UNIVERSITY OF CALIFORNIA
Los Angeles

Balance Tests as a Learning Problem:
Assessing 3,000 Lotteries with Machine Learning

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Fernando Barros de Mello

2023

ABSTRACT OF THE THESIS

Balance Tests as a Learning Problem:
Assessing 3,000 Lotteries with Machine Learning

by

Fernando Barros de Mello
Master of Science in Statistics
University of California, Los Angeles, 2023
Professor Chad J. Hazlett, Chair

This thesis proposes a way to look beyond mean balance tests. Balance tests are a key component for plausible causal identification. Both in experimental designs and natural experiments, the standard practice is to use statistical tests with the null hypothesis of no difference between the pre-treatment covariates across treatment and control groups. The main idea is that the distributions of pretreatment variables should be roughly balanced between treatment and control groups. In recent decades, presenting evidence for the quality of the causal research designs became standard in social science. While observational researchers normally focus on evidence of balance for the covariates included in their model, experimental researchers provide randomization tests for balance on pretreatment covariates.

The thesis asks if any functions of the observed covariates, linear or otherwise, is able to predict who wins lotteries used to distribute houses to low-income citizens in Brazil. My final dataset contains 1,777,385 observations (winners and losers of the lotteries), distributed among 3,012 lotteries. By using a random forest algorithm that attempts to predict who will win the lottery, I extract from each lottery an out-of-sample estimate of model fit, including the AUC, the prediction

R^2 (the correlation of the estimated probability of winning with the actual indicator for winning, squared), and corresponding p-value. The distribution of these estimates across lotteries are shown.

A large-scale housing program in Brazil is used to assess the use of machine learning as a balance test tool. The *Minha Casa Minha Vida program* (MCMV) awards heavily subsidized mortgages to low-income citizens through a lottery system at the municipal level. The thesis focuses on a bracket of the program in which housing assignments are made by lottery. Between 2009 and 2017, the MCMV Faixa 1 program resulted in contracts for 1.4 million housing units in almost 3000 municipalities. With over R$74 billion (USD 18.5 billion) spent in Faixa 1 benefits, it is one of the largest lottery-based housing projects in the world, according to the Brazilian government, and the largest housing program ever implemented in Latin America.

The thesis of Fernando Barros de Mello is approved.

Mark S. Handcock

Jeffrey B. Lewis

Chad J. Hazlett, Committee Chair

University of California, Los Angeles

2023

To Fabiola, who optimizes happiness

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Chad Hazlett is a true scientist: dedicated, committed to the pursuit of the truth, and always pushing for methodical advancements. Chad has encouraged me to learn methods that I never thought I would be able to. He also knows that commitment to the truth does not need to translate into harshness. His kindness has a huge effect on those how are not naturally comfortable with methods. I owe a lot to Chad.

I am grateful to Mark Handcock, who has guided me through the master's program, and advised me over the years. I have taken all the available classes offered by Professor Handcock, whose suggestions helped me not only in the MS, but also in my P.h.D. Since the beginning of my studies, Jeff Lewis has taught me how to navigate in the world of statistics, and encouraged me to pursue different projects. Even when I just needed some advice for an initial idea of research, Jeff was always available and provide feedback.

VITA

2023	Instructor of Record and Research Fellow, Government Department, Georgetown University.
2022	D.Phil in Political Science — University of California, Los Angeles
2020	Instructor of Record, Political Science Department, UCLA
2017–2021	Teaching Assistant, Political Science Department, UCLA
2015	MA in Latin American Studies and Government — Georgetown University

PUBLICATIONS

Seligman, Milton, and **Fernando B. Mello**. "Lobbying Uncovered: Corruption, Democracy and Public Policy in Brazil." *Wilson Center* , 2020.

Colomer, Josep, David Banerjea, and **Fernando B. Mello**. "To democracy through anocracy." *Democracy Society* , 2016.

CHAPTER 1

Introduction

Balance tests are a key component for plausible causal identification. Both in experimental designs and natural experiments, the standard practice is to use statistical tests with the null hypothesis of no difference between the pre-treatment covariates across treatment and control groups. The main idea is that the distributions of pretreatment variables should be roughly balanced between treatment and control groups. In recent decades, presenting evidence for the quality of the causal research designs became standard in social science (B. B. Hansen 2008; Hartman and Hidalgo 2018). While observational researchers normally focus on evidence of balance for the covariates included in their model, experimental researchers provide randomization tests for balance on pretreatment covariates.

But how similar should these covariates be? Previous research (Ben B. Hansen and Bowers 2008) asked a series of questions such as: Can descriptive comparisons meaningfully be paired with significance tests? Should there be several balance tests, one for each pretreatment variable, or an omnibus test? Which tests of balance would be optimal? Dunning (2010) identifies an increasing concern in social science with foundational issues of research design,

This thesis uses a large-scale housing program in Brazil to assess the use of machine learning as a balance test tool. [1] The *Minha Casa Minha Vida program* ("My House, My Life", henceforth MCMV) awards heavily subsidized mortgages to low-income citizens through a lottery system at the municipal level. The thesis focuses on a bracket of the program in which housing assignments are made by lottery. This bracket targets families with monthly income below R$1800 (USD 400).

[1] The Pre-Analysis Plan is registered at OSF

For lottery winners, 90% or the cost of a home is subsidized, with the remaining 10% paid interest free over 10 years.[2]

Between 2009 and 2017, the MCMV Faixa 1 program resulted in contracts for 1.4 million housing units in almost 3000 municipalities. With over R$74 billion (USD 18.5 billion) spent in Faixa 1 benefits, it is one of the largest lottery-based housing projects in the world, according to the Brazilian government, and the largest housing program ever implemented in Latin America (Biderman, Ramos, and Hiromoto 2018).

1.1 Balance tests and natural experiments

Researchers use randomization to eliminate confounding (Duflo, Glennerster, and Kremer 2006; Fisher 1935). But depending on the question, experimental research can be expensive, impractical, or unethical. Natural experiments are increasingly used to identify real world situations with as-if random assignment (T. Dunning and Brady 2010; Thad Dunning 2012; Gerber and Green 2012; Gerber, Green, and Larimer 2008; Sekhon 2009). In these cases, assignment happens due to social and political processes, and researchers look for opportunities to use natural experiments in the analysis of observational data.

In natural experiments, outcomes are compared across subjects exposed to treatment and control conditions (Collier et al. 2009; T. Dunning and Brady 2010). Nevertheless, subjects are almost always assigned to the treatment not at random, but rather as-if at random. Thus, the treatment manipulation was not controlled by the researcher – the reason why the study can be considered observational. The responsibility to make a credible claim that the assignment of subjects to treatment and control conditions is as-if random falls upon the academic.

Collier et al. (2009) and T. Dunning and Brady (2010) describe the paradigmatic example of

[2]This bracket is called Faixa 1. There are three other brackets: Faixa 1.5, 2 and 3, which have higher interest rates and cover families with higher income. Because the government subsidy is lower in these other brackets, they are not administered by lottery; they are essentially low-interest mortgages.

natural experiments. The 19th century epidemiologist Snow demonstrated that cholera was a waterborne infectious disease. To do that, Snow employed a natural experiment, which compared households that received water from two different companies. The Southwark and Vauxhall company distributed contaminated water, and households served by it had a death rate between eight and nine times as great as in the houses supplied by the Lambeth company, which supplied relatively pure water.

The first task was to support the claim that the allocation of water had occurred as-if at random. The study is based on the idea that distribution of water did not follow a systematic plan, that close houses would not receive water from the same company, and households would not decide where to live depending on the water company. After that, Snow's analysis was made by comparing the incidence of cholera per 10,000 houses considering different water supply companies. Collier et al. (2009) argues that before deciding the quantitative study, Snow followed key qualitative steps:

1. Investigating if conventional theories were wrong
2. formulating the water hypothesis
3. Noticing that in 1852, the Lambeth company moved its intake pipe to obtain relatively pure water, while Southwark and Vauxhall continued to draw heavily contaminated water.

1.2 Housing programs and natural experiments

MCMV is not the only housing program that provides a natural experiment for researchers. Previously, Galiani, Gertler, and Schargrodsky (2005) and Galiani and Schargrodsky (2004) provide credible causal claims on the effects of property rights and land titles on the development of poor communities in Argentina. Between 1981 and 1982, 2,000 families occupied over two square kilometers of vacant land in the province of Buenos Aires, dividing and allocating the land to individual families. After the country transitioned back to democracy in 1984, a new law transferred the titles to the squatters. However, in some cases, original landowners challenged the measure

in court. As a result, while some squatters faced long delays in the transfer of titles, others were granted the titles immediately.

Therefore, in this case, there was not a government sanctioned lottery to distributed houses, but the social situation created a treatment group of squatters (those for whom the titles were granted) and a control group (to whom titles were not granted). The authors compare the two groups across different social development indicators: average housing investment, household structure, and educational attainment of children.

Galiani and Schargrodsky (2004) present various kinds of evidence to support the key claim that land titles were assigned as-if at random. First, before the occupation of the land, transfer of titles to squatters could not have been predicted – after all Argentina was still a dictatorship during that time. Moreover, titled and untitled parcels occupied the area, side by side, the families had similar characteristics. The paper compares average parcel characteristics for the group that was offered property rights and the group that was not. The variables are parcel surface (in squared meters), distance to a nearby creek (in blocks), and a dummy for whether the parcel is on a corner of a block.

In addition, the authors test the null hypotheses of absence of differences between these two groups for a set of pre-treatment household characteristics: age of the household head, gender of the household head, nationality of the household head, years of education of the household head, nationality of the father of the household head, years of education of the father of the household head, nationality of the mother of the household head, and years of education of the mother of the household head. Finally, using interviews, the authors argue that factors unrelated to the characteristics of squatters or their parcels explain some owners? decisions to challenge expropriation.

Natural experiments and as-if random treatment assignment may stem from different sources (T. Dunning and Brady 2010; Thad Dunning 2012):

1. A procedure specifically designed to randomize (i.e., a lottery)

2. The nonsystematic implementation of certain interventions

3. Arbitrary division of units by jurisdictional borders

The plan of the thesis is the following. **Chapter 2** describes the lottery system of MCMV. **Chapter 3** reviews recent discussions on types of balance tests. **Chapter 4** then describes the data construction and the identification strategy for the lottery-level balance testing on pre-treatment covariates and sorting estimates by reliability of the lottery. The identifying assumption is not that winning the lottery is ignorable in a pooled dataset; rather, it is ignorable only by each lottery. **Chapter 5** present the results of using machine learning methods to predict treatment assignment in each lottery. Finally, **Chapter 6** describes risks for future inference of using untrustworthy lotteries, and **Chapter 7** discusses the advantages of using a non-parametric model as balance test.

MCMV program has already been used to test social science hypotheses (Bueno, Nunes, and Zucco 2022). But any causal claim needs to be based on the confidence we have on the random assignment of the lotteries. That is the key goal of this thesis. The basic question that must be asked: Is this a valid natural experiment?

CHAPTER 2

The Minha Casa Minha Vida Program

Governments around the world have funded massive housing programs, giving away or heavily subsidizing homes for low-income citizens. Since the 1990s the World Bank points out to the importance of policy differences in shaping housing sector outcomes is supported by recent data on 52 countries collected by the Housing Indicators Program, a joint program of the United Nations Centre for Human Settlements and the World Bank. According to the McKinsey Global Institute, currently, the affordable housing gap stands at $650 billions a year and current trends suggest that there could be 106 million more low-income urban households by 2025 (Woetzel et al. 2014).

In this project, I use the Brazilian large-scale housing program to assess how useful is to use machine learning as a balance test tool. MCMV is a housing program that randomly awards houses to low-income citizens in a municipality. The program was created in May 2009, when the then President Dilma Rousseff, from the Workers' Party (PT), vetoed the part of the law establishing the use of the lottery. However, in March 2010, the Ministry responsible for MCMV, following the determination of the auditing court, established the use of lotteries again [1]. The program works as follows: The mayors present projects to the federal government, which has the power to decide where to allocate the budget. Once a project is approved for a municipality, the local government organizes a list of people who wish to enter into a lottery to get a subsided house. The federal government subsidizes 90% of the cost of the house, but people still need to guarantee monthly payments for over 10 years. The average monthly installment is R$ 80 (US$ 20) and the maximum

[1]The Ministerial Orders establishing lotteries are: 163, from 05/06/2016; 412, from 08/06/2015; 595, from 12/18/2013; 610, from 12/26/2011 and 140, from 03/05/2010.

is R$ 270 (US$ 90). There is no interest rate charged in the installments. Selling or renting the house is forbidden. In other words, the ministry has the power to decide where the houses are going to be built, but not who gets them.

Even so, the Minister of municipalities still has some discretionary power to decide which municipalities will receive the program, although technical studies and the support of the national bank are necessary to move a project forward. Official data show that 96% of the municipalities in the country received houses from the program (Consultoria de Orentos, Fiscaliza e Controle, 2016).

2.1 Use of lotteries: a natural experiment

MCMV can be classified, at least in part, as programmatic policy. Not only does it distribute the houses using a lottery, but all the legislation involving it has been approved by the majority of parties, and most of the bills were approved unanimously, including the opposition. The original bill, for instance, was converted from an executive order in the Chamber of Deputies and in the Senate with voice votes, when no politician voted against the law. In the Chamber of Deputies, there were discussions and three versions of the bill before a consensus was achieved. The politician in charge of the discussions and who bargained for the approval was Henrique Eduardo Alves, then the speaker of the House, and the leader of the largest party in the Chamber, PMDB. Other bills that updated the program were also approved unanimously [2]. When the first bill was voted, in 2009, Mr. Alves defended the implementation of the lottery and declared: "We must avoid any politicking or interference of any party or power" (Agencia Camara 2009). The program is largely financed by a state-owned federal bank, Caixa Econa Federal (CEF).

[2]The first bill was voted in the Chamber on 05/20/2009. After three drafts of the bill a consensus was achieved and all parties sup-ported the text. The project was discussed by eight deputies from the main government and opposition coalition parties: Dep. Ema-nuel Fernandes (PSDB-SP), Dep. Arnaldo Jardim (PPS-SP), Dep. Fernando Chucre (PSDB-SP), Dep. Fernando Coruja (PPS-SC), Dep. Luiz Carlos Hauly (PSDB-PR), Dep. Joliveira (DEM-TO), Dep. Ivan Valente (PSOL-SP), Dep. Lincoln Portela (PR-MG), Dep. Josno (PT-SP) e Dep. Vicentinho (PT-SP).

By the rules of the program, politicians cannot use discretionary power to exclude people from participating in the lottery, but there are different criteria for eligibility. As discussed in the introduction, in this thesis, I analyze the lowest bracket of MCMV, which contains families with monthly income bellow R$ 1,800 (US$ 450). Called bracket 1, this is the only part of the program that allocates houses through lotteries, which happen at the municipal level. The federal government defines bracket 1 as: "families with monthly gross income of up to R$ 1,800, as established in Ordinance No. 99, of March 30, 2016, which are fully retained from the General Budget of the Union and therefore does not constitute a financing, but a subsidy transfer program."[3]

Citizens need to register to participate in the lottery. In fact, many municipalities create a unique dataset to conduct the lotteries. In large capital municipalities, these lists have tens of thousands of people. In addition, the vast majority of municipalities use the numbers from the weekly federal lottery (known as *Mega Sena*, which distributes cash awards) to allocate the houses. Sometimes, the lottery is conducted in live events, in the presence of politicians and the population. Different lotteries can be find on Youtube videos. When the lottery is conducted by the city hall, external actors need to be present, such as representatives from the civil society and from auditing courts.

There are also some criteria to give priority in a lottery. Elderly citizens, women who are in charge of their family or people with some types of handicap are gathered in different polls to get their own share of houses. Even if politicians cannot take the house from a beneficiary, it does not mean that a person cannot lose the house. There are also rules that regulate when a house can be taken back by the federal government. First, selling or renting the houses is forbidden. In addition, a person can lose their house when having three or more late installments. The cases are analyzed individually and depend on the federal bank, CEF, to confiscate the house. Between 2009 and 2016, the program awarded 1,137,547 houses in the lower-income bracket. The amount spent by the federal government for this group was R$ 76.3 billion (US$ 25 billion). Between 2009

[3] Answer to a FOIA request submitted in 2017.

and 2016, only 1,261 houses were taken back and there were more 1,205 litigations in process . Therefore, only 0.21% of the total houses distributed in the same period (1,137,547) were either taken back or in the litigious process, which means that the probability of losing the house is very small. [4]

2.2 Qualitative evidence

Before applying statistical methods, to check the fairness of the lottery, I have interviewed two of the responsible for auditing the program in the Federal Court of Accounts (TCU).[5] The court has found no evidence about frauds in the lotteries, which does not guarantee the fairness of the lotteries. In a different audit, the Comptroller General of the Union (CGU) has investigated 195 projects in 110 cities from 20 Brazilian states from January 2012 and February 2014. The sample represents a universe of 688 projects, with a total investment of R$8.3 billion.

The auditing has not pointed to any problems with the fairness of the lottery, but indicated some issues in terms of transparency. In at least 52.5% of the cases, the result of the selection of the beneficiaries was disclosed to the population. On the selection of beneficiaries, the CGU has pointed to some positive characteristics of the program such as (i) clear guidelines in selection criteria; (ii) mandatory publication of the register of candidates for beneficiaries; and (iii) transparency of lotteries and dissemination of the results of the selection. In addition, the audit revealed that, considering all the beneficiaries audited through the data, the number of records with evidence of inconsistency was very low (1,258 out of 186,271, or about 0.7% of the total). In 2016, the Federal Prosecution Service pressured the federal government to create an online tool to gather the infor-

[4]Primary information collected using FOIA request.

[5]The Federal Court of Accounts is the external control institution of the federal government that supports the National Congress with the mission of overseeing the budget and the financial execution. TCU is responsible for evaluating and judging the accounts of administrators and of other individuals responsible for federal public money, assets and values. TCU also evaluates and judges the ac-counts of those who have caused loss, misappropriation or other irregularity resulting in losses to the public treasury.

mation about the participants of the lottery and to promote lotteries using computers. "Now, the draw is done by the electronic system among the beneficiaries who are automatically prioritized. So there is no possibility of irregular prioritization or turning the lottery into a political-electoral event", said the prosecutor responsible for the system in an interview for the dissertation.[6]

As discussed by T. Dunning and Brady (2010) and Thad Dunning (2012), the quality of the assertion of as-if random assignment from policy implementation may vary. Dunning emphasizes that qualitative evidence is central to validating the natural experiment. He suggests interviews with key officials to understand if the rules are manipulated or respected. TCU has not yet conducted any test to assess the presence or fairness of the lottery. In 2010, one year after the beginning of the program, the court audited some municipalities and recommended that the Office of the Comptroller General of the Union include, within the scope of the municipal inspections carried out through a lottery, the verification of points relevant to the selection of beneficiaries of the MCMV to increase the isonomy, impersonality, publicity, and effectiveness of the subsidies granted.[7]

The goal of the thesis is to provide more systematic tests to assess the quality of the lotteries.

[6]For more information: http://www.mpf.mp.br/pgr/noticias-pgr/minha-casa-minha-vida-atuacao-do-mpf-resulta-em-criacao-de-cadastro-nacional-do-programa

[7]Decision by the TCU on the case TC-028.461/2010-0.

CHAPTER 3

Existing Approaches to balance tests

One of the reasons randomized experiments are appealing to social scientists is that, in expectation, they achieve "balance" on all pre-treatment-assignment variables (covariates), both measured and unmeasured. This means that the distributions of covariates differ only randomly between the treatment and control units. For that reason, careful procedures to check the assumptions justifying a causal design are as important as those used to estimate causal effects (Rubin 2008). For example, as evidence in favor of their designs, observational researchers are expected to show covariate balance, while experimental researchers present randomization checks for balance on pretreatment covariates.

For many years, causal inference based on randomized experiments (Cochran and G. M. Cox 1992; D. R. Cox 1958; Fisher 1935; Kempthorne 1952) was considered to be a distinct endeavor than causal inference based on observational data (Blalock 1964; Campbell 1970; Cook 1979; Kenny 1979). According to Rubin (2008), this changed in the 1970?s when potential outcomes started to be used to define causal effects in both randomized experiments and observational studies. This chapter discusses the use of existing approaches to balance tests in observational studies, mainly natural experiments.

3.1 Pre-treatment difference-in-means

Given the fact that natural experiments claim that as-if randomness is present in the study, it becomes necessary that studies evaluate the plausibility of such claims. In expectation, randomness

or as-if randomness should ensure that assignment is statistically independent of factors that could influence outcomes. In other words, before making any causal claim, it is necessary to test the assumption that the treatment assignment is unconfounded. Even in observational settings, researchers seek evidence that they are using data consistent with the identifying assumptions (Imbens and Rubin 2015).

Causal research designs follow certain testable implications. Theoretically, unconfoundedness means the same distributions of the potential outcomes for treatment and control groups. It is impossible to test this distribution of the potential outcomes. In other words, causal inference requires assumptions that cannot be verified by the data (Pearl 2009). Yet it is possible to test how similar the two groups are considering pretreatment covariates, which is commonly called a "balance test." The most common analysis for pre-treatment balance tests (or randomization checks in the experimental design literature) is a simple and transparent comparison of distributions or of means, to check if pretreatment variables are approximately the same among treatment and control units.

A large number of pretreatment covariates balanced across treatment and control group is often used to test the credibility of the design.[1] However, T. Dunning and Brady (2010) argues that the credibility of the statistical model is not inherent in all studies that claim to use natural experiments. After analyzing 29 studies that claimed to use natural experiment, the author finds out that more than 40% did not use simple, unadjusted difference-of-means test to evaluate the null hypothesis of no effect of treatment on pre-treatment covariates. **Table 3.1** reproduces some of the papers used in Dunning's analysis.

As discussed by (Hartman and Hidalgo 2018), the balance test standard employs statistical tests with a null hypothesis of no differences between treatment and control groups as an indirect way of testing whether the data are consistent with an unconfounded design. In the standard practice, a valid research design is supported if the statistical test fails to provide evidence in favor of a

[1]Even in experiments – where unconfoundedness is expected to be achieved via randomization– it is standard practice to check for "bad draws," when imbalances on covariates could bias the results.

Authors	Source of alleged Natural Experiment	Standard Natural Experiment	Simple difference of means test
Angrist and Lavy (1999)	Discontinuities introduced by enrollment ceilings on class sizes	No	Yes
Ansolabehere, Snyder, and Stewart (2000)	Electoral redistricting	Yes	Yes
Banerjee and Iyer (2005)	Land tenure patterns instituted by British in colonial India	Yes	No
Berger (2009)	The division of northern and southern Nigeria	Yes	No
Blattman (2008)	As-if random abduction of children by the Lord‘s Resistance Army	Yes	No
Brady and McNulty (2004)	Precinct consolidation in California gubernatorial recall election	Yes	Yes
Card and Krueger (1994)	Differential exposure to minimum-wage laws among fast-food restaurants on the New Jersey-Pennsylvania border	Yes	Yes
Chattopadhyay and Duflo (2004)	Random assignment of quotas for village council presidencies	Yes	Yes
Cox, Rosenbluth, and Thies (2000)	Cross-sectional and temporal variation in institutional rules in Japanese parliamentary houses	Yes	Yes
Doherty, Green, and Gerber (2006)	Random assignment of lottery winnings, among lottery players	Yes	No
Ferraz and Finan (2008)	Release of randomized corruption audits in Brazil	Yes	Yes
Galiani and Schargrodsky (2004)	Judicial challenges to transfer of property titles to squatters	Yes	Yes
Ho and Imai (2008)	Randomized ballot order under alphabet lottery in California	Yes	Yes
Titiunik (2008)	Random assignment of U.S. state senate seats to two or four year terms after reapportionment	Yes	Yes

Table 3.1: A non-exhaustive list of studies in political science using natural experiment shows how even simple pre-treatment difference of means test is not always used (Dunning 2010).

difference (i.e., a large p-value). However, balance tests risk to confound low power with the acceptance of the null hypothesis of no differences between the groups. One solution is to conduct permutation-based inference. Randomization inference on pre-treatment covariates, allow to test the probability that the differences between groups could be explained by chance.

A different approach is to conduct omnibus tests for overall balance of the pre-treatment covariates (Caughey, Dafoe, and Seawright 2017). Nonparametric combination assesses the joint probability of observing the theoretically predicted pattern of results under the sharp null of no effects. This approach accounts for the dependence among the component tests without relying on modeling assumptions or asymptotic approximations.

Finally, placebo tests are used to test the effect of a treatment on a outcome that is know by the researchers to be unaffected by the dependent variable. If the placebo test shows a correlation with the placebo outcome, the validity of the analyses is called into question.

3.2 Recent approaches for balance tests

In this section, I will discuss two recent approaches that can be used to strength the analyses of causal research designs. Gross (2015) and Rainey (2014) discuss if cases with no statistically significant difference between groups is not sufficient evidence for showing substantive insignificance. These authors recommend the use of the $100(1 - 2 * \alpha)$ confidence interval and suggest evaluating whether confidence range of the parameter lies within ("negligible") or outside ("substantively significant") the null effect range.

Hartman and Hidalgo (2018) propose the use of equivalence approaches when conducting balance tests. In their view, researchers should "begin with the initial hypothesis that the data are inconsistent with a valid research design, and only reject this hypothesis if they provide sufficient statistical evidence in favor of data consistent with a valid design" , instead of a null hypothesis of no difference between treatment and control groups.

The most important proposed difference between equivalence testing and tests of difference is

whether or not the researcher needs to make an ex ante decision over what range of values to define as "similar" versus "different." In equivalence tests, an equivalence range is set, defining values within which the difference between the two variables is substantively inconsequential. Choosing appropriate values for the upper and lower bounds is the most important aspect of equivalence testing. The test is conducted using two one-sided t-tests, and the null of difference is rejected in favor of equivalence if the p-value for both one-sided tests is less than α.

These different tests of design focus on the analysis of observable imbalance. In addition, they help to understand how, under certain assumptions, the imbalance is able to impact estimates. However, these tests do not offer any information about unobservable imbalance. That is why a different line of research indicates the importance of conducting sensitivity analyses (Cinelli and Hazlett 2020; Rosenbaum and Silber 2009), which can also be applied to further improve balance tests. In this framework, the question to be asked is how sensitive results are to potential unobserved confounding. In other words, how fragile a result is against the possibility of unobserved confounding, or imbalanced not measure for some of the reasons discussed above, After all, even if pre-treatment covariates are balanced, unobserved variables may be both imbalanced and related to the outcome in ways that will bias analyses of interest. Thus, sensitivity analyses can complement a credible argument of ignorability or as-if randomness nature of the treatment assignment.

CHAPTER 4

Identification Strategy: evidence of within-lottery randomization

As discussed in **Chapter 1**, the thesis focuses on lotteries used to distribute houses for low-income citizens in Brazil. While the program (Faixa 1 of MCMV) is controlled at the national level, the lotteries are organized at the municipality level. Some municipalities implement their lotteries using the numbers drawn from the federal lottery for simplicity and transparency. Other municipalities promote the lottery by having public drawings. Even when the lottery is not public, municipalities are expected to have representatives from civil society and auditors from the accounting courts to verify the drawings.

The most important issue is that each municipality may run numerous lotteries over time. In principle, both the probability of winning and the characteristics of the participants could vary across these lotteries within a municipality. Thus, it is not possible to aggregate lotteries to the municipal level as if all were randomized there. Rather, each lottery event acts as a randomization with fixed probability. Each lottery is associated to one group of houses, which is identifiable by a unique "APF number."[1] For all of those reasons, the identifying assumption is not that winning the lottery is ignorable in a pooled dataset; rather, it is ignorable only APF. Specifically unbiased estimation of the pre-treatment covariates will be achievable under the assumption that:

[1] Note that the group of houses associated with each lottery or APF number may also differ from the larger 'housing projects' in a geographical cluster. For example, in the state of Acre, there is a large housing project from MCMV called CONJUNTO HABITACIONAL CIDADE DO POVO. Though this is one large housing complex, it contains 12 different APFs/lotteries.

$$\{X_{0i}, X_{1i}\} D_i \mid APF_i$$

where Y_{0i} and Y_{1i} are unit i's (potential) outcomes on some pre-treatment variable X under treatment or non-treatment; D_i indicates winning the lottery, and APF_i indicates the APF unique number for each lottery in which unit i was eligible for the lottery.

In summary, the identification strategy relies on the following assumptions, which are derived from the nature of the data construction:

- Each lottery event acts as a randomization with fixed probability. $Pr(apf|D = 1)$ is taken as fixed because the number of treated units per municipality is set by policy and not as part of the randomization process.
- Each lottery is associated to one group of houses, which is identifiable by a unique number.
- Identifying assumption is not that winning the lottery is ignorable in a pooled dataset; rather, it is ignorable only by APF (lottery number).
- $\{X_{0i}, X_{1i}\} \perp D_i \perp APF_i$.

4.1 ID strategy: Conditioning on APF

The identification strategy is necessary to analyze both pre-treatment covariates and eventual treatment effects. It is possible to analyze the presence of four potential types of people who are selected to be in the treatment or control groups: compliers, never takers, always takers and defiers. Due to the design of the program, the dataset contains only compliers: people selected by the lottery actually receive the house and those who lost the lottery did not receive the house. In other words, compliers are individuals who (always) comply with their assignment, that is, take the treatment if assigned to it and not take it if assigned to the control group. By design, always-takers and defiers seem to be impossible to happen in, which allows researchers to recover unbiased estimators.

People who lost the lottery cannot receive a unit from the government (always-takers and some defiers). It is possible that some people who were sampled in the lottery to receive a house choose not to accept it (never takers and defiers), but they would not be in the dataset collected by the government, since this house would have been distributed to someone else. Therefore, in order to identify any effect of receiving a house on a different set of variables, the identification strategy takes into the consideration the design of the lotteries in different municipalities.

That is the reason why any estimation of possible effects of receiving a house need to consider if there is any systematic difference between treatment and control group, conditioned on APF. This approach comports with the knowledge that lotteries are occurring in different municipalities, among different populations, and at different times. One way to falsify the randomization within lotteries is to find "imbalances" on pre-treatment covariates, conditionally on APF. That is, for pre-treatment covariates X_i I try to falsify the claim that:

$$\mathbf{X}_i D_i \mid \text{APF}_i$$

where APF_i indicated the APF unique number (lottery) of unit i. The pre-treatment covariates to be included in $\mathbf{X}_i$ are the information that participants of the lotteries were expected to provide to the government. In the registered Pre-Analysis Plan I had identified six variables that can confidently be classified as pre-treatment: age, sex, race, marriage status, (pre-treatment) monthly house rental expenditure, and self-reported family income at time of registration. For all the pre-treatment data, I used information provided by the Cadastro Unico before the treatment assignment. Unfortunately, most families did not report information about pre-treatment rental expenditures, which is dropped from the model. In **Chapter 5**, I propose to examine balance tests in the following ways:

Non-parametric, per-APF treatment modeling. One way to look beyond mean balance is to ask if any functions of the observed covariates, linear or otherwise, is able to predict who wins the lottery. Specifically, within each APF I will train a random forest algorithm that attempts to predict

who will win the lottery. I will then extract from each APF an out-of-sample estimate of model fit, including the AUC, the prediction R^2 (the correlation of the estimated probability of winning with the actual indicator for winning, squared), and corresponding p-value. The distribution of these estimates across APFs will be shown.

Mean balance: equivalence by lottery. In checking "mean balance", investigate whether the means of each X appear to be equal in the winning and losing groups, conditionally upon APF (lottery). The thesis employs equivalence tests (Hartman and Hidalgo 2018) within lottery and simple pre-treatment difference-in-means. Because these tests are run "per-APF", the results will be shown as distributions that summarize the results from across APFs.

Sensitivity as balance testing Finally, while inferential procedures ask whether there is evidence for (im)balance, this is unrelated to the question of whether imbalance is bad enough to worry about. Sensitivity analysis offers another lens, asking "would confounders as strongly related to winning the lottery and to an outcome of interest (or k times as strongly related) generate worrying degrees of bias in the resulting estimates?" The distinction between this and the inferential approach becomes increasingly important as the amount of data grows: in a very large dataset, even tiny imbalances can appear to be highly significant, but would in fact not generate bias, nor would confounders like them or many times stronger than them. I calculate and report the partial r square for every pre-treatment covariate: $R^2 = t^2/(t^2 + DF)$.

The general implementation of the identification strategy follows four steps:

Step 1: Loop over APFS: The process starts by looping over the 3012 lotteries for each I was able to collect data. There is difference in the sizes of these lotteries. Some were used to distribute just few houses, while others have thousands of participants and beneficiaries.

Step 2: Split the data: After that, 50% of the data for each APF is used to train a random forest model, which uses the pre-treatment covariates as the predictors for treatment assignment.

For each APF, I will use cross-validation and 5 folds.

Step 3: Test the data: In every iteration of the loop, extract out-of-sample estimates for the AUC.

Step 4: Different approaches: Calculate the prediction R^2 (the correlation of the estimated probability of winning with the actual indicator for winning, squared) and corresponding p-value, difference-in-means for all the pre-treatment covariates, t-statistics and the partial r-squares for each pre-treatment covariates .

4.2 Data construction

The dataset is constructed by matching three different sources of administrative information from the Brazilian government. First, a table with all the projects constructed in the program, including the project APF (the unique identifier of each project building), and the total number of units (houses or apartments) available in each project. The APF number is the identifier for the lotteries. This table was useful to compare the number of units in each project and the number of beneficiaries.

The number of beneficiaries – the treatment group – came from a different data source. This dataframe contains all the people selected to receive a house. The APF number and a personal identifier number were used to merge the two datasets. The third, and final dataset for this part of the project, is a list of participants in the selection process and their family members – such as spouses and children. This list contains not only people who were selected to receive a unit (treatment group), but a sample of participants who were not selected (control group). This dataset is called Sitah and aggregates all the participants of the lotteries and their family members. The key variable here is the one that defines if a person is a participant in the lottery (1), a spouse (2) or another family member (3).

One important observation is necessary. The federal government has provided the complete list of winners from each lottery, but among the non-winners, the available data contains only a

sample. The fraction of non-winners for whom data were sent can vary by municipality and by APF. This creates variation in the probability of winning across APFs, in the observed data, and that variation could in principle be related to characteristics of the participants. This problem is addressed by conditioning on APF since we expect the randomization to hold (only) within each APF. These data were obtained using Freedom of Access of Information (FOIA) act requests, in 2016 and 2017. In 2019, I undertook an extensive series of meetings and communiques with administrative staff of state-owned bank Caixa Economica Federal (CEF) in order to complete the dataset.

After thorough outreach to various bureaucrats in the national state-owned bank (CEF) two datasets were provided. One contains all the winners of the lotteries, while the other contains a sample of participants who did not get the house. After conducting the lotteries, municipality halls should send the information about all the winners to the national bank. In addition, municipalities were expected to send a random sample of people who entered the lottery and lost. The samples vary in length, but the expectation was that municipalities should share at least 25% the length of the winners list. The reason for this requirement is that the list of "losers" is intended as a backup list in case those on the winning list do not accept the home or are found to be ineligible upon further investigation.

The final step was to merge these data to the Brazilian Unified Registry data. This dataset was introduced after *Bolsa Familia* (a cash transfer program that constitutes the largest social program in the world), and is used across agencies to identify the poorest segment of the population and to deliver programs and services to them. Both datasets contain unique identifiers including social security numbers, which we use for merging purposes then strip from the data. This provides the pre-treatment variables used in the balance tests.

During the data construction, I had to deal with some challenges. First, around 36.000 people appear as recipients of the house in the treatment dataset, but in the Sitah dataset they were classified as spouses of family members. This means that something happened between the lottery and the contract signing that the family substituted the person who signed the contract. According to

bureaucrats who work at Caixa with the program, this substitution could happen, but the reasons are unclear.

Fortunately, I was able to collect the unique identifier for individual families, and this allowed to add the correct family member to the treatment group. Thus, we are only considering the correct people who were selected in the lotteries, and not the ones who actually signed the contract. Finally, there were around 18.000 people who appear in the winners dataset, but are not in the Sitah dataset. The only possibility was to drop these people because there is no information about their participation in the lotteries.

The final dataset contains 1,777,385 observations, distributed among 3,012 lotteries.

CHAPTER 5

Balance tests as a learning problem: assessing 3,000 MCMV lotteries

In this chapter, I present the results for the random forest models, and compare with different methods of assessing the balance between treatment and control covariates. In **Chapter 4**, I have discussed how the identifying assumption is not that winning the lottery is ignorable in a pooled dataset, but it is ignorable only in each APF. Any analysis involving the pre-treatment covariates must be achievable under the assumption that: $\{X_{0i}, X_{1i}\}D_i \mid APF_i$.

First, I present the distribution of the AUCs for all the lotteries (Hastie 2009; James 2021; Melo 2013; Murphy 2012). AUC stands for "Area under the ROC Curve", and ranges in value from 0 to 1. If a model provides 100% wrong predictions it will have an AUC of 0, while for those which predictions are 100% correct have an AUC of 1. AUC is both scale-invariant (measures how well predictions are ranked), and classification-threshold-invariant (it measures the quality of the model's predictions irrespective of what classification threshold is chosen).

5.1 AUC for all the lotteries

Considering the lotteries to distribute houses, a good way of interpreting AUC is the probability that the model ranks a random treatment more highly than a random control. In other words, the AUC is the probability that a randomly selected positive is ranked higher than a randomly selected negative. All in all, an AUC of 0.5 means that the model does not rank positives higher than negatives, nor the opposite. If all lotteries are truly random, the random forest model with the

pre-treatment variables should not be able to predict the treatment assignment and AUCs should be centered at 0.5.

Figure 5.1 indicates that results are close to the target, but not perfect. The average AUC is 0.507 and the median is 0.514, with a minimum of 0.04 and maximum of 1. This means that most lotteries are close to the truth, but there are some signs that the model is able to predict who is going to win the house, based on the pre-treatment covariates. In total, there are 1638 lotteries with AUC greater or equal to 0.5 (55%), 483 with AUC greater or equal 0.6 (16.3%), 76 with AUC greater or equal 0.7 (2.5%), and 2 with AUC greater or equal 0.9.

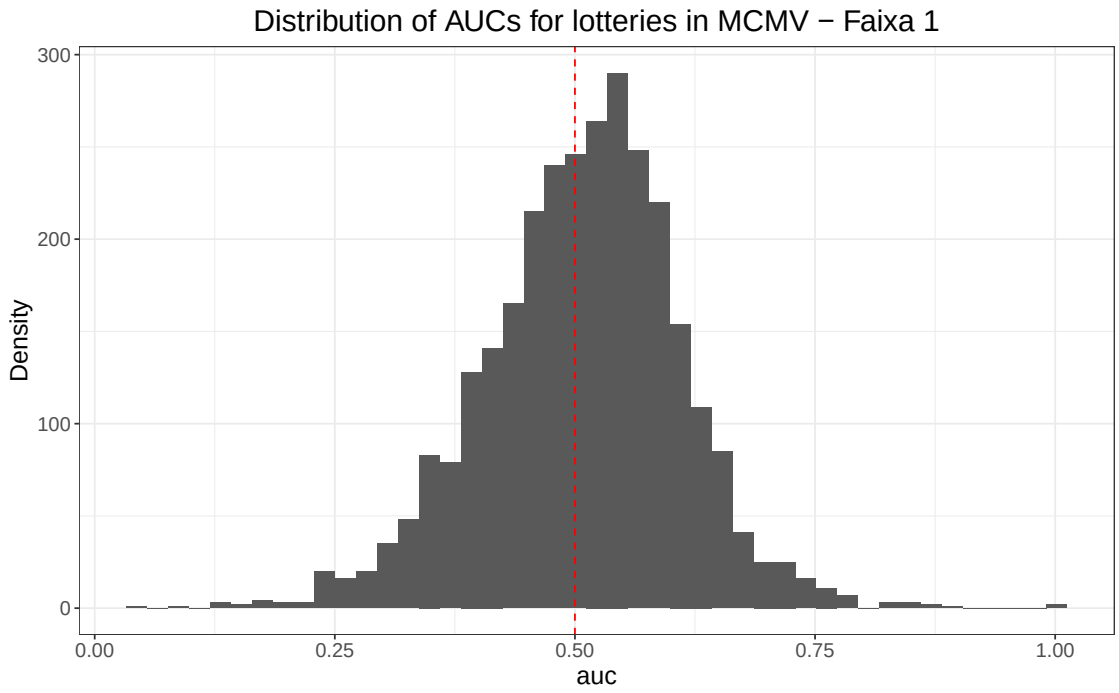

Figure 5.1: The distribution of AUCs for all the 3,000 lotteries for which it was possible to collect information for.

5.2 ROC curves for all the lotteries

AUC is the area underneath the entire ROC (Receiver Operating Characteristic) curve, which provides a visual way to observe how changes in the model's classification thresholds affect its performance. The ROC curve plots the model True-Positive Rate (TPR) and its False-Positive Rate (FPR) across all possible classification thresholds. For the lotteries case, while TPR is the probability that a house recipient is correctly predicted in the treatment class, FPR is the probability that a non-recipient is incorrectly predicted in the treatment class.

For each lottery, rather than making simple classification of treatment and control, the selected model gives probability scores for the treatment assignment. Using certain cutoff or threshold values, it is possible to dichotomize treatment and control predictions, and calculate the described metrics. **Figure 5.2** plot the ROC curves for all the lotteries. As expected, they are centered around the 45 degree line, and there are fewer lotteries where the models were able to provide good classifications.

Figure 5.3 exemplifies two real roc curves for different lotteries, where the treatment assignment seems to be random. On the one hand, a 45 degree line implies that TPR=FPR for every classification threshold. In other words, the classifier is just making random guesses, which in this context would be evidence of the randomness of the lottery.

On the other hand, if the lottery is not random, the line might be above or below the 45 degree line, as exemplified in **Figure 5.4**.A perfect classifier ROC goes along the outer-left and top of the chart, implying that the classifier will always have a TPR=1, regardless of the FPR.[1]. Finally, for some lotteries (i.e., 301084) the model is making more incorrect predictions than correct ones for the treatment assignment. An interactive GIF-version for all the lotteries' AUCs is available in my

[1]The ROC curve consists of a vertical line (x=0) and a horizontal line (y=1)

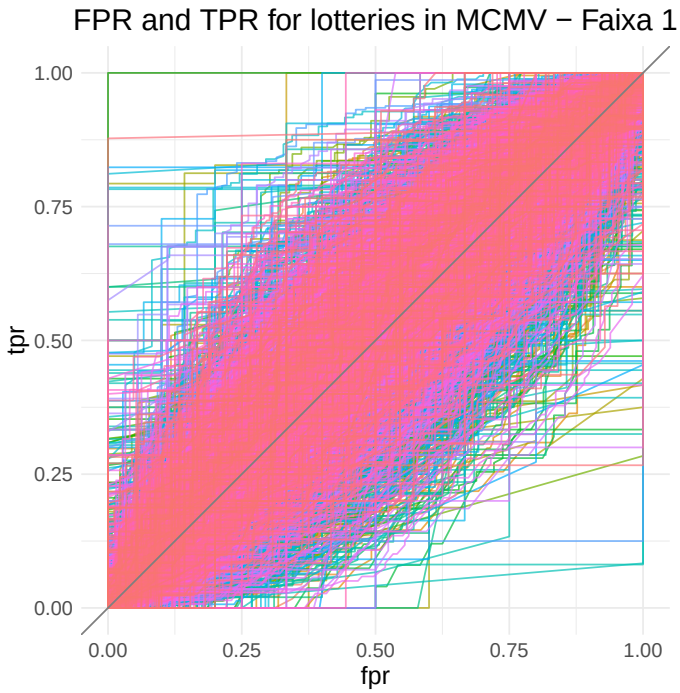

Figure 5.2: The distribution of ROC curves for all the 3,000 lotteries for which it was possible to collect information for.

website.[2].

[2] https://www.fernandobmello.com/methods-papers

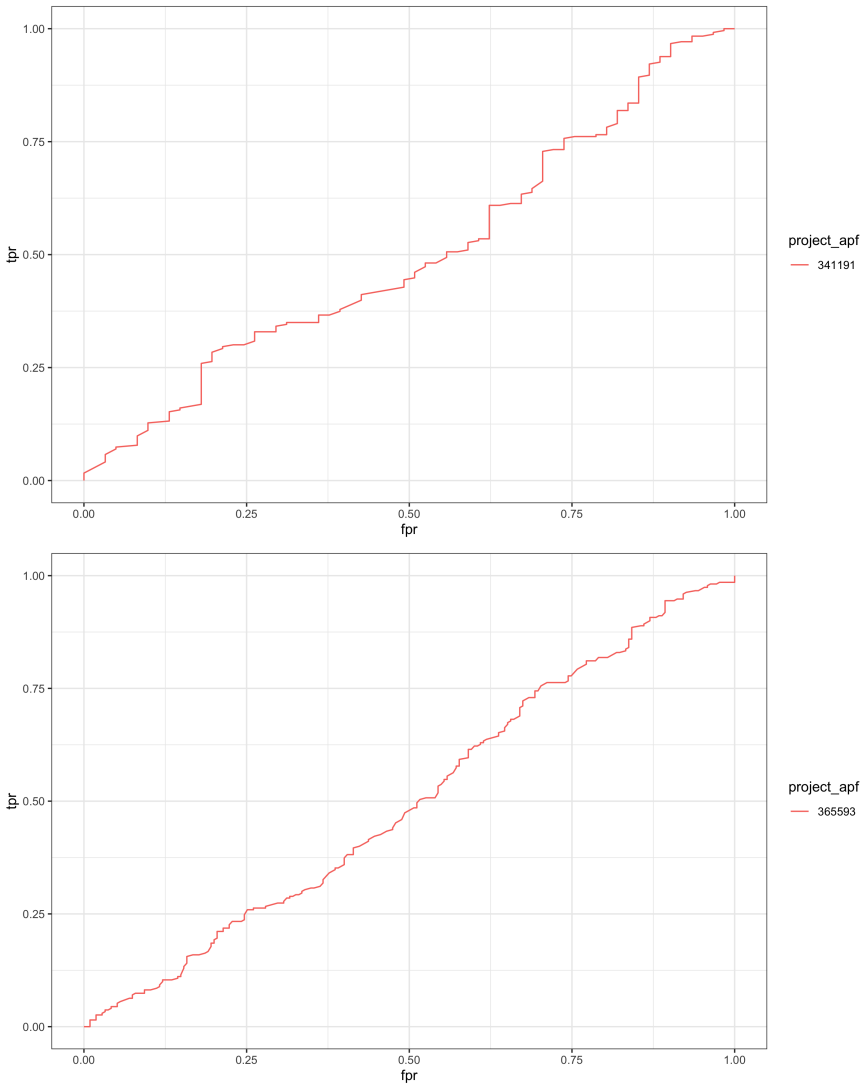

Figure 5.3: Two examples of lotteries with AUCs of 0.5.

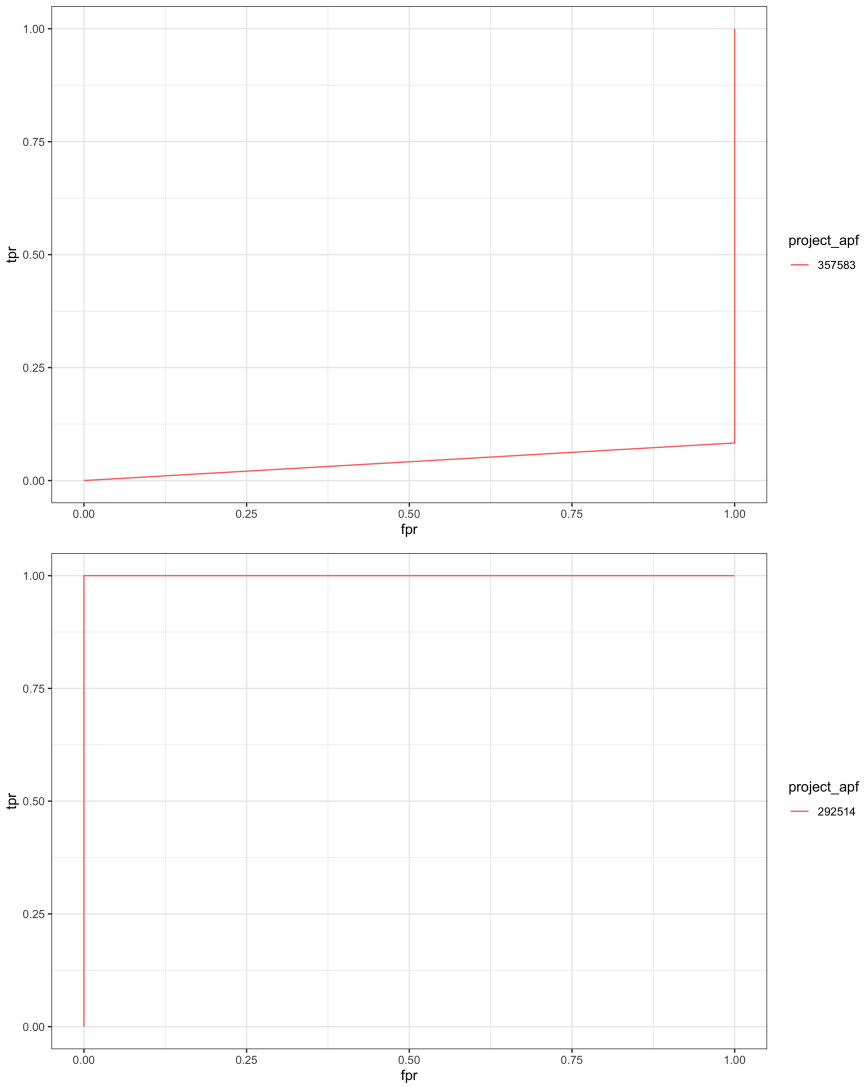

Figure 5.4: Lottery 357583 has a higher FPR than TPR, while lottery 292514 has a perfect prediction for treatment assignment.

5.3 Comparing results by size and percentage of control in the lotteries

One possible explanation for the difference in the trustworthiness of the lotteries is their sizes. The minimum number of participants (people selected to receive the houses + non-selected) is 17, the median 439, and the maximum 11398. It should be expected that imbalances happen more often in smaller lotteries, just because of the luck of the draw in a small lottery. However, the results show

that lotteries with a larger n of participants have a higher average AUC. There are 91 lotteries with 2,000 participants or more (only 3% of the total number of lotteries). As depicted in **Figure 5.5**, the AUC distribution for bigger lotteries is more to the right to the expected cut point of 50%. Thus, unexpectedly, in the lotteries with a larger numbers of total participants pre-treatment covariates show a higher predicted power on the treatment assignment.

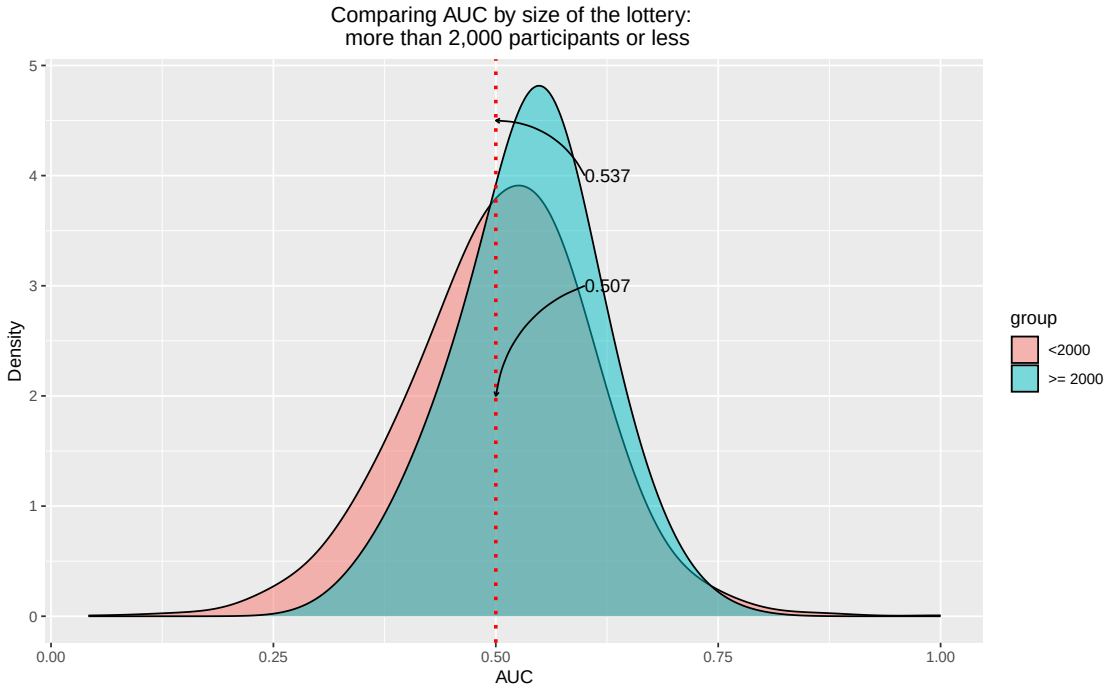

Figure 5.5: The distribution of AUC curves conditional on the size of the lottery.

A better explanation for the distribution of AUCs comes from the number of control units available for each lottery. As discussed in **Chapter 4**, the lotteries were conducted at the municipal level, which were expected to provide information to the national state-owned bank (CEF), who manages MCMV. According to the rules, municipality halls should send the information about all the winners to the national bank, and a random list of non-selected participants with at least 25% the length of the winners list. Yet, the analysis show that some lotteries did not provide the necessary information about control participants. The minimum percentage of control units is 2%,

the first quartile 24%, the median 34%, and maximum 99%.

Therefore, there is variation in terms of information about the number of controls informed by municipalities. The distribution of AUC curves conditional on the quantity of controls shows how lotteries with less than the expected number of non-selected participants of the lotteries are less trustworthy than lotteries with the expected number of control units. While the first group presents an average AUC of 0.54, the average AUC for the latter group is the expected 0.5.

A direct conclusion from these distributions is that lotteries that did not follow the expected rules regarding the number of non-recipients are also the ones with more predictive power for treatment assignment. It is impossible to conclude if there were any irregularities in these lotteries, but they should not be used for inferences.

Figure 5.6: The distribution of AUC curves conditional on the quantity of control.

5.4 Comparing T-values for all pre-treatment covariates and all lotteries

In this section, I compare the difference-in-means for all the selected pre-treatment covariates. Because there are over 3,000 lotteries, the traditional plot with the difference-in-means and 95% confidence intervals are not feasible. Instead, I present the distributions of the T values for each pre-treatment covariates. I also describe the percentage of cases that are outside the traditional boundaries of -2.05 and $+2.05$.

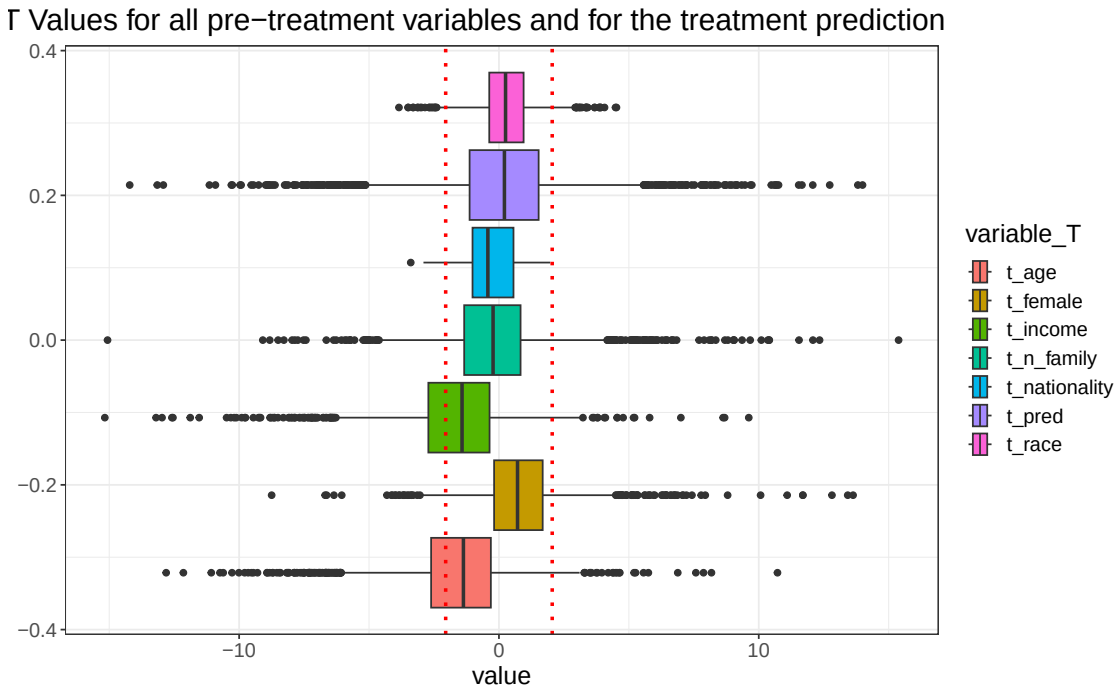

Figure 5.7: T values for all pre-treatment covariates in all the lotteries.

In general, people in the treatment group tends to be younger, poorer, and women. However, **Figure 5.7** shows how there is a lot of variation in the imbalance, not only by pre-treatment variables, but also by lottery. MCMV has an important gender component. For instance, the government tried to incentivize women to enter the lotteries. Women also keep the houses in case of divorce.

As described in **Chapter 4**, after constructing the dataset, I identified around 36,000 people who were classified as recipients of the house in the treatment dataset, but the lottery information classified them as spouses of family members. Over 95% of these recipients where women, whose husbands had entered the lottery. I was able to use the correct participants of the lotteries for these cases and, thus, these cases could not explain any difference in terms of treatment and control variables.

The simple difference-in-means also helps to understand possible imbalance by covariates. **Figure 5.8** indicates how age and income are the pre-treatment covariates with most differences across lotteries. In both cases, over 35% of the lotteries have a statistically significant difference between treatment and control groups, considering a 95% confidence in.

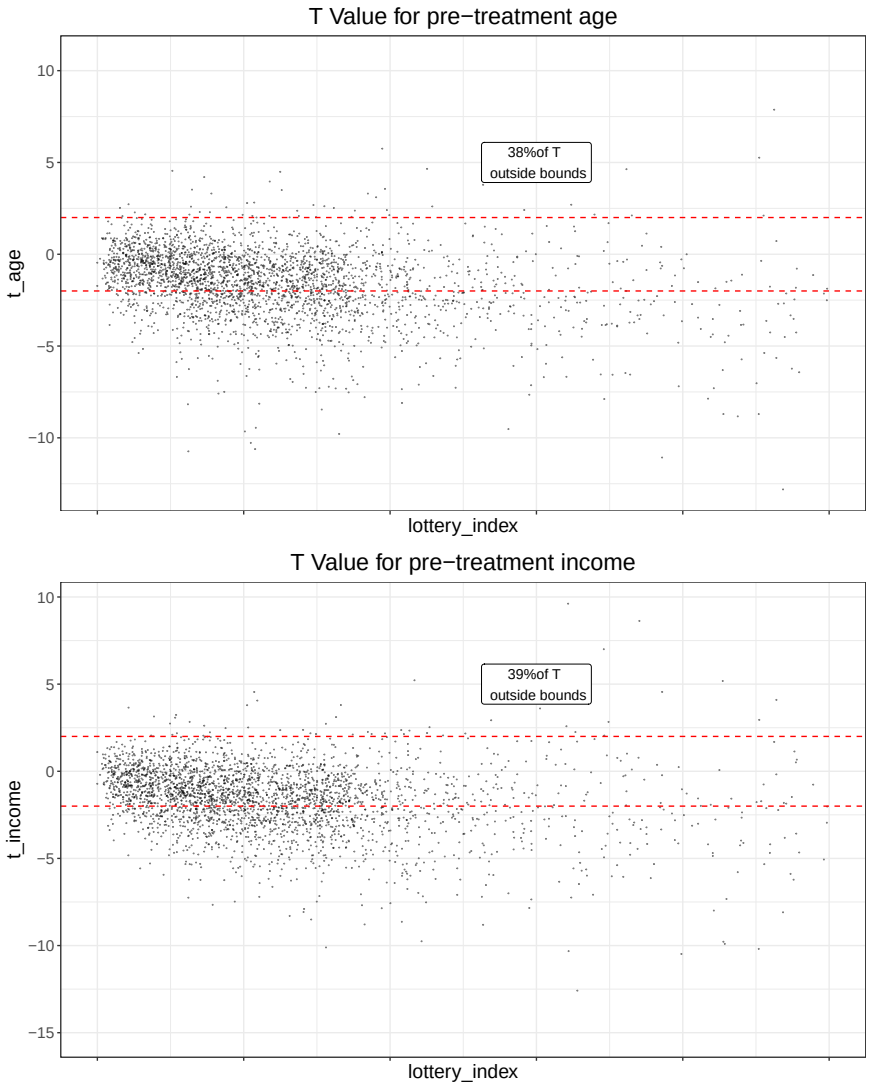

Figure 5.8: Age and income are the pre-treatment covariates with most differences across lotteries.

The variables number of members in the family and race show less differences across lotteries. Gender is coded as a dummy variable for females, since women are the majority of participants in the lotteries. In 21% of the lotteries, there are statistically significant differences between the proportion of women in the treatment and control groups. The number of members in each family was coded by grouping all the family members with the same family codes in each APF.

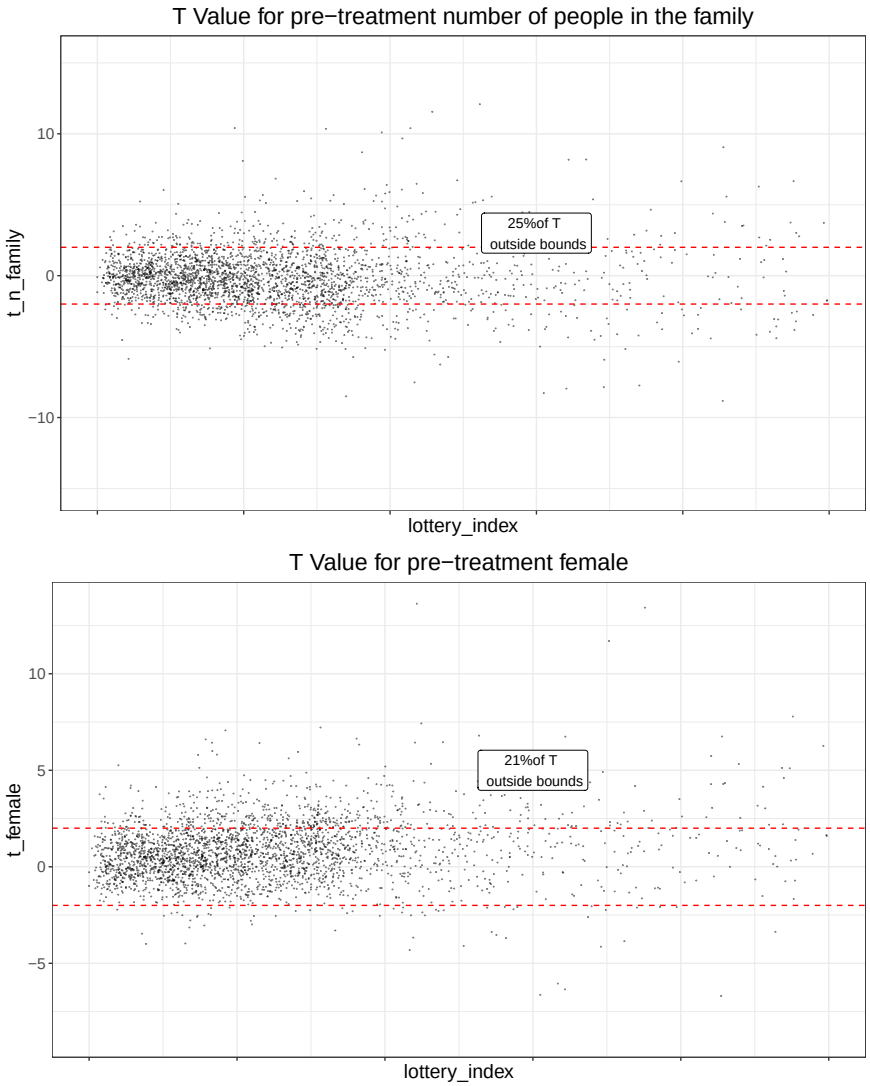

Figure 5.9: Gender and number of members in the family show less differences across lotteries.

Finally, race and nationality are only different in a small number of municipalities. In the case of nationality, there are a minority of participants who are not Brazilians, but more than 98% are born in the country. As a result, there is much less variation in this variable and the difference-in-means is almost always negligible. The variable for race was also coded as a binary variable, grouping black and brown people in the same category. In this case, only 7% of the lotteries have a statistically significant difference between the two groups.

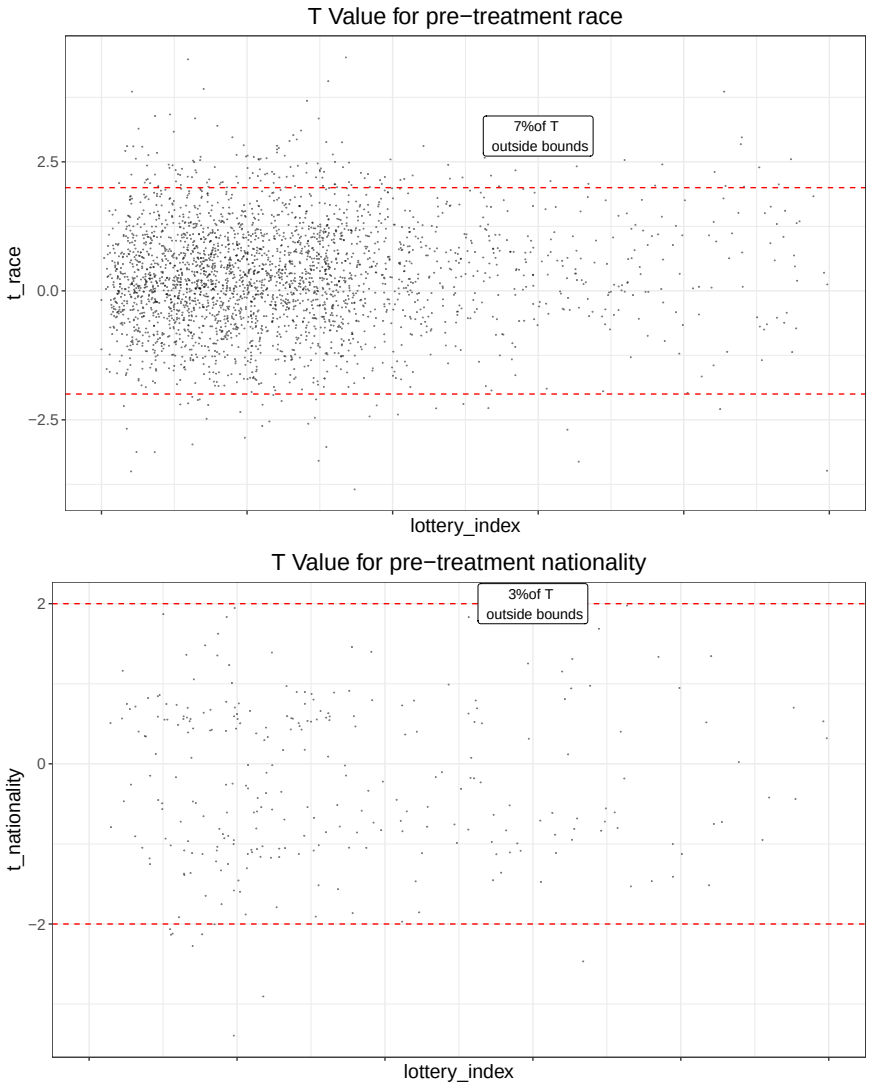

Figure 5.10: Race and nationality has less variation across lotteries.

5.5 Sensitivity as balance testing

Inferential procedures ask whether there is evidence for (im)balance between treatment and control groups, but we should also ask whether imbalance is bad enough to worry about. This is an application for sensitivity analysis. As described by (Cinelli and Hazlett 2020), the partial R^2 of the treatment with the outcome shows how strongly confounders explaining all the residual

outcome variation would have to be associated with the treatment to eliminate the estimated effect. Sensitivity analysis is also useful in the balance approach.

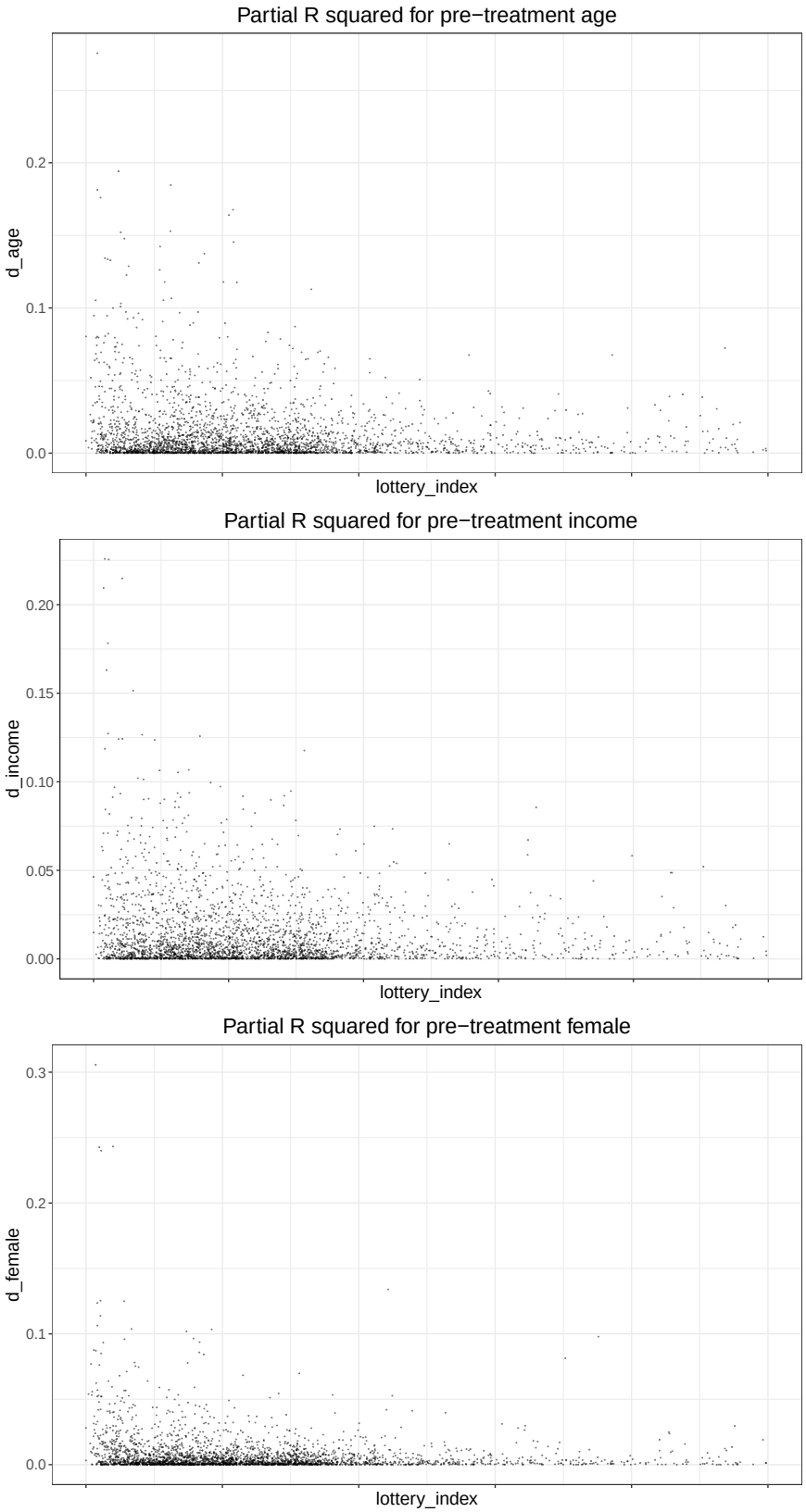

Figure 5.11: Are confounders strongly related to winning the lottery and to an outcome of interest?

The distinction between sensitivity and the inferential approach becomes increasingly important as the amount of data grows: in a very large dataset, even tiny imbalances can appear to be highly significant, but would in fact not generate bias, nor would confounders like them or many times stronger than them.

This section reports the partial R^2 for every pre-treatment covariate. In general, even when the pre-treatment covariates are not balanced, their impact are small and should not generate worrying degrees of bias in any resulting estimates. **Figure 5.11** indicates how, even when the variables have some statistically significant difference-in-means, they are responsible for small R^2 of the treatment. This provides further evidence to support any causal claim using the dataset because it is not going to generate large degree of bias.

5.6 Comparing methods

Does the use of machine learning methods provide any extra benefit for researchers than previous approach? In this section, I compare the results of simple difference-in-means, equivalence tests, and the capacity of random forest to predict treatment assignment. The main question is: does machine learning provide harder tests for balance assumptions?

To answer this question, I compare the results of the random forest model with the simple difference-in-means and the equivalence tests. For example, Hartman and Hidalgo (2018) apply the equivalence test on pre-treatment covariates in ten natural experiments. Nine out of the ten natural experiments reported difference-in-means t-test failing to reject the null hypothesis of no mean difference. When the equivalence test was used, in only for five of the ten studies it was possible to reject the null hypothesis of a mean difference.

It is to be expected that these different measures are related to each other. For instance, by operationalizing a dependent variable as the AUC, we can describe the relationship between AUC, difference-in-means and equivalence tests. **Table 5.1** indicates that an extra pre-treatment variable with signs of imbalance (statistically significant difference-in-means) represent a change of 3.7

points in the AUC, while one extra variable that fails to reject the null hypothesis of difference between treatment and control group (equivalence test) represents an increase of 1.3 points in the AUC.[3]

	Dependent variable:
	auc2
equivalence	1.313***
	(0.129)
'difference-in-means'	3.710***
	(0.137)
Constant	42.101***
	(0.439)
Observations	2,958
R^2	0.227
Adjusted R^2	0.227
Residual Std. Error	9.048 (df = 2955)
F Statistic	434.673*** (df = 2; 2955)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5.1: The relationship between AUC, difference-in-means and equivalence tests

This is consistent the description by (Hartman and Hidalgo 2018), according to which equivalence tests provide harder tests for pre-treatment balance evaluation. As shown in **Figure 5.12** equivalence tests provide more evidence for imbalances when compared to simple difference-in-means. The use of machine learning also provides harder tests for researchers. In 10% of the lotteries, there was no statistically significant difference between control and treatment groups in all pre-treatment covariates, but the model detected AUCs bigger than 0.5 – thus, providing some signaling of imbalance. In 5% of the lotteries, there are zero or one covariate which failed the

[3]The equivalence interval is based on the suggested pattern of 0.2 standard deviations and the use of the population mean difference.

equivalence tests, but have a AUC bigger than 0.5% – and in 1% of the lotteries the AUC is greater than 0.5 even if there is no covariate that fails the equivalence test.

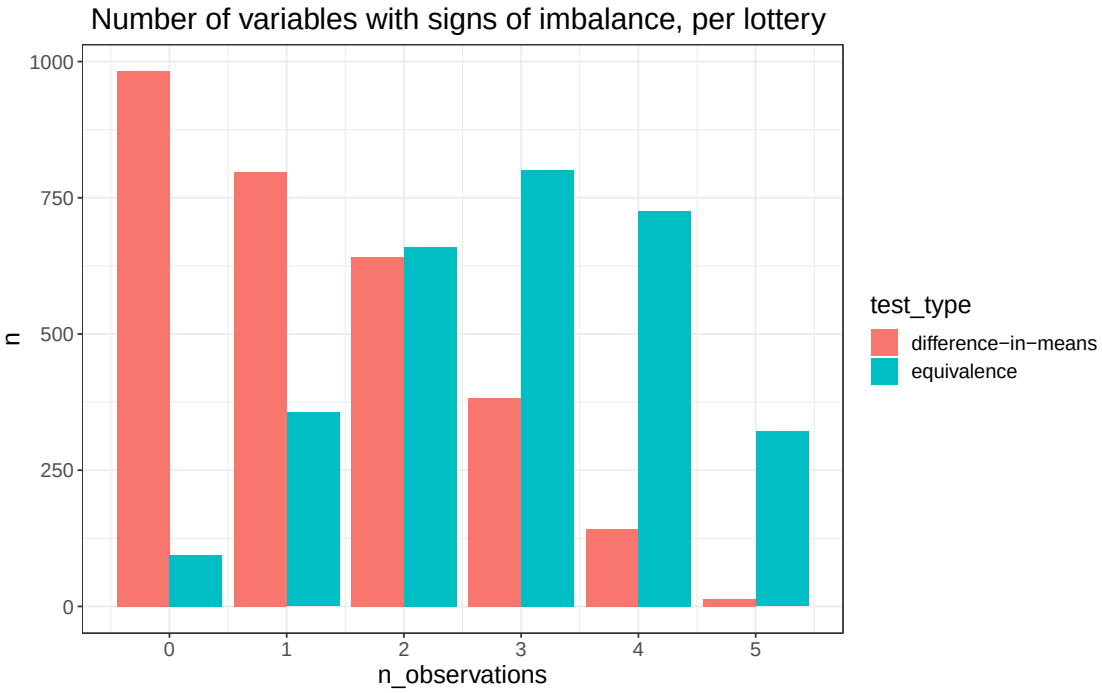

Figure 5.12: The number of pre-treatment variables with signals of imbalance.

But what these results mean in terms of evidence for the identifying assumption? Equivalence tests are suggested as the best toll to avoid conflating low power with similarity between treatment and control groups.However, equivalence test also provide assessment per variable, and do not generate an easily interpretable value. I propose the use of AUC for a straightforward reason. It not only provides one value that allows researchers to make a call about the treatment assignment. The use of AUC as the main result to assess balance in a natural experiment provides some advantages:

1. It is a single number with easy interpretation
2. It does not assume any functional form
3. AUC is scale-invariant

4. AUC is classification-threshold-invariant

5. It provides harder tests for balance

All in all, I consider that the use of machine learning models as balance tests increases the necessary evidence for the identifying causal assumptions for natural experiments.

CHAPTER 6

The risks for inference: an initial approach

The goal of this dissertation is to use random forest to provide further balance tests and evaluate the trustworthiness of over 3,000 lotteries used to distribute houses in Brazil. This is important because, when making causal claims from observational data, researchers are expected to provide evidence that results are not due to confounding. Using unbalanced lotteries to provide any inference about the effects of a housing program such as MCMV, could lead to erroneous conclusions.

In this chapter, I present only initial evidence that using lotteries with higher AUC would bias inferences using the MCMV data. These results come from an ongoing project with Chad Hazlett, for which the Pre-Analysis Plan is registered here https://osf.io/fs8uw/. The goal of this project is to investigate if winning a house causes any improvement in the lives of families and individuals. The project asks if house programs make life better for the recipients. In what ways?

Considering the purpose of this dissertation, **Figure 6.1** shows how all the results would be stronger by using the unbalanced lotteries for inference. In every possible outcome, the bias would not decrease the effect, but increase it, which would led the researchers to make stronger claims than allowed by the data. The direction of the biases is the same in the case of outcomes for the families, as described in **Figure 6.2**:

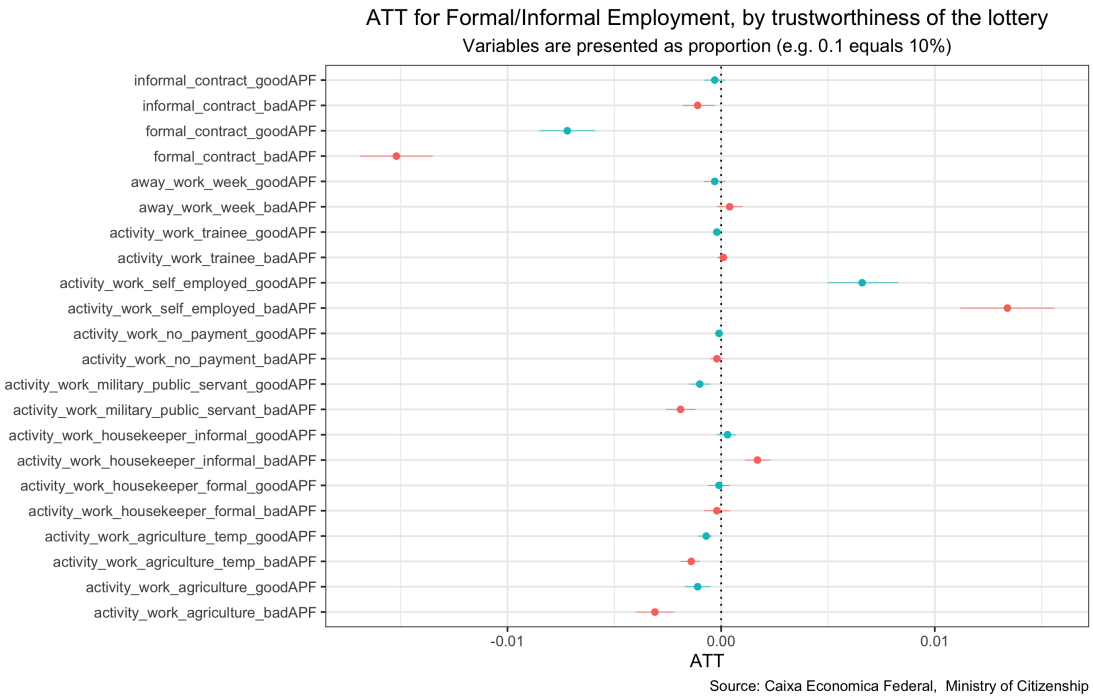

Figure 6.1: Use of unbalanced lotteries biases the results for individual outcomes.

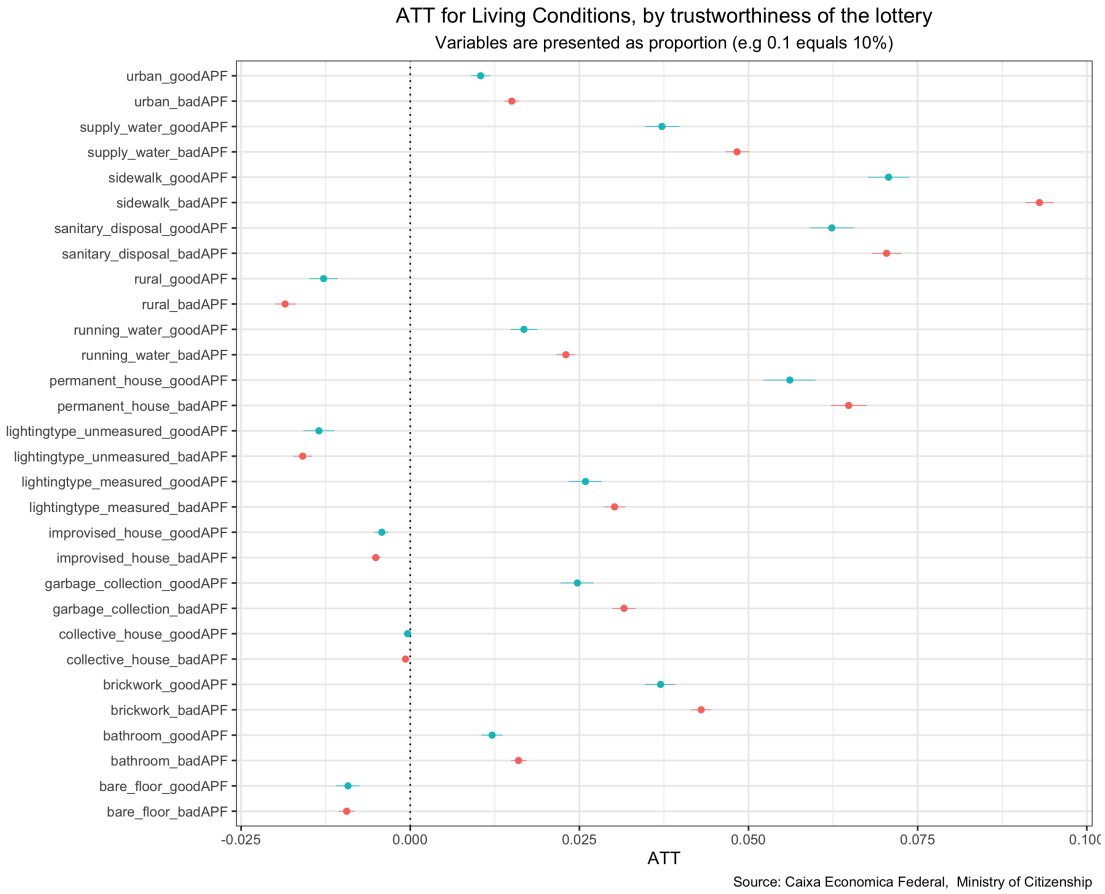

Figure 6.2: Use of unbalanced lotteries biases the results for family outcomes.

CHAPTER 7

Conclusion

This thesis discusses how balance tests are a key component for plausible causal identification. While observational researchers normally focus on evidence of balance for the covariates included in their model, experimental researchers provide randomization tests for balance on pretreatment covariates. By investigating a large-scale housing program in Brazil, I propose the use of machine learning as a balance test tool.

Natural experiments are increasingly used to identify real world situations with as-if random assignment (T. Dunning and Brady 2010; Thad Dunning 2012; Gerber and Green 2012; Gerber, Green, and Larimer 2008; Sekhon 2009). When analyzing an "as-if" situation, researchers should ask if pre-treatment covariates are able to predict the treatment assignment. After all, if the assignment is truly random, it should be impossible to predict it.

When the assignment of the treatment is not controlled by the researchers themselves, the responsibility to make a credible claim that the assignment of subjects to treatment and control conditions is as-if random falls upon the academic. As show in this thesis, by answering a simple question about the prediction of treatment assignment, researcher can make stronger claims about their assumptions.

REFERENCES

Biderman, Ciro, Frederico Ramos, and Martha Hiromoto (Mar. 2018). *THE BRAZILIAN HOUSING PROGRAM-MINHA CASA MINHA VIDA-EFFECT ON URBAN SPRAWL*.

Blalock, Hubert M. (1964). *Causal Inferences in Nonexperimental Research*. Chapel Hill Books. 10. Chapel Hill: Univ. of North Carolina Press.

Bueno, Natália S., Felipe Nunes, and Cesar Zucco (Oct. 2022). "Making the Bourgeoisie? Values, Voice, and State-Provided Home Ownership." In: *The Journal of Politics* 84.4, pp. 2064–2079. ISSN: 0022-3816. DOI: 10.1086/719275.

Campbell, Donald T. (Donald Thomas) (1970). *Experimental and Quasi-Experimental Designs for Research*. Chicago: R. McNally.

Caughey, Devin, Allan Dafoe, and Jason Seawright (Apr. 2017). "Nonparametric Combination (NPC): A Framework for Testing Elaborate Theories." In: *The Journal of Politics* 79.2, pp. 688–701. ISSN: 0022-3816. DOI: 10.1086/689287.

Cinelli, Carlos and Chad Hazlett (Feb. 2020). "Making Sense of Sensitivity: Extending Omitted Variable Bias." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.1, pp. 39–67. ISSN: 13697412. DOI: 10.1111/rssb.12348.

Cochran, William G. and Gertrude M. Cox (May 1992). *Experimental Designs*. Wiley. ISBN: 978-0-471-54567-5.

"On Types of Scientific Inquiry" (2009). "On Types of Scientific Inquiry: The Role of Qualitative Reasoning." In: *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Ed. by David Collier, David A. Freedman, Jasjeet S. Sekhon, and Philip B. Stark. Cambridge: Cambridge University Press, pp. 337–356. ISBN: 978-0-521-19500-3. DOI: 10.1017/CBO9780511815874.022.

Cook, Thomas D. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin. ISBN: 978-0-395-30790-8.

Cox, D. R. (1958). *Planning of Experiments | Wiley*.

Duflo, Esther, Rachel Glennerster, and Michael Kremer (Dec. 2006). *Using Randomization in Development Economics Research: A Toolkit*. Tech. rep. t0333. Cambridge, MA: National Bureau of Economic Research, t0333. DOI: 10.3386/t0333.

Dunning, T. and Henry E. Brady (2010). "Design-Based Inference: Beyond the Pitfalls of Regression Analysis?" In.

Dunning, Thad (2012). *Natural Experiments in the Social Sciences: A Design-Based Approach*. Strategies for Social Inquiry. Cambridge: Cambridge University Press. ISBN: 978-1-107-01766-5. DOI: 10.1017/CBO9781139084444.

Fisher, Ronald A. (1935). *The Design of Experiments*. 1st edition. New York: Macmillan Pub Co. ISBN: 978-0-02-844690-5.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky (Feb. 2005). "Water for Life: The Impact of the Privatization of Water Services on Child Mortality." In: *Journal of Political Economy* 113.1, pp. 83–120. ISSN: 0022-3808. DOI: 10.1086/426041.

Galiani, Sebastian and Ernesto Schargrodsky (Dec. 2004). "Effects of Land Titling on Child Health." In: *Economics and Human Biology* 2.3, pp. 353–372. ISSN: 1570-677X. DOI: 10.1016/j.ehb.2004.10.003.

Gerber, Alan S. and Donald P. Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. W.W. Norton.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer (Feb. 2008). "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." In: *American Political Science Review* 102.1, pp. 33–48. ISSN: 1537-5943, 0003-0554. DOI: 10.1017/S000305540808009X.

Gross, Justin H. (July 2015). "Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance." In: *American Journal of Political Science* 59.3, pp. 775–788. ISSN: 1540-5907.

Hansen, B. B. (May 2008). "The Essential Role of Balance Tests in Propensity-Matched Observational Studies: Comments on 'A Critical Appraisal of Propensity-Score Matching in the Medical Literature between 1996 and 2003' by Peter Austin, Statistics in Medicine." In: *Statistics in Medicine* 27.12, 2050–2054, discussion 2066–2069. ISSN: 0277-6715. DOI: 10.1002/sim.3208.

Hansen, Ben B. and Jake Bowers (May 2008). "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." In: *Statistical Science* 23.2, pp. 219–236. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/08-STS254.

Hartman, Erin and F. Daniel Hidalgo (Oct. 2018). "An Equivalence Approach to Balance and Placebo Tests." In: *American Journal of Political Science* 62.4, pp. 1000–1013. ISSN: 1540-5907.

Hastie, Trevor (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York: Springer. ISBN: 978-0-387-84857-0.

Imbens, Guido W. and Donald B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press. ISBN: 978-0-521-88588-1. DOI: 10.1017/CBO9781139025751.

James, Gareth (2021). *An Introduction to Statistical Learning: With Applications in R*. Second edition. Springer Texts in Statistics. New York, NY: Springer. ISBN: 978-1-07-161417-4.

Kempthorne, Oscar (May 1952). "The Design and Analysis of Experiments:" in: *Soil Science* 73.5, p. 415. ISSN: 0038-075X. DOI: 10.1097/00010694-195205000-00012.

Kenny, David A. (Jan. 1979). *Correlation and Causality*. 1st Edition. New York: John Wiley & Sons. ISBN: 978-0-471-02439-2.

Melo, Francisco (2013). "Area under the ROC Curve." In: *Encyclopedia of Systems Biology*. Ed. by Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota. New York, NY:

Springer, pp. 38–39. ISBN: 978-1-4419-9863-7. DOI: 10.1007/978-1-4419-9863-7_209.

Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. Cambridge, Mass: MIT Press. ISBN: 978-0-262-01802-9.

Pearl, Judea (2009). *Causality*. Second. Cambridge: Cambridge University Press. ISBN: 978-0-521-89560-6. DOI: 10.1017/CBO9780511803161.

Rainey, Carlisle (2014). "Arguing for a Negligible Effect." In: *American Journal of Political Science* 58.4, pp. 1083–1091. ISSN: 1540-5907. DOI: 10.1111/ajps.12102.

Rosenbaum, Paul R. and Jeffrey H. Silber (2009). "Sensitivity Analysis for Equivalence and Difference in an Observational Study of Neonatal Intensive Care Units." In: *Journal of the American Statistical Association* 104.486, pp. 501–511. ISSN: 0162-1459.

Rubin, Donald B. (Sept. 2008). "For Objective Causal Inference, Design Trumps Analysis." In: *The Annals of Applied Statistics* 2.3. ISSN: 1932-6157. DOI: 10.1214/08-AOAS187.

Sekhon, Jasjeet S. (June 2009). "Opiates for the Matches: Matching Methods for Causal Inference." In: *Annual Review of Political Science* 12.1, pp. 487–508. ISSN: 1094-2939, 1545-1577. DOI: 10.1146/annurev.polisci.11.060606.135444.

Woetzel, Jonathan, Sangeeth Ram, Jan Mischke, Nicklas Garemo, and Shirish Sankhe (2014). *Tackling the World's Affordable Housing Challenge*. Tech. rep.