UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

RAFAEL ZIMMERLE DA NÓBREGA

Causal Inference in Sampling From Finite Populations

Recife

2022

RAFAEL ZIMMERLE DA NÓBREGA

Causal Inference in Sampling From Finite Populations

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador: Cristiano Ferraz

Coorientador: Marcel de Toledo Vieira

Recife

2022

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

N754c Nóbrega, Rafael Zimmerle da
Causal inference in sampling from finite populations / Rafael Zimmerle da Nóbrega. – 2022.
64 f.: il., fig., tab.

Orientador: Cristiano Ferraz.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2022.
Inclui referências e apêndices.

1. Estatística aplicada. 2. Estudos observacionais. 3. Amostragem balanceada. 4. Calibração. I. Ferraz, Cristiano (orientador). II. Título.

310 CDD (23. ed.) UFPE- CCEN 2022 - 63

RAFAEL ZIMMERLE DA NÓBREGA

" CAUSAL INFERENCE IN SAMPLING FROM FINITE POPULATIONS "

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 18 de fevereiro de 2022.

BANCA EXAMINADORA

Prof. Dr. Cristiano Ferraz
DE/UFPE

Prof. Dr. Vinícius Quintas Souto Maior
DE/UFPE

Prof.Dr. Maurício Teixeira Leite Vasconcellos
ENCE/IBGE

To my family and friends.

To Maria Luiza, a great friend and mother, in memoriam.

ACKNOWLEDGEMENTS

The road that leads me here was built by many arms and many hearts. Especially in these strange pandemic times we've all been through, with distance and difficulties, I was always supported by the ones I love. I want to thank my parents, Ana and Ricardo, for all their kindness and encouragement, for all the opportunities they gave me: I know that they measure no efforts to see me happy; my siblings, Ricardo and Carol, for all the moments of warmth, which are, now, brightest parts of my memories; all my friends, brothers and sisters in life and in soul, for all the laughter and the uncountable moments of joy that we've shared; my grandparents, Vovó Nardinha, Vovô Ramon and Vovó Ana Maria, for teaching me so much about the world and about life; Maria Luiza, my second mother and great friend, for all the conversations we had, for all love and care... I miss you so much; my sister-in-law, Milla and my brother-in-law Jorge, for always believing me and for all the good moments; my brother-in-law, Luiz, for the partnership of everyday; my best friend, my soulmate, Laísa, who colors my days and supports me and makes me happy. I love you all, very much.

I would like to express special thanks to Wilson Fusco and Morvan Moreira, who were my first mentors in research, and who undoubtedly inspired me to see science from a much broader point of view.

I would like, also, to thank all my master's colleagues and professors, for teaching me a lot. Especially, I want to thank Professor Cristiano, who was there to help me, whenever I needed, and guided me to the realm of Statistics – in particular to Survey Sampling, what definitely changed my ways of seeing and practicing science – and Professor Marcel, for all the patient and supportive advises. Thank you for believing me.

I want to thank Professor Maurício Teixeira Leite Vasconcellos and Professor Vinícius Quintas Souto Maior for their beneficial comments and suggestions regarding this work.

At last, I want to thank CNPq, for the financial aid.

“Equânime, o Poema se ignora. Leopardo ponderando-se no salto, que é da presa, pluma de som, evasiva gazela dos sentidos?” (CAMPOS, 1992, pp. 39-40)

ABSTRACT

Causal inference deals with estimating the effects of specific interventions on a response variable. The estimation strategy involves comparing units exposed to intervention factor's levels, forming a treatment group, with those units not exposed, forming a control group. The control group serves as the base to estimate the counterfactual response of the treatment group. In observational studies, a major concern when building such groups is to ensure their comparability, controlling for characteristics others than the treatment itself, that may cause undesired interference on causal effects estimates, leading to systematic bias. Although the theory behind observational studies has advanced with methods to reduce such bias using conditional inference, in several of these studies data is obtained through complex probability sampling designs seldom taken into account in the estimation process. This thesis considers that, beyond representing a source of variability that must be incorporated in the analysis, sample design and estimation techniques can have a central role to estimate causal effects efficiently. Studies are carried out to investigate the use of balanced samples to ensure comparability between treatment and control groups with respect to the distributions of covariates, and the use of calibration estimates for the control group average response, improving estimates of the average counterfactual treatment response. The methods are compared with those already available in the literature, via Monte Carlo simulation.

Keywords: observational studies; sampling; balanced sampling; calibration.

RESUMO

A inferência causal lida com a estimação do efeito de intervenções específicas sobre uma variável de resposta. A estratégia de estimação envolve a comparação de unidades expostas a níveis de fatores de intervenção, com unidades não expostas, as quais formam um grupo de controle. O grupo de controle serve como base para estimar o contrafactual da resposta no grupo de tratamento. Em estudos observacionais, uma grande preocupação na construção desses grupos é garantir a comparabilidade entre eles, a partir do controle de outras características que não o próprio tratamento, as quais podem causar interferência indesejada sobre estimativas dos efeitos causais, provocando um viés sistemático. Embora a teoria por trás de estudos observacionais tenha avançado com métodos para reduzir esse viés, os dados utilizados em diversos desses estudos são obtidos por meio de amostragem probabilística complexa raramente levados em consideração no processo de estimação. A presente dissertação considera que, além de representar uma fonte de variabilidade que deve ser incorporada na estimação de efeitos causais, planos e técnicas de estimação de amostragem podem ter um papel central para estimar efeitos causais de forma eficiente. São realizados estudos para investigar o uso de amostras balanceadas que garantam a comparabilidade entre grupos de tratamento e controle, no que diz respeito às distribuições das covariáveis, e de estimadores para a média da variável de resposta no grupo de controle baseados em calibração, a fim de melhorar as estimativas da resposta média contrafactual do grupo de tratamento. Comparam-se esses métodos com aqueles já disponíveis na literatura, por meio de simulações de Monte Carlo.

Palavras-chaves: estudos observacionais; amostragem; amostragem balanceada; calibração.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

Causal inference has a vast and well-established literature that roots back to writings of Neyman and Fisher. Early results were mainly in the direction of setting sound basis to types of statistical investigation labeled experimental studies, and of making distinctions of those studies from the ones which were named observational. Both of those types of study share a common aim: the estimation of effects of particular interventions on a response variable of interest, measured on a given set of units. However, they represent hugely different situations, especially regarding the control of the investigator over the data generating process and the assumptions needed to be made in order to draw statistical inference.

Frequently, the intervention considered is dichotomous, dividing the set of units into two partitions: one characterized by the truly activeness; the other by the absence of the intervention. Very often in the literature those partitions are called treated and comparison – or control – groups. The fundamental aim is to estimate *if* and *to what level* the intervention affects the response considered. In order to do that, the researcher must contrast the response measured in the treatment group with the response measured in a suitable group of comparison units. It is of greatest importance to consider how similar the groups are, to the extent of all factors – others than the intervention itself – that, although not affected by the intervention, may influence response values, because differences with respect to those factors may lead to systematic biases in causal effects estimates. We call those factors covariates. When the treatment and control group are similar in terms of covariates, they are said to be balanced.

The fundamental difference of experimental situations, when compared to the observational ones, is that, when in the former, the investigator (in that case also called experimenter) completely controls the treatment assignment mechanism over the units in question – what in general is done randomly. In fact, the physical process of randomization allows inferences to be made with respect to the design implemented in the study and in addition guarantees important properties regarding the distributions of covariates over the set of experimental units: in expectation, the groups compared tend to be balanced on these factors, what leads to derivations of causal effect estimators of simple forms and good properties.

On the other hand, in observational situations, the researcher does not control the intervention assignment mechanism. Moreover, it is possible, and, indeed, very frequently occurs, the distributions of covariates to be different in the treated and control groups.

To illustrate, consider the example addressed in Cochran (1968), using data of death rates comparisons of men classified in three of the following categories: non-smokers, smokers of cigarettes only and smokers of cigars, pipes or both. The results presented by Cochran refer to three different studies, performed in Canada, Britain and United States, whose data were supplied to the U.S. Surgeon General's Committee.

Table 1 reproduces the results brought in his paper Cochran (1968, p. 297).

Table 1 – Death rates per 1,000 person-years by smoking classification (Cochran, 1968).

	Study		
Smoking Group	Canadian	British	U. S.
Non-smokers	20.2	11.3	13.5
Cigarettes only	20.5	14.1	13.5
Cigars and/or pipe	35.5	20.7	17.4

Source: Cochran (1968)

At a first glance, the results may lead to the conclusion that death rates in the cigar-pipe group is, by a large amount, greater than in the other groups. The reliance on that results may lead one to claim smokers of cigars and pipes to belong to a much more vulnerable group, in the sense that the probability of death associated with them is greater than to the other groups considered and argue that this is caused by their smoking habit.

However, if one finds out that the ages of the individuals, for instance, in the cigar-pipe group have a different distribution from the other groups, then a part of the effect once attributed to the smoking habit may be due to the difference in age. That reasoning comes from the perception that age is a factor which relates to the probability of dying and that is not affected by the treatment.

Table 2 – Mean ages in years.

	Study		
Smoking Group	Canadian	British	U. S.
Non-smokers	54.9	49.1	57.0
Cigarettes only	50.5	49.8	53.2
Cigars and/or pipe	65.9	55.7	59.7

Source: Cochran (1968)

As discussed in Cochran's paper, the mean age of individuals in the cigar-pipe group is higher than in the other groups. This is shown in Table 2. The bias arising from the utilization of a simple mean comparison, as may be drawn from Table 1, is an example of what Rosenbaum

(2002, p. 71) calls an overt bias. In fact, causal inference in observational studies may be subject to systematic *overt* or *hidden* biases.

For the first case, the literature has advanced in the direction of proposing methods to reduce such biases and to propose estimators for treatment effects. Among the methods proposed, matching methods have become an increasingly popular one, with applications in fields such as Economics, Political Sciences and Medicine. The idea is that, given some assumptions about the treatment assignment mechanism, by comparing treated and control units with similar covariates values – or some *special* functions of the covariates –, the mean difference of the response values between treated and control groups is an unbiased estimator of the average treatment effect in the population.

In this work, we suppose that the population of interest is a finite population, that is: the population variables are not treated as random variables, but as fixed quantities. Also, the treatment assignment is viewed as fixed: in the finite population, each unit belongs either to the treated or the control group. The aim is to estimate the effect of having been assigned to the treated group, instead of being assigned to the control group, on a response variable of interest. In order to do that, probabilistic samples are drawn from the finite population. Consider Cochran's example (1968) again. Take, for instance, the Canadian study: in a loosely manner, the finite population may be viewed as all men in Canada in a given period of time. The number of smoking men is a fixed quantity, the death rates of smoking men is a fixed quantity, and so on.

Many times, investigators in applied fields use survey data obtained from surveys with probabilistic complex designs. The nature of the sampling selection are, though, rarely taken into account. Besides, there are few papers discussing good practices when making causal inference with survey sampling data, especially in what concerns to the way in which the sampling process is to be accounted for.

This thesis' arguments are that, beyond being a source of variation that has to be taken into account when using already available methods in observational studies, sampling can be used as a standard tool for estimating causal effects. To this end, sampling theory results are used to justify the employment of widespread methods from design and analysis of sampling from finite populations to observational studies. In addition, the randomization-based inference perspective are adopted, as it is not usual in the literature, where frequently one resorts on model-based or bayesian perspectives. Our position allows us to make inferences without making assumptions as the ones needed in the last two cited inference perspectives.

The aim of this thesis is to propose and discuss ways to cope with the difficulties of causal inference in observational studies, with the aid of methods developed in Sampling theory, making use of randomization-based inference, as it is usual in sampling from finite populations. Specifically, we intend to: discuss the main approaches already available in the literature of observational studies, especially *matching methods* and how they relate to sampling from finite populations; assess the performance of *rejective sampling* technique as a tool for obtaining samples where treatment and control groups are balanced in terms of covariates; propose a causal effect estimator based on the theory of *calibration* in survey sampling and assess the performance of the calibration estimator when compared to matching estimators.

This thesis is organized as follows: in Chapter 2 we present a theoretical framework of causal inference in observational studies. In Section 2.1, we define causal effects in terms of the model of potential outcomes, commonly known as the Neyman-Rubin Causal Model. In Section 2.2, we present the approach of Cochran and Rubin (1973) and discuss the appealing of matching methods. In Sections 2.3 and 2.4, we present the methods of Propensity Score Matching and Genetic Matching, respectively. In Chapter 3, we discuss the role of sampling in observational studies. In Section 3.1 we discuss notions of sampling theory. In Section 3.2 we discuss design-based considerations about the approach in Cochran and Rubin (1973). In Section 3.4, we discuss the role of auxiliary variables in sampling theory, particularly focusing on the methods of calibration and balanced sampling and how they can be useful for causal inference in observational studies. Specifically, we introduce a treatment effect estimator based on calibration technique. In Chapter 4 we present the design of the Monte Carlo simulations conducted to compare the performances of the matching estimators and the calibration estimator and discuss the results. In Chapter 5, we conclude.

2 CAUSAL EFFECTS IN OBSERVATIONAL STUDIES

The model of potential outcomes, widely known as Neyman-Rubin Causal Model or just Rubin Causal Model, is the basis on which a large extent of the literature in observational studies were developed, either theoretically or practically.

Neyman (1923) was the first to employ the concept of potential outcomes in the context of causal inference. In this seminal work, Neyman makes use of an urn model to draw random samples from a finite population with units corresponding to potential yields of particular varieties on plots of an agricultural field. The idea of random assignment mechanism of interventions is present, albeit implicitly, behind his usage of an urn model.

In fact, while Neyman uses random assignment as a mean to make statistical inferences possible, the idea that the physical process of randomization is a *sine qua non* to experimental results validation is attributed to Ronald Fisher (REID, 1998).

A more complete formulation of the model of potential outcomes is often attributed to Donald Rubin – see Holland (1986), Sekhon (2009), Imbens and Rubin (2015) –, in a series of important papers (Rubin (1974), Rubin (1977), Rubin (1978), Rubin (1980)). Nevertheless, most of the ideas used in Rubin's formulation were already currently presented in the formulation of experimental studies. For instance, the very concept of potential outcomes has essentially the same meaning as the concept of "true" or "conceptual" response, as discussed, for instance, in Kempthorne (1955), Wilk and Kempthorne (1955) and Hinkelmann and Kempthorne (2007, p. 157), in the tradition of Kempthorne (1952).

Rubin's formulation of the potential outcomes framework is an attempt to extend terms and concepts from experimental to observational studies, and thenceforth represents an effort to a synthesis comprising different situations of causal inference. In order to do this, several assumptions are needed to be made, so as to interpret the observational situation "as if" it were generated by a randomized mechanism, although unknown to the investigator. The Neyman-Rubin Framework (NRF) thus became the cornerstone on which methods for estimating causal effects in observational studies were developed.

In this Chapter, we present an overview of the NRF, discussing the role of the treatment assignment mechanism and the definition of causal effects (Section 2.1). Next, we present an approach developed by Cochran and Rubin (1973), in which the relationship between the response variables and the covariates is assumed to follow a regression structure, what makes it clear to visualize the appealing of matching methods (Section 2.2).

2.1 OVERVIEW, THE TREATMENT ASSIGNMENT MECHANISM AND THE DEFINITION OF CAUSAL EFFECT

A general presentation of Rubin's framework is made in Holland and Rubin (1980), Holland and Rubin (1983), Rosenbaum (1984) and Holland (1986), among others. A population U, consisting of N units, indexed by $k, k \in U$ is considered. The units are all subject to an "experimental manipulation" (HOLLAND; RUBIN, 1980, p. 3)– a mechanism through which *treatment conditions*, or *levels*, are assigned to each unit in the population. Throughout this text, as it is usual in the literature, the terms *treatment levels* or simply *treatments* are used interchangeably. Moreover, without loss of generality, only two treatment levels are considered.

The treatment levels define two groups in the population: if unit k is exposed to the intervention, it is said that it belongs to the *treatment* group; otherwise, it belongs to the *control* group. We may refer to the units in the treatment group as *treated units* or simply *treated* and to the units in control group as *control units* or simply *controls*. The population groups defined by the treatment levels will be denoted by U_T, with $T \in \{0, 1\}$, so that if unit k is treated, then $k \in U_1$; if it is not, then $k \in U_0$.

We define a random vector $\mathbf{T} = (T_1, \ldots, T_k, \ldots, T_N)$ which assumes a realized value $\mathbf{t} = (t_1, \ldots, t_k, \ldots, t_N)$ such that if unit $k \in U_T$, than $t_k = T$. We call the set of all possible vectors $\mathbf{t}$ the treatment space and denote it by $\mathcal{T} = \{0, 1\}^N$. Thus, the treatment assignment mechanism is viewed as a probability distribution in $\mathcal{T}$, denoted by $\lambda(\cdot)$ such that $\lambda(\mathbf{t}) = \Pr(\mathbf{T} = \mathbf{t})$.

It is noteworthy that behind the idea of experimental manipulation is the premise that each unit in the population has a non-zero probability of receiving the treatment. When this premise holds, the treatment assignment mechanism is said to be *probabilistic* (IMBENS; RUBIN, 2015). Thus, the concept of treatment in the NRF is that of an intervention that can potentially be applied to any unit in the population considered, so that, in this sense, exposition to radiation is an example of treatment, while an unit's attribute, such as gender or race, is not (HOLLAND; RUBIN, 1983; ROSENBAUM, 1984; HOLLAND, 2008). Moreover, it stands out that, by referring to an experimental manipulation, the model does not restricts itself to experimental contexts, but encompasses observational studies, comprised as situations in which a random treatment assignment mechanism – although out of investigator's control – is assumed to occur. This is considered an important contribution of Rubin's framework, and has its roots on the oft-quoted motto "no causation without manipulation" (HOLLAND, 1986, p. 959).

In addition, a set of pre-treatment characteristics is associated with each unit in the population. These characteristics must not be affected by the treatment, but are possibly related to the response variable. We call them covariates. Without loss of generality, we consider throughout this text that there is only one single covariate, which we denote X_k for the kth unit in the population. The N-vector containing all values X_k will be denoted by $\mathbf{X}$.

The response variable of interest is defined as Y. Each unit has a different version of Y corresponding to each treatment level. These versions are properly what is comprised as potential outcomes. In this sense, to each unit $k \in U$ there is a corresponding vector of potential outcomes given by (HOLLAND; RUBIN, 1980):

$$\mathbf{Y}_k = \begin{pmatrix} Y_{0k} & Y_{1k} \end{pmatrix}^{\mathrm{T}}$$

where Y_{tk} is the response realized for unit k when $t_k = t$, $t \in \{1, 0\}$. We denote by $\mathbf{Y}$ the $2 \times N$ population matrix with all $\mathbf{Y_k}$. Likewise, we denote by $\mathbf{Y_t} = (Y_{t1}, \ldots, Y_{tk}, \ldots, Y_{tN})$ the matrix with all response values in population for a given treatment level T, such that $\mathbf{Y} = \begin{pmatrix} \mathbf{Y_0} & \mathbf{Y_1} \end{pmatrix}^{\mathrm{T}}$.

In general, the function $\lambda(\mathbf{t})$ can possibly depend on the values of the covariates and the potential outcomes. For instance, as mentioned in Rubin (1977), children may be assigned to a treatment – for example, a compensatory reading program – on the basis of a reading test score – see also Imbens and Rubin (2015), for a further discussion with examples of treatment assignments depending on covariates and potential outcomes. For this reason, $\lambda(\mathbf{t})$ is frequently denoted by $\lambda(\mathbf{t}|\mathbf{X}, \mathbf{Y})$.

The unit-level causal effect, in the case of two treatment levels, is defined as the difference between the potential outcomes

$$\tau_k = Y_{1k} - Y_{0k} \tag{2.1}$$

Once that, at a time, only one treatment level is attributed to each unit, only one potential outcome is observed. Therefore, the unit-level causal effect 2.1 can never be observed in practice.Holland and Rubin (1980) and Holland (1986) call this the Fundamental Problem of Causal Inference.

Commonly, the interest is to estimate the population average treatment effect (PATE), given in Definition

Definition 2.1 (PATE). *The population average treatment effect is given by*

$$\bar{\tau} = \mathrm{E}\Big(Y_{1k} - Y_{0k}\Big) \tag{2.2}$$

where $\mathrm{E}(\cdot)$ denotes the expectation over the distribution of $Y_{1k} - Y_{0k}$ in the population.

Note that the PATE 2.2 involves two unobserved quantities, because

$$\bar{\tau} = \mathrm{E}\Big(Y_{1k} - Y_{0k}\Big) = \mathrm{E}\Big(Y_{1k} - Y_{0k}|U_1\Big) + \mathrm{E}\Big(Y_{1k} - Y_{0k}|U_0\Big) =$$

$$= \mathrm{E}\Big(Y_{1k}|U_1\Big) - \mathrm{E}\Big(Y_{0k}|U_1\Big) + \mathrm{E}\Big(Y_{1k}|U_0\Big) - \mathrm{E}\Big(Y_{0k}|U_0\Big) \tag{2.3}$$

In Equation 2.3, $\mathrm{E}\Big(Y_{0k}|U_1\Big)$ and $\mathrm{E}\Big(Y_{1k}|U_0\Big)$ are unobserved since we only observe responses conditional on at most one treated level.

Sometimes, the interest of a study is to estimate the effect of the intervention only on the treated units. For this reason, another common estimand is the population average treatment effect on the treated (PATT), given by Definition 2.2 below.

Definition 2.2 (PATT). *The population average treatment effect on the treated is given by*

$$\bar{\tau}_1 = \mathrm{E}\Big(Y_{1k} - Y_{0k}\Big|U_1\Big) \tag{2.4}$$

By a similar argument as the one developed in Equation 2.3, it can be easily seen that there is only one unobserved quantity involved in Equation 2.4.

In experimental situations, the investigator can make use of the known treatment assignment mechanism, induced by the experimental plan, as a source of randomness under which inferential results can be provided. In addition, under that perspective of inference, estimators with simple forms, such as the difference in means between treated and control responses, are unbiased for the PATE in Equation 2.2. Importantly, under the experimental plan, systematic sources of error tend to be made random, and consequently cancel out on average, a result that echoes back to Fisher (1992) (COCHRAN; RUBIN, 1973). For a discussion with examples, see Hinkelmann and Kempthorne (2007) and Rubin (1974). For this reason, differences in covariates distributions between the treated and control groups are of minor concern. Moreover, the experimental plan gives the inferential basis to the estimation of confidence intervals and hypothesis testing (HINKELMANN; KEMPTHORNE, 2007; COX; REID, 2000).

In observational studies, however, the investigator's ignorance about the treatment assignment mechanism does not allow him to make use of that inference perspective. In addition, there is a major concern about systematic biases caused by differences in covariates distributions. That point will be stated more clearly in Section 2.2. In order to cope with these difficulties, several methods were developed, unavoidably relying in assumptions either about the treatment assignment mechanism or the relationship between the response variables, the treatment effects and the covariates.

2.2 THE APPROACH OF COCHRAN AND RUBIN (1973)

In order to assess biases incurred in the estimation of causal effects in observational studies, a common approach is to assume a relationship between the response variable, the treatment effect and the covariates. This approach has been commonly used in texts since before Rubin Causal Model was established and generally with no mention to a treatment assignment mechanism (see, for instance, Cochran (1968), Cochran and Rubin (1973), Rubin (1973a), Rubin (1973b)).

In general, the relationship considered is supposed to follow a regression structure. In the simplest case, the regression equations have only one single covariate, on which the equation is linear. Moreover, the equations in each treatment group are parallel to each other. The treatment effect is supposed to be constant and additive for each unit.

Here, we follow the developments presented in Cochran and Rubin (1973). Consider that, for the kth unit in the treatment level U_t of the population, the response variable can be formulated as

$$Y_{tk} = \mu_t + \beta(X_{tk} - \eta_t) + \varepsilon_{tk} \tag{2.5}$$

where μ_t and η_t are the population means of Y and X in the treatment level U_t, respectively, and $\mathrm{E}(\varepsilon_{tk}|X_{tk}) = 0$ and $\mathrm{E}(\varepsilon_{tk}^2|X_{tk}) = \sigma_t^2$.

Using Equations 2.2 and 2.4, it is easy to see that, in this case,

$$\bar{\tau} = \bar{\tau}_1 = \tau = \mu_1 - \mu_0 - \beta(\eta_1 - \eta_0) \tag{2.6}$$

what is achieved by considering the difference in the expectations of Y between the two treatment levels conditional on X values. Note that, under the assumptions of the formulation in Equation 2.5, treatment effects are constant for all $k \in U$.

Cochran and Rubin (1973) argue that, if one draws random samples from the population, then – using lower case letters to denote sample values –,

$$\mathrm{E}_r(\bar{y}_1 - \bar{y}_0) = \mu_1 - \mu_0 = \tau + \beta(\eta_1 - \eta_0) \tag{2.7}$$

where $\mathrm{E}_r(\cdot)$ denotes expectation over random samples, $\bar{y}_t$ denotes the sample mean of y_{tk} and η_t is the expectation of $\bar{x}_t$, so that the bias arising when using the simple difference in means between y values equals $\beta(\eta_1 - \eta_0)$.

As a measure of initial bias, Cochran and Rubin (1973) propose the standardized difference-in-means given by

$$D = \frac{\eta_1 - \eta_0}{\sqrt{\frac{\sigma_1^2 + \sigma_0^2}{2}}} \tag{2.8}$$

reporting that values $D \geq 1$ are considered large.

One way by which the bias in Equation 2.7 can be reduced is by matching control and treated units. Essentially, in matching, one subsets the sample of control and treated units in such a way that the units in the resulting matched set have similar values in the covariates – or some special functions of covariates (see Section 2.3). The efficacy of matching techniques for reducing bias has been studied for a long time, examples including – besides Cochran and Rubin (1973) – Cochran (1953), Greenberg (1953), Billewicz (1965), Rubin (1973a) and Rubin (1973b).

Equations 2.6 and 2.7 reveal the appealing of matching methods. Denote $\mathscr{S}$ a sample from U. Consider $\mathscr{M} \subset \mathscr{S}$ the set of matched units. Suppose further that only the control sample is subset in matching. It is easily seen that if for every $k \in \mathscr{S}$ we have $x_{1k} = x_{0j}$ for some $j \in \mathscr{M}$, then the expected value of the difference between response means in treated and control groups over matched samples equals

$$\begin{aligned}\mathrm{E}_m(\bar{y}_1 - \bar{y}_0) &= \mathrm{E}_m\{\mu_1 + \beta(\bar{x}_1 - \eta_1)\} - \mathrm{E}_m\{\mu_0 + \beta(\bar{x}_0 - \eta_0)\} \\ &= \tau + \beta\{\eta_1 - \mathrm{E}_m(\bar{x}_0)\} = \tau \end{aligned} \tag{2.9}$$

where $\mathrm{E}_m(\cdot)$ denotes expectation over matched samples.

In fact, matching can be implemented with and without replacement: that is, one may or may not allow a single control unit to be matched with different treated units. Matching with replacement has advantages particularly in what concerns to bias reduction, but also disadvantages, especially with regard to the variance – and its estimation – of the matching estimator: on the one hand, the fact that generally few control units will be used in matching, it is expected the variance of the estimator to increase; on the other hand, matching with replacement induces correlations between matched pairs that share the same control unit, what makes variance estimation not as straightforward as in the without-replacement case (see Imbens and Rubin (2015, Chapter 18), for instance). As our primary aim in this thesis is not to discuss variance estimation, we will explore the differences between matching with and without replacement no further, but we include simulation results for both the cases (see Section 4).

2.3 PROPENSITY SCORE MATCHING

Initial investigations on the use of matching to remove bias in observational studies were, in general, in the direction of assessing matching on covariates values. However, for cases with multiple and continuous covariates, it becomes virtually impossible to find exact matches (ROSENBAUM, 2002; IMBENS; RUBIN, 2015). In order to overcome this, Rosenbaum and Rubin (1983) proposed the use of special functions of the covariates, known as balancing scores.

A balancing score is a function of the observed covariates $b(X_k)$ such that

$$T_k \perp\!\!\!\perp X_k | b(X_k), \forall k \in U \tag{2.10}$$

where $\perp\!\!\!\perp$ holds for independence (DAWID, 1979) and T_k is the indicator of treatment assignment (see Section 2.1). It is easy to see that the simplest balancing score is given by $b(X_k) = X_k$.

Rosenbaum and Rubin (1983) showed that, under regularity assumptions about the treatment assignment (see Appendix B), an important balancing score is the propensity score. Briefly, the propensity score is the unit-level probability of being treated, conditional on the values of covariates.

Definition 2.3 (Propensity Score)**.** *Given a treatmant assignment mechanism $\lambda(\mathbf{t}|\mathbf{X}_k)$, we call*

$$\theta_k(\mathbf{X}_k) = \sum_{\mathbf{t} \in \mathcal{T} | (t_k = 1)} \lambda(\mathbf{t}|\mathbf{X}_k) \tag{2.11}$$

the propensity score.

As an important contribution, Rosenbaum and Rubin (1983) also showed that, under regularity assumptions, if one conditions the Y values on a balancing score, then one can get an unbiased estimator of the population treatment effects. This result is presented in Result C.3, in Appendix C.

The appealing of using the propensity score is that it is the "*coarsest*" balancing score, in the sense that it is a function every other balancing score. This result is stated in Result C.1, in Appendix C. In this sense, using the propensity score allows one to overcome the problem of dimensionality and makes it easier to get exact matches.

After defining the balancing score on which matching will be made, the next element needed is a distance metric to compare alternative potential units to be matched. In general, the method of Nearest Available Matching (COCHRAN; RUBIN, 1973; ROSENBAUM; RUBIN, 1985b; IMBENS; RUBIN, 2015) is implemented. By this method, each treated unit is matched with the nearest (in terms of a pre-defined distance function) control. A common approach is to use either Euclidean Distance or Mahalanobis Distance (ROSENBAUM; RUBIN, 1985b; DIAMOND; SEKHON, 2013; SEKHON, 2009). For the case of multiple covariates, given two vectors of balancing scores $\mathbf{b}_k$ and $\mathbf{b}_j$, for any $k, j \in U$, the Mahalanobis distance is defined as

$$MD(\mathbf{b}_k, \mathbf{b}_j) = \sqrt{(\mathbf{b}_k - \mathbf{b}_j)^{\mathrm{T}} V^{-1} (\mathbf{b}_k - \mathbf{b}_j)} \tag{2.12}$$

where V is a covariance matrix.

For the case of a single covariate, or when one uses a single balancing score, a simplified version of the Mahalanobis Distance is obtained as

$$MD(b(X_k), b(X_j)) = \sqrt{\frac{(b(X_k) - b(X_j))^2}{\mathrm{V}(\mathbf{b})}} \tag{2.13}$$

In particular, $b(X_k)$ can equal the covariate value for kth unit, X_k. In general, the vector $\mathbf{b}_k$ can also include functions of X_k and other balancing scores, such as the propensity score. Rosenbaum and Rubin (1985b) define V, for the multivariate case, as the covariance matrix of the balancing scores in the control group. More generally, we can consider V to take into account both treated and control groups as, in our case of a single covariate,

$$V = (\mathbf{b} - \bar{\mathbf{b}})^{\mathrm{T}} (\mathbf{b} - \bar{\mathbf{b}}) \tag{2.14}$$

where

$$\mathbf{b} = \Big(b(X_1), \ldots, b(X_k), \ldots, b(X_N)\Big)^{\mathrm{T}}$$

and $\bar{\mathbf{b}}$ denotes the mean of $\mathbf{b}$.

In particular, there are many forms by which propensity scores can be combined with Mahalanobis distance. For instance, Rosenbaum and Rubin (1985b) suggest using Mahalanobis Distance with the vectors $\mathbf{b}_k$ and $\mathbf{b}_j$ in Equation 2.12 including both the units' covariates and propensity scores, and using Mahalanobis Distance within groups defined by similar values of the propensity score. The rationale behind this suggestion is that Mahalanobis Distance matching are generally more successful in producing balance in covariates than in the propensity score.

If one's option is to use the propensity score as a matching criterion, it needs to be estimated, since its true form is almost always unknown. The usual proceeding is to apply binary variable regressions models in the class of generalized linear models – such as *logit* and *probit* models (see McCullagh and Nelder (2019) and Agresti (2003)) – on the realized treatment assignment vector $\mathbf{t}$ as a function of the observed covariates $\mathbf{X}$. Furthermore, as Sekhon (2009) points out, if the propensity score is estimated by logit or probit models, it is generally better to match units on the basis of the linear predictors, what avoids matching based on values between 0 and 1. This whole process is, however, subject to misspecifications of the propensity score model.

The property that the propensity score is a balancing score (see Result C.1) is frequently used, in practice, as a guideline so as to accept a given specification of the estimated propensity score. A step-by-step procedure is described, for instance, in Rosenbaum and Rubin (1985b): a propensity score model is estimated with an initial specification – say, with covariates main effects as explanatory variables, only – and matches are found using the estimated model. Next, balance in covariates is assessed for the matched sample, using some criterion: if balance is judged acceptable, the matched sample is maintained and one can proceed to analysis; otherwise, one reformulates the propensity score model, possibly including functions of the covariates, such as quadratic and interactions terms. This process is repeated until balance is achieved.

Whenever applying this procedure, the investigator faces the question of how balance should be assessed, a question that has no definitive answer in the literature. There are many commonly suggested balance measures, including the *standardized difference-in-means*

(COCHRAN; RUBIN, 1973; IMBENS; RUBIN, 2015), *two sample t-statistics* (ROSENBAUM; RUBIN, 1985b; DIAMOND; SEKHON, 2013), *F-ratios* Rosenbaum and Rubin (1985b) and, more recently, *Kolmogorov-Smirnov statistics* (DIAMOND; SEKHON, 2013). Besides, Imbens and Rubin (2015) suggest computing *log standard deviations ratios* as a measure of dispersion comparison between treated and control covariates distributions and the frequency of one group's units whose covariate values lie in the tail – defined by arbitrarily choosing an α-quantile – of the other group's covariates empirical distribution function.

In addition to the large number of balancing assessment measures, there doesn't seem to exist a definite guidance on what values of these measures should be accepted as sufficiently good in practice. Therefore, it is not astounding that many studies fail to report any measure of balancing quality as discussed in Diamond and Sekhon (2013) and Austin (2008).

Moreover, the process of "manually" checking balance may not be optimum (DIAMOND; SEKHON, 2013; SEKHON, 2009). One's judgement about the quality of balance is a quiet relative measure. Thus, it is not infrequent that proceeding like this do not produce sufficient covariate balance.

2.4 GENETIC MATCHING

As a mean to overcome the problems faced in Propensity Score Matching, Diamond and Sekhon (2013) proposed a different matching method, based on a generalization of the Mahalanobis Distance. They consider a weighting version of Equation 2.12 given by

$$GD(\mathbf{b}_k, \mathbf{b}_j) = \sqrt{(\mathbf{b}_k - \mathbf{b}_j)^{\mathrm{T}}(S^{-\frac{1}{2}})^{\mathrm{T}} W S^{-\frac{1}{2}}(\mathbf{b}_k - \mathbf{b}_j)} \tag{2.15}$$

where W is a positive definite square matrix with dimension equal to the length of $\mathbf{b}_k$ and such that every element lying outside the main diagonal is set to equal zero. The W matrix is, therefore, a weight matrix, that weights every coordinate of $\mathbf{b}_k - \mathbf{b}_j$. In turn, $S^{-\frac{1}{2}}$ is such that $V = S^{-\frac{1}{2}}(S^{-\frac{1}{2}})^{\mathrm{T}}$ – that is, it corresponds to the Cholesky decomposition of V.

It is easily seen that Mahalanobis Distance 2.12 is a particular case of distance 2.15, with the former expression being obtained if the matrix W, in the later expression, is the identity matrix.

Distance 2.15 is combined with an iterative, genetic optimization algorithm, developed by Sekhon and Mebane (1998) (see also Mebane and Sekhon (2011)). The combination of distance 2.15 with genetic optimization has been called the Genetic Matching method, or simply GenMatch (DIAMOND; SEKHON, 2013). As stated by Sekhon (2009), GenMatch attempts to find matches of treated and control units so as to maximize covariate balance, by minimizing some loss function. To this end, at each step of the algorithm, the weight matrix W is automatically adjusted: for example, if $\mathbf{B_k}$ includes both the propensity score and the vector of covariates and if the optimal balance is achieved using only the covariates, then the element of W corresponding to the propensity score is set to 0, while the elements in W corresponding to the covariates is set to a positive number, found by the algorithm so as to maximize balance.

In short, GenMatch optimizes the balance on covariates distributions between treated and control matches given pre-defined measures of balance. As default measures of balance, the algorithm uses *Kolmogorov-Smirnoff statistics* and *Paired t-statistics*, minimizing the largest individual discrepancy, as measured by the *p*-value of the tests "for all variables that are being matched on" (DIAMOND; SEKHON, 2013, p. 934). As emphatically highlighted by Sekhon (2008) and Diamond and Sekhon (2013), the tests are not used to conduct formal hypothesis inferences, but only as measures of discrepancy between matched treated-control groups.

A description of the algorithm is given by Diamond and Sekhon (2013):

1. Set an initial value for the W matrix in Equation 2.15;

2. Create a generation composed of g different W matrices, where g is a number that may be specified by the user;

3. For each W matrix in generation, match units based on the distance 2.15;

4. For each of the g matched samples, compute the value of the loss function;

5. Get the W matrix from the minimum loss sample and check optimization criterion. If satisfied, go to next step; otherwise, go back to 2;

6. Use distance 2.15 with optimal W matrix to find the final matched sample.

The algorithm will converge to the optimal matched sample asymptotically in the size of g (DIAMOND; SEKHON, 2013; MEBANE; SEKHON, 2011; SEKHON, 2008). Besides the genetic optimization algorithm, which seeks to guarantee maximum balance, GenMatch works exactly as a method of Nearest Available Matching, with the generalized distance function 2.15.

3 THE ROLE OF SAMPLING IN OBSERVATIONAL STUDIES

Insofar, we have not considered the role of sampling in observational studies. This is, however, an important task: first because many observational studies, in particular in social and biomedical sciences, utilize survey data as sources of information; secondly, because that will be the basis for the main propositions in this thesis. In the present chapter, we give an outline of sampling theory and discuss how observational studies relate to it.

3.1 NOTIONS OF SAMPLING THEORY

In this section, we discuss notions of sampling. In particular, we consider that $U = (\mathbf{Y}, \mathbf{X})$ is a finite population. By finite population, we mean that the variables $\mathbf{Y}$ and $\mathbf{X}$ are viewed as fixed, rather than random, quantities. Again, U has N units, $\mathbf{X}$ denotes an N-vector of covariates and $\mathbf{Y} = \begin{pmatrix} \mathbf{Y_0} & \mathbf{Y_1} \end{pmatrix}^{\mathrm{T}}$.

We do not mention the treatment assignment mechanism in this section, supposing, thus, that the finite population is given under a fixed treatment assignment. We want to reinforce that, although each individual in the population has a true, conceptual, Y value both under treatment and control levels, once we suppose the treatment assignment as given, we can only observe at most one of these values. In other words, although we include both potential outcomes for each population unit as components of the finite population, we do not mean that they are simultaneously observed.

We denote by $\Omega = \{0, 1\}^N$ the sample space. Note that a sample is represented by a vector of sample inclusion indicator variables, in an analogous way as the treatment assignment vector in Chapter 2. In a similar way, we define $\mathbf{S} = (S_1, \ldots, S_k, \ldots, S_N)^{\mathrm{T}}$, a random sample vector that assumes a particular value $\mathrm{s} = (s_1, \ldots, s_k, \ldots, s_N)^{\mathrm{T}} \in \Omega$, so that, if the kth unit in the population is selected in the sample, $s_k = 1$; otherwise $s_k = 0$. Let $p(\cdot)$ be a probability distribution on Ω that ascribes the probability $\Pr(\mathbf{S} = \mathrm{s}) = p(\mathrm{s})$. We call $p(\mathrm{s})$ a *sampling design*. For a value s, we define the set of sampled units $\mathscr{S} \subseteq U$ such that $k \in \mathscr{S}$ if $s_k = 1$. Likewise, we define $\mathscr{S}_1 \subseteq U_1$ and $\mathscr{S}_0 \subseteq U_0$ the samples of treated and control units, respectively.

Every population unit k has a sampling selection probability, given by the sampling design $p(\mathrm{s})$, known by the investigator, which we denote by π_k. These probabilities are called first-order inclusion probabilities. Note that π_k is the probability of $S_k = 1$. The first-order inclusion probability is the sum of the probabilities of all sample assignment vectors $\mathbf{s} = (s_1, \ldots, s_k, \ldots, s_N)^{\mathrm{T}} \in \Omega$ such that $s_k = 1$.

Definition 3.1 (First-Order Inclusion Probability). *For all $k \in U$, we call*

$$\pi_k = \sum_{\mathbf{s} \in \Omega | (s_k = 1)} p(\mathbf{s}) \tag{3.1}$$

the first-order inclusion probability for unit k.

Our aim is to estimate population average treatment effects. Here, we focus on the population average treatment effect on the treated (PATT). If we recall Equation 2.4, it is easy to see that in a finite population, the PATT can be written as a difference in means. For ease of notation, denote

$$\bar{Y}_{11} = N_1^{-1} \sum_{k \in U_1} Y_{1k} \tag{3.2}$$

$$\bar{Y}_{10} = N_1^{-1} \sum_{k \in U_1} Y_{0k} \tag{3.3}$$

$$\bar{Y}_{01} = N_0^{-1} \sum_{k \in U_0} Y_{0k} \tag{3.4}$$

$$\bar{Y}_{00} = N_0^{-1} \sum_{k \in U_0} Y_{0k} \tag{3.5}$$

Then, we can define the finite population average treatment effect on the treated (FPATT), as in Definition 3.2 below.

Definition 3.2 (Finite PATT). *In a finite population U, the population average treatment effect on the treated is given by*

$$\bar{\tau}_1 = \bar{Y}_{11} - \bar{Y}_{10} \tag{3.6}$$

where $\bar{Y}_{11}$ and $\bar{Y}_{10}$ is given by Equations 3.2 and 3.3, respectively.

A design-unbiased estimator for $\bar{Y}_{11}$ is given by

$$\hat{\bar{Y}}_{11_\pi} = N_1^{-1} \sum_{k \in \mathscr{S}_1} d_k Y_{1k} \tag{3.7}$$

where

$$d_k = \pi_k^{-1}$$

is called the basic design weight.

The estimator in Equation 3.7 is commonly called the Horvitz-Thompson (HT) estimator of the population mean as a reference to Horvitz and Thompson (1952). Next, we show the design-unbiasedness of the HT estimator, using the concept of design expectation. The design expectation is the expected value of a statistic taken over all possible realizations of the vector $\mathbf{S} = (S_1, \ldots, S_k, \ldots, S_N)^{\mathrm{T}}$ – that is, over all samples possibly drawn by the sampling design.

Result 3.1 (Design-Unbiasedness of HT Estimator)**.** *Suppose we want to estimate* $\bar{Y}_{11}$ *. Then*

$$\hat{\bar{Y}}_{1_\pi} = N_1^{-1} \sum_{k \in \mathscr{S}_1} d_k Y_{1k} \tag{3.8}$$

is a design-unbiased estimator.

Proof. We can rewrite

$$\hat{\bar{Y}}_{1_\pi} = N_1^{-1} \sum_{k \in \mathscr{S}_1} d_k Y_{1k} = N_1^{-1} \sum_{k \in U_1} d_k Y_{1k} S_k \tag{3.9}$$

Since population values and weights are fixed, the only random variable in Equation 3.9 is S_k. Now, taking design expectations, we have

$$\mathrm{E}_p(\hat{\bar{Y}}_{1_\pi}) = N_1^{-1} \sum_{k \in U_1} d_k Y_{1k} \times \mathrm{E}_p(S_k) = N_1^{-1} \sum_{k \in U_1} Y_{1k} \tag{3.10}$$

because S_k is a Bernoulli variable with probability of success π_k. □

3.2 DESIGN-BASED CONSIDERATIONS ABOUT THE APPROACH OF COCHRAN AND RUBIN (1973)

An important question to be made: is the sample mean a design-unbiased estimator of the population mean? The answer is: generally not. This point is discussed, for instance, in Skinner and Wakefield (2017) and Haziza and Beaumont (2017). Particularly, Haziza and Beaumont (2017) show that the bias incurred in using the sample mean as an estimator for the sample mean will only disappear if the first-order inclusion probability π_k is not correlated with the response variable Y_{1k}, for all $k \in U$.

Recall the approach discussed in Section 2.2. It is important to note that, even though Cochran and Rubin (1973) do not mention it, their approach to inference imply assumptions about the sampling mechanism. In particular, they assume that the sample mean is an unbiased estimator of the population mean. Consider that $\mathscr{S} \subseteq U$ is the set of sampled units. This assumption is equivalent to state that

$$\mathrm{E}_r(\bar{y}_t|k \in \mathscr{S}) = \mathrm{E}_r(\bar{y}_t|k \notin \mathscr{S}) \tag{3.11}$$

where, as in Section 2.2, $\bar{y}_T$ is the sample mean of Y for sampled units in treatment level $t \in \{1, 0\}$.

This corresponds to the notion of *non-informative sampling* (SKINNER; WAKEFIELD, 2017). Specifically, for $\bar{y}_{tk}$, under the formulation of the problem as given by Equation 2.5 this assumption holds by construction, as each Y_{tk}, $k \in U$, is independently and identically distributed with mean μ_t – that is, because treatment effects are supposed constant. In practice, however, this can be nothing but an assumption which the investigator may or may not be willing to make. For this reason, using the HT estimator, defined in Section 3.1, instead of sample means, is a good way of preventing against assumptions failures arising from non-informative sampling designs. The HT estimator is unbiased under any sampling design.

Even if we imagine that the random samples, as referred to by Cochran and Rubin (1973), are obtained via some non-informative sampling procedure, still matching can produce bias on causal effects estimates: in the same sense as pointed out in last paragraph for the case of sampling mechanisms, matching can be informative.

It is not always possible to find good matches for all treated units in a study. This is the case when the covariates distributions in the treated group do not completely overlap the support of the covariates distribution in the control group. When this occurs, matches will only be found for those treated units within the common support between treatment and control covariates and the other, non-matched units, will generally be dropped from analysis. Unless the specification of the model in Equation 2.5 holds, with constant treatment effects, the quantity being estimated will change and will concern to treatment effects on the restricted set of matched units. Rosenbaum and Rubin (1985a) call this *incomplete matching*.

It raises a question about *external vs. internal validity* of the inferences one can draw: whereas the result of analysis is reliable – in the sense of bias reduction and precision increasing – for the matched units in the study, it does not generalize itself, under broad assumptions, to a larger population of interest. This issue has been discussed in a long-standing debate in experimental studies, receiving increasingly attention (see Campbell and Stanley (2015) and Yang, Qu and Li (2021)), but has been relatively less addressed in observational contexts (as exceptions, see Imbens and Rubin (2015, p. 359) and Khandker, Koolwal and Samad (2009, p. 59), for instance).

To illustrate, consider that the population model has the form as in Equation 3.12 below

$$Y_{tk} = \mu_{tk} + \beta(X_{tk} - \eta_t) + \varepsilon_{tk} \tag{3.12}$$

with all terms being equal to those in Equation 2.5, except for the fact that the treatment effect is not constant anymore. Now, we have that the average treatment effect is given by

$$\bar{\tau} = \mathrm{E}\Big(Y_{1k} - Y_{0k}\Big) = \mathrm{E}\Big(\mu_{1k} - \mu_{0k}\Big) - \beta\Big(\eta_1 - \eta_0\Big) \tag{3.13}$$

where $\mathrm{E}(\cdot)$ denotes expectation over the conceptual response population.

Consider now that we construct a matched sample but, instead of subsetting the sampled control group, only, we have – for lack of overlap – to subset the treatment group, as well. Again, denote $\mathscr{M} \subset \mathscr{S}$ the set of matched units. Suppose also that, within $\mathscr{M}$ we have perfect matches, so that for all x_{1k} there is a x_{0j} such that $x_{1k} = x_{0j}$. Then,

$$\begin{aligned}\mathrm{E}_m(\bar{y}_1 - \bar{y}_0) &= \mathrm{E}_m\{\bar{\mu}_1 + \beta(\bar{x}_1 - \eta_1)\} - \mathrm{E}_m\{\bar{\mu}_0 + \beta(\bar{x}_0 - \eta_0)\} \\ &= \mathrm{E}_m(\bar{\mu}_1 - \bar{\mu}_0) - \beta(\eta_1 - \eta_0) \end{aligned} \tag{3.14}$$

and bias may arise whenever $\mathrm{E}_m(\bar{\mu}_1 - \bar{\mu}_0) \neq \mathrm{E}(\bar{\mu}_1 - \bar{\mu}_0)$.

In summary, even if good matches can be found for all units in a study, it is indispensable that the investigator correctly take sampling weights into account when doing analysis. This is discussed, for example, by Ridgeway et al. (2015), DuGoff, Schuler and Stuart (2014) and Austin, Jembere and Chiu (2018). Moreover, whenever matches are found only for a subset of the sampled treated treated units, it will not be possible to estimate the average effect of the treatment for the entire population, unless the treatment effect is constant for every unit.

3.3 MATCHING ESTIMATION AS A DOMAIN ESTIMATION PROBLEM

If we restrict ourselves to situations where matching is non-informative – or if we resign ourselves to estimate causal effects only for a subset of treated units, one way by which we can tackle the estimation problem is by viewing each matched treatment group as a sample from a domain.

In general terms, a domain is any subpopulation for which we may want to produce estimates. The sample size corresponding to a domain is not necessarily controlled by the sampling design: sometimes, a sampling frame informing which population units belong to each domain of interest is not available; at other times, a specific subpopulation is only perceived to be of interest after the field work has been completed. When this occurs, the sample size corresponding to a given domain is a random variable and, in practice, can be very small. For these reasons, estimation for domains is a topic in survey sampling with its own challenges.

Here, we only give a brief outline of estimation for domains, relating it to matching estimation, in the hope of inciting a discussion that may be treated with more complexity – possibly with fruitful outcomes – later on. We follow the developments of Särndal, Swensson and Wretman (1992, Chapter 10), to whom we refer for further details.

For convenience, suppose that, for our purposes, we only need to subset the control group, in matching, leaving the treatment group intact. The subset of the population control group composed of potential matches to the treatment units is our subpopulation of interest. Of course, for each matching metric we will have a corresponding subset. Thus, our subpopulation can, in general terms, be defined as

$$M_0 = \{k \mid d(b_k, b_j) < \gamma \text{ for some } j \in U_1\} \subseteq U_0 \tag{3.15}$$

where γ is an arbitrarily chosen non-negative value, $d(\cdot)$ is any distance metric – such as the Euclidean, Mahalanobis and Genetic distances – and b_k is the value of a balancing score for the kth unit, as before (see Sections 2.3 and 2.4).

In a finite population – for a given metric – we can define fixed variables indicating whether a unit belongs to the domain of interest. We denote it by

$$Z_k = \begin{cases} 1, & \text{if } k \in M_0 \\ 0, & \text{otherwise.} \end{cases} \tag{3.16}$$

Then, the subpopulation size is given by

$$N_{M_0} = \sum_{k \in U_0} Z_k \tag{3.17}$$

In a similar fashion, the total of the response variable Y in the domain is given by

$$t_{Y_{M_0}} = \sum_{k \in U_0} Z_k Y_k \tag{3.18}$$

If we have information so as to identify, at the design stage, the subset of control potential matches, we can treat that subpopulation as a stratum, hence controlling the number of sampled units which belong to that domain, in a similar manner as proposed by Ferraz and Vieira (2013). Instead, if we do not know beforehand which units belong to the domain – as is frequently the case –, the domain mean for the response variable is defined as a ratio of two unknown quantities

$$\bar{Y}_{M_0} = \frac{t_{Y_{M_0}}}{N_{M_0}} = \frac{\sum_{k \in U_0} Z_k Y_k}{\sum_{k \in U_0} Z_k} \tag{3.19}$$

In this case, we use the Horvitz-Thompson principle to get an approximately design-unbiased estimator as

$$\hat{\bar{Y}}_{M_{0\pi}} = \frac{\sum_{k \in \mathscr{S}_0} d_k Y_k}{\sum_{k \in \mathscr{S}_0} d_k Z_k} = \frac{\sum_{k \in \mathscr{M}_0} d_k Y_k}{\sum_{k \in \mathscr{M}_0} d_k Z_k} \tag{3.20}$$

where

$$\mathscr{M}_0 = M_0 \cap \mathscr{S}_0$$

Then, a matching estimator for the FPATT can be given as

$$\hat{\bar{\tau}}_{1_M} = \hat{\bar{Y}}_{11_\pi} - \hat{\bar{Y}}_{M_{0\pi}} \tag{3.21}$$

where $\hat{\bar{Y}}_{11_\pi}$ is given by equation 3.7.

We want to make two important remarks. Firstly, it is easy to see that, in this case, the number of sampled units which belong to the domain is a random variable and can be denoted by

$$n_{\mathscr{M}_0} = \sum_{k \in U_0} Z_k S_k \tag{3.22}$$

where, as before, S_k denotes the sample membership indicator. Hence, over repeated samples, the expected number of units sampled from the domain is given by

$$\mathrm{E}_p(n_{\mathscr{M}_0}) = \sum_{k \in U_0} Z_k \pi_k = \sum_{k \in M_0} \pi_k \tag{3.23}$$

As a consequence, assuming the number of potential matches in the control population is non-negligible, to guarantee a reasonable sample fraction of them, the sampling design must assign relatively higher inclusion probabilities to the individuals which belong to that subpopulation. We consider a way to do that in Section 3.4, making use of auxiliary information.

Secondly, the estimator 3.20 should be used very carefully. The reason is that the indicator variables Z_k are defined as population quantities and so, in order to use them from a sample, we must guarantee that their respective sample values are correctly measured, so as to properly identify the sample units which belong to the domain, as defined in equation 3.15. This is crucial when the matching metric to be used includes functions of the measured variables, such as the propensity score.

To clarify, let us illustrate it briefly. Suppose that our domain of interest is defined using the Euclidean distance and the propensity scores. Then, from equation 3.15, we have

$$M_0 = \{k \mid \|\theta_k - \theta_j\|_2 < \gamma \text{ for some } j \in U_1\} \subseteq U_0 \tag{3.24}$$

where $\|\cdot\|_2$ denotes the Euclidean distance function and θ_k denotes the propensity score for the kth unit, as in section 2.3. To retrieve the population indicator variables from the sample, it is required that either one knows the true propensity scores for the sampled units or one correctly estimates their true values. Thus, it is imperative that sampling design information is taken into account. This reinforces, using different arguments, the results presented by Ridgeway et al. (2015), who conclude that sampling weights must be used both when estimating the propensity scores and when estimating causal effects. Besides, an additional variance is expected for the estimator, as a logical consequence of using estimated values (of the propensity score, in that case) to compute the domain indicator variables in the sample.

3.4 THE ROLE OF AUXILIARY VARIABLES

Auxiliary information is, in a broad concept, any information about the population available independently of the sampling process itself (TILLÉ, 2020, p. 9). In general, auxiliary information is used to increase the precision of sampling estimates. There are many ways whereby one can use these information, either in the sampling design or analysis. Particularly important for our purposes are balanced sampling and calibration techniques, that use auxiliary information in design and analysis, respectively. In this section, we present an overview of these techniques. For the case of balanced sampling, we suppose that we have information on the covariate values for each unit in the population, treating them as our auxiliary variables. Even in the case of a single covariate, we denote the the set of auxiliary information for the kth unit in the population as

$$\mathbf{X}_k = (1 \quad X_k)^{\mathrm{T}}, \forall k \in U$$

For the general case of calibration and balanced sampling, doing so permits one to, on the one hand, unbiasedly estimate the population size or, on the other hand, control to some extent the actual sample size. For more details, the reader is referred to Tillé (2011) and Tillé (2020).

3.4.1 Calibration

Differently from the other case considered in this thesis – in balancing samples (see Section 3.4.4) –, the calibration technique uses auxiliary variables after the sample is drawn. The formalization of calibration technique was made by Deville and Särndal (1992). Here, we use as references the presentations of Silva (2004), Särndal (2007) and Tillé (2020), as well.

Originally, the idea behind the calibration method consists of changing the basic design weights in order to obtain an estimator *consistent* with the population totals and means. Mathematically, it corresponds to finding weights w_k such that

$$\sum_{k \in \mathscr{S}} w_k \mathbf{X}_k = \mathbf{T}_X \tag{3.25}$$

where $\mathbf{X}_k$ is a vector of auxiliary variables and $\mathbf{X} = \sum_{k \in U} \mathbf{X}_k$ is the vector of auxiliary variables totals.

In general, the weights w_k are sought in such a way that they are not so distant from the basic design weights. Define the function

$$G_k(w_k, d_k)$$

We want $G_k(d_k, w_k)$ to be positive, strictly convex, differentiable with respect to w_k and such that $G_k(d_k, d_k) = 0$.

Then, the problem of finding w_k may be stated as follows:

$$\begin{aligned} \min_{w_k} \quad & \sum_{k \in \mathscr{S}} G_k(d_k, w_k) \\ \text{s.t.} \quad & \sum_{k \in \mathscr{S}} w_k \mathbf{X}_k = \mathbf{T}_X \end{aligned} \tag{3.26}$$

Writing the Lagrange function as

$$\mathscr{L}(w_k, \boldsymbol{\lambda}) = \sum_{k \in \mathscr{S}} G_k(d_k, w_k) - \boldsymbol{\lambda}^{\mathrm{T}} \Big(\sum_{k \in \mathscr{S}} w_k \mathbf{X}_k - \mathbf{T}_X \Big) \tag{3.27}$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_j, \ldots, \lambda_p)^{\mathrm{T}}$, p being the dimension of $\mathbf{X}_k$.

Differentiating Equation 3.27 with respect to w_k and equaling to zero, we obtain

$$\frac{\partial \mathscr{L}(w_k, \lambda_p)}{\partial w_k} = g_k(d_k, w_k) - \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{X}_k = 0 \tag{3.28}$$

where

$$g_k(d_k, w_k) = \frac{\partial G_k(d_k, w_k)}{\partial w_k}$$

We have, then

$$w_k = d_k F_k(\mathbf{X}_k^{\mathrm{T}} \boldsymbol{\lambda}) \tag{3.29}$$

where $F_k(\cdot)$ is the inverse of the function $g_k(\cdot, d_k)$ (TILLÉ, 2020, p. 238). The function $F_k(\cdot)$ is called the calibration function. The calibration weight are found by solving the calibration equation

$$\sum_{k \in \mathscr{S}} d_k \mathbf{X}_k F_k(\mathbf{X}_k^{\mathrm{T}} \boldsymbol{\lambda}) = \mathbf{T}_X \tag{3.30}$$

for $\boldsymbol{\lambda}$. We denote the $\boldsymbol{\lambda}$ value satisfying Equation 3.30 by $\boldsymbol{\lambda}^*$. Then, the final calibration weights are given by

$$w_k = d_k F_k(\mathbf{X}_k^{\mathrm{T}} \boldsymbol{\lambda}^*) \tag{3.31}$$

Definition 3.3. *(Calibration Estimator) Let Y be a population quantity, the total of which we want to estimate. Then*

$$\hat{Y}_w = \sum_{k \in \mathscr{S}} w_k Y_k \tag{3.32}$$

where w_k is defined as in Equation 3.31, is called the calibration estimator of the Y total.

Many distance functions have been proposed in the literature, each one leading to a different calibration function. Here, we consider the simplest case, where

$$G_k(w_k, d_k) = \frac{(w_k - d_k)^2}{2d_k}$$

with corresponding calibration equation given by

$$F_k(\mathbf{X}_k^{\mathrm{T}}\boldsymbol{\lambda}) = 1 - \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{X}_k \tag{3.33}$$

Substituting Equation 3.33 in Equation 3.30, we get the corresponding calibration equation

$$\sum_{k\in\mathscr{S}} d_k\mathbf{X}_k(1 - \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{X}_k) = \hat{\mathbf{T}}_X + \sum_{k\in\mathscr{S}} d_k\mathbf{X}_k\mathbf{X}_k^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{T}_X$$

Now, if the matrix $\sum_{k\in\mathscr{S}} d_k\mathbf{X}_k\mathbf{X}_k^{\mathrm{T}}$ is non-singular, then we get

$$\boldsymbol{\lambda}^* = (\sum_{k\in\mathscr{S}} d_k\mathbf{X}_k\mathbf{X}_k^{\mathrm{T}})^{-1}\mathbf{T}_X - \hat{\mathbf{T}}_X \tag{3.34}$$

and, substituting Equation 3.34 in 3.31, and applying altogether in Equation 3.32 we have

$$\begin{aligned}\hat{Y}_w = \sum_{k\in\mathscr{S}} w_kY_k &= \hat{Y} + (\mathbf{T}_X - \hat{\mathbf{T}}_X)^{\mathrm{T}}(\sum_{k\in\mathscr{S}} d_k\mathbf{X}_k\mathbf{X}_k^{\mathrm{T}})^{-1}\sum_{k\in\mathscr{S}} d_k\mathbf{X}_kY_k \\ &= \hat{Y} + (\mathbf{T}_X - \hat{\mathbf{T}}_X)^{\mathrm{T}}\hat{\mathbf{B}} \end{aligned} \tag{3.35}$$

The estimator given in Equation 3.35 is equal to the regression estimator, which is very well discussed in Särndal, Swensson and Wretman (1992). Note that the calibration estimators of totals and means of the auxiliary variables have null variances. In addition, if some variable of interest, say Y, have the form $Y_k = \mathbf{X}_k^{\mathrm{T}}\mathbf{B}$ for any constant vector $\mathbf{B}$, then the calibrated estimators of the total and mean of Y have null variances, as well. In other words, the regression estimator explores the relationship between the Y variable and the auxiliary variable, using it to *assist* inferences about the finite population.

It is worthy to emphasize this perspective, as it is in the core of the justification of the approach presented in the next chapters. To cite Särndal, Swensson and Wretman (1992): “We do not require the model to be ‘true’ in the sense of correctly depicting some process by which the population data have been generated. We only believe that the population data can be fairly well described by the model” (SÄRNDAL; SWENSSON; WRETMAN, 1992, p. 239).

For a discussion on the use of different distance functions to obtain calibration weights, see Tillé (2020, Chapter 12). For a more detailed description of the regression estimator and its properties, see Särndal, Swensson and Wretman (1992, Chapter 6).

3.4.2 Calibration and The Analogy Between NRF and Survey Non-Response

As mentioned in Section 3.4.1, the primarily aim of using auxiliary variables in sampling design and analysis is to obtain efficiency gains in estimation. However, calibration techniques have been proposed as a method to handle problems arising from survey non-response, as well. In that case, calibration is used to limit the bias incurred when some sampled units fail to respond survey questionnaires.

In Section 3.4.3, we discuss how the ideas developed in that context can fit the problem of estimating treatment effects in observational studies. We explore the frequently mentioned analogy between the Fundamental Problem of Causal Inference (see Section 2.1) and survey estimation in the presence of non-response (see, for instance, Imbens and Rubin (2015) and Rubin (1990), among many others). But first, it is important to discuss the analogy between the NRF and the theory of survey non-response. We will only discuss notions of survey non-response theory to the extent that it will be helpful for our purposes.

Usually, the theoretical approach to handling non-response is conceived under the assumption that response has a probabilistic nature: every unit in population has a probability of responding the survey. This probability, often is viewed as depending on the corresponding unit's characteristics. Thus, the sampling process can be decomposed into two phases, one corresponding to the sample selection and the other corresponding to response selection. In the former, the sampled units are drawn with probabilities given by the sampling design, whereas in the latter, the probabilities are given by a response mechanism. The unit's response probability is called the response propensity.

As we discuss in Section 2.1, under the assumptions of the NRF, the problem in estimating causal effects is that we cannot observe both potential outcomes for every unit in the population. This is exemplified in Equations 2.2 and 2.4. The unobserved potential outcomes for each unit, thus, can be viewed as missing values with the response probabilities given by the treatment assignment mechanism (RUBIN, 1974; RUBIN, 1990; IMBENS; RUBIN, 2015, p. 14).

The ideas and terms used in the traditional formulation of non-response handling have a lot in common with the ones in the Neyman-Rubin Framework. This is partially due to the influence of Rubin (1976), for instance, who proposed a classification of response mechanisms in a very similar way as the one employed in the NRF. For this reason, the terms *ignorable* and *unconfounded* response mechanism are common in survey non-response jargon, and are defined in an analogous way as in NRF.

Exploring this analogy, in the regular conditions (see Appendix B), we have that the propensity score (see Definition 2.3) works as an analogous concept as the response propensity relative to the unit's potential outcome when allocated to the treatment group.

Nevertheless, there are differences between the two approaches. In particular, in the context of survey non-response, the classic approach suppose that the response selection occurs after the sampling selection (TILLÉ, 2020, p. 336). In the NRF, specifically in observational studies, the response selection is made previous to – and, thus, independently of – the sampling selection.

As a consequence, whereas in handling survey non-response, the inferential process has to take into account the response mechanism as a source of randomness (see the discussion in Haziza and Lesage (2016) and Lesage, Haziza and D'Haultfœuille (2019), for instance), in observational studies we can take the response assignment – that is, the treatment assignment – as given. In other words, in the classic two-phase approach to survey non-response, whenever a sample is drawn, two random selections occur, and this must be taken into account. Otherwise, in observational studies, our position is that the fundamental aim is to make inferences about a finite population with a fixed response structure.

As references to a more complete discussion of the theory involved in survey non-response, we cite Särndal, Swensson and Wretman (1992, Chapter 15) and Tillé (2020, Chapter 16). Here, our development will also be based on the approach brought in Lundström and Särndal (1999), Särndal and Lundström (2005) and Särndal (2007). More recent debates include Haziza and Lesage (2016) and Lesage, Haziza and D'Haultfœuille (2019).

3.4.3 Calibration as Tool for Reducing Non-Response Bias

Consider that we have a finite population U, described in the same way as in Section 3.1. Suppose we draw a sample $\mathscr{S}$ from U by a probabilistic sampling design $p(\mathbf{s})$, giving first-order inclusion probabilities π_k, for $k \in U$. Denote, as previously, $d_k = \pi_k^{-1}$, the basic design weight. Suppose further we want to estimate

$$\sum_{k\in U} N^{-1}Y_{0k} \tag{3.36}$$

Let δ_{0k} denote a variable, indicating whether the kth unit in the population was assigned to the control group, that is

$$\delta_{0k} = \begin{cases} 1, & \text{if } k \in U_0 \\ 0, & \text{otherwise} \end{cases}$$

We can think of δ_{0k} as an indicator variable denoting whether the kth unit *responds* the value Y_{0k} or not.

We have that

$$\Pr(\delta_{0k} = 1) = 1 - \theta_k(X_k) = \phi_k \tag{3.37}$$

where $\theta_k(X_k)$ is the unit-level assignment probability, or, equivalently (as the regularity conditions discussed in Appendix B holds), the unit's propensity score. We call ϕ_k the kth unit response propensity.

It is easy to see that

$$\mathrm{E}_p(\sum_{k\in\mathscr{S}} d_k Y_{0k}) = \sum_{k\in U} \delta_{0k}Y_{0k} = \sum_{k\in U_0} Y_{0k} \tag{3.38}$$

However, we can propose a calibration-based estimator for $\sum_{k\in U_1} N^{-1}Y_{0k}$, using a similar approach as the one adopted by Lundström and Särndal (1999), Deville (2002) and Särndal and Lundström (2005), who advocate the usage of a calibration approach to limit the bias due to non-response.

If we assume that that the bias of the HT estimator in estimating $\sum_{k\in U} Y_{0k}$ comes solely from the differences in the covariates values between the treatment-control groups, then a calibration estimator for $\sum_{k\in U_1} Y_{0k}$ may be expressed as

$$\hat{Y}_{10_w} = \sum_{k\in\mathscr{S}_0} w_k Y_{0k} \tag{3.39}$$

with

$$w_k = d_k\Big\{1 + (\sum_{k\in\mathscr{S}_1} d_k\mathbf{X}_k - \sum_{k\in\mathscr{S}_0} d_k\mathbf{X}_k)^{\mathrm{T}}(\sum_{k\in\mathscr{S}_0} d_k\mathbf{X}_k\mathbf{X}_k^{\mathrm{T}})^{-1}\mathbf{X}_k\Big\}$$

We now propose a expression for the bias of the estimator in Equation 3.39, based on the development given in Lundström and Särndal (1999) and Särndal and Lundström (2005). Unlike them, however, we only consider the design bias, whereas they consider the bias under the response mechanism as well. As we have discussed in Section 3.4.2, that position comes from the fact that we do not intend to make inferences about populations under different realizations of the treatment assignments and, thus, treat the response mechanism as given.

Result 3.2 (Bias of the Calibration Estimator)**.** *Let $\hat{Y}_{0_w}$ be an estimator as given in Equation 3.39. Then*

$$\mathrm{E}_p(\hat{Y}_{0_w}) - \Big(\sum_{k\in U} Y_{0k} - \sum_{k\in U_0} Y_{0k}\Big) = \sum_{k\in U_0} E_k - \sum_{k\in U_1} E_k \tag{3.40}$$

where

$$E_k = Y_k - \mathbf{X}_k^{\mathrm{T}}\mathbf{B}_{\mathbf{U_0}}$$

and

$$\mathbf{B}_{\mathbf{U_0}} = \Big(\sum_{k\in U_0} \mathbf{X}_k\mathbf{X}_k^{\mathrm{T}}\Big)^{-1} \sum_{k\in U_0} \mathbf{X}_k Y_{0k}$$

Proof. See Appendix A.

If we believe that the relationship given by E_k in Result 3.2 represents a good description of the Y variables in the population, in particular, that Y_k and X_k have a linear relationship for every unit in both treatment levels in population, and that the response curves in each treatment group are parallel to each other – that is, that both treatment and control groups share a common constant value $\mathbf{B}_{\mathbf{U}} = \mathbf{B}_{\mathbf{U_0}} = \mathbf{B}_{\mathbf{U_1}}$ –, then we can get an unbiased estimator of the FPATT as

$$\hat{\bar{\tau}}_{1_w} = \hat{\bar{Y}}_{11_\pi} - \hat{\bar{Y}}_{10_w} \tag{3.41}$$

where, from Equations 3.7 and 3.39,

$$\hat{\bar{Y}}_{11_\pi} = N_1^{-1} \sum_{k\in\mathscr{S}_1} d_k Y_{1k}$$

$$\hat{\bar{Y}}_{10_w} = N_1^{-1} \sum_{k\in\mathscr{S}_0} w_k Y_{0k}$$

with

$$w_k = d_k \Big\{ 1 + (\sum_{k \in \mathscr{S}_1} d_k \mathbf{X}_k - \sum_{k \in \mathscr{S}_0} d_k \mathbf{X}_k)^{\mathrm{T}} (\sum_{k \in \mathscr{S}_0} d_k \mathbf{X}_k \mathbf{X}_k^{\mathrm{T}})^{-1} \mathbf{X}_k \Big\}$$

Note that, in order to use the calibration estimator in Equation 3.41 we do not need to have access to any population level auxiliary information. This situation is equivalent to the case described in Lundström and Särndal (1999) and Särndal and Lundström (2005), where they denote auxiliary information at the sample level as InfoS (see Särndal and Lundström (2005, p. 54)).

However, if we have access to population-level auxiliary information on the covariates values, we can use a double-calibration approach: instead of calibrating the control group covariate mean estimator on the mean estimator of the treated group covariate (see Equation 3.39), only, we can also calibrate the control group covariate mean estimator to make it consistent with its corresponding population mean, as in the traditional usage of calibration technique. The rationale behind this is that, as the arguments discussed in Section 3.4.1 suggest, we can decrease the variance of our estimator.

In this case, we proceed in two steps:

1. Find first-step weights w_k^1, for every $k \in \mathscr{S}_t$, such that

$$w_k^1 = d_k \Big\{ 1 + (\sum_{k \in U_t} \mathbf{X}_k - \sum_{k \in \mathscr{S}_t} d_k \mathbf{X}_k)^{\mathrm{T}} (\sum_{k \in \mathscr{S}_t} d_k \mathbf{X}_k \mathbf{X}_k^{\mathrm{T}})^{-1} \mathbf{X}_k \Big\} \tag{3.42}$$

2. For $k \in \mathscr{S}_0$, find second-step weights w_k^2 such that

$$w_k^2 = w_k^1 \Big\{ 1 + (\sum_{k \in \mathscr{S}_1} w_k^1 \mathbf{X}_k - \sum_{k \in \mathscr{S}_0} w_k^1 \mathbf{X}_k)^{\mathrm{T}} (\sum_{k \in \mathscr{S}_0} w_k^1 \mathbf{X}_k \mathbf{X}_k^{\mathrm{T}})^{-1} \mathbf{X}_k \Big\} \tag{3.43}$$

Our double-calibration estimator will, then, be

$$\hat{\hat{\tau}}_{1_w^2} = \hat{\hat{Y}}_{11_w^1} - \hat{\hat{Y}}_{10_w^2} \tag{3.44}$$

where

$$\hat{\hat{Y}}_{11_w^1} = N_1^{-1} \sum_{k \in \mathscr{S}_1} w_k^1 Y_{1k}$$

$$\hat{\hat{Y}}_{10_w^2} = N_1^{-1} \sum_{k \in \mathscr{S}_0} w_k^2 Y_{0k}$$

It is important to remark that regression estimators have been used for a long time in observational studies. However, usually they are derived from a model-based inference perspective. Imbens and Rubin (2015) show concerns about employing regression estimators to estimate causal effects. They argument that regression estimators can be very sensitive with relation to model mispecifications. For instance, in Imbens and Rubin (2015, pp. 335-336), they use the example of LaLonde's experimental data (LALONDE, 1986) to estimate the outcome of the treated in the absence of treatment, comparing experimental with non-experimental data. To this end, they use a series of polynomial regressions and show that the point estimators as well as the variances of these estimators, are very unstable, presenting significantly different values from one model specification to another, when using non-experimental data. This occurs in contrast with the experimental situation, where increasing the degree of the polynomial used in the model, besides mildly increasing variances, does not significantly affect the point estimators. For a detailed description of the problem, see Imbens and Rubin (2015), in the referred pages.

The evidence we produce in the simulation results (see Chapter 4) points to the reasoning that, at least from a design-based, model-assisted inference perspective, the estimated bias and variance of the calibration estimators tend to be stable. In particular, if we include covariate higher-order terms as auxiliary variables in the regression estimator – even if they are absent in the true population relationship between the response variable and the covariate –, the point estimators and their respective variances show to be pretty stable. See Section 4. This can support the view that, at least for models in a "neighbourhood" of the true model – considering a sequence of nested models – the estimates produced are stable, a question that needs further investigation. An interesting question, that was not addressed here, is whether using model-based regression estimators would result in significantly different estimates under the same model-misspecification setup.

3.4.4 Balanced Sampling

Throughout survey sampling history, the term "balanced sample" has been given many meanings. For a brief historical perspective, see Tillé (2020) and Tillé (2011). Here, we use the definition given in Deville and Tillé (2004).

Definition 3.4 (Balanced Sample)**.** *A sample $\mathscr{S}$ that satisfies*

$$\sum_{k \in \mathscr{S}} d_k \mathbf{X}_k = \mathbf{T}_X \tag{3.45}$$

is said to be balanced on the totals of the covariates variables. Here, d_k is the basic design weight as defined in Section 3.1.

Of course, if a sample is balanced on the totals of the auxiliary vector variables, then it is balanced on their means. To see this, just divide both sides of Equation 3.45 by N. Equation 3.45 is referred to as the balancing equation, whereas X_k is referred to as the vector of balancing variables.

Moreover, if a sampling design $p(\mathbf{s})$ is such that, for a given vector of values X_k, Equation 3.45 is always satisfied, then $p(\mathbf{s})$ is said to be a balanced sampling design. Deville and Tillé (2004) give various examples of commonly used sampling designs which are particular cases of balanced sampling designs, for other balancing variables. Importantly, any sampling design that gives fixed sized samples are balanced, with the balancing variable corresponding, for instance, to $\mathbf{X}_k = \mathbb{1}_{1 \times N}$, where $\mathbb{1}_{1 \times N}$ is any N-vector whose elements all equals 1. In fact, $\mathbf{X}_k$ could be an N-vector whose elements equals any constant.

Tillé (2011) discusses some advantages of balanced samples. Among them, in a similar manner as discussed in Section 3.4.1, if the balancing equations are fully satisfied, then the variances of estimators under balanced sampling can be largely reduced.

It is not always possible to find exactly balanced samples, especially if the dimension of the auxiliary vector is large. For this reason, the many methods proposed try to find samples which is, as best as it is possible, approximately balanced.

A way to do that was proposed by Fuller (2009). He considers a rejective sampling procedure, by which a sample is selected only if the difference between the sample estimates and the respective population quantities lies within a pre-defined threshold. In his case, Fuller (2009) considers the case where the sample is accepted if

$$\left(\hat{\bar{\mathbf{X}}} - \bar{\mathbf{X}}\right)^{\mathrm{T}} V_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1} \left(\bar{\mathbf{X}} - \bar{\mathbf{X}}\right) < \gamma \tag{3.46}$$

where $\hat{\bar{\mathbf{X}}} = N^{-1} \sum_{k \in \mathscr{S}} d_k \mathbf{X}_k$, $\bar{\mathbf{X}} = N^{-1} \mathbf{T}_X$ and

$$V_{\bar{\mathbf{X}}\bar{\mathbf{X}}} = N^{-2} \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \mathbf{X}_k^{\mathrm{T}} \mathbf{X}_k d_k d_l$$

the HT estimator design variance (see Särndal, Swensson and Wretman (1992) and Lohr (2009) as references).

The sampling procedure goes as follows: by a sampling design $p(\mathbf{s})$, repeated samples are drawn. At each repetition, the condition given by the metric in inequality 3.46 is assessed. If verified, the sample is accepted; otherwise, the process continues.

The appealing of using inequality 3.46 as a metric of balance is that, if the original sampling design has a central limit theorem, the left side in inequality 3.46 asymptotically follows a χ^2 distribution with degrees of freedom corresponding to the length of $\mathbf{X}$ (FULLER, 2009; LEGG; YU, 2010). Thus, we can have a guess on the rejection rate of the procedure by looking at the probability of the corresponding χ^2 distribution, associated with the quantile equal to γ.

While, originally, the idea behind drawing balanced samples was to make sample estimates of auxiliary variables consistent with their respective population values, to our purposes, the question to be answered will be: is it possible to obtain samples of treated and control units that are balanced on covariates through probabilistic sampling design? If so, do they provide good estimates of population treatment effects?

In order to do that we may stipulate a different criterion for sampling rejection. We can, for instance, draw samples requiring that the difference between the sample means in each sampled group are less than a pre-specified threshold, as follows

$$n_1^{-1} \sum_{k \in \mathscr{S}_1} X_k - n_0^{-1} \sum_{k \in \mathscr{S}_0} X_k < \gamma \tag{3.47}$$

where n_t is the size of each sampled group $\mathscr{S}_t$.

Note that, if we were to substitute each sample mean in inequality 3.47 for a design-unbiased estimator of the covariates means in each population group – as the HT estimator –, then the value γ would be virtually limited to the actual difference in means in the population, and no bias could be reduced.

We can combine the rejective rule in inequality 3.47 with an initial sampling design that ascribes greater inclusion probabilities for the units in one treatment group whose covariates values are closer to the other groups' covariates means. The rationale behind this is that units with covariate values closer to the other group's mean will tend to be oversampled and the restriction in inequality 3.47 will be more easily satisfied.

For instance, we can attribute π_k values such that

$$\pi_k = \begin{cases} \pi_k \propto \frac{1}{(X_k - \bar{X}_1)^2}, & \text{if } k \in U_0 \\ \pi_k \propto \frac{1}{(X_k - \bar{X}_0)^2}, & \text{if } k \in U_1 \end{cases} \tag{3.48}$$

where $\bar{X}_t$ is the mean of the covariate value in the treatment level U_t.

A concern about rejective procedures is that it alters the units initial inclusion probabilities. In the case presented in Fuller (2009), using the criterion in inequality 3.46, units have a greater chance of being included in the sample when their respective $\mathbf{X}_k$ values are closer to the population mean. Legg and Yu (2010) presented Monte Carlo simulation results which corroborate this rationale. However, both Fuller (2009) and Legg and Yu (2010) point out that the regression estimator (see Section 3.4.1) have practically the same properties under the rejective design and the original design. Furthermore, they show that the variance estimator of the expansion estimator will also be biased, and propose, in a similar manner, the use of regression estimators as a way to obtain efficient estimates.

In general, we expect the rejective sampling to perform well in terms of bias reduction only if the treatment effect is constant or if combined with one of the other methods discussed in this text. In other words, we do not expect that any direct estimator can be derived from the rejective sampling procedure, under non-constant treatment effects. The reason for that is that, although we expect rejective sampling to work well in the task of finding more balanced samples – which is very beneficial for the other methods –, if we were to use any design-unbiased estimator with the balanced samples obtained, we would eventually retrieve the respective population quantities. To illustrate, suppose that the HT estimator is design-unbiased under rejective sampling. This is not the case, as we have discussed, but it will serve as an illustration. If we estimate the treatment effect as

$$\hat{\bar{t}}_{rj} = N_1^{-1}\Big(\sum_{k \in \mathscr{S}_1} d_k Y_{1k}\Big) - N_0^{-1}\Big(\sum_{k \in \mathscr{S}_0} d_k Y_{0k}\Big) \tag{3.49}$$

then, obviously, we would have that

$$\mathrm{E}_p(\hat{\bar{t}}_{rj}) = \bar{Y}_{11} - \bar{Y}_{00} \tag{3.50}$$

Unavoidably, even if we do succeed in drawing balanced samples, we need to condition our estimators on the covariate values to get unbiased estimates of treatment effects. It is easy to use the rejective sampling in conjunction with the calibration estimator, for example. Once that the HT estimator is biased under rejective sampling, but the regression estimator is not, we can use a double-calibration approach as described previously in Section 3.4.3. As we will see in Chapter 4, balanced samples combined with double-calibration estimator can reduce the estimates variance.

4 SIMULATION DESIGN AND RESULTS

A Monte Carlo simulation was implemented to evaluate the performance of the different estimators presented in the last chapters. We generated a finite population U of size $N = 6000$, with $N_1 = 1000$ and $N_0 = 5000$. Each unit k in the population has values recorded for a single covariate X_{tk} and for the response variable Y_{tk}. We consider that X_{tk} were generated by Normal distributions with different means for each treatment level, as follows:

- $X_{1k} \sim \mathcal{N}(23, 1), \forall k \in U_1$
- $X_{0k} \sim \mathcal{N}(20, 1), \forall k \in U_0$

With these parameters, the value of initial bias, as calculated by the measure in Equation 2.8 is $D = 2.9905$.

We generated the response variable accordingly to the model

$$Y_{tk} = 100 + \tau_k t_k + X_{tk} + \varepsilon_k$$

where t_k is the treatment indicator of the kth unit, each ε_k is independently and identically distributed as $\varepsilon_k \sim \mathcal{N}(0, 1)$ and with τ_k defined in two ways

1. Constant, where $\tau_k = \tau = 10, \forall k \in U_1$;
2. Non-constant, where $\tau_k = -10 + 10X_{1k}$.

Next, we draw $R = 100$ repeated samples from U. To this end, three sampling designs were considered:

1. Simple Random Sampling Without Replacement – SRSWOR;
2. Poisson Sampling – POI;
3. Rejective Sampling with Initial Poisson Design – REJPOI.

Both SRSWOR and POI designs are very well-known sampling designs. For references, see Särndal, Swensson and Wretman (1992, p. 66) and Tillé (2020, p. 27), for SRSWOR, and Särndal, Swensson and Wretman (1992, p. 85) and Tillé (2020, p. 92), for POI.

Both in POI and REJPOI, the first-order inclusion probabilities were generated as described in Equation 3.48

$$\pi_k = \begin{cases} \pi_k \propto \frac{1}{(X_k - \bar{X}_1)^2}, & \text{if } k \in U_0 \\ \pi_k \propto \frac{1}{(X_k - \bar{X}_0)^2}, & \text{if } k \in U_1 \end{cases}$$

where $\bar{X}_t$ is the mean of the covariate value in the treatment level U_t.

In REJPOI, the rejection rule given by inequality 3.47 was set to be

$$n_1^{-1} \sum_{k \in \mathscr{S}_1} X_k - n_0^{-1} \sum_{k \in \mathscr{S}_0} X_k < 0.2$$

In total, six scenarios were created, defined by combinations of the two definitions of τ_k and the three sampling designs.

Samples of size $n_1 = 250$ and $n_0 = 500$ were drawn from U_1 and U_0, respectively, under SRSWOR. In the case of POI, these were set to be the expected sample sizes in each group, by the construction of the inclusion probabilities.

The REJPOI design showed to be very sensitive to the expected sample sizes, so that no samples could be drawn using the specifications for POI, for example. Because of this, the expected sample size when using REJPOI was set to be $\mathrm{E}_p(n_t) = 100$ for both treatment levels. Moreover, it is noteworthy that no samples could be drawn for distributions with larger differences in covariates means.

For each sample drawn, we compute the estimate of $\bar{\tau}_1 = N_1^{-1} \sum_{k \in U_1} \tau_k$, the Finite Population Average Treatment Effect on the Treated (FPATT).

We compute the estimated absolute bias for each estimator $\hat{\hat{\tau}}$ as

$$\hat{\mathrm{B}}(\hat{\hat{\tau}}) = R^{-1} \sum_{i \in R} \hat{\hat{\tau}}_i - N_1^{-1} \sum_{k \in U_1} \tau_k$$

The estimated variance for each estimator was computed as

$$\hat{\mathrm{V}}(\hat{\hat{\tau}}) = \Big(R - 1\Big)^{-1} \sum_{i \in R} \Big(\hat{\hat{\tau}}_i - R^{-1} \sum_{i \in R} \hat{\hat{\tau}}_i\Big)^2$$

At last, the estimated mean-squared error (MSE) of each estimator was computed as

$$\mathrm{M\hat{S}E}(\hat{\hat{\tau}}) = \hat{\mathrm{V}}(\hat{\hat{\tau}}) + \hat{\mathrm{B}}(\hat{\hat{\tau}})^2$$

We enumerate the scenarios as follows

1. SRSWOR

    a) Constant treatment effect;

   b) Non-constant treatment effects;

2. POI

   a) Constant treatment effect;

   b) Non-constant treatment effects;

3. REJPOI

   a) Constant treatment effect;

   b) Non-constant treatment effects;

We evaluated the performance of six estimators:

- Propensity Score Matching Without Replacement estimator – PSMwor;
- Propensity Score Matching With Replacement estimator – PSMwr;
- Genetic Matching Without Replacement estimator – GMwor;
- Genetic Matching With Replacement estimator – GMwr;
- Calibration estimator – CALIB;
- Double-Calibration estimator - DCALIB.

Note, again, that the calibration estimator was not computed in the case of REJPOI.

We now recall the form of each one of the estimators.

For both matching methods, the treatment effect estimator is given by the difference between the matched sample means in treated and control group. Let $\mathscr{M} \subset \mathscr{S}$ be the set of matched units in the sample, either by Propensity Score Matching or Genetic Matching. Let M denote the cardinality of $\mathscr{M}$. Then, the matching estimator for $\bar{\tau}_1$ is given by

$$\hat{\bar{\tau}}_{\mathscr{M}} = M^{-1}\Big(\sum_{k\in\mathscr{M}} Y_{1k} - \sum_{j\in\mathscr{M}} Y_{0j}\Big)$$

Note that the forms of with and without-replacement matching estimators are identical, with the difference in them corresponding to the units composing the matching subset.

The calibration estimator is given by

$$\hat{\hat{\tau}}_{1_w} = N_1^{-1}\Big(\sum_{j\in\mathscr{S}_1} d_j Y_{1j} - \sum_{k\in\mathscr{S}_0} w_k Y_{0k}\Big)$$

where

$$w_k = d_k\Big\{1 + (\sum_{k\in\mathscr{S}_1} d_k\mathbf{X}_k - \sum_{k\in\mathscr{S}_0} d_k\mathbf{X}_k)^{\mathrm{T}}(\sum_{k\in\mathscr{S}_0} d_k\mathbf{X}_k\mathbf{X}_k^{\mathrm{T}})^{-1}\mathbf{X}_k\Big\}$$

In turn, the double-calibration estimator is given by

$$\hat{\hat{\tau}}_{1^2_w} = N_1^{-1}\sum_{k\in\mathscr{S}_1} w_k^1 Y_{1k} - N_1^{-1}\sum_{k\in\mathscr{S}_0} w_k^2 Y_{0k}$$

where

$$w_k^1 = d_k\Big\{1 + (\sum_{k\in U_t} \mathbf{X}_k - \sum_{k\in\mathscr{S}_t} d_k\mathbf{X}_k)^{\mathrm{T}}(\sum_{k\in\mathscr{S}_t} d_k\mathbf{X}_k\mathbf{X}_k^{\mathrm{T}})^{-1}\mathbf{X}_k\Big\}$$
$$w_k^2 = w_k^1\Big\{1 + (\sum_{k\in\mathscr{S}_1} w_k^1\mathbf{X}_k - \sum_{k\in\mathscr{S}_0} w_k^1\mathbf{X}_k)^{\mathrm{T}}(\sum_{k\in\mathscr{S}_0} w_k^1\mathbf{X}_k\mathbf{X}_k^{\mathrm{T}})^{-1}\mathbf{X}_k\Big\}$$

A summary of the performance of each estimator under each scenario is given in Table 3.

From Table 3, we can see that the CALIB estimator performs well in all scenarios, except under POI with non-constant treatment effects, where its variance is very large, even though its absolute bias is relatively small. As expected, in this situation, the DCALIB estimator represents a way whereby one can prevent the variance to increase.

In what matching estimators are concerned, we can see that, as expected, matching with replacement performs better in terms of bias reduction, comparing with the without-replacement case, with an increase in the variance. Anyway, the variances of all matching estimators are very small relatively to the true parameter values, and that increase in variance should not be a problem when drawing inferences. In general, considering matching without replacement, GM and PSM performances are similar. Considering matching with replacement, otherwise, GM performs much better than PSM. As expected, PSM and GM perform better when treatment effects are constant. Differently from what expected, however, even with non-constant treatment effects, under SRSWOR, both matching estimators do not seem to perform worse than in the case of constant treatment effects. This is probably due to the fact that the population distributions are not sufficiently apart from one group to another and, in addition, that the average treatment effect for the matched units are not very far from the average treatment

Table 3 – Estimated Absolute Bias and Variance of PATT Estimators in Each Simulated Scenario.

		Scenario					
		1a.	**1b.**	**2a.**	**2b.**	**3a.**	**3b.**
PSMwor	**Bias**	2.2813	2.3046	0.5902	-6.6947	0.3567	-10.2572
	Variance	0.0151	0.3588	0.0099	0.2902	0.0330	3.9743
	MSE	5.2194	5.6700	0.3582	45.1092	0.1602	109.1845
GMwor	**Bias**	2.2818	2.3052	0.5886	-6.6963	0.1934	-11.6209
	Variance	0.0151	0.3584	0.0100	0.2880	0.0229	0.7171
	MSE	5.2217	5.6723	0.3564	45.12843	0.0603	135.7624
PSMwr	**Bias**	1.4676	1.4910	0.6185	-6.6664	0.1611	-11.6522
	Variance	0.0483	0.3626	0.0092	0.2673	0.0303	0.6236
	MSE	2.20215	2.5856	0.3917	44.7082	0.0562	136.3974
GMwr	**Bias**	0.6047	0.6281	0.2642	-7.0207	0.1090	-11.7043
	Variance	0.2650	0.5466	0.0056	0.2467	0.0275	0.5857
	MSE	0.6307	0.9411	0.0754	49.5369	0.0394	137.5763
CALIB	**Bias**	0.0923	0.1157	0.2609	1.8270	-	-
	Variance	0.0215	0.2849	0.1982	288.6043	-	-
	MSE	0.0300	0.2983	0.6200	291.9422	-	-
DCALIB	**Bias**	0.1043	0.1277	0.1214	0.1255	0.1841	-0.0401
	Variance	0.0950	0.6456	0.0356	0.2270	0.1580	0.1656
	MSE	0.1059	0.6619	0.0503	0.2428	0.1919	0.1672

Source: The author (2022)

effect for the entire population. We expect that, if the individual-level treatment effect vary more, the performance of matching will be poorer.

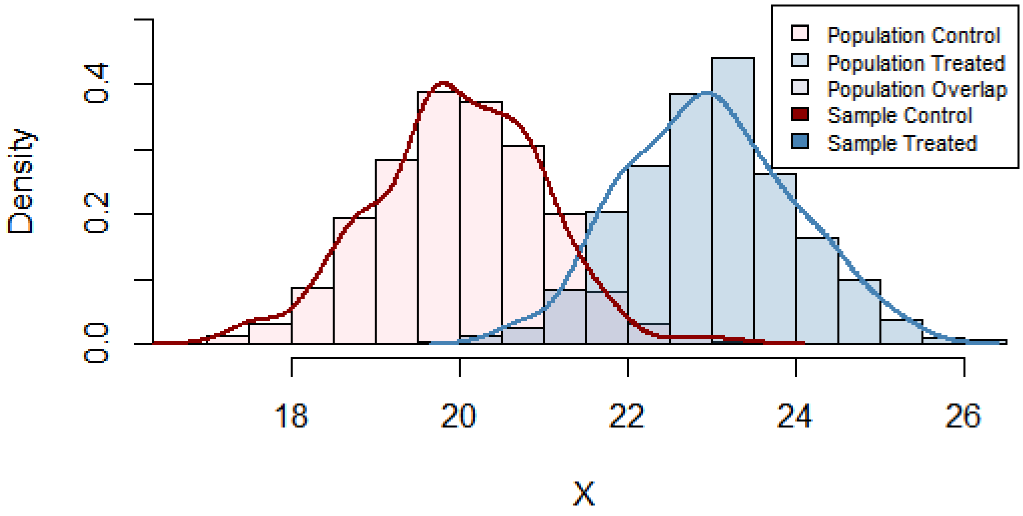
Nonetheless, once we consider informative sampling designs with non-constant treatment effects (scenarios 2b. and 3b.), both matching methods perform very poorly. This fact shows the importance of taking the sampling design into account, using sampling weights, for example.

An interesting aspect is that, although matching without replacement performs poorly in terms of bias even for the case of constant treatment effects in SRSWOR, as we change the sampling design appropriately, their performance get better and better. The rationale behind that fact is that the sampling distributions of the covariate in the treatment and control groups tend to overlap each other more and more, as we change from scenarios 1 to 3, as we discuss below, allowing for better matches to be found.

We expected that the rejective rule in REJPOI would improve the performances of the estimators. We see that, in the case of non-constant treatment effects, even with less than a half the sample sizes considered in the other scenarios, the estimated bias and variance of the double-calibration estimator are very small.

Moreover, we can see that, in fact, REJPOI achieves a greater covariate balance between the treated and control sampled groups. In Figures 1, 2 and 3, we show the distributions of treated and control covariates both at the population-level – showed as the histograms – and in one of the $R = 100$ repeated samples, arbitrarily chosen, respectively for SRSWOR, POI and REJPOI. The pink histogram represents the control distribution and the blue histogram represents the treated distribution at the population-level. Likewise, the red line represents the control distribution and the blue line represents the treated distribution in the sample.

Figure 1 – Distributions of covariate values in the population and in SRSWOR samples, for treated and control groups.

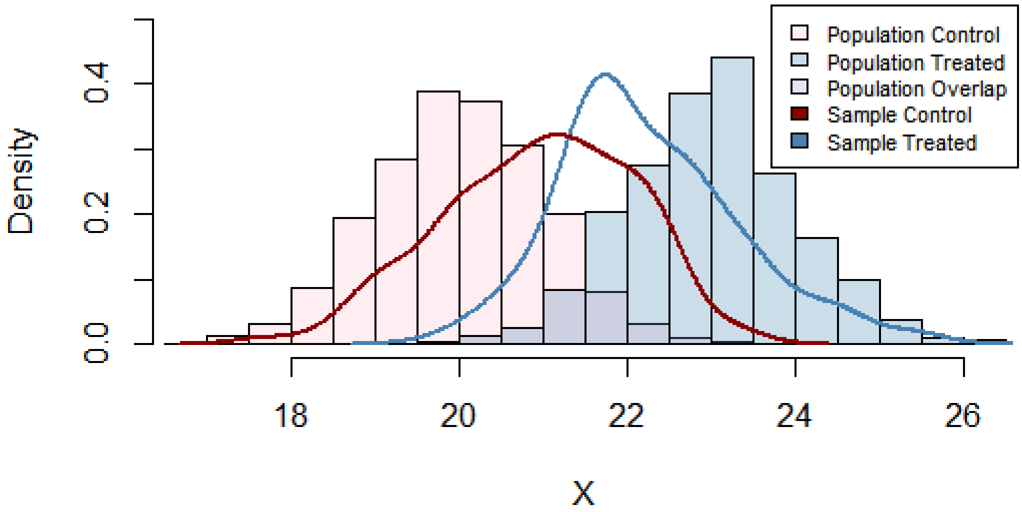


Source: The author (2022)

As expected, in the case of SRSWOR, Figure 1, the sampling design preserves the balancing structure seen at the population level. We can see that in the case of REJPOI, Figure 3, the sample distributions of the covariates practically overlap one another. The case of POI, Figure 2, shows an intermediary degree of balance, in comparison with SRSWOR and REJPOI.

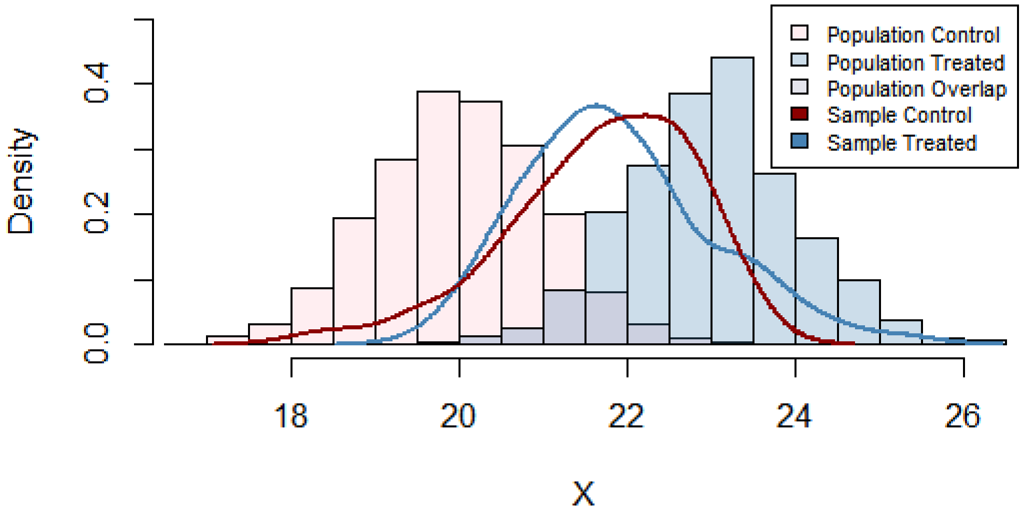
These results suggest that, if matching procedures correctly incorporate sampling weights, maybe they could produce reliable estimates even if the sampling design is informative and if treatment effects vary more across the population units. In particular, with REJPOI, we expect the procedure of matching to be facilitated.

Figure 2 – Distributions of covariate values in the population and in POI samples, for treated and control groups.



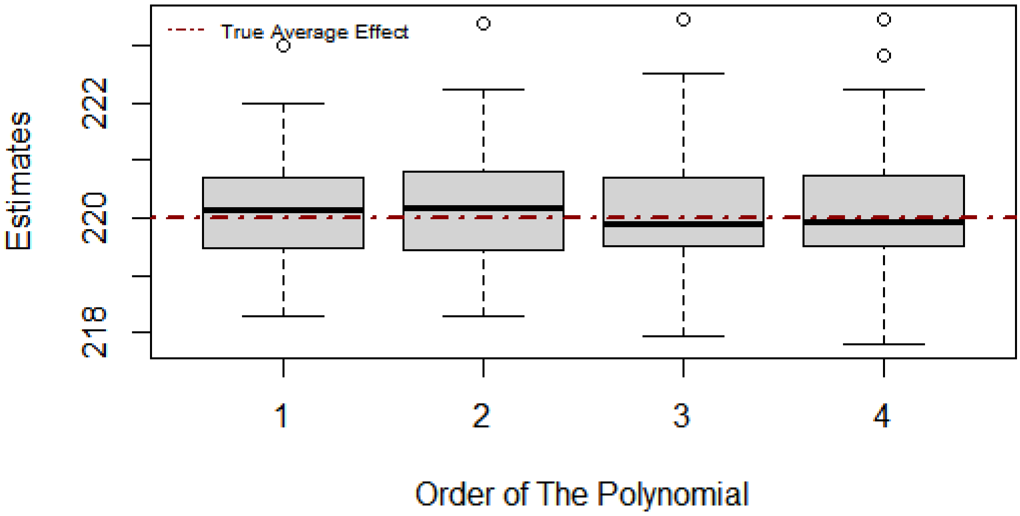
Source: The author (2022)

Figure 3 – Distributions of covariate values in the population and in REJPOI samples, for treated and control groups.



Source: The author (2022)

Figure 4 – Boxplots of double-calibration estimator with higher order terms included for non-constant treatment effects under SRSWOR.

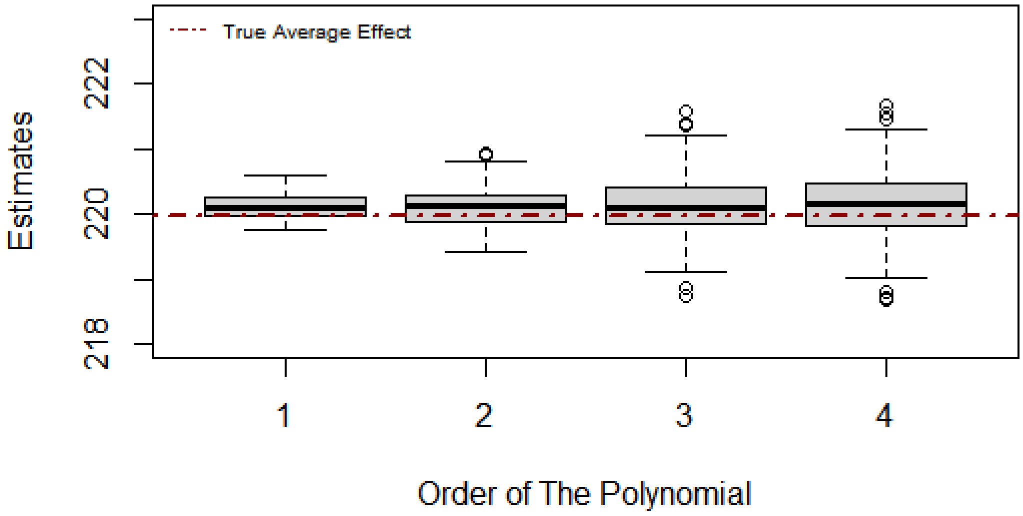


Source: The author (2022)

Another aspect we want to assess, via simulation, is the stability of the calibration estimator in face of model mispecification. In order to do that, we included in the specification of the calibration weights higher order terms of the covariate values. Particularly, we included terms of order $m = 1, 2, 3, 4$.

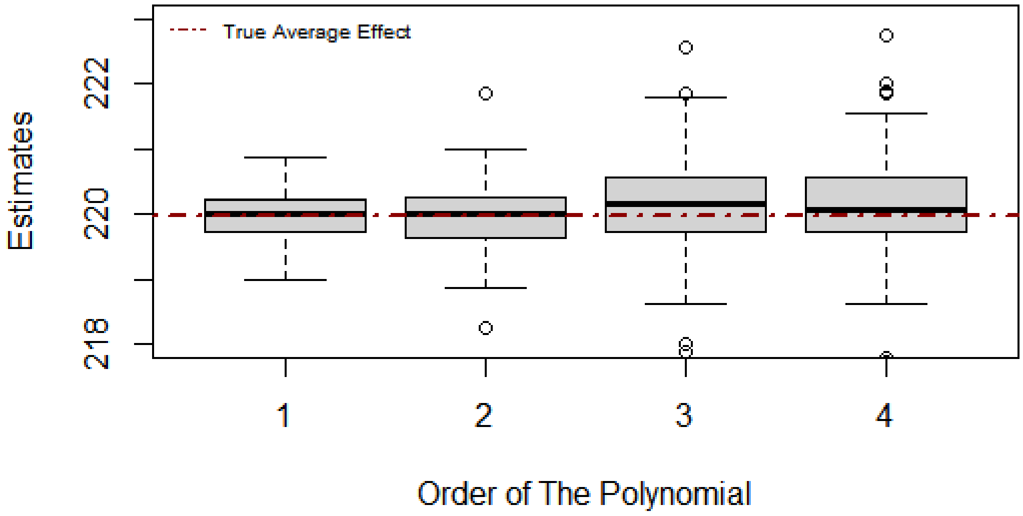
Figures 4, 5 and 6 show boxplots of the double-calibration estimator under each one of the specifications, for SRSWOR, POI and REJPOI, respectively. As we mentioned in Section 3.4.3, we can see that the expected point estimates and variances are very stable for the three scenarios: the median estimate, for each model, in each scenario, is very close to the true parameter; besides, the lengths of the intervals spanned by the estimates, for each model, in each scenario, practically overlap one another. Moreover, we can see that, albeit including additional terms in the model implies in additional variability among the estimates, this increase in the variance of the estimators is small relatively to the true parameter value. In summary, we can expect that including higher-order terms as auxiliary variables in the calibration estimator do not significantly change the inference one can draw, either in terms of point or interval estimates of the true parameter. The performances of the estimators for the other cases and scenarios were similar and were not presented here for the sake of brevity.

Figure 5 – Boxplots of double-calibration estimator with higher order terms included for non-constant treatment effects under POI.



Source: The author (2022)

Figure 6 – Boxplots of double-calibration estimator with higher order terms included for non-constant treatment effects under REJPOI.



Source: The author (2022)

5 CONCLUSION

This work discussed some of the main approaches available for estimation of causal effects in observational studies. We showed that, when the study regards to estimating causal effects in a finite population, the role of sampling is an important factor, that has to be accounted for. In particular, when using common approaches such as matching, if the investigator fails to incorporate sampling design information, such as sampling weights, causal effects estimates can be very biased, specifically when one is not able to match all treated units and treatment effects are not constant across the population units.

Apart from being a source of randomness that has to be taken into account, the sampling process can be used in order to obtain accurate and unbiased estimates of causal effects. From the sampling design perspective, the combination of rejective sampling with an appropriate initial design can improve balance between treated and control groups in the sample. This suggests that, if the analyst is able to appropriately incorporate sampling weights in the analysis, causal effects may be estimated with less bias and more precision.

From the sampling analysis perspective, the calibration and double-calibration estimators proposed in this work showed to perform very well in the estimation of causal effects. In particular, one has a lot to gain in using those estimators as alternatives to model-based regression estimators, which largely depends on model assumptions. The calibration estimators showed to be pretty stable even under model misspecification. Besides, the calibration approach represents a very straightforward way of using sampling weights.

Of course, only a simple case was addressed, that of a response variable with linear regression structures, single covariate and additive effects. The literature of calibration, however, has largely advanced in the direction of proposing calibrated weights that can encompass more complex relationships between the auxiliary variables and the response variable. Whereas the first formulations of the calibration method were presented in order to make estimators consistent with known population totals and means, there has been an increasing interest in deriving weight systems to make estimators consistent with other population quantities of interest. In particular, we mention the works of Harms and Duchesne (2006), who considers calibration on quantiles, Wu and Sitter (2001), who proposed model calibration, and Goga and Ruiz-Gazen (2014), who proposed nonparametric calibration. These are approaches that can be explored in future works.

REFERENCES

AGRESTI, A. *Categorical data analysis*. [S.l.]: John Wiley & Sons, 2003.

AUSTIN, P. C. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine*, Wiley Online Library, v. 27, n. 12, p. 2037–2049, 2008.

AUSTIN, P. C.; JEMBERE, N.; CHIU, M. Propensity score matching and complex surveys. *Statistical methods in medical research*, SAGE Publications Sage UK: London, England, v. 27, n. 4, p. 1240–1257, 2018.

BILLEWICZ, W. The efficiency of matched samples: an empirical investigation. *Biometrics*, JSTOR, p. 623–644, 1965.

CAMPBELL, D. T.; STANLEY, J. C. *Experimental and quasi-experimental designs for research*. [S.l.]: Ravenio Books, 2015.

CAMPOS, H. de. *Os melhores poemas de Haroldo de Campos*. [S.l.]: Global Editora, 1992.

COCHRAN, W. G. Matching in analytical studies. *American Journal of Public Health and the Nations Health*, American Public Health Association, v. 43, n. 6_Pt_1, p. 684–691, 1953.

COCHRAN, W. G. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, JSTOR, p. 295–313, 1968.

COCHRAN, W. G.; RUBIN, D. B. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, JSTOR, p. 417–446, 1973.

COX, D. R.; REID, N. *The theory of the design of experiments*. [S.l.]: CRC Press, 2000.

DAWID, A. P. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 41, n. 1, p. 1–15, 1979.

DEVILLE, J.-C. La correction de la non-réponse par calage généralisé. *Actes des journées de méthodologie statistique*, Insee-Méthodes, p. 4–20, 2002.

DEVILLE, J.-C.; SÄRNDAL, C.-E. Calibration estimators in survey sampling. *Journal of the American statistical Association*, Taylor & Francis, v. 87, n. 418, p. 376–382, 1992.

DEVILLE, J.-C.; TILLÉ, Y. Efficient balanced sampling: the cube method. *Biometrika*, Oxford University Press, v. 91, n. 4, p. 893–912, 2004.

DIAMOND, A.; SEKHON, J. S. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, The MIT Press, v. 95, n. 3, p. 932–945, 2013.

DUGOFF, E. H.; SCHULER, M.; STUART, E. A. Generalizing observational study results: applying propensity score methods to complex surveys. *Health services research*, Wiley Online Library, v. 49, n. 1, p. 284–303, 2014.

FERRAZ, C.; VIEIRA, M. D. T. Sample design for impact evaluation of welfare programs: The yemen case. In: *59th ISI World Statistics Congress, 2013, Hong Kong*. [S.l.: s.n.], 2013. Único, p. 5498-5502.

FISHER, R. A. Statistical methods for research workers. In: *Breakthroughs in statistics*. [S.l.]: Springer, 1992. p. 66–70.

FULLER, W. A. Some design properties of a rejective sampling procedure. *Biometrika*, Oxford University Press, v. 96, n. 4, p. 933–944, 2009.

GOGA, C.; RUIZ-GAZEN, A. Efficient estimation of non-linear finite population parameters by using non-parametrics. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, JSTOR, p. 113–140, 2014.

GREENBERG, B. G. The use of analysis of covariance and balancing in analytical surveys. *American Journal of Public Health and the Nations Health*, American Public Health Association, v. 43, n. 6_Pt_1, p. 692–699, 1953.

HARMS, T.; DUCHESNE, P. On calibration estimation for quantiles. *Survey methodology*, v. 32, n. 1, p. 37, 2006.

HAZIZA, D.; BEAUMONT, J.-F. Construction of weights in surveys: A review. *Statistical Science*, Institute of Mathematical Statistics, v. 32, n. 2, p. 206–226, 2017.

HAZIZA, D.; LESAGE, É. A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, Statistics Sweden (SCB), v. 32, n. 1, p. 129, 2016.

HINKELMANN, K.; KEMPTHORNE, O. *Design and analysis of experiments, volume 1: Introduction to experimental design*. [S.l.]: John Wiley & Sons, 2007.

HOLLAND, P. W. Statistics and causal inference. *Journal of the American statistical Association*, Taylor & Francis, v. 81, n. 396, p. 945–960, 1986.

HOLLAND, P. W. Causation and race. *White logic, white methods: Racism and methodology*, Rowman & Littlefield New York, NY, p. 93–109, 2008.

HOLLAND, P. W.; RUBIN, D. B. *Causal Inference in Prospective and Retrospective Studies*. [S.l.], 1980.

HOLLAND, P. W.; RUBIN, D. B. On lord's paradox. *Principals of modern psychological measurement*, Erlbaum Hillsdale, NJ, p. 3–25, 1983.

HORVITZ, D. G.; THOMPSON, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, Taylor & Francis Group, v. 47, n. 260, p. 663–685, 1952.

IMBENS, G. W.; RUBIN, D. B. *Causal inference in statistics, social, and biomedical sciences*. [S.l.]: Cambridge University Press, 2015.

KEMPTHORNE, O. The design and analysis of experiments. Wiley, 1952.

KEMPTHORNE, O. The randomization theory of experimental inference. *Journal of the American Statistical Association*, Taylor & Francis, v. 50, n. 271, p. 946–967, 1955.

KHANDKER, S. R.; KOOLWAL, G. B.; SAMAD, H. A. *Handbook on impact evaluation: quantitative methods and practices*. [S.l.]: World Bank Publications, 2009.

LALONDE, R. J. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, JSTOR, p. 604–620, 1986.

LEGG, J. C.; YU, C. L. A comparison of sample set restriction procedures. *Survey Methodology*, v. 36, n. 1, p. 69–79, 2010.

LESAGE, É.; HAZIZA, D.; D'HAULTFŒUILLE, X. A cautionary tale on instrumental calibration for the treatment of nonignorable unit nonresponse in surveys. *Journal of the American Statistical Association*, Taylor & Francis, v. 114, n. 526, p. 906–915, 2019.

LOHR, S. L. *Sampling: design and analysis*. [S.l.]: Nelson Education, 2009.

LUNDSTRÖM, S.; SÄRNDAL, C.-E. Calibration as a standard method for treatment of nonresponse. *Journal of official statistics*, Statistics Sweden (SCB), v. 15, n. 2, p. 305, 1999.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. [S.l.]: Routledge, 2019.

MEBANE, W. R.; SEKHON, J. S. Genetic optimization using derivatives: the rgenoud package for r. *Journal of Statistical Software*, v. 42, n. 1, p. 1–26, 2011.

NEYMAN, J. S. On the application of probability theory to agricultural experiments. essay on principles. section 9. (translated and edited by d. m. dabrowska and t. p. speed (1990), 5, 465-480). *Annals of Agricultural Sciences*, v. 10, p. 1–51, 1923.

REID, C. *Neyman*. [S.l.]: Springer Science & Business Media, 1998.

RIDGEWAY, G.; KOVALCHIK, S. A.; GRIFFIN, B. A.; KABETO, M. U. Propensity score analysis with survey weighted data. *Journal of causal inference*, De Gruyter, v. 3, n. 2, p. 237–249, 2015.

ROSENBAUM, P. *Observational Studies*. [S.l.]: Springer, 2002.

ROSENBAUM, P. R. From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association*, Taylor & Francis, v. 79, n. 385, p. 41–48, 1984.

ROSENBAUM, P. R.; RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, Oxford University Press, v. 70, n. 1, p. 41–55, 1983.

ROSENBAUM, P. R.; RUBIN, D. B. The bias due to incomplete matching. *Biometrics*, JSTOR, p. 103–116, 1985.

ROSENBAUM, P. R.; RUBIN, D. B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, Taylor & Francis, v. 39, n. 1, p. 33–38, 1985.

RUBIN, D. B. Matching to remove bias in observational studies. *Biometrics*, JSTOR, p. 159–183, 1973.

RUBIN, D. B. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, JSTOR, p. 185–203, 1973.

RUBIN, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, American Psychological Association, v. 66, n. 5, p. 688, 1974.

RUBIN, D. B. Inference and missing data. *Biometrika*, Oxford University Press, v. 63, n. 3, p. 581–592, 1976.

RUBIN, D. B. Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, Sage Publications Sage CA: Thousand Oaks, CA, v. 2, n. 1, p. 1–26, 1977.

RUBIN, D. B. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, JSTOR, p. 34–58, 1978.

RUBIN, D. B. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, JSTOR, v. 75, n. 371, p. 591–593, 1980.

RUBIN, D. B. Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference*, Elsevier, v. 25, n. 3, p. 279–292, 1990.

SÄRNDAL, C.-E. The calibration approach in survey theory and practice. *Survey methodology*, v. 33, n. 2, p. 99–119, 2007.

SÄRNDAL, C.-E.; LUNDSTRÖM, S. *Estimation in surveys with nonresponse*. [S.l.]: John Wiley & Sons, 2005.

SÄRNDAL, C.-E.; SWENSSON, B.; WRETMAN, J. *Model assisted survey sampling*. [S.l.]: Springer Science & Business Media, 1992.

SEKHON, J. S. Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software, Forthcoming*, 2008.

SEKHON, J. S. Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, Annual Reviews, v. 12, p. 487–508, 2009.

SEKHON, J. S.; MEBANE, W. R. Genetic optimization using derivatives. *Political Analysis*, Cambridge University Press, v. 7, p. 187–210, 1998.

SILVA, P. L. do N. *Calibration Estimation: When and Why, How Much and How*. 2004. Working paper, Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro. Available at: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv66414.pdf>.

SKINNER, C.; WAKEFIELD, J. Introduction to the design and analysis of complex survey data. *Statistical Science*, Institute of Mathematical Statistics, v. 32, n. 2, p. 165–175, 2017.

TILLÉ, Y. Ten years of balanced sampling with the cube method: an appraisal. *Survey methodology*, Statistics Canada, v. 37, n. 2, p. 215–226, 2011.

TILLÉ, Y. *Sampling and estimation from finite populations*. [S.l.]: John Wiley & Sons, 2020.

WILK, M. B.; KEMPTHORNE, O. Fixed, mixed, and random models. *Journal of the American Statistical Association*, Taylor & Francis, v. 50, n. 272, p. 1144–1167, 1955.

WU, C.; SITTER, R. R. A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, Taylor & Francis, v. 96, n. 453, p. 185–193, 2001.

YANG, Z.; QU, T.; LI, X. Rejective sampling, rerandomization, and regression adjustment in survey experiments. *Journal of the American Statistical Association*, Taylor & Francis, p. 1–15, 2021.

APPENDIX A – PROOF OF RESULT 3.2

Proof. We have

$$\hat{Y}_{0_w} = \sum_{k\in\mathscr{S}_0} d_k Y_{0k} + (\sum_{k\mathscr{S}_1} d_k \mathbf{X}_k - \sum_{k\in\mathscr{S}_0} d_k \mathbf{X}_k)^{\mathrm{T}} \hat{\mathbf{B}} \tag{A.1}$$

where

$$\hat{\mathbf{B}} = (\sum_{k\in\mathscr{S}_0} d_k \mathbf{X}_k \mathbf{X}_k^{\mathrm{T}})^{-1} \mathbf{X}_k$$

Taking design expectations on each term of 3.32 and using δ_{0k} as the control assignment indicator (see Section 3.4.3), we have

$$\begin{aligned} \mathrm{E}_p(\sum_{k\in\mathscr{S}_0} d_k Y_{0k}) &= \sum_{k\in U} \delta_{0k} Y_{0k} \\ \mathrm{E}_p(\hat{\mathbf{B}}) = \Big(\sum_{k\in U} \delta_{0k} \mathbf{X}_k \mathbf{X}_k^{\mathrm{T}}\Big)^{-1} \sum_{k\in U} \delta_{0k} \mathbf{X}_k Y_{0k} &= \mathbf{B}_{\mathbf{U_0}} \\ \mathrm{E}_p(\sum_{k\mathscr{S}_0} d_k \mathbf{X}_k) &= \sum_{k\in U} \delta_{0k} \mathbf{X}_k \\ \mathrm{E}_p(\sum_{k\mathscr{S}_1} d_k \mathbf{X}_k) &= \sum_{k\in U} (1-\delta_{0k}) \mathbf{X}_k \end{aligned}$$

Thus,

$$\hat{Y}_{0_w} = \sum_{k\in U} \delta_{0k} Y_{0k} + \sum_{k\in U} (1-\delta_{0k}) \mathbf{X}_k^{\mathrm{T}} \mathbf{B}_{\mathbf{U_0}} - \sum_{k\in U} \delta_{0k} \mathbf{X}_k^{\mathrm{T}} \mathbf{B}_{\mathbf{U_0}} \tag{A.2}$$

Now, subtracting $\sum_{k\in U} Y_{0k} - \sum_{k\in U_0} Y_{0k}$ from A.2, rearranging and using $E_k = Y_{0k} - \mathbf{X}_k^{\mathrm{T}} \mathbf{B}_{\mathbf{U_0}}$,

$$\begin{aligned} \hat{Y}_{0_w} - \Big(\sum_{k\in U} Y_{0k} - \sum_{k\in U_0} Y_{0k}\Big) &= 2\Big(\sum_{k\in U} \delta_{0k} E_k\Big) - \sum_{k\in U} E_k \\ &= 2\Big(\sum_{k\in U} \delta_{0k} E_k\Big) - \Big(\sum_{k\in U} (1-\delta_{0k}) E_k + \sum_{k\in U} \delta_{0k} E_k\Big) \\ &= \sum_{k\in U_0} E_k - \sum_{k\in U_1} E_k \end{aligned}$$

□

APPENDIX B – NEYMAN-RUBIN FRAMEWORK REGULARITY ASSUMPTIONS

Assumption B.1 (Stable Unit Treatment Value). *For each unit in the population, there is only one version of the potential outcome for each treatment level. Moreover, one unit's potential outcomes is not affected by other unit's treatment status.*

Assumption B.2 (Individualistic Treatment Assignment Mechanism). *The Treatment Assignment Mechanism $\lambda(\mathbf{T}|\mathbf{X},\mathbf{Y})$ is individualistic if. That is, for some function $g(\cdot) \in [0,1]$*

$$\theta_k(\mathbf{X},\mathbf{Y}) = g(\mathbf{X}_k, \mathbf{Y_k}) \tag{B.1}$$

and

$$\lambda(\mathbf{T}|\mathbf{X},\mathbf{Y}) = \prod_{k \in U} g(\mathbf{X}_k, \mathbf{Y_k})^{t_k} (1 - g(\mathbf{X}_k, \mathbf{Y_k}))^{(1-t_k)} \tag{B.2}$$

where

$$\theta_k(\mathbf{X},\mathbf{Y}) = \sum_{\mathbf{t} \in \mathcal{T}|(t_k=1)} \lambda(\mathbf{t}|\mathbf{X},\mathbf{Y}) \tag{B.3}$$

is the individual-level assignment probability.

Then, we can rewrite $\theta_k(\mathbf{X},\mathbf{Y}) = \theta_k(\mathbf{X}_k, Y_k)$.

Assumption B.3 (Unconfounded Treatment Assignment Mechanism). *The Treatment Assignment Mechanism is unconfounded. That is, it satisfies*

$$Y_{1k}, Y_{0k} \perp\!\!\!\perp T_k | \mathbf{X}_k \tag{B.4}$$

and we can rewrite $\theta_k(\mathbf{X}_k, Y_k) = \theta_k(\mathbf{X}_k)$.

Assumption B.4 (Probabilistic Treatment Assignment Mechanism). *The Treatment Assignment Mechanism is probabilistic. That is,*

$$0 < \theta_k(\mathbf{X}_k) < 1, \forall k \in U \tag{B.5}$$

APPENDIX C – THEORETICAL RESULTS CONCERNING BALANCING SCORES

Result C.1. *For every unit $k \in U$*

$$\mathbf{X}_k \perp\!\!\!\perp T_k | \theta_k(\mathbf{X}_k)$$

Moreover,

$$\mathbf{X}_k \perp\!\!\!\perp T_k | b(\mathbf{X}_k) \iff \theta_k(\mathbf{X}_k) = f\{b(\mathbf{X}_k)\}$$

for some function $f(\cdot)$.

Proof. see Theorems 1 and 2 in Rosenbaum and Rubin (1983, p. 44).

Result C.2. *For a given treatment assignment mechanism, if*

$$Y_{1k}, Y_{0k} \perp\!\!\!\perp T_k | \mathbf{X}_k, \forall k \in U$$

then

$$Y_{1k}, Y_{0k} \perp\!\!\!\perp T_k | b(\mathbf{X}_k), \forall b(\mathbf{X}_k) | k \in U$$

Proof. see Theorem 3 in Rosenbaum and Rubin (1983, p. 45).

Result C.3. *Assuming unconfoundedness of treatment assignments, we have*

$$\mathrm{E}(Y_1 | b(\mathbf{x}), U_1) - \mathrm{E}(Y_0 | b(\mathbf{x}), U_0) = \mathrm{E}(Y_1 | b(\mathbf{x})) - \mathrm{E}(Y_0 | b(\mathbf{x}))$$

Moreover,

$$\begin{aligned} \mathrm{E}_{b(\mathbf{x})}\Big\{\mathrm{E}(Y_1 | b(\mathbf{x}), U_1) - \mathrm{E}(Y_0 | b(\mathbf{x}), U_0)\Big\} &= \mathrm{E}_{b(\mathbf{x})}\Big\{\mathrm{E}(Y_1 | b(\mathbf{x})) - \mathrm{E}(Y_0 | \mathbf{X}_k)\Big\} \\ &= \mathrm{E}(Y_1) - \mathrm{E}(Y_1) \end{aligned}$$

for every balancing score $b(\mathbf{x})$.

Proof. The first part is straightforward, by the definition of unconfoundedness, which holds by assumption, and by Result C.2. The second part holds by applying iterative expectations. □