-.03(.04)	.65*(.03)	.65*(.03)
Race			
White	Ref.	Ref.	Ref.
Black	.07*(.03)	-.07*(.02)	-.07*(.02)
Multiracial	.05*(.02)	-.03*(.01)	-.03*(.01)
Asian	.04(.09)	-.13(.08)	-.13(.08)
Indigenous	.01(.16)	-.05(.11)	-.05(.11)
Undeclared	-.06(.10)	-.21*(.09)	-.21*(.09)
Region			
North	Ref.	Ref.	Ref.
Northeast	.03(.03)	.01(.02)	.01(.02)
Southeast	.01(.03)	.03(.02)	.03(.02)
South	.01(.04)	-.06(.03)	-.06(.03)
Center	.01(.03)	-.01(.02)	-.01(.02)
Urban	-.11*(.02)	.08*(.02)	.08*(.02)

* denotes that the estimate is statistically significant at .05.

Table 1.19: Instrumental Variables Result for Grade Repetition by Gender: Male

Dep Var	First Stage (Recipients)	Second Stage (Grade Rep)	Reduced Form (Grade Rep)
Registered	.517*(.014)	-	.067*(.022)
Recipient	-	.129*(.042)	-
Age			
6	Ref.	Ref.	Ref.
7	-.05(.04)	.05(.03)	.04(.03)
8	.00(.03)	.06*(.03)	.06*(.03)
9	.01(.03)	.07*(.02)	.07*(.02)
10	.01(.03)	.06*(.02)	.06*(.02)
11	.04(.03)	.03(.03)	.04(.02)
12	.01(.03)	.04(.02)	.04(.02)
13	.05(.03)	.04(.03)	.05*(.03)
14	.00(.03)	.04(.03)	.04(.03)
15	.00(.04)	.06*(.03)	.06*(.03)
Race			
White	Ref.	Ref.	Ref.
Black	.02(.03)	.05*(.02)	.05*(.02)
Multiracial	.01(.02)	.02*(.01)	.03*(.01)
Asian	.18(.10)	-.08(.05)	-.06(.05)
Indigenous	-.29(.21)	-.07(.19)	.03(.18)
Undeclared	-.20*(.09)	.19*(.06)	.17*(.06)
Region			
North	Ref.	Ref.	Ref.
Northeast	.10*(.03)	-.00(.02)	.01(.02)
Southeast	.04(.03)	-.04*(.02)	-.03*(.02)
South	.06(.04)	.03(.03)	.04(.03)
Center	.07*(.03)	.04(.02)	.05(.02)
Urban	-.06*(.02)	-.03(.02)	-.04*(.02)

* denotes that the estimate is statistically significant at .05.

Table 1.20: Instrumental Variables Result for Grade Repetition by Gender: Female

Dep Var	First Stage (Recipients)	Second Stage (Grade Rep)	Reduced Form (Grade Rep)
Registered Recipient	.515*(.015)	-	.015(.020)
Age	-	.029(.040)	-
6	Ref.	Ref.	Ref.
7	-.01(.04)	.05(.03)	.05(.03)
8	-.06(.03)	.06*(.02)	.06*(.02)
9	-.04(.03)	.03(.02)	.03(.02)
10	-.01(.03)	.00(.02)	.00(.02)
11	-.03(.04)	.00(.02)	.00(.02)
12	-.02(.03)	.01(.02)	.01(.02)
13	-.03(.04)	.01(.02)	.01(.02)
14	-.06(.04)	.01(.02)	.01(.02)
15	-.03(.04)	.02(.02)	.02(.02)
Race			
White	Ref.	Ref.	Ref.
Black	.07*(.03)	.06*(.02)	.06*(.02)
Multiracial	.05*(.02)	.02(.01)	.02(.01)
Asian	.04(.09)	.12(.08)	.12(.08)
Indigenous	.01(.16)	-.02(.07)	-.02(.07)
Undeclared	-.06(.10)	.22*(.08)	.21*(.08)
Region			
North	Ref.	Ref.	Ref.
Northeast	.03(.03)	.03(.01)	.03*(.01)
Southeast	.01(.03)	-.01(.01)	-.01(.01)
South	.01(.04)	.05(.03)	.05(.03)
Center	.01(.03)	.02(.02)	.02(.02)
Urban	-.11*(.02)	-.04*(.02)	-.05*(.02)

* denotes that the estimate is statistically significant at .05.

Table 1.21: Instrumental Variables Result for Dropout: Male

Dep Var	First Stage (Recipients)	Second Stage (Dropout)	Reduced Form (Dropout)
Registered	.512*(.012)	-	.004(.01)
Recipient	-	.009(.017)	-
Age			
6	Ref.	Ref.	Ref.
7	.01(.03)	.00(.01)	.00(.01)
8	.01(.03)	.00(.01)	.00(.01)
9	.02(.03)	-.01(.01)	-.01(.01)
10	.03(.03)	.00(.01)	.00(.01)
11	.06*(.03)	.00(.01)	.00(.01)
12	.01(.03)	.00(.01)	.00(.01)
13	.07*(.03)	.00(.01)	.00(.01)
14	.01(.03)	.01(.01)	.01(.01)
15	.01(.03)	.06*(.01)	.06*(.01)
Race			
White	Ref.	Ref.	Ref.
Black	.02(.03)	.00(.01)	.00(.01)
Multiracial	.01(.02)	.00(.00)	.00(.00)
Asian	.12(.09)	-.01(.03)	.01(.03)
Indigenous	.09(.18)	-.02*(.01)	-.02*(.01)
Undeclared	-.19*(.09)	.00(.02)	.00(.02)
Region			
North	Ref.	Ref.	Ref.
Northeast	.08*(.02)	-.01(.01)	-.01(.01)
Southeast	.02(.02)	-.02*(.01)	-.02*(.01)
South	.07(.04)	-.02*(.01)	-.02*(.01)
Center	.04(.03)	-.02*(.01)	-.02*(.01)
Urban	-.07*(.02)	.00(.01)	.00(.00)

* denotes that the estimate is statistically significant at .05.

Table 1.22: Instrumental Variables Result for Dropout: Female

Dep Var	First Stage (Recipients)	Second Stage (Dropout)	Reduced Form (Dropout)
Registered	.511*(.014)	-	-.008(.010)
Recipient	-	-.016(.020)	-
Age			
6	Ref.	Ref.	Ref.
7	.03(.03)	.00(.01)	.00(.01)
8	-.04(.03)	-.01(.01)	-.01(.01)
9	-.03(.03)	-.01*(.01)	-.01(.01)
10	.01(.03)	.00(.01)	.00(.01)
11	-.02(.03)	-.01(.01)	-.01(.01)
12	-.01(.03)	.00(.01)	.00(.01)
13	-.02(.03)	.01(.01)	.01(.01)
14	-.05(.03)	.02*(.01)	.02*(.01)
15	-.01(.03)	.05*(.01)	.05*(.01)
Race			
White	Ref.	Ref.	Ref.
Black	.05(.03)	.00(.01)	.00(.01)
Multiracial	.04*(.02)	.01(.00)	.01(.00)
Asian	.06(.08)	.00(.02)	.00(.02)
Indigenous	.04(.14)	.03(.05)	.03(.05)
Undeclared	-.06(.09)	.02(.04)	.02(.03)
Region			
North	Ref.	Ref.	Ref.
Northeast	.03(.02)	.00(.01)	.00(.01)
Southeast	.01(.02)	-.01(.01)	-.01(.01)
South	.03(.04)	-.01(.01)	-.01(.01)
Center	.01(.03)	-.01*(.01)	-.01(.01)
Urban	-.11*(.02)	.00(.01)	.00(.01)

* denotes that the estimate is statistically significant at .05.

Table 1.23: Instrumental Variables Result for Dropout for Age 16-17:
Male

Dep Var	First Stage (Recipients)	Second Stage (Dropout)	Reduced Form (Dropout)
Registered Recipient	.450*(.024)	-	.097*(.027)
Age	-	.215*(.061)	-
16	Ref.	Ref.	Ref.
17	.00(.03)	.03(.02)	.03(.02)
Race			
White	Ref.	Ref.	Ref.
Black	-.01(.06)	.00(.04)	.00(.04)
Multiracial	-.01(.04)	.01(.02)	.01(.02)
Asian	.04(.15)	.08(.16)	.09(.14)
Undeclared	.22(.16)	-.17*(.05)	-.12*(.03)
Region			
North	Ref.	Ref.	Ref.
Northeast	.04(.05)	-.02(.04)	-.01(.03)
Southeast	-.02(.04)	-.02(.03)	-.02(.03)
South	.15(.08)	-.08(.05)	-.05(.05)
Center	.00(.06)	-.07(.04)	-.07(.04)
Urban	-.20*(.04)	.10*(.03)	.05*(.03)

* denotes that the estimate is statistically significant at .05.

Table 1.24: Instrumental Variables Result for Dropout for Age 16-17:
Female

Dep Var	First Stage (Recipients)	Second Stage (Dropout)	Reduced Form (Dropout)
Registered Recipient	.484*(.022)	-	-.011(.040)
Age	-	-.022(.082)	-
16	Ref.	Ref.	Ref.
17	-.06*(.03)	.04(.02)	.04(.02)
Race			
White	Ref.	Ref.	Ref.
Black	.09(.06)	-.04(.03)	-.04(.03)
Multiracial	.07(.04)	.01(.02)	.01(.02)
Asian	-.22(.15)	.16(.16)	.16(.16)
Indigenous	.34(.26)	-.11*(.04)	-.12*(.03)
Undeclared	-.12(.14)	-.11*(.03)	-.11*(.02)
Region			
North	Ref.	Ref.	Ref.
Northeast	.07(.05)	.00(.03)	.00(.03)
Southeast	.02(.05)	-.01(.03)	-.01(.03)
South	.01(.08)	.06(.05)	.06(.06)
Center	-.03(.06)	.02(.04)	.02(.04)
Urban	.04(.04)	-.05(.03)	-.06(.03)

* denotes that the estimate is statistically significant at .05.

Chapter 2

The Impact of Cash Transfer Participation on Unhealthy Consumption in Brazil

2.1 Introduction

Conditional Cash-transfer (CCT) programs were created by the government to alleviate poverty and improve the nutrition of low-income families, among other goals. Literature shows that families that participate in these programs use this aid to purchase more food, improving their caloric consumption, but the quality of the food purchased is still low. For instance, Ramírez-Silva et al. (2013) find that participants of the program *Oportunidades* in Mexico improved their dietary intake of the micronutrients examined (iron, zinc, and vitamin A). However, they conclude that this effect is due to the consumption of a food supplement that is part of this program, and not to the improvement of their diet.

In a systematic review of the dietary quality of Supplemental Nutrition Assistance Program (SNAP) participants in the United States, Andreyeva et al. (2015) find that participants consume as many calories as non-participants. However, most of the reviewed studies show that SNAP participants have a lower dietary quality than the comparison groups. The authors suggest that SNAP participants fulfill their caloric needs by purchasing high energy-dense foods, but the type of food they consume is poor in nutrients.

Given this trend, in this chapter, I examine if the increase in food expenses

among the participants of *Bolsa Família* translates into a higher expenditure in high-caloric/low-nutritious food. While program participants are purchasing more food, it is important to understand if they are consuming more ultra-processed foods. After all, over-consumption of energy, fat, and sugar leads to diseases such as obesity, hypertension, allergies, and cancer (Moran et al., 2019), which could also lead to high costs to the government in the long-run. Therefore, my motivation for this study is to investigate if participants of the cash-transfer program are increasing their unhealthy consumption expenditures. The objective of this study is to evaluate the impact of *Bolsa Família* participation on expenses with ultra-processed food, alcohol, and smoking products.

For these analyses, I use the method of Propensity Score Matching. The propensity score estimation relies mostly on logistic regression (Thoemmes and Kim, 2011), which requires the model to be linear that can lead to biased estimates. Recent studies suggest using machine learning techniques to enhance the predictability of econometric models (see Athey et al., 2019). In this work, I compare five methods (Logistic Regression, Random Forests, Gradient Boosting, Support Vector Machines, and Neural Networks) and analyze their predictability power in estimating the probability of being a *Bolsa Família* recipient. Then, I proceed with the Propensity Score Matching to estimate the effects of the treatment at three margins: extensive margin, intensive margin, and the overall effect.

My results contribute to the literature by using non-parametric modeling to estimate the propensity score. The application of a machine learning approach in econometric models is recently growing in the literature, and its application in propensity score estimation is still not vastly explored. No studies using real data that compare multiple of these techniques in this topic were found.

2.2 Food consumption among *Bolsa Família* participants

A critical outcome of the *Bolsa Família* program is that the participating families use the aid to increase their standards of living, mainly by raising their food consumption. Menezes et al. (2008) report that 87% of families use the money from *Bolsa Família* primarily to purchase food and that, on average, beneficiaries spend 56% of their household income on food. De Bem Lignani et al. (2011) show that among five thousand households selected from the *Bolsa Família* registry, families reported higher consumption of cereals, processed foods, meat, milk and dairy, beans, and sugar.

An increase in food consumption for low-income households is a positive outcome since it means that they are consuming more calories, and it might lead to a reduction in food insecurity. However, along with income development comes potential risks for unhealthy food consumption and an increase in BMI above the normal level.

Studies show an increase in food expenses among *Bolsa Família* participants, but the program does not offer instructions or education on good nutrition practices. I hypothesize that these families are buying more food, but maintaining a poor nutrition standard. Program participants could be increasing general food consumption and also increasing their consumption of ultra-processed foods (UPFs).

2.3 Ultra-processed Foods

Ultra-processed foods (UPFs) are foods processed such as they become convenient, cheap, and flavorful, but contain a high amount of calories, fat, sugar, and/or salt. UPFs are predominant in high-income countries, and their popularity is fast increasing in middle-income countries (Popkin et al., 2012). As defined by Monteiro et al. (2010), UPFs are “durable, accessible, convenient, and palatable ready-to-eat or ready-to-heat food products liable to be consumed as snacks or replace home-prepared dishes” (p.7). Typical examples of UPFs are bread, chips, soft drinks, and processed meat.

Ultra-processed foods are particularly attractive to households with children and

to those of lower-income and education (Moran et al., 2019). In the United States, households participating in the SNAP program have higher spending on UPFs and lower spending on healthy foods such as fruit and vegetables, compared to purchases without the SNAP benefit (Franckle et al., 2017). In Mexico, people in households participating in the CCT program *Oportunidades* are associated with a higher BMI and a higher prevalence of excess weight and obesity (Fernald et al., 2008).

Using data from three household budget surveys across three decades (1987, 1995, and 2003), Monteiro et al. (2010) show that the consumption of ultra-processed food products increases among both lower and upper-income groups in Brazil. The most recent survey reveals that ultra-processed food products represent 28% of the total energy (i.e., 418kcal *per capita*) a Brazilian household purchases. The survey also documented an increase in sugar, saturated fat, and sodium compared to the past decades. Using Household Budget Survey data from 2009, Canella et al. (2014) find that the availability of ultra-processed products in the household is positively associated with the average BMI and with the prevalence of excess weight and obesity.

In Brazil, UPFs consumption is increasing at a fast-pace, and therefore, the number of calories in the Brazilian diet is also increasing, even though calories consumed from *in natura* or minimally processed foods is decreasing (Martins et al., 2013).

Since the consumption of UPFs is increasing among the Brazilian population, I investigate if participants of *Bolsa Família* are increasing their unhealthy consumption, which is measured by the household purchase of ultra-processed foods. I also analyze if there is a change in the purchase of alcohol and smoking products.

This study captures the relevant effects of the outcomes at three different margins – extensive margin, intensive margin, and the overall effect. The extensive margin examines differences in participation and the intensive margin refers to the intensity (Saez, 2002). More specifically, in this study, the extensive margin (participation) refers to the probability that a household that participates in the program purchases a product. The intensive margin (intensity) considers participant households that

purchased a product and measure whether they spend more. The overall effect is a combination of the extensive and intensive margins.

2.4 Data

I use the most recent data from the Household Budget Survey (*Pesquisa de Orçamentos Familiares* - POF), conducted by the Brazilian Institute of Geography and Statistics (*Instituto Brasileiro de Geografia e Estatística* - IBGE). The institute collected information from 2008 to 2009 throughout the country, covering 55,970 households, totaling 190,159 individuals.

The main objective of the POF survey is to investigate the pattern of consumption and expenditure of the Brazilian population. These data serve as input for the construction of consumption baskets used to estimate IBGE's consumer price indexes, such as the IPCA (the main consumer price index in Brazil). POF provides information on individuals (age, level of education, and income), housing (size, the existence of sewage, and type of walls), expenditure for each household (habitation, clothing, health, and food), and source of income (total income, income from social programs, income from *Bolsa Família*).

2.4.1 Study Sample

For the present study, households with the reference person (head of the household) below 18 years old were excluded. Extreme observations were also excluded, such as total monthly income below R\$1. The data also contains inconsistent information, such as families who claim to participate in the *Bolsa Família* program and have a monthly wage *per capita* of above R\$6,000 (approximately US\$3,000¹). To balance the sample, I select households that receive up to R\$465 (US\$233) – the minimum monthly wage in 2009. The final study sample comprises 33,395 households.

¹The conversion rate of the values from the dataset refers to 2009 dollars: <https://data.oecd.org/conversion/exchange-rates.htm>

2.4.2 Variables

The main variables of investigation are expenses with unhealthy consumption products. More specifically, the variables selected for this study are household expenses with soda, cookies, packaged foods, food away from home, alcohol, and smoking products² in the month before the survey was conducted. The variable food away from home was divided into ‘unhealthy’, which includes pastries and snacks, and ‘total’ that consists of all food purchased away from home, including the unhealthy options mentioned, and also meals and drinks like coffee, milk, and hot chocolate. The variables ‘total food expense’ and ‘household food expense’ were selected for the estimation of the increased expenditure with food among the program participants.

Before investigating these variables, I perform a correlation analysis among them. If two or more variables are highly correlated, they are assumed to be part of a broader dimension, and then these dimensions should be identified by performing factor analysis. Factor Analysis identifies underlying factors or constructs that reflect what the variables share in common. Table 2.1 shows that the variables are correlated no higher than .25, therefore factor analysis is not necessary³ (Hair et al., 1998).

Table 2.1: Correlation Matrix

	Soda	Cookies	Packaged foods	FAH unhealthy	FAH total	Alcohol	Smoking
Soda	1						
Cookies	.204	1					
Packaged foods	.234	.144	1				
Food away from home unhealthy	.158	.119	.113	1			
Food away from home total	.139	.111	.094	.651	1		
Alcohol	.148	.079	.099	.195	.246	1	
Smoking	.012	.005	-.002	.041	.075	.147	1

²Smoking products include cigarettes, hand-rolled, cigar, pipes, and cigarillos.

³With the exception of total food away from home and unhealthy food away from home that have a correlation of .65. Since one variable is included in the other, it does not make sense to group them.

Since the expenditure graphs are right-skewed, the outcome variables had to be corrected. To capture the effect at the intensive margin, which considers households that are already purchasing a product, I use the log-transformation. By using the log-transformation, I am restricted to comparing the percentage difference in expenditure among households who purchased the product. To capture the overall effect, I need a transformation that allows the analysis across the households who made a purchase or not. I use the inverse hyperbolic sine (ihs)-transformation to be able to analyze the percentage difference expenditure across all households (Rogers et al., 2018). Following the log- and ihs-transformation, the skewness of all variables lied between the ideal range of -0.5 and 0.5 for the distribution to be considered approximately symmetric, except for total food expense (-.87) and household food expense (-.69).

The treatment variable is whether the family participates in the *Bolsa Família* program. Covariates that could predict participating in the program are the characteristics of the household reference person, such as gender, age, age squared, highest schooling degree (4 years of schooling or less, 4–7 years of schooling, 8–10 years, 11–14 years, and 15 years of schooling or more), and race (white, black, Asian, multiracial, and indigenous). Among covariates is also included the characteristics of the household, such as the number of people in the household, number of children, income, region (North, Northeast, Southeast, South, and Center), area (urban or rural), whether the household is considered below the poverty line, whether it is below the extreme poverty line, whether someone in the household is pregnant, and whether someone is breastfeeding.

2.5 Descriptive Statistics

In my study sample, 26.4% of households participate in the *Bolsa Família* program (8,829 households). These families receive from the program an average of R\$86.32 per month. Their average monthly income without government assistance is R\$609.22. By participating in this program, the average increase in income is about 14%.

Table 2.2 shows a summary of the dependent variables in this study, presenting the percentage of households with non-zero expenditures and the average household monthly expenditure, in Brazilian Reals. For example, 94.2% claim to have purchased food in the previous month, spending an average of R\$323.90, 27.1% of households bought soda, spending an average of R\$22.50, and 22% had any expenses with smoking products, spending an average of R\$42.02.

Table 2.2: Average Expenditure by Household

	Any expenditure	Expenditure
Total food	94.2%	R\$323.90
Total food at home	91.0%	R\$263.16
Soda	27.1%	R\$22.50
Cookies	36.2%	R\$19.25
Packaged Food	9.0%	R\$37.21
Food away from home (Unhealthy)	38.2%	R\$43.22
Food away from home (Total)	59.4%	R\$104.58
Alcohol	13.4%	R\$67.18
Smoking	22.0%	R\$42.02

Summary statistics of the variable used in the analysis are shown in Table 2.3. The average age of the household reference person is 46, being 31% female. The most prevalent race is multiracial (51%), followed by white (37%). In terms of schooling, 70% studied for up to 8 years. The average family size is 3.8, with 2 children (among households with a non-zero number of children). The average household has a monthly income *per capita* of R\$232. They are predominantly located in the Northeast (37%), followed by the Southeast (35%), and 77% live in urban areas. A percentage of 24% of families are under the poverty line, with 11% in a situation of extreme poverty.

2.6 Methodology

On the methodological front, my analyses rely on the method of Propensity Score Matching (PSM). The PSM method analyzes the causal effect of treatment from observational data, reducing selection biases in program evaluation (Guo et al., 2006).

Table 2.3: Descriptive Statistics of the Study Sample

Household reference person		Household	
Age	46	Family size	3.75
Female (%)	0.31	Number of children	2.04
Race (%)		Household income (R\$)	232.05
White	0.37	Region (%)	
Black	0.11	North	0.09
Asian	0.00	Northeast	0.37
Multiracial	0.51	Southeast	0.35
Indigenous	0.01	South	0.12
Undeclared	0.00	Center	0.07
Schooling (%)		Urban area (%)	0.77
<4 years	0.38	Extreme Poverty (%)	0.11
4-7 years	0.32	Poverty (%)	0.24
8-10 years	0.13	Pregnant (%)	0.04
11-14 years	0.15	Breastfeeding (%)	0.08
15+ years	0.02		

Recent literature has started exploiting machine learning methods to develop econometrics models (see Athey, 2018). Inspired by these recent applications, I select classification algorithms in Machine Learning to apply in the present study. In classification models, the algorithm learns from the data input and classifies a new observation that can be a bi-class or a multi-class prediction. Since the goal is to estimate a propensity score that predicts whether a family participates in the program, I select models that give a bi-class prediction as the outcome. These are Random Forests, Gradient Boosting Machine, Support Vector Machines, and Neural Networks.

2.6.1 Propensity Score Matching Method

Propensity Score Matching, introduced by Rosenbaum and Rubin (1983), is a technique that estimates the effect of a treatment conditional on covariates that could predict receiving the treatment. Given a vector of these covariates, the model estimates a score that is the predicted probability of receiving the treatment. Condition-

ing the probability of treatment on individual covariates assures that the treatment is independent of covariates.

Given an individual exposed to the treatment and another not exposed with roughly similar propensity scores, assuming the treatment is independent of confounding variables, these two individuals are matched by their propensity scores. By matching treated with untreated individuals, the model estimates the average causal effect of the treatment on the outcome.

The propensity score estimation reduces selection bias while providing a good precision comparing to traditional models such as regression analysis. However, propensity score estimation does not control for unobserved or unmeasured covariates (Joffe and Rosenbaum, 1999).

In my analyses, the treatment is participation in the *Bolsa Família* program. Controlling for variables that can predict being a recipient (such as family size, region, household income, among others), the model estimates the propensity score. Then, it estimates the average treatment effect of the outcomes – household expenses with each unhealthy product – by comparing program participants and non-participants that have similar propensity scores.

In a systematic review of propensity score methods analyzing 86 studies, Thoemmes and Kim (2011) find that logistic and probit regressions are the predominant methods for propensity score estimation, used by 90% of their sample, while the remaining 10% did not specify the method used.

While logistic regression is a reliable model for propensity score estimation, other relatively recent techniques can produce better results. These advanced models have advantages over logistic regression, such as the use of an algorithm that does not assume linear relationships.

In the following section, I examine some of these methods from the machine learning literature, as well as discuss the advantages and disadvantages of logistic regression and the machine learning methods that can be used for propensity score estimation.

2.7 Logistic Regression

Logistic regression was introduced in the 19th century to model population growth (Cramer, 2002), and it is still widely used for several reasons. It is a reasonably easy model to understand and interpret for researchers in multiple areas. For instance, logistic regression is regularly used for classification purposes since it generates a linear combination of variables with coefficients that indicate variable importance when the independent variables are normalized, the relationship between explanatory and response variables, and the significance of these relationships. Its use for probability prediction is appealing given its mathematical constraint of estimating probabilities and converging efficiently on parameter estimates (Westreich et al., 2010). Furthermore, regression models are very accessible to use, being available in virtually all statistical programs.

However, logistic regression relies on the assumption of linearity of the logarithm of the odds of the response variable, and researchers fail to assess this assumption, leading to a poor model fit and a biased effect estimate (D’Agostino Jr, 1998). Even if the linearity assumption is checked, Breiman et al. (2001) explain that goodness-of-fit tests do not reject linearity unless the lack of fit is extreme. Additionally, residual analysis is unreliable, since it does not reject linearity if testing an equation with more than four or five explanatory variables (Breiman et al., 2001).

For this reason, non-parametric modeling has been indicated as more efficient in multidimensional classification problems compared to parametric regression models and may be more applicable to estimate propensity scores (Westreich et al., 2010).

While logistic regression models have advantages such as simplicity and interpretability, they have downsides such as lack of accuracy. Improving accuracy usually requires more complex predictive models. However, since I am estimating a propensity score for matching purposes, the goal is not interpretability, but an accurate prediction.

2.8 Machine Learning Models

Breiman et al. (2001) suggest that there are two cultures in the use of statistical modeling. The first culture is called ‘Data Modeling’, which uses data models to reach conclusions from data, such as linear regression. These models are validated by goodness-of-fit tests and residual analysis. He claims that the statistical community is mostly restricted to the use of data models that can lead to questionable conclusions and irrelevant theory. The second culture is called ‘Algorithm Modeling’, which assumes an unknown data mechanism and employs algorithmic models that can be used in more complex datasets and therefore leads to a more accurate alternative to data modeling. Examples of Algorithm Modeling are decision trees and neural networks. These models are validated by their predictive accuracy.

Athey (2018) defines machine learning as “a field that develops algorithms designed to be applied to datasets, with the main area of focus being prediction (regression), classification, and clustering or grouping tasks”. What Breiman defined in 2001 as an “Algorithm Modeling Culture” is nowadays widely known as “Machine Learning Modeling”.

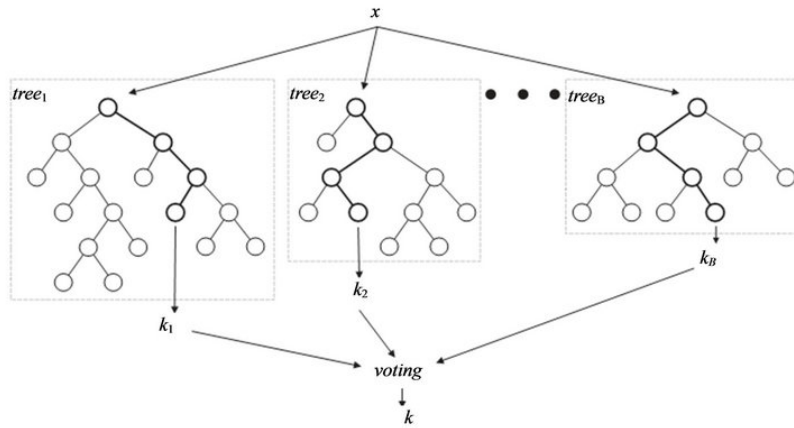
In this work, I analyze four machine learning methods that are best suitable for estimating a propensity score, since they are binary classification models. The methods are Random Forests, Gradient Boosting, Support Vector Machines, and Neural Networks, and are summarized below.

Literature shows an improvement of the prediction accuracy by using some of these methods compared to logistic regression, but not comparing these methods among each other. Few studies compare multiple machine learning models using simulated data, but these methods have not previously been used in this context to model observational data.

2.8.1 Random Forests

The Random Forests method was introduced by Breiman (2001) who describes it as an ensemble method of uncorrelated trees, using the procedure of Classification and Regression Trees (CART) that combines random node optimization and the bagging (bootstrap aggregation)⁴ algorithm. The trees are trained in parallel and each tree does not depend on the other trees. The trees are not pruned like in a decision tree and they are created using subsets of data and averaged together, which improves the prediction accuracy comparing to a single model approach. Figure 2.1 shows trees independently derived from a test sample input. Each tree has its best predictor, and the Random Forests predictor is the average of these predictors.

Figure 2.1: Random Forests Structure. x represents the test sample input, k_i is the prediction from each tree, and k the random forests prediction that is the average of all trees' predictions (Nguyen et al., 2013).



The trees composing a forest are different because they are constructed from a sample without replacement from all observations. Also, the rules for splitting a node are randomly selected from all input variables. This procedure averages the predictions of the individual trees to predict an observation (Breiman, 2001). The theory behind this approach is that averaging the predicted probabilities of trees with different

⁴Bootstrapping is a technique to train a model by drawing predictors used in a large number of samples and estimating models for each sample. These estimates are combined providing the best-estimated coefficients (Hair et al., 1998).

training samples is more robust than a prediction on one training sample.

Many studies suggest using Random Forests to estimate a propensity score to yield better results than using logistic regression. Among many advantages, Zhao et al. (2016) discuss its capacity to handle up to 40% of missing data, Lee et al. (2010) show that the Random Forests approach is well suited to balancing covariates between treatment groups and reduced the bias on effect estimation, and Watkins et al. (2013) discuss the flexibility of Random Forests in incorporating interactions and nonlinear forms.

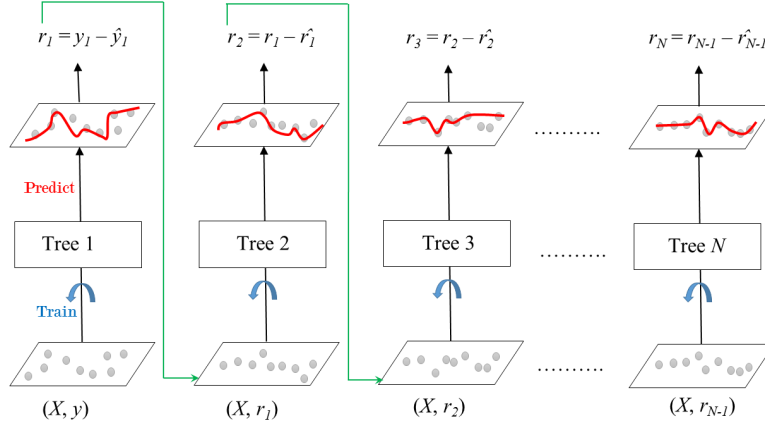
2.8.2 Gradient Boosting Machine

Friedman (2001) introduced the machine learning technique Gradient Boosting Machine with origins from the AdaBoost algorithm of Freund and Schapire (1996). The Boosting algorithm works as an ensemble of “weak” learning algorithms over the training data and combines the classifiers to produce a classification or prediction rule.

Similarly to Random Forests, Gradient Boosting generates more accurate predictions by using an ensemble of decision tree models. The difference is that Random Forests uses bagging instead of boosting. As seen in Figure 2.2, Gradient Boosting starts with an initial model and updates it by successively adding a sequence of regression trees in a step-wise manner. In this sequence, each tree is produced by using the residuals from the previous tree as the input (Sarma, 2017). Each sequential tree of this model slowly reduces the overall error of the previous trees, enabling Gradient Boosting to have a higher predictive power.

Since this technique is based on an ensemble of regression trees, an advantage is that an expertise with the algorithm is not required. Another is that given the ensemble learning over the training data, this method is less susceptible to overfitting. However, since each successive tree uses the residuals of the previous tree, this model has lower interpretability.

Figure 2.2: Gradient Boosting Structure. Subsequent trees are built using the residual r_i from the previous tree. Adapted from Kawerk (2020).



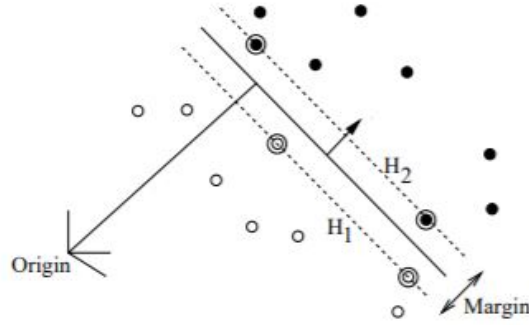
The use of Gradient Boosting Machines for propensity score estimation has seen a growing popularity in the past few years since it outperforms logistic regression for this estimation (Griffin et al., 2017; Jacobsen et al., 2016). Other advantages of the Gradient Boosting Method over logistic regression include handling any type of variable, including missing data. It also captures non-linear and interaction terms and works well with high-dimensional data, even if most covariates are correlated or are unrelated to the treatment variable (McCaffrey et al., 2004).

2.8.3 Support Vector Machines

Support Vector Machines (SVMs) perform classification and regression prediction based on supervised learning models. As with logistic regression, SVMs calculate a set of coefficients for variables based on a transformation of the covariates. However, instead of modeling the probability of the outcomes, SVMs use hyperplanes to divide the observations into class memberships. It was introduced by Vapnik in 1979 and it is used on pattern recognition problems, such as object recognition, speaker identification, face detection in images, and text categorization (Burges, 1998).

SVMs are based on a linear combination of the data points that defines a hy-

Figure 2.3: Linear separating hyperplanes for the separable case. The support vectors are circled. Adapted from Burges (1998).



perplane of a given space, separating the data points into classes (Cristianini et al., 2000). Figure 2.3 shows a simplistic example of a hyperplane splitting black and white dots into two classes.

Since there are multiples ways for this split, a margin is defined and optimization will maximize this margin to improve the accuracy of this separation. The dots that define the margin are called support vectors. The size of the margin carries out a trade-off between correct classification and generalization. A wide margin generates more misclassification but generalizes better, while a narrow margin fits the training better but might overfit the training data. For this reason, the best accuracy will come from the ability of the machine to learn any training set with the lowest error, which is called the principle of structural risk minimization (Joachims, 1998).

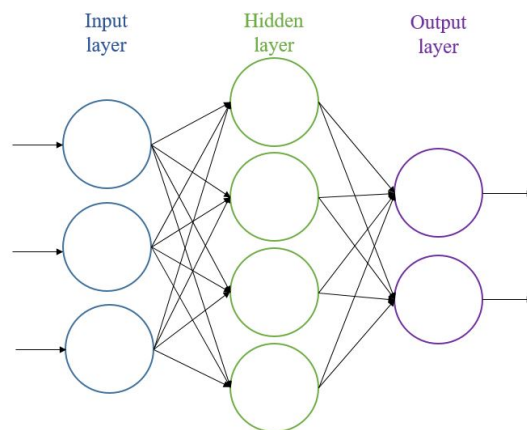
Support Vector Machines use a kernel function for separation in a higher-dimensional space. The kernel function transforms complex data spaces into a form that can be more easily separated. The selected kernel and its associated parameters have a significant effect on how well the resulting model properly classifies the data.

Few studies use SVMs to estimate propensity scores. Ratkovic (2014) shows that the method performs well in theory and practice. Keller et al. (2013) used both SVMs and Neural Networks for propensity score estimation and found that Neural Nets outperforms SVMs under all scenarios studied.

2.8.4 Neural Networks

Neural Networks “can be regarded as a nonlinear mathematical function which transforms a set of input variables into a set of output variables” (Bishop, 1994, p. 1804). Also called Artificial Neural Networks (ANN), this algorithm is inspired by the structure of the nervous system that has an interconnected set of neurons. Neural networks are formed by an input layer, hidden layers, an output layer, and several interconnected nodes that contain an activation function. In Figure 2.4, the circles are nodes that represent neurons and the arrows show how the nodes are interconnected. The process starts from the input layer, communicating to the hidden layer where the process is run via a system of weighted connections, going to the output layer where the prediction is made. A network of interconnected neurons can perform complex learning tasks like classification and pattern recognition (Larose, 2015).

Figure 2.4: A Neural Network with one hidden layer.



Neural networks are useful when the prediction is more important than interpretability and a lot of training data are available.

The advantages of neural networks are that this model works well with high dimensional data and its nonlinearity can approximate any polynomial function (Barron, 1994). The disadvantage of this model is the network training process since there are no rules for selecting the number of hidden nodes and avoid overfitting (Bishop,

1994).

Compared to logistic regression, studies show that neural networks have demonstrated superior performance in propensity score estimation, producing less bias and mean square error (Keller et al., 2015; Setoguchi et al., 2008). However, Farrell et al. (2018) acknowledge that neural networks have inferior performance compared to other machine learning methods mainly because of their limited empirical performance and challenging optimization.

Table 2.4 shows the comparison of the five methods discussed, with their main advantages and limitations.

Table 2.4: Machine Learning methods comparison.

Method	Advantages	Limitations
Logistic Regression	Easy to understand and interpret, accessible	Does not handle missing data, assumes linear relationships, unreliable fit tests
Random Forests	Non-parametric, handles missing data well, incorporates interactions, strong predictive power, robust to outliers	Interpretability, requires computational resources
Gradient Boosting	Non-parametric, less susceptible to overfitting, can handle high-dimensional data, handles any variable type and missing data, robust to outliers	Fairly complex, interpretability, requires computational resources
Support Vectors	Non-parametric, high-dimensional data, incorporates interactions, robust to overfitting	Does not perform well with noisy data, not suitable for large datasets, trade-off accuracy vs. generalization, kernel selection
Neural Networks	Non-parametric, can handle high-dimensional data, handles missing data	Requires inputs for hidden layer and training procedures, can overfit

2.9 Propensity Score Matching Results

The Propensity Score Matching (PSM) method is divided into two stages. The first stage calculates the propensity score, which is the probability of being a *Bolsa Família* recipient. The propensity score is estimated by the Machine Learning method that provides the best prediction of being a recipient. In the second stage, the observations are matched by similar propensity scores – likelihoods of being assigned the treatment conditions.

2.9.1 Model selection

For the proposed estimation, I use the statistical software SAS Enterprise Miner version 15.1 (SAS Enterprise Miner, 2018) for Logistic Regression, SVM, and Neural Networks, and the statistical software Salford Systems version 8.3.0 (Salford Systems, 2018) for Random Forests and Gradient Boosting.

One of the goals of machine learning models is the ability to generalize. To avoid over-fitting, it is necessary to train the model. The training technique divides the dataset into a subset to “train” the model and the complementary subset is used to “test” the model. As a practice in model building and evaluating, the dataset was split in 80% for training and 20% testing. The results presented were validated by the testing dataset.

The target variable should be balanced, i.e., the sum of all weights in each class are equal. Since the target variable participation is unbalanced (about 26% of the households participate in the program), I used balanced class weights in all models.

The specifics for each method used are the following:

- Logistic Regression: standard logistic regression with the main effect for each covariate.
- Random Forests: number of trees: 500, number of predictors: square root of the number of eligible predictors.

- Gradient Boosting: number of trees: 500, with a maximum node of 6 per tree.
- Support Vector Machine: polynomial kernel function.
- Neural Network: the network architecture used was the multilayer perceptron that has no direct connections and the number of hidden neurons is data-dependent.

One of the most common model evaluation methods for classification performance is to analyze the area under the ROC curve (AUC). However, I do not use this metric for the following reasons. First, the ROC curve chart shows the rate of false-positive against the rate of true-positive. In my work, the rate of false-positive can be misleading since there are people who are eligible to participate in the program and could be predicted as so but do not receive the benefit because the municipality they live in has reached the quota, for example. Therefore, the focus of this work is to analyze the best model by comparing the rate of true positive (sensitivity) since what is important is the accuracy of predicting the probability of receiving the benefit (propensity score). Second, analyzing the area under the ROC curve for comparison between classification models is inappropriate since it uses different metrics to evaluate different classifiers (Hand, 2009). Also, this metric is not reliable as an indication of a correctly specified propensity score (Austin, 2009).

Instead, I evaluate the models by sensitivity. Equation 2.1 shows that sensitivity is the number of true positives over the total positive. In the context of my analyses, this is the probability of correctly predicting a program participant among all program participants.

$$Sensitivity = \frac{Number\ of\ true\ positives}{Total\ actually\ positive} \quad (2.1)$$

Table 2.5 shows the model comparison by sensitivity and misclassification. Misclassification is the rate of wrong predictions for both classes. Gradient Boosting shows superior performance, predicting 93.5% correctly among program participants,

followed by Random Forests 83.3% and Neural Networks 77.6%, while SVM and Logistic Regression predict 76.6% and 76.5% respectively.

Table 2.5: Model Comparison.

	Sensitivity	Misclassification
Gradient Boosting	0.935	0.235
Random Forests	0.833	0.237
Neural Networks	0.776	0.246
Support Vector	0.766	0.258
Logistic Regression	0.765	0.258

A Lift chart is a graphical method for evaluating and comparing classification models and their improvement compared to a random guess. As seen in Equation 2.2, the lift for the full dataset measures the change in terms of a lift score, which can be defined as the proportion of true positives divided by the proportion of positives in the full data set (Larose, 2015).

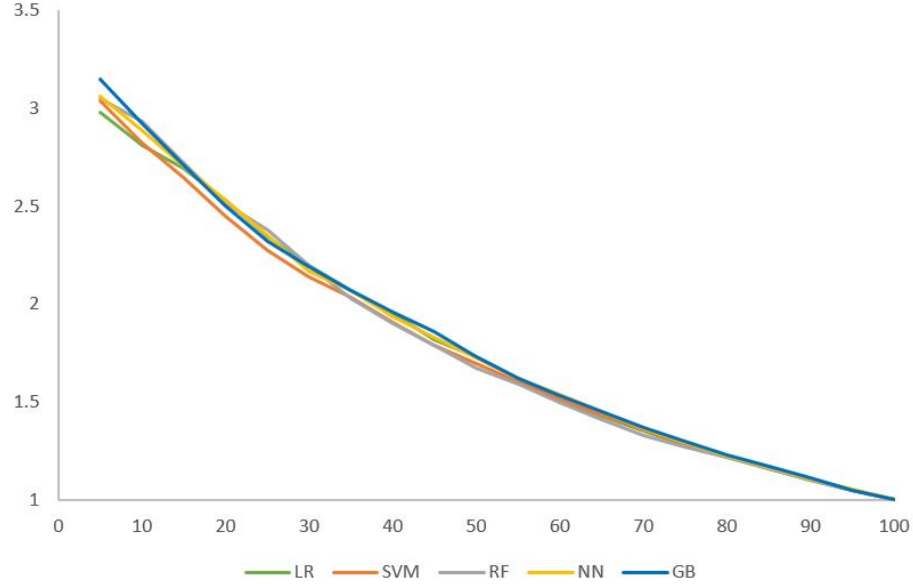
$$Lift = \frac{\text{Proportion of true positives}}{\text{Proportion of positive hits}} \quad (2.2)$$

Lift is typically calculated by sorting the observations by the probability of being classified positive. The lift is then calculated for every decile as the proportion of recipients in the decile divided by the overall proportion of recipients. A chart is then produced that graphs lift against the percentile of the data set.

Figure 2.5 shows the lift chart of the models studied and we can observe that Gradient Boosting (GB) has a higher lift overall, except for between the 20th and 30th percentiles, where Neural Networks and Random Forests provide a slightly higher lift.

From Table 2.5 and Figure 2.5, we can observe that the model with the best prediction power is Gradient Boosting. This is the model selected to perform the propensity score matching.

Figure 2.5: Lift chart - model comparison.



2.9.2 Propensity Score Balance

After estimating the propensity scores and before proceeding to the results, the common support should be checked. Common support is the overlap in the range of propensity scores across *Bolsa Família* participants and non-participants and it is visually assessed in the graph of propensity scores across these groups (Garrido et al., 2014). Figure 2.6 shows the density plots and the overlapping of the probabilities calculated for recipients (red line) and non-recipients (blue line). The first graph shows the distributions across the full sample data (“raw”) and the second graph across the matched sample. The algorithm uses only observations in which the two distributions overlap (“matched”).

A check for the balance in matched samples can be examined by box plots, which graphically evaluates the quality of the match. Figure 2.7 shows the box plots for recipients (red) and non-recipients (blue). In the full sample data (“raw”) we can observe a great disparity between these groups, while the matched sample shows a balance between recipients and non-recipients.

Last, I check the balance of covariates via the standard percentage bias. It is the

Figure 2.6: Density plots for the propensity score.

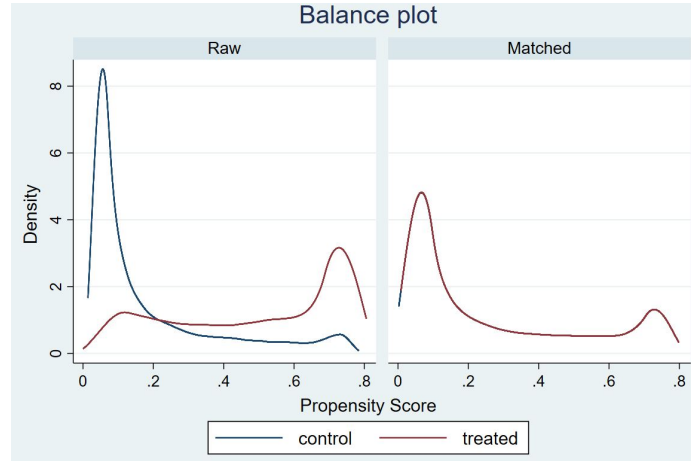
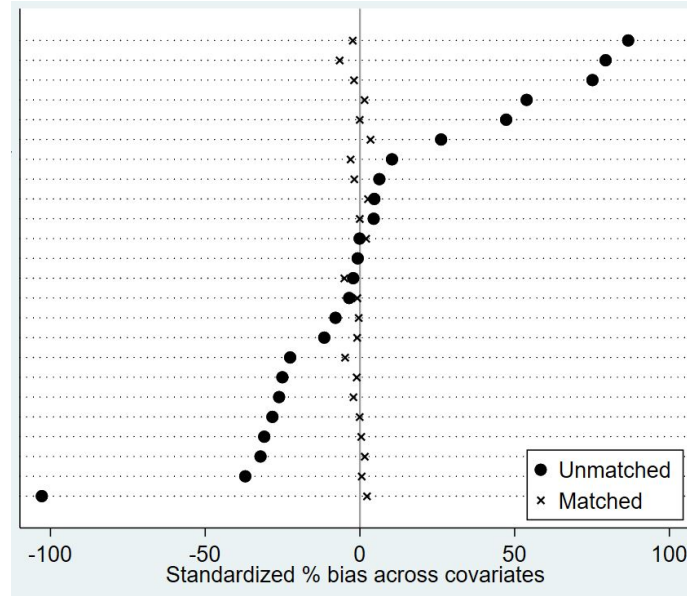


Figure 2.7: Box plots for the propensity score.



percentage difference of the sample means among recipients and non-recipients (in both full and matched samples) as a percentage of the square root of the average of the sample variances in both groups. For a propensity score, the maximum standardized differences of covariates should not be higher than 25 percent (Garrido et al., 2014). Figure 2.8 shows the standardized percentage differences across covariates. In the full sample, the percentage goes from -103% to 87%. In the matched data sample, it goes from -7% to 4%. After verifying that the propensity score is balanced, I proceed to the matching results.

Figure 2.8: Standardized percentage bias across covariates.



2.9.3 Propensity Score Matching

For the matching, I use the software Stata version 15 (Stata Statistical Software, 2017). The method ‘teffects psmatch’ was derived by Abadie and Imbens in 2012 and estimates the standard errors of the estimator that matches on estimated treatment probabilities (Abadie and Imbens, 2016). The matching algorithm used is Nearest Neighbor, in which the observation from the comparison group is chosen as a matching partner for a treated observation that is closest in terms of propensity score (Caliendo and Kopeinig, 2008).

Three PSM models were estimated to capture the effects of the outcomes at the extensive margin, intensive margin, and the overall effect (Table 2.6):

- Model 1: The extensive margin was estimated using a dummy variable denoting having purchased the product in the past month. This model estimates the average treatment effect (ATE) of the outcomes among all households.
- Model 2: The intensive margin was estimated using the log-transformation of the monetary expenditure. This model estimates the average treatment effect (ATE) of the outcomes among the households with a positive expense on the

outcome.

- Model 3: The overall effect was estimated using the inverse hyperbolic sine (ihs)-transformation of the monetary expenditure. This model estimates the average treatment effect (ATE) of the outcomes across all households.

Table 2.6: Marginal Effects on food, alcohol, and smoking spending

	Model 1		Model 2		Model 3	
	ATE	SE	ATE	SE	ATE	SE
Total food	.020*	.004	.145*	.038	.282*	.048
Total food at home	.033*	.006	.104*	.039	.342*	.054
Soda	-.033*	.012	-.058	.042	-.111*	.045
Cookies	.048*	.021	.041	.029	.217*	.103
Packaged Food	-.010	.012	-.126	.090	-.029	.046
Food away from home (Unhealthy)	.075*	.022	-.016	.040	.321*	.095
Food away from home (Total)	.057*	.022	-.034	.042	.258*	.098
Alcohol	.003	.012	.019	.122	.022	.050
Smoking	.005	.014	-.094	.054	.020	.063

Note: each ATE cell represents a separate regression. * denotes that the estimate is statistically significant at .05. Model 1: extensive margin; Model 2: intensive margin; Model 3: overall effect.

Extensive margin results (Model 1) show that the percentage points (pp) of spending with overall food purchase, particularly cookies and food away from home increase among households participating in *Bolsa Família* program, while the purchase of soda decreases. Spending on food in the previous month increases by 2 pp, and food purchased for home by 3.3 pp. Expenses for cookies increase by 4.8 pp, unhealthy food away from home by 7.5 pp, and any food away from home by 5.7 pp. Expenses for soda decrease by 3.3 pp.

Intensive margin results (Model 2) show that among those who purchased food in the last month, *Bolsa Família* participants increase their total spending with food by 14.5% and food purchased for home 10.4%. There is no significant difference in monetary expenses on the products analyzed between program participants and non-

participants.

Overall effect results (Model 3) show that the dollar spent with food in general increases by 28.2%, while food purchase for home increases by 34.2%. Expenses with cookies increase by 21.7%, unhealthy food away from home by 32.1%, and all food away from home by 25.8%. Expenses with soda decrease by 11.1%.

2.9.4 Robustness Check

As a robustness check, the average treatment effects (ATEs) are estimated via Nearest Neighbor PSM with a caliper, which matches only if the difference between two propensity scores is within the caliper size. The caliper size is calculated as a quarter of the standard deviation of the propensity score variable (Rosenbaum and Rubin, 1983). The standard deviation of the propensity score calculated by the Gradient Boosting Machine method is .2943, so the caliper used is .0735.

Using caliper as a robustness check, I obtained the same effects as those estimated by the PSM, meaning that the results obtained are reliable and not dependent on the particular method chosen.

2.10 Conclusion

When compared to a similar household that does not receive CCT from the government, a *Bolsa Família* participant household displays some differences in their expenses behavior. Program participants spend more purchasing food, but they do not spend more on unhealthy products, among families that already buy those. However, the probability of consuming cookies and food away from home increases.

The overall effect shows that they increase their expenses on cookies and food away from home, considering the whole sample. Interestingly, there are no significant differences in the probability of purchasing or the amount spent on alcohol and smoking products. Another finding is that there are no significant differences in the probability of purchasing or amount spent on packaged foods, which are not only ultra-processed

products but are in the rise in many countries, including Brazil. Furthermore, another product type that is rising in many Latin American countries is soda, but program participants are decreasing their household expenses on this product.

These findings show that *Bolsa Família* participants are using the cash-transfer to purchase more food but not necessarily that their diets are getting worse. Overall, they increase snack consumption like cookies and out-of-home pastries.

The present study contributes to the literature by analyzing if Bolsa Família participants are using the cash-transfer to purchase more unhealthy products than non-recipients. Furthermore, I contribute to the machine learning and econometrics literature by using bi-class machine learning classification models to estimate a propensity score. These models are generally criticized by their non-trivial interpretability, however, in estimating a propensity score for matching purposes, interpretation is not essential. In this case, it is important to have a higher prediction accuracy, and I show that in the context of this data, all machine learning models have better accuracy than the model that is widely used in the Propensity Score Matching method, logistic regression.

There are some limitations to this study. First, household expenses are self-reported, thus the items purchased are not accurate. Second, the data contain household purchases, but there is no information on household consumption. A suggestion for a future study is the estimation of the consumption by each member of the family, for example, to analyze whether children, in particular, are consuming more unhealthy products. Last, the data used are the most recent Household Budget Survey from 2009, so these results might look different now. I plan to perform the same analyses once IBGE makes available the Household Budget Survey data collected in 2018, scheduled at the end of 2020.

Chapter 3

Direct and Indirect Associations between Body Image Perception, Depression, and Risk Behavior among Brazilian Adolescents

3.1 Introduction

Adolescence is the period of life in which great emotional, social, and physical changes occur. Adolescents are influenced by key determinants, such as relationships, cultures, and economic conditions, that can be referred to as risk or protective factors. According to the World Health Organization (WHO)¹, “risk factors are conditions or variables associated with a lower likelihood of positive outcomes and a higher likelihood of negative or socially undesirable outcomes. Protective factors have the reverse effect: they enhance the likelihood of positive outcomes and lessen the likelihood of negative consequences from exposure to risk.” Risk behaviors in adolescents include anxiety, poor social skills, conduct disorder, substance use, lack of adult supervision, and aggression (OConnell et al., 2009).

During adolescence, there is an increase in psychosocial problems associated with body image (Laus et al., 2011). For instance, body dissatisfaction increases suicidal behaviors among American adolescents, with a strong impact on girls (Dave and

¹https://www.who.int/hiv/pub/me/en/me-prev_ch4.pdf

Rashad, 2009).

This phenomenon is not restricted to the United States. Edmonds (2007) documents the value associated with the body image as a status symbol among Brazilians. In fact, Brazil's beauty culture is part of the national identity. The country ranks second in the number of cosmetic surgery procedures, only behind the United States². There is a perception of getting better job opportunities and social benefits if one has better aesthetics. As a result, body dissatisfaction is increasing among Brazilian adolescents.

In Brazil, risk behaviors are also increasing in this group, such as unsafe sex, domestic violence, involvement in gunfights, and bullying victimization (Azeredo et al., 2019). For example, when asked about reasons for bullying, 18.6% of Brazilian adolescents claim their body appearance. This is the most frequent reason for bullying among those who replied. Other reasons were facial appearance 16.2%, race 6.8%, sexual orientation 2.9%, religion 2.5%, and region of origin 1.7% (Oliveira et al., 2015).

Current literature presents associations among body dissatisfaction, depression, and risk behaviors. For instance, Connell et al. (2009) find a strong association between substance use and risky sexual behavior among adolescents. Other studies show associations between adolescent depression and sex and drug behaviors (Hallfors et al., 2004), health risk behaviors (Paxton et al., 2007), and negative body image and negative feelings of self-worth (McGrath et al., 2009). In Brazil, studies show a relationship between body image, gender and nutritional status (Laus et al., 2013), and an association between bullying and risk behaviors to the adolescents' physical and psychological health (Silva et al., 2012).

The objective of this study is to identify direct and indirect associations between negative body image perception, depression, and risk behaviors among adolescents in Brazil. Analyzing negative body image perception is appropriate since studies sug-

²<https://www.worldatlas.com/articles/the-number-of-cosmetic-surgery-procedures-by-country.html>

gest that the way adolescents perceive their body image may be more relevant than their actual weight in affecting depression or other problematic behaviors (Dave and Rashad, 2009).

3.2 Methodology

In order to obtain a better understanding of the direct and indirect associations among the variables of interest, several statistical techniques are employed. First, the groups of behavioral variables that are indicators of risk behaviors, depression, and body image perception undergo an Exploratory Factor Analysis to identify latent constructs from the observed variables. Second, a Confirmatory Factor Analysis tests the relationship between the variables and their constructs.

With the latent constructs identified, a Directed Acyclic Graphs (DAG) analysis is executed to identify possible directed associations among the constructs. Then, I construct a Structural Equations Model (SEM) with the associations indicated by DAG to obtain the effects among the constructs.

3.2.1 Exploratory Factor Analysis

Exploratory Factor Analysis (EFA) is a statistical technique that identifies a small number of common factors that explain the correlation between the measured variables. It identifies a structure between observed data and underlying latent constructs (Ferguson and Cox, 1993). EFA has been used to explore and define the underlying factor structure of a group of variables with no need to enforce a framework structure *a priori* (Cattell, 2012).

3.2.2 Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) is a statistical model used to verify a known factor structure of a set of observed variables. CFA empirically estimates and confirms the

relationship between the observed variables and their underlying latent constructs that were identified *a priori*, during the EFA phase (Child, 1990).

3.2.3 Directed Acyclic Graphs

Directed Acyclic Graphs (DAG) is a method that estimates directional links among constructs. In this study, DAG can shed light on the direct and indirect effects among variables.

A directed graph is a path of arrows and vertices that represents the direct relationship among a set of variables (Bessler and Loper, 2001). The graph is acyclic if it contains no cycle, i.e., no path goes back to a variable that has passed already (Haughton and Haughton, 2011). For example, the path $A \rightarrow B \rightarrow C \rightarrow A$ would denote a cyclic graph since it goes back to A after C.

A directed acyclic graph can be used to represent conditional independence relations in a probability distribution (Spirtes et al., 2000). Pearl et al. (2009) posit that direct effects in graphical models can be identified if the graph is acyclic, satisfying a Causal Markovian Condition, and the error terms are jointly independent.

If there is a direct arrow from node A to B, then A is a parent of B. Let v_i denote the n variables in the dataset, pa_i denote the set of parents of v_i , and P denote the probability of an event involving the variables. The Causal Markovian Condition theorem holds the following.

Theorem 1 (Causal Markovian Condition). *Any distribution generated by a Markovian model can be factorized as:*

$$P(v_1, v_2, \dots, v_n) = \prod_i P(v_i | pa_i)$$

The product is calculated over all variables v_i and each term in the product refers to the conditional density of v_i given its parents (Pearl, 2009). Since the DAG represents a joint distribution it can be said that each variable is independent of its

non-descendants given its parents (Haughton et al., 2006).

3.2.4 Structural Equations Modeling

Structural Equations Modeling (SEM) constructs models in which the equations are directed relationships (Goldberger, 1972). As a combination of factor analysis and multiple regression, they model the relationships between unobservable latent constructs and observed variables. SEM allows for a representation of unobserved concepts each with several indicators, assessment of measurement error, multiple and interrelated dependence equations, and for the dependent variable in one equation to be independent in another.

3.3 Data

The data used in this study arise from the most recent Brazilian National Survey of School Health (*Pesquisa Nacional de Saúde do Escolar* – PeNSE), collected in 2015 at public and private schools from the 26 state capitals and the Federal District, ensuring representativeness of Brazil’s population. The validity of the survey instrument was evaluated by several studies (e.g., Tavares et al., 2014). The data contain 100,497 observations, each representing an adolescent attending the last year of middle school (ninth grade).

The main objective of the PeNSE survey is to assess the risk and protective factors in the adolescent’s health. Topics surveyed include sociodemographic characteristics, dietary habits, parental involvement, alcohol and substance use, sexual behavior, violence, and body image³.

³<https://www.ibge.gov.br/estatisticas/sociais/saude/9134-pesquisa-nacional-de-saude-do-escolar.html>

3.3.1 Variables

The main variables of interest are negative body image perception, depression, and risk behaviors. Given the differences among adolescents' problems by gender (Leadbeater et al., 1999), I analyze the association for boys and girls separately. Control variables are age, race, and region.

The variables selected to analyze negative body image perception are the questions (1) "In relation to your body, you feel", with answer options 'very thin', 'thin', 'normal', 'fat', and 'very fat'; (2) "Are you trying to change your weight?", with answer options 'no', 'trying to lose', 'trying to gain', 'trying to maintain'; (3) 'Have you ever vomited or taken laxatives to lose weight?', with answer options 'yes' and 'no'; (4) "In the past 30 days, did you take any medicine or product to lose weight that was not recommended by your doctor?", with answer options 'yes' and 'no'; (5) "In the past 30 days, did you take any medicine or product to gain weight or lean mass that was not recommended by your doctor?", with answer options 'yes' and 'no'.

Depression is estimated by the following questions (6) "In the last 12 months, how often have you felt lonely" and (7) "In the last 12 months, how often were you unable to sleep at night because something worried you a lot?", both with answer options 'never', 'rarely', 'sometimes', 'mostly', and 'always', and the question (8) "How many close friends do you have?", with answer options 'none', '1 friend', '2 friends', and '3 or more friends'.

The variables associated with risk behaviors are the questions: "In the past 30 days, how many days did you use" (9) cigarette, (10) tobacco, (11) alcohol⁴, (12) heavy drinking episode, (13) marijuana, and (14) crack, with answer options "none", "1 or 2 days", "3 to 5 days", "6 to 9 days", "10 or more in the past 30 days"; (15) were physically assaulted by an adult family member, with answer options "none", "once", "2 or 3 times", "4 or 5 days", "6 or 7 times", "8 or 9 times", "10 or 11 times",

⁴Consumption of alcohol and tobacco are considered as risky activities as illegal substances since these products are illegal for purchase and consumption for those younger than 18 years old, per Brazilian law.

“12 or more times in the past 30 days”; “In the past 12 months, how many times you” (16) were physically assaulted, (17) have you been in a fight, (18) have you been seriously injured, with answer options “none”, “once”, “2 to 3 times”, “4 to 5 times”, “6 to 7 times”, “8 to 9 times”, “10 to 11 times”, “12 or more times in the past 12 months”; (19) “Have you had sex?”, (20) “How many sexual partners have you had?”, (21) “Did you use a condom in your first sexual intercourse?”; “In the past 30 days, how often did your parents or guardians” (22) check your homework, (23) are aware of what you do in your free time, (24) understand your problems and concerns, with answer options ‘never’, ‘rarely’, ‘sometimes’, ‘mostly’, and ‘always’.

All variables were scaled in such a way that the lower value represents low risk and the higher value represents a high risk. Since these are ordinal variables with limited values, the data do not contain outliers. For this same reason, there is no need to standardize the data.

3.3.2 Descriptive Statistics

Summary statistics of the sociodemographic variables are shown in Table 3.1. The total sample contains 100,497 adolescents attending ninth grade in school, 49% boys (46,202) and 51% girls (50,511).

In Brazil, the age range varies in each year of school, since failing and repeating the grade is somewhat common. In the sample data, boys are 13 years old and below (16%), 14 (49%), 15 (22%), and 16 and above (13%). The most prominent race is multiracial (40%) followed by white (37%) and black (15%). Most live in the Southeast (44%) and Northeast (26%) regions. Among girls, the age distribution is 13 years old and below (20%), 14 (53%), 15 (17%), and 16 and above (9%). The most prominent race is multiracial (46%) followed by white (35%) and black (11%). Most live in the Southeast (42%) and Northeast (29%) regions.

Table 3.2 shows the statistics of the variables selected for the constructs of negative body image perception, depression, and risk behaviors by gender. The variables are

Table 3.1: Descriptive Statistics of the Adolescents

Variable	Boys (49%)	Girls (51%)
Age: 13-	0.16	0.20
Age: 14	0.49	0.53
Age: 15	0.22	0.17
Age: 16+	0.13	0.09
White	0.37	0.35
Black	0.15	0.11
Asian	0.04	0.05
Multiracial	0.40	0.46
Indigenous	0.04	0.03
North	0.09	0.09
Northeast	0.26	0.29
Southeast	0.44	0.42
South	0.12	0.12
Center-west	0.08	0.07

described in section 3.3.1 and can be identified by their reference number. The first two variables (bodyperc and changew) were re-coded from the description of the variables, in which 0 means perceive the body as normal or does not want to change weight, while 1 means perceive the body as thin or fat or wants to change weight.

All variables show a statistical difference between genders, except for the frequency in use of tobacco and alcohol.

For the variables related to negative body perception, 47% of the girls and 41% of the boys perceive their body other than ‘normal’, 46% of the girls and 39% of the boys are trying to change their weight, 7% of girls and 6% of boys take laxative or vomit for weight loss, 5% of girls and 6% of boys take medicine or product for weight loss, and 5% of girls and 8% of boys take medicine or product for weight gain.

For the variables related to depression, most girls feel lonely and cannot sleep because are worried sometimes while most boys rarely, and most boys and girls have more than 3 close friends.

Among boys, 19% smoked cigarettes in the past month at least once, 21% tobacco, 53% alcohol, 9% marijuana, 9% crack, 13% were physically assaulted by an adult

Table 3.2: Descriptive Statistics of the Variables

Variables		Min	Max	Mean Boys	SE Boys	Mean Girls	SE Girls
1	bodyperc	0	1	0.41	.002	0.47	.002
2	changew	0	1	0.39	.002	0.46	.002
3	laxat	0	1	0.06	.001	0.07	.001
4	prodlose	0	1	0.06	.001	0.05	.001
5	prodgain	0	1	0.08	.001	0.05	.001
6	lonely	1	5	2.03	.005	2.64	.005
7	nosleep	1	5	1.91	.005	2.40	.005
8	closefriends	1	4	3.65	.004	3.61	.003
9	cigarette	0	7	0.33	.004	0.27	.003
10	tobac	0	7	0.35	.004	0.35	.004
11	alcohol	0	7	0.93	.006	0.95	.005
12	heavydrink	0	5	0.88	.005	0.85	.005
13	marijuana	0	4	0.17	.003	0.13	.002
14	crack	0	4	0.10	.002	0.08	.001
15	aggrefam	1	8	1.36	.006	1.33	.005
16	physaggre	1	8	1.45	.006	1.40	.005
17	fight	1	8	1.67	.007	1.30	.004
18	injur	1	8	1.30	.005	1.18	.003
19	sex	0	1	0.36	.002	0.19	.002
20	qtsex	0	6	1.17	.009	0.41	.005
21	condom	0	2	0.52	.004	0.25	.002
22	checkhw	1	5	3.08	.007	3.31	.007
23	aware	1	5	2.30	.006	2.14	.006
24	underst	1	5	2.71	.007	2.92	.007

family member, 18% were physically assaulted, 30% have been in a fight, and 13% have been seriously injured. Among girls, 17% smoked cigarettes in the past month at least once, 25% tobacco, 55% alcohol, 8% marijuana, 8% crack, 15% were physically assaulted by an adult family member, 18% were physically assaulted, 16% have been in a fight, and 10% have been seriously injured.

A percentage of 36% of the boys have had sex, 46% of them had at least 2 partners, and 56% of them used a condom in the first sexual intercourse, while 19% of the girls have had sex, 47% of them had at least 2 partners, and 70% of them used a condom

in the first sexual intercourse.

Parents or guardians check the homework of boys sometimes and more often of girls, are rarely aware of what boys and girls do in their free time, and sometimes understand their concerns.

3.4 Results

3.4.1 Exploratory Factor Analysis

To construct a model based on risk behaviors among teenagers, I first conduct an Exploratory Factor Analysis to identify underlying constructs or factors that group variables in common. The selected variables were checked for appropriateness by performing a correlation. All variables should be correlated to at least another by 30% or more, otherwise, they should be excluded from the analysis (Hair et al., 1998). Some of the variables selected did not have the desired correlation. Those excluded were ‘how many close friends do you have’, ‘heavy drinking episode’, ‘are you trying to change your weight’, ‘in relation to your body do you feel’ for both boys and girls and ‘used product for weight or lean mass gain’ for girls only.

Exploratory Factor Analysis was performed on the statistical program SPSS (SPSS, 2012) with the maximum likelihood extraction method. The factor analysis indicators of fit presented adequate levels. The measures of sampling adequacy (MSA) of all variables are above 0.5. Another measure of sampling adequacy is the Kaiser-Meyer-Olkin test that shows indices above 0.8. Communalities show the variances of original variables explained by extracted factors. All variables are close to or above .3, which is ideal.

Using the varimax rotation to minimize the number of variables that load high on each factor, the analysis suggests six factors based on eigenvalues greater than one. The exploratory factor analysis for the extraction of six factors was able to explain 64% of the variance of the boys sample and 65% of the girls sample. The variables in

each factor with loadings above .3 for the boys sample are the following.

- Factor 1 (*Sexual behavior*): has had sex, number of sexual partners, and condom use
- Factor 2 (*Illegal substance use*): marijuana, crack, cigarette, tobacco, and alcohol
- Factor 3 (*Aggression*): physical assault, physical assault by an adult family member, have been injured, have been in a fight
- Factor 4 (*Depression*): feel lonely, cannot sleep worried
- Factor 5 (*Negative body image*): use of laxatives or induced vomit, use products for weight loss, use products for weight gain
- Factor 6 (*Parental involvement*): parents understand my problems and concerns, parents check my homework, parents are aware of what I do in my free time

For the girls sample, the factors from the explanatory factor analysis are the same, except for the factor “Body image distortion”, which does not contain the variable ‘use products for weight gain’ that was excluded before the analysis since it was not correlated with any other variables analyzed.

3.4.2 Confirmatory Factor Analysis

After performing the exploratory factor analysis, it is necessary to validate the proposed factorial structure and explore whether any significant changes are needed. Using the statistical program Amos (SPSS Amos, 2012), these can be verified by performing a confirmatory factor analysis (CFA). For both boys and girls samples, the variables discarded were those associated with the factor ‘Parental involvement’ that demonstrated to be inadequate in the search for good consistency in representing the construct. Therefore this factor is excluded from the model. The derived models showed acceptable levels of fit.

Table 3.3: Factors for the boys sample.

	Factors - Boys				
	Sexual behavior	Illegal substance use	Aggression	Depression	Negative body image
Has had sex	.964				
Condom use	.901				
Number sex partners	.763				
Marijuana		.945			
Crack		.843			
Cigarette		.583			
Tobacco		.476			
Alcohol		.386			
Physical aggression			.716		
Aggression by family			.611		
Injured			.573		
Fight			.516		
Feel lonely				.989	
Cannot sleep				.414	
Laxatives					.684
Product for weight loss					.549
Product for weight gain					.503

Boys sample: CFI (comparative fit index) = 0.930, IFI (incremental fit index) = 0.930, and TLI (Tucker-Lewis index) = 0.914 are adequate since all these indices are higher than 0.9. The RMSEA (root mean square error of approximation) = 0.067 shows the “badness of fit index” and it is adequate since it is below 0.08 (Hair et al., 1998). Chi-square test was not analyzed since it is deemed unreliable for large samples (Hair et al., 1998). The final scale identified was composed of 17 indicators that make up the five dimensions of the construct. Table 3.3 shows the dimensions and attributes found, with assigned names.

Girls sample: CFI = 0.930, IFI = 0.930, and TLI = 0.913. RMSEA = 0.061. The final scale identified was composed of 16 indicators that make up the five dimensions of the construct. The factors and attributes are similar to the boys sample, with differences in the variable loadings (Table 3.4).

Table 3.4: Factors for the girls sample.

	Factors - Girls				
	Sexual behavior	Illegal substance use	Aggression	Depression	Negative body image
Has had sex	.963				
Condom use	.922				
Number sex partners	.744				
Marijuana		.967			
Crack		.860			
Cigarette		.554			
Tobacco		.447			
Alcohol		.336			
Physical aggression			.744		
Aggression by family			.696		
Injured			.444		
Fight			.408		
Feel lonely				.869	
Cannot sleep				.466	
Laxatives					.568
Product for weight loss					.563

3.4.3 Directed Acyclic Graphs

With the five factors defined, Directed Acyclic Graphs (DAG) is the method recommended for identifying direct associations prior to performing a structural equation modeling (Haughton et al., 2006).

Directed Acyclic Graphs were performed in the program GeNIe Modeler (GeNIe Modeler, 2019), using the learning algorithm Partial Correlation (PC) that is the most popular algorithm for its high computational efficiency property (Spirtes et al., 2000; Kalisch and Bühlmann, 2007). It uses independences observed in data to infer the structure that has generated them. The recommended parameters were selected - maximum adjacency size of 8, a significance level of 0.05, and no time limit for performing the algorithm.

The model can be interpreted in the following way. For the Girls sample (Fig-

ure 3.1), illegal substance use, sexual behavior, negative body image, and depression are directly associated with aggression. Both sexual behavior and depression are directly associated with negative body image. For the Boys sample (Figure 3.2), illegal substance use, sexual behavior, negative body image, and depression are directly associated with aggression. Illegal substance use is directly associated with sexual behavior. Sexual behavior, depression, and illegal substance use are directly associated with negative body image.

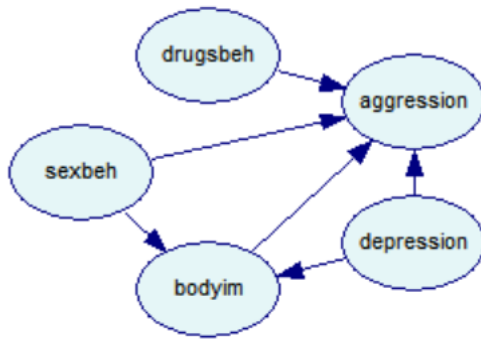


Figure 3.1: DAG: Girls

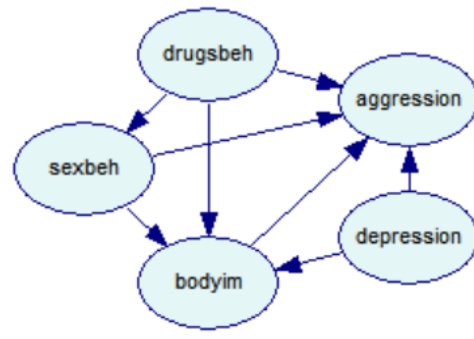


Figure 3.2: DAG: Boys

3.4.4 Structural Equations Modeling

Given the relationships identified by the Directed Acyclic Graphs, I proceed to construct a structural equation model (SEM) using the statistical program Amos (SPSS Amos, 2012). For both the Boys and the Girls samples, I input the direct and indirect relationships as identified by the DAG, in addition to controlling for race, age, and region.

The Bootstrap approach was performed to obtain robust statistics in Structural Equations Modeling (Yung and Bentler, 1996). This method is not grounded on any assumptions of the population's distribution or any covariance structure model for the data (Nevitt and Hancock, 2001). Amos offers bootstrap-derived robust statistics as an alternative to normal theory hypothesis testing methods, providing both standard errors and an adjusted model test statistic p-value. The bootstrap is performed with

200 samples and 95% bias-corrected confidence intervals.

Table 3.5: Structural Equations Modeling Estimates.

Parameter			Boys				Girls			
			Est.	SE	LCI	UCI	Est.	SE	LCI	UCI
sexbeh	<–	drugsbeh	.35	.005	.34	.36	-	-	-	-
body	<–	sexbeh	.18	.007	.16	.19	.09	.007	.08	.11
body	<–	depress	.07	.009	.06	.09	.28	.009	.26	.30
body	<–	drugsbeh	.15	.011	.12	.17	-	-	-	-
aggres	<–	drugsbeh	.18	.011	.16	.20	.16	.012	.13	.18
aggres	<–	sexbeh	.08	.007	.07	.10	.06	.007	.04	.07
aggres	<–	depress	.20	.008	.18	.21	.26	.008	.25	.28
aggres	<–	body	.37	.012	.35	.40	.19	.012	.17	.22
tobac	<–	drugsbeh	.52	.010	.50	.54	.48	.011	.46	.49
cigarette	<–	drugsbeh	.64	.008	.63	.66	.61	.008	.60	.63
marijua	<–	drugsbeh	.93	.003	.92	.93	.96	.003	.95	.96
alcohol	<–	drugsbeh	.48	.008	.46	.50	.42	.007	.41	.43
crack	<–	drugsbeh	.87	.004	.86	.88	.89	.003	.89	.90
fight	<–	aggres	.59	.008	.57	.60	.48	.011	.45	.50
aggrefam	<–	aggres	.65	.009	.63	.66	.71	.009	.68	.73
physaggre	<–	aggres	.68	.008	.66	.70	.73	.008	.72	0.75
injur	<–	aggres	.65	.009	.63	.67	.48	.012	.45	.50
sex	<–	sexbeh	.98	.001	.98	.99	.98	.001	.98	.98
condom	<–	sexbeh	.93	.001	.93	.94	.95	.001	.95	.95
qtsex	<–	sexbeh	.81	.002	.81	.81	.79	.002	.78	.79
nosleep	<–	depress	.71	.014	.69	.74	.70	.007	.69	.71
lonely	<–	depress	.60	.011	.58	.62	.65	.007	.63	.66
laxat	<–	body	.58	.010	.56	.59	.69	.013	.66	.71
prodlose	<–	body	.68	.009	.66	.70	.56	.012	.53	.58
prodgain	<–	body	.56	.009	.54	.57	-	-	-	-

Table 3.5 shows the estimates of the relationship between measured variables and latent constructs, and among the latent constructs that have directional links between them. The table is divided into Boys and Girls samples and displays their standardized parameter estimates (Est.), standard errors (SE), and 95% confidence intervals (LCI and UCI).

From the SEM results, the sexual behavior construct (sexbeh) has the highest

impact on negative body image perception (body) among boys (.18), followed by illegal substance use (drugsbeh) (.15) and depression (depress) (.07). Among girls, negative body image perception is highly impacted by the depression construct (.28), followed by sexual behavior (.09).

Among boys, aggression (aggress) is highly impacted by having a negative body image perception (.37), followed by depression (.20), illegal substance use (.18), and sexual behavior(.08). Among girls, aggression is impacted by depression (.26), negative body image perception (.19), illegal substance use (.16), and sexual behavior(.06).

There is a high impact of illegal substance use on sexual behavior (.35) among boys, while this association was not found among girls.

3.5 Discussion

Table 3.5 shows that there is no statistical difference on the direct effect of illegal substance use on aggression between boys (.18) and girls (.16) or on the direct effect of risky sexual behavior on aggression between boys (.08) and girls (.06). This result is consistent to the literature. In a longitudinal study of an African American young adults, Friedman et al. (1996) found that drug use predicted violent behavior for both men and women. Pepler et al. (2002) found no gender differences in the strength of association between the risk of being aggressive and substance use among Canadian adolescents.

Another finding consistent with the literature is that, among girls, depression has the highest impact on aggression and on negative body image (the latter, about four times higher than for boys). In a study with adolescents, Knox et al. (2003) found that while women with depression show more aggression, men with depression show less aggression. These authors point out that their results are different than the findings from Gjerde (1995), who found that depressed males are more prone to aggression than depressive females, but their results might have the cohort effect since adolescent aggression has increased in recent years.

Also, among girls, there is no direct effect of illegal substance use on negative body image perception, while this effect exists among boys. This can be explained by the variable “take a product or medication for weight gain” that is present on the model of the boys sample but not of the girls sample since this variable has a low correlation to other variables. Literature shows that boys can become steroid users if they have a body image disorder (Keane, 2005).

The impact of sexual behavior on negative body image perception in boys is about two times higher than in girls. Studies on the relationship between body image and risky sexual behaviors find that boys who are more confident with their bodies engage in riskier sexual behaviors (Gillen et al., 2006). However, my results show the opposite relationship. Boys with a negative body image are more susceptible to have risky sexual behaviors.

Furthermore, I find that the impact of negative body image perception on aggression is also about two times higher for boys than for girls. However, there is no literature on this relationship. Published works link body image to bullying, but in the present study, I do not use bullying since this variable was not enough correlated to any other. Some works link aggression to BMI (Gallup and Wilson, 2009). However, BMI is also not used in this work, since it is not a reliable measure for overweight or obesity, and the adolescent’s perception of their body is more important to understand their mental health than their body size. Therefore, this is an area to explore in this literature.

3.6 Conclusion

In this work, I use Directed Acyclic Graphs (DAG) and Structural Equations Modeling (SEM) to identify direct and indirect effects of negative body image perception, depression, and risk behaviors among adolescents in Brazil. Using Exploratory and Confirmatory Factor Analysis, constructs found for risk behaviors are aggression, illegal substance use, and sexual behavior. All the analyses were done for the boys

sample and the girls sample separately.

Performing DAG, I find that an indirect effect between two constructs is also explained by a direct effect. For example, in Figure 3.1, sexual behavior has an indirect effect on aggression through the negative body image perception construct. However, sexual behavior has also a direct effect on aggression. This relation can be observed in all nodes that have an indirect effect.

By performing SEM, it can be observed the effects among the constructs for each of the samples, and the differences between them, as seen in the discussion session.

A limitation of this study is that causality cannot be inferred from the findings. Although Pearl (2009) claims that DAG can infer causality, the assumptions are very restrictive and are rarely satisfied (Haughton et al., 2014). Another limitation is the use of data from a survey that was not specifically designed for the objectives of this research, resulting in the exclusion of potential variables that had problems with variable measurements.

Despite these limitations, this paper makes several contributions. First, the results of the present study show that the DAG is a reliable model to reveal directed links among variables since the findings are consistent with those in the literature. Therefore, I show that the DAG method is validated to help set up an SEM in the context of understanding adolescent behavior, in Brazil or elsewhere, even though its model is not based on theory. Second, few studies have examined body image and sexual behavior, and even fewer focus on risky sexual behavior. Since results are not consistent, this is an interesting area for future research. Third, I find an association that was not studied in the literature. Future research should continue to investigate the impact of negative body image perception on aggression, particularly among boys.

References

- Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.
- Andreyeva, T., Tripp, A. S., and Schwartz, M. B. (2015). Dietary quality of americans by supplemental nutrition assistance program participation status: a systematic review. *American journal of preventive medicine*, 49(4):594–604.
- Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist’s companion*. Princeton university press.
- Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.
- Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.
- Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107.
- Azeredo, C. M., de Rezende, L. F., Mallinson, P. A. C., Ricardo, C. Z., Kinra, S., Levy, R. B., and Barros, A. J. (2019). Progress and setbacks in socioeconomic inequalities in adolescent health-related behaviours in brazil: results from three cross-sectional surveys 2009–2015. *BMJ open*, 9(3):e025338.
- Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133.
- Becker, G. S. and Tomes, N. (1979). An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of political Economy*, 87(6):1153–1189.
- Behrman, J. R., Gaviria, A., Székely, M., Birdsall, N., and Galiani, S. (2001). Intergenerational mobility in latin america [with comments]. *Economia*, 2(1):1–44.
- Bessler, D. A. and Loper, N. (2001). Economic development: evidence from directed acyclic graphs. *The Manchester School*, 69(4):457–476.
- Bishop, C. M. (1994). Neural networks and their applications. *Review of scientific instruments*, 65(6):1803–1832.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

- Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.
- Brollo, F., Kaufmann, K., and La Ferrara, E. (2017). The political economy of program enforcement: Evidence from brazil. *Journal of the European Economic Association*.
- Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.
- Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72.
- Canella, D. S., Levy, R. B., Martins, A. P. B., Claro, R. M., Moubarac, J.-C., Baraldi, L. G., Cannon, G., and Monteiro, C. A. (2014). Ultra-processed food products and obesity in brazilian households (2008–2009). *PloS one*, 9(3):e92752.
- Cattell, R. (2012). *The scientific use of factor analysis in behavioral and life sciences*. Springer Science & Business Media.
- Child, D. (1990). *The essentials of factor analysis*. Cassell Educational.
- Connell, C. M., Gilreath, T. D., and Hansen, N. B. (2009). A multiprocess latent class analysis of the co-occurrence of substance use and sexual risk behavior among adolescents. *Journal of studies on alcohol and drugs*, 70(6):943–951.
- Cramer, J. S. (2002). The origins of logistic regression. *Tinbergen Institute Discussion Paper*.
- Cristianini, N., Shawe-Taylor, J., et al. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.
- D’Agostino Jr, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine*, 17(19):2265–2281.
- Dave, D. and Rashad, I. (2009). Overweight status, self-perception, and suicidal behaviors among adolescents. *Social Science & Medicine*, 68(9):1685–1691.
- De Bem Lignani, J., Sichieri, R., Burlandy, L., and Salles-Costa, R. (2011). Changes in food consumption among the programa bolsa família participant families in brazil. *Public health nutrition*, 14(5):785–792.
- De Brauw, A., Gilligan, D. O., Hoddinott, J., and Roy, S. (2015). The impact of bolsa familia on schooling. *World Development*, 70:303–316.
- Edmonds, A. (2007). the poor have the right to be beautiful: cosmetic surgery in neoliberal brazil. *Journal of the Royal Anthropological Institute*, 13(2):363–381.

- Farrell, M. H., Liang, T., and Misra, S. (2018). Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*.
- Ferguson, E. and Cox, T. (1993). Exploratory factor analysis: A users guide. *International journal of selection and assessment*, 1(2):84–94.
- Fernald, L. C., Gertler, P. J., and Hou, X. (2008). Cash component of conditional cash transfer program is associated with higher body mass index and blood pressure in adults. *The Journal of nutrition*, 138(11):2250–2257.
- Franckle, R. L., Moran, A., Hou, T., Blue, D., Greene, J., Thorndike, A. N., Polacsek, M., and Rimm, E. B. (2017). Transactions at a northeastern supermarket chain: differences by supplemental nutrition assistance program use. *American journal of preventive medicine*, 53(4):e131–e138.
- Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *In Machine Learning: Proceedings of the Thirteenth International Conference*, volume 96, pages 148–156. Morgan Kaufman, San Francisco.
- Friedman, A. S., Kramer, S., Kreisher, C., and Granick, S. (1996). The relationships of substance abuse to illegal and violent behavior, in a community sample of young adult african american men and women (gender differences). *Journal of substance abuse*, 8(4):379–402.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- Gallup, A. C. and Wilson, D. S. (2009). Body mass index (bmi) and peer aggression in adolescent females: An evolutionary perspective. *Journal of Social, Evolutionary, and Cultural Psychology*, 3(4):356.
- Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., and Aldridge, M. D. (2014). Methods for constructing and assessing propensity scores. *Health services research*, 49(5):1701–1720.
- GeNIe Modeler, BayesFusion, L. (2019). *BayesFusion, LLC*. University of Pittsburgh.
- Gillen, M. M., Lefkowitz, E. S., and Shearer, C. L. (2006). Does body image play a role in risky sexual behavior and attitudes? *Journal of Youth and Adolescence*, 35(2):230–242.
- Gjerde, P. F. (1995). Alternative pathways to chronic depressive symptoms in young adults: Gender differences in developmental trajectories. *Child Development*, 66(5):1277–1300.
- Glewwe, P. and Kassouf, A. L. (2012). The impact of the bolsa escola/familia conditional cash transfer program on enrollment, dropout rates and grade promotion in brazil. *Journal of development Economics*, 97(2):505–517.

- Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001.
- Griffin, B. A., McCaffrey, D. F., Almirall, D., Burgette, L. F., and Setodji, C. M. (2017). Chasing balance and other recommendations for improving nonparametric propensity score models. *Journal of causal inference*, 5(2).
- Guo, S., Barth, R. P., and Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review*, 28(4):357–383.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L., et al. (1998). *Multivariate data analysis*, volume 5. Prentice hall Upper Saddle River, NJ.
- Hallfors, D. D., Waller, M. W., Ford, C. A., Halpern, C. T., Brodish, P. H., and Iritani, B. (2004). Adolescent depression and suicide risk: association with sex and drug behavior. *American journal of preventive medicine*, 27(3):224–231.
- Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning*, 77(1):103–123.
- Haughton, D. and Haughton, J. (2011). *Living standards analytics: Development through the lens of household survey data*. Springer Science & Business Media.
- Haughton, D., Hua, G., Jin, D., Lin, J., Wei, Q., and Zhang, C. (2014). Optimization of the marketing mix in the health care industry. *arXiv preprint arXiv:1403.7971*.
- Haughton, D., Kamis, A., and Scholten, P. (2006). A review of three directed acyclic graphs software packages: Mim, tetrad, and winmine. *The American Statistician*, 60(3):272–286.
- Jacobsen, J. P., Levin, L. M., and Tausanovitch, Z. (2016). Comparing standard regression modeling to ensemble modeling: How data mining software can improve economists predictions. *Eastern Economic Journal*, 42(3):387–398.
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.
- Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: propensity scores. *American journal of epidemiology*, 150(4):327–333.
- Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636.
- Kawerk, E. (2020). *Gradient Boosting*. DataCamp.
- Keane, H. (2005). Diagnosing the male steroid user: Drug use, body image and disordered masculinity. *Health*, 9(2):189–208.

- Keller, B., Kim, J.-S., and Steiner, P. M. (2015). Neural networks for propensity score estimation: Simulation results and recommendations. In *Quantitative psychology research*, pages 279–291. Springer.
- Keller, B. S., Kim, J.-S., and Steiner, P. M. (2013). Data mining alternatives to logistic regression for propensity score estimation: Neural networks and support vector machines. *Multivariate behavioral research*, 48(1):164–164.
- Knox, M., Carey, M., and Kim, W. J. (2003). Aggression in inpatient adolescents: The effects of gender and depression. *Youth & Society*, 35(2):226–242.
- Larose, D. T. (2015). *Data mining and predictive analytics*. John Wiley & Sons.
- Laus, M. F., Costa, T. M. B., and Almeida, S. S. (2011). Body image dissatisfaction and its relationship with physical activity and body mass index in brazilian adolescents. *Jornal Brasileiro de Psiquiatria*, 60(4):315–320.
- Laus, M. F., Miranda, V. P. N., Almeida, S. S., Braga Costa, T. M., and Ferreira, M. E. C. (2013). Geographic location, sex and nutritional status play an important role in body image concerns among brazilian adolescents. *Journal of health psychology*, 18(3):332–338.
- Leadbeater, B. J., Kuperminc, G. P., Blatt, S. J., and Hertzog, C. (1999). A multivariate model of gender differences in adolescents’ internalizing and externalizing problems. *Developmental psychology*, 35(5):1268.
- Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346.
- Martins, A. P. B., Levy, R. B., Claro, R. M., Moubarac, J. C., and Monteiro, C. A. (2013). Increased contribution of ultra-processed food products in the brazilian diet (1987-2009). *Revista de saude publica*, 47(4):656–665.
- McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.
- McGrath, R. J., Julie, W., and Caron, R. M. (2009). The relationship between resilience and body image in college women. *The Internet Journal of Health*.
- Menezes, F., Gadelha, E., Santarelli, M., Billo, R., Costa, D. M., Curado, J. C., Burlandy, L., Magalhães, R., da Costa, R. S., de Magalhães, I. B., et al. (2008). Repercussões do programa bolsa família na segurança alimentar e nutricional das famílias beneficiadas. *Rio de Janeiro: IBASE*.
- Monteiro, C. A., Levy, R. B., Claro, R. M., de Castro, I. R. R., and Cannon, G. (2010). Increasing consumption of ultra-processed foods and likely impact on human health: evidence from brazil. *Public health nutrition*, 14(1):5–13.

- Moran, A. J., Khandpur, N., Polacsek, M., and Rimm, E. B. (2019). What factors influence ultra-processed food purchases and consumption in households with children? a comparison between participants and non-participants in the supplemental nutrition assistance program (snap). *Appetite*, 134:1–8.
- Nevitt, J. and Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural equation modeling*, 8(3):353–377.
- Nguyen, C., Wang, Y., and Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*, 6(5).
- OConnell, M. E., Boat, T., Warner, K. E., et al. (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities*, volume 7. Washington, DC: National Academies Press.
- Oliveira, W. A. d., Silva, M. A. I., Mello, F. C. M. d., Porto, D. L., Yoshinaga, A. C. M., and Malta, D. C. (2015). The causes of bullying: results from the national survey of school health (pense). *Revista latino-americana de enfermagem*, 23(2):275–282.
- Paxton, R. J., Valois, R. F., Watkins, K. W., Huebner, E. S., and Drane, J. W. (2007). Associations between depressed mood and clusters of health risk behaviors. *American journal of health behavior*, 31(3):272–283.
- Pearl, J. (2009). *Causality*. Cambridge university press.
- Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.
- Pepler, D. J., Craig, W. M., Connolly, J., and Henderson, K. (2002). Bullying, sexual harassment, dating violence, and substance use among adolescents. *The violence and addiction equation: Theoretical and clinical issues in substance abuse and relationship violence*, pages 153–168.
- Popkin, B. M., Adair, L. S., and Ng, S. W. (2012). Global nutrition transition and the pandemic of obesity in developing countries. *Nutrition reviews*, 70(1):3–21.
- Ramírez-Silva, I., Rivera, J. A., Leroy, J. L., and Neufeld, L. M. (2013). The oportunidades program’s fortified food supplement, but not improvements in the home diet, increased the intake of key micronutrients in rural mexican children aged 12–59 months. *The Journal of nutrition*, 143(5):656–663.
- Ratkovic, M. (2014). Balancing within the margin: Causal effect estimation with support vector machines. *Department of Politics, Princeton University, Princeton, NJ*.

- Rogers, E. S., Dave, D. M., Pozen, A., Fahs, M., and Gallo, W. T. (2018). Tobacco cessation and household spending on non-tobacco goods: results from the us consumer expenditure surveys. *Tobacco control*, 27(2):209–216.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Saez, E. (2002). Optimal income transfer programs: intensive versus extensive labor supply responses. *The Quarterly Journal of Economics*, 117(3):1039–1073.
- Salford Systems, M. (2018). *Minitab LLC*. San Diego, CA, USA.
- Sarma, K. S. (2017). *Predictive modeling with SAS Enterprise Miner: Practical solutions for business applications*. SAS Institute.
- SAS Enterprise Miner, S. I. (2018). *SAS Institute Inc*. Cary, NC, USA.
- Schaffland, E. (2011). Conditional cash transfers in brazil: Treatment evaluation of the ‘bolsa família’ program on education. Technical report, Courant Research Centre: Poverty, Equity and Growth-Discussion Papers.
- Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555.
- Shei, A., Costa, F., Reis, M. G., and Ko, A. I. (2014). The impact of brazils bolsa família conditional cash transfer program on childrens health care utilization and health outcomes. *BMC international health and human rights*, 14(1):10.
- Silva, R. A. d., Cardoso, T. d. A., Jansen, K., Souza, L. D. d. M., Godoy, R. V., Cruzeiro, A. L. S., Horta, B. L., and Pinheiro, R. T. (2012). Bullying and associated factors in adolescents aged 11 to 15 years. *Trends in psychiatry and psychotherapy*, 34(1):19–24.
- Simões, A. A. and Sabates, R. (2014). The contribution of bolsa família to the educational achievement of economically disadvantaged children in brazil. *International Journal of Educational Development*, 39:141–156.
- Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., and Richardson, T. (2000). *Causation, prediction, and search*. MIT press.
- SPSS, I. S. S. (2012). *IBM SPSS Statistics*. Armonk, NY, USA.
- SPSS Amos, I. S. S. (2012). *IBM SPSS Statistics*. Armonk, NY, USA.
- Stata Statistical Software, S. (2017). *StataCorp LLC*. College Station, TX, USA.
- Tavares, L. F., Castro, I. R. R. d., Levy, R. B., Cardoso, L. O., Passos, M. D. d., and Brito, F. d. S. B. (2014). Validade relativa de indicadores de práticas alimentares da pesquisa nacional de saúde do escolar entre adolescentes do rio de janeiro, brasil. *Cadernos de Saúde Pública*, 30:1029–1041.

- Thoemmes, F. J. and Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate behavioral research*, 46(1):90–118.
- Watkins, S., Jonsson-Funk, M., Brookhart, M. A., Rosenberg, S. A., O’Shea, T. M., and Daniels, J. (2013). An empirical comparison of tree-based methods for propensity score estimation. *Health services research*, 48(5):1798–1817.
- Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology*, 63(8):826.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.
- Yung, Y.-F. and Bentler, P. M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. *Advanced structural equation modeling: Issues and techniques*, pages 195–226.
- Zhao, P., Su, X., Ge, T., and Fan, J. (2016). Propensity score and proximity matching using random forest. *Contemporary clinical trials*, 47:85–92.

Vita

Fernanda M. Araujo Maciel graduated from Universidade Federal do Rio de Janeiro (UFRJ) in Brazil with a Bachelor of Science in Statistics in 2009. She earned her Master of Science in Business Administration and Marketing from Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) in 2011. In 2014, she earned her Master of Science in Marketing Analytics from Bentley University. After a total of 5 years of working experience in areas such as database marketing, analytics research, and consulting, in 2015 she joined the Bentley University Ph.D. program.

This manuscript was typed by the author.