UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Ana Kolar

Velikost vzorca pri ocenjevanju vzročnega učinka po metodi nagnjenja

Sample Size Considerations When Using Propensity Score Methods To Estimate Causal Effect

Doktorska disertacija

Ljubljana, 2013

UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Ana Kolar

Mentor: prof. dr. Vasja Vehovar
Somentor: Donald B. Rubin, John L. Loeb Professor of Statistics

Velikost vzorca pri ocenjevanju vzročnega učinka po metodi nagnjenja

Sample Size Considerations When Using Propensity Score Methods To Estimate Causal Effect

Doktorska disertacija

Ljubljana, 2013

To my mothers – Mari and Mimi

Posvečeno mojima materama – Mari in Mimi

Acknowledgments

I would like to sincerely thank my thesis advisors, Vasja Vehovar and Don Rubin for guiding me through the process of obtaining my PhD degree. I would like to express my sincere gratitude to Don Rubin for all the vast knowledge about propensity score methods that he has been sharing with me, for all the valuable discussions we had when working together in different parts of the world, for all the advices and support. At the same time, I would like to sincerely thank Peter Steiner for all the valuable advices, fruitful discussions, support and knowledge-sharing. I also sincerely thank Roman Luštrik for sharing his vast knowledge on programming in R, when setting up the simulation study. Sincere gratitude goes to my family, particularly my endlessly supportive uncle Črtomir Polanc, and to my precious friends, Tamara Gnilšek, Marjana Kos, Mateja Poljšak, Darko Grašič, and Mičo Labohar, for all their love, support, understandings, and for simply being who they are.

Univerza
v Ljubljani Fakulteta
za družbene vede

Kardeljeva ploščad 5
1000 Ljubljana, Slovenija
telefon 01 58 05 120

IZJAVA O AVTORSTVU
doktorske disertacije

Podpisani/-a Ana Kolar, z vpisno številko 7400080, sem avtor/-ica doktorske disertacije z naslovom:
Velikost vzorca pri ocenjevanju vzročnega učinka po metodi nagnjenja / Sample Size Considerations When Using Propensity Score Methods To Estimate Causal Effect.

S svojim podpisom zagotavljam, da:

- je predložena doktorska disertacija izključno rezultat mojega lastnega raziskovalnega dela;
- sem poskrbel/-a, da so dela in mnenja drugih avtorjev oz. avtoric, ki jih uporabljam v predloženem delu, navedena oz. citirana v skladu s fakultetnimi navodili;
- sem poskrbel/-a, da so vsa dela in mnenja drugih avtorjev oz. avtoric navedena v seznamu virov, ki je sestavni element predloženega dela in je zapisan v skladu s fakultetnimi navodili;
- sem pridobil/-a vsa dovoljenja za uporabo avtorskih del, ki so v celoti prenesena v predloženo delo in sem to tudi jasno zapisal/-a v predloženem delu;
- se zavedam, da je plagiatorstvo – predstavljanje tujih del, bodisi v obliki citata bodisi v obliki skoraj dobesednega parafraziranja bodisi v grafični obliki, s katerim so tuje misli oz. ideje predstavljene kot moje lastne – kaznivo po zakonu (Zakon o avtorski in sorodnih pravicah (UL RS, št. 16/07-UPB3, 68/08, 85/10 Skl.US: U-I-191/09-7, Up-916/09-16), prekršek pa podleže tudi ukrepom Fakultete za družbene vede v skladu z njenimi pravili;
- se zavedam posledic, ki jih dokazano plagiatorstvo lahko predstavlja za predloženo delo in za moj status na Fakulteti za družbene vede;
- je elektronska oblika identična s tiskano obliko doktorske disertacije ter soglašam z objavo doktorske disertacije v zbirki »Dela FDV«.

V Ljubljani, dne 13.05.2013 Podpis avtorja/-ice: ________________________

Povzetek

Metode nagnjenja (propensity score methods), katerih temelj je pristop na osnovi načrta zasnove Rubinovega modela vzročnosti (RCV), dandanes prevladujejo na področju vzročnega sklepanja z opazovalnimi podatki. Metode se obsežno uporabljajo na področju ekonomije, medicine, izobraževanja, politologije, psihologije, upravljanja in poslovanja. Cilj pristopa na osnovi načrta zasnove pri vzročnem sklepanju v opazovalnih študijah je popraviti opazovalno (neslučajno) zasnovo na način, da se le-ta približa slučajni zasnovi poskusa. Tako je slučajna zasnova poskusa temelj pristopa na osnovi načrta zasnove.

V disertaciji razširimo uporabo metod nagnjenja tudi na opazovalne študije, kjer nam narava opazovanih podatkov ne dovoljuje zanesljive ocene vzročnih učinkov, zato lahko v teh primerih ocenjujemo le pogojne asociacije. Z aplikacijo na realnih podatkih pokažemo, da je uporaba pristopa na osnovi načrta zasnove (design-based approach) veliko bolj zanesljiv način ocenjevanja pogojnih asociacij (predvsem v primeru majhnih vzorcev), kot uporaba pristopa na osnovi modela (model-based approach) (t.i. regresijska analiza).

Metode nagnjenja sestavljajo: (i) faza zasnove, ki izključuje uporabo podatkov izida in katere namen je uravnotežiti študijsko zasnovo glede na opazovane sospremenljivke (npr., odprava pristranskosti v opazovalnih študijah z namenom imitacije slučajne zasnove poskusa), in; (ii) faza analize, v kateri uporabimo podatke izida, z namenom ocene vzročnih učinkov ali pogojnih odvisnosti, izvedemo dodatna uravnavanja sospremenljik (covariate adjustments) in analizo občutljivosti (sensitivity analysis) ocen vzročnih učinkov.

Razvoj metod nagnjenja je v zadnjih treh desetletjih postregel s smernicami za ocenjevanje vzročnih učinkov z velikimi vzorci opazovalnih podatkov. Vendar pa vprašanje, »kako velik« vzorec se zahteva, ostaja v večjem delu neodgovorjeno. Ker se tako v družboslovju, kot tudi v medicini pogosto srečujemo z majhnimi vzorci, ostaja raziskovanje minimalnih zahtev glede velikosti vzorca z namenom zanesljive ocene vzročnih učinkov zelo pomembno področje. Glede na objavljene raziskave o metodah nagnjenja z majhnimi vzorci, obravnavani vzorci manjši od 100 enot še niso bili dovolj raziskani.

Doktorska disertacija tako proučuje minimalne zahteve glede velikosti vzorca, s katerimi še lahko zanesljivo ocenimo vzročne učinke. S tem namenom smo izvedli serijo simulacijskih študij, ki proučujejo različne velikosti majhnih obravnavanih vzorcev (vzorci manjši kot 100), različna razmerja med skupinama (t.j., med skupino enot, ki niso bile obravnavane – kontrolna skupina in med skupino enot, ki so bile obravnavane – obravnavana skupina), različne mehanizme izbire (selection

mechanism) (t.j., raven začetnih neuravnoteženj v sospremenljivkah med obravnavano in kontrolno skupino), različno število opazovanih sospremenljivk in učinek različnih algoritmov usklajevanja (matching algorithm) (t.j., požrešen (greedy) in optimalen). Hkrati preučujemo tudi vlogo srednje velikih obravnavanih vzorcev (t.j., obravnavani vzorci v velikosti 200 in 500 enot) z namenom primerjave obnašanja majhnih v. srednje velikih vzorcev v metodah nagnjenja.

Dve dodatni simulacijski študiji raziskujeta: (i) če različna korelacijska struktura (šibkejša v primerjavi z močnejšo) med opazovanimi sospremenljivkami in spremenljivko izida vpliva na študijo nagnjenja, in; (ii) če ima različen tip spremenljivke izida (dihotomna v. zvezna) drugačen vpliv na ocene vzročnih učinkov z majhnimi vzorci, kot s srednje velikimi vzorci.

Rezultati simulacijskih študij kažejo, da je uspeh metod nagnjenja (t.j., uspešna odprava pristranskosti) z majhnimi obravnavanimi vzorci, primarno odvisen od zadostne velikosti skupine kontrolnih enot. Kakorkoli, zahtevana velikost vzorca kontrolnih enot je odvisna od: (i) velikosti obravnavane skupine; (ii) števila opazovanih sospremenljivk; in (iii) moči mehanizma izbire, ki je merjen z začetno neuravnoteženostjo opazovanih sospremenljivk med obravnavano in kontrolno skupino.

Ugotovitve simulacijske študije kažejo, da so majhni obravnavani vzorci (tako majhni, kot je $n_t = 8$) enako uspešni pri odpravljanju pristranskosti iz opazovalnih študij, kot srednje veliki vzorci (t.j., n_t od 200 ali 500), če je razmerje med skupinama le dovolj veliko in je mehanizem izbire strogo pogojno neodvisen (strongly ignorable). Seveda pa so standardne napake ocen vzročnih učinkov z majhnimi vzorci po pričakovanju veliko večje, kot v primeru srednje velikih vzorcev. Zato so ocene vzročnih učinkov z majhnimi vzorci veliko manj natančne, kot ocene vzročnih učinkov s srednje velikimi vzorci.

Kakorkoli, velikost standardnih napak ocen vzročnih učinkov z majhnimi vzorci ni odvisna samo od velikosti celotnega vzorca, ampak tudi od števila opazovanih sospremenljivk in algoritma usklajevanja, ki je uporabljen pri odpravi pristranskosti v študiji nagnjenja. Po drugi strani pa število opazovanih sospremenljivk (t.j., $p = 10, 15, 20, 30$) in algoritem usklajevanja nimata vpliva na velikost standardnih napak ocen vzročnih učinkov s srednje velikimi vzorci.

Uporaba različnih algoritmov usklajevanja nima vpliva na minimalno zahtevano razmerje med skupinama v opazovalnih študijah s srednje velikimi vzorci, vendar pa ima vpliv v opazovalnih študijah z majhnimi vzorci (t.j., optimalni algoritem usklajevanja je v povprečju boljši – za odpravo pristranskosti v povprečju zahteva manjša razmerja med skupinama). Hkrati je študija nagnjenja v primeru uporabe optimalnega algoritma usklajevanja z majhnimi vzorci rezultirala v povprečju v malenkost manjših standardnih napakah vzorčnih učinkov, kar pa se ni izkazalo v primeru srednje velikih vzorcev.

Aplikacija rezultatov simulacijske študije na realnih opazovalnih podatkih (Lalonde 1986) dodatno potrjuje naše zaključke glede delovanja metod nagnjenja z majhnimi vzorci. Medtem, ko druga aplikacijska študija z realnimi opazovalnimi podatki (Luthar, et al. 2011) podaja primer študije, kjer raziskovalna vprašanja morda so vzročna, vendar pa zaradi narave opazovalnih podatkov ni mogoče zanesljivo oceniti vzorčnih učinkov. Tako lahko, glede na posebej zanimive sospremenljivke, ocenimo le pogojne asociacije med indikator spremenljivko in spremenljivko izida. Ta aplikacija je še posebej zanimiva, ker: (i) pokaže kako so lahko metode nagnjenja uporabljene z namenom ocenjevanja pogojnih asociacij, in; (ii) pokaže, zakaj je pri ocenjevanju pogojnih asociacij pristop na osnovi načrta zasnove veliko bolj zanesljiv v primerjavi s pristopom na osnovi modela (e.g., regresijska analiza) – še posebej, ko imamo opravka z majhnimi vzorci.

Ključne besede: metode nagnjenja, vzročno sklepanje, Rubinov model vzorčnosti, opazovalne študije, pogojne asociacije, majhni in srednje veliki vzorci

Abstract

Propensity score methods, whose foundation is the design-based approach of the Rubin Causal Model, prevail today in the causal inference field for observational studies. The methods are largely applied in the fields of economics, medicine, education, political science, psychology, management and business. Such a design-based approach for causal inference from observational studies aims to mend an observational (non-randomised) design so as to approximate a randomised experiment design and, in this sense, it strictly follows the rationale of experimental designs.

We extend the use of propensity score methods also to observational designs where the nature of observed data does not allow us to estimate reliably causal effects; thus, we can only estimate conditional associations. We show, with an application, that the use of the design-based approach, particularly for small sample studies, is more trustworthy than model-based approaches (i.e., regression analyses), when estimating conditional associations.

Propensity score methods consist of: (i) the design phase, which is outcome free and aims to balance a study design with respect to observed covariates (e.g. removes selection bias in observational designs in order to mimic randomised experiment designs when estimating causal effects), and; (ii) the analysis phase, which uses the outcome data in order to estimate causal effects or conditional associations, to perform additional covariate adjustments, and to carry out sensitivity analysis for obtained causal effect estimates.

The development of propensity score methods over the past three decades has resulted in guidance on estimating causal effects from large observational data sets. However, the question of "how large" a data set should be, remains mostly unanswered. Because the social sciences and medical research often face relatively small samples, the investigation of minimum sample size requirements for reliably estimating causal effects from observational designs, remains a highly important topic. Based on the published research on propensity score methods with small samples, treated samples that consists of less than 100 units have not yet been sufficiently investigated.

The thesis thus investigates minimum sample size requirements that enable a reliable estimation of causal effects. We carry out a series of simulation studies examining a variety of small treated sample sizes (samples smaller than 100), group ratios (i.e., ratio between the samples of control and treated units), different selection mechanisms (i.e., the level of initial covariate imbalances between treated and control groups), different numbers of observed covariates, and the performance of different matching algorithms (i.e., greedy and optimal). We also examine performance of moderately large treated samples (i.e., samples consisting of 200 and 500 units) in order to compare the behaviour of small versus moderately large treated samples with propensity score methods.

Two simulation study extensions are performed to investigate: (i) whether different correlation structures (weaker versus stronger) between the observed covariates and the outcome variable has an impact on propensity score studies; and (ii) whether different classes of the outcome variable (binary versus continuous) affect treatment effect estimates with small samples differently from larger samples.

The simulation results show that the success of propensity score study (i.e., successful removal of selection bias) with small treated samples primarily depends on a sufficiently large pool of control units. However, the required size of a control group depends on: (i) the size of a treated group; (ii) the number of observed covariates; and (iii) the strength of the selection mechanism, measured by the initial imbalances in observed covariates between the treated and control groups.

The simulation study's findings demonstrate that small treated samples (as small as $n_t = 8$) perform as good as moderately large treated samples (i.e., n_t of 200 or 500) at removing selection bias from observational study designs as long as the group ratio is sufficiently large, and the treatment assignment mechanism is strongly ignorable. Of course, the treatment effect estimates derived from small treated samples are much less precise in comparison to those obtained by moderately large treated samples, due to much larger treatment effect standard errors in cases of small treated samples.

However, the size of treatment effect standard errors in small sample studies does not depend only on the overall sample size but also on the number of observed covariates and the matching algorithm used in the propensity score study. On the other hand, the number of observed covariates (i.e., $p = 10, 15, 20, 30$) and the matching algorithm used has a negligible effect on the size of treatment effect standard errors in moderately large sample studies.

The use of different matching algorithms has a negligible effect on the minimum required group ratio for removing selection bias from observational designs with moderately large samples, whereas a tiny effect with small samples (i.e., optimal matching algorithm on average performs slightly better – on average requiring smaller group ratios for removing selection bias). At the same time, the propensity score study with optimal matching algorithm for small samples resulted on average in slightly smaller treatment effect standard errors, which was not the case for moderately large samples.

An application of the simulation study results with real observational data (LaLonde 1986) additionally supports our conclusions regarding the performance of propensity score methods with small samples. The second application of real observational data (Luthar, et al. 2011) provides an example of a study where research questions might be causal but due to the nature of observed data, we are not able to estimate, reliably, causal effects. Hence, only conditional associations between an indicator variable and an outcome variable, given a set of substantively interesting covariates, can be estimated. This application is particularly interesting

because: (i) it shows the use of propensity score methods to estimate conditional associations, and; (ii) it shows why design-based approaches are much more trustworthy, in comparison to the model-based approaches (e.g., regression analysis), when estimating conditional associations, particularly in small sample studies.

Keywords: propensity score method, causal inference, Rubin Causal model, observational studies, conditional associations, small and moderately large samples

Contents

Chapter 1

Introduction

During the last decades, the awareness that correlation does not imply causation has been stimulating some major methodological developments with causal inference for observational study designs. The statistical methods developed for causal inference with randomised experiments have been known not to be appropriate for the use with observational designs. Thus, it was long advised that within non-randomised settings we can only provide descriptions of observed associations (Cochran 1965) without being able to talk definitively about any causal quantities and hence to draw, trustworthy, causal conclusions.

Under the definition of observational designs we classify non-experimental designs (i.e., surveys), study designs where treatment conditions are not randomly assigned to units (i.e., quasi-experimental designs (Shadish, et al. 2002)) and study designs where complete randomisation of treatment conditions fails due to missing data or noncompliance of the assigned treatment unit, i.e., broken randomised designs (Barnard, et al. 2003). The key difference between observational designs and randomised experiments is hence in the selection procedure, i.e., how treatments are assigned to units – what is the process that dictates to which units a treatment is applied and to which it is not applied.

In randomised experiments, the process of assigning treatments to units is controlled by the experimenter. In this sense the experimenter tries to assure that units that receive different treatments[1] are comparable (i.e., they share the same characteristics – in expectation, their covariate distributions are identical). This comparability is ensured by randomly assigning different treatments. In this way

[1] A unit to which a treatment is applied is denoted as a treated unit belonging to a treated group, whereas a unit to which treatment is not applied is denoted as a control unit and hence belongs to the control group.

observed, but also unobserved covariates, tend to be comparable between groups that receive different treatments (i.e., distributions of observed and unobserved covariates are on average the same for the treated and control groups). Hence, a study design is said to be statistically balanced with regard to covariate distributions between the two groups, whereas some possible imbalances in covariate distributions are only due to a chance rather than systematic selection procedures that potentially induce selection bias (Rosenbaum 2002, 21).

In contrast, the assignment process in observational designs is only partially controlled by the investigator or not controlled at all. Hence, randomisation (i.e., randomly assigning different treatments) is not feasible, which typically results in incomparability of the units that receive different treatments (i.e., the distributions of observed covariates between the treated and control group are on average different - selection bias). Therefore, estimation of causal effects in such study designs requires a special approach – the selection bias has to be removed from a study design in order to estimate, unbiasedly, casual effects. There are different approaches for removing selection bias from a study design (i.e., design-based and model-based approaches (Section 1.3) and the development of these approaches has been motivated by the fact that most study designs are indeed observational because randomised experiments are not only expensive but often not even feasible due to ethical norms.

To illustrate the aforesaid with an example; imagine that you would like to assess how harmful smoking is to health (i.e., what is the effect of smoking on health). Could you take a population of non-smokers and simply randomise them into two groups, i.e., a treatment group where you would require participants to start smoking and smoke for an extended period of time (e.g., 20 years or so) and a control group where people would not be allowed to smoke for the same period of time? It should surely be possible to create a control group, but it would be much harder to create a treatment group. Imagine if some people's health in the treatment group would, after some time, be seriously threatened due to smoking, and their doctors would advise them to stop smoking. Could you possibly demand them not to follow their doctor's advice because otherwise your experiment would confront a noncompliance issue? Could you request them to keep on smoking so that you can investigate when and how they are going to die (i.e., would they develop a lung cancer or some other smoking related disease)? No! That would be highly unethical, and this is the reason why in so many situations randomised experiments are not realistic.

Examples like this can be found in many fields of social (e.g., education, economics, and politics) and medical (e.g., epidemiology, immunology, pharmacology) sciences. Thus, being able to investigate causal effects in observational designs highly contributes to the well-being of society.

1.1 Causality

The idea of causality is very old and goes back to philosophers such as Plato, Aristotle, Hume, Mill and some others. According to Hulswit (2002) it was Plato who first formulated the principle of causality by stating: "everything that becomes or changes must do so owing to some cause; for nothing can come to be without a cause". These early philosophers thus look at the causality by trying to find the cause of an effect that is seen.

In contrast, the statistical community looks at causality from a different angle: units are manipulated by a known intervention (i.e., the cause is known) and we try to estimate, unbiasedly, an effect caused by such an intervention. Thus, a cause is seen as an active intervention that is applied to some units at a particular point of time in order to investigate how differently these units behave in comparison to the behaviour of the units from which the intervention is withheld.

Having that said, causal effects that we aim to estimate statistically are the effects caused by an intervention being posed to some of the units in a population. However, the units to which an intervention is applied have to be comparable to the units to which the intervention is not applied. The way units, to which the intervention is applied, are affected is considered as an effect of the intervention and this effect is denoted as a treatment effect (i.e., an effect caused by applying specific treatment – intervention – to some of the units) that is, a causal effect. Both terms will be interchangeably used.

1.1.1. Observational designs versus Randomised Experiments

As mentioned previously, the main difference between observational designs and randomised experiments for estimating causal effects is in a selection procedure (i.e., how/based on which criteria a treatment status is assigned to units). Observational designs are characterised by self- or third-person selection procedures whereas randomised experiments are characterised by randomised selection procedures. The consequence of the non-randomised selection procedure is often selection bias (i.e., observed covariates of a treated and a control group are not effectively balanced). The bigger the imbalances between observed covariates of the treated and control groups, the higher the selection bias. A study design which is prone to different levels of selection bias, such as an observational design, can also be denoted as an unbalanced study design – a study design where treated and control groups are not comparable (i.e., their covariate distributions do not perfectly overlap – units in the groups do not share common characteristics).

In randomised experiments, the selection procedure, which defines the assignment mechanism, is known (i.e., the probability structure for units to be selected in either group is controlled by the experimenter). Consequently, the assignment mechanism is unconfounded (i.e., there is no dependence between the assignment mechanism and the potential outcomes - units assigned to either treated or control group are independent of the unobserved potential outcomes (Rubin 1974; Rubin 1978; Rubin 1990). Furthermore, the assignment mechanism is also probabilistic (i.e., each unit has a positive probability to be assigned to a treated or to a control group). An unconfounded and a probabilistic assignment mechanism is said to be strongly ignorable (Rosenbaum and Rubin 1983a) and, thus, causal effects can be in general estimated without bias.

On the other hand, non-randomised assignment mechanisms are generally unknown or only partly known. Thus, in order to obtain unbiased estimates of treatment effects, where by "unbiased" we mean unbiased or approximately unbiased, we are required to posit an assignment mechanism or redesign an existing non-randomised assignment mechanism[2] in a way to mimic a randomised assignment mechanism. However, this can be done only if all the relevant covariates are observed[3] because the treatment status of units assigned either to the treated or control group is then independent of the unobserved potential outcomes. If each unit also has a positive probability of being treated or untreated (i.e., in the control condition), the assignment mechanism also becomes ignorable. Because it is impossible to test whether all the relevant covariates are observed, almost all estimates of treatment effects in observational designs are based on a strong ignorability assumption (i.e., we assume that all the relevant covariates are observed). Hence, sensitivity analysis of obtained treatment effect estimates should always be performed to assess how conclusions would change when the strong ignorability assumption is not met.

By posing or redesigning a non-randomised assignment mechanism to approximate a randomised one under unconfoundedness we are balancing an observational study design (i.e., removing selection bias), hence, unbiased estimates of treatment effects can be obtained. Once a balanced design is obtained, we may proceed with methods used for estimating causal effects of randomised experiments because our observational design, due to the redesigned assignment mechanism, closely approximates a randomised experiment design. However, redesigning non-randomised assignment mechanisms to well-approximate randomised assignment

[2] A well modelled assignment mechanism is a crucial part for estimating unbiased causal effects. The fundamental tool for redesigning a non-randomised assignment mechanism to approximate closely a randomised assignment mechanism is a propenstiy score. The redesign process is presented in Chapter 2.

[3] Selection of relavant covariates, often called baseline covariates, is described in Section 2.3.2.

mechanisms might sometimes also affect the structure of our sample (i.e., the sample after the redesign phase might not correctly represent the original population of interest). Thus, we should be cautious about causal claims we are making. The process of redesigning a non-randomised assignment mechanism to approximate an assignment mechanism of randomised experiments is presented in Chapter 2.

1.1.2. Causal Inference Notation

Causal inference is the process of drawing a conclusion about an effect that is caused as a consequence of some intervention being applied. Before proceeding with the causal inference, causal questions should always be clearly and precisely defined. The reason for doing it so is twofold. First, a precisely defined causal question will give us an idea whether the nature of collected data, or the data that we attempt to collect, will enable us to estimate, unbiasedly, causal effects and provide an answer to our causal question. When the intervention cannot be precisely defined (as described in Section 2.1), we should not proceed with the estimation of causal effects but rather accept the fact that causal effects cannot be estimated. Hence, we can only estimate associations conditional on the substantively interesting covariates[4]. Second, based on the defined causal question an appropriate causal quantity is selected for estimation.

Two common causal quantities are: (i) average treatment effect - ATE, and; (ii) average treatment effect on the treated - ATT. The main difference between those two causal quantities is the population for which we aim to derive causal claims. ATE is the average treatment effect for the overall population whereas ATT is the average treatment effect for the subpopulation of those units to which the treatment is applied. The choice of the estimated causal quantity depends on the nature of the field in which we are aiming to derive causal claims and should thus reflect our causal question.

[4] Which variables comprise the substantivelly interesting covariates is up to the investigator to define and defend as interesting, and surely depends upon the context.

A typical example for estimating ATT is an effect of smoking on health because we are interested only in estimating the effect for the subpopulation of people who smoke. It would not be ethical to force non-smokers to smoke; thus, in cases like this, estimating ATE would not be a reasonable choice because such estimation would heavily rely on extrapolating our estimates to the subpopulation of non-smokers.

The notation that we are using for estimating causal effects from observational designs is the following:

n_t	sample size of the treated group
n_c	sample size of the control group
$Y(1)$	the potential outcome of the treated group
$Y(0)$	the potential outcome of the control group
X	observed covariates, also called baseline covariates
W	treatment indicator indicating whether a unit received the treatment ($W = 1$) or whether thee treatment was withheld from a unit ($W = 0$)
τ	treatment effect

1.2 Developments of Causal Inference for Observational designs

The development of formal causal inference methods for observational designs started seriously being considered only in the late 1970's and the early 1980's with the work of Rubin (1974; 1977; 1978; 1980), Rosenbaum and Rubin (1983a), Holland and Rubin (1988), Angrist, Imbens and Rubin (1996) and Rosenbaum (2002). The foundation of their work is based on the randomised experiment framework and it is, thus, a continuation of ideas of the work of Neyman (1923), Fisher (1925), Kempthorne (1952), Cochran and Cox (1957) and Cox (1958).

The foundation follows Neyman's (1923) notation of potential outcomes that he developed in the field of experiments where the causal inference aims to provide an answer to the question: what would the outcome of the treated group have been if it had not been treated and vice versa. The causal effect is then defined as the difference between both hypothetical obtained outcomes:

$$\tau = Y(1) - Y(0).$$

In 1974 Rubin extended Neyman's notation of potential outcomes also to observational designs in his paper on "Estimating causal effects of treatments in randomized and non-randomized studies". Further on Rubin (1975; 1978) also formally incorporated the information of an assignment mechanism in the potential outcomes notation. This extension of the potential outcomes notation later became known as the potential outcomes framework (Morgan and Winship 2007) – the foundation for causal inference in observational study designs.

Due to Rubin's large contribution, this framework is often referred to as Rubin's model for causal inference (Holland 1986; Rubin and Imbens 2008) or simply the Rubin Causal Model (RCM) which is also the foundation of propensity score methods – the most widely applied class of methods for estimating causal effects in observational designs in the last decades.

1.3 Role of Propensity Score Methods in Causal Inference

We distinguish between two common approaches for causal inference in observational designs: (i) the standard model-based approach, sometimes based on ordinary-least-squares; and (ii) the design-based approaches based on Neyman (1923), Fisher (1925) or Rubin Causal Model - the foundation of propensity score methods. The main difference between the approaches is in how they remove selection bias from an observational design.

The model-based approach (i.e., ordinary-least-squares or structural-equation-modelling) requires the outcome variable, Y, in the process of removing selection bias from a design. Hence, the removal of selection bias and the estimation of treatment effects is done simultaneously by modelling the outcome variable (i.e., regressing the outcome variable, Y, on all the observed covariates, X, and on the treatment indicator, W). This approach thus relies on a very strong assumption: The assumption of a correctly specified outcome model through which selection bias is removed.

The model-based approach in observational designs is problematic from three perspectives: (i) in cases of severely imbalanced designs, such an approach heavily relies on extrapolation (King and Zeng 2007; Ho, et al. 2007); (ii) in cases of a misspecified outcome model, causal estimates are generally biased (Dehejia and Wahba 1999; Dehejia and Wahba 2002) but because it is impossible to test whether the model is correctly specified, we essentially must assume that it is correctly specified based on substantial knowledge we have about the field in which we are estimating causal quantities; (iii) because the removal of selection bias and estimation of treatment effects occur simultaneously in this model-based approach (due to modelling the outcome variable), chances exist that the outcome model might be manipulated in a direction to produce estimates of causal quantities the researcher would like to obtain; hence, such results are less objective and less trustworthy.

However, using the model-based approach is not wrong as long as we know how to specify our model correctly or at least objectively. Because we never know the correct model and because it is impossible to assess how objectively we specified it, the model-based approach may not be very trustworthy within observational data settings. Hence, we do not recommend it to be used with observational designs, particularly in cases of largely unbalanced designs, unless it is implemented very cautiously.

On the other hand, in the design-based approach of the Rubin Causal Model (RCM), the selection bias is removed without the outcome variable in sight (i.e., the outcome variable should be literally removed from the data set until the study design is fixed – until the selection bias, due to observed covariates, is removed) (Rubin 2007). In this sense, this design-based approach of the RCM strictly follows the rational of experimental designs (i.e., when designing a randomised experiment we do not have the outcome data, Y, available because we measure Y only after the study design is fixed) and this safeguards against possible manipulations of treatment effect estimates in a direction a researcher might want to see.

Although the design-based approach does not model the outcome variable in order to remove selection bias, the approach does consist of some modelling when using propensity score methods[5]. We are required to model the assignment mechanism (described in Chapter 2) in order to mend an observational assignment mechanism to mimic an assignment mechanism of a randomised experiment design. Such modelling requires information on observed covariates, X, and the treatment assignment, W, and relies on the strong ignorability assumption (i.e., all the relevant covariates are observed/measured). The modelling in the design-based approach is thus used solely in the design phase (i.e., the process of redesigning or posing the unknown or partly known assignment mechanism of an observational design in a way to well approximate a randomised assignment mechanism). However, model misspecifications in the design phase will not bias causal effect

[5] There are methods which follow the design-based framework and do not require any modelling (e.g., multivariate matching).

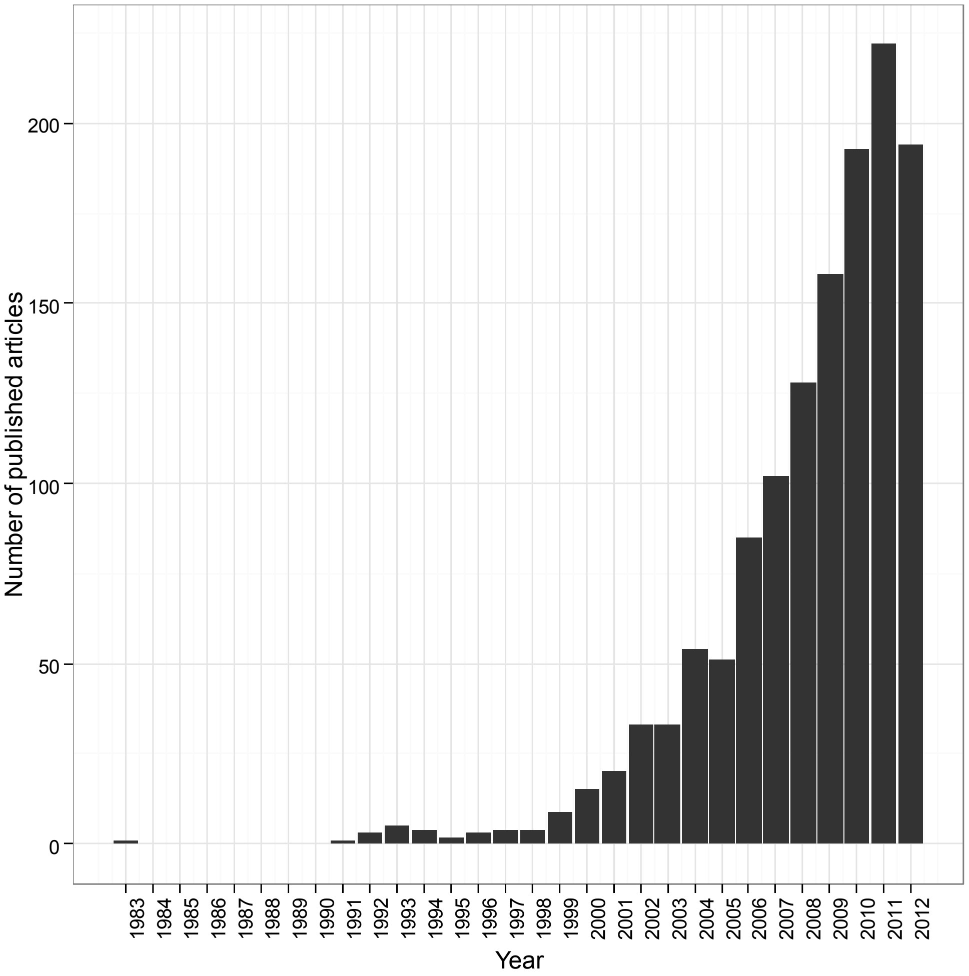
estimates for most of the covariate adjustment methods used within the framework of the propensity score methods (Drake 1993; Waernbaum 2010).

Furthermore, according to Dehejia and Wahba (1999; 2002) and Zhao (2004), a misspecification of the outcome model when using a model-based approach will result in much more biased causal effect estimates than those obtained by using the design-based approach where the model for the assignment mechanism is misspecified. Based on these findings, the design-based approach is not more appropriate only because it follows the rational of randomised experiments, but it can also be seen as more trustworthy approach for investigating causal effects in observational design. Thus, it is not surprising to see that propensity score methods, whose foundation is the design-based approach of the Rubin Causal Model, prevail today in the causal inference field of observational studies.

Thus, with developments of propensity score methods for estimating causal effects from observational designs, the loud mantra “correlation does not reveal causation” slowly became better heard. In the last decade, more and more published articles written by statisticians are greatly encouraging researchers to unglue from various correlation/regression methods and to start applying the methods founded on the design-based approach, such as propensity score methods, when estimating causal effects from observational designs.

More and more researchers are today indeed using propensity score methods when estimating causal effects within observational design. Figure 1.1 shows an exponential growth of the number of published articles in the Web of Science (Thomson Corporation 2012) with the keywords: “propensity score”, “causal effect” and/or “treatment effect”. Altogether, since 1983 until November 2012, there are 1,324 published articles which include the aforementioned keywords.

Figure 1.1: Number of published articles with keywords: “propensity score”, “causal effect” and/or “treatment effect”



Source: Web of Science ®, Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC.

Furthermore, Table 1.1 presents fields of science, according to the Web of Science Categories, where most of the articles are published. As we can see, propensity score methods are largely applied in the fields of economics, public environmental occupational health, cardiology, social and health science, pharmacology, psychiatry, medicine in general, education, political science, psychology, management and business.

Table 1.1: Number of published articles with keywords: “propensity score”, “causal effect” and “treatment effect” by the field – Web of Science Categories

Field: Web of Science Categories	Record Count	% of 1324
ECONOMICS	267	20,17%
STATISTICS PROBABILITY	172	12,99%
PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH	137	10,35%
CARDIAC CARDIOVASCULAR SYSTEMS	117	8,84%
SOCIAL SCIENCES MATHEMATICAL METHODS	96	7,25%
HEALTH CARE SCIENCES SERVICES	93	7,02%
MATHEMATICAL COMPUTATIONAL BIOLOGY	85	6,42%
PHARMACOLOGY PHARMACY	84	6,34%
HEALTH POLICY SERVICES	79	5,97%
PSYCHIATRY	65	4,91%
MEDICINE GENERAL INTERNAL	59	4,46%
MEDICAL INFORMATICS	57	4,31%
MEDICINE RESEARCH EXPERIMENTAL	55	4,15%
ONCOLOGY	53	4,00%
SURGERY	48	3,63%
RESPIRATORY SYSTEM	45	3,40%
MATHEMATICS INTERDISCIPLINARY APPLICATIONS	43	3,25%
EDUCATION EDUCATIONAL RESEARCH	37	2,80%
CLINICAL NEUROLOGY	34	2,57%
UROLOGY NEPHROLOGY	29	2,19%
SOCIOLOGY	27	2,04%
PERIPHERAL VASCULAR DISEASE	26	1,96%
BUSINESS FINANCE	24	1,81%
POLITICAL SCIENCE	24	1,81%
CRITICAL CARE MEDICINE	22	1,66%
AGRICULTURAL ECONOMICS POLICY	20	1,51%
BIOLOGY	20	1,51%
PSYCHOLOGY CLINICAL	20	1,51%
SOCIAL SCIENCES INTERDISCIPLINARY	18	1,36%
ENVIRONMENTAL STUDIES	17	1,28%
MANAGEMENT	17	1,28%
NEUROSCIENCES	17	1,28%
PLANNING DEVELOPMENT	17	1,28%
PSYCHOLOGY MULTIDISCIPLINARY	17	1,28%
PSYCHOLOGY DEVELOPMENTAL	14	1,06%

Source: Web of Science ®, Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC.

Nevertheless, even though propensity score methods were primarily developed for estimating causal effects of observational designs, we extend their framework also to studies when the nature of collected data does not allow us to estimate causal effects (as briefly explained in Section 1.1.3 and with a more detailed description in Section 2.1); hence, only associations between variables conditional on the substantively interesting covariates can be estimated. The second real data application in Chapter 6 reveals why the use of design-based approaches, also when estimating conditional associations, is more trustworthy than the use of model-based approaches, particularly in studies consisting of small samples.

1.4 Thesis Objectives

Propensity score methods for estimating causal effects of observational designs have been known to work well with large samples, and according to Rubin (1997) the methods in general perform better with large samples. However, Rubin does not provide any particular insight on "how large" samples should be.

Because the social sciences and medical studies often face relatively small samples (e.g., the number of students in classrooms, number of schools, or number of patients with a rare disease), an investigation of minimum sample size requirements for estimating unbiased causal effects from observational designs remains a highly important topic.

Based on the published research on propensity score methods, treated samples that consists of less than 100 units have not yet been sufficiently investigated. Thus, this thesis aims to illuminate the role of sample size (i.e., small versus moderately large samples) when estimating causal effects of observational designs when using propensity score methods.

This thesis predominately focuses on examining how well propensity score methods perform with small samples in comparison to moderately large samples. The investigation is based on findings of the limited past research regarding small samples in propensity score methods and on an extensive simulation study which examines performance of the methods when applied to data sets with different sample sizes.

Furthermore, the simulation study investigates the influence that the number of observed covariates has, in combination with different sample sizes and different levels of initial imbalances, on the performance of propensity score methods. The simulation study also examines the performance of two of the most widely used matching algorithms (i.e., greedy versus optimal), different classes of an outcome variable (i.e., continuous versus binary) and different correlation structure between the outcome variable and covariates (i.e., weaker versus stronger).

In addition, the simulation study investigates the performance of propensity score methods when implemented with true versus estimated propensity scores to study the behaviour of the methods with different sample sizes in perfect (i.e., theoretical but unrealistic) versus real world scenarios when true propensity scores are unknown.

The findings of the simulation study results, with regard to small sample sizes, are furthermore applied to real observational data (Chapter 6 – Real Data Set 1), for which causal effect estimate of a randomised experiment exists, in order to evaluate how reliable our simulation results are for practise.

Apart from investigating small sample properties, in propensity score methods for estimating causal effects of observational designs this thesis offers two other scientific contributions. The first contribution is in developing a precise definition of propensity score methods for estimating causal effects that will help to clarify confusing literature based on which, the methods have often been misused.

With the second contribution we extend the definition of the use of propensity score methods also to observational studies where the nature of observed data does not allow estimation of causal effects; thus, only associations conditional on the substantively interesting covariates can be estimated. We provide an application of such a study (Chapter 6 – Real Data Set 2) and show that the use of propensity score methods, in comparison to regression methods, when estimating conditional associations appears to be a more trustworthy approach, particularly when dealing with small samples.

1.5 Plan of Thesis

Chapter 2 provides a definition of propensity score methods, introduces estimation of conditional association within the methods, presents the foundation of the methods – the Rubin Causal Model, describes design and analytic phase of propensity score methods and lists available software for estimating causal effects or conditional associations with propensity score methods. Chapter 3 studies the role of sample size for causal inference with propensity score methods by reviewing past research on small samples within propensity score methods, and by providing theoretical and past research findings to address which factors should be examined in our simulation study. Chapter 4 presents our simulation study design, the investigated factors and procedures for the analysis of our simulated data. Chapter 5 presents results. Chapter 6 presents two applications. The first application evaluates how reliable are our simulation results of small samples for practise. The second application provides an example of real data where, due to the nature of data, causal effects cannot be estimated. Thus, the framework of propensity score methods is used for estimating conditional associations. Chapter 7 concludes with a discussion of our results and recommendations for future research.

Chapter 2

Propensity Score Methods

This Chapter provides the definition of propensity score methods, presents the foundation of the methods, and describes tools to be used within the framework of propensity score methods for: causal inference in observational designs and conditional associations from observed data. We complete this chapter by listing available software to be used with propensity score methods.

2.1 Definition and Usage

Propensity score methods are founded on the design-based approach and their implementation consists of two important parts: (i) **design phase** where we balance a study design with respect to observed covariates. The design phase is “outcome free” (i.e., the outcome of our interest is out of sight) and consists of balancing tools and balance assessment tools, and; (ii) **analysis phase** where we use the outcome data to estimate desired quantities, perform additional statistical adjustments and sensitivity analysis.

Although propensity score methods were primarily developed for estimating causal effects from observational designs, we extend the use of their framework also to observational studies which hope to deal with causal research questions, but the nature of their data does not allow us to estimate causal effects. Hence, the framework of propensity score methods can be used for estimating associations conditional on the substantively interesting covariates. To illustrate such a nature of observed data, we provide first a definition of what is causal, and based on this definition, we derive the definition for conditional association.

When estimating the causal effect of treatment versus control, we must be able to define (1) an intervention that could have been applied to all "treated" units that would convert them to "control" units (e.g., instead of medical pill, we would give them placebo pill), and (2) a similar intervention that could have been applied to all "control" units that would convert them into "treated" units (e.g., instead of placebo pill we would give a medical pill). All such real or hypothetical versions of (1) and (2) must lead to the same potential outcome of each "treated" unit, $Y(1)$, and of each "control" unit, $Y(0)$, in order for the following two assumptions to be satisfied: (i) treatment applied to one unit does not affect the outcome of the other units – no interference between units (Cox, The Planning of Experiments 1958); and (ii) there is only one type of treatment or control available for each unit[6]. All measurements that are made, or at least determined, before either of interventions (i.e., (1) or (2)) is "assigned" to each unit, are baseline covariates, and all measurements that are determined after the intervention is "assigned" is an outcome variable.

When the nature of our observed data does not allow us to formulate convincingly the intervention as described above, we cannot estimate causal effects of "treatment" versus "control". Thus, we should aim for estimating conditional associations between a binary variable Z and another variable Y given a fixed value of the conditioned variables X, which is a vector of covariates selected by the investigator as substantively interesting. An example of such observed data is an estimation of the effect of minority status on students' educational attainment – whether they attend college. We have two groups of students (e.g., white and black) and we cannot convert black students into white or vice versa. Thus, only conditional associations can be estimated.

[6] The above described assumptions constitute in the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980; Rubin 1990) (Section 2.2).

The main aim of the design phase when estimating causal effects of observational designs is to remove successfully selection bias from a study design. Study designs where selection bias is present are called unbalanced designs due to covariate imbalances between groups that receive different treatments (e.g., a group of units to which treatment is applied ($W = 1$) – a treated group, and a group of units to which treatment is not applied ($W = 0$) – a control group). The level of selection bias can thus be viewed as a level of covariate imbalance between treated and control group.

On the other hand, the main aim of the design phase when estimating conditional associations from observed data is to effectively control for the substantively interesting covariates without using the outcome data (i.e., the outcome of our interest, Y). Conditioning on X means finding sets of units with $Z = 1$ and sets of units with $Z = 0$ with the identical values of X or sets of $Z = 1$ units with the same distribution of X values as a set of $Z = 0$ units. The definition can be loosened further by adding "under explicitly stated assumptions" such as all we care about is that the two sets of units should have the same average X values for each variable in Z. The more the distributions of X in the $Z = 1$ units and in the $Z = 0$ units are similar, without the need for any explicitly stated assumptions, the more X has been successfully controlled in the comparison. In this sense, we are balancing our study design based on X but because the selection procedure did not take place in such a study, the imbalances due to conditioning on X should not be viewed as due to selection bias.

For the sake of clarity, we present propensity score methods with the terminology and notation used when estimating causal effects. The binary variable Z when estimating conditional associations plays a similar role as the treatment indicator W when estimating causal effects, in the sense that both Z and W are indicators specifying to which group a unit belongs. Although, the W indicates to which unit a treatment is applied and to which it is not, the Z indicates which units belong to arbitrarily created group 1 and which to arbitrarily created group 2, where the names of those groups can be of any kind (e.g., black and white group when comparing populations of black and white people). At the same time, we are using the term "removal of selection bias" when presenting propensity score methods, although such a term should never be used when estimating conditional associations because there is no selection process in the observed data from which conditional associations are estimated. The term "removal of selection bias" can thus be viewed as removal of covariate imbalances between group 1 and group 2 when estimating conditional associations.

2.2 Rubin Causal Model and Potential Outcomes

The formal use of the potential outcomes approach was introduced by Neyman (1923), but the notation was used only in the context of randomised experiments. Half a century later Rubin (1974) extended Neyman's potential outcome approach to non-randomised studies and, with this approach, defined causal effect at the unit level for randomised and non-randomised studies. Such a causal effect definition does not include information about an assignment mechanism (i.e., how units are assigned to treatment or control group – whether a unit selection is random or non-random). Thus, there is no difference in causal effect definition between randomised and non-randomised studies.

Incorporating the information of an assignment mechanism in the potential outcomes approach makes the estimation of causal effects between randomised and non-randomised designs distinctive. Such incorporation was done by Rubin (1975; 1978) and ever since the assignment mechanism has been formulated in general mathematical terms using the potential outcomes framework. This was a large contribution to the development of causal inference within observational study designs. Due to this contribution and Rubin's other work related to the causal inference within observational study designs, the potential outcomes framework is often referred to as Rubin's model for causal inference (Holland 1986) or simply the Rubin Causal Model.

The Rubin Causal Model consists of two essential parts. Part one addresses causal effects by defining ***units***, ***treatments*** and ***potential outcomes***, whereas part two addresses the ***assignment mechanism*** (Rubin 2008). A unit is defined as a physical object (e.g., student, patient) at a particular place and point in time, and a treatment is an intervention that can be imposed or withheld from that unit at a particular place and point in time. If the treatment is imposed, such a unit is denoted as a treated unit with a treatment indicator, $W = 1$, and if the treatment is not imposed, such a unit is denoted as a control unit with a treatment indicator, $W = 0$. Each of these units has its own value of some outcome measure, Y (e.g., income, health status, test score).

Thus, in the potential outcomes framework for a dichotomous treatment variable, each unit has a pair of potential outcomes: the potential treatment outcome $Y_i(1)$, which we would observe if treatment is applied, and the potential control outcome $Y_i(0)$, which we would observe under the control condition (i.e., if the treatment is not applied).

By using the simple potential outcomes notation described above, we accept the following two assumptions: (i) treatment applied to one unit does not affect the outcome of the other units – no interference between units (Cox, The Planning of Experiments 1958); and (ii) there is only one type of treatment or control available for each unit – for example, if we are testing whether aspirin removes headache then each person will have only one aspirin pill available or not.

The above assumptions constitute the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980) (Rubin 1990). SUTVA is an untestable assumption even in the controlled randomised experiments and according to Rubin (1991), it can be replaced with some other assumptions.

Under SUTVA, the average causal effect is then defined as the difference between the average potential treatment outcomes and average potential control outcomes over all units:

$$\tau = E[Y(1)] - E[Y(0)]$$

Because we can never observe both potential outcomes at the same time (i.e., each unit can only be exposed to one condition but never to both conditions simultaneously) causal inference is, at its core, a missing data problem – the fundamental problem of causal inference (Holland 1986; Rubin 1978).

If observed and missing outcomes are defined as:

$$Y_i^{obs} = Y_i(W_i) = \begin{cases} Y_i(0) & if \quad W_i = 0 \\ Y_i(1) & if \quad W_i = 1 \end{cases}$$

$$Y_i^{miss} = Y_i(1 - W_i) = \begin{cases} Y_i(0) & if \quad W_i = 1 \\ Y_i(1) & if \quad W_i = 0\,, \end{cases}$$

we can define potential outcomes in terms of observed and missing outcomes:

$$Y_i(0) = \begin{cases} Y_i^{miss} & if \quad W_i = 1, \\ Y_i^{obs} & if \quad W_i = 0. \end{cases} \qquad \text{and} \qquad Y_i(1) = \begin{cases} Y_i^{miss} & if \quad W_i = 0, \\ Y_i^{obs} & if \quad W_i = 1. \end{cases}$$

The second part of the Rubin Causal Model is the ***assignment mechanism*** denoted as the conditional probability of being treated given the observed covariates, X, and potential outcomes, $Y(1)$ and $Y(0)$ (Rubin 1975):

$$\Pr[W \mid X, Y(1), Y(0)].$$

In order to obtain unbiased estimates of treatment effects, usually the assignment mechanism is required to be unconfounded and probabilistic. In an unconfounded assignment mechanism, the selection probabilities do not depend on the potential outcomes (Rubin, 1978),

$$\Pr\left[W \mid X, Y(1), Y(0)\right] = \Pr\left(W \mid X\right),$$

and a probabilistic assignment mechanism assures that each unit has a positive probability of being assigned to either the treatment or control condition,

$$0 < \Pr\left[W_i \mid X, Y(1), Y(0)\right] < 1.$$

An assignment mechanism that is unconfounded and probabilistic is said to be strongly ignorable (Rosenbaum and Rubin 1983a). A treatment effect estimated under such an assignment mechanism can be unbiasedly estimated.

Randomised experiment designs are unconfounded and probabilistic due to an effective (i.e., perfectly implemented) randomisation process. Thus, randomised experiment designs are known to have observed all covariates that are simultaneously related to treatment assignment, W, and the outcome variable, Y. Due to an effective randomisation process, the distribution of observed and unobserved covariates between the treated and control group are, in expectation, balanced.

In contrast, observational designs that lack the randomised selection process typically result in unbalanced covariate distributions and, consequently, in selection bias (due to a confounded assignment mechanism of an observational design). By measuring all the observed covariates[7] we can unconfound[8] an observational assignment mechanism and thus remove selection bias from a study design. Such

[7] A variable selection for observed covariates that are required in the proces of unconfounding the assignment mechanism is described in Section 2.3.2.

[8] By unconfounding an assignment mechanism of an observational design we mean to redesign the assignment mechanism of an observational design in a way to mimic an unconfounded assignment mechanism of randomised experiments, i.e., by posing an assignment mechanism that is unconfounded.

an unconfounded observational assignment mechanism hence closely approximates an assignment mechanism of randomised experiments, thus, unbiased treatment effect estimates can be obtained.

2.3 Design phase

The design phase of propensity score methods consist of **balancing tools** and **balance assessment tools**. The balancing tools consist of methods and techniques used to either remove selection bias[9] from a study design when estimating causal effects or to balance a study design based on the substantively interesting[10] covariates when estimating conditional associations. In both cases the aim is to obtain a balanced design with respect to observed covariates. The balance assessment tools are used to assess the obtained covariate balance and should be used during the design phase of propensity score methods.

The main element of the balancing tools is the propensity score (Rosenbaum and Rubin 1983a), which is a balancing score and thus an essential ingredient in the process of balancing a study design (i.e., removing selection bias from a study design when estimating casual effects or removing covariate imbalances between two groups of our interest when estimating conditional associations) (Section 2.3.1).

In randomised experiment designs, the propensity score is a known function (i.e., an accepted specification for the propensity score exists), whereas in observational designs the propensity score function is mostly unknown (i.e., there is no accepted specification for the propensity score) (Rosenbaum and Rubin 1983a). Thus, observational designs require an estimation of propensity scores from the

[9] Selection bias is removed by unconfounding a confounded assignment mechanism of an observational design; hence, an assignment mechanism which closely approximates an assignment mechanism of a randomised experiment (i.e., an unconfounded and probabilistic assignment mechanism) is obtained.

[10] The selection of the substantivelly interesting covariates is up to the investigator to define and defend as interesting, and surely depends upon the context.

observed data (i.e., observed covariates and the treatment indicator, W, when estimating causal effects or observed covariates and the binary variable, Z, when estimating conditional associations).

The estimation of propensity scores requires thoughtful selection of variables that serve as baseline covariates[11] for which we attempt to balance treatment and control group when estimating causal effects (Section 2.3.2) or selection of the substantively interesting covariates for which we attempt to balance two groups of our interest when estimating conditional associations.

We would like to notify the reader again that, for the sake of clarity, the terminology and notation of propensity score methods in the following sections is for estimating causal effects and should not be mistaken with the terminology and notation that should be used when estimating conditional associations (as explained in Section 2.1).

There are assumptions in the propensity score methods framework that have to be met for being able to estimate, unbiasedly, causal effects, but are not required for successful estimation of conditional associations. The reader will be informed whenever the requirements of the methods when estimating causal effects can be loosened for the estimation of conditional associations.

Again, terms such as: selection bias, treatment indicator - W, treated or control group and baseline covariates (as defined in Section 2.3.2) are used only when estimating causal effects and can be replaced by terms such as: covariate imbalance between two groups, group indicator (i.e., to which group a unit belongs - Z), group 1 or group 2 (or any other name, which depends upon the context of the observed data) and substantively interesting covariates (as defined in the footnote of Section 2.3).

[11] The baseline covariates are the covariates that are observed before treatment is applied to any of the units, thus, often they are called simply as observed covariates, because by definition all covariates are pre-treatment.

2.3.1 Propensity Score

The term propensity score was coined by Rosenbaum and Rubin in 1983 in the paper: The Central Role of the Propensity Score in Observational Studies for Causal Effect. This is a highly cited paper with thousands of citations. The propensity score has two important uses: (i) addressing the dimensionality problem in studies with many covariates; and (ii) the core element in the process of balancing a study design (i.e., removing selection bias when estimating causal effects or removing covariate imbalances when estimating conditional associations).

To balance designs in studies with many covariates can often be difficult, if not even an impossible task. The propensity score can address the dimensionality problem and can make balancing possible regardless of the number of observed covariates. The propensity score, as such, integrates information about the observed covariates and summarises it to a single value on the interval between 0 and 1 for each unit. The numerous observed covariates are thus reduced to a single covariate – the propensity score.

Such a transformation is crucial for studies with high dimensional covariate structure when aiming to remove covariate imbalances between two groups of units. In some special cases, where the number of observed covariates is low and the class of observed covariates is primarily continuous, the originally observed covariates together with the Mahalanobis distance could replace the role of propensity score in removing covariate imbalances between two groups of units.

The reason why the propensity score is characterised as the core element in the process of removing covariate imbalances between two groups of units (i.e., balancing a study design) derives from propensity score definition where the propensity score, $e(X)$, is defined as a balancing score, $b(X)$ – a function of the observed covariates, X, where the conditional distribution of observed covariates given the balancing score is the same for the units in both groups (i.e., ($W=1$) and ($W=0$) when estimating causal effects and ($Z=1$) and ($Z=0$) when estimating conditional associations). This relation can be mathematically shown by using (or abusing) the Dawid's (1979) notation:

$$X \perp W \mid b(X).$$

If treatment indicator, W, is strongly ignorable given observed covariates, X, then the difference between the average potential treatment outcomes and average potential control outcomes at each value of a balancing score is an unbiased estimate of the treatment effect at that value.

On the other hand, the strong ignorability assumption is not required for the group indicator, Z, because when estimating conditional associations we do not deal with the missing data problem, thus, estimation of conditional associations has nothing to do with potential outcomes.

The propensity score adjustment methods, such as matching and subclassification on a balancing score, in general produce unbiased estimates of the average treatment effect (Rosenbaum and Rubin 1983a). If the propensity score is not only a balancing score but it also correctly presents probabilities for a unit to be selected in either treated or control group, then also weighting on the inverse of a balancing score can produce unbiased estimates of treatment effects.

When estimating causal effects, the propensity score, $e(X)$, is defined as the conditional probability of being treated, $W=1$, given the observed covariates:

$$e(X) = pr(W=1 \mid X).$$

Rosenbaum and Rubin (1983a) showed that the propensity score is a balancing score because the conditional distribution of observed covariates given the propensity score is the same for treated ($W=1$) and control ($W=0$) units. This is the critical property of a balancing score because if the treatment assignment is unconfounded given the full set of observed covariates, then it is also unconfounded conditioning only on a balancing score (Rosenbaum and Rubin 1983a). Accordingly, treated and control units with (approximately) the same propensity score have (approximately) identical covariate distributions in expectation. The average treatment effect (ATE) can thus be defined as the difference in conditional expectations of treatment and control group's outcomes at X, averaging all values of X:

$$\tau = E\{E(Y \mid W = 1, X)\} - E\{E(Y \mid W = 0, X)\},$$

where the inner expectations refer to the expected potential outcomes at a given value of observed covariates, X, and the outer expectations average the expected potential outcomes across the distribution of observed covariates in the population.

Because

$$\begin{aligned} & E\{E(Y \mid W = 1, X)\} \\ & = E\{E(Y(1) \mid W = 1, X)\} \\ & = E\{E(Y(1) \mid X)\} \\ & = E[Y(1)], \end{aligned}$$

and similarly

$$\begin{aligned} & E\{E(Y \mid W = 0, X)\} \\ & = E\{E(Y(0) \mid W = 0, X)\} \\ & = E\{E(Y(0) \mid X)\} \\ & = E[Y(0)], \end{aligned}$$

we obtain the average treatment effect as the difference between the average potential treatment outcomes and average potential control outcomes:

$$\tau = E[Y(1)] - E[Y(0)].$$

The same can be shown also for the treatment effect on the treated (ATT) but then the outer expectations do not average the expected potential outcomes across the distribution of all the observed covariates but only across the distribution of observed covariates for the treated group (i.e., the group of units to which treatment is applied).

In observational designs, true propensity scores are unknown and thus, we are required to estimate them from observed data, X and W. The main aim is not to obtain the best propensity score estimates in terms of minimising the difference between the true and the estimated propensity scores, but to obtain propensity score estimates that create balance on the observed covariates between treated and control groups. Thus, often we would prefer to use estimated propensity scores instead of true propensity scores (true propensity scores are known in randomised experiments) (Rubin and Thomas 1992a) because it is not necessary that true propensity scores, which are population propensity scores, would balance our sample equally well as they would balance the population.

It is needless to say that, in real world examples of observational designs, estimated propensity scores are the only choice we have. However, how preferable an estimated propensity score is in case we could choose between estimated or true propensity scores mainly depends on: (i) the propensity score adjustment method used for removing selection bias, and; (ii) the size of selection bias (i.e., the level of covariate imbalance between the treated and control group) – in general the use of estimated propensity scores will remove more selection bias (Rubin and Thomas 1996).

The use of true propensity scores is preferable in cases of heavily unbalanced designs (i.e., large selection bias), and in cases where propensity score weighting adjustment method is used in order to avoid a possibility of misspecifying propensity score model (a misspecified propensity score model can highly bias causal estimates even if the covariate balance based on poorly estimated propensity scores is obtained (Waernbaum 2010)).

On the other hand, estimated propensity scores are preferable for cases where propensity score matching adjustment method is used to obtain well-matched samples because the matching approach can account for random imbalances between covariate distributions in a similar way as those that would arise from randomised experiments (Hirano et al. 2003; Rubin and Thomas 1992b; Rubin and Thomas 1996; Rubin and Thomas 2000).

2.3.2 Propensity Score Model Specification and Selection of Covariates (Variable Selection)

In order to balance an observational design (i.e., removing covariate imbalances between two groups of units), we must first select observed covariates based on which a propensity score model is specified and propensity scores estimated. There are no fundamental rules for selecting covariates when estimating conditional associations. The investigator selects covariates that he/she defines and defends as substantively interesting and that surely depends upon the topic of a study. Having that said, no matter how we specify the propensity score model, when estimating conditional associations, it will be correctly specified as long as obtained propensity score estimates are balancing scores.

On the other hand, when estimating causal effects, the correct specification of the propensity score model matters fundamentally because highly misspecified propensity score models (i.e., modelling propensity scores with only demographic variables) can bias causal effect estimates. The within-study-comparison of Cook et al. (2008) showed that propensity score models that include only demographic variables have failed on a regular basis to reproduce causal effect estimates of experiments.

Thus, we should be cautious about which observed covariates are included in the propensity score model. Observed covariates can be classified as: (i) covariates that are simultaneously related to both, the outcome and the treatment indicator; (ii) covariates that are related only to the outcome variable and not to the treatment indicator; and (iii) covariates that are related only to the treatment indicator, but not to the outcome variable except via its influence on the treatment (i.e., instrumental variable).

Only covariates that are simultaneously related to both – the outcome variable and the treatment indicator – are accountable for removing selection bias. Covariates that are only related to the treatment indicator should be included in the propensity score model only if the functional form of the model is correctly specified, otherwise inclusion of such an instrumental variable decreases precision of treatment effect estimates.

Rubin and Thomas (1996) suggest that all the covariates, that are simultaneously related to both the outcome and the treatment (i.e., treatment indicator, W) should be included in the propensity score model even if a covariate is only slightly related to the outcome variable. Rubin (1997) further claims that excluding such a variable from the propensity score model could result in a more biased treatment effect estimate in comparison to the level of lost efficiency (precision of treatment effect estimates) when we would include such a variable in the propensity score model.

Based on the work of Rubin and Thomas (1996), Rubin (1997), Brookhart et al. (2006) and Austin et al. (2007) the best covariates to be included in the propensity score model are those that are simultaneously related to the outcome variable and to the treatment indicator. However, Brookhart et al.'s expansion of the simulation study to small samples contradicts Rubin's (1997) conclusion regarding the inclusion of covariates that are only weakly related to the outcome variable but simultaneously related to the treatment indicator. Brookhart et al. findings show that inclusion of such a covariate results in a decrease of precision of the estimated treatment effect, while removing only a tiny amount of bias. More research should be done in this area before we could conclude about the role of variable selection for modelling propensity scores with small samples.

Once we successfully select observed covariates for modelling propensity scores, we should start by first setting up a simple model based on substantive knowledge of the problem we are addressing. For example, if a political orientation and residence location are two of our observed covariates, which we include in the propensity score model, and if based on our substantive knowledge, there are fundamental differences in political orientation between different locations, we should consider including an interaction term of those two covariates in our propensity score model.

The estimation of propensity scores can then be performed by using either binomial regression methods (e.g., a linear probability model, logistic or probit regression) or other classification methods (e.g., classification trees, boosted regression, random forest: Mccaffrey et al. 2004; Siroky 2009; Westreichab et al. 2010).

It is often advisable to transform propensity scores, which are in fact estimated probabilities, to the logit scale and thus use linearised propensity scores (Rubin and Thomas 1992a; Rubin 2001),

$$\hat{\mathrm{l}} = \log\left[\frac{\hat{\mathrm{e}}(\mathrm{X})_{\mathrm{i}}}{1-\hat{\mathrm{e}}(\mathrm{X})_{\mathrm{i}}}\right].$$

In this sense those linearised propensity scores are nearly linear in the original covariates and their squares and products (which is important for cases where we include interaction terms or squares of covariates in the propensity score model).

Rubin (2001) lists three reasons for such a transformation: (i) the linear propensity score (i.e., propensity score logit) assesses the efficiency of linear modelling adjustments more adequately in comparison to estimated probabilities; (ii) propensity score logit tends to be more normally distributed (i.e., similar variances and more symmetry) because they are weighted averages; and (iii) propensity score logit is more related to the benchmarks in the literature on covariate adjustments which are based on linearity and normality assumptions.

Once we obtain propensity score estimates, we use different propensity score adjustment methods (Section 2.3.3) to adjust for covariate imbalances between the treated and control groups. The obtained balance is then assessed with different balance assessment tools (Section 2.4). In case balance is not achieved (i.e., substantial covariate imbalances between the two groups of units still exist), we should respecify the propensity score model, repeat the balancing procedure (i.e., apply the propensity score adjustment methods) and assess the balance again.

We should iterate back and forth between the step of specifying the propensity score model, adjusting for covariate imbalances and assessing the obtained balance until selection bias is removed (i.e., until our design is balanced). Only once the study design is successfully balanced (with respect to observed covariates), should the outcome data be examined and causal quantities or conditional associations be estimated.

Even though our main aim, when specifying propensity score model, is to specify it in a way that enables us to obtain a balanced design based on the observed covariates (i.e., remove covariate imbalances and obtain comparable groups), the treatment effect estimators can be sensitive to the specification of the propensity score model. Yet, the level of sensitivity of treatment effects estimators to the correct specification of the propensity score model largely depends on the adjustment method used in removing selection bias.

The three main propensity score adjustment methods are: (i) matching on propensity score – propensity score matching; (ii) using propensity scores to create stratas/subclasses within which we then can use other adjustment methods – propensity score subclassification/stratification; and (iii) weighting by the inverse of the propensity score (i.e., Horvitz-Thompson weighting methods) – propensity score weighting.

Waernbaum (2010; 2011) compared the impacts of a misspecified propensity score model when removing selection bias with propensity score matching and propensity score weighting. His findings are that after obtaining a supposedly balanced design with both adjustment methods, the design obtained with propensity score weighting resulted in biased estimates of treatment effects, whereas this was not the case with the design obtained using propensity score matching. The same criticism regarding propensity score weighting can be found in earlier research (Kang and Schafer 2007; Schafer and Kang 2008; Stuart 2010).

As long as covariate balance is achieved (i.e., differences in covariate distributions between treated and control groups are eliminated – selection bias is removed), the correct specification of the propensity score model, with propensity score matching or propensity score subclassification, will likely matter less than in cases when we use propensity score weighting.

2.3.3 Balancing tools

The balancing tools consist of propensity score adjustment methods with which we adjust for covariate imbalances in observational designs when estimating causal effects or conditional associations. Once a balanced design, with respect to observed covariates, is obtained, an unbiased or approximately unbiased treatment effects can be estimated only under SUTVA (as defined in Section 2.2) and under the assumption that the treatment assignment is strongly ignorable (i.e., treatment assignment, W, and the potential outcomes, $Y(0)$ and $Y(1)$, are conditionally independent given the observed covariates, X:

$$\Pr(W \mid X, Y(0), Y(1)) = \Pr(W \mid X).$$

On the other hand, we can reliably estimate conditional associations without having to consider any of these assumptions. This section describes adjustment methods that can be used in the propensity score methods framework for balancing a study design when aiming to estimate either conditional associations or causal effects.

PROPENSITY SCORE MATCHING

Matching is one of the most widely applied adjustment methods for balancing observational designs, and it has been used as an adjustment method even before the development of propensity scores (Cochran and Rubin 1973). However, the method relies on having a moderately large pool of control units.

Furthermore, propensity score matching is most appropriate for causal inference settings where our aim is in estimating average treatment effects on a treated (i.e., when we are estimating treatment effects on the subpopulation of treated units only) versus population average treatment effect. Nevertheless, in the same way we can also estimate an effect on untreated (i.e., we take the subpopulation of control units). Combining these two effects can then also provide us with an average treatment effect (i.e., population average treatment effect) by taking the weighted average of both estimates where the weights are defined by the number of treated and control units.

The key idea of this method is to find matched pairs of treated and control units that are comparable (i.e., units that share the same covariate values – covariate distributions of the treated and control group are overlapping). The units from each group are matched based on their values of their estimated propensity scores (i.e., a treated and a control unit that have approximately the same values of the estimated propensity score create a matched pair).

There are different ways of how units can be matched. Propensity score matching can be done with or without replacement (i.e., the unit that was already matched can be used to be matched again) and we can also match more control units to one treated unit (k-to-one matching) or vice versa.

The main drawback when matching without replacement is a decrease in efficiency since we would only use the control units which would provide close matches with treated units while the rest of the control units would be discarded. In contrast, matching with replacement enables us to obtain closer matched pairs and hence more balanced designs.

The k-to-one matching (in case we have more control units than treated units) results in less balanced matched pairs but with an increase in efficiency in comparison to one-to-one matching, which discards all the unmatched control units and thus obtains better matched pairs, while facing a decrease in efficiency.

The three main matching approaches proposed by Rosenbaum and Rubin (1985) are: (i) nearest neighbour matching on the estimated propensity score - greedy matching; (ii) Mahalanobis metric matching including the propensity score; and (iii) nearest available Mahalanobis metric matching with callipers defined by the propensity score. Apart from those three there, is also an optimal pair matching (Rosenbaum 1989; Rosenbaum 1991) which creates not only well-matched groups but also well-matched pairs within a group, a genetic matching (Diamond and Sekhon in press), and full optimal matching (Rosenbaum 1986).

The main difference between the nearest neighbour and optimal matching is that optimal matching is based on some optimality criterion of the whole matched sample, whereas the nearest neighbour matching is not. To illustrate in the case of one-to-one matching without replacement, nearest neighbour matching consecutively takes each treated unit (treated units are ordered in decreasing order of estimated propensity score) and matches it to the control unit based on the minimal difference in their propensity score values. Once the match is found, those two units are removed from the pools and we proceed with matching the next treated unit to a control counterpart. In this sense, on average we obtain well-matched groups but not necessarily also well-matched pairs within a group.

On the other hand, optimal matching looks back at the units that were matched and re-matches them in a way that, in addition to obtaining well-matched groups, we obtain well-matched pairs within the group. From this perspective, optimal matching should give us better matched pairs which could result in more comparable groups. However, to the best of our knowledge, no research comparing these two matching approaches has been done except Gu and Rosenbaum (1993), whose findings show that the optimal matching in comparison to the greedy matching can produce closer matched pairs, but does not have an effect on producing better balance of the matched samples. Optimal full matching is an advanced version of the optimal pair matching where treated units can be matched to many control units or vice versa (Hansen 2004; Rosenbaum 1991; Stuart and Green 2008). According to Gu and Rosenbaum (1993), optimal full matching can frequently perform better than optimal matching alone, but we are not aware of a convincing example.

The genetic matching proposed by Diamond and Sekhon in 2005 (Diamon and Sekhon in press) is a generalization of propensity score and Mahalanobis distance matching, and it is based on a genetic algorithm (Sekhon and Mebane 1998; Mebane and Sekhon 1998). Genetic matching does not depend on the estimated propensity score; however, its inclusion does improve it. The algorithm has been used in different applications with large data sets: (Gilligan and Sergenti 2008; Gordon and Huber 2007; Henderson and Chatfield 2009; Herron and Wand 2007; Morgan and Harding 2006; Raessler and Rubin 2005; Sekhon and Grieve 2011) and it appears that it can be superior to other matching methods (e.g., greedy or optimal matching) in obtaining a good balance between treated and control groups (i.e., in obtaining comparable groups), when it is computationally feasible.

One of the most important factors within propensity score matching is the group ratio,

$$R = n_c / n_t \ ,$$

indicating the number of control units per treated unit, where n_c denotes the sample of control units and n_c denotes the sample size of treated units. The group ratio plays a crucial role in propensity score matching because it defines the size of the pool of control units that can be matched to treated units. As mentioned earlier, the success of matching heavily relies on having a moderately large pool of control units in comparison to the pool of treated units. The bigger the pool of control units, the easier it is to find close matches of treated and control units, where a close match is defined with regard to the difference in the observed covariates or the estimated propensity score. For samples with a large group ratio, say $R > 10$, we are more likely to find close matches for each treated unit than for samples with a smaller group ratio.

Another important factor with propensity score matching tells us how many control units are matched to one treated unit $-m$. If we combine both factors (i.e., R and m) into one formula, we obtain the so-called matching ratio, R/m (Rubin 1996). Generally, a bigger matching ratio with one-to-one matching (i.e., $m = 1$) delivers the best matched pairs because closer matches of treated and control units can be found thus, removing the most of selection bias (Rosenbaum and Rubin 1983a).

On the other hand, in cases of k-to-one matching ($m = k$) the group ratio has to be increased in order to achieve the same bias reduction as with one-to-one matching (e.g., two-to-one matching, $m = 2$, requires the group ratio to double) (Rubin 1996). Even though one-to-one matching creates better balance, we are facing a loss in precision because all control units that have not been matched are discarded from future analysis. Thus, there is always a trade-off between the level of removed selection bias and the loss in precision.

Another approach that can be also used with propensity score matching is caliper matching. Caliper matching gives a restriction to how close a treated and control unit have to be with regard to their propensity score values so that such a possible match is considered a close match. For example, if the absolute difference between propensity score values of the treated and control units is bigger than 0.1 standard deviations of propensity scores but the caliper is set to 0.1 standard deviations, such a possible match would not be considered a match. The use of calipers thus enables us to get closer matches, but on the other hand, because of discarding units which do not result in close enough matches, it results in decreased efficiency of estimated treatment effects, and changes the target population by discarding some treated units.

That is, because we are discarding units (treated and control) that do not satisfy the caliper criteria, the structure of our sample changes and it might not reflect the target population of interest anymore. As a consequence, we should be careful with causal statements we are making because the treatment effect estimates might be affected (Crump, et al. 2009). The use of calipers is popular in the presence of a strong selection procedure, when the initial covariate imbalances are big (i.e., study design is heavily unbalanced – the selection bias is large) as in cases where only limited numbers of treated and control units in a sample share a common support (i.e., covariate distributions of the treated and control group are overlapping).

PROPENSITY SCORE SUBCLASSIFICATION (STRATIFICATION)

Subclassification has been used for adjusting imbalances in designs before the development of propensity score methods (Cochran 1968; Cochran and Rubin 1973; Rubin 2006, 7-29). Such simple subclassification faces a major challenge when the number of observed covariates increases because then the number of subclasses grows substantially. For example, if there are ten observed covariates, and each has only two categories, then that would require from us to create $2^{10} = 1024$ subclasses. Creating that many subclasses would result in most of the subclasses not containing both treated and control units. Thus, in order to avoid this problem, we can subclassify on the propensity score

Propensity score subclassification is thus the second main adjustment method. Here control and treated units are divided into subclasses based on the values of the propensity score (Rosenbaum and Rubin 1983a; Rosenbaum and Rubin 1984; Rosenbaum and Rubin 1985). Subclasses should not only be homogenous, but it is also usually desired that subclasses are about the same size. It is common to use five or six subclasses (Rubin 2006); however, with large data sets, we can consider ten or more subclasses as well (Lunceford and Davidian 2004; Rubin and Waterman 2006).

The research conducted before the development of the propensity score methods investigated the required number of subclasses by using only one observed covariate of a continuous variable. Cochran (1968) showed that using only five subclasses, at least 90% of the initial bias usually can be removed. Cochran and Rubin (1973) showed that usually in order to remove 80%, 90% and 95% of the initial bias, three, five and ten classes need to be created, respectively.

To proceed with propensity score subclassification, we must first obtain propensity score estimates, $\hat{e}(X)$, and then form K subclasses based on the sample quantiles of the propensity scores where *j*th sample quantile, $\hat{q}_j$, $j=1,\ldots,K$, is created according to the proportion of $\hat{e}_i \leq \hat{q}_j$ and it is approximately j/K, $\hat{q}_0 = 0$, and $\hat{q}_K = 1$ (Lunceford and Davidian 2004).

PROPENSITY SCORE WEIGHTENING

The propensity score weighting adjustment method also called Inverse-Propensity Weighting (IPW) (Horvitz and Thompson 1952) uses the inverse of the estimated propensity score as a weight where $1/\hat{e}(X)$ is the weight applied to a treated unit and $1/(1-\hat{e}(X))$ is the weight applied to a control unit (Czajka, et al. 1992; Imbens 2000).

The propensity score weighting requires the most cautious of all the main three propensity score adjustment methods. Beside the strong ignorability assumption and SUTVA, the weighting approach requires also for the propensity score model to be nearly correctly specified in order to obtain unbiased or approximately unbiased estimates of treatment effects.

Schafer and Kang (2008) also showed that weighting as an adjustment method often can be poor. It appears to be, in general, much less efficient than any other adjustment methods and quite sensitive to misspecification of the propensity score model. The misspecification of the estimated propensity score model can result in highly biased estimates of treatment effects when the propensity score weighting design is applied (Kang and Schafer 2007; Stuart 2010; Waernbaum 2010). Even if we manage to balance our design using the estimated propensity score, but due to the misspecified propensity score model, estimated propensity scores do not accurately reflect the selection probabilities, the treatment effect estimates obtained by using weighting approach can be badly biased.

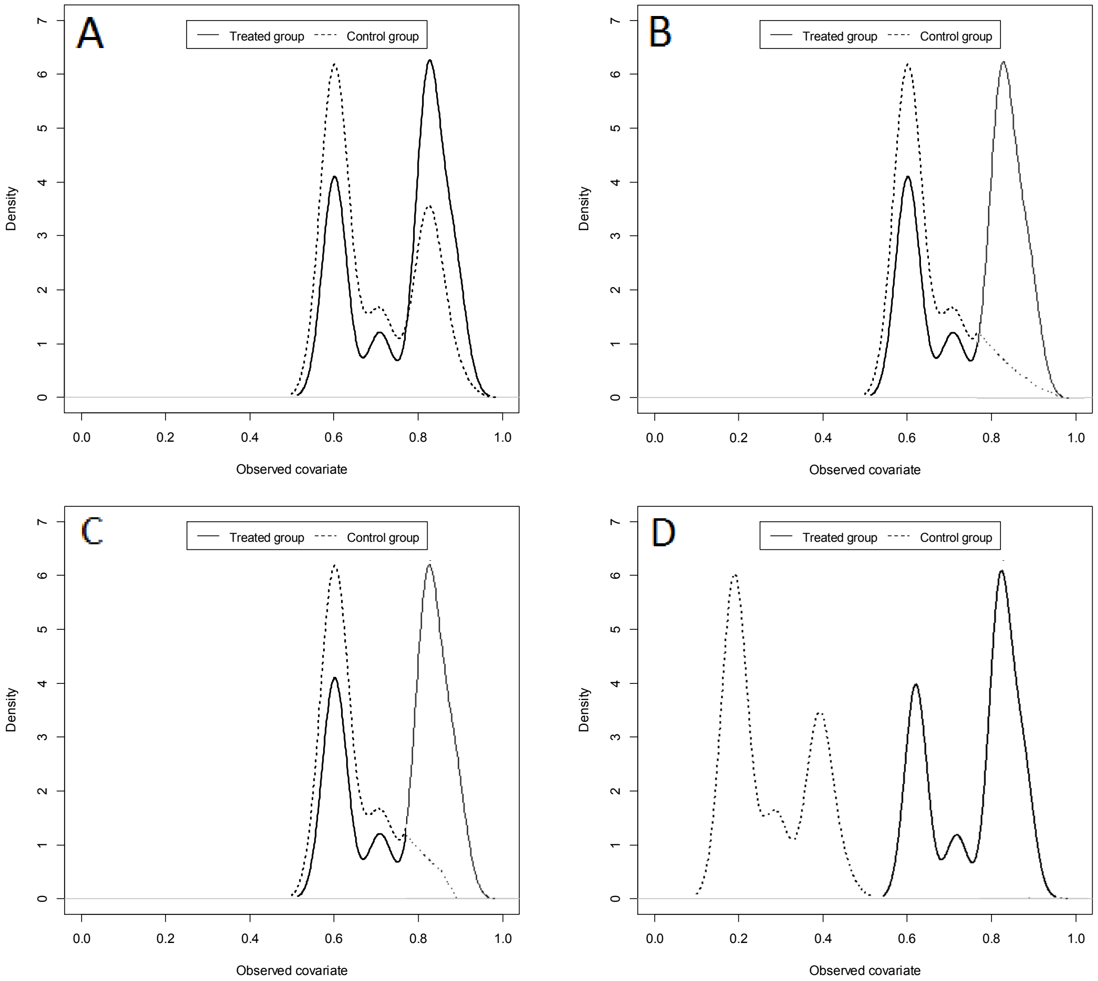
Hirano and Imbens (2001) used propensity score weighting in combination with regression adjustment and obtained stable results. Yet, Waernbaum (2011) showed with his simulation study that in cases of misspecified propensity score models, propensity score matching still performs better than the combination of the propensity score weighting and additional regression adjustment.

2.3.4 Balance assessment tools

Balance assessment tools are used to assess how well distributional forms of the covariate distributions in the treated and control group (or any other two arbitrarily created groups in conditional association studies) overlap. However, besides the **overlap assessment**, we should also assess the **common support issue**, defined as the area where covariate distributions of both groups overlap in geometric terms.

Figure 2.1 provides three illustrations of different overlap and common support levels. Figure A shows an example of a good overlap and a good common support. Figure B shows a lack of the overlap but a good common support. Figure C shows a lack of the overlap combined with the lack of common support; thus, we should discard the units for which covariate distributions do not share a common support in order to avoid extrapolations when balancing a study design. Figure D shows no overlap and no common support either.

Figure 2.1: Illustrations for different levels of overlap and common support



There are two main types of tools to be used for assessing covariate balance: quantitative and graphical tools. These tools should be used before and during the design phase of propensity score methods. The design phase is completed only when the balance assessment tools show negligible differences in covariate distributions for the treated and control group (or any other two arbitrarily created groups when estimating conditional associations).

Assessing covariate imbalances before proceeding with the design phase is important for two reasons. First, it is important to learn whether covariate distributions overlap. If covariate distributions do not overlap, then consequently there may be no common support for the two covariate distributions. In this case we should not proceed with the propensity score study, because treatment effects can only be estimated by heavily relying on extrapolations.

Second, if there is overlap, but we lack a common support (in covariate distributions for the two groups) for some of the units in the sample (as in Figure 2.1 - C), we should discard those units in the design phase to avoid undesired extrapolations and consequently possibly biased treatment effect estimates (Heckman et al. 1997; Dehejia and Wahba 1999).

QUANTITATIVE TOOLS

Some of the most used quantitative tools for assessing covariate balance are: (i) standardised difference in covariates, or logits of propensity score means, between treated and control groups (Rosenbaum and Rubin 1985); (ii) the difference in the means of the propensity score logit between the two groups (Rubin 2001); (iii) the ratio of the variances of the propensity scores logit of the two groups and the ratios of covariates orthogonal to the propensity score (Rubin 2001); (iv) comparing interactions between treated and control groups to determine the similarity of covariances (Ho, et al. 2007); (v) the covariate Mahalanobis distance measure (Rubin 1976).

The standardised mean differences and the Mahalanobis distance measure provide us with information on how many standard deviations apart the treated and control groups are with respect to the covariates or propensity scores. They all share an important property that its values are not sensitive to the measurement scale in the sense of being affinely invariant. The standardised mean difference in covariates,

$$\hat{\Delta}_{tc} = \frac{\overline{X}_t - \overline{X}_c}{\sqrt{\left(s_c^2 + s_t^2\right)/2}},$$

is the difference in average covariate values, X, normalised by an average standard deviation of the covariates, where s_c^2 and s_t^2 are the sample variances of the covariates in the control and treated group, respectively.

The standardised mean difference in estimated propensity score logit is thus defined as

$$\hat{\Delta}_{ct}^{l} = \frac{\bar{l}_t - \bar{l}_c}{\sqrt{\left(s_{l,c}^2 + s_{l,t}^2\right)/2}}$$

where $\bar{l}_c$ and $\bar{l}_t$ are the average values for the linearised propensity scores for control and treated units, and $s_{l,c}^2$ and $s_{l,t}^2$ their corresponding sample variances.

The standardised mean difference can be used as a type of selection bias measure for assessing both: the initial and the remaining bias. Although there are no precise rules of acceptable values of standardised differences after propensity score adjustment, an absolute standardised difference of 0.2 or more might be of concern (Stuart and Rubin 2007). However, absolute standardised differences with values of less than 0.1 usually indicate acceptable covariate balance (i.e., there are negligible differences in covariate distributions between both groups) (Austin 2011; Cochran and Rubin 1973; Cochran 1968; Steiner and Cook in press). Thus, an absolute standardised mean difference of 0.1 or less can often be considered to indicate negligible remaining bias (i.e., approximately unbiased treatment effects can be estimated).

The Mahalanobis distance measure is a similar measure to standardised mean differences with an addition that it also incorporates information about the variance-covariance matrix:

$$B^2 = (\mu_t - \mu_c)' \sum\nolimits_c^{-1} (\mu_t - \mu_c)$$

with μ_t and μ_c denoting the mean value of the treated and control group and $\sum_c$ denoting the variance-covariance matrix of the control group.

The ratio of the variances, VR, of the propensity scores logit of the two samples,

$$VR = \frac{s_{l,t}^2}{s_{l,c}^2},$$

indicates the similarity of the covariate distributions' variance between the treated and control group. The variance ratio should be close to one, whereas values of 0.5 or below and 2 or above are too extreme (Rubin, 2001).

It is important to keep in mind that the metrics that are used to diagnose the balance needed to be aware of with the adjustment method that is used to balance the design. For instance, when using propensity score subclassification, the balance should be assessed within each subclass, or when using propensity score weighting the weights should be incorporated into the calculation of the balance measure.

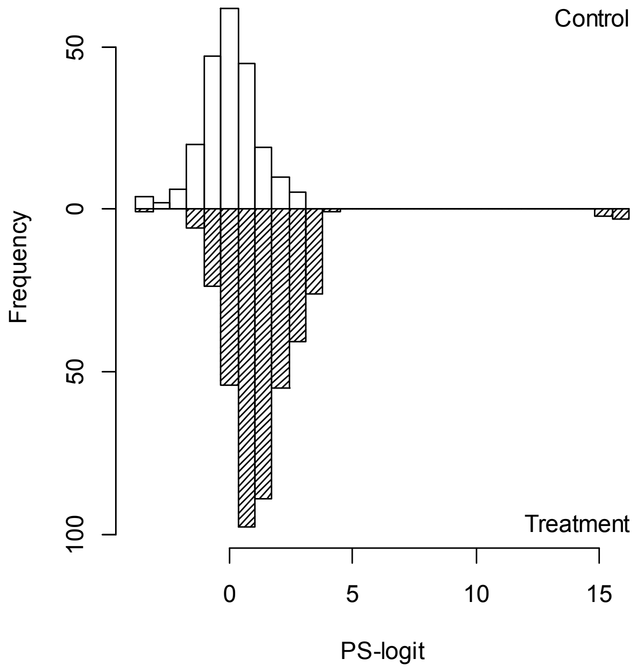
GRAPHICAL TOOLS

Graphical tools are mainly used to get an idea on how balanced a study design is based on the observed covariates in both groups (i.e., treatment and control groups). For a brief assessment of covariate balance we can use some graphical diagnostics tools presented in Figure 2.2 – 2.6.

Figures 2.2 and 2.3 depict covariate distributions of both groups (i.e., treated and control) where the group of units to which treatment is applied is denoted by dashed bars below the horizontal line which represents zero frequency and the group of units to which treatment is not applied is denoted by empty bars above the horizontal line representing zero frequency.

Figure 2.2 depicts a large regime of common support for covariate distributions of treated and control units because there is only a small number of treated and control units for which distributions of observed covariates do not share a common support. Only the units presented in the second and the third bar for the control group and in the last four bars for the treated group do not share a common support in observed covariate distributions. In order to avoid extrapolations, it is recommended to consider discarding those units from a sample when balancing a study design

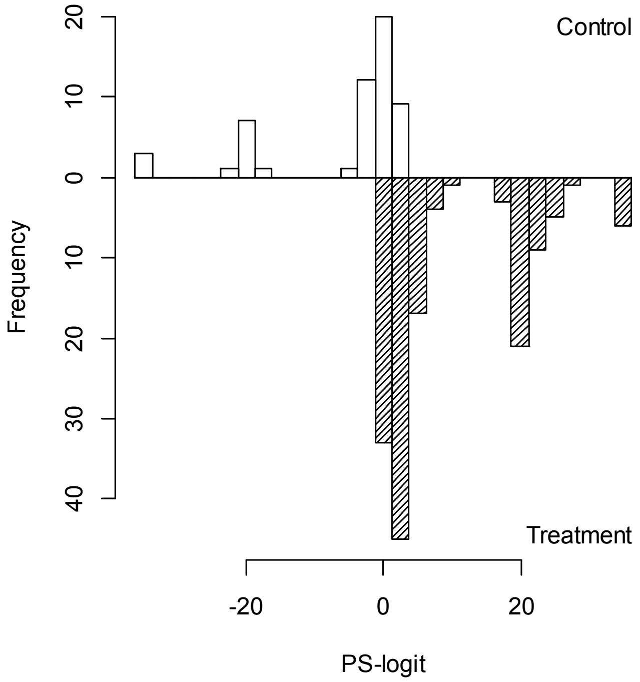
Figure 2.2: Distribution of observed covariates for the treated and control groups – relative large common support (i.e., observed covariates overlap for most of the treated and control units).



On the other hand, Figure 2.3 depicts a very small common support for covariate distributions of treated and control units, because there is a large number of treated and control units for which distributions of observed covariates do not share a common support. Such designs can be called as heavily unbalanced study designs.

As we can see from the Figure 2.3, only covariate distributions of units in the first and second bar of the treated group share a common support. In order to avoid extrapolations, we would have to discard a significant number of treated and control units, which could severely change the structure of the sample (i.e., the sample might not correctly represent our target population anymore).

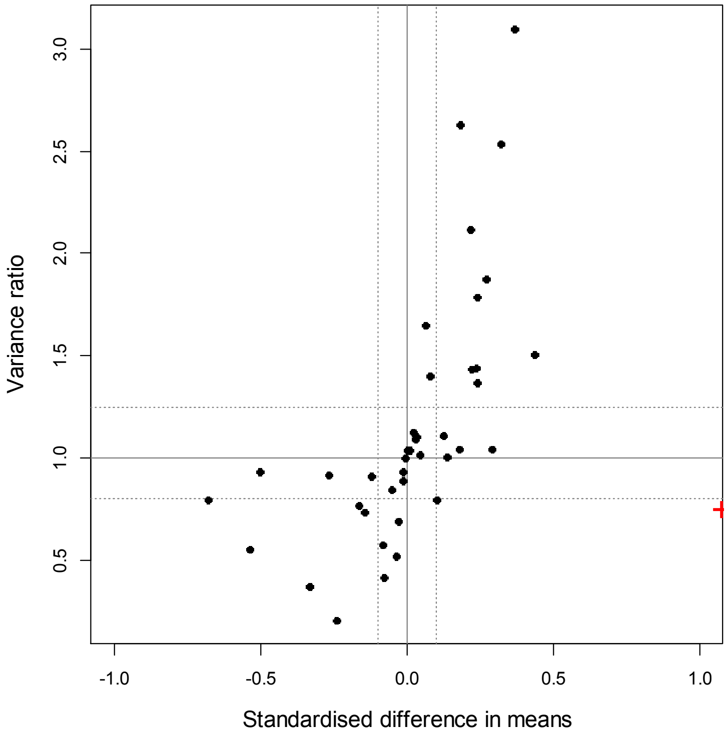
Figure 2.3: Distribution of observed covariates for the treated and control group - small common support (i.e., observed covariates overlap only for some of the treated and control units).



In Figure 2.4 and 2.5, each dot presents the level of covariate balance between the treated and control group for each observed covariate. The balance is measured with standardised differences in means of observed covariates for the treated and control group and with variance ratios for the observed covariates for the treated and control groups.

The dashed horizontal and vertical lines denote “acceptable” levels of the covariate balance (i.e., there are negligible differences in covariate distributions between both groups) where the absolute value of the standardised mean difference should be smaller than 0.1 and where the variance ratio should not be smaller than 0.5 or bigger than 2. The red cross indicates covariate (im)balance in propensity score.

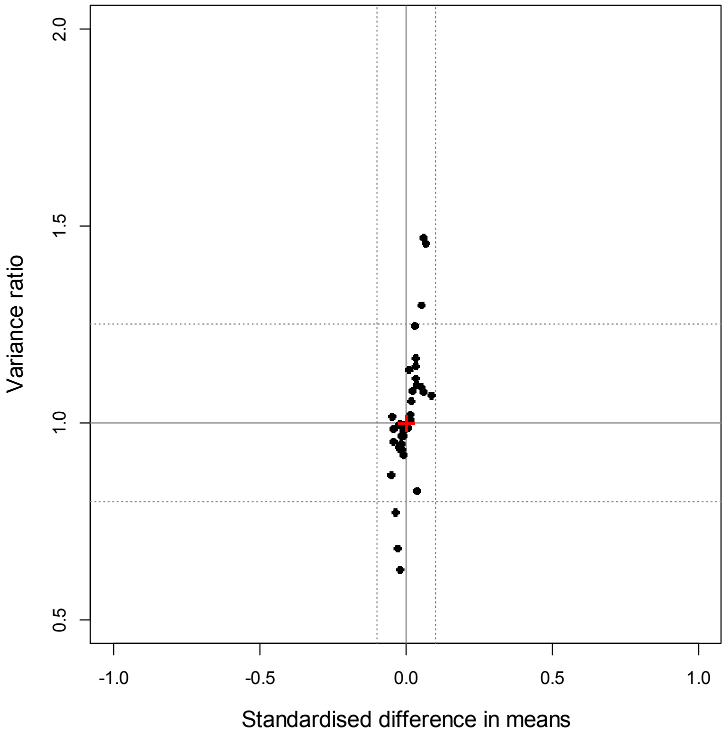
Figure 2.4: An unbalanced design based on the standardised mean difference and the variance ratio diagnostics



Source: Steiner et. al., On the Importance of Reliable Covariate Measurement in Selection Bias Adjustments Using Propensity Scores, 2011.

Figure 2.4 shows an imbalanced design where most of the observed covariates do not satisfy the criterion of acceptable covariate balance. This criterion is not satisfied also for the linearised propensity score (i.e., propensity score logit), indicated by the red cross, which appears to be far off the boundaries of the “acceptable” balance. On the other hand, Figure 2.5 shows an “acceptable” level of the covariate balance for all the observed covariates and also for the propensity score logit.

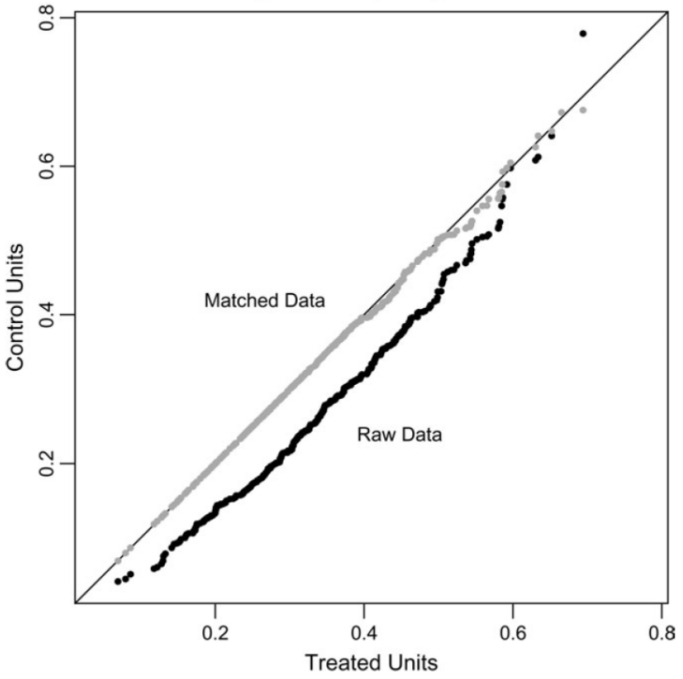
Figure 2.5: A balanced design based on the standardised mean difference and the variance ratio diagnostics



Source: Steiner et. al., On the Importance of Reliable Covariate Measurement in Selection Bias Adjustments Using Propensity Scores, 2011.

Figure 2.6 shows quantile-quantile plot of propensity scores for the raw data (i.e., before the removal of selection bias) and for the matched data (i.e., after the removal of selection bias with propensity score matching adjustment method). The matched data set, where selection bias was removed, clearly shows a more balanced covariate structure between the treated and control group (grey line) than the raw data (black line).

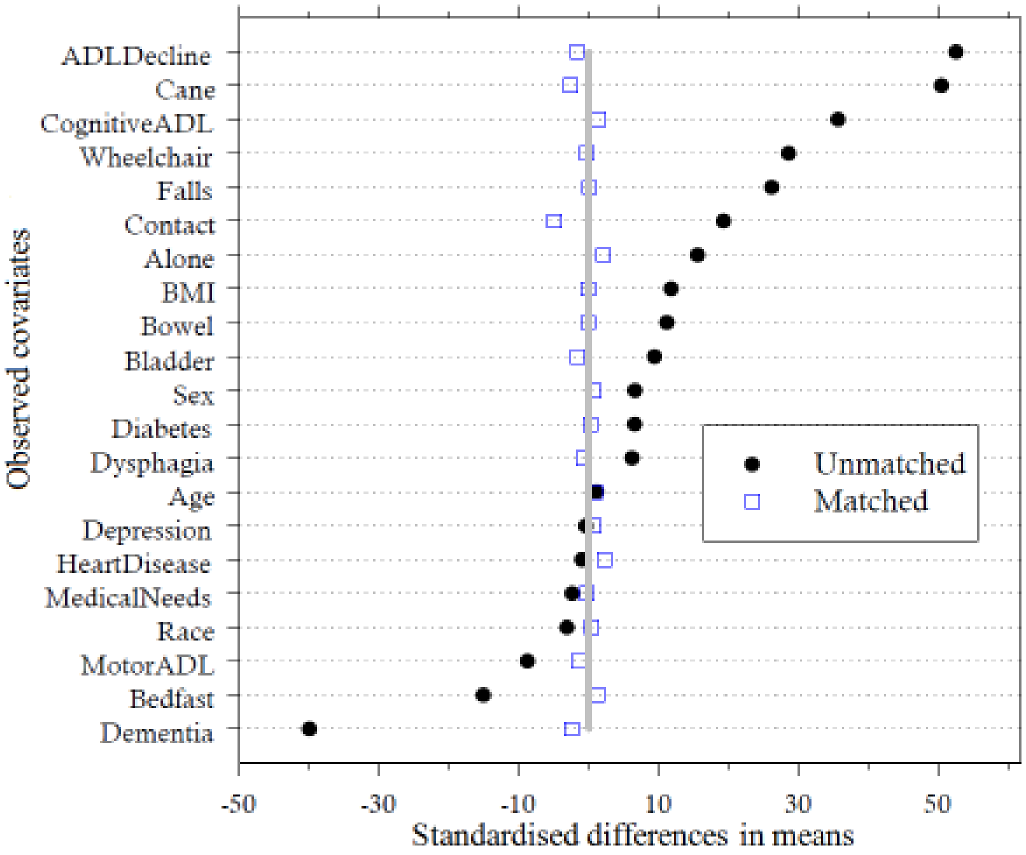
Figure 2.6: Quantile-quantile (QQ) plot of propensity scores



Source: Ho D. E., Imai, King, & Stuart, Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference, 2007

Figure 2.7 shows another version of graphical display for assessing covariate balance in observed data (Love 2002). The horizontal axis presents values for standardised differences in covariate means. The vertical axis presents observed covariates based on which a study design is balanced. The vertical grey line presents a balanced design. The black full circles present standardised differences in covariate means before balancing design (i.e., before proceeding with the design phase of propensity score methods) and the empty blue squares present standardised differences after design was balanced (i.e., after completing the design phase).

Figure 2.7: A balancing plot displying covariate (im)balances of the (un)matched cases



Source: Love, 2002

2.4 Analysis phase

This section presents propensity score estimators, how to proceed with an additional covariate adjustment in order to remove residual covariate imbalances (i.e., the imbalances left after completion of the design phase), and how to perform sensitivity analysis. [12]

[12] Note that when estimating conditional associations there is no need to perform sensitivity analysis.

2.4.1 Propensity Score Estimators

A variety of propensity score estimators can be used for estimating causal effects with propensity score methods. This section presents propensity score estimators that can be used with each of the three main propensity score adjustment methods when estimating causal effects. When estimating conditional associations, the approach is more straightforward (i.e., we take the average difference in outcome values, Y , of the two groups and not the values of potential outcomes as in estimating causal effects), and their estimates will be denoted with the term: conditional comparison estimates.

PROPENSITY SCORE MATCHING ESTIMATOR

The propensity score matching estimator – the estimator used to estimate causal effects when removing selection bias with propensity score matching adjustment method – is, in its the most basic form (i.e., one-to-one matching without replacement), defined as follows:

$$\hat{\tau}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i(1) - \hat{Y}_i(0)) \ ,$$

$$\begin{aligned} \hat{\tau}_{ATT} &= \frac{1}{N_T}\sum_{i \in T}(\hat{Y}_i(1) - \hat{Y}_i(0)) \\ &= \frac{1}{N_T}\sum_{i \in T}(\hat{Y}_i(1) - \hat{Y}_i(0)) \end{aligned}$$

where $i = 1, \ldots, N$ indicates the number of matched pairs for the whole sample, N_T indicates the number of matched pairs for the sample of treated units (Abadie and Imbens 2002).

The $\hat{Y}_i(1)$ and $\hat{Y}_i(0)$ denote imputed potential treatment and control outcomes defined as:

$$\hat{Y}_i^1 = \begin{cases} Y_i & if\ W_i = 0 \\ \frac{1}{M}\sum_{j \in J_M(i)} Y_j & if\ W_i = 1 \end{cases}$$

$$\hat{Y}_i^0 = \begin{cases} \frac{1}{M}\sum_{j \in J_M(i)} Y_j & if\ W_i = 0 \\ Y_i & if\ W_i = 1 \end{cases}.$$

where M denotes the number of matched pairs and $J_M(i)$ denotes that unit j belongs to the group of matched pairs, M, of the unit i. The $\hat{\tau}_{ATE}$ estimates the average treatment effect, whereas the $\hat{\tau}_{ATT}$ estimate the average treatment effect on the treated.

PROPENSITY SCORE SUBCLASSIFICATION ESTIMATOR

When using the propensity score subclassification adjustment method, we first obtain causal effects for each subclass,

$$\hat{\tau}_j = \overline{Y}_{Tj} - \overline{Y}_{Cj}$$

where $j = 1, \ldots, K$ indexes classes with K subclasses. The average treatment effect is then calculated as the weighted average of subclass-specific treatment effect estimates across subclasses.

The subclassification estimators are defined as follows:

$$\hat{\tau}_{ATE} = \sum_{j=1}^{K} w_j \hat{\tau}_j \text{ with weights, } w_j = (N_{Cj} + N_{Tj}) / N,$$

$$\hat{\tau}_{ATT} = \sum_{j=1}^{K} w_{Tj} \hat{\tau}_j \text{ with weights, } w_{Tj} = N_{Tj} / N_T,$$

where subscripts T and C represents treated and control units, respectively and w denotes weights which are calculated as:

$$w_j = n_j / \sum_j^K n_j \ .$$

PROPENSITY SCORE WEIGHTING ESTIMATOR

With the propensity score weighting adjustment methods, the causal estimator is defined as follows:

$$\hat{\tau}_{ATE} = \frac{\sum\limits_{i \in T} w_i Y_i}{\sum\limits_{i \in T} w_i} - \frac{\sum\limits_{i \in C} w_i Y_i}{\sum\limits_{i \in C} w_i}$$

with weights, $w_i = 1/\hat{e}(X_i)$ for the treated units, $i \in T$, and $w_i = 1/(1-\hat{e}(X_i))$ for the control units, $i \in C$. The ATT is calculated with the same equation, the only difference is in the calculation of weights. For the treated units weights are $w_{Ti} = 1$ and for the control units $w_{Ci} = \hat{e}(X_i)/(1-\hat{e}(X_i))$.

Each propensity score adjustment method with its own propensity score estimator requires its own variance estimation of estimated treatment effects. When we would use a combination of different adjustment methods (i.e., subclassification and matching or matching and covariate regression adjustment, etc.) the estimation of variances of different effects is a complex issue. This thesis does not address the topic of how to estimate treatment effect sampling variances, but the following literature does (Abadie and Imbens 2004; Hill and Jerome 2006; Imbens 2004; Rubin and Thomas 1996; Schafer and Kang 2008).

2.4.2 Covariate Regression Adjustment

Covariate regression adjustment can be employed after we completed the design phase of propensity score study. The main reason for employing regression adjustment after completing the design phase of propensity score methods is to remove residual covariate imbalances in our study design – the minor covariate imbalances that could not be removed in the design phase of the use of propensity score methods.

Also, the regression adjustment is trustworthy when the magnitude of the difference in the distributions of the observed covariates between the two groups satisfies at least the following two conditions: (i) the absolute standardised mean difference of the propensity scores in the two groups is smaller than 0.5; (ii) the ratio of the propensity score variances in the two groups is close to one – values of 0.5 and 2 are too extreme. For more in-depth discussion, on the required conditions, please refer to Rubin (2001, 173-174).

The combination of propensity score matching with later covariate regression adjustment is known to be, in general, superior to propensity score matching alone (Rubin 2006, 234) and usually produces less biased treatment effect estimates (Rubin 1973, 185). The same holds when combining the propensity score subclassification with later regression adjustment (Imbens 2004).

As aforementioned, covariate regression adjustment is not part of the design phase because it requires inclusion of the outcome variable. Thus, when using regression adjustment, in addition to some other propensity score adjustment method, the treatment effect is estimated by:

$$\hat{\tau} = \left(\overline{Y}_t - \overline{Y}_c\right) - \hat{\beta}\left(\hat{X}_t - \hat{X}_c\right),$$

where $\overline{Y}_t - \overline{Y}_c$ is the difference in the outcomes of the matched pairs (i.e., matched treated and control units) and the $\hat{X}_t - \hat{X}_c$ is the difference in covariate values for the treated and control group. The latter can also be replaced by $\hat{l}_t - \hat{l}_c$ denoting the difference in propensity scores logit for the treated and control group.

The regression coefficient, $\hat{\beta}$, can be obtained in number of ways (Rubin 1973): (i) as the pooled estimate of the regression coefficient from a one-way analysis of variance, $\hat{\beta}_p$; (ii) as the regression coefficient of differences, $\hat{\beta}_d$, obtained from the regression of

$$Y_{dj} = Y_{tj} - Y_{cj} \text{ on } \hat{X}_{dj} = \hat{X}_t - \hat{X}_c \text{ (Rubin, 1979)}$$

with Y_{dj} being the difference in the outcomes of the matched pairs (i.e., matched treated and control units) and $\hat{X}_{dj}$ the differences in the covariate values of the matched pairs – recommended in pair match settings; (iii) as the regression coefficient from the treated group, $\hat{\beta}_T$, obtained from regressing outcomes of treated units, Y_{tj}, on X of treated units; and (iv) as the regression coefficient from the control group, $\hat{\beta}_C$, obtained from regressing outcomes of control units, Y_{cj}, on X of control units, $\hat{l}_c$.

The decision of which regression coefficient to estimate has to be based on the data characteristics (note that all of the proceeding regression coefficients are estimated with the outcome data we obtain after we balance the design with propensity score adjustment methods). The last two listed regression coefficients, $\hat{\beta}_T$ and $\hat{\beta}_C$ should be avoided when the response surfaces (i.e., outcome variables from both groups – treated and control) are parallel (Rubin 1979).

When the variances of covariates between both groups are approximately equal and their distributions approximately symmetric, using $\hat{\beta}_p$ might result in a slightly less biased results of treatment effects than using, $\hat{\beta}_d$. In cases when control group is at least twice as big as the treated group and propensity score matching method is used using, $\hat{\beta}_d$, will most likely result in the least biased treatment effect estimates (Rubin 1973).

The additional regression adjustment can also be employed by simply regressing Y on some observed covariates for which we believe that they will remove residual imbalances between observed covariates of the treated and control groups (an example is provided with an application in Chapter 6 – Real Data Set 2).

However, some more complex regressions can be done as well, for example, using non-linear terms such as squares or even splines, and in cases when propensity score subclassification adjustment method is employed, we could use an indicator function within subclasses. However, these more extensive covariate regression adjustments would be tough or even impossible to be done in small sample studies.

For more details on how to employ covariate regression adjustments, after completing the design phase of propensity score methods when estimating causal effects or conditional associations, please refer to (Gutman and Rubin 2012; Schafer and Kang 2008; Steiner 2012).

2.4.3 Sensitivity Analysis

Sensitivity analysis is the last step when estimating causal effects with propensity score methods. By using propensity score adjustment methods, when estimating causal effects, we attempt to unconfound the confounded assignment mechanism and thus remove selection bias, which results from a non-randomised selection procedure. By doing so, we remove imbalances in observed covariates; however, we have no control over the hidden bias that might also be part of a study that does not randomly select units in the treated or control group.

In a randomised experiment design, the randomisation takes care of balancing unmeasured covariates (at least in expectation), but this is not the case in a non-randomised design. The hidden bias exists if, for example, there are two units with the same observed covariates, but they have a different chance of receiving or not receiving a treatment (i.e., they have a different chance of assignment to treatment) (Rosenbaum 2010). Because it is a hidden bias, we cannot directly measure it; however, we can use techniques that enable us to study how sensitive our causal claims might be to possible hidden biases.

Sensitivity analysis can be done in a parametric or a non-parametric framework. Parametric sensitivity analysis was developed back in early eighties with the work of Rosenbaum and Rubin (1983b) and also applied in Rosenbaum (1986), whereas the non-parametric sensitivity analysis was developed by Rosenbaum (2005). The most recent approach for analysing sensitivity of causal claims to possible hidden biases can be performed by the Enhanced Tipping-Point Displays which were developed by Liublinska and Rubin (2012) for investigating sensitivity to nonignorable reasons for missing data

In this chapter we present the Rosenbaum (2005) and Liublinska&Rubin (2012) approaches to sensitivity analysis. The former approach is particularly appealing because Keele (2011) developed software (an R package called rbounds) to perform sensitivity analysis based on the Rosenbaum approach. The latter approach is a novelty in the field of sensitivity analyses, and it has been recently accepted by the U.S. FDA (i.e., Food and Drug Administration) as a sufficient approach for studying sensitivity of causal claims to missing outcome data in one example. However, the software to perform Enhanced Tipping-Point Displays is still under development.

ROSENBAUM SENSITIVITY ANALYSIS IN OBSERVATIONAL STUDIES (2005)

The Rosenbaum sensitivity analysis starts with the question of how could our causal claims change in presence of hidden biases? Thus, we try to understand how big the hidden bias should be for our causal claims to change. This is closely related to Cornfield et al. (1958), which may be the first sensitivity analysis in an observational study.

Let say we have two units j and k with the same observed covariates, X, but different probabilities to be assigned to either a treatment or a control group. In order to control for an overt bias (i.e., bias due to imbalances in observed covariates - selection bias) we would create matched pairs with units j and k for which observed covariates are the same, $X_{[j]} = X_{[k]}$, but the chance for each of this unit to get assigned to a treatment group is possibly different, $\pi_{[j]} \neq \pi_{[k]}$. The odds that each of these units get assigned to a treatment is equal to $\pi/(1-\pi)$ for j and k respectively. The odds ratio of j and k units is thus:

$$\frac{\pi_{[j]}(1-\pi_{[j]})}{\pi_{[k]}(1-\pi_{[k]})}.$$

If this odds ratio is equal to one, then $\pi_{[j]} = \pi_{[k]}$ whenever $X_{[j]} = X_{[k]}$. Such studies are hence free of hidden bias. In case this odds ratio equals two, it means that unit j is twice as likely to receive the treatment as unit k.

Rosenbaum (2002) formulates the sensitivity analysis by considering several possible values of Γ so that

$$\frac{1}{\Gamma} \leq \frac{\pi_{[j]}(1-\pi_{[k]})}{\pi_{[k]}(1-\pi_{[j]})} \leq \Gamma$$

for all j, k with $X_{[j]} = X_{[k]}$, and then look at how the causal inference might change with different magnitudes of Γ.

Causal inference of observational study designs assumes strong ignorability which means that all the covariates are observed. In this sense we assume that there is no hidden bias thus the odds ratio should be equal to one,

$$\frac{\pi_{[j]}(1-\pi_{[j]})}{\pi_{[k]}(1-\pi_{[k]})}=1.$$

In case values of Γ close to one lead in a very different causal inference than those obtained assuming strong ignorability (i.e., assuming that the study is free of hidden bias) the study is sensitive to a hidden bias.

If even extreme values of Γ do not change our causal claims, the study is not sensitive to hidden bias, and thus we can be comfortable making strong causal claims.

ENHANCED TIPPING-POINT DISPLAYS

The foundation of the Enhanced Tipping-Point displays is the "tipping-point" analysis introduced by Yan et al. (2009) but anticipated by others. The main objective of the "tipping-point" analysis is to assess if our conclusions about causal claims would have been different under a variety of plausible assignment mechanisms that we would pose for our observational design in order to mimic a randomised experiment design (Cochran and Rubin 1973).

Typically an assignment mechanism for an observational design would be posed under the missing at random assumption (MAR). The Enhanced Tipping-Point displays thus assess sensitivity of our causal claims to unobserved covariates by posing a variety of assignment mechanisms for an observational design under the missing not at random assumption (MNAR). These assignment mechanisms are hence posed based on substantial knowledge of the study field.

Figure 2.7 shows an example of the Enhanced Tipping-Point display where displayed cells represent p-value from a hypothesis test (i.e., H_0: there is a difference in the outcomes between the units to which treatment was applied and units to which treatment was not applied). Although Figure 2.7 shows a missing data mechanism, it could also be used as assignment mechanism; thus, we are going to explain it that way.

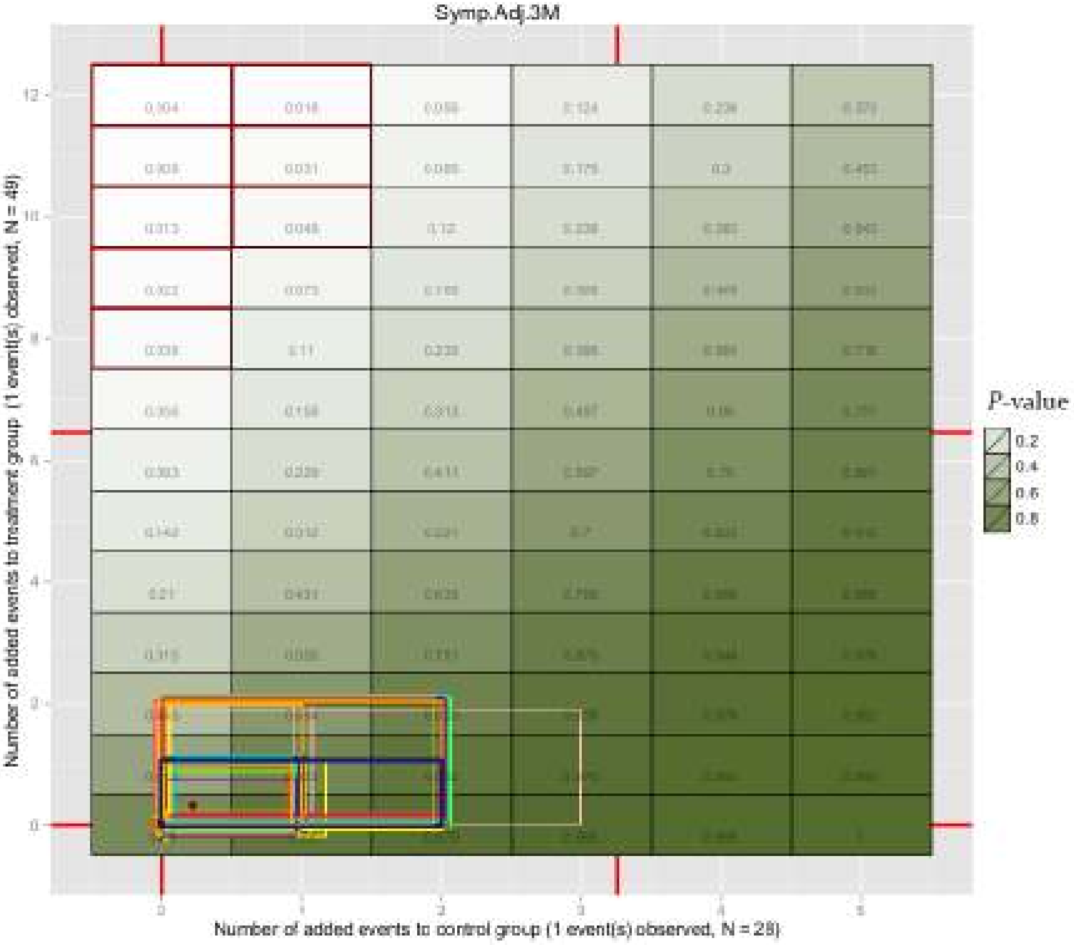
The dark blue square presents an assignment mechanism for observational design being posed under the MAR assumption while the remaining 32 squares present 32 other plausible (more or less extreme) models of assignment mechanisms posed under the MNAR assumption.

The dark blue square is in the area of high p-values denoting that there is a difference in the outcomes between the units to which treatment was applied and units to which treatment was not applied). Accordingly, the tipping-points denoted by the red contour show an area with p-values smaller than 0.05 (i.e., no differences in outcomes between treated and control group).

As long as all the 32 assignment mechanisms posed under the MNAR assumption and presented with 32 squares of different colours are far from the area of cells denoting very small p-values, our conclusions regarding causal claims are not sensitive to unobserved covariates.

Beside p-values, the displayed cells can also represent estimated treatment effect (i.e., specific values of treatment effect) or bounds on interval estimates (Rubin and Liublinska 2012).

Figure 2.7: An example of the Enhanced Tipping-Point display



Source: Rubin and Liublinska, 2012

2.5 Software

There has been a variety of software available to study causal questions within the propensity score method framework; some are presented in Table 2.1 – 2.4.

Table 2.1: Software for STATA

match	It can be used for k:1 matching with and without replacement for estimating: ATT, ATE and robust variances.	Abadie et al., Implementing matching estimators for average treatment effects in Stata 2004 http://www.economics.harvard.edu/faculty/imbens/software_imbens
pscore	It can be used for k:1 matching, matching with a caliper and also when performing propensity score subclassification adjustments for estimating ATT.	Becker and Ichino, Estimation of average treatment effects based on propensity scores 2002 http://www.lrz-muenchen.de/~sobecker/pscore.html
psmatch2	It can be used for k:1 propensity score matching, Mahalanobis matching and kernel weighting for estimating ATT and ATE. It also includes tools for assessing the balance.	Leuven and Sianesi, psmatch2 2003 http://econpapers.repec.org/software/bocbocode/s432001.htm
rbounds	Performs Rosenbaum (2005) sensitivity analysis for ATT	Markus Gangl http://econpapers.repec.org/software/bocbocode/s438301.htm
mhbounds	Performs Rosenbaum (2005) sensitivity analysis	Sascha O. Becker, Marco Caliendo http://ideas.repec.org/p/diw/diwwpp/dp659.html
sensatt	Performs a simulation-based sensitivity analysis for matching estimators	Tommaso Nannicini http://ideas.repec.org/c/boc/bocode/s456747.html http://www.tommasonannicini.eu/Portals/0/sensatt_wp_4.pdf

Table 2.2: Software for R

Matching	It performs multivariate and propensity score matching based on a genetic search algorithm. It includes a variety of univariate and multivariate tests to assess balance	Sekhon, Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R 2011 http://sekhon.berkeley.edu/matching
MatchIt	It includes a variety of matching procedures (i.e., nearest neighbour, Mahalanobis distance, caliper, exact, full, optimal, subclassification) and diagnostic tools for assessing balance	Ho et al., MatchIt: Nonparametric Preprocessing for Parametric Causal Inference 2011 http://gking.harvard.edu/matchit
optmatch	It performs optimal and full matching	Hansen and Klopfer, Optimal full matching and related designs via network flows 2006 http://cran.r-project.org/web/packages/optmatch/index.html
PSAgraphics	It includes a variety of function for assessing balance	Helmreich and Pruzek, PSAgraphics: Propensity score analysis graphics 2009 http://cran.r-project.org/web/packages/PSAgraphics/index.html
rbounds	Performs Rosenbaum (2002) sensitivity analysis	Keele, rbounds: Perform Rosenbaum bounds sensitivity tests for matched and unmatched data 2011 http://cran.r-project.org/web/packages/rbounds/index.html
twang	It provides functions for propensity score estimating and weighting, nonresponse weighting, and diagnosis of the weights	Ridgeway et al., Toolkit for weighting and analysis of non-equivalent groups 2012 http://cran.r-project.org/web/packages/twang/index.html

Table 2.3: Software for SAS

SAS usage note	How to use SAS for matching on propensity scores	http://support.sas.com/kb/30/971.html
gmatch macro	It provides functions for k:1 matching using greedy algorithm	Kosanke and Bergstralh, gmatch 2004 http://mayoresearch.mayo.edu/mayo/research/biostat/upload/gmatch.sas
vmatch macro	It provides functions to perform matching with optimal matching algorithm	Kosanke and Bergstralh, gmatch 2004 http://mayoresearch.mayo.edu/mayo/research/biostat/upload/vmatch.sas
1:1 Mahalanobis matching within propensity score calipers	It provides functions for matching on propensity scores and Mahalanobis distance	Feng, W. W., Jun, Y. and Xu, R. 2005 www.lexjansen.com/pharmasug/2006/publichealthresearch/pr05.pdf
Greedy 1:1 matching	It provides functions for one-to-one matching with the greedy matching algorithm	Parsons, L. S. 2005 http://www2.sas.com/proceedings/sugi25/25/po/25p225.pdf
weighting	It provides functions for propensity score weighting adjustment method	Leslie, S. and Thiebaud, P. 2006 http://www.lexjansen.com/wuss/2006/Analytics/ANL-Leslie.pdf

Table 2.4: Software for SPSS

Propensity score matching in SPSS	It provides functions for propensity score matching with the greedy matching algorithm (i.e., k:1 matching, matching with caliper, matching with or without replacement) and functions to assess balance	Thoemmes, F. 2012: Propensity score matching in SPSS http://sourceforge.net/projects/psmspss/files/ http://sourceforge.net/projects/psmspss/files/

Chapter 3

Sample Size Concerns with Propensity Score Methods

The development of propensity score methods in the last decades resulted in guidance on estimating causal effects from large observational data sets. However, the question of "how large" the treated and control samples should be, or what minimum sample sizes are required for a successful implementation of the propensity score methods with estimated propensity scores, remains mostly unanswered. By using the term "successful implementation" we mean being able to balance a study design, with respect to observed covariates, to the level that possible residual imbalances in observed covariates can be considered as negligible (Section 2.3.4). Nevertheless, more research regarding small sample properties within propensity score methods is important also when the methods are used for estimating conditional associations on substantively interesting covariates – the covariates used to balance a study design.

3.1 Sample Size and Causal Inference with Propensity Score Methods

An overall sample size, n, with a dichotomous treatment variable (i.e., $W=1$ if treatment is applied to a unit, and $W=0$ if treatment is not applied to a unit) is defined as a sum of the units to which treatment is applied (i.e., a treated sample, n_t) and the units to which treatment is not applied (i.e., a control sample, n_c). Thus the overall sample size is $n=n_t+n_c$.

In randomised experiments n_t and n_c are often equal and the samples are balanced with respect to observed and unobserved covariates, thus the size of n_t and n_c matters only from the perspective of statistical power and efficiency (i.e., standard errors).

On the other hand, in observational designs, n_t and n_c are typically not balanced, with respect to the observed and unobserved covariates (which results in selection bias when estimating causal effects), and often not of the same size. Therefore, we are required to first balance such an unbalanced design (with respect to observed covariates), before proceeding with the estimation of causal effects or conditional associations.

The topic of removing selection bias from observational designs, when estimating causal effects, was researched already in the sixties, when Cochran (1965) indicated that samples selected for study should be “large enough” in order to be able to minimise differences in units’ characteristics between the treated and control groups (i.e., to balance a study design – to obtain comparable groups).

In 1996 Rubin and Thomas showed that by using propensity score matching, when estimating causal effects, selection bias can be removed from observational designs with moderately large samples. However, it is unclear from their paper, what sample sizes are considered as moderately large.

Three simulation studies that include small and large samples (Rubin and Thomas 1996; Zhao 2004; Luellen 2007) confirm that successful implementation of propensity score methods (i.e., comparable groups are obtained) depends on used sample sizes. This research thus indicates that guidelines established for large data sets may not be appropriate with small data sets.

3.2 Small Samples with Propensity Score Methods

Small samples are common in various fields of social sciences and medical research (e.g., the number of students in classrooms, number of schools, or number of patients with a rare disease). Thus, the statistical community has been encouraged to conduct more research regarding small sample properties within propensity score methods (Shadish and Steiner 2010).

To our knowledge, only a small number of publications have investigated how successfully selection bias can be removed from observational designs with propensity score methods, when only small samples are available (Rubin and Thomas 1996; Zhao 2004; Luellen 2007). These studies suggest that successful implementation of the methods depends on: (i) the sample size; (ii) the method used to estimate propensity scores, and; (iii) the propensity score adjustment method used to remove selection bias.

Some within-study-comparisons (Shadish et al. 2008; Pohl, et al. 2009), where causal effect estimates from a randomised experiment are compared to those obtained from corresponding non-randomised design, suggest that small samples are in fact capable of approximating results of randomised designs. Yet, more research is required to be able to generalise their conclusions regarding small sample studies.

3.2.1 Issues with Small Samples

One of the main issues with small samples in general in statistical inference is the issue of large standard errors (i.e., the efficiency issue). It has been widely known that statistical inference with small samples is less precise than with large samples. However, the “efficiency issue” is not the only issue that small samples face when estimating causal effects or conditional associations with propensity score methods.

In order to remove covariate imbalances from observational designs, we must first estimate propensity scores. The aim of the estimated propensity scores is to act as balancing scores. The smaller is the sample, the less precise are the propensity score estimates and precision is decreasing, when the number of observed covariates increases. Thus, balancing a study design with less precisely estimated propensity scores might result in more residual imbalances (i.e., the covariate imbalances that remain after completing the design phase of propensity score methods).

Furthermore, observational designs with small samples may also suffer more from the lack of overlap or lack of common support (as defined in Section 2.3.4), than observational designs with large samples. Consequently, due to small sample sizes the lack of a good overlap makes the balancing process more difficult, because it is harder to find comparable units, if only a limited number of units is available. At the same time, by discarding units that do not share a common support, we further reduce our sample size to avoid extrapolations.

Once a balanced design in small sample observational studies is obtained, estimators (i.e., causal estimators or conditional comparison estimators) are much less efficient, due to larger standard errors, than those obtained with large samples.

3.2.2 Past Research on Small Samples when Estimating Causal Effects

When reviewing propensity score publications with small samples, we were interested in publications that investigated how well covariate imbalances can be removed in small sample observational designs, and what is the most appropriate propensity score adjustment method to be used.

The most comprehensive study of propensity score methods was performed by Luellen (2007). Luellen investigated: (i) two different treated sample sizes (i.e., $n_1 = 100$ and $n_2 = 500$) with the group ratio of one; (ii) different methods for estimating propensity scores (i.e., logistic regression, classification trees and ensemble methods such as bootstrap aggregating, boosted regression, and random forest); and (iii) the main propensity score adjustment methods (i.e., matching, subclassification and weighting) performed independently or in combination with an additional regression adjustment.

His simulation study used only one set of observed covariates, $p = 20$ (all continuous), a binary treatment variable (whether a treatment was assigned to a unit or not) and a continuous outcome variable. His simulation results show, that all the investigated factors have an impact on how successfully selection bias can be removed from the observational design (i.e., how effectively treated and control groups can be balanced with respect to the observed covariates).

His findings show that the effect of sample size is sensitive to both: the propensity score adjustment method used to balance an observational study design and to the method used to estimate propensity scores. With the smallest treated sample size $n_1 = 100$, only logistic method for estimating propensity scores, and the propensity score matching adjustment method, implemented with one-to-one matching, performed well.

The ensemble methods for estimating propensity scores are meant to be used with large data sets, therefore, it is not surprising that these methods did not perform well with the smallest sample. The same holds for the subclassification adjustment method (it is meant to be used with large data sets), in order to ensure that each created subclass consists of a sufficient number of units. Hence, in small sample studies we might not even be able to follow Cochran's (1968) advice that at least five to six subclasses should be created. For example, with treated samples, $n_t < 100$, less than 20 treated units would be included in one subclass, which would result in a lack of power, when estimating treatment effects within each subclass.

Luellen's simulation results also show that propensity score weighting performs the worst of all the adjustment methods. Thus, he does not recommend the weighting approach, regardless of the used sample sizes. It is important to note here, that Luellen's study simulates real observed data, thus, he was unable to know whether the model, that he is using to estimate propensity scores, is correctly specified. As explained in Section 2.3.3, the propensity score weighting requires a correctly specified propensity score model in order to estimate, unbiasedly, treatment effects. Luellen's simulation study also confirms previous findings from Rubin (2006, 234), and Hirano and Imbens (2001), that the combination of propensity score adjustment methods with an additional covariate regression adjustment, performs better than any of the adjustment methods alone.

Rubin and Thomas (1996) analytically investigated the role of the group ratio, $R = n_c / n_t$, when employing propensity score matching with one-to-one matching. Their analytical results are based on the moderately large treated samples, however, they tested their analytical findings with a simulation study using small samples. Their findings show, that in cases of moderately large treated samples and the initial bias[13], B, of 0.5, 1.0, and 1.5, group ratios, R, of 2, 3, and 6 are required to eliminate differences in covariate distributions of the treated and control groups. Thus, the greater the difference in the treated and control groups' covariate distributions, the larger the initial bias, and consequently, the more control units per treated unit are required. At the same time, Rubin and Thomas noted, that in cases of smaller treated samples, even larger group ratios are required, but without suggesting how large.

[13] The initial bias, B, is defined in terms of the Mahalanobis distance, $B^2 = (\mu_t - \mu_c)' \sum_c^{-1} (\mu_t - \mu_c)$ with μ_t and μ_c denoting the covariate mean values of the treated and control group and $\sum_c$ denoting the variance-covariance matrix of the control group.

Rubin and Thomas tested their analytical findings with a simulation study of ellipsoidal data by using treated samples of 25 and 50 units (these are, to our knowledge, the smallest treated samples ever investigated in propensity score methods) with 5 and 10 observed covariates, group ratios, R, of 2, 5 and 10, and different level of initial bias, B, of 0.0, 0.25, 0.5, 0.75, 1.0 and 1.5. For the initial bias of $B = 0.5$, their simulation results show, that with a group ratio of $R = 5$ or $R = 10$, essentially all the selection bias is removed whereas for the larger initial biases, the selection bias could not be removed with these group ratios. Similar results were also achieved with their real data simulation. Yet, they do not provide any insights whether different number of observed covariates has an influence on the level of selection bias that can be removed.

Zhao's simulation study investigated small and moderately large treated samples when comparing four different matching estimators[14] (i.e., one-to-one propensity score matching, covariate-Mahalanobis distance matching, covariate-and-propensity-score matching and covariate-and-outcome matching). The smallest investigated sample has 100 treated and 400 control units, indicating a group ratio of four. His results show that the propensity score matching estimator (i.e., one-to-one propensity score matching) most effectively removes selection bias.

3.3 Conclusion

Based on the published research, regarding the propensity score estimation techniques and the propensity score adjustment methods, when using small samples, and by incorporating the theoretical background of the propensity score methods, we can conclude the following: (i) treated samples smaller than 100 have not yet been sufficiently investigated; (ii) the most sensible propensity score estimation method to be used, in cases of small samples, is the logistic regression;

[14] Two of the most widely applied matching estimators are: (i) the propensity scores as defined in the Chapter 2.3.1; and (ii) the covariate matching estimator that uses the Mahalanobis distance in order to balance covariates of both groups.

(ii) the use of the propensity score weighting can be too speculative in cases when the true model of the propensity score is not known, therefore, such an adjustment method might not be the best choice regardless of the used sample size; (iii) the propensity score subclassification is not a realistic option, in cases of small samples, when the smallest treated sample can be as small as consisting of only eight units; (iv) one-to-one propensity score matching appears to perform the best of all the propensity score adjustment methods, in terms of balancing observational study designs; (v) the importance of the size of the group ratio, with small treated samples, has not been investigated beyond the simulation study of Rubin and Thomas (1996) which only used two treated sample sizes (i.e., $n_t = 25$ and $n_t = 50)$, five and ten observed covariates, and three group ratios (i.e., $R = 2, 5, 10$), showing that only $R \geq 5$ removes on average all the selection bias in both of the studied treated samples; (vi) the use of the covariate regression adjustment, after successfully completing the design phase of propensity score methods (i.e., the covariate imbalances are removed to negligible levels) is highly recommended (Rubin 2006, 234; Rubin 2001, 173-174). Such an adjustment can further remove possible residual covariate imbalances that remain after completing the design phase.

Chapter 4

Simulation Study

Based on the discussion in Section 3.3, our simulation study investigates small sample properties of propensity score methods using logistic regression for estimating the unknown propensity score, a one-to-one propensity score matching approach to balance a study design, and an additional covariate regression adjustment for further removing residual bias to estimate the average treatment effect on the treated (ATT).

In particular, our focus is on studying the required sizes of control groups when dealing with small treated samples. Thus, our primary interest lies in examining small sizes of treated samples and the corresponding required sizes of control samples to estimate approximately unbiased treatment effects from observational data. In this sense, our study aims to define a minimum required group ratio, $R^* = n_c / n_t$, for treated samples of size $n_t = 8, 10, 15, 20, 25, 30, 50, 100$, and to compare the findings with the minimum required group ratios for moderately large treated samples of $n_t = 200, 500$. Such a comparison is important from two perspectives.

First, minimum required group ratios for moderately large treated samples were investigated by Rubin and Thomas (1996); thus, inclusion of moderately large treated samples in our study enables us to check the consistency of our results regarding moderately large treated samples to the results obtained by Rubin and Thomas. A high consistency of these two sets of findings increases the reliability of our results for small treated samples. Second, our simulation study also investigates the influence of the number of observed covariates on the minimum required group ratio.

As when studying small treated samples and the required minimum group ratios for satisfactorily removing selection bias from observational designs consisting of small samples, to the best of our knowledge, no research has investigated the influence of the number of observed covariates on the minimum required group ratio in cases of small or moderately large treated samples. Thus, the simulation study focuses on a unique aspect that no previous research has systematically investigated; treated samples that consist of fewer than 100 units, nor the minimum group ratio required for small treated samples, nor the influence of the number of observed covariates on the minimum required group ratio in cases of small and moderately large treated samples.

We present two main simulation studies and two extensions to these. The first main simulation study investigates small sample properties by using estimated propensity scores when applying the matching approach, whereas the second one uses true propensity scores. Although, according to the previous research, balancing designs with estimated propensity scores often is preferable to using true propensity scores, these findings are established only for moderately large and large treated samples. We believe that the behaviour of small treated samples may be quite different; thus such a comparison will have a value for real world scenarios, which often face small treated samples.

Both of the main simulation studies consist of three separate sub-studies covering selection mechanisms of different strengths (i.e., the level of initial imbalances in study design). The first simulation investigates a selection mechanism, which results in an initial squared bias, calculated in terms of the Mahalanobis distance of $B = 0.5$ where,

$$B^2 = \left(\mu_t - \mu_c\right)' \sum\nolimits_c^{-1} \left(\mu_t - \mu_c\right)$$

with μ_t and μ_c denoting the mean value of X in the treated and control group, respectively and $\sum_c$ denoting the variance-covariance matrix of X in the control group.

The second and third sub-studies investigate stronger selection mechanisms with initial biases of 1 and 1.5, respectively.

The first extension of the simulation study assesses whether different correlation structures between the outcome variable and observed covariates (i.e., weaker versus stronger) have an impact on the minimum required group ratio with small treated samples to estimate, unbiasedly, treatment effects. The second extension of the simulation study uses an outcome variable that is binary instead of continuous like in the rest of simulation studies.

This chapter is organised as follows. The first section presents the simulation design for two main simulation studies by describing the data generation process for factors that are known or estimable by the investigator at the design stage (i.e., sample sizes, initial imbalances, the number of observed covariates, etc.), and by presenting measures on quality of the procedures (i.e., remaining bias and variance ratio).

4.1 Simulation – Factors Known or Estimable at the Design Phase

The factors known or estimable by the investigator are the factors that we observe (i.e., treated sample size, group ratio and number of observed covariates), the factors that are estimable (i.e., initial bias) or chosen (i.e., matching algorithm) prior to the implementation of propensity score study. Table 4.1 displays the abovementioned factors together with their levels in the simulation study.

Table 4.1: Factors known or estimable at the design phase

The examined factors	Each Factor's Levels for the Small treated sample study	Each Factor's Levels for the Moderately large treated sample study
n_t - treated sample size	{8, 10, 15, 20, 25, 30, 50, 100}	{200, 500}
R - group ratio	{1:100}	{1:9}
p - number of observed covariates	{10, 15, 20, 30}	{10, 15, 20, 30}
B^2 - initial bias	{0.5, 1.0, 1.5}	{0.5, 1.0, 1.5}
Matching algorithm	{greedy, optimal}	{greedy, optimal}
Factor design	**8 x 100 x 4 x 3 x 2 = 19200**	**2 x 9 x 4 x 3 x 2 = 432**

2.1.1 Data Generation

The simulation design is based on a target population of $N = 1{,}125{,}000$ units from which we draw repeated samples of investigated sizes without replacement. In order to investigate the influence of the number of observed covariates on the minimum required group ratio, we generate four such target populations, each representing a different sized covariate set with p covariates, $p \in \{10, 15, 20, 30\}$.

The observed covariates, X, are generated as independent and normally distributed variables:

$$X \sim N(0,1).$$

The outcome variable, Y, is generated as:

$$Y = \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon \quad \text{with} \quad \varepsilon \sim N(0,1)$$

for each set of p observed covariates. No treatment effect is added, hence, the outcome variable for the treated and for the control group has the same functional form in the treatment and control groups.

The beta coefficients for each set of p covariates are calculated as:

$$\beta_p = \sqrt{\frac{Q}{p}},$$

where the factor Q denotes the covariance between the outcome variable and the linear combination of observed covariates

$$Q = Cov\left(Y, \beta \sum X_i\right),$$

and is derived as follows.

With $X_i, \varepsilon \sim \mathrm{N}(0,1)$ and $Y = \sum \beta X_i + \varepsilon$, it follows that

$$Y = \beta \sum X_i + \varepsilon \ .$$

The covariance structure between a linear combination of X_i, $\underline{X} = \beta \sum X_i$, and Y is then

$$Cov\left(Y, \beta \sum X_i\right)$$

$$= Cov\left(\beta \sum X_i + \varepsilon, \beta \sum X_i\right)$$

$$= Cov\left(\beta \sum X_i, \beta \sum X_i\right) + \underbrace{Cov\left(\varepsilon, \beta \sum X_i\right)}_{\substack{\text{is 0 due to the independence} \\ \text{of the error term}}}$$

$$= \beta^2 Cov\left(\sum X_i, \sum X_i\right) + 0 .$$

Because of the independently normally distributed observed covariates, $X_i \sim N(0, 1)$, it follows:

$$Cov\left(\sum X_i, \sum X_i\right) = Var\left(\sum X_i\right) = \sum Var(X_i) = p .$$

As a result we obtain that

$$Cov\left(Y, \beta \sum X_i\right) = \beta^2 p$$

$$= Q .$$

The calculation of such beta coefficients fixes the correlation structure between the linear combination of observed covariates and the outcome variable for all the covariate sets. We set the factor Q to be 0.35 and hence obtain the correlation structure between observed covariates and the outcome variable we would typically observe in practise:

$$R^2_{Y,\underline{X}} = \frac{Cov\left(Y, \beta \sum X_i\right)}{\sqrt{Var(Y)Var(\beta \sum X_i)}} = \frac{Q}{\sqrt{(Q+1)Q}} = \frac{0.35}{\sqrt{(0.35+1)\cdot 0.35}} = 0.51$$

The numerator in the above equation denotes the covariance between covariates and the outcome variable, $Cov(Y, \underline{X})$, and the denominator denotes the square root variance of the outcome variable, $Var(Y)$ times the variance of $\beta \sum X$, $Var(\beta \sum X)$.

The 1+Q is hence derived from $Var(Y)$ as follows:

$$Var(Y) = Var\left(\beta \sum X_i + \varepsilon\right)$$

$$= \beta^2 Var\left(\sum X_i\right) + \underbrace{Var\left(\varepsilon\right)}_{1}$$

$$= \beta^2 \sum \underbrace{Var\left(X_i\right)}_{1} + 1$$

$$= \beta^2 p + 1$$

$$= Q + 1.$$

Based on $Q = 0.35$, we calculate the values of the beta coefficients, β_i, for each covariate set, p, as follows:

$$\beta_{p=10} = \sqrt{\frac{0.35}{10}} = 0.19$$

$$\beta_{p=15} = \sqrt{\frac{0.35}{15}} = 0.15$$

$$\beta_{p=20} = \sqrt{\frac{0.35}{20}} = 0.13$$

$$\beta_{p=30} = \sqrt{\frac{0.35}{30}} = 0.11 .$$

After generating the target populations for each covariate set, we calculate true propensity scores, $e(X)$, for each target population of size, N_{p_i}, and each strength of the selection mechanism, i.e., initial bias, $B^2 \in \{0.5, 1.0, 1.5\}$. The true propensity scores, $e(X)$, are calculated as

$$\text{logit}^{-1}(e(X)) = \gamma(X_1 + \ldots + X_p),$$

where p is the number of observed covariates and γ is the coefficient determining the strength of the selection mechanism, that is, the size of the initial squared bias B^2. Table 4.2 displays the coefficients used for different selection mechanisms and different numbers of observed covariates.

Table 4.2: Gamma coefficient for calculating true propensity scores, $e(X)$, as a function of B^2 and p.

B^2	$p = 10$	$p = 15$	$p = 20$	$p = 30$
0.5	0.24	0.19	0.17	0.14
1.0	0.35	0.29	0.25	0.21
1.5	0.46	0.37	0.33	0.27

For each target population, the treatment indicator variable, W_i, is generated as a random draw from a Bernoulli distribution with the true propensity score, $e(X)$, representing the probability of being selected:

$$W_i \sim \text{Bernoulli}(\text{prob} = e(X)).$$

2.1.2 Data Generation – Simulation Study Extensions

The first extension of the simulation study only changes the value of the factor Q in order to achieve a stronger correlation structure between the outcome variable and the observed covariates. The second extension of the simulation study replaces only the measurement type of the outcome variable from (i.e., the one in the main simulation studies) from continuous to binary.

SIMULATION STUDY EXTENSION – Stronger Correlation Structure

The only part of the data generation process that changes for this simulation study is the value of the factor Q. In order to generate a stronger correlation structure between the outcome variable and the observed covariates, we increase the value of Q to 1.5. By doing so, the correlation between the outcome variable and the observed covariates is as follows:

$$R^2_{Y,\underline{X}} = \frac{Cov\left(Y, \beta \sum X_i\right)}{\sqrt{Var(Y)Var(\beta \sum X_i)}} = 0.78,$$

which is stronger correlation structure than in the main simulation studies:

$$R^2_{Y,\underline{X}} = 0.51.$$

The remaining parts of the data generation process are the same as in the two main simulation studies; however, the study was done only with the true propensity score – the same as the first main simulation study.

SIMULATION STUDY EXTENSION – Binary Outcome

The only part of the data generation process that changes for this simulation study is the generation of the outcome variable. The outcome variable is generated in two steps. In step one we generate the outcome variable, Y, the same way as in the main simulation studies:

$$Y = \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon \quad \text{with} \quad \varepsilon \sim \text{N(0,1)}$$

for each set of p observed covariates and also without a treatment effect (i.e., the outcome variable for the treated and for the control group has the same functional form).

In step two we discretise the continuous outcome values into binary values according to the following rule:

$$Y_i = \begin{cases} Y_i = 0 & if \quad Y_i < 0, \\ Y_i = 1 & if \quad Y_i \geq 0. \end{cases}$$

The remaining parts of the data generation process are the same as in the two main simulation studies; however, this study is performed only with the true propensity score – the same as the second main simulation study.

4.2 Simulation – Measures of Quality of the Procedure

The quality of a procedure is assessed by two measures: the remaining bias, RB, and the variance ratio, VR. These two measures give us an idea of how well a study design is balanced, after adjusting for covariate imbalances using propensity score matching adjustment methods.

The remaining bias, RB, measures how much of the selection bias is still left (i.e., residual selection bias) after we balance the design with a propensity score matching approach, and it is defined as the absolute standardised mean difference of the logits of propensity scores,

$$RB = \frac{|l_t - l_c|}{\sqrt{\left(s_{lt}^2 + s_{lc}^2\right)/2}}$$

where l_t and l_c are the mean values of the propensity score logit of the treated group and control group and, s_{lt}^2 and s_{lc}^2 are the variances of the propensity score logit for the treated group and for the control group, respectively, where logit of propensity score is calculated as

$$l = \log\left\lfloor \frac{e(X)_i}{1 - e(X)_i} \right\rfloor.$$

The calculation of the remaining bias in the first main simulation study is performed with the estimated propensity scores, $\{\hat{e}(X)_i\}$, whereas in the second main simulation study, it is performed with true propensity scores, $\{e(X)_i\}$.

The variance ratio, VR, is the ratio of the variances of propensity score logits between the treated group and control group, s_{lt}^2 / s_{lc}^2, also calculated using estimated propensity scores in the first main simulation study and with true propensity scores in the second main simulation study.

Because it is rarely possible to remove 100 per cent of the selection bias, previous research (Austin 2011; Cochran 1968; Rosenbaum and Rubin 1983a; Rosenbaum 2010; Rubin 1979; Steiner and Cook (in press)) suggests that the remaining bias, RB, of 0.1 or smaller indicates an acceptable covariate balance (i.e., there are negligible differences in covariate distributions between both groups – approximately balanced study design), whereas with values bigger than 0.1 but smaller than 0.2 an approximately unbiased treatment effect estimates can be obtained provided that an additional statistical adjustment, possibly a covariate regression adjustment, is applied to further remove the residual selection bias. The variance ratio, VR, of propensity scores between treated and control groups should be close to one (values smaller than 0.5 or bigger than 2 are, according to Rubin (2001), considered too extreme).

4.3 Simulated Data

This section describes how the data were simulated (Section 4.3.1) and how we analysed the simulated data (Section 4.3.2). All the simulation studies were performed with 1,000 replications.

The simulation study is programmed and analysed with R (R Core Team 2012). The package MatchIt (Ho, et al. 2011) is used for matching treated and control units with the greedy matching algorithm, and the package optmatch (Hansen and Klopfer 2006) is used to perform matching with the optimal matching algorithm.

4.3.1 Data simulation

The data were simulated by drawing repeated samples without replacement from the target populations (generated as described in Section 4.1.1). We draw 1000 such samples, and in each sample, we perform one-to-one propensity score matching without replacement (i.e., we match one control unit to one treated unit) on the logit of the estimated and true propensity scores, respectively. After obtaining a balanced design, we estimated the treatment effect on the treated by using an additional propensity score regression adjustment.

The simulated data are then summarised by: (i) averaging values of 1000 simulation replications for the remaining bias, RB, the variance ratio, VR, and the treatment effect estimate, $A\hat{T}T$; and (ii) by using the standard deviation of 1000 simulation replications for $A\hat{T}T$ in order to calculate standard errors of $A\hat{T}T$, which are used for constructing confidence intervals of $A\hat{T}T$.

The matching is performed by one of two matching algorithms: greedy or optimal matching. In the study with estimated propensity scores, the propensity scores are estimated via logistic regression according to the equation:

$$logit(W) = \lambda_0 + \lambda_1 X_1 + \ldots + \lambda_p X_p,$$

where W denotes the treatment indicator (i.e., whether treatment was applied to a unit, $W = 1$, or it was not applied to a unit, $W = 0$) and X_p denotes observed covariates for each covariate set, $p \in \{10, 15, 20, 30\}$.

When estimating the average treatment effect on the treated, ATT, we also perform an additional covariate regression adjustment on the matched data, where the regression adjusted treatment effect is calculated as:

$$\hat{\tau} = \left(\overline{Y}_t - \overline{Y}_c\right) - \hat{\beta}_d\left(\hat{l}_t - \hat{l}_c\right),$$

where the regression coefficient, $\hat{\beta}_d$, is obtained from the regression of

$Y_{dj} = Y_{tj} - Y_{cj}$ on $\hat{l}_{dj} = \hat{l}_{tj} - \hat{l}_{cj}$ (Rubin 1979)

with Y_{dj} being the difference in the outcomes of the matched pairs (i.e., matched treated and control units) and the $\hat{l}_{dj}$ are the differences in the propensity score logits of the matched pairs.

According to Rubin (1979), the covariate regression adjustment with $\hat{\beta}_d$ is the most natural one in pair match settings, usually producing the least biased treatment effect estimates in particular when bigger group ratios are used.

4.3.2 Analysis of simulated data

The results of the simulated data are anaysed by performing an analysis of variance – ANOVA, in order to investigate which factors known or estimable by the investigator (i.e., n_t, R, p, B^2 and the matching algorithm) impact measures on quality of the procedure the most (i.e., RB and VR).

Furthermore, descriptive statistical analyses are performed to display how the examined factors (i.e., n_t, R, p, B^2) that lead to balanced design by the criteria of $RB < 0.15$ and $0.5 < VR < 2$ interact. We allow for a bit bigger remaining bias (i.e., $RB < 0.15$) in comparison to the negligable remaining bias (i.e., $RB < 0.10$) because we combine the matching approach in the design phase with an additional propensity score regression adjustment to estimate average treatment effect on the treated – ATT.

The descriptive analysis of the simulated data is then performed by finding the *minimum required group ratio*, R^*, i.e., the smallest group ratio for which the remaining bias (in absolute terms) on the propensity score logit, RB, is smaller than 0.15 standard deviations and that the variance ratio, VR, of the propensity score logit is between 0.5 and 2. More formally:

$$R^* = \min\{R : RB(R) < 0.15 \text{ and } VR(R) < 2\},$$

where $RB(R)$ and $VR(R)$ indicate that the remaining bias and variance ratio is a function of the group ratio, R.

Both of the analyses (i.e., ANOVA and descriptive) are presented in Chapter 5 supported by tables and graphical depictions.

Chapter 5

Results

This Chapter provides results of the analyses for the two main simulation studies summarised by analyses of variance – ANOVAs, and descriptive statistics. Additionally, we present the results of two simulation study extensions. The first simulation extension examines the impact that different correlation structures between observed covariates and the outcome variable have (weaker versus stronger correlation). The second simulation extension examines the impact of a continuous versus a binary outcome variable.

The ANOVA suggests which factors, that are known or estimable in the design phase of propensity score study (i.e., n_t, R, p, B^2), explain the results of measures of quality of the procedures (i.e., RB and VR) the most. Based on the ANOVA findings, we then provide the analysis of descriptive statistics and show how the bias of treatment effect estimates is affected by the levels of the factors known or estimable in the design phase.

5.1 Simulation with Estimated Propensity Scores

This section presents results of the analyses performed on the simulated data when propensity scores are estimated, thereby reflecting real world situations when true propensity scores are unknown.

5.1.1 Analysis of Variance

The simulation consists of a 8 x 100 x 4 x 3 x 2 factorial design (19,200 cells) for the small treated sample study and a 2 x 9 x 4 x 3 x 2 factorial design (432 cells) for the moderately large treated sample study (Table 4.1).

However, the simulated data of the small treated sample study resulted in some empty cells for to the following reason: when estimating propensity scores with very small treated samples (i.e., n_t of 8, 10, 15, 20, 25) and using group ratio, $R=1$, the logistic regression resulted in extreme values of 0 and 1 for all sets of observed covariates, i.e., $p=\{10, 15, 20, 30\}$ all simulation replications. We consider that such behaviour in the simulation violates the probabilistic part of the strong ignorability assumption. Thus, we excluded those from further propensity score analysis.

As a result we have some empty cells in the small treated sample study design. Yet, this problem for very small treated samples did not occur only for the group ratio, R of 1 but also for some larger group ratios when 30 covariates are observed. Table 5.1 displays treated samples, group ratios and the number of observed covariates for which logistic regression resulted in extreme values of 0 and 1.

Each row in Table 5.1 (except the first row) presents treated samples where logistic regression resulted in extreme values of 0 and 1 for each set of observed covariates, i.e., $p=\{10, 15, 20, 30\}$, respectively and for the group ratios, R, specified in the first row of the table. The empty cells represent that for a specific group ratio and the number of observed covariates, the logistic regression did not result in extreme values of zero and one.

Table 5.1: Factors and their levels for which logistic regression, for estimating propensity score, resulted in extreme values of 0 and 1.

	$R=1$	$R=2$	$R=3,4,5,6$	$R=7,8,9,10,11,12$
$p=10$	$n_t=8,10,15,20,25$	/	/	/
$p=15$	$n_t=8,10,15,20,25$	/	/	/
$p=20$	$n_t=8,10,15,20,25$	/	/	/
$p=30$	$n_t=8,10,15,20,25$	$n_t=8,10,15$	$n_t=8,10$	$n_t=8$

p – number of observed covariates; R – group ratio; n_t– treated sample size.

Based on the aforementioned constraints and to perform ANOVA on a balanced design, we carried out three ANOVA analyses with the factor designs presented in the Table 5.2.

Table 5.2: Factors known or estimable by the investigator at the design phase

Factors	Factor's Levels **Small treated sample study 1**	Factor's Levels **Small treated sample study 2**	Factor's Levels **Moderately large treated sample study**
n_t	{8, 10, 15, 20, 25, 30, 50, 100}	{20, 25, 30, 50, 100}	{200, 500}
R	{13:100}	{2:100}	{1:9}
p	{10, 15, 20, 30}	{10, 15, 20, 30}	{10, 15, 20, 30}
B^2	{0.5, 1.0, 1.5}	{0.5, 1.0, 1.5}	{0.5, 1.0, 1.5}
Method	{greedy, optimal}	{greedy, optimal}	{greedy, optimal}
Factor design	**8 x 88 x 4 x 3 x 2 = 16896**	**5 x 99 x 4 x 3 x 2 = 11880**	**2 x 9 x 4 x 3 x 2 = 432**

n_t – treated sample size; R – group ratio; p – number of observed covariates; B^2 – initial squared bias; Method - the matching algorithm used (greedy or optimal).

The ANOVA analyses are performed for both the remaining bias, RB, and the variance ratio, VR, and we include main effects as well as all the interactions (i.e., up to five-way interactions). The results are presented in Tables 5.3 and 5.4 where the factors known or estimable in the design phase are sorted by decreasing order of the mean sum of squares explained - MSS.

ANOVA – REMAINING BIAS

The ANOVA analyses for the remaining bias, RB, with **small treated samples** (the two small treated sample studies) show that the most influential factors in propensity score studies with small treated samples are the treated sample size, n_t, the number of observed covariates, p, the initial squared bias, B^2, the group ratio, R, and the two-way interactions of the treated sample size and the number of observed covariates, $n_t : p$. However, the ANOVA results show no influence of higher levels of the interactions (i.e., above the two-way interaction level) in small treated sample studies.

The most influential factors for **moderately large treated sample** study are the group ratio, R, the initial imbalance, B^2, and the interaction between group ratio and initial imbalances, $R : B^2$. The main difference, in the most influential factors, between small and moderately large treated sample studies is in the most influential two-way interaction factors. In the small treated sample study, the interaction of n_t and p is the most influential, whereas in the moderately large treated sample study, this interaction cannot be considered as influential. On the other hand, in the moderately large treated sample study, the interaction between R and B^2 is the most influential, whereas such an interaction does not appear to have a strong impact on the propensity score study with small treated samples.

However, we should be cautious with such a direct comparison of the most influential factors in small and moderately large treated sample studies because, as is displayed in Table 4.1, the factorial designs of these studies are not fully comparable. Nevertheless, the differences in ANOVA results between small and moderately large treated sample studies can still serve as an indicator that small treated samples behave differently in propensity score studies from moderately large treated samples.

Table 5.3: ANOVA table for Small treated sample study 1 and 2 with estimated propensity scores for the remaining bias measure, RB

Small treated sample study 1			**Small treated sample study 2**		
n_t	{8, 10, 15, 20, 25, 30, 50, 100}		n_t	{20, 25, 30, 50, 100}	
R	{13:100}		R	{2:100}	
Factor	DF	MSS	Factor	DF	MSS
n	7	4.60	B	2	2.26
p	3	2.53	R	98	1.48
B	2	2.37	n	4	1.29
n:p	21	0.32	p	3	0.96
R	87	0.18	n:p	12	0.09
n:B	14	0.09	R:B	196	0.05
p:B	6	0.03	n:B	8	0.04
method	1	0.02	R:n	392	0.04
R:n	609	0.01	R:p	294	0.03
R:B	174	0.01	p:B	6	0.02
R:p	261	0.01	method	1	0.01
n:p:B	42	0.01	R:n:p	1176	0.00
R:method	87	0.00	n:p:B	24	0.00
p:method	3	0.00	R:n:B	784	0.00
R:n:p	1827	0.00	R:p:B	588	0.00
n:method	7	0.00	n:method	4	0.00
method:B	2	0.00	R:method	98	0.00
R:n:method	609	0.00	R:n:p:B	2352	0.00
R:p:method	261	0.00	p:method	3	0.00
n:p:method	21	0.00	method:B	2	0.00
R:n:B	1218	0.00	R:n:method	392	0.00
R:p:B	522	0.00	R:p:method	294	0.00
R:method:B	174	0.00	n:p:method	12	0.00
n:method:B	14	0.00	R:method:B	196	0.00
p:method:B	6	0.00	n:method:B	8	0.00
R:n:p:method	1827	0.00	p:method:B	6	0.00
R:n:p:B	3654	0.00	R:n:p:method	1176	0.00
R:n:method:B	1218	0.00	R:n:method:B	784	0.00
R:p:method:B	522	0.00	R:p:method:B	588	0.00
n:p:method:B	42	0.00	n:p:method:B	24	0.00
R:n:p:method:B	3654	0.00	R:n:p:method:B	2352	0.00

n - treated sample size, n_t; R – group ratio, R ; p – number of observed covariates, p; method – the matching algorithm used; B – initial squared bias; DF – degrees of freedom; MSS – mean sum of squares explained.

Table 5.4: ANOVA table for Moderately large treated sample study with estimated propensity scores for the remaining bias measure, RB

Moderately large treated sample study		
n_t	{200, 500}	
R	{1:9}	
Factor	DF	MSS
R	8	5.29
B	2	1.80
R:B	16	0.10
n	1	0.07
p	3	0.02
R:n	8	0.01
n:p	3	0.00
R:p	24	0.00
method	1	0.00
R:method	8	0.00
n:method	1	0.00
p:method	3	0.00
n:B	2	0.00
p:B	6	0.00
method:B	2	0.00
R:n:p	24	0.00
R:n:method	8	0.00
R:p:method	24	0.00
n:p:method	3	0.00
R:n:B	16	0.00
R:p:B	48	0.00
n:p:B	6	0.00
R:method:B	16	0.00
n:method:B	2	0.00
p:method:B	6	0.00
R:n:p:method	24	0.00
R:n:p:B	48	0.00
R:n:method:B	16	0.00
R:p:method:B	48	0.00
n:p:method:B	6	0.00
R:n:p:method:B	48	0.00

n - treated sample size, n_t; R – group ratio, R ; p – number of observed covariates, p; method – the matching algorithm used; B – initial squared bias; DF – degrees of freedom; MSS – mean sum of squares explained.

ANOVA – VARIANCE RATIO

The ANOVA analyses for the variance ratio, VR, are presented in the Tables 5.5 and 5.6. The results are very similar regarding the most influential factors to the results of the analyses with the remaining bias, RB. However, the treated sample size, n_t, and the number of observed covariates, p, appear to be far more influential regarding the size of the MSS than the rest of the most influential factors (i.e., B^2, $n_t : p$) in small sample studies.

Yet, in the moderately large sample study, the interaction between group ratio and initial imbalances, $R : B^2$, is not a very influential factor, although it is in the ANOVA with the remaining bias. The number of observed covariates, p, is the second most influential factor in the ANOVA for the variance ratio, although this factor has a negligible influence in the ANOVA with the remaining bias in the moderately large treated sample study. The discussion on the obtained results with ANOVA analyses is presented in Section 5.1.3 together with findings obtained from descriptive analysis.

Table 5.5: ANOVA table for Small treated sample study 1 and 2 with estimated propensity scores for the variance ratio measure, VR

Small treated sample study 1			**Small treated sample study 2**		
n_t	{8, 10, 15, 20, 25, 30, 50, 100}		n_t	{20, 25, 30, 50, 100}	
R	{13:100}		R	{2:100}	
Factor	DF	MSS	Factor	DF	MSS
n	7	36.88	p	3	6.20
p	3	33.37	n	4	4.85
B	2	3.30	B	2	1.41
n:p	21	2.95	n:p	12	0.60
R	87	0.25	R	98	0.12
n:B	14	0.08	p:B	6	0.10
p:B	6	0.08	n:B	8	0.06
n:p:B	42	0.02	R:p	294	0.01
R:n	609	0.01	R:n	392	0.01
R:p	261	0.01	n:p:B	24	0.00
method	1	0.00	R:B	196	0.00
R:method	87	0.00	R:n:p	1176	0.00
n:method	7	0.00	method	1	0.00
p:method	3	0.00	R:method	98	0.00
R:B	174	0.00	n:method	4	0.00
method:B	2	0.00	p:method	3	0.00
R:n:p	1827	0.00	method:B	2	0.00
R:n:method	609	0.00	R:n:method	392	0.00
R:p:method	261	0.00	R:p:method	294	0.00
n:p:method	21	0.00	n:p:method	12	0.00
R:n:B	1218	0.00	R:n:B	784	0.00
R:p:B	522	0.00	R:p:B	588	0.00
R:method:B	174	0.00	R:method:B	196	0.00
n:method:B	14	0.00	n:method:B	8	0.00
p:method:B	6	0.00	p:method:B	6	0.00
R:n:p:method	1827	0.00	R:n:p:method	1176	0.00
R:n:p:B	3654	0.00	R:n:p:B	2352	0.00
R:n:method:B	1218	0.00	R:n:method:B	784	0.00
R:p:method:B	522	0.00	R:p:method:B	588	0.00
n:p:method:B	42	0.00	n:p:method:B	24	0.00
R:n:p:method:B	3654	0.00	R:n:p:method:B	2352	0.00

n - treated sample size, n_t; R – group ratio, R ; p – number of observed covariates, p; method – the matching algorithm used; B – initial squared bias; DF – degrees of freedom; MSS – mean sum of squares explained.

Table 5.6: ANOVA table for Moderately large treated sample study with estimated propensity scores for the variance ratio measure, VR

Moderately large treated sample study		
n_t	{200, 500}	
R	{1:9}	
Factor	DF	MSSx100
n	1	0.50
p	3	0.20
R	8	0.20
B	2	0.10
R:n	8	0.00
n:p	3	0.00
p:B	6	0.00
n:B	2	0.00
n:p:B	6	0.00
R:p	24	0.00
R:B	16	0.00
R:n:p	24	0.00
R:p:B	48	0.00
R:n:p:B	48	0.00
R:n:B	16	0.00
method	1	0.00
R:method	8	0.00
n:method	1	0.00
p:method	3	0.00
method:B	2	0.00
R:n:method	8	0.00
R:p:method	24	0.00
n:p:method	3	0.00
R:method:B	16	0.00
n:method:B	2	0.00
p:method:B	6	0.00
R:n:p:method	24	0.00
R:n:method:B	16	0.00
R:p:method:B	48	0.00
n:p:method:B	6	0.00
R:n:p:method:B	48	0.00

n - treated sample size, n_t; R – group ratio, R ; p – number of observed covariates, p; method – the matching algorithm used; B – initial squared bias; DF – degrees of freedom; MSS – mean sum of squares explained.

5.1.2 Descriptive Analysis

The simulated data are analysed as explained in Chapter 4.3, and the results are presented in Tables 5.7 – 5.10. The first two tables present results of propensity score matching with the greedy matching algorithm whereas the other two present results of the matching with the optimal matching algorithm.

The first column in each table presents the treated sample size, n_t, followed by the minimum required group ratio, R^*, (as defined in Section 4.3.2), the absolute value of the remaining bias, RB, the variance ratio of the propensity score logits between the treated and control groups ($VR = s_t^2 / s_c^2$), the 99% confidence intervals of the estimated average treatment effect of the treated - $A\widehat{T}T$, and the simulation standard errors of the $A\widehat{T}T$.

The results show that the minimum required group ratio, R^*, is decreasing when the treated sample size, n_t, is increasing, which means that with more treated units a smaller pool of control units is required, relative to the size of the treated sample, in order to estimate, unbiasedly, treatment effects. This agrees with theoretical results in Rubin (1973).

Furthermore, bigger samples of treated units, as expected, provide us with more precise estimates of treatment effects, as a result of smaller standard errors, SE, of treatment effect estimates for bigger treated samples. For example, with an initial squared bias of $B^2 = 0.5$ and p of 10 or 15, standard errors of the estimated treatment effect (i.e., the average treatment effect on the treated – $A\widehat{T}T$) are almost seven times bigger (i.e., 0.02/0.003=6.7 – Table 5.7) for the smallest treated sample $n_t = 8$ in comparison to the moderately large treated samples, n_t of 200 or 500. This ratio (SE of the $A\widehat{T}T$ between small and moderately large treated samples) increases further with larger initial squared bias, B^2, or with more observed covariates, p. For instance, when $B^2 = 1.5$ and p is 10 or 15, the SE of $A\widehat{T}T$ for $n_t = 8$ are ten times bigger (i.e., 0.02/0.002=10 –

Table 5.7) in comparison to the SE of $A\widehat{T}T$ for $n_t = 500$. With more observed covariates (e.g., $p = 30$) the SE of $A\widehat{T}T$ for $n_t = 8$ are more than ten times bigger in comparison to the SE of $A\widehat{T}T$ for $n_t = 500$ (i.e., 0.027/0.002=13.5 – Table 5.8).

On the other hand, the minimum required group ratio, R^*, for small treated samples (i.e., $n_t < 100$) is greatly influenced by the increased number of observed covariates, p. For instance, when the number of observed covariates increases from $p = 10$ to $p = 30$, the minimum required group ratio increases by more than four times for the treated sample size $n_t = 8$, whereas it barely changes for $n_t \geq 100$.

Table 5.7: Minimum required group ratios for investigated treated samples for $p = 10$ and $p = 15$ - greedy matching algorithm*

$\boldsymbol{p = 10}$						$\boldsymbol{p = 15}$					
n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$	n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$
$\boldsymbol{B^2 = 0.5}$						$\boldsymbol{B^2 = 0.5}$					
8	13	0.149	1.57	[-0.08,0.03]	0.020	8	22	0.142	1.47	[-0.03,0.07]	0.020
10	10	0.144	1.46	[-0.03,0.06]	0.017	10	15	0.149	1.51	[-0.02,0.06]	0.017
15	7	0.123	1.38	[-0.02,0.05]	0.014	15	9	0.139	1.43	[-0.02,0.05]	0.014
20	5	0.138	1.40	[-0.03,0.04]	0.013	20	7	0.125	1.37	[-0.01,0.05]	0.012
25	5	0.107	1.31	[-0.03,0.02]	0.010	25	6	0.115	1.33	[-0.01,0.04]	0.011
30	4	0.128	1.35	[-0.03,0.02]	0.010	30	5	0.122	1.35	[-0.01,0.04]	0.010
50	3	0.137	1.35	[-0.01,0.03]	0.008	50	4	0.101	1.28	[-0.01,0.02]	0.008
100	3	0.086	1.24	[-0.01,0.02]	0.005	100	3	0.102	1.28	[-0.01,0.02]	0.005
200	3	0.061	1.18	[-0.01,0.01]	0.003	200	3	0.066	1.19	[-0.01,0.01]	0.004
500	2	0.140	1.33	[-0.01,0.01]	0.003	500	2	0.148	1.35	[-0.00,0.01]	0.003
$\boldsymbol{B^2 = 1}$						$\boldsymbol{B^2 = 1}$					
8	19	0.149	1.53	[-0.04,0.06]	0.019	8	32	0.149	1.53	[-0.07,0.03]	0.020
10	15	0.137	1.47	[-0.01,0.08]	0.017	10	23	0.144	1.51	[-0.06,0.03]	0.017
15	10	0.137	1.42	[-0.02,0.05]	0.014	15	13	0.142	1.46	[-0.04,0.03]	0.014
20	8	0.131	1.41	[-0.04,0.02]	0.011	20	10	0.137	1.42	[-0.04,0.02]	0.012
25	7	0.130	1.39	[-0.03,0.02]	0.010	25	8	0.143	1.43	[-0.04,0.01]	0.011
30	6	0.136	1.40	[-0.01,0.04]	0.009	30	7	0.139	1.41	[-0.02,0.04]	0.009
50	5	0.131	1.38	[-0.02,0.02]	0.008	50	6	0.114	1.34	[-0.03,0.02]	0.007
100	4	0.138	1.39	[-0.01,0.02]	0.005	100	5	0.101	1.30	[-0.01,0.01]	0.005
200	4	0.112	1.33	[-0.01,0.01]	0.004	200	4	0.119	1.34	[-0.02,0.01]	0.004
500	4	0.101	1.30	[-0.00,0.01]	0.002	500	4	0.104	1.30	[-0.02,0.00]	0.002
$\boldsymbol{B^2 = 1.5}$						$\boldsymbol{B^2 = 1.5}$					
8	27	0.149	1.54	[-0.01,0.09]	0.020	8	47	0.149	1.52	[-0.08,0.02]	0.020
10	22	0.146	1.50	[-0.04,0.05]	0.017	10	33	0.149	1.52	[-0.06,0.03]	0.017
15	14	0.149	1.51	[-0.03,0.04]	0.014	15	20	0.149	1.50	[-0.03,0.04]	0.014
20	12	0.137	1.45	[-0.04,0.03]	0.012	20	15	0.149	1.49	[-0.03,0.03]	0.012
25	10	0.141	1.46	[-0.01,0.05]	0.011	25	13	0.138	1.45	[-0.03,0.03]	0.010
30	9	0.142	1.44	[-0.02,0.03]	0.010	30	11	0.143	1.46	[-0.04,0.01]	0.010
50	8	0.124	1.39	[-0.03,0.02]	0.007	50	9	0.130	1.41	[-0.01,0.02]	0.007
100	7	0.149	1.45	[-0.03,0.01]	0.005	100	7	0.135	1.41	[-0.01,0.02]	0.005
200	6	0.134	1.41	[-0.01,0.01]	0.004	200	6	0.140	1.42	[-0.00,0.02]	0.004
500	6	0.121	1.38	[-0.00,0.01]	0.002	500	6	0.127	1.39	[-0.00,0.01]	0.002

$*$ n_t – treated samples size; RB – the remaining bias; $s^2{}_{ps_t}/s^2{}_{ps_c}$ – variance ratio (VR); R^* - the minimum required group ratio, for each investigated treated sample which satisfies: $\mathrm{RB} < 0.15$ and $0.5 < VR < 2$. The 99% confidence intervals of the estimated ATT, calculated with standard errors (i.e., simulation SE of $\hat{ATT}$) designed by using the standard deviation of treatment effect estimates across 1,000 iterations, $\mathrm{SE}_{\hat{\mathrm{ATT}}} = \mathrm{s}_{\hat{\mathrm{ATT}}}/\sqrt{1000}$, show that the true value of the treatment effect ($\tau = 0$) is within the interval boundaries of the estimated treatment effect. This indicates that treatment effect estimates are unbiased (i.e., selection bias is successfully removed).

Table 5.8: Minimum required group ratios for investigated treated samples for $p=20$ and $p=30$ - greedy matching algorithm*

$p = 20$						$p = 30$					
n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$	n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$
$\boldsymbol{B^2 = 0.5}$						$\boldsymbol{B^2 = 0.5}$					
8	33	0.149	1.50	[-0.10,0.01]	0.026	a8	58	0.149	1.43	[-0.09,0.04]	0.049
10	23	0.144	1.46	[-0.06,0.04]	0.022	a10	45	0.149	1.46	[-0.04,0.07]	0.033
15	13	0.147	1.45	[-0.04,0.03]	0.017	15	22	0.145	1.46	[-0.07,0.00]	0.023
20	9	0.134	1.39	[-0.02,0.04]	0.015	20	14	0.139	1.41	[-0.03,0.03]	0.019
25	7	0.134	1.39	[-0.02,0.03]	0.013	25	10	0.141	1.41	[-0.03,0.03]	0.018
30	6	0.130	1.37	[-0.02,0.04]	0.011	30	8	0.142	1.41	[-0.03,0.02]	0.016
50	4	0.136	1.37	[-0.02,0.02]	0.009	50	5	0.133	1.37	[-0.02,0.01]	0.011
100	3	0.118	1.32	[-0.02,0.01]	0.006	100	4	0.087	1.25	[-0.01,0.02]	0.007
200	3	0.076	1.22	[-0.00,0.01]	0.004	200	3	0.093	1.26	[-0.00,0.01]	0.005
500	3	0.051	1.16	[-0.00,0.00]	0.002	500	3	0.058	1.18	[-0.00,0.01]	0.003
$\boldsymbol{B^2 = 1}$						$\boldsymbol{B^2 = 1}$					
8	45	0.149	1.47	[-0.05,0.06]	0.021	[a] 8	79	0.157	1.45	[-0.07,0.07]	0.027
10	32	0.143	1.48	[-0.04,0.05]	0.018	[a] 10	65	0.140	1.39	[-0.05,0.04]	0.019
15	19	0.137	1.44	[-0.04,0.04]	0.014	15	31	0.148	1.47	[-0.01,0.07]	0.014
20	13	0.143	1.43	[-0.04,0.02]	0.012	20	20	0.145	1.45	[-0.03,0.04]	0.013
25	10	0.146	1.44	[-0.04,0.01]	0.011	25	15	0.140	1.42	[-0.02,0.04]	0.011
30	9	0.136	1.41	[-0.02,0.02]	0.010	30	12	0.141	1.43	[-0.04,0.01]	0.010
50	6	0.139	1.40	[-0.01,0.02]	0.008	50	8	0.129	1.39	[-0.01,0.03]	0.008
100	5	0.115	1.33	[-0.01,0.01]	0.005	100	5	0.145	1.42	[-0.02,0.01]	0.005
200	4	0.131	1.37	[-0.01,0.00]	0.004	200	4	0.149	1.42	[-0.00,0.01]	0.004
500	4	0.111	1.32	[-0.01,0.00]	0.002	500	4	0.113	1.33	[-0.01,0.01]	0.002
$\boldsymbol{B^2 = 1.5}$						$\boldsymbol{B^2 = 1.5}$					
8	65	0.147	1.50	[-0.06,0.05]	0.021	[a]8	98	0.148	1.42	[-0.11,0.03]	0.027
10	48	0.149	1.52	[-0.04,0.05]	0.018	[a]10	75	0.149	1.45	[-0.08,0.03]	0.020
15	27	0.144	1.46	[-0.05,0.02]	0.014	15	43	0.149	1.45	[-0.05,0.03]	0.014
20	18	0.149	1.48	[-0.02,0.04]	0.011	20	29	0.146	1.45	[-0.04,0.02]	0.012
25	15	0.144	1.45	[-0.01,0.05]	0.010	25	22	0.143	1.44	[-0.03,0.02]	0.011
30	13	0.140	1.43	[-0.02,0.03]	0.010	30	18	0.142	1.44	[-0.02,0.03]	0.010
50	9	0.141	1.43	[-0.03,0.02]	0.007	50	12	0.136	1.42	[-0.02,0.02]	0.008
100	7	0.144	1.44	[-0.02,0.01]	0.005	100	8	0.139	1.42	[-0.01,0.02]	0.005
200	6	0.147	1.44	[-0.00,0.02]	0.004	200	7	0.127	1.39	[-0.01,0.01]	0.004
500	6	0.129	1.40	[-0.00,0.01]	0.002	500	6	0.133	1.40	[-0.01,0.01]	0.002

$*$ n_t – treated samples size; RB – the remaining bias; $s^2{}_{ps_t}/s^2{}_{ps_c}$ – variance ratio (VR); R^* - the minimum required group ratio, for each investigated treated sample which satisfies: $\mathrm{RB} < 0.15$ and $0.5 < VR < 2$. The 99% confidence intervals of the estimated ATT, calculated with standard errors (i.e., simulation SE of $\hat{ATT}$) designed by using the standard deviation of treatment effect estimates across 1,000 iterations, $\mathrm{SE}_{\mathrm{A\hat{T}T}} = \mathrm{s}_{\mathrm{A\hat{T}T}}/\sqrt{1000}$, show that the true value of the treatment effect ($\tau = 0$) is within the interval boundaries of the estimated treatment effect. This indicates that treatment effect estimates are unbiased (i.e., selection bias is successfully removed).

[a] Approximately 30% of simulation replications where logistic regression for estimating propensity scores results in extreme values of zero and one.

Table 5.9: Minimum required group ratios for investigated treated samples for $p = 10$ and $p = 15$ - optimal matching algorithm*

$\boldsymbol{p = 10}$						$\boldsymbol{p = 15}$					
n_t	R*	RB	$\frac{s^2_{ps_t}}{s^2_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$	n_t	R*	RB	$\frac{s^2_{ps_t}}{s^2_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$
$\boldsymbol{B^2 = 0.5}$						$\boldsymbol{B^2 = 0.5}$					
8	13	0.141	1.56	[-0.07,0.03]	0.020	8	21	0.143	1.49	[-0.02,0.08]	0,019
10	10	0.135	1.45	[-0.03,0.06]	0.017	10	15	0.141	1.49	[-0.03,0.05]	0,016
15	6	0.143	1.44	[-0.02,0.05]	0.013	15	9	0.131	1.41	[-0.05,0.04]	0,013
20	5	0.127	1.38	[-0.02,0.04]	0.012	20	7	0.117	1.35	[-0.02,0.04]	0,011
25	5	0.147	1.42	[-0.02,0.04]	0.010	25	5	0.148	1.43	[-0.02,0.03]	0,010
30	4	0.117	1.34	[-0.02,0.03]	0.009	30	5	0.113	1.33	[-0.02,0.03]	0,009
50	3	0.127	1.34	[-0.01,0.03]	0.007	50	4	0.093	1.27	[-0.02,0.02]	0,007
100	3	0.079	1.23	[-0.01,0.03]	0.005	100	3	0.095	1.27	[-0.01,0.01]	0,005
200	3	0.056	1.17	[-0.01,0.00]	0.003	200	3	0.062	1.19	[-0.01,0.01]	0,003
500	2	0.139	1.33	[-0.00,0.01]	0.002	500	2	0.147	1.35	[-0.00,0.01]	0,002
$\boldsymbol{B^2 = 1}$						$\boldsymbol{B^2 = 1}$					
8	18	0.148	1.54	[-0.01,0.08]	0.018	8	32	0.146	1.50	[-0.07,0.03]	0.023
10	14	0.146	1.48	[-0.05,0.04]	0.017	10	22	0.146	1.52	[-0.07,0.01]	0.020
15	9	0.148	1.46	[-0.01,0.06]	0.013	15	13	0.136	1.44	[-0.05,0.01]	0.016
20	8	0.124	1.39	[-0.01,0.04]	0.010	20	10	0.131	1.40	[-0.04,0.02]	0.013
25	7	0.122	1.38	[-0.01,0.04]	0.010	25	8	0.136	1.42	[-0.03,0.02]	0.012
30	6	0.129	1.39	[-0.02,0.04]	0.008	30	7	0.132	1.40	[-0.02,0.02]	0.011
50	5	0.125	1.37	[-0.01,0.02]	0.007	50	5	0.149	1.43	[-0.02,0.02]	0.008
100	4	0.135	1.39	[-0.02,0.02]	0.005	100	4	0.148	1.42	[-0.02,0.02]	0.005
200	4	0.110	1.32	[-0.01,0.00]	0.003	200	4	0.118	1.34	[-0.01,0.01]	0.004
500	4	0.101	1.30	[-0.00,0.01]	0.002	500	4	0.103	1.30	[-0.01,0.00]	0.002
$\boldsymbol{B^2 = 1.5}$						$\boldsymbol{B^2 = 1.5}$					
8	27	0.147	1.52	[-0.02,0.08]	0.019	8	47	0.149	1.48	[-0.06,0.05]	0.020
10	21	0.147	1.52	[-0.07,0.01]	0.016	10	33	0.147	1.49	[-0.05,0.04]	0.017
15	14	0.144	1.49	[-0.03,0.04]	0.013	15	20	0.146	1.48	[-0.03,0.04]	0.013
20	12	0.131	1.44	[-0.02,0.04]	0.011	20	15	0.146	1.47	[-0.03,0.03]	0.011
25	10	0.136	1.45	[-0.01,0.04]	0.010	25	12	0.149	1.48	[-0.03,0.03]	0.010
30	9	0.137	1.43	[-0.02,0.01]	0.009	30	11	0.139	1.45	[-0.02,0.03]	0.009
50	8	0.120	1.38	[-0.02,0.01]	0.007	50	8	0.149	1.47	[-0.02,0.02]	0.007
100	6	0.148	1.44	[-0.02,0.02]	0.005	100	7	0.133	1.41	[-0.01,0.02]	0.005
200	6	0.133	1.41	[-0.02,0.00]	0.004	200	6	0.140	1.42	[-0.01,0.01]	0.003
500	6	0.121	1.38	[-0.00,0.01]	0.002	500	6	0.126	1.39	[-0.01,0.01]	0.002

$*\, n_t$ – treated samples size; RB – the remaining bias; $s^2_{ps_t}/s^2_{ps_c}$ – variance ratio (VR); R^* - the minimum required group ratio, for each investigated treated sample which satisfies: $\mathrm{RB} < 0.15$ and $0.5 < VR < 2$. The 99% confidence intervals of the estimated ATT, calculated with standard errors (i.e., simulation SE of $\hat{ATT}$) designed by using the standard deviation of treatment effect estimates across 1,000 iterations, $\mathrm{SE}_{\hat{\mathrm{ATT}}} = \mathrm{s}_{\hat{\mathrm{ATT}}}/\sqrt{1000}$, show that the true value of the treatment effect ($\tau = 0$) is within the interval boundaries of the estimated treatment effect. This indicates that treatment effect estimates are unbiased (i.e., selection bias is successfully removed).

Table 5.10: Minimum required group ratios for investigated treated samples for $p = 20$ and $p = 30$ - optimal matching algorithm *

$\boldsymbol{p = 20}$						$\boldsymbol{p = 30}$					
n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$	n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$
$\boldsymbol{B^2 = 0.5}$						$\boldsymbol{B^2 = 0.5}$					
8	33	0.148	1.49	[-0.10,0.01]	0.021	[a] 8	58	0.149	1.43	[-0.08,0.05]	0.025
10	22	0.146	1.46	[-0.06,0.03]	0.017	[a] 10	45	0.149	1.46	[-0.05,0.05]	0.019
15	13	0.142	1.43	[-0.03,0.04]	0.013	15	22	0.145	1.46	[-0.06,0.01]	0.013
20	9	0.127	1.38	[-0.03,0.03]	0.012	20	14	0.139	1.41	[-0.03,0.03]	0.011
25	7	0.126	1.38	[-0.03,0.03]	0.011	25	10	0.141	1.41	[-0.03,0.03]	0.010
30	6	0.122	1.36	[-0.03,0.02]	0.010	30	8	0.142	1.41	[-0.02,0.03]	0.009
50	4	0.130	1.36	[-0.02,0.02]	0.007	50	5	0.133	1.37	[-0.02,0.01]	0.007
100	3	0.113	1.31	[-0.02,0.01]	0.005	100	4	0.087	1.25	[-0.01,0.01]	0.005
200	3	0.073	1.21	[-0.01,0.01]	0.003	200	3	0.093	1.26	[-0.00,0.01]	0.004
500	3	0.050	1.15	[-0.00,0.01]	0.002	500	3	0.058	1.18	[-0.00,0.01]	0.002
$\boldsymbol{B^2 = 1}$						$\boldsymbol{B^2 = 1}$					
8	45	0.149	1.43	[-0.06,0.04]	0.021	[a] 8	79	0.157	1.45	[-0.07,0.07]	0.028
10	31	0.148	1.48	[-0.04,0.05]	0.017	[a] 10	64	0.140	1.39	[-0.05,0.04]	0.019
15	18	0.145	1.44	[-0.03,0.03]	0.013	15	31	0.148	1.47	[-0.02,0.05]	0.014
20	13	0.138	1.41	[-0.03,0.03]	0.011	20	19	0.145	1.45	[-0.02,0.04]	0.012
25	10	0.140	1.42	[-0.04,0.02]	0.010	25	14	0.140	1.42	[-0.03,0.03]	0.010
30	9	0.131	1.40	[-0.04,0.01]	0.009	30	11	0.141	1.43	[-0.03,0.02]	0.009
50	6	0.134	1.39	[-0.02,0.03]	0.007	50	8	0.129	1.39	[-0.02,0.02]	0.007
100	5	0.112	1.33	[-0.02,0.02]	0.005	100	5	0.145	1.42	[-0.01,0.01]	0.005
200	4	0.129	1.37	[-0.01,0.01]	0.003	200	4	0.149	1.42	[-0.01,0.01]	0.004
500	4	0.110	1.32	[-0.01,0.01]	0.002	500	4	0.113	1.33	[-0.00,0.01]	0.002
$\boldsymbol{B^2 = 1.5}$						$\boldsymbol{B^2 = 1.5}$					
8	65	0.149	1.45	[-0.06,0.11]	0.021	[a] 8	98	0.148	1.41	[-0.10,0.04]	0.019
10	46	0.147	1.47	[-0.08,0.06]	0.017	[a] 10	81	0.147	1.38	[-0.10,0.01]	0.013
15	26	0.148	1.46	[-0.04,0.06]	0.013	15	44	0.145	1.42	[-0.04,0.03]	0.012
20	18	0.147	1.47	[-0.03,0.05]	0.011	20	29	0.149	1.44	[-0.05,0.01]	0.011
25	15	0.141	1.44	[-0.04,0.04]	0.010	25	21	0.140	1.42	[-0.02,0.03]	0.009
30	13	0.137	1.42	[-0.03,0.04]	0.009	30	18	0.133	1.41	[-0.02,0.03]	0.007
50	9	0.138	1.42	[-0.01,0.04]	0.007	50	12	0.138	1.42	[-0.02,0.02]	0.005
100	7	0.143	1.43	[-0.02,0.02]	0.005	100	8	0.127	1.39	[-0.01,0.01]	0.003
200	6	0.146	1.44	[-0.01,0.01]	0.003	200	7	0.132	1.40	[-0.01,0.01]	0.002
500	6	0.129	1.40	[-0.01,0.01]	0.002	500	6	0.148	1.31	[-0.01,0.00]	0.019

$*$ n_t – treated samples size; RB – the remaining bias; $s^2{}_{ps_t}/s^2{}_{ps_c}$ – variance ratio (VR); R^* - the minimum required group ratio, for each investigated treated sample which satisfies: $\mathrm{RB} < 0.15$ and $0.5 < VR < 2$. The 99% confidence intervals of the estimated ATT, calculated with standard errors (i.e., simulation SE of $\hat{ATT}$) designed by using the standard deviation of treatment effect estimates across 1,000 iterations, $\mathrm{SE}_{\mathrm{A\hat{T}T}} = \mathrm{s}_{\mathrm{A\hat{T}T}}/\sqrt{1000}$, show that the true value of the treatment effect ($\tau = 0$) is within the interval boundaries of the estimated treatment effect. This indicates that treatment effect estimates are unbiased (i.e., selection bias is successfully removed).

[a] Approximately 30% of simulation replications where logistic regression for estimating propensity scores results in extreme values of zero and one.

Our initial descriptive-analysis findings are furthermore supported with some graphical depictions to see how strongly treated sample sizes depend on the adequate size of the control group in removing a sufficient amount of selection bias (i.e., $RB(R) < 0.15$ and $VR(R) < 2$), before using an additional covariate regression adjustment to estimate ATT.

In the beginning of the descriptive analysis section, we already mentioned that the minimum required group ratio, R^*, decreases when the size of a treated sample, n_t, increases. We also mentioned that the minimum required group ratios for small treated samples, $n_t < 100$, increases when the number of observed covariates increases, p.

Figure 5.1 shows this relation between the minimum required group ratio, R^*, and the size of the treated sample, n_t, on the log scale, $\log(n_t)$ for the initial squared bias of $B^2 = 1$ (other B^2 values produce very similar depictions) with the lines representing the four covariate sets.

Figure 5.1: Comparison of the number of observed covariates, p, versus the treated sample sizes and their correspondingly required minimum group ratios, R^* with $B^2 = 1$

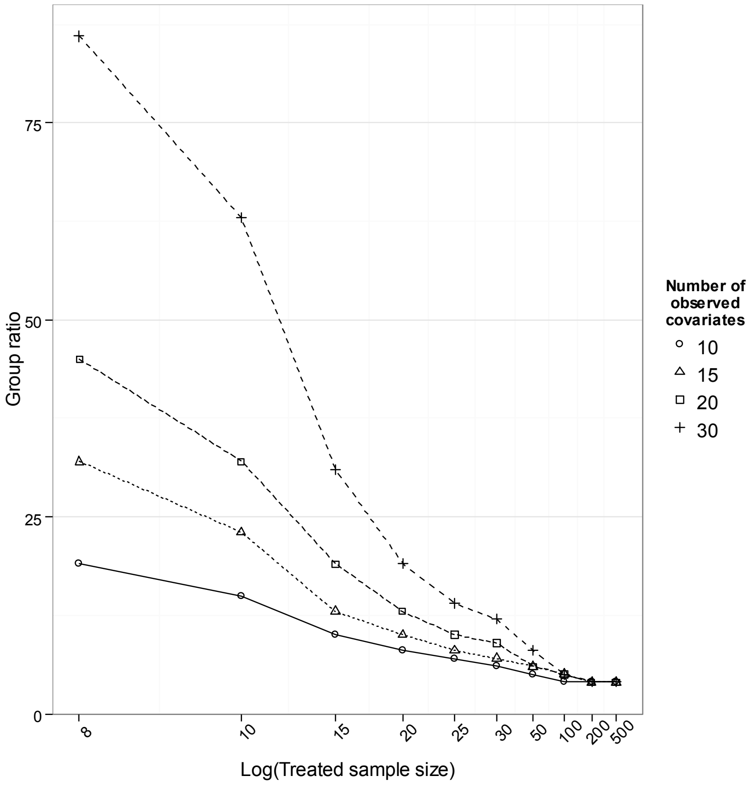
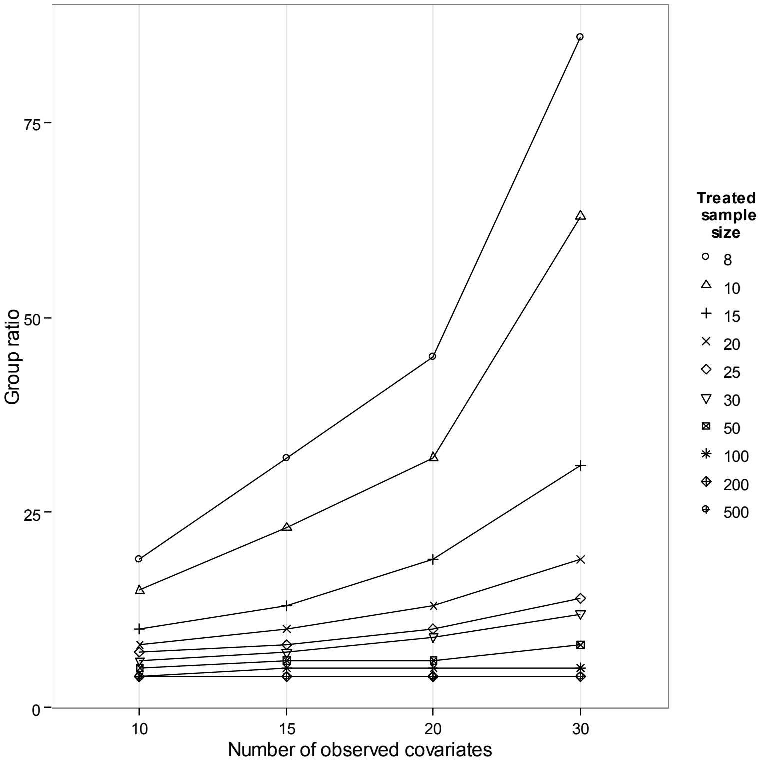


Figure 5.1 demonstrates that the minimum required group ratio increases when the size of the treated sample decreases. Furthermore, the figure also shows that the minimum required group ratio does depend, not only on the size of the treated sample, but also on the number of observed covariates (each line represents each covariate set). However, this finding is particularly applicable for treated samples smaller than 100. A treated sample $n_t = 8$ with $p = 10$ requires a minimum group ratio of $R^* = 19$, whereas with $p = 30$, requires almost $R^* = 80$. On the other hand, a treated sample of $n_t = 100$ with $p = 10$ requires $R^* = 4$, whereas with $p = 30$ it requires only $R^* = 5$.

The relation between the minimum required group ratio, R^*, and the number of observed covariates, p, is depicted in Figure 5.2, where lines denote different treatment sample sizes. The figure shows that treated samples with $n_t < 100$ are sensitive to the number of observed covariates. Particularly with treated samples of $n_t < 20$, the relationship between R^* and p shows a strong exponential functional form, whereas with treated samples of n_t >100, the number of covariates has a negligible effect on the group ratio.

Figure 5.2: Relationship between the number of observed covariates, p, and the minimum required group ratio, R^*, for different treated samples (presented with lines) when $B^2 = 1$ (other B^2 produce very similar depictions).

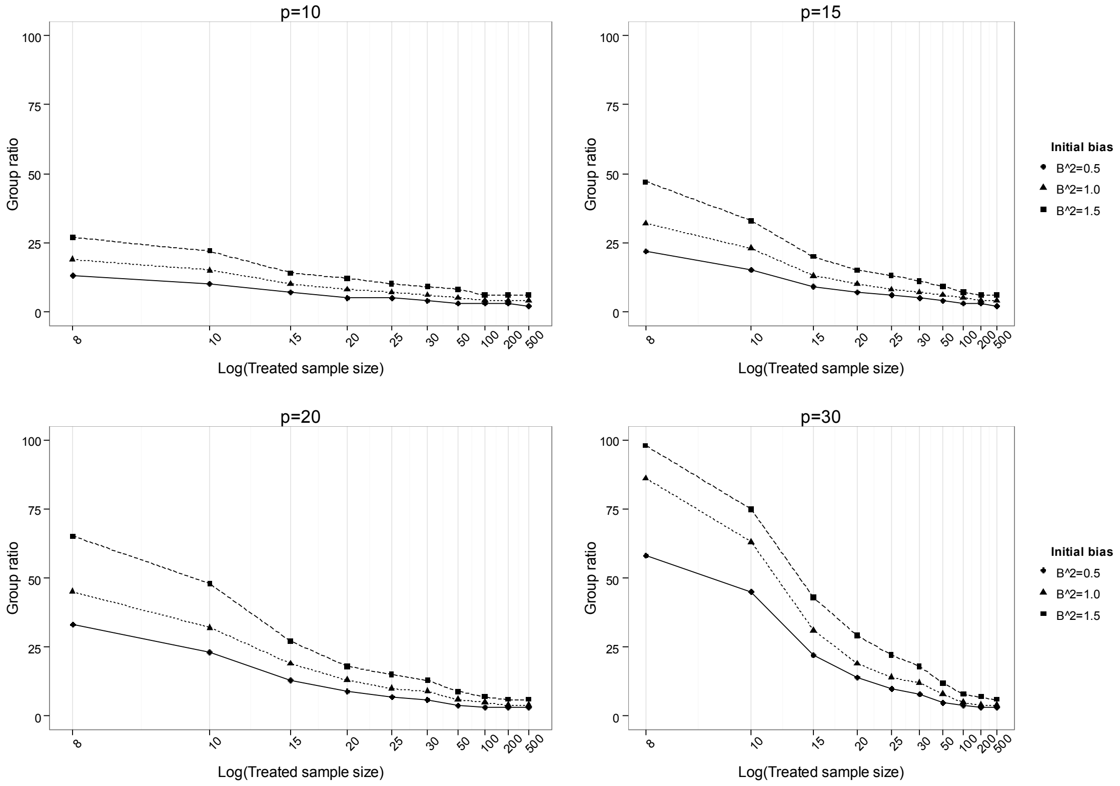


Treated sample, n_t, of 200 and 500 are presented with the same line (the first line from bottom-top) due to the same values of the minimum required group ratios.

Besides the vast impact that the number of observed covariates has on the minimum required group ratio with small treated samples, the number of observed covariates plays another important role when estimating propensity scores with small samples. The simulated data show that with 30 observed covariates, the smallest investigated treated samples of $n_t = 8$ and $n_t = 10$, with a required minimum group ratio of 79, and 65, respectively, when the initial squared bias is $B^2 = 1$, result in over 30 per cent of simulation replications where we estimate that the probabilistic part of the strong ignorability assumption appears to be violated (in the sense that the logistic regression for estimating propensity scores results in extreme values of zero and one). Thus, the group ratios in the above mentioned settings should be even larger, $R^* > 100$.

Furthermore, the level of initial imbalances in study designs has its impaction with the minimum required group ratio as well. The minimum required group ratio increases with an increasing level of initial imbalances for small treated samples. For example, in the case of $p = 10$ covariates, and an initial squared bias of $B^2 = 0.5$, our results demonstrate that the treated sample of $n_t = 8$ requires at least a group ratio of $R^* = 13$. If the initial squared bias of $B^2 = 1$ or $B^2 = 1.5$, a group ratio of at least 19, and 27, respectively, is required. The relation between different levels of initial squared biases and the minimum required group ratios is depicted in Figure 5.3, where each graph within the figure presents the relation between the initial imbalances (different initial biases, i.e., B^2 of 0.5, 1 and 1.5 are depicted with lines) and minimum required group ratio for each treated sample, n_t and each covariate set, p.

Figure 5.3: Relationship between different initial squared biases, B^2, and the minimum required group ratio, R^*, for different treated samples and different number of observed covariates, p.



Later, we investigate possible differences in results of propensity score study when matching is performed with the greedy or the optimal matching algorithm. The evaluation is done by comparing mean-squared-errors (MSE) of the estimated treatment effects obtained using different matching algorithms for all the treated samples, number of observed covariates and initial squared biases. The MSE of an estimator $\hat{\tau}_{ATT}$ with respect to the estimated parameter τ_{ATT} is defined as:

$$MSE(\hat{\tau}_{ATT}) = E\left((\hat{\tau}_{ATT} - \tau_{ATT})^2\right)$$

and it is thus equal to the sum of variance and the squared bias of the estimator

$$MSE(\hat{\tau}_{ATT}) = Var(\hat{\tau}_{ATT}) + (Bias(\hat{\tau}_{ATT}, \tau_{ATT}))^2.$$

The results of the MSE of the estimated treatment effects are based on the propensity score study with the minimum required group ratios for each treated sample, covariate set and level of initial imbalances, respectively (Table 5.1).

Although the optimal matching algorithm performs slightly better, in comparison to the greedy matching algorithm with regard to the MSE, the difference in MSE is not significant. Yet, the difference is bigger for small treated samples in comparison to the moderately large treated samples.

Table 5.11: MSE of $\hat{ATT}$ when matching is performed with greedy and optimal algorithm

		$B^2 = 0.5$		$B^2 = 1$		$B^2 = 1.5$		$MSE_{Greedy} - MSE_{optimal}$		
n_t	p	Greedy	Optimal	Greedy	Optimal	Greedy	Optimal	$B^2 = 0.5$	$B^2 = 1$	$B^2 = 1.5$
8	10	0.42	0.39	0.38	0.33	0.41	0.36	0.03	0.05	0.05
10	10	0.31	0.29	0.30	0.30	0.29	0.27	0.02	0.00	0.02
15	10	0.20	0.17	0.19	0.18	0.20	0.18	0.03	0.01	0.02
20	10	0.16	0.14	0.12	0.11	0.13	0.12	0.02	0.01	0.01
25	10	0.11	0.10	0.11	0.10	0.12	0.11	0.01	0.01	0.01
30	10	0.10	0.08	0.09	0.08	0.10	0.09	0.02	0.01	0.01
50	10	0.06	0.05	0.06	0.05	0.05	0.05	0.01	0.01	0.00
100	10	0.03	0.02	0.03	0.02	0.03	0.02	0.01	0.01	0.01
200	10	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00
500	10	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.01	0.01
8	15	0.38	0.35	0.40	0.36	0.39	0.38	0.03	0.04	0.01
10	15	0.28	0.26	0.29	0.29	0.30	0.28	0.02	0.00	0.02
15	15	0.19	0.18	0.19	0.17	0.20	0.17	0.01	0.02	0.03
20	15	0.15	0.13	0.15	0.13	0.14	0.13	0.02	0.02	0.01
25	15	0.12	0.10	0.11	0.10	0.10	0.10	0.02	0.01	0.00
30	15	0.09	0.08	0.09	0.08	0.09	0.08	0.01	0.01	0.01
50	15	0.06	0.05	0.06	0.05	0.05	0.05	0.01	0.01	0.00
100	15	0.03	0.02	0.03	0.03	0.03	0.02	0.01	0.00	0.01
200	15	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00
500	15	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.01	0.01

Table 5.11 (continues): MSE of $A\hat{T}T$ when matching is performed with greedy and optimal algorithm

		$B^2 = 0.5$		$B^2 = 1$		$B^2 = 1.5$		$MSE_{Greedy} - MSE_{optimal}$		
n_t	p	Greedy	Optimal	Greedy	Optimal	Greedy	Optimal	$B^2 = 0.5$	$B^2 = 1$	$B^2 = 1.5$
8	20	0.41	0.39	0.42	0.38	0.40	0.39	0.02	0.04	0.01
10	20	0.36	0.29	0.30	0.30	0.31	0.28	0.07	0.00	0.03
15	20	0.20	0.17	0.20	0.17	0.19	0.18	0.03	0.03	0.01
20	20	0.16	0.15	0.15	0.13	0.13	0.13	0.01	0.02	0.00
25	20	0.13	0.11	0.12	0.10	0.11	0.10	0.02	0.02	0.01
30	20	0.10	0.09	0.09	0.08	0.10	0.08	0.01	0.01	0.02
50	20	0.06	0.05	0.06	0.05	0.06	0.05	0.01	0.01	0.01
100	20	0.03	0.03	0.03	0.02	0.03	0.02	0.00	0.01	0.01
200	20	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00
500	20	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.01	0.01
8	30	0.42	0.38	0.44	0.44	0.39	0.30	0.04	0.00	0.09
10	30	0.35	0.31	0.31	0.30	0.32	0.30	0.04	0.01	0.14
15	30	0.20	0.17	0.19	0.18	0.20	0.18	0.03	0.01	0.06
20	30	0.16	0.13	0.17	0.15	0.15	0.14	0.03	0.02	0.03
25	30	0.12	0.11	0.13	0.11	0.12	0.12	0.01	0.02	0.03
30	30	0.11	0.09	0.10	0.09	0.09	0.09	0.02	0.01	0.04
50	30	0.06	0.05	0.06	0.05	0.06	0.05	0.01	0.01	0.03
100	30	0.03	0.02	0.03	0.03	0.03	0.03	0.01	0.00	0.02
200	30	0.01	0.01	0.02	0.01	0.01	0.01	0.00	0.01	0.01
500	30	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.01

5.1.3 Summary

The descriptive analyses complement the ANOVA, which reveals the most influential factors for propensity score study design for small and moderately large treated samples. The most influential factors obtained by performing ANOVA for small (i.e., n_t, R, p, B^2, and the two-way interaction, $n_t : p$) and moderately large treated sample studies (i.e., R, B^2, and the two-way interaction, $R : B^2$) are well in line with the findings of descriptive analyses.

Both treated sample sizes (i.e., small and moderately large) can equally well remove a sufficient amount of selection bias from observational studies to estimate, unbiasedly, treatment effects, given a strongly ignorable assignment mechanism. However, there are four main differences between propensity score studies performed with small versus large treated samples.

First, small treated samples require a substantially bigger group ratio (i.e., more control units per treated unit). This is not elucidated only by descriptive analyses, but also with results of ANOVA showing that the interaction between the group ratio and treated sample size is one of the influential factors in propensity score study with small samples.

Second, the number of observed covariates has a vast impact on small treated samples, with regard to the minimum required group ratio ($n_t : p$ being another influential factor), whereas the interaction $n_t : p$ is not an influential factor with moderately large treated samples.

Third, the number of observed covariates has an influence on the standard errors of small treated samples, whereas it barely affects standard errors of moderately large treated samples, i.e., when the number of observed covariates increases, treatment effect's standard errors, $SE_{A\hat{T}T}$, of small treated samples increase considerably (e.g., for $n_t = 8$ and $B^2 = 1$ the $SE_{A\hat{T}T}$ for $p = 10$ is 0.019, whereas for the $p = 30$ the $SE_{A\hat{T}T}$ is 0.027).

5.2 Simulation with True Propensity Scores

This section presents results of the simulated analyses performed on the data using true propensity scores. The purpose of such a study is purely theoretical because in real world examples of observational designs, true propensity scores are unknown. However, the results serve as a benchmark for the results obtained in the previous sections, when the simulated data reflect situations we would observe in practice, where we have to estimate propensity scores.

5.2.1 Analysis of Variance

The simulation is a 8 x 100 x 4 x 3 x 2 factorial design (19,200 cells) for the small treated sample study, and a 2 x 9 x 4 x 3 x 2 factorial design (432 cells) for the moderately large treated sample (Table 4.1). Due to the propensity scores being known (i.e., true propensity scores), there are no empty cells as in the previous simulation study with estimated propensity scores. The strong ignorability assumption is thus not violated, although it appeared to be in the previous simulation study, when propensity scores were estimated.

However, in order to compare these results to the results obtained with estimated propensity scores, the ANOVA is performed on both small treated sample study designs as in the previous section (Table 5.2), and in addition also on the full factorial design (19,200 cells), which includes group ratios, R, from 1 to 100.

The ANOVA is performed for both measures of quality of the procedure: the remaining bias, RB, and the variance ratio, VR, and it includes main effects and all the interaction effects (i.e., up to five-way interactions). The results are presented in Tables 5.12 and 5.13 where the factors known or estimable in the design phase of propensity score methods are sorted by their decreasing order of the mean sum of squares explained - MSS.

ANOVA – REMAINING BIAS

The ANOVA results of small treated sample studies with the true propensity scores are a bit different from what we obtained in the previous section with the estimated propensity scores. On the other hand, the ANOVA results of moderately large sample study with the true propensity scores are comparable to those obtained in the previous section with the estimated propensity scores.

The most influential factors for **small treated sample** studies with true propensity scores are initial squared bias, B^2, group ratio, R, treated sample size, n_t, and a two-way interaction between the group ratio and initial squared bias, $R:B^2$. Although the orders of the most influential factors do not perfectly match between the three small treated sample studies, due to the slight differences in the factorial designs of the studies, the most important – the type of the most influential factors – matches across all the simulation studies.

The most influential factors for the **moderately large treated sample** study are: group ratio, R, initial imbalance, B^2, the interaction between group ratio and initial imbalance, $R:B^2$, and the treated sample size, n_t.

Table 5.12: ANOVA table for Small treated sample study 1 and 2 with true propensity scores for the remaining bias measure, RB

Small treated sample study 1			**Small treated sample study 2**		
n_t	{8, 10, 15, 20, 25, 30, 50, 100}		n_t	{20, 25, 30, 50, 100}	
R	{13:100}		R	{2:100}	
Factor	DF	MSS	Factor	DF	MSS
B	2	0.40	B	2	0.81
n	7	0.05	R	98	0.27
R	87	0.02	n	4	0.03
n:B	14	0.01	R:B	196	0.03
method	1	0.00	method	1	0.01
R:B	174	0.00	n:B	8	0.00
p:B	6	0.00	R:n	392	0.00
method:B	2	0.00	p:B	6	0.00
R:n	609	0.00	p	3	0.00
p	3	0.00	R:method	98	0.00
R:p	261	0.00	n:method	4	0.00
n:p	21	0.00	method:B	2	0.00
R:method	87	0.00	R:p	294	0.00
n:method	7	0.00	n:p	12	0.00
p:method	3	0.00	p:method	3	0.00
R:n:p	1827	0.00	R:n:p	1176	0.00
R:n:method	609	0.00	R:n:method	392	0.00
R:p:method	261	0.00	R:p:method	294	0.00
n:p:method	21	0.00	n:p:method	12	0.00
R:n:B	1218	0.00	R:n:B	784	0.00
R:p:B	522	0.00	R:p:B	588	0.00
n:p:B	42	0.00	n:p:B	24	0.00
R:method:B	174	0.00	R:method:B	196	0.00
n:method:B	14	0.00	n:method:B	8	0.00
p:method:B	6	0.00	p:method:B	6	0.00
R:n:p:method	1827	0.00	R:n:p:method	1176	0.00
R:n:p:B	3654	0.00	R:n:p:B	2352	0.00
R:n:method:B	1218	0.00	R:n:method:B	784	0.00
R:p:method:B	522	0.00	R:p:method:B	588	0.00
n:p:method:B	42	0.00	n:p:method:B	24	0.00
R:n:p:method:B	3654	0.00	B	2	0.81

n - treated sample size, n_t; R – group ratio, R ; p – number of observed covariates, p; method – the matching algorithm used; B – initial squared bias; DF – degrees of freedom; MSS – mean sum of squares explained.

Table 5.13: ANOVA table for Moderately large treated sample study with true propensity scores for the remaining bias measure, RB

Small treated sample study 3			**Moderately large treated sample study**		
n_t	{8, 10, 15, 20, 25, 30, 50, 100}		n_t	{200, 500}	
R	{1:100}		R	{1:9}	
Factor	DF	MSS	Factor	DF	MSS
R	99	2.35	R	8	4.53
B	2	2.27	B	2	1.68
n	7	0.14	R:B	16	0.11
R:B	198	0.09	n	1	0.00
method	1	0.01	p	3	0.00
n:B	14	0.01	method	1	0.00
R:n	693	0.00	R:n	8	0.00
p:B	6	0.00	R:p	24	0.00
p	3	0.00	n:p	3	0.00
R:method	99	0.00	R:method	8	0.00
n:method	7	0.00	n:method	1	0.00
method:B	2	0.00	p:method	3	0.00
n:p	21	0.00	n:B	2	0.00
n:p:B	42	0.00	p:B	6	0.00
R:p	297	0.00	method:B	2	0.00
p:method	3	0.00	R:n:p	24	0.00
R:n:p	2079	0.00	R:n:method	8	0.00
R:n:method	693	0.00	R:p:method	24	0.00
R:p:method	297	0.00	n:p:method	3	0.00
n:p:method	21	0.00	R:n:B	16	0.00
R:n:B	1386	0.00	R:p:B	48	0.00
R:p:B	594	0.00	n:p:B	6	0.00
R:method:B	198	0.00	R:method:B	16	0.00
n:method:B	14	0.00	n:method:B	2	0.00
p:method:B	6	0.00	p:method:B	6	0.00
R:n:p:method	2079	0.00	R:n:p:method	24	0.00
R:n:p:B	4158	0.00	R:n:p:B	48	0.00
R:n:method:B	1386	0.00	R:n:method:B	16	0.00
R:p:method:B	594	0.00	R:p:method:B	48	0.00
n:p:method:B	42	0.00	n:p:method:B	6	0.00
R:n:p:method:B	4158	0.00	R:n:p:method:B	48	0.00

n - treated sample size, n_t; R – group ratio, R ; p – number of observed covariates, p; method – the matching algorithm used; B – initial squared bias; DF – degrees of freedom; MSS – mean sum of squares explained.

The main difference in the most influential factors between the small and moderately large treated sample studies is the treated sample size, n_t, which is one of the most influential factors for the small treated sample studies but it is not considered as an influential factor in the moderately large treated sample study. Although this finding does provide some insights on the role of sample sizes in propensity score studies, we remain cautious in drawing a strong conclusion due to the fact that the factorial designs of the small treated sample studies and moderately large treated sample studies are not fully comparable (as displayed in Table 4.1). Further descriptive analyses might provide more details on such insights.

ANOVA – VARIANCE RATIO

The ANOVA results for the variance ratio, VR, are presented in Tables 5.14 and 5.15. The results are a bit different, with respect to the most influential factors (for small and moderately large treated sample studies), from the results of the analyses with the remaining bias, RB, as described below. The most influential factors in the small treated sample study are: all the factors identified as the most influential already in the ANOVA with the remaining bias, number of observed covariates, p, and a two-way interaction between the number of observed covariates and initial squared bias, $p:B^2$.

The differences in results between ANOVA with the remaining bias and ANOVA with the variance ratio, with respect to the most influential factors, are even bigger for the moderately large treated sample study. The most influential factors in ANOVA with the variance ratio are: initial squared bias, B^2, group ratio, R, a two-way interaction between the number of observed covariates and initial squared bias, $p:B^2$, a two-way interaction between the treated sample size and initial squared bias, $n_t:B^2$, number of observed covariates, p, method (i.e., nearest versus greedy matching algorithm), a two-way interaction between the method and initial squared bias, $method:B^2$, treated sample size, n_t, and a two-way interaction between the group ratio and treated sample size, $R:n_t$. The discussion on the obtained results with ANOVA analyses is presented in Section 5.2.3 together with findings obtained from descriptive analysis.

Table 5.14: ANOVA table for Small treated sample study 1 and 2 with true propensity scores for the variance ratio measure, VR

Small treated sample study 1			**Small treated sample study 2**		
n_t	{8, 10, 15, 20, 25, 30, 50, 100}		n_t	{20, 25, 30, 50, 100}	
R	{13:100}		R	{2:100}	
Factor	DF	MSS	Factor	DF	MSS
B	2	0.101	B	2	0.054
p:B	6	0.064	p:B	6	0.037
n	7	0.008	p	3	0.006
n:B	14	0.004	n:B	8	0.004
n:p:B	42	0.003	n	4	0.004
n:p	21	0.003	n:p:B	24	0.003
p	3	0.003	R	98	0.003
R	87	0.001	n:p	12	0.000
R:B	174	0.000	R:n	392	0.000
R:p:B	522	0.000	R:p:B	588	0.000
R:p	261	0.000	R:p	294	0.000
R:n	609	0.000	R:B	196	0.000
R:n:p	1827	0.000	R:n:p	1176	0.000
R:n:p:B	3654	0.000	R:n:B	784	0.000
R:n:B	1218	0.000	R:n:p:B	2352	0.000
n:method:B	14	0.000	method	1	0.000
n:method	7	0.000	method:B	2	0.000
R:method	87	0.000	p:method	3	0.000
p:method	3	0.000	p:method:B	6	0.000
R:p:method	261	0.000	n:method	4	0.000
n:p:method	21	0.000	n:method:B	8	0.000
R:method:B	174	0.000	n:p:method	12	0.000
p:method:B	6	0.000	n:p:method:B	24	0.000
R:p:method:B	522	0.000	R:method	98	0.000
n:p:method:B	42	0.000	R:n:method	392	0.000
method	1	0.000	R:p:method	294	0.000
R:n:method	609	0.000	R:method:B	196	0.000
R:n:p:method	1827	0.000	R:n:p:method	1176	0.000
R:n:method:B	1218	0.000	R:n:method:B	784	0.000
R:n:p:method:B	3654	0.000	R:p:method:B	588	0.000
method:B	2	0.000	R:n:p:method:B	2352	0.000

n - treated sample size, n_t; R – group ratio, R ; p – number of observed covariates, p; method – the matching algorithm used; B – initial squared bias; DF – degrees of freedom; MSS – mean sum of squares explained.

Table 5.15: ANOVA table for Small treated sample study 1 and 2 with true propensity scores for the variance ratio measure, VR

Small treated sample study 3			**Moderately large treated sample study**		
n_t	{8, 10, 15, 20, 25, 30, 50, 100}		n_t	{200, 500}	
R	{1:100}		R	{1:9}	
Factor	DF	MSS	Factor	DF	MSSx1000
B	2	0.124	B	2	0.39
p:B	6	0.077	R	8	0.24
n	7	0.062	p:B	6	0.21
R	99	0.059	n:B	2	0.19
R:n	693	0.006	p	3	0.13
n:B	14	0.004	method	1	0.05
n:p:B	42	0.004	method:B	2	0.04
p	3	0.003	n	1	0.04
n:p	21	0.003	R:n	8	0.04
R:B	198	0.001	n:p:B	6	0.03
R:p	297	0.000	R:n:p:B	48	0.03
R:p:B	594	0.000	R:B	16	0.02
R:n:p	2079	0.000	n:p:method	3	0.02
R:n:p:B	4158	0.000	n:p:method:B	6	0.02
R:n:B	1386	0.000	R:p:B	48	0.02
n:method:B	14	0.000	p:method	3	0.02
n:method	7	0.000	p:method:B	6	0.02
R:method	99	0.000	n:method	1	0.01
p:method	3	0.000	n:method:B	2	0.01
R:p:method	297	0.000	R:p	24	0.01
n:p:method	21	0.000	n:p	3	0.01
R:method:B	198	0.000	R:n:p	24	0.01
p:method:B	6	0.000	R:n:B	16	0.00
R:p:method:B	594	0.000	R:n:p:method	24	0.00
n:p:method:B	42	0.000	R:n:p:method:B	48	0.00
method	1	0.000	R:p:method	24	0.00
R:n:method	693	0.000	R:p:method:B	48	0.00
R:n:p:method	2079	0.000	R:method:B	16	0.00
R:n:method:B	1386	0.000	R:method	8	0.00
R:n:p:method:B	4158	0.000	R:n:method	8	0.00
method:B	2	0.000	R:n:method:B	16	0.00

n - treated sample size, n_t; R – group ratio, R ; p – number of observed covariates, p; method – the matching algorithm used; B – initial squared bias; DF – degrees of freedom; MSS – mean sum of squares explained.

5.2.2 Descriptive Analysis

The simulated data are analysed as explained in Section 4.3 and the results are presented in Tables 5.16 – 5.19. The first two tables present results of the propensity score matching with the greedy matching algorithm whereas the other two present results of the matching with the optimal matching algorithm.

The first column in each table presents the treated sample size, n_t, followed by the minimum required group ratio, R^*, (as defined in Section 4.3), the absolute value of the remaining bias, RB, the variance ratio of the propensity score logit between the treated and control group ($VR = s_t^2 / s_c^2$), the 99% confidence intervals for the estimated average treatment effect of the treated - $\widehat{ATT}$ and the simulation standard errors of the $\widehat{ATT}$.

The results show that the minimum required group ratio, R^*, is decreasing when the treated sample size, n_t, is increasing which means that with more treated units, we can have a smaller pool of control units, relative to the size of the treated sample to estimate, unbiasedly, treatment effects.

The number of observed covariates, p, does not have an impact on either the minimum required group ratio, or the standard errors of the treatment effect estimates. When we observe more covariates, for example, when the number of observed covariates increases from $p = 20$ to $p = 30$, the minimum required group ratio remains the same for all the treated samples. Moreover, such an increase in the observed covariates does not have an impact on the standard errors either.

Table 5.16: Minimum required group ratios for investigated treated samples for $p = 10$ and $p = 15$ - greedy matching algorithm *

$p = 10$						$p = 15$					
n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $A\hat{T}T$ based on the simulation SE	Simulation SE of $A\hat{T}T$	n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $A\hat{T}T$ based on the simulation SE	Simulation SE of $A\hat{T}T$
$\boldsymbol{B^2 = 0.5}$						$\boldsymbol{B^2 = 0.5}$					
8	4	0.129	1.40	[-0.03,0.06]	0.018	8	4	0.123	1.38	[-0.04,0.05]	0.018
10	4	0.113	1.35	[-0.03,0.04]	0.016	10	4	0.102	1.31	[-0.01,0.07]	0.016
15	3	0.129	1.34	[-0.04,0.03]	0.013	15	3	0.127	1.33	[-0.02,0.04]	0.013
20	3	0.112	1.30	[-0.03,0.03]	0.012	20	3	0.110	1.29	[-0.02,0.04]	0.011
25	3	0.103	1.28	[-0.03,0.02]	0.010	25	3	0.100	1.26	[-0.02,0.03]	0.010
30	3	0.102	1.28	[-0.02,0.03]	0.010	30	3	0.092	1.25	[-0.02,0.03]	0.009
50	3	0.076	1.21	[-0.02,0.02]	0.007	50	3	0.073	1.20	[-0.00,0.03]	0.007
100	3	0.148	1.34	[-0.01,0.02]	0.006	100	3	0.146	1.34	[-0.01,0.02]	0.006
200	2	0.138	1.33	[-0.02,0.01]	0.004	200	2	0.135	1.32	[-0.01,0.01]	0.004
500	2	0.131	1.31	[-0.00,0.02]	0.003	500	2	0.129	1.31	[-0.01,0.01]	0.003
$\boldsymbol{B^2 = 1}$						$\boldsymbol{B^2 = 1}$					
8	6	0.146	1.49	[-0.03,0.06]	0.018	8	6	0.145	1.49	[-0.07,0.03]	0.019
10	6	0.130	1.45	[-0.04,0.04]	0.015	10	6	0.127	1.41	[-0.04,0.04]	0.016
15	5	0.132	1.40	[-0.04,0.03]	0.012	15	5	0.129	1.39	[-0.04,0.03]	0.013
20	5	0.118	1.36	[-0.04,0.02]	0.011	20	5	0.116	1.34	[-0.02,0.04]	0.011
25	4	0.147	1.42	[-0.02,0.03]	0.010	25	4	0.147	1.42	[-0.02,0.04]	0.011
30	4	0.137	1.39	[-0.02,0.03]	0.010	30	4	0.138	1.39	[-0.02,0.03]	0.009
50	4	0.120	1.34	[-0.01,0.03]	0.007	50	4	0.121	1.35	[-0.02,0.02]	0.007
100	4	0.106	1.31	[-0.01,0.02]	0.005	100	4	0.105	1.31	[-0.01,0.02]	0.005
200	4	0.103	1.31	[-0.01,0.01]	0.004	200	4	0.098	1.29	[-0.01,0.01]	0.004
500	4	0.099	1.29	[-0.01,0.00]	0.002	500	4	0.095	1.28	[-0.01,0.00]	0.002
$\boldsymbol{B^2 = 1.5}$						$\boldsymbol{B^2 = 1.5}$					
8	9	0.149	1.53	[-0.02,0.07]	0.019	8	9	0.143	1.56	[-0.07,0.03]	0.018
10	9	0.132	1.45	[-0.03,0.06]	0.016	10	9	0.136	1.50	[-0.05,0.03]	0.016
15	8	0.137	1.48	[-0.03,0.03]	0.013	15	8	0.147	1.50	[-0.04,0.03]	0.013
20	7	0.144	1.47	[-0.02,0.04]	0.011	20	7	0.133	1.45	[-0.02,0.03]	0.011
25	7	0.138	1.44	[-0.03,0.02]	0.010	25	7	0.143	1.46	[-0.02,0.04]	0.010
30	7	0.126	1.40	[-0.02,0.03]	0.010	30	7	0.137	1.45	[-0.00,0.04]	0.009
50	6	0.141	1.44	[-0.01,0.03]	0.007	50	6	0.147	1.46	[-0.01,0.03]	0.007
100	6	0.128	1.40	[-0.02,0.01]	0.005	100	6	0.132	1.41	[-0.01,0.02]	0.005
200	6	0.121	1.38	[-0.01,0.01]	0.004	200	6	0.123	1.38	[-0.01,0.01]	0.004
500	6	0.118	1.36	[-0.00,0.01]	0.002	500	6	0.120	1.37	[-0.01,0.01]	0.002

$*$ n_t – treated samples size; RB – the remaining bias; $s^2{}_{ps_t}/s^2{}_{ps_c}$ – variance ratio (VR); R^* - the minimum required group ratio, for each investigated treated sample which satisfies: $\mathrm{RB} < 0.15$ and $0.5 < VR < 2$. The 99% confidence intervals of the estimated ATT, calculated with standard errors (i.e., simulation SE of $A\hat{T}T$) designed by using the standard deviation of treatment effect estimates across 1,000 iterations, $\mathrm{SE}_{\mathrm{A\hat{T}T}} = \mathrm{s}_{\mathrm{A\hat{T}T}}/\sqrt{1000}$, show that the true value of the treatment effect ($\tau = 0$) is within the interval boundaries of the estimated treatment effect. This indicates that treatment effect estimates are unbiased (i.e., selection bias is successfully removed).

Table 5.17: Minimum required group ratios for investigated treated samples for $p = 20$ and $p = 30$ - greedy matching algorithm*

$p = 20$						$p = 30$					
n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{AT}T$ based on the simulation SE	Simulation SE of $\hat{AT}T$	n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{AT}T$ based on the simulation SE	Simulation SE of $\hat{AT}T$
$\boldsymbol{B^2 = 0.5}$						$\boldsymbol{B^2 = 0.5}$					
8	4	0.131	1.39	[-0.07,0.02]	0.019	8	4	0.133	1.39	[-0.05,0.04]	0.019
10	4	0.112	1.33	[-0.09,-0.01]	0.016	10	4	0.109	1.33	[-0.02,0.06]	0.015
15	3	0.132	1.33	[-0.04,0.02]	0.013	15	3	0.133	1.34	[-0.05,0.02]	0.014
20	3	0.114	1.29	[-0.03,0.03]	0.011	20	3	0.117	1.31	[-0.05,0.01]	0.011
25	3	0.103	1.26	[-0.02,0.03]	0.010	25	3	0.101	1.26	[-0.04,0.01]	0.010
30	3	0.097	1.26	[-0.01,0.04]	0.009	30	3	0.090	1.24	[-0.02,0.02]	0.009
50	3	0.074	1.20	[-0.02,0.02]	0.007	50	3	0.072	1.20	[-0.02,0.01]	0.007
100	3	0.148	1.35	[-0.02,0.01]	0.006	100	3	0.143	1.33	[-0.02,0.01]	0.005
200	2	0.134	1.32	[-0.01,0.01]	0.004	200	2	0.136	1.32	[-0.01,0.01]	0.004
500	2	0.130	1.31	[-0.01,0.01]	0.003	500	2	0.131	1.32	[-0.00 ,0.01]	0.003
$\boldsymbol{B^2 = 1}$						$\boldsymbol{B^2 = 1}$					
8	6	0.146	1.51	[-0.05,0.04]	0.018	8	6	0.143	1.50	[-0.06,0.04]	0.019
10	6	0.132	1.45	[-0.02,0.07]	0.017	10	6	0.130	1.42	[-0.04,0.04]	0.016
15	5	0.136	1.44	[-0.01,0.06]	0.013	15	5	0.127	1.37	[-0.03,0.04]	0.013
20	5	0.123	1.40	[-0.01,0.05]	0.011	20	5	0.112	1.34	[-0.03,0.03]	0.011
25	4	0.114	1.36	[-0.01,0.04]	0.010	25	4	0.140	1.40	[-0.02,0.03]	0.010
30	4	0.150	1.42	[-0.03,0.02]	0.010	30	4	0.135	1.38	[-0.03,0.02]	0.010
50	4	0.131	1.37	[-0.03,0.01]	0.007	50	4	0.121	1.36	[-0.02,0.02]	0.007
100	4	0.111	1.32	[-0.01,0.02]	0.005	100	4	0.106	1.31	[-0.01,0.02]	0.005
200	4	0.100	1.30	[-0.01,0.01]	0.004	200	4	0.102	1.30	[-0.02,0.00]	0.003
500	4	0.100	1.30	[-0.01,0.01]	0.002	500	4	0.097	1.29	[-0.00,0.01]	0.002
$\boldsymbol{B^2 = 1.5}$						$\boldsymbol{B^2 = 1.5}$					
8	9	0.147	1.54	[-0.02,0.08]	0.019	8	9	0.150	1.55	[-0.06,0.03]	0.018
10	9	0.138	1.50	[-0.02,0.06]	0.016	10	9	0.136	1.49	[-0.06,0.02]	0.016
15	8	0.135	1.46	[-0.03,0.04]	0.013	15	8	0.138	1.45	[-0.04,0.03]	0.013
20	7	0.147	1.48	[-0.02,0.04]	0.012	20	7	0.148	1.47	[-0.04,0.02]	0.012
25	7	0.138	1.44	[-0.02,0.04]	0.011	25	7	0.139	1.43	[-0.03,0.02]	0.010
30	7	0.127	1.41	[-0.02,0.03]	0.009	30	7	0.134	1.42	[-0.02,0.03]	0.010
50	6	0.137	1.42	[-0.03,0.01]	0.007	50	6	0.145	1.44	[-0.02,0.02]	0.008
100	6	0.128	1.40	[-0.02,0.00]	0.005	100	6	0.132	1.40	[-0.01,0.01]	0.005
200	6	0.121	1.38	[-0.01,0.01]	0.004	200	6	0.121	1.38	[-0.01,0.01]	0.004
500	6	0.121	1.38	[-0.01,0.00]	0.002	500	6	0.119	1.37	[-0.01,0.01]	0.002

* n_t – treated samples size; RB – the remaining bias; $s^2{}_{ps_t}/s^2{}_{ps_c}$ – variance ratio (VR); R^* - the minimum required group ratio, for each investigated treated sample which satisfies: $\mathrm{RB} < 0.15$ and $0.5 < VR < 2$. The 99% confidence intervals of the estimated ATT, calculated with standard errors (i.e., simulation SE of $\hat{AT}T$) designed by using the standard deviation of treatment effect estimates across 1,000 iterations, $\mathrm{SE}_{\mathrm{A\hat{T}T}} = \mathrm{s}_{\mathrm{A\hat{T}T}}/\sqrt{1000}$, show that the true value of the treatment effect ($\tau = 0$) is within the interval boundaries of the estimated treatment effect. This indicates that treatment effect estimates are unbiased (i.e., selection bias is successfully removed).

Table 5.18: Minimum required group ratios for investigated treated samples for $p = 10$ and $p = 15$ - optimal matching algorithm*

$p = 10$						$p = 15$					
n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of ATT	n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$
$\boldsymbol{B^2 = 0.5}$						$\boldsymbol{B^2 = 0.5}$					
8	4	0.112	1.39	[-0.04,0.05]	0.018	8	4	0.102	1.36	[-0.03,0.06]	0.017
10	3	0.139	1.42	[-0.05,0.03]	0.016	10	3	0.130	1.42	[-0.01,0.07]	0.015
15	3	0.108	1.32	[-0.03,0.03]	0.012	15	3	0.104	1.31	[-0.02,0.05]	0.012
20	3	0.093	1.28	[-0.03,0.03]	0.011	20	3	0.091	1.27	[-0.01,0.05]	0.010
25	3	0.085	1.26	[-0.04,0.01]	0.010	25	3	0.083	1.25	[-0.01,0.04]	0.009
30	3	0.086	1.26	[-0.02,0.03]	0.009	30	3	0.075	1.23	[-0.01,0.03]	0.008
50	3	0.063	1.19	[-0.02,0.02]	0.007	50	3	0.061	1.18	[-0.01,0.03]	0.007
100	2	0.139	1.33	[-0.01,0.02]	0.005	100	2	0.138	1.33	[-0.01,0.02]	0.005
200	2	0.135	1.32	[-0.01,0.00]	0.003	200	2	0.131	1.31	[-0.01,0.01]	0.003
500	2	0.130	1.31	[-0.00,0.01]	0.002	500	2	0.127	1.31	[-0.00,0.01]	0.002
$\boldsymbol{B^2 = 1}$						$\boldsymbol{B^2 = 1}$					
8	6	0.134	1.51	[-0.05,0.05]	0.018	8	6	0.136	1.49	[-0.07,0.03]	0.018
10	5	0.141	1.47	[-0.02,0.06]	0.016	10	5	0.148	1.49	[-0.04,0.04]	0.016
15	5	0.150	1.44	[-0.02,0.05]	0.013	15	5	0.118	1.37	[-0.03,0.03]	0.013
20	5	0.142	1.41	[-0.01,0.04]	0.011	20	5	0.145	1.42	[-0.02,0.03]	0.011
25	4	0.139	1.40	[-0.01,0.04]	0.010	25	4	0.137	1.40	[-0.02,0.03]	0.010
30	4	0.127	1.38	[-0.02,0.03]	0.009	30	4	0.128	1.37	[-0.02,0.02]	0.009
50	4	0.113	1.33	[-0.01,0.03]	0.007	50	4	0.114	1.33	[-0.02,0.02]	0.007
100	4	0.102	1.30	[-0.01,0.01]	0.005	100	4	0.101	1.30	[-0.01,0.01]	0.005
200	4	0.102	1.30	[-0.01,0.00]	0.003	200	4	0.096	1.29	[-0.01,0.01]	0.003
500	4	0.098	1.29	[-0.01,0.00]	0.002	500	4	0.095	1.28	[-0.01,0.00]	0.002
$\boldsymbol{B^2 = 1.5}$						$\boldsymbol{B^2 = 1.5}$					
8	9	0.141	1.52	[-0.03,0.06]	0.018	8	9	0.148	1.59	[-0.06,0.03]	0.018
10	9	0.145	1.54	[-0.04,0.04]	0.016	10	9	0.143	1.54	[-0.06,0.02]	0.016
15	8	0.147	1.50	[-0.02,0.04]	0.013	15	8	0.138	1.49	[-0.03,0.03]	0.013
20	7	0.138	1.46	[-0.02,0.04]	0.011	20	7	0.146	1.50	[-0.02,0.03]	0.011
25	7	0.132	1.44	[-0.03,0.02]	0.010	25	7	0.136	1.45	[-0.01,0.04]	0.010
30	7	0.146	1.46	[-0.02,0.03]	0.009	30	7	0.131	1.43	[-0.02,0.03]	0.009
50	6	0.136	1.43	[-0.01,0.03]	0.007	50	6	0.142	1.45	[-0.01,0.03]	0.007
100	6	0.126	1.39	[-0.02,0.01]	0.005	100	6	0.130	1.40	[-0.01,0.01]	0.005
200	6	0.120	1.38	[-0.01,0.01]	0.003	200	6	0.122	1.38	[-0.01,0.01]	0.003
500	6	0.118	1.36	[-0.00,0.01]	0.002	500	6	0.120	1.37	[-0.01,0.00]	0.002

* n_t – treated samples size; RB – the remaining bias; $s^2{}_{ps_t}/s^2{}_{ps_c}$ – variance ratio (VR); R^* - the minimum required group ratio, for each investigated treated sample which satisfies: $\mathrm{RB} < 0.15$ and $0.5 < VR < 2$. The 99% confidence intervals of the estimated ATT, calculated with standard errors (i.e., simulation SE of $\hat{ATT}$) designed by using the standard deviation of treatment effect estimates across 1,000 iterations, $\mathrm{SE}_{\hat{\mathrm{ATT}}} = \mathrm{s}_{\hat{\mathrm{ATT}}}/\sqrt{1000}$, show that the true value of the treatment effect ($\tau = 0$) is within the interval boundaries of the estimated treatment effect. This indicates that treatment effect estimates are unbiased (i.e., selection bias is successfully removed).

Table 5.19: Minimum required group ratios for investigated treated samples for $p = 20$ and $p = 30$ - optimal matching algorithm*

$p = 20$						$p = 30$					
n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$	n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$
$\boldsymbol{B^2 = 0.5}$						$\boldsymbol{B^2 = 0.5}$					
8	4	0.112	1.38	[-0.08,0.01]	0.018	8	4	0.117	1.38	[-0.05,0.04]	0.018
10	3	0.143	1.44	[-0.09,-0.01]	0.016	10	3	0.138	1.43	[-0.03,0.05]	0.016
15	3	0.108	1.30	[-0.04,0.03]	0.012	15	3	0.111	1.32	[-0.05,0.02]	0.013
20	3	0.094	1.27	[-0.03,0.02]	0.011	20	3	0.097	1.28	[-0.04,0.01]	0.010
25	3	0.086	1.24	[-0.02,0.02]	0.009	25	3	0.084	1.24	[-0.04,0.00]	0.009
30	3	0.081	1.24	[-0.01,0.03]	0.008	30	3	0.073	1.21	[-0.02,0.02]	0.009
50	3	0.062	1.18	[-0.01,0.02]	0.007	50	3	0.146	1.34	[-0.02,0.01]	0.007
100	2	0.140	1.34	[-0.02,0.00]	0.005	100	2	0.135	1.32	[-0.01,0.01]	0.005
200	2	0.130	1.31	[-0.01,0.01]	0.004	200	2	0.132	1.32	[-0.01,0.01]	0.003
500	2	0.129	1.31	[-0.01,0.00]	0.002	500	2	0.130	1.31	[-0,00,0.01]	0.002
$\boldsymbol{B^2 = 1}$						$\boldsymbol{B^2 = 1}$					
8	6	0.131	1.49	[-0.05,0.04]	0.017	8	6	0.132	1.49	[-0.06,0.04]	0.018
10	5	0.150	1.52	[-0.01,0.08]	0.016	10	5	0.148	1.50	[-0.04,0.04]	0.016
15	5	0.124	1.43	[-0.02,0.04]	0.013	15	5	0.113	1.36	[-0.03,0.03]	0.012
20	5	0.111	1.38	[-0.01,0.05]	0.011	20	5	0.140	1.41	[-0.02,0.03]	0.011
25	4	0.143	1.44	[-0.01,0.04]	0.010	25	4	0.129	1.38	[-0.03,0.02]	0.010
30	4	0.136	1.40	[-0.01,0.03]	0.009	30	4	0.125	1.37	[-0.02,0.03]	0.009
50	4	0.118	1.35	[-0.02,0.01]	0.007	50	4	0.114	1.34	[-0.02,0.02]	0.007
100	4	0.106	1.31	[-0.02,0.01]	0.005	100	4	0.102	1.30	[-0.01,0.01]	0.005
200	4	0.098	1.30	[-0.01,0.01]	0.004	200	4	0.101	1.30	[-0.01,0.01]	0.003
500	4	0.095	1.28	[-0.01,0.01]	0.002	500	4	0.094	1.28	[-0.01,0.00]	0.002
$\boldsymbol{B^2 = 1.5}$						$\boldsymbol{B^2 = 1.5}$					
8	9	0.138	1.53	[-0.02,0.08]	0.018	8	9	0.138	1.54	[-0.06,0.03]	0.017
10	9	0.150	1.56	[-0.03,0.05]	0.015	10	9	0.144	1.53	[-0.05,0.03]	0.015
15	8	0.148	1.51	[-0.03,0.03]	0.013	15	8	0.129	1.44	[-0.04,0.03]	0.013
20	7	0.141	1.47	[-0.02,0.04]	0.011	20	7	0.140	1.46	[-0.03,0.02]	0.011
25	7	0.132	1.43	[-0.01,0.04]	0.010	25	7	0.132	1.42	[-0.03,0.02]	0.010
30	7	0.148	1.47	[-0.02,0.03]	0.009	30	7	0.128	1.41	[-0.02,0.03]	0.009
50	6	0.133	1.42	[-0.02,0.02]	0.007	50	6	0.142	1.44	[-0.01,0.03]	0.007
100	6	0.126	1.40	[-0.02,0.01]	0.005	100	6	0.130	1.40	[-0.01,0.02]	0.005
200	6	0.121	1.38	[-0.01,0.01]	0.003	200	6	0.120	1.37	[-0.01,0.01]	0.003
500	6	0.120	1.38	[-0.01,0.00]	0.002	500	6	0.119	1.37	[-0.01,0.00]	0.002

* n_t – treated samples size; RB – the remaining bias; $s^2{}_{ps_t}/s^2{}_{ps_c}$ – variance ratio (VR); R^* - the minimum required group ratio, for each investigated treated sample which satisfies: $\mathrm{RB} < 0.15$ and $0.5 < VR < 2$. The 99% confidence intervals of the estimated ATT, calculated with standard errors (i.e., simulation SE of $\hat{ATT}$) designed by using the standard deviation of treatment effect estimates across 1,000 iterations, $\mathrm{SE}_{\mathrm{A\hat{T}T}} = \mathrm{s}_{\mathrm{A\hat{T}T}}/\sqrt{1000}$, show that the true value of the treatment effect ($\tau = 0$) is within the interval boundaries of the estimated treatment effect. This indicates that treatment effect estimates are unbiased (i.e., selection bias is successfully removed).

Our initial descriptive analyses findings are furthermore supported with some graphical depictions. The aim is to display how strongly treated sample sizes depend on the adequate size of the control group, in order to remove a sufficient amount of selection bias (i.e., $RB(R) < 0.15$ and $VR(R) < 2$) before using an additional covariate regression adjustment and estimating ATT.

In the beginning of the descriptive-analysis section we noted that the minimum required group ratio, R^*, decreases when the size of a treated sample, n_t, increases. The relationship between the minimum required group ratio and treated sample size, together with the number of observed covariates is presented in Figure 5.4.

Figure 5.4: Relationship between the number of observed covariates, p, and the minimum required group ratio, R^*, for different treated samples (presented with lines) when $B^2 = 1$ (other B^2 produce very similar depictions).

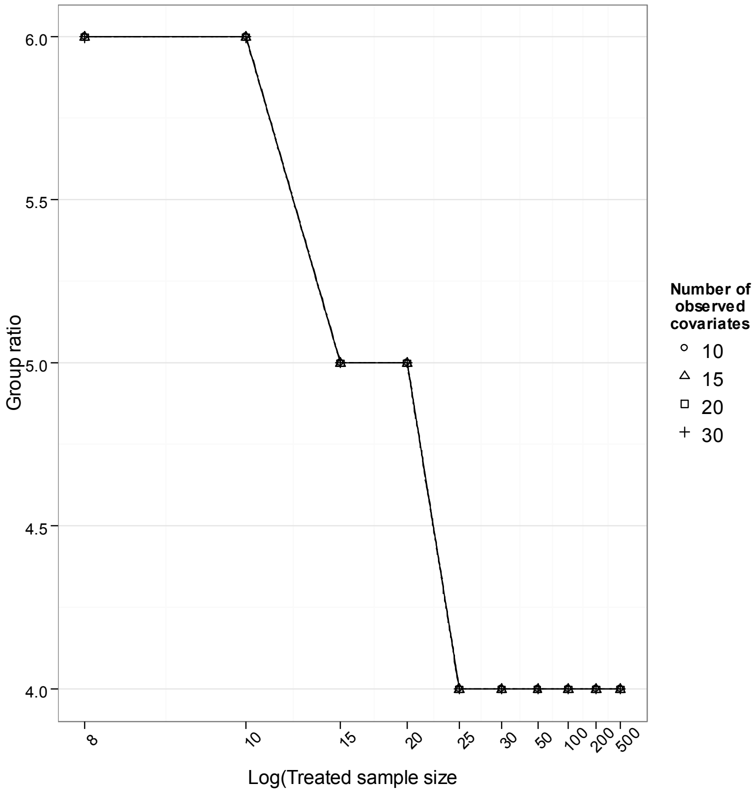
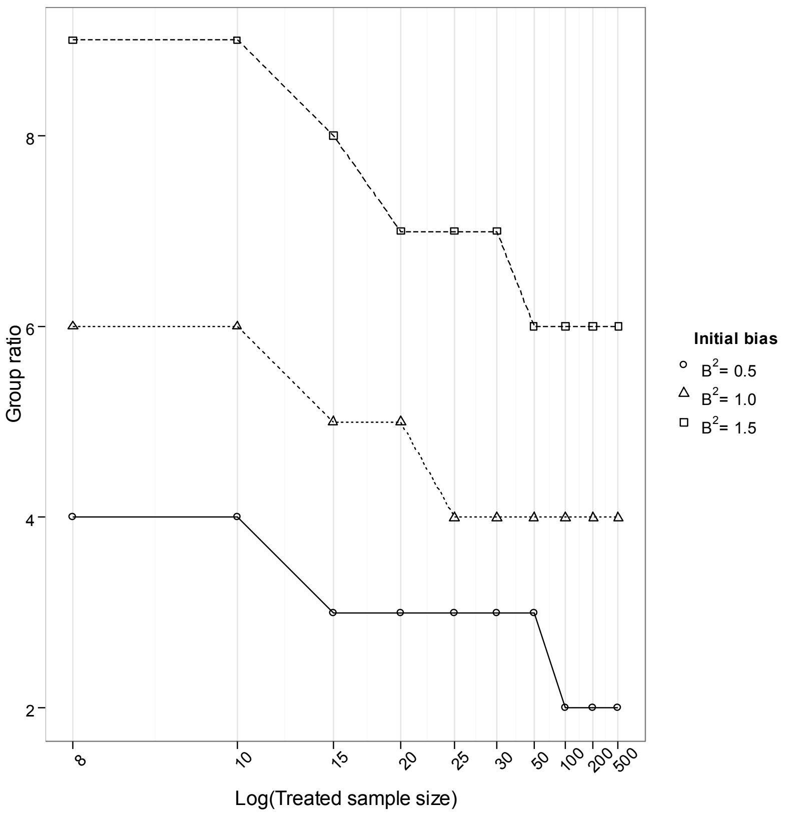


Figure 5.4 shows that the minimum required group ratio increases when the size of the treated sample decreases. Additionally, the figure also shows that the minimum required group ratio in propensity score studies with true propensity scores does not depend on the number of observed covariates (lines which present each covariate set are on the top of each other, thus we can see only one line representing the relation between the treated sample size and minimum required group ratio).

Moreover, the level of initial imbalance in a study design does have an influence on the minimum required group ratio. For example, with an initial squared bias of $B^2 = 0.5$, results show that a treated sample of $n_t = 8$ requires at least a group ratio of $R^* = 4$. If the initial squared bias increases to $B^2 = 1$ or $B^2 = 1.5$, a group ratio of at least 6 and 7, respectively, is required. The relation between different levels of initial bias and the minimum required group ratios for the treated sample, n_t, is shown in Figure 5.5, where the lines represent different initial squared biases, i.e., B^2 of 0.5, 1 and 1.5.

Figure 5.5: Relationship between the minimum required group ratio, R^*, and initial squared biases, B^2, for different sizes of treated samples.



Also in this study, with true propensity scores, we investigated possible differences in results of propensity score studies when matching is performed with the greedy or with the optimal matching algorithm. The evaluation is again performed by comparing mean-squared-errors (MSE) of $\widehat{ATT}$ obtained when using different matching algorithms (i.e., greedy and optimal) for all the treated samples, number of observed covariates, and initial squared biases.

The results are presented in Table 5.20 and show that in most of the cases, the optimal matching algorithm does perform slightly better with respect to MSE than greedy matching algorithm; however, the differences are not significant.

Table 5.20: MSE of $\widehat{ATT}$ when matching is performed with greedy or optimal algorithms

		$B^2 = 0.5$		$B^2 = 1$		$B^2 = 1.5$		$MSE_{Greedy} - MSE_{optimal}$		
n_t	p	Greedy	Optimal	Greedy	Optimal	Greedy	Optimal	$B^2 = 0.5$	$B^2 = 1$	$B^2 = 1.5$
8	10	0.35	0.32	0.35	0.32	0.32	0.33	0.03	0.00	0.02
10	10	0.27	0.23	0.26	0.25	0.25	0.26	0.02	-0.02	0.00
15	10	0.18	0.16	0.16	0.16	0.18	0.16	0.02	-0.02	0.00
20	10	0.14	0.12	0.13	0.12	0.12	0.12	0.02	0.00	0.01
25	10	0.10	0.11	0.10	0.09	0.10	0.09	0.01	0.01	0.01
30	10	0.09	0.10	0.09	0.08	0.07	0.09	0.01	0.03	0.00
50	10	0.05	0.05	0.05	0.05	0.05	0.05	0.00	0.00	0.00
100	10	0.03	0.03	0.03	0.03	0.02	0.02	0.00	0.01	0.01
200	10	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00
500	10	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.01
8	15	0.35	0.35	0.34	0.30	0.33	0.32	0.05	0.02	0.02
10	15	0.25	0.27	0.26	0.23	0.26	0.25	0.02	0.01	0.01
15	15	0.18	0.18	0.17	0.15	0.16	0.16	0.03	0.02	0.01
20	15	0.12	0.13	0.12	0.11	0.12	0.12	0.01	0.01	0.00
25	15	0.10	0.11	0.11	0.09	0.10	0.10	0.01	0.01	0.01
30	15	0.08	0.09	0.09	0.07	0.08	0.08	0.01	0.01	0.01
50	15	0.05	0.05	0.05	0.04	0.05	0.05	0.01	0.00	0.00
100	15	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01
200	15	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00
500	15	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.01
8	20	0.36	0.33	0.35	0.33	0.31	0.34	0.03	0.02	0.01
10	20	0.25	0.30	0.26	0.26	0.28	0.24	-0.01	0.02	0.02
15	20	0.18	0.17	0.18	0.15	0.16	0.17	0.03	0.01	0.01
20	20	0.12	0.12	0.14	0.11	0.11	0.13	0.01	0.01	0.01
25	20	0.10	0.10	0.11	0.09	0.10	0.11	0.01	0.00	0.00
30	20	0.08	0.10	0.08	0.07	0.08	0.08	0.01	0.02	0.00
50	20	0.05	0.05	0.06	0.04	0.04	0.05	0.01	0.01	0.01
100	20	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01
200	20	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00
500	20	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.01

Table 5.20 (continues): MSE of $\widehat{ATT}$ when matching is performed with greedy or optimal algorithms

		$B^2 = 0.5$		$B^2 = 1$		$B^2 = 1.5$		$MSE_{Greedy} - MSE_{optimal}$		
n_t	p	Greedy	Optimal	Greedy	Optimal	Greedy	Optimal	$B^2 = 0.5$	$B^2 = 1$	$B^2 = 1.5$
8	30	0.38	0.38	0.32	0.35	0.34	0.29	0.03	0.04	0.03
10	30	0.24	0.27	0.25	0.25	0.26	0.24	-0.01	0.01	0.01
15	30	0.19	0.16	0.18	0.17	0.15	0.17	0.02	0.01	0.01
20	30	0.13	0.13	0.14	0.11	0.12	0.13	0.02	0.01	0.01
25	30	0.10	0.10	0.11	0.09	0.09	0.10	0.01	0.01	0.01
30	30	0.08	0.10	0.09	0.08	0.09	0.09	0.00	0.01	0.00
50	30	0.05	0.06	0.06	0.05	0.05	0.05	0.00	0.01	0.01
100	30	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01
200	30	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00
500	30	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01

5.2.3 Summary

The descriptive analyses complement the ANOVA results, which reveal the most influential factors in propensity score studies with true propensity scores for small and moderately large treated samples.

The most influential factors revealed by performing ANOVA for small (i.e., B^2, R, n_t and a two-way interaction $R:B^2$) and moderately large treated sample studies (i.e., R, B^2, and the two-way interaction, $R:B^2$) accord well with the findings of descriptive analyses.

Both treated sample sizes (i.e., small and moderately large) remove a sufficient amount of selection bias from observational studies to estimate, unbiasedly, treatment effects given a strongly ignorable assignment mechanism. However, there are two main differences between propensity score studies performed with small versus large treated samples: (i) small treated samples require slightly bigger group ratios (i.e., more control units per treated unit); (ii) moderately large treated samples produce much more precise treatment effect estimates.

5.3 Discussion on Simulation Results: True versus Estimated Propensity Score

Our comparison between using the true propensity scores and the estimated propensity scores starts by comparing the findings of the analyses performed by ANOVA (where the dependent variable is the remaining bias, RB, and the independent variables are factors known or estimable in the design phase of a propensity score study: treated sample size, n_t, number of observed covariates, p, group ratio, R, and initial squared bias, B^2)[15] and continues with the comparison of the results of the descriptive analyses for the simulation study with true and estimated propensity scores.

Although the ANOVA results for the moderately large treated samples, with true and estimated propensity scores are fully comparable, there are some major differences in the ANOVA results between true versus using estimated propensity scores for small treated sample studies. Table 5.21 presents ANOVA results for fully comparable factorial designs of the Small treated sample study 1 with estimated and true propensity scores. All the factors that have MSS values larger than zero are displayed to two decimal places, sorted by decreasing order of MSS.

The main difference between the two studies (i.e., the estimated versus true propensity score study) is in the factors that appear as the most influential. The ANOVA results of the most influential factors show that the number of observed covariates, p, is an influential factor when using estimated propensity scores, whereas the number of observed covariates does not appear as an influential factor when using true propensity scores. Furthermore, as we can see from Table 5.21, there are some more factors that appear as influential in studies with estimated propensity scores, but are not influential in studies with true propensity scores (e.g., $n_t : p$, $p : B^2$, $method$, $R : n_t$, $R : B$, $R : p$ and $n_t : p : B^2$).

[15] The ANOVA for VR produces comparable results to the ANOVA for RB for both: true and estimated propensity score study.

Table 5.21: Comparison of ANOVA - RB results for the **Small treated sample study 1** with estimated and true propensity scores for $n_t =$ {8, 10, 15, 20, 25, 30, 50, 100} and $R = \{13:100\}$

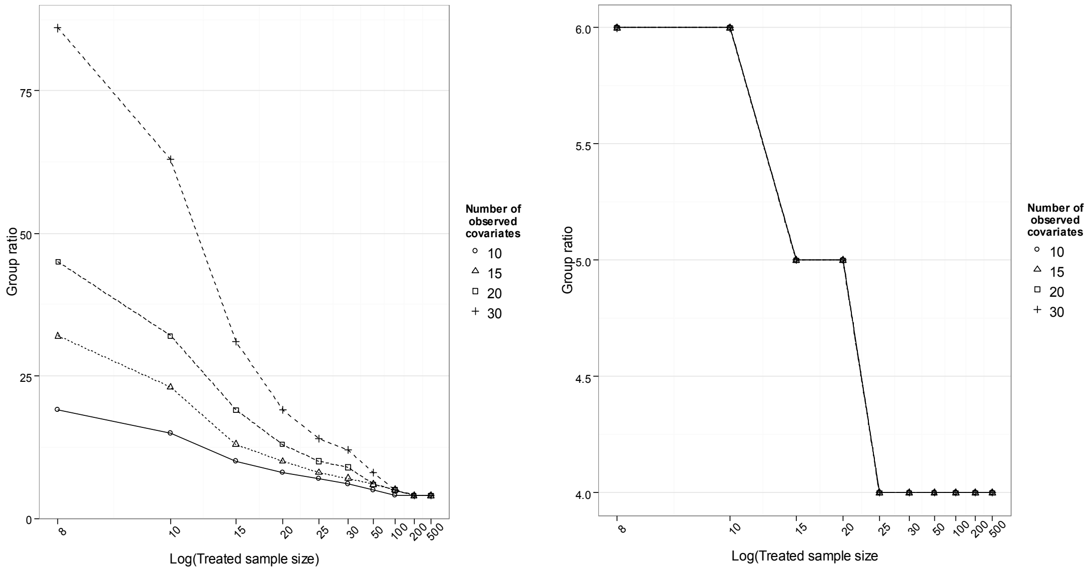
Estimated propensity scores			***True propensity score***		
Factor	DF	MSS	Factor	DF	MSS
n	7	4,60	B	2	0,40
p	3	2,53	n	7	0,05
B	2	2,37	R	87	0,02
n:p	21	0,32	n:B	14	0,01
R	87	0,18			
n:B	14	0,09			
p:B	6	0,03			
method	1	0,02			
R:n	609	0,01			
R:B	174	0,01			
R:p	261	0,01			
n:p:B	42	0,01			

n – treated sample size, n_t; R – group ratio, R ; p – number of observed covariates, p; method – the matching algorithm used; B – initial squared bias; DF – degrees of freedom; MSS – mean sum of squares explained.

In small sample studies with estimated propensity scores, the sample sizes of the treated and control groups (the latter is in our study defined based on the group ratio) play a very important role, when the number of observed covariates increases. In cases of very small treated samples and a larger number of observed covariates, a propensity score study requires substantially larger pools of control units for the estimated propensity scores to be effective balancing scores, and hence, being able to balance an observational study design (i.e., to remove selection bias). Therefore, factors such as $R:p$ and $n_t:p$ which are influential in studies with estimated propensity scores, do not have an influence in studies with true propensity scores when the propensity scores are known thus they are automatically effective balancing scores. For the same reason, other factors that are influential in studies with estimated propensity scores such as, $p:B^2$, $method$, $R:n_t$, $R:B$, and $n_t:p:B^2$, do not appear to be influential in studies with true propensity scores.

The descriptive analyses of our studies complement the findings of ANOVA. The minimum required group ratio, in the study with estimated propensity scores, increases severely with an increasing number of observed covariates for small treated sample; however, this is not the case in the study with true propensity scores (Figure 5.6) because propensity scores are known. Again, the reason for this distinctness is in the estimation process of propensity scores (i.e., to obtain effective balancing scores).

Figure 5.6: Comparison of the number of observed covariates, p, (presented with lines) versus the treated sample sizes and their correspondingly required minimum group ratios with $B^2 = 1$. Study with estimated propensity scores (left) and study with true propensity scores (right).



For the graph on the right: lines which present each covariate set are on the top of each other; thus, we can only see one line representing the relation between the treated sample size and minimum required group ratio – the number of observed covariates does not have an impact on the minimum required group ratio when true propensity scores are used.

As mentioned in the beginning of this section, there are no differences in the most influential factors obtained with ANOVA for moderately large treated samples between the studies using estimated propensity scores (Table 5.4) versus true propensity scores (Table 5.13). This indicates that different sets of observed covariates, p=10, 15, 20, 30, do not materially affect the estimation of propensity scores with moderately large treated sample (i.e., n_t of 200 and 500) as they do in studies with small treated samples (i.e., sample sizes are large enough to estimate effectively propensity scores that are precise balancing scores).

5.4 Extensions

Two simulation study extensions are carried out with true propensity scores, for only one level of the initial squared bias factor: $B^2 = 1$ and the greedy matching algorithm. The aim of these simulation extensions is to get an idea of some other factors that might have an impact on propensity score studies with small treated samples. Only descriptive analyses are provided for the two simulation study extensions.

5.4.1 Stronger Correlation between the Outcome Variable and Covariates

This simulation extension is conducted by using true propensity scores and considers a stronger correlation structure ($R^2 = 0.78$) between the outcome variable and the observed covariates, in comparison to our main simulation study with the covariate structure between the outcome variable and the observed covariates of $R^2 = 0.51$. The results are displayed in Table 5.22.

The first column in each table presents the treated sample size, n_t, followed by the minimum required group ratio, R^*, (as defined in Section 4.3.2), the absolute value of the remaining bias, RB, the variance ratio of the propensity score logit between the treated and control group ($VR = s_t^2 / s_c^2$), the confidence intervals of the estimated average treatment effect of the treated - $\widehat{ATT}$ and the simulation standard errors of $\widehat{ATT}$.

Table 5.22: The minimum required group ratios for investigated treated samples and observed covariates p =10, 15, 20, 30 with greedy matching algorithm*

$\boldsymbol{B^2 = 1}$											
n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$	n_t	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$
			$\boldsymbol{p = 10}$						$\boldsymbol{p = 15}$		
8	6	0,132	1,48	[-0.06,0.03]	0,019	8	6	0,145	1,49	[-0.07,0.03]	0,019
10	6	0,136	1,49	[-0.07,0.02]	0,016	10	6	0,127	1,41	[-0.04,0.04]	0,016
15	5	0,137	1,43	[-0.05,0.02]	0,013	15	5	0,129	1,39	[-0.04,0.03]	0,013
20	5	0,121	1,37	[-0.05,0.02]	0,012	20	5	0,116	1,34	[-0.02,0.04]	0,011
25	4	0,113	1,35	[-0.04,0.01]	0,011	25	4	0,147	1,42	[-0.02,0.04]	0,011
30	4	0,145	1,41	[-0.02,0.03]	0,009	30	4	0,138	1,39	[-0.02,0.03]	0,009
50	4	0,124	1,36	[-0.03,0.01]	0,008	50	4	0,121	1,35	[-0.02,0.02]	0,007
100	4	0,108	1,31	[-0.01,0.02]	0,005	100	4	0,105	1,31	[-0.01,0.02]	0,005
200	4	0,100	1,29	[-0.01,0.01]	0,004	200	4	0,098	1,29	[-0.01,0.01]	0,004
500	4	0,096	1,29	[-0.01,0.00]	0,002	500	4	0,095	1,28	[-0.01,0.03]	0,002
			$\boldsymbol{p = 20}$						$\boldsymbol{p = 30}$		
8	6	0,139	1,46	[-0.09,0.00]	0,018	8	6	0,146	1,49	[-0.03,0.07]	0,020
10	6	0,127	1,41	[-0.08,0.00]	0,016	10	6	0,128	1,43	[-0.04,0.05]	0,017
15	5	0,128	1,39	[-0.05,0.02]	0,013	15	5	0,133	1,40	[-0.01,0.06]	0,014
20	5	0,115	1,34	[-0.04,0.02]	0,011	20	5	0,113	1,34	[-0.02,0.04]	0,012
25	4	0,145	1,41	[-0.03,0.02]	0,011	25	4	0,143	1,40	[-0.02,0.04]	0,011
30	4	0,140	1,39	[-0.02,0.03]	0,010	30	4	0,135	1,38	[-0.03,0.02]	0,010
50	4	0,120	1,34	[-0.02,0.02]	0,007	50	4	0,121	1,36	[-0.02,0.02]	0,007
100	4	0,108	1,32	[-0.02,0.01]	0,005	100	4	0,106	1,31	[-0.01,0.02]	0,005
200	4	0,100	1,30	[-0.01,0.01]	0,004	200	4	0,102	1,30	[-0.02,0.00]	0,004
500	4	0,096	1,29	[-0.00,0.01]	0,002	500	4	0,099	1,29	[-0.01,0.01]	0,002

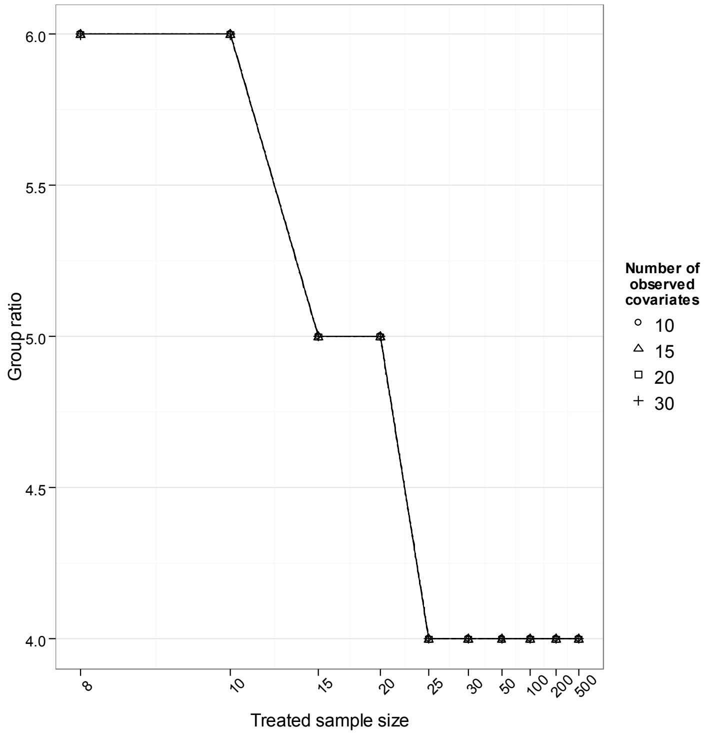
$*$ The 99% confidence intervals of the estimated ATT, calculated with standard errors (i.e., simulation SE of $\hat{ATT}$) designed by using the standard deviation of treatment effect estimates across 1,000 iterations, $\mathrm{SE_{ATT}} = \mathrm{s_{ATT}}/\sqrt{1000}$, show that the true value of the treatment effect ($\tau = 0$) is within the interval boundaries of the estimated treatment effect. This indicates that treatment effect estimates are unbiased (i.e., selection bias is successfully removed).

The results show exactly the same structure of the minimum required group ratios for the investigated treated samples and covariate sets as in our main simulation study, performed with true propensity scores and a weaker correlation structure (Tables 5.16 - 5.17).

Furthermore, Figure 5.7 shows the required group ratios for the investigated treated sample sizes and covariate sets. The depiction is essentially the same as the one presented in Section 5.2.2 – descriptive analyses of the simulation study with true propensity scores.

Additionally, we also obtained similar values of treatment effect simulation standard errors and the treatment effect confidence intervals. Based on that, we can conclude that the strength of the correlation structure between observed covariates and the outcome variable does not play a major role in propensity score studies – neither with small, nor with moderately large treated samples.

Figure 5.7: Comparison of the number of observed covariates, p, versus the treated sample sizes and their correspondingly required minimum group ratios with $B^2 = 1$ (the lines denoting observed covariates are on the top of each other).



5.4.2 Binary outcome variable

The following simulation extension uses a binary outcome variable. The propensity score matching is conducted with true propensity scores and performed with the greedy matching algorithm. The additional propensity score regression adjustment, for removing the residual selection bias, was not performed because the investigation is beyond the scope of this thesis.

Thus, we present results of the treatment effects estimates obtained from matched pairs as a difference in proportions between treated and control group. We present the results of the binary outcome simulation study together with the results of the main simulation study with continuous outcome, where additional regression adjustment was not performed in order to show consistency in obtained results between these two simulations (Table 5.23).

The first column in each table presents the treated sample size, n_t, followed by the minimum required group ratio, R^*, (as defined I Section 4.3.2), the absolute value of the remaining bias, RB, the variance ratio of the propensity score logit between the treated and control group ($VR = s_t^2 / s_c^2$), the 99% confidence intervals of the estimated average treatment effect of the treated - $A\widehat{TT}$ and the simulation standard errors of $A\widehat{TT}$. The confidence intervals of $A\widehat{TT}$, calculated with standard errors (i.e., simulation SE of $A\widehat{TT}$) using the standard deviation of treatment effect estimates across 1,000 iterations, $\mathrm{SE}_{\mathrm{ATT}} = \mathrm{s}_{\mathrm{ATT}} / \sqrt{1000}$. If the true value of the treatment effect ($\tau = 0$) is within the interval boundaries of $A\widehat{TT}$ than our $A\widehat{TT}$ is unbiased or approximately unbiased.

Table 5.23: The minimum required group ratios for investigated treated samples, initial squared bias, $B^2 = 1$, and observed covariates without additional propensity score regression adjustment, when the outcome is continuous or binary

Continuous outcome				$\hat{ATT}$ without additional regression adjustment		Binary outcome				$\hat{ATT}$ without additional regression adjustment	
n_t	R*	RB	$\frac{s^2_{ps_t}}{s^2_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$	n_t	R*	RB	$\frac{s^2_{ps_t}}{s^2_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE	Simulation SE of $\hat{ATT}$
$p = 10$						$p = 10$					
8	6	0.146	1.49	[0.04,0.12]	0.015	8	6	0,144	1,51	[-0.01,0.03]	0,007
10	6	0.130	1.45	[0.03,0.10]	0.013	10	6	0,126	1,42	[-0.01,0.03]	0,007
15	5	0.132	1.40	[0.03,0.09]	0.011	15	5	0,124	1,37	[-0.01,0.02]	0,005
20	5	0.118	1.36	[0.03,0.08]	0.010	20	5	0,111	1,34	[-0.00,0.02]	0,005
25	4	0.147	1.42	[0.05,0.10]	0.009	25	4	0,146	1,42	[0.01,0.03]	0,004
30	4	0.137	1.39	[0.05,0.09]	0.008	30	4	0,143	1,41	[0.01,0.03]	0,004
50	4	0.120	1.34	[0.05,0.08]	0.006	50	4	0,122	1,35	[0.01,0.03]	0,003
100	4	0.106	1.31	[0.04,0.07]	0.005	100	4	0,109	1,32	[0.01,0.02]	0,002
200	4	0.103	1.31	[0.04,0.06]	0.003	200	4	0,103	1,30	[0.01,0.02]	0,001
500	4	0.099	1.29	[0.04,0.05]	0.002	500	4	0,095	1,28	[0.01,0.01]	0,001
$p = 15$						$p = 15$					
8	6	0.145	1.49	[0.01,0.09]	0.016	8	6	0,145	1,49	[-0.01,0.03]	0,007
10	6	0.127	1.41	[0.02,0.09]	0.014	10	6	0,127	1,41	[-0.01,0.03]	0,006
15	5	0.129	1.39	[0.03,0.09]	0.011	15	5	0,129	1,39	[-0.00,0.02]	0,005
20	5	0.116	1.34	[0.04,0.09]	0.010	20	5	0,116	1,34	[-0.00,0.02]	0,005
25	4	0.147	1.42	[0.05,0.09]	0.009	25	4	0,147	1,42	[0.01,0.03]	0,004
30	4	0.138	1.39	[0.05,0.09]	0.008	30	4	0,138	1,39	[0.01,0.03]	0,004
50	4	0.121	1.35	[0.04,0.07]	0.006	50	4	0,121	1,35	[0.01,0.02]	0,003
100	4	0.105	1.31	[0.04,0.06]	0.004	100	4	0,105	1,31	[0.01,0.02]	0,002
200	4	0.098	1.29	[0.04,0.05]	0.003	200	4	0,098	1,29	[0.01,0.01]	0,002
500	4	0.095	1.28	[0.04,0.05]	0.002	500	4	0,095	1,28	[0.01,0.01]	0,001
$p = 20$						$p = 20$					
8	6	0.146	1.51	[0.03,0.11]	0.015	8	6	0,149	1,50	[0.00,0.03]	0,006
10	6	0.132	1.45	[0.05,0.12]	0.015	10	6	0,124	1,40	[-0.00,0.03]	0,006
15	5	0.136	1.44	[0.05,0.11]	0.011	15	5	0,128	1,39	[-0.00,0.02]	0,005
20	5	0.123	1.40	[0.04,0.09]	0.010	20	5	0,117	1,36	[-0.00,0.02]	0,004
25	4	0.114	1.36	[0.04,0.09]	0.009	25	4	0,109	1,33	[-0.00,0.02]	0,004
30	4	0.150	1.42	[0.05,0.09]	0.008	30	4	0,140	1,40	[0.02,0.04]	0,004
50	4	0.131	1.37	[0.04,0.07]	0.006	50	4	0,123	1,35	[0.02,0.03]	0,003
100	4	0.111	1.32	[0.05,0.07]	0.005	100	4	0,109	1,32	[0.02,0.03]	0,002
200	4	0.100	1.30	[0.04,0.06]	0.003	200	4	0,101	1,30	[0.01,0.02]	0,001
500	4	0.100	1.30	[0.04,0.05]	0.002	500	4	0,099	1,29	[0.01,0.02]	0,001
$p = 30$						$p = 30$					
8	6	0.143	1.50	[0.02,0.11]	0.016	8	6	0,146	1,49	[0.00,0.03]	0,006
10	6	0.130	1.42	[0.02,0.09]	0.014	10	6	0,128	1,43	[0.00,0.03]	0,006
15	5	0.127	1.37	[0.03,0.09]	0.011	15	5	0,133	1,40	[0.01,0.03]	0,005
20	5	0.112	1.34	[0.03,0.09]	0.010	20	5	0,113	1,34	[0.01,0.03]	0,004
25	4	0.140	1.40	[0.05,0.09]	0.009	25	4	0,143	1,40	[0.01,0.03]	0,004
30	4	0.135	1.38	[0.05,0.09]	0.009	30	4	0,135	1,38	[0.01,0.03]	0,004
50	4	0.121	1.36	[0.04,0.08]	0.007	50	4	0,121	1,36	[0.01,0.03]	0,003
100	4	0.106	1.31	[0.04,0.07]	0.005	100	4	0,106	1,31	[0.01,0.02]	0,002
200	4	0.102	1.30	[0.04,0.06]	0.003	200	4	0,100	1,30	[0.01,0.01]	0,002
500	4	0.097	1.29	[0.04,0.05]	0.002	500	4	0,096	1,29	[0.01,0.02]	0,001

The results obtained with the simulation study for binary outcome variable are fully comparable, with regard to the minimum required group ratio, to the results obtained with the simulation study for continuous outcome. This means that selection bias is equally well removed, regardless of the class of the outcome variable. However, the results are different with respect to the treatment effect standard errors (i.e., the simulation study with binary outcome variable produces smaller treatment effect standard errors) and with respect to the unbiasedness of $A\widehat{T}T$.

With a continuous outcome variable the ATT estimated without the additional regression adjustment is biased (i.e., the true value of the treatment effect $(\tau = 0)$ is not within the confidence interval boundaries of the estimated treatment effect). On the other hand, with a binary outcome variable the ATT estimated without the additional regression adjustment for the smallest treated samples, n_t of 8, 10, 15 and 20, and with observed covariates, p of 10, 15 and 20 are unbiased, whereas biased with p of 30. The ATT estimated without the additional regression adjustment, for all the remaining investigated treated samples, n_t of 25, 30, 50, 100 and 200, are biased regardless of the number of observed covariates.

These differences in results are due to the outcome variable, Y, not being generated as a linear function but as an approximation of a linear model. Thus, small samples normal distributions are not a really good approximation of a normal distribution. Consequently, the bias is lost in the noise of the estimated simulation standard errors for small samples. The noise of the estimated simulation standard errors decreases once the sample size increases; hence, results of large samples with binary outcome variable approximate the results of simulation study with continuous outcome variable.

Chapter 6

Applications

This Chapter presents two applications. The first application simulates real observational data (Lalonde data (1986)) based on the results of the descriptive analysis of our theoretical simulation study with estimated propensity scores (Section 5.1). The purpose of such an application is to evaluate how reliable our results of the descriptive analysis are for practice.

The aim of the second application is twofold: (i) we want to show an example of real data where the nature of data set does not allow the estimation of causal effects; thus, we estimate conditional associations; (ii) to provide an example of why the use of propensity score methods is so important also when studying conditional association questions (i.e., why model-based approaches, (e.g., regression methods) cannot be trusted and or are not reliable when dealing with small samples).

6.1 Real Data Set 1 – The Lalonde Data

To evaluate the reliability of the descriptive analysis results, we apply our findings of the minimum required group ratios to a real data setting by using the within-study comparison of Lalonde (1986).

Lalonde examined the effect of labour market training programmes on earnings. In his within-study comparison, he compared the results of a randomised experiment (the National Support Work Demonstration, NSW) with results obtained from a non-randomised experiment (i.e., where the randomised control group is substituted with a non-equivalent comparison group from the Current Population Survey (CPS) and the Panel Study of Income Dynamics survey (PSID)).

Lalonde then used least squares regressions, an instrumental variable approach and Heckman’s (1979) two-step procedure in order to remove a selection bias in the non-randomised experiment.

This very influential study concluded that the statistical adjustments of the observational data failed to replicate the results of the experimental data. Later on, Dehejia and Wahba (1999) re-evaluated Lalonde’s study by applying propensity score methods and concluded that the use of propensity score methods succeeded in replicating the results of Lalonde’s randomised experiment. Our simulation study of real data uses the same data as Dehejia and Wahba (1999), which can be found in the MatchIt package of the statistical software R (Ho, et al. 2011).

The Lalonde data consist of 445 observations with a treated sample, n_t, of 185 and a control sample, n_c, of 260 with measurements on 10 covariates, an earnings outcome and an assignment variable (of having participated in the labour market programme or not). We conducted a simulation with the Lalonde data based on the minimum required group ratio for each treated sample size, obtained from our theoretical simulation study with estimated propensity scores. Hence, we evaluate whether our descriptive analysis results, of the theoretical simulation, regarding the minimum required group ratio of corresponding treated samples enable to estimate, unbiasedly, treatment effects also within real data settings.

6.1.1.Simulation study

Since Lalonde’s disposable sample indicates an initial squared bias of $B^2 = 0.19$, we conducted an additional theoretical simulation with an initial squared bias of $B^2 = 0.19$ and 10 covariates, because our theoretical simulations covered only initial squared bias, B^2, of 0.5, 1 and 1.5.

The data are generated as explained in Chapter 4, but in order to obtain an initial bias of 0.19, the true propensity scores calculated with a logistic model are following the function:

$$\text{logit}^{-1}(e(X)) = \gamma(X_1 + \ldots + X_{10}),$$

where the gamma coefficient, γ, is 0.09.

The results are evaluated based on the same criteria as in our theoretical simulation study: (i) the absolute value of the remaining bias, RB, calculated as absolute standardised difference in means of propensity score logit is less than 0.15; and (ii) the variance ratio, VR, of propensity score logit variances should be between 0.5 and 2.

Table 6.1 presents results of descriptive analysis, obtained with our theoretical simulation with estimated propensity scores, for 10 observed covariates and an initial squared bias, B^2, of 0.19.

Table 6.1: The minimum required group ratio, R^*, for $p = 10$, and $B^2 = 0.19$

n	R*	RB	$\frac{s^2{}_{ps_t}}{s^2{}_{ps_c}}$	99% CI of $\hat{ATT}$ based on the simulation SE of $\hat{ATT}$	Simulation SE of $\hat{ATT}$
8	11	0.14	1.42	[-0.05,0.08]	0.021
10	8	0.13	1.39	[-0.11,0.02]	0.019
15	5	0.13	1.38	[-0.08,0.02]	0.015
20	4	0.13	1.34	[-0.07,0.02]	0.013
25	4	0.09	1.27	[-0.06,0.02]	0.011
30	3	0.12	1.32	[-0.06,0.01]	0.011
50	3	0.07	1.19	[-0.03,0.02]	0.007
100	2	0.09	1.23	[-0.03,0.01]	0.006
200	2	0.05	1.15	[-0.02,0.01]	0.004
500	2	0.03	1.10	[-0.01,0.01]	0.002

p – number of covariates; R* – minimum group ratio; RB – remaining bias. The confidence interval of the estimated ATT includes the value of zero which represents the true treatment effect.

As expected, the initial squared bias of $B^2 = 0.19$ results in smaller minimum required group ratios for investigated treated samples than observational data with larger initial squared biases, B^2, of 0.5, 1.0 and 1.5. The smallest investigated treated sample of $n_t = 8$ only requires a minimum group ratio of 11 when $B^2 = 0.19$, whereas with an initial squared bias of $B^2 = 1$, a minimum group ratio of 19 is required. Table 6.1 provides the minimum required group ratios, of the additional theoretical simulation, for each investigated treated sample with initial imbalances of $B^2 = 0.19$. Based on these group ratios the Lalonde data are simulated.

Lalonde's disposable sample of 445 units acts, in this case, as the target population from which samples of desired sizes are drawn. Based on disposable treated ($n_t = 185$) and control ($n_c = 260$) units of the Lalonde sample, we can only simulate small treated samples with 100 or less units. We thus investigated the following treated sample sizes: $n_t = \{8, 10, 15, 20, 25, 30, 50, 100\}$ with group ratios (obtained from Table 6.1), R^*, of 11, 8, 5, 4, 4, 3, 3, and 2, respectively. With regard to the investigated treated sample sizes and group ratios, the sample sizes of the control group are: $n_c = \{88, 80, 75, 80, 100, 90, 150, 200\}$.

We use the same logistic regression model for estimating propensity scores as in the theoretical simulation study: $\textit{logit}(W) = \lambda_0 + \lambda_1 X_1 + \ldots + \lambda_{10} X_{10}$ where W denotes a treatment indicator (i.e., whether treatment was applied to a unit, $W = 1$, or it was not applied to a unit, $W = 0$). The simulation consists of 1,000 replications and the propensity score logit is used for computing the balancing measures (e.g., remaining bias and variance ratio).

6.1.2. Simulation Results

The results of the Lalonde simulation aim to evaluate the consistency with the results of the theoretical simulation regarding the balancing diagnostics (i.e., remaining bias, variance ratio) and thus provide answers to three main questions: (i) is the remaining bias in absolute terms, RB, smaller than 0.15; (ii) is the variance ratio, VR, between 0.5 and 2; and (iii) is the estimated ATT close to the treatment effect of the randomised experiment provided in Lalonde (1986), $\tau = 1794$.

The Lalonde simulation results are provided in Table 6.2, and are consistent with the results of the theoretical simulation, regarding the balance diagnostics. The absolute value of the remaining bias is below 0.15 and the variance ratio does not go below 0.5 or above 2 for all of the investigated treated samples.

Table 6.2: Simulation results of the Lalonde data set.

n	R*	RB	$\frac{s^2_{ps_t}}{s^2_{ps_c}}$	$\hat{ATT}$	99% CI of $\hat{ATT}$ based on the simulation SE of $\hat{ATT}$	Simulation SE of $\hat{ATT}$	$\hat{ATT} - \tau$
[a] 8	r11	0.11	1.74	1380	[577,2183]	230	-414
10	r8	0.11	1.69	1908	[1314,2501]	183	114
15	r5	0.12	1.56	1760	[1400,2121]	121	-34
20	r4	0.11	1.50	1807	[1560,2054]	86	13
25	r4	0.08	1.30	1767	[1570,1963]	70	-27
30	r3	0.11	1.37	1776	[1606,1945]	62	-18
50	r3	0.05	1.15	1876	[1770,1982]	40	82
100	r2	0.05	1.12	1789	[1723,1855]	25	-5

p – number of covariates; R* – minimum required group ratio; RB – remaining bias. The confidence interval of the estimated ATT includes the value of 1794 which represents the true treatment effect (LaLonde 1986).

[a] the logistic regression used for estimating propensity scores resulted in extreme values of 0 and 1 for 32% of simulation replications

Furthermore, for all the investigated treated samples, the Lalonde simulation results show even slightly less remaining bias in comparison to our theoretical simulation results. For treated sample sizes of $n_t \geq 30$, the Lalonde simulation results demonstrate even better balancing also regarding the variance ratio (VR is closer to 1 than in our theoretical simulation). In addition, the ATT estimate

($\hat{\tau} = 1789$ with $n_t = 100$ and $R = 2$) is very close to the treatment effect estimate of the randomised experiment ($\tau = 1794$, Lalonde (1984)) and to the treatment effect estimate of Dehejia and Whaba (1999) - $\hat{\tau} = 1788$.

The only inconsistency found, is with the smallest investigated treated sample, $n_t = 8$. We estimate that in the Lalonde simulation with eight treated units, too many simulation replications (32%) violated the estimated probabilistic part of the strong ignorability assumption. This was not the case in our theoretical simulation when the percentage was only 0.5% for the treated sample consisting of eight units. Such an inconsistency might result from the fact that none of observed covariates of the Lalonde data set are normally distributed.

6.2 Real Data Set 2 – The Role of Cultural Capitals in Production of Good Health

The observational data for investigating the role of cultural capitals in production of good health consist of small and moderately large samples (15 propensity score studies are performed). The main aims of such studies are: (i) to provide an example of observed data where our research questions are causal in some vague sense but the nature of the data does not allow us to formulate the intervention as defined in Section 2.1. thus, we cannot estimate causal effects of “treatment” versus “control” but only conditional associations; (ii) to show the unreliability of model-based approaches (e.g., regression analysis) when estimating conditional associations with small samples.

A major part of this section was published in the Slovenian Journal of Public Health, co-authored with Kamin and Steiner (2013).

6.2.1 Propensity Score Study

> *The main research questions of our study are: (i) How do different states of the cultural capital (i.e., institutionalised, objectified and incorporated) associated with the self-assessed health[16]? (ii) Do objectified and incorporated cultural capital associated with self-assessed health if we condition on institutionalised cultural capital, i.e., on education levels (i.e., do objectified and incorporated cultural capital explain something in addition to the institutionalised cultural capital)? (iii) Is there a difference between women and men in how cultural capital is associated with their self-assessed health?*

We are aiming to estimate conditional associations between a binary variable Z, which denotes here two different levels of cultural capital that a person can possess (i.e., low versus high level of cultural capital) and Y, which represents the self-assessed health given the observed covariates, X, on which we are conditioning to obtain a balanced study design (i.e., group of respondents with low level of cultural capital is comparable to the group of respondents with high level of cultural capital).

> *We conceptualise cultural capital according to the theory of the French sociologist Pierre Bourdieu* (Bourdieu 1986)*, who operationalises cultural capital in three mutually dependent states: (i) institutionalised cultural capital, which represents formal education and qualifications; (ii) incorporated cultural capital, which stands for embodied knowledge, cognitive abilities, skills, taste, and competencies; and (iii) objectified cultural capital, which represents material forms and representations of knowledge, social recognition, and cultural goods.*

[16] The self-assessed health is a good subjective measure for the (objective) health status. Its relationship with morbidity and mortality is well-researched and proves to be a good measure for the health status: self-assessed poor health is related to higher risk of poor health outcomes (i.e., higher mortality and morbidity) (Idler and Benyamini 1997)

DATA

The data were collected between 1st of December 2009 and 15th of February 2010. The target population comprises to adults aged 18 years or older with a permanent address either in Ljubljana or Maribor (Kamin, et al. 2013).

> *Units were randomly sampled from the Central Population Register of Slovenia. Simple random sampling was used for this type of population, as recommended by the Slovenian Statistical Office, based on the prior experience of their sampling professionals. 820 face-to-face interviews were successfully completed. (...) The collected data have been weighted based on gender and age via poststratification adjustment using raking method [* (Lavrakas 2010)*]*.
>
> *The survey data consist of 105 variables. For our analyses we selected three types of variables: (i) self-assessed health as our dependent variable (...); (ii) variables related to cultural capital states in order to construct cultural capital indexes; and (iii) variables that serve as covariates for balancing group differences between units belonging to different cultural capital levels (low vs. high).*

The variables that were selected to control for group differences are the following: respondent's gender, age and nationality, the job of respondent's father when the respondent was 15 years old, education of respondent's father, the job of respondent's mother when the respondent was 15 years old, education of respondent's mother, the political party to which the respondent relates, the political orientation (i.e., left, right, middle) and the residence location (i.e., Maribor, Ljubljana). These variables appear as substantively interesting because they are related to both the self-assessed health and the cultural capital states (Kamin, et al. 2013).

The aim was to investigate conditional associations between different cultural capital states and self-assessed health, thus, three different cultural capital indexes were constructed: objectified, incorporated and overall cultural capital. The selection of variables to create the three indexes follows Bourdieu's theory (1986) (Kamin, et al. 2013).

The institutionalised cultural capital includes only one variable – education, which denotes the highest educational degree attained. The objectified cultural capital index includes six variables: possession of a number of: (i) books; (ii) original music LPs and CDs; (iii) Music in mp3; (iv) original Art; (v) possession of PC (yes/no); and (vi) access of Internet connection in the household (yes/no). According to the theory, these variables represent material forms and representations of knowledge, social recognition and cultural goods. The incorporated cultural capital index includes 3 variables: (i) self-reported foreign language skills (English, Ex-Yugoslavian, Other); (ii) self-assessed competencies of Internet use (scale from 1 to 7); and (iii) self-assessed knowledge about art (five levels of agreement: I don't agree at all, I don't agree, neither-neither, I agree, I completely agree). All of these variables are intrinsic to a person (i.e., they present embodied knowledge, perceptions, cognitive abilities, skills, and competencies). The overall cultural capital index consists of all variables included in the incorporated and objectified cultural capital index, and the institutionalised cultural capital. The institutionalised cultural capital is represented by a three-level education variable: (i) low educational level (11 or less completed schooling years) (ii) middle education (between 12 and 14 completed schooling years) and (iii) high education level (15 and more completed schooling years).

Some observations in our data set were missing due to item-nonresponse. However, none of our variables contained more than 10% missing values. The majority of variables had less than 5% missing values. We imputed the few missing values in our data set [according to Rubin's theory (1987; 1996)] using chained equations as implemented in the R-package mice (van Buuren and Groothuis-Oudshoorn 2011)*, which uses linear regression for continuous* *variables, logistic regression for dummy variables and polytomous [(unordered)] regression for discrete variables with more than 2 levels.*

[After imputing the missing values,] we constructed the objectified, incorporated and the overall cultural capital indexes by first transforming all the continuous variables to a zero-to-one scale (with minimum of 0 and maximum of 1), and then by taking the average value of all the variables included in the index. For each continuous index, we then created a dummy variable, indicating a low vs. high cultural capital status, where the cut-off point was defined by the median value of each index, respectively.

We used these dichotomous cultural capital indexes as dependent variables for estimating propensity scores. Table (…) [6.3] shows mean values and standard deviations for low and high levels of cultural capital indexes, as well as standardised mean differences, which indicate the difference between the two groups in terms of the underlying continuous indexes.

As we can see [from Table 6.3], by switching from low to high cultural capital there is a shift of 1.49 standard deviations (SD) in the objectified cultural capital, 1.64 standard deviations in the incorporated cultural capital and 1.63 standard deviations in the overall cultural capital. This information helps when interpreting the conditional comparison estimates (i.e., the estimates obtained by estimating conditional associations) for cultural capital and self-assessed health presented later in Tables (…) [6.4, 6.5 and 6.6].

Table 6.3: Descriptive statistics for objectified, incorporated and overall cultural capital (CC)

Cultural Capital (CC)	Mean for low CC	Standard deviation for low CC	Mean for high CC	Standard deviation for high CC	Standardised mean difference
Objectified CC	0.24	0.15	0.50	0.07	1.49
Incorporated CC	0.34	0.11	0.62	0.09	1.64
Overall CC	0.30	0.12	0.60	0.09	1.63

Source: Kamin, et al. 2013, 111.

Once the database required for estimating conditional associations was obtained, the self-assessed health variable was removed from the database in order to follow the design phase procedure of propensity score methods (i.e., no outcome variable in sight).

The design phase of this propensity score study was performed using propensity score matching adjustment by using optimal full matching algorithm (Hansen and Klopfer 2006) with a calliper of 0.1 standard deviations of the logit of the propensity score. The optimal full matching algorithm exhausts the complete dataset; thus, we discarded only the units for which propensity score logit values were outside the specified caliper (Kamin, et al. 2013).

PROPENSITY SCORE ESTIMATION AND COVARIATE BALANCE EVALUATION

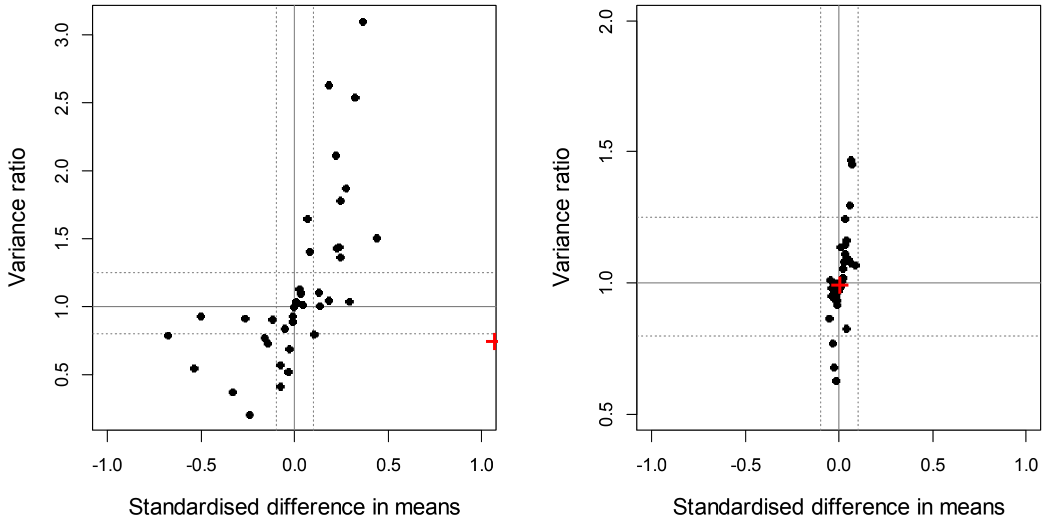
> *In order to balance the study design and obtain comparable cultural capital groups on the observed covariates (i.e., the 10 covariates that were selected) we first estimated, for each cultural capital index, the respondents' propensities for being in the high vs. low group (for the three-valued education variable, we estimated the propensity scores for all three group comparisons).*

Propensity scores were estimated using logistic regression. The specification of logistic models used to estimate propensity scores is included in the footnotes when displaying results of the estimated conditional associations (Tables 6.4 – 6.6).

The balance diagnostics were performed graphically by depicting levels of quantitative balance diagnostics measures, such as, standardised difference in means and variance ratios. Figure 6.1 shows such a balance plot of observed covariates before and after the employment of propensity score matching for the objectified cultural capital group. The balance diagnostics for propensity score studies for other cultural capital states show analogous graphical depictions (i.e., study designs after adjusting for imbalances in covariates are sufficiently balanced), thus, we are not including their graphs.

The dashed horizontal and vertical lines in Figure 6.1 denote the "acceptable" levels of covariate balance (i.e., there are negligible differences in covariate distributions between the groups) where the absolute value of the standardised mean difference should be smaller than 0.1, and the variance ratio should not be smaller than 0.5 or bigger than 2. The red cross indicates covariate (im)balance in the propensity score logit.

Figure 6.1: Balance plots for objectified cultural capital: initial imbalance in observed covariates (left plot) and balance in observed covariates after matching (right plot). (Categorical variables are included as 0/1 indicator variables).



From Figure 6.1 we can see how heavily imbalanced observed covariates between the two groups are in the original study design – the left plot in Figure 6.1 shows that, respondents in the high and low objectified cultural capital group considerably differ in several observed covariates. Using such a design for estimating conditional associations can potentially result in severely biased (unrealistic) conditional comparison estimates.

After matching respondents with low and high objectified cultural capital on the estimated logit of the propensity scores, nearly all covariate differences between the low and high objectified cultural capital groups have been removed (right plot in Figure 6.1). ”All standardised mean differences are close to zero (within 0.1 standard deviations) and for most covariates, the variance ratios between the low and high objectified cultural capital groups are within 4/5 and 5/4. Thus, the plot indicates a good balance in the observed covariates [(i.e., a balanced study design)]” (Kamin, et al. 2013, 112-113).

ESTIMATION OF CONDITIONAL ASSOCIATIONS

Once the study design was balanced (i.e., the design phase was completed), conditional associations were estimated by using weighted least-squares regression with an additional regression covariate adjustment as stated in (Kamin, et al. 2013, 113):

$$\text{Self-assessed health} = \beta_0 + \beta_1 \text{cultural capital state} + \beta_2 \text{location} + \beta_3 \text{sex} + \beta_4 \text{age} + \varepsilon,$$

> *with individual case weights derived from the matching structure. We included covariates sex, age and the residence location (i.e., Maribor, Ljubljana) in order to remove residual imbalances (i.e., the imbalances left after employing propensity score matching), and to increase the precision of our estimates.*

RESULTS

As mentioned before, several propensity score studies were conducted to investigate conditional associations between different states of cultural capital and self-assessed health.

Conditional comparison estimates between the self-assessed health and institutionalised, objectified, incorporated, and overall cultural capital, respectively

Table 6.4 presents the conditional comparison estimates between self-assessed health and different levels of institutionalised cultural capital (i.e., education), respectively. The estimates refer to comparisons of low vs. high education level, low vs. medium education level and medium vs. high education level. Table 6.5 presents: (i) the conditional comparison estimate between self-assessed health and objectified cultural capital; (ii) the conditional comparison estimate between self-assessed health and incorporated cultural capital; and (iii) the conditional comparison estimate between self-assessed health and overall cultural capital (Kamin, et al. 2013, 113).

Table 6.4: Conditional comparison estimates between self-assessed health and institutionalised, objectified, incorporated and overall cultural capita, respectively

	Estimate	Std. Error	p -value	n	Preserved (effective n)
Low vs. high institutionalised CC[17]	0.44	0.09	0.000	419	70.64% (n=295)
Low vs. medium institutionalised CC[18]	0.19	0.07	0.006	620	93.4% (n=579)
Medium vs. high institutionalised CC[19]	0.11	0.07	0.120	599	88.16% (n=528)

Source: Kamin, et al. 2013, 113

[17]

$$Prob(W_i = 1) = F(\text{sex, age, nationality, father's work, father's education, mother's work, mother's education, political party, political orientation, respondent's location, political party * father's work, age * political orientation, location * political party, age * political party})$$

[18]

$$Prob(W_i = 1) = F(\text{sex, age, nationality, father's work, father's education, mother's work, mother's education, political party, political orientation, respondent's location, location * political party, sex * political party})$$

[19]

$$Prob(W_i = 1) = F(\text{sex, age, nationality, father's work, father's education, mother's work, mother's education, political party, political orientation, respondent's location, social class, age * mother's education, sex * political party, location * political party})$$

The first investigation with respect to the conditional comparison estimates for different levels of **institutionalised cultural capital** and self-assessed health are significantly positive for the low vs. high contrast and the low vs. medium contrast, but insignificantly positive for the medium vs. high contrast. “This means that an individual with higher institutionalised cultural capital assesses his/her health as being better than an individual with a lower level of institutionalised cultural capital” (Kamin, et al. 2013, 113).

Table 6.5: Conditional comparison estimates between self-assessed health and institutionalised, objectified, incorporated and overall cultural capita, respectively

	Estimate	Std. Error	p -value	n	Preserved (effective n)
Objectified CC (OCC)[20]	0.21	0.06	0.000	819	96.1% (n=787)
Incorporated CC (ICC)[21]	0.06	0.06	0.370	819	87.2% (n=714)
Cultural capital (institutional + objectified + incorporated CC) (CCI)[22]	0.36	0.06	0.000	819	88.5% (n=724)

Source: Kamin, et al. 2013, 113.

The conditional comparison estimate for objectified cultural capital and the self-assessed health is significantly positive (0.21) (Kamin, et al. 2013, 113).

[20]

$$\mathrm{P}rob(W_i = 1) = F(\text{sex, age, nationality, father's work, father's education, mother's work, mother's education,}$$
$$\text{social class, political party, political orientation, respondent's location, age * mother's work,}$$
$$\text{sex * political party, location * political party)}$$

[21]

$$\mathrm{P}rob(W_i = 1) = F(\text{sex, age, nationality, father's work, father's education, mother's work, mother's education,}$$
$$\text{political party, political orientation, respondent's location, location * political party, sex * political party,}$$
$$\text{age * political party, mother's work * age, father's work * political party, father's education * location)}$$

[22]

$$\mathrm{P}rob(W_i = 1) = F(\text{sex, age, nationality, father's work, father's education, mother's work, mother's education,}$$
$$\text{political party, political orientation, respondent's location, location * political party, sex * political party,}$$
$$\text{mother's work * mother's education, political orientation * location,}$$
$$\text{father's work * political orientation, father's work * age)}$$

> *Individuals that possess more objectified cultural capital assess their health better than individuals with less objectified cultural capital. However, from a subject-matter point of view, the average 0.21 increase on the 5-point scale of self-assessed health represents only a moderate effect, given that the contrast between the low and high objectified cultural capital status is rather strong (1.49 standard deviations) as Table 6.3 shows; the 0.21 increase then translates into an effect size of 0.15 standard deviations). The conditional comparison estimate for incorporated cultural capital and self-assessed health is positive but small and also insignificant.*

The conditional comparison estimate for **overall cultural capital**, which includes all three states of cultural capital (i.e., institutionalised, objectified and incorporated), and the self-assessed health of an individual is significantly positive (0.36) (Kamin, et al. 2013, 114).

> *Individuals that possess more cultural capital, assess their health better than individuals with less cultural capital. However, from a subject-matter point of view, the average of 0.36 increase on the 5-point scale of self-assessed health represents only a moderate effect (effect size of 0.23 standard deviations) because the contrast between the low and high overall cultural capital status is rather strong (1.6 standard deviations as Table (…) [6.3] shows).*

Conditional comparison estimates between self-assessed health and objectified cultural capital, and incorporated cultural capital, respectively within each level of institutionalised cultural capital

The previous section presented the conditional comparison estimates for each cultural capital index without controlling for the other cultural capital indexes. Thus, the positive conditional association between objectified cultural capital and self-assessed health, and the positive but insignificant conditional association between incorporated cultural capital and self-assessed health might be due to objectified and incorporated cultural capital's correlation with institutionalised cultural capital rather than the objectified and incorporated cultural capital on their own. Hence, the second analysis controls for the institutionalised cultural capital (education variable), when estimating conditional associations between the self-assessed health and objectified cultural capital, and between the self-assessed health and incorporated cultural capital (Kamin, et al. 2013, 114).

A stratification approach was used to conduct a separate propensity score analysis with optimal full matching in each of the three levels of institutionalised cultural capital. By stratifying on the educational levels, the effect of education was removed from both the objectified and incorporated cultural capital. Tables 6.6 and 6.7 show the results from the stratification analyses (Kamin, et al. 2013, 115).

Table 6.6: Conditional associations between self-assessed health and objectified cultural capital investigated within each level of institutionalised cultural capital

	Estimate	Standard Error	p -value	n	Preserved (effective n)
Low institutionalised CC level – objectified CC[23]	0.18	0.14	0.20	220	67.7% (n=148)
Medium institutionalised CC level – objectified CC[24]	0.13	0.07	0.08	400	91.75% (n=367)
High institutionalised CC – objectified CC[25]	0.22	0.15	0.15	199	61.8% (n=122)

Source: Kamin, et al. 2013, 114

[23] $\mathrm{P}rob(W_i = 1) = F(\text{sex, age, nationality, father's work, father's education, mother's work, mother's education, political party, political orientation, respondent's location, location * political orientation, sex * political orientation political party * father's education})$

[24] $\mathrm{P}rob(W_i = 1) = F(\text{sex, age, nationality, father's work, father's education, mother's work, mother's education, political party, political orientation, respondent's location, location * political orientation})$

[25] $\mathrm{P}rob(W_i = 1) = F(\text{sex, age, nationality, father's work, father's education, mother's work, mother's education, political party, political orientation, respondent's location, age * mother's work, location * political party, father's work * political orientation})$

Objectified cultural capital shows a positive conditional association with the self-assessed health even after controlling for levels of institutionalised cultural capital. The conditional comparison estimate of the objectified cultural capital and self-assessed health for the low level of institutionalised cultural capital is 0.18, for the medium level 0.13, and for the high level it is 0.22 (Table 6.6).

However, the conditional comparison estimates for the low and medium levels of institutionalised cultural capital are somewhat smaller than the conditional comparison estimates for objectified cultural capital and self-assessed health, when we are not controlling for the education (0.21, see Table 6.4). At the same time, the three estimates (Table 6.6) are no longer significant, which is likely due to the reduced sample sizes within each educational stratum. Nonetheless, the pattern of results across the three educational levels suggests, that objectified cultural capital is associated with self-assessed health in addition to education.

Table 6.7 shows results for the incorporated cultural capital, where the conditional comparison estimate of the incorporated cultural capital and self-assessed health decreases, as the educational level increases. For the low educational level, the estimate amounts to 0.27, for the medium level to 0.20, and for the high educational level the conditional comparison estimate is slightly negative (-0.07).

Table 6.7: Incorporated cultural capital investigated within each level of institutionalised cultural capital

	Estimate	Standard Error	p -value	n	Preserved (effective n)
Low institutionalised CC level – incorporated CC[26]	0.27	0.15	0.07	220	67.2% (n=147)
Medium institutionalised CC level – incorporated CC[27]	0.20	0.08	0.01	400	89.3% (n=357)
High institutionalised CC level – incorporated CC[28]	-0.07	0.13	0.60	199	82.9% (n=164)

Source: Kamin, et al. 2013, 114

However, the conditional comparison estimate for incorporated cultural capital and self-assessed health is significant for the medium educational level, but insignificant for the low and high educational levels, which is likely the consequence of the reduced sample sizes[29].

Thus, if we compare these results to the results where we are not controlling for the education levels (i.e., conditional comparison estimates for incorporated cultural capital and self-assessed health are insignificant – see Table 6.4), we can conclude that incorporated cultural capital is associated with self-assessed health, but only for persons with a medium, and maybe for persons with a low educational level. (Note that results from Table 6.4 and 6.7 are not fully comparable because the underlying sample sizes differ.)

[26]

$$\begin{aligned}&\mathrm{P}rob(W_i = 1) = F(\text{sex, age, nationality, father's work, father's education, mother's work, mother's education,}\\&\text{political party, political orientation, respondent's location, location * political orientation, age * political orientation,}\\&\text{political party * father's education})\end{aligned}$$

[27]

$$\begin{aligned}&\mathrm{P}rob(W_i = 1) = F(\text{sex, age, nationality, father's work, father's education, mother's work, mother's education}\\&\text{political party, political orientation, respondent's location, location * political orientation,}\\&\text{age * political party, age * mother's work})\end{aligned}$$

[28]

$$\begin{aligned}&\mathrm{P}rob(W_i = 1) = F(\text{sex, age, nationality, father's work, father's education, mother's work, mother's education,}\\&\text{political party, political orientation, respondent's location, age * mother's work, location * political party,}\\&\text{father's work * political orientation})\end{aligned}$$

[29] In order to obtain a balanced design some units were discarded.

Gender-specific analysis

A gender-specific analysis was performed in order to investigate possible differences between women and men in conditional comparison estimates for different cultural capital states and self-assessed health (Kamin, et al. 2013, 115).

Table 6.8: Conditional comparison estimates for the self-assessed health and overall cultural capital, objectified, and incorporated cultural capital, respectively when controlling for sex

	Estimate	Standard Error	p -value	n	Preserved (effective n)
Women – overall CC [30]	0.34	0.08	0.000	450	74% (n=333)
Men – overall CC	0.23	0.10	0.015	369	82% (n=301)
Women – objectified CC [31]	0.21	0.08	0.008	450	91% (n=409)
Men – objectified CC	0.16	0.09	0.080	369	79% (n=291)
Women – incorporated CC [32]	0.19	0.08	0.020	450	90% (n=405)
Men – incorporated CC	-0.08	0.08	0.370	369	90% (n=332)

Source: Kamin, et al. 2013, 115

Table 6.8 indicates that for women, the conditional comparison estimates for the self-assessed health and overall cultural capital, objectified, and incorporated cultural capital, respectively are significantly positive. For males, only the conditional comparison estimate of self-assessed health and overall cultural capital is significantly positive, whereas the conditional comparison estimate of self-

[30] $\mathrm{P}rob(W_i = 1) = F(\text{age, nationality, father's work, father's education, mother's work, mother's education,}$
$\text{political party, political orientation, respondent's location, location * political party,}$
$\text{mother's work * mother's education, political orientation * location, father's work * political orientation,}$
$\text{father's work * age})$

[31] $\mathrm{P}rob(W_i = 1) = F(\text{age, nationality, father's work, father's education, mother's work, mother's education}$
$\text{social class, political party, political orientation, respondent's location, age * mother's work,}$
$\text{location * political party})$

[32] $\mathrm{P}rob(W_i = 1) = F(\text{age, nationality, father's work, father's education, mother's work, mother's education}$
$\text{political party, political orientation, respondent's location, location * political party, age * political party,}$
$\text{mother's work * age, father's work * political party, father's education * location})$

assessed health and objectified, and the incorporated cultural capital, respectively are insignificant.

We also made an attempt to investigate conditional associations between the self-assessed health and different cultural capital states within each level of education (i.e., institutionalised cultural capital) for the gender-specific analysis. Unfortunately, due to very small samples and group ratios, we were unable to obtain a sufficient balance on observed covariates for those designs. Hence, conditional associations could not be reliably estimated. Table 6.9 presents small sample sizes and corresponding group ratios with which we could not obtain a balanced design.

Table 6.9: Small sample sizes and group ratios with which we could not obtain a balanced design

	$R = \frac{n_{Women}(Z=0)}{n_{Women}(Z=1)}$	$R = \frac{n_{Men}(Z=0)}{n_{Men}(Z=1)}$
Low institutionalised CC level – objectified CC	$R = \frac{59}{35} = 1.7$	$R = \frac{99}{27} = 3.7$
Medium institutionalised CC level – objectified CC	$R = \frac{79}{95} = 0.8$	$R = \frac{117}{27} = 4.3$
High institutionalised CC – objectified CC	$R = \frac{24}{77} = 0.3$	$R = \frac{30}{68} = 0.4$

However, if we would use a model-based approach (e.g., regression methods) instead, we would not be aware of the fact, that obtained estimates of conditional associations with such small samples (in combination with not sufficiently large group ratios) are heavily relying on extrapolation, and are thus less trustworthy. In this sense, the use of propensity score methods safeguards us, because the main aim of the methods is to balance a study design first, and only once a balanced design is obtained, we proceed with the estimation of desired quantities.

6.2.2 Conclusion

The results suggest that cultural capital is associated with self-assessed health: “persons with a high CC [cultural capital] assess their health better than persons with low CC [cultural capital], even after controlling for many background characteristics” (Kamin, et al. 2013, 115).

The uniqueness of this application is twofold. First, it provides an example of observed data, where the nature of data does not allow us to estimate causal effects of “treatment” versus “control” because the intervention cannot be formulated as defined in Section 2.1 (i.e., the level of cultural capital, that a respondent possesses, is what it is, and there is no intervention that could at particular point of time change the level of respondents’ cultural capital (from high to low or vice versa).

Second, the initial covariate imbalances in this study design are much larger in comparison to what we investigated with our theoretical simulation studies. Additionally, ratios between the units with low level of cultural capital and units with high level of cultural capital are much smaller than the group ratios that would be required (based on the results of our theoretical simulation study) for balancing a study design. Thus, many of units had to be discarded from our sample in order to obtain comparable groups (i.e., balanced design). Consequently, the application includes small and moderately large samples. Thus, it shows how important it is, particularly when dealing with small samples, to use propensity score methods to estimate a valid conditional associations, because the estimates will be more trustworthy (i.e., the estimates can be obtained only if a study design is balanced, whereas with regression analyses, the estimates can always be obtained due to linear extrapolation).

Chapter 7

Conclusion

This thesis defines propensity score methods in terms of their implementation and usage with observational study designs. The methods can be used when estimating causal effects or conditional associations from observed data. In Section 2.1 we provide the definition of what is causal and based on this definition, we offer a definition for conditional association. In Section 6.2 we provide an example of the observed data where only conditional associations can be investigated, and further show the importance of the design-based approach (i.e., the foundation of propensity score methods) for obtaining trustworthy estimates of desired quantities (e.g., causal effects or conditional comparison estimates).

Propensity score methods that are founded on the design-based approach consist of two important parts: (i) the design phase, which is "outcome free" and consist of balancing and balance assessment tools with which we balance a study design with respect to observed covariates, and; (ii) the analysis phase, which uses the outcome data to perform additional statistical adjustments when estimating causal effects or conditional associations. The analysis phase also consists of sensitivity analyses, which should always be performed when estimating causal effects from observational data.

The model-based approach requires the outcome data in the process of balancing a study design and, thus, removes the imbalances in a study design simultaneously with the estimation of desired quantities. However, such an approach is problematic because it relies on strong assumptions: the outcome model, through which we balance a study design and estimate desired quantities, is correctly specified.

Additionally, we show that the model-based approach is even more untrustworthy when the observational data involves small samples. Our Real Data Set 2 application (Section 6.2) shows that for some instances, when sample sizes are small, group ratios small and initial imbalances severe, we are not able to obtain a balance design that is required for reliable estimation of desired quantities. Thus, we should not proceed with the analysis phase of propensity score methods, but accept the fact that, due to insufficient sample sizes, desired quantities (i.e., causal effects or conditional comparison estimates) cannot be reliably estimated.

In contrast, the model-based approach would not give us any warning, but it would simply provide us with some estimates that would rely on extrapolation, and then it is up to the investigator to decide how much to trust the obtained estimates. This approach certainly gives a lot of room for unhealthy data manipulation (i.e., obtaining estimates that one would like to see).

Furthermore, the study of small sample properties with propensity score methods reveals many findings, previously unknown to the research society, which employs these methods, for estimation of causal effects in observational designs. We would like to note here that all our simulation study findings regarding small and moderately large samples in propensity score methods are applicable to causal inference in observational designs when the assignment mechanism is strongly ignorable and SUTVA is satisfied. Yet, these results can be directly applicable to conditional association inference without having to consider either of the above mentioned assumptions.

Our findings evidently show that propensity score methods perform differently when small samples are used, in comparison to moderately large samples. By examining theoretical properties of propensity score methods and by incorporating the findings of the previous research, with regard to sample size concerns in propensity score studies for estimating causal effects, we conclude that propensity score matching adjustment is often the most suitable approach to be used when dealing with small samples. Propensity score matching adjustment method is also one of the most widely applied adjustment method for removing initial covariate imbalances in observational designs. Yet, the method greatly relies on having a pool of control units that is moderately larger from the pool of treated units, in order to balance a study design with respect to observed covariates.

In accordance with the aim of our study: to find how well propensity score methods perform in cases of small samples, and what are the smallest possible treated samples with which the methods can effectively remove initial covariate imbalances from observational designs, we carried out a variety of simulation studies examining different scenarios. The range of simulation scenarios includes simulations performed for small and moderately large treated samples. Our results of the moderately large treated sample study, with respect to the minimum required group ratios, show consistency with the results obtained by Rubin and Thomas (1996). Such a consistency increases the reliability of our simulation results obtained for small treated samples.

Primarily, the simulation studies examined how factors, such as the number of observed covariates and the level of initial covariate imbalances in a study design impact the sample size requirements for propensity score methods to remove initial covariate imbalances from observational design. The sample size requirements are examined by studying different sizes of treated samples and different levels of group ratios, which define the number of control units per treated unit, and hence reveal the required size of a control sample.

Furthermore, we studied the differences of the results obtained when propensity score matching adjustment method is performed with true versus estimated propensity scores. In this sense, we examine a deviation in results between a “perfect scenario” and a “real world scenario”. Additionally, we also examined possible differences in performance of the methods when the two main matching algorithms (i.e., greedy versus optimal matching algorithm) are used with propensity score matching.

The analyses of simulated data are performed descriptively and by the analysis of variance. The results show that small samples behave somehow differently, when using propensity score methods, from moderately large treated samples. Yet, small samples perform as good as moderately large treated samples in removing initial covariate imbalances from observational designs, but under different conditions than moderately large treated samples. The success of a propensity score study (i.e., the successful removal of initial covariate imbalances) with small treated samples primarily depends on a sufficiently large pool of control units. However, the required size of a control group depends also on the number of observed covariates and the level of initial covariate imbalances.

The level of the minimum required group ratio (i.e., ratio between the samples of control and treated units) with small treated samples predominantly depends on the treated sample size, on the number of observed covariates and on the level of the initial imbalances in covariates between the treated and control groups (e.g., the initial bias). The smaller the treated samples, the more observed covariates and the larger the initial covariate imbalances (i.e., the more heterogeneous the two groups are), the larger are the group ratios required to balance a study design.

On the contrary, the level of the required group ratio with moderately large treated samples mainly depends on the level of the initial covariate imbalances, whereas the number of observed covariates, at least those that we have investigated ($p = 10,15,20,30$), has a negligible impact on the required group ratio with moderately large treated samples. The fact that the number of observed covariates plays such a major role with small treated samples has to do with the estimation of propensity scores. The more observed covariates we have, and the smaller the overall sample is, the harder is to estimate propensity scores with high enough precision, so that they would act as good balancing scores in the process of removing covariate imbalances from observational design.

Estimating propensity scores with high enough precision is not the only issue that small treated samples are facing. With very small overall samples $n \leq 50$, group ratio $R = 2$ and by having 10 observed covariates, the logistic regression used for estimating propensity scores resulted in extreme values of 0 and 1 (i.e., we estimate that the probabilistic part of the strong ignorability assumption is violated); thus, it might not be wise to proceed with a propensity score study.

However, according to our results, the “success” of logistic regression (i.e., logistic regression not resulting in extreme values of 0 and 1) primarily depends on the size of the treated group and the number of observed covariates. With 30 observed covariates, an overall sample size $n = 104$ and the group ratio $R = 12$ (i.e., $n_t = 8$ and $n_c = 96$), logistic regression resulted in extreme values of 0 and 1 for all investigated levels of the initial squared bias and in all the simulation replications. In contrast, with 30 observed covariates, an overall sample size $n = 120$ and the group ratio $R = 3$ (i.e., $n_t = 30$ and $n_c = 90$) the logistic regression was “not successful”[33] in only 10 per cent of simulation replications but merely for the strongest selection mechanism $B^2 = 1.5$. These findings are confirmed also by the results obtained with ANOVA where the treated sample size appears as the most influential factor in the small treated sample study (Table 5.21).

Although the simulation study’s findings demonstrate that small treated samples (as small as $n_t = 8$) can perform as good as moderately large treated samples (i.e., n_t of 200 or 500) in removing covariate imbalances from observational designs, as long as the group ratio is sufficiently big and the treatment assignment mechanism is strongly ignorable, the treatment effect estimates with small treated samples are obviously much less precise.

[33] The logistic regression resulting in extreme values of 0 and 1.

The lack of precision in treatment effect estimates for small treated samples is, as expected from the basic standard error calculations, due to larger standard errors in comparison to the standard errors of treatment effect estimates with moderately large treated samples. The treatment effect standard errors with small treated samples can be many times bigger (up to 15 times bigger) than those obtained with moderately large treated samples.

However, large standard errors in small sample studies are not solely a consequence of the small treated samples used, but are affected also by the number of observed covariates. Hence, the estimated treatment effect's standard errors in small treated sample studies increase with an increasing number of observed covariates, even though the overall sample size increases with an increasing number of observed covariates, due to an increase in the minimum group ratio required to balance a study design. Thus, the number of observed covariates with small treated samples does not have an impact only on the required group ratio, but it affects also standard errors of treatment effect estimates. Although these conclusions are founded based on the simulation standard errors of estimated treatment effects, we believe that the use of standard error estimators, as proposed by Imbens 2004; Rubin and Thomas 1996; Schafer and Kang 2008, would likely not change these findings. Nevertheless, the investigation of appropriately estimated treatment effect standard errors, in cases of small samples, is beyond the scope of this thesis.

Additionally, our simulation results also show that the choice of a matching algorithm (i.e., greedy versus optimal) matters more with small treated samples than with moderately large treated samples. Yet, these results might be due to the fact that small treated samples require substantially larger pools of control units than moderately large treated samples; hence, closer matched pairs can be obtained with optimal matching algorithm when using small treated samples. Moreover, the use of different matching algorithms has a tiny effect on the minimum required group ratio for removing covariate imbalances from observational designs with small treated samples (i.e., optimal matching algorithm on average requires smaller group ratios for removing covariate imbalances than the greedy matching algorithm), whereas this effect is negligible for the cases of moderately large treated samples. Besides, the simulation results with small treated samples also show on average smaller treatment effect standard errors when optimal matching algorithm is used, whereas this is not the case for moderately large treated samples.

The results of the simulation study extensions show that the correlation between the observed covariates and the outcome variable does not play a role in propensity score study when the treatment assignment mechanism is strongly ignorable. Also different class of the outcome variable (binary versus continuous) does not play a role in removing covariate imbalances from observational designs, according to the balance diagnostics used and described in Section 4.3.2. Yet, an additional regression adjustment, in the case of a binary outcome variable, has to be performed differently from the one performed in the case of a continuous outcome variable. Because it is unclear how to perform additional regression adjustment in the most optimal way, when the outcome variable is binary, and because such an investigation is beyond the scope of this thesis this would be our first recommendation for the future research of small sample properties in propensity score methods.

The Real Data Set 1 application (Section 6.1), which is based on the Lalonde data (1986), shows a high level of consistency with the results obtained with our theoretical simulation study (with respect to the small treated samples and minimum required group ratios) and the simulation study using real data. Hence, the application supports our conclusions regarding the performance of propensity score methods with small samples.

Future research on this topic should also include mixed types of covariates (continuous and discrete), investigate other important matching estimators like Mahalanobis distance matching on originally observed covariates, and use different types of response surfaces (e.g., parallel or non-linear). Additionally, future research should investigate the selection of variables for estimating propensity scores with small samples and how to estimate appropriately treatment effect standard errors in cases of small samples.

References

Abadie, A., Drukker, D., Herr, J. L. and Imbens, G. W. 2004. Implementing matching estimators for average treatment effects in Stata. *The Stata Journal*, 4: 290-311.

Abadie, A. and Imbens, G. W. 2002. *Simple and bias-corrected matching estimators,* Berkeley: Department of Economics, University of California.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91: 444-472.

Austin, P. C. 2011. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46, (3): 399-424.

Austin, P. C., Grootendorst, P. and Anderson, G. M. 2007. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine*, 26: 734-753.

Becker, S. and Ichino, A. 2002. Estimation of average treatment effects based on propensity scores. *The Stata Journal*, 2: 358-377.

Bourdieu, P. 1986. The forms of capiatl. In: *Education: Culture, Economy, and Society*, 46-58. Oxford: Oxford University Press.

Brookhart, M. A. et al. 2006. Variable Selection for Propensity Score Models. *American Journal of Epidemiology*, 163, (12): 1149-1156.

Buuren, S. v. and Groothuis-Oudshoorn, K. 2011. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45, (3).

Cochran, G. W. 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24: 295-313.

--- 1965. The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society*, 128: 234-255.

Cochran, G. W. and Rubin, D. B. 1973. Controlling bias in observational studies: A review. *Sankhya*, 35: 471-446.

Cochran, W. G. and Cox, G. M. 1957. *Experimental design.* New York: Wiley.

Cook, T. D., Shadish, W. R. and Wong, V. C. 2008. Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons. *Journal of Policy Analysis and Management*, 27, (4): 724-750.

Cornfield, J. et al. 1958. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22: 173-203.

Cox, D. R. 1958. *The Planning of Experiments.* New York: Wiley.

Crump, R., Hotz, V. J., Imbens, G. W. and Mitnik, O. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96, (1): 187-199.

Czajka, J. C., Hirabayashi, S. M., Little, R. J. and Rubin, D. B. 1992. Projecting from advance data using propensity modeling. *Journal of Business and Economics Statistics*, 10: 117-131.

Dawid, A. P. 1979. Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society*, 41: 1-31.

Dehejia, H. R. and Wahba, S. 1999. Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of American Statistical Association*, 94: 1053-1062.

Dehejia, R. H. and Wahba, S. 2002. Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84, (1): 151-161.

Drake, C. 1993. Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. *Biometrics*, 49, (4): 1231-1236.

Fisher, R. A. 1925. *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Fisher, R. A. 1935. *The design of experiments.* Edinburgh: Oliver and Boyd.

Gilligan, M. J. and Sergenti, E. J. 2008. Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference. *Quarterly Journal of Political Science*, 3: 89-122.

Gordon, S. C. and Huber, G. A. 2007. The Effect of Electoral Competitiveness on Incumbent Behavior. *Quarterly Journal of Political Science*, 2: 107-138.

Gutman, R. and Rubin, D. B. 2012. Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. *Statistics in Medicine*, 32, (11): 1795-1814.

Gu, X. S. and Rosenbaum, P. R. 1993. Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *Journal of Computational and Graphical Statistics*, 2, (4): 405-420..

Hansen, B. B. 2004. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99, (467): 609-618.

Hansen, B. B. and Klopfer, S. O. 2006. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15, (3): 609-627.

Heckman, J. J., Hidehiko, H. and Todd, P. 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Review of Economic Studies, 64: 605-654.

Helmreich, J. E. and Pruzek, R. M. 2009. PSAgraphics: Propensity score analysis graphics. *Journal of Statistical software*, 29.

Herron, M. C. and Wand, J. 2007. Assessing partisan bias in voting technology: The case of the 2004 New Hampshire recount. *Electoral Studies*, 26: 247-261.

Hill, J. and Jerome, R. P. 2006. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, 25: 2230-2256.

Hirano, K. and Imbens, G. 2001. Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Hear Catherization. *Health Services and Outcome Research Methodology*, 2: 259-278.

Hirano, K., Imbens, G. W. and Ridder, G. 2003. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, 74, (4): 1161-1189.

Ho, D. E., Imai, K., King, G. and Stuart, E. A. 2011. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, 42, (8): 1-28.

Holland, P. W. 1986. Statistic and Causal Inference. *Journal of the American Statistical Association*, 81, (396): 945-960.

Holland, P. W. and Rubin, D. B. 1988. Causal Inference in Retrospective Studies. *Evaluation Review*, 12, (3): 203-231.

Horvitz, D. G. and Thompson, D. J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47: 663-685.

Hulswit, M. 2002. *From Cause to Causation. A Peircean Perspective.* Dordrecht: Kluwer Publishers.

Idler, E. L. and Benyamini, Y. 1997. Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies. *Journal of Health and Social Behavior*, 38, (1): 21-37.

Imbens, G. W. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87: 706-710.

Imbens, G. W. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86, (1): 4-29.

Joffe, M. M., Ten Have, T. R., Feldman, H. I. and Kimmel, S. E. 2004. Model selection, confounder control, and marginal structural models. *The American Statistician*, 58, (4): 272-279.

Kamin, T., Kolar, A. and Steiner, P. M. 2013. The role of cultural capital in producing good health: a propensity score study / Vpliv kulturnega kapitala na zdravje: študija nagnjenja. *Slovenian Journal of Public Health*, 52, (2): 108–118.

Kang, J. D. and Schafer, J. L. 2007. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22: 523-539.

Keele, L. J. 2011. *rbounds: Perform Rosenbaum bounds sensitivity tests for matched and unmatched data*: R package version 0.9.

Kempthorne, O. 1952. *The Design and Analysis of Experiments.* New York: Wiley.

LaLonde, J. R. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, 76, (4): 604-620.

Lavrakas, P. J. ed. 2010. *Encyclopedia of Survey Research Methods.* New York: Sage Publications.

Leuven, E. and Sianesi, B. 2003. *psmatch2.*

Liublinska, V. and Rubin, D. B. 2012. *Enhanced Tipping-Point Displays.* San Diego, American Statistical Association.

Love, T. E. 2002. *Displaying Covariate Balance After Adjustment for Selection Bias.* Available at: http://www.chrp.org/love/JSM_Aug11_TLove.pdf (January 20, 2013).

Luellen, J. K. 2007. *A Comparison of Propensity Score Estimation and Adjustment Methods on Simulated Data" unpublished doctoral dissertation,* University of Memphis, Dep. of Psychology.

Lunceford, K. J. and Davidian, M. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparitive study. *Statistics in Medicine*, 23: 2937-2960.

Morgan, S. L. and Harding, D. J. 2006. Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice. *Sociological Methods and Research,* 35, (1): 3-60.

Morgan, S. L. and Winship, C. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research).* Edinburgh: Cambridge University Press.

Neyman, J. S. 1923. Statistical problems in agricultural experiments. *Journal of the Royal Statistical Society*, 2*:* 107-180.

R Core Team, 2012. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing.*

Raessler, S. and Rubin, D. B. 2005. *Complications When Using Nonrandomized Job Training Data to Draw Causal Inferences.* Sydney, Proceedings of the International Statistical Institute.

Ridgeway, G., McCaffrey, D. and Morral, A. 2012. *Toolkit for weighting and analysis of non-equivalent groups,* R package version 1.2-5.

Rosenbaum, P. R. 1986. Dropping out high school in the United States: An observational study. *Journal of Educational Statistics*, 11: 207-224.

--- 1989. Optimal Matching in Observational Studies. *Journal of the American Statistical Associations*, 84: 1024-1032.

--- 1991. A characteriztion of optimal design for observational studies. *Journal of the Royal Statistical Society, Series B (Methodological)*, 53: 597-610.

--- 2002. Covariance Adjustment in Randomized Experiments and Observational Studies. *Statistical Science*, 17, (3): 286-327.

--- 2002. *Observational Studies.* New York: Springer.

--- 2005. Sensitivity Analysis in Observational Studies. In: B. S. Everitt and D. C. Howell, eds. *Encyclopedia of Statistics in Behavioral Science.* Chichester: John Wiley and Sons: 1809-1814.

--- 2010. *Design of Observational Studies.* New York: Springer.

Rosenbaum, P. R. and Rubin, D. B. 1983a. The Central Role of the Propensity Score in Observational Studies for Causal Effect. *Biometrika*, 70, (1): 41-55.

--- 1983b. Assessing sensitivity to an unobserved binary covariates in an observational study with binary outcome. *Journal of the Royal Statistical Society, B,* 45, (2): 212-218.

--- 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79: 516-524.

--- 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39: 33-38.

Rubin, D. B. 1973. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics*, 29, (1): 184-203.

--- 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66*:* 688-701.

--- 1975. Bayesian Inference for Causal Effects: The importance of Randomization.The Annals of Statistics, 6, (1): 233-239.

--- 1976. Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics*, 32: 109-120.

--- 1977. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2: 1-26.

--- 1978. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6: 34-58.

--- 1979. Using Multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the Royač Statistical Society*, 41: 318-328.

--- 1980. Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *Journal of the American Statistical Association*, 74*:* 318-328.

--- 1987. *Multiple imputation for nonresponse in surveys.* New York: Wiley.

--- 1990. Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25: 279-292.

--- 1991. Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism. *Biometrics*, 47, (4): 1213-1234.

--- 1996. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, 91, (434): 473-489.

--- 1997. Estimating causal effects from large data sets using the propensity score. *Annals of Internal Medicine*, 127: 757-763.

--- 2001. Using propensity score to help design observational studies: application to the tobacco litigation. *Health ServicesandOutcomes Research Methodology*, 2: 169-188.

--- 2006. *Matched Sampling for Causal Effects.* New York: Cambridge University Press.

--- 2007. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26: 20-36.

--- 2008. For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics*, 2, (3): 808-840.

Rubin, D. B. and Imbens, G. 2008. "Rubin Causal Model". In: S. M. Durlauf and C. E. Blume, eds. *The New Palgrave Dictionary of Economics.* New York: Palgrave McMillan: 255-262.

Rubin, D. B. and Liublinska, V. 2012. *Enhanced Tipping-Point Displays,* Ljubljana. Available at: http://www.fdvinfo.net/uploadi/editor/1355130373SFO_2012_2_EnhancedTippingPointDisplays_Oct2.pdf

Rubin, D. B. and Thomas, N. 1992a. Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions. *Biometrika*, 79: 797-809.

--- 1992b. Affinely Invariant Matching Methods with Ellipsoidal Distributions. *The Annals of Statistics*, 2, (20): 1079-1093.

--- 1996. Matching using estimated propensity scores, relating theory to practice. *Biometrics*, 52: 249-264.

--- 2000. Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. *The Journal of the American Statistical Association*, 95, (450): 573-585.

Rubin, D. B. and Waterman, R. P. 2006. Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology. *Statistical Science*, 21, (2): 206-222.

Schafer, J. L. and Kang, J. 2008. Average Causal Effects From Nonrandomized Studies: A Practical Guide and Simulated Example. *Psychological Methods*, 13, (4): 279-313.

Sehkon, J. S. and Grieve, R. 2011. A Nonparametric Matching Method for Covariate Adjustment with Application to Economic Evaluation (Genetic Matching). *Health Economics*, 21, (6): 695-714.

Sekhon, J. and Mebane, W. 1998. Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models. *Political Analysis*, 7: 189-213.

Sekhon, J. S. 2011. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. *Journal of Statistical Software*, 42, (7): 1-52.

Sekhon, J. S. and Diamond, A., Forthcoming. Genetic Matching fo Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics.*

Shadish, W. R., Clark, M. H. and Steiner, P. M. 2008. Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *Journal of the American Statistical Association*, 103, 484: 1334-1343.

Shadish, W. R., Cook, D. T. and Campbell, D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

Shadish, W. R. and Steiner, P. M. 2010. A primer on Propensity Score Analysis. *Elsevier*, 10, (1): 19-26.

Siroky, D. S. 2009. Navigating Random Forests and related advances in algorithmic modeling. *Statistics Surveys*, 3: 147-163.

Steiner, P. M. 2012. Design-Based and Model-Based Analysis of Propensity Score Designs. *Working paper.*

Steiner, P. M. and Cook, T. D., (in press). Matching and Propensity Scores. In: T. D. Little, ed. *The Oxford Handbook of Quantitative Methods.* New York: Oxford University Press.

Steiner, P. M., Cook, T. D. and Shadish, W. R. 2011. On the Importance of Reliable Covariate Measurement in Selection Bias Adjustments Using Propensity Scores. *Journal of Educational and Behavioral Staitstics*, 36, (2): 213-236.

Stuart, E. A. 2010. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25, (1): 1-21.

Stuart, E. A. and Green, K. M. 2008. Using full matching to estimate causal effects in non-experimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Development Psychology*, 44, (2): 395-406.

Stuart, E. A. and Rubin, D. B. 2007. Best practices in quasi-experimental designs: Matching methods for causal inference. In: *Best Practices in Quantitative Methods,* 155-176. New York: Sage Publications.

Thoemmes, F. J. and Kim, S. E. 2011. A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research*, 46: 90-118.

van Buuren, S. and Groothuis-Oudshoorn, K. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45, (3): 1-67.

Waernbaum, I. 2010. Propensity score model specification for estimation of average treatment effects. *Journal of Statistical Planning and Inference*, 140, (7): 1948-1956.

--- Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Statistics in Medicine*, 31, (15): 1572-1581

Westreichab, D., Lesslerc, J. and Jonsson-Funk, M. 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63, (8): 826-833.

Yan, X., Lee, S. and Li, N. 2009. Missing data handling methods in medical device clinical trials. *Journal of Biopharmaceutical Statistics*, 19, (6): 1085-1098.

Zhao, Z. 2004. Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. *The Review of Economics and Statistics*, 86, (1): 91-107.

Index

Povzetek disertacije v slovenskem jeziku

Klasične statistične metodologije za vzročno sklepanje so bile razvite v eksperimentalnih študijah in njihova uporaba v opazovalnih študijah ni priporočljiva. Pred razvojem metod nagnjenja (propensity score methods) je bilo v opazovalnih študijah, pri razlagi povezav med različnimi dogodki, dolgo časa priporočljivo uporabljati le opisno statistiko, kajti kakršnokoli vzročno sklepanje bi, zaradi narave podatkov (neslučajna selekcija enot, katerim obravnava je ali ni dodeljena), ne podalo zaupanja vrednih ocen (Cochran 1965).

Pod definicijo »opazovalne študije« razumemo vse ne-eksperimentalne študijske zasnove (t.j., anketiranje), študijske zasnove kjer obravnava (treatment) ni slučajno dodeljena enotam (t.j., kvazi-eksperimentalne študije (Shadish, et al. 2002)) in študijske zasnove kjer popolno slučajenje (randomisation) obravnavanega stanja (treatment condition) spodleti, t.j., kršena (broken) slučajna zasnova poskusa (Barnard, et al. 2003). Glavna razlika med opazovalnimi študijami in slučajno zasnovo poskusa je tako v procesu selekcije (selection procedure) – kako je obravnava dodeljena posamezni enoti – kakšen je proces, ki narekuje, kateri enoti je obravnava dodeljena in kateri enoti ni dodeljena.

Pri slučajni zasnovi poskusa je proces dodeljevanja obravnav enotam kontroliran s strani raziskovalca. Tako raziskovalec poskuša zagotoviti, da so enote, katerim so dodeljene različne obravnave, primerljive (imajo enake karakteristike – njihove porazdelitve sospremenljivk so identične). Ta primerljivost je zagotovljena z naključnim dodeljevanjem različnih obravnav enotam, kajti slučajenje (randomisation) lahko v večini primerov zagotovi, da obravnava, uporabljena pri eni enoti, ne vpliva na izid (outcome) pri drugi enoti, pri kateri obravnava je ali ni bila dodeljena. Pomembna lastnost slučajnega dodeljevanja obravnav enotam je tako stroga neodvisnost med mehanizmom dodeljevanja[34] (assignment mechanism) in izidom.

Tako so opazovane in neopazovane sospremenljivke, pri slučajni zasnovi poskusa, primerljive med skupinama, katerima so dodeljene različne obravnave (t.p., porazdelitve opazovanih in neopazovanih sospremenljivk, med obravnavano in kontrolno skupino[35], so v povprečju enake). Takšna študijska zasnova je statistično uravnotežena (statistically balanced), glede na porazdelitve sospremenljivk med obema skupinama. Možna neuravnoteženost, v porazdelitvi sospremenljivk med skupinama, je tako zgolj naključna in ni posledica sistematične izbire (systematic selection procedure), ki bi lahko povzročila pristranskost ocen vzročnih učinkov (Rosenbaum 2002, 21).

V opazovalnih študijah je proces dodeljevanja obravnave le delno kontroliran s strani raziskovalca, ali sploh nekontroliran. Tako slučajenje (naključna dodelitev različnih obravnav enotam) ni izvedljivo, kar običajno vodi v neprimerljivost enot, katerim je dodeljena različna obravnava (porazdelitve opazovanih sospremenljivk med obravnavano in kontrolno skupino so v povprečju različne – statistično neuravnotežena študijska zasnova). V takšnih primerih zahteva ocenjevanje vzročnih učinkov poseben pristop: da bi lahko nepristransko ocenili vzročne učinke, moramo iz študijske zasnove najprej odpraviti pristranskost.

Ločimo med dvema takšnima pristopoma: (i) pristop na osnovi načrta zasnove (design-based approach) – temelj metod nagnjenja; in (ii) pristop na osnovi modela (model-based approach). Razvoj teh pristopov je motiviran glede na dejstvo, da je večina študijskih zasnov opazovalnih, ker slučajne zasnove poskusov ne samo, da so stroškovni zalogaj, ampak v večini primerov niso izvedljive.

[34] Mehanizem dodeljevanja nam poda informacijo o tem, katerim enotam je obravnava dodeljena in katerim ni.

[35] Obravnavana skupina (treatment group) je skupina enot, kateri je obravnava (treatment) dodeljena. Kontrolna skupina (control group) je skupina enot, katerim obravnava ni dodeljena.

1. Vzročnost in razvoj metod nagnjenja

Ideja vzročnosti je zelo stara in gre nazaj v čas velikih filozofov kot so Platon, Aristotel, Hume, Mill in drugi. Glede na Hulswita (2002) je bil Platon tisti, ki je prvi formuliral princip vzročnosti s stavkom: «vse kar se zgodi oziroma spremeni, se spremeni zaradi nekega vzroka; nič se ne more zgoditi brez vzroka«. Ti zgodnji filozofi tako gledajo na vzročnost s perspektive, da je potrebno najti vzrok učinka, ki ga vidimo.

Nasprotno pa statistična družba gleda na vzročnost z drugega zornega kota: enote so manipulirane z neko znano intervencijo (vzrok je znan) in cilj je nepristransko oceniti učinek, povzročen s takšno intervencijo. Vzrok je tako predstavljen z aktivno intervencijo, ki je dodeljena nekaterim enotam v določenem času, zato da lahko raziščemo, kako drugače bi se te enote obnašale od enot, ki niso bile manipulirane s to isto intervencijo. Tako ocena učinka intervencije predstavlja vzročni učinek oziroma učinek obravnave (treatment effect), t.j., učinek, povzročen z dodelitvijo specifične obravnave – intervencije – nekaterim enotam. Oba strokovna izraza (t.j., vzročni učinek in učinek obravnave) se tako izmenično uporabljata.

Razvoj metod vzročnega sklepanja v opazovalnih študijah se je resno začel šele v sedemdesetih in začetku osemdesetih let z delom Rubina (1974; 1977; 1978; 1980), Rosenbauma in Rubina (1983a), Holland-a in Rubin-a (1988), Angrista, Imbensa in Rubina (1996), in Rosenbauma (2002). Osnova njihovih del temelji na slučajni zasnovi poskusov in je tako nadaljevanje idej del Neymana (1923), Fisherja (1925), Kempthorneja (1952), Cochrana in Coxa (1957), in Coxa (1958).

Neymanova notacija možnih izidov je temelj razvoja metod vzročnega sklepanja v opazovalnih študijah. Osnovni cilj je odgovoriti na vzročno vprašanje: kakšen bi naj bil izid (outcome) obravnavane skupine, če le-ta ne bi bila obravnavana, in vice versa.

Vzročni učinek je tako definiran kot razlika med obema hipotetično dobljenima izidoma:

$$\tau = Y(1) - Y(0),$$

kjer τ označuje vzročni učinek, $Y(1)$ je možni izid za enote, katerim obravnava je dodeljena, in $Y(0)$ je možni izid za enote, katerim obravnava ni dodeljena.

V sedemdesetih, Rubin (1974, 1975, 1978) razširi Neymanovo notacijo možnih izidov na opazovalne študije s tem, ko v notacijo možnih izidov tudi formalno vključi informacijo o mehanizmu dodeljevanja (assignment mechanism). Glede na Rubinov velik znanstveni prispevek, se tako okvir možnih izidov pogosto imenuje Rubinov model vzročnosti (Holland 1986) – osnova metod nagnjenja – dandanes najbolj razširjene metode za ocenjevanje vzročnih učinkov v opazovalnih študijah.

2. Cilji doktorske disertacije

Razvoj metod nagnjenja je v zadnjih desetletjih rezultiral v jasnih smernicah ocenjevanja vzročnih učinkov z velikimi vzorci, vendar pa vprašanje »kako velik« vzorec je potreben za »uspešno implementacijo«[36] metod nagnjenja, ostaja neodgovorjeno. Majhni vzorci so pogosti v družbenih vedah (npr., izobraževanje: število študentov v razredu, število šol; medicina: število bolnikov z redko boleznijo), zato je osvetlitev delovanja metod nagnjenja z majhnimi vzorci zelo pomembna (Shadish in Steiner 2010).

Cilj prvega izvirnega prispevka k razvoju področja metod nagnjenja je tako raziskati vlogo velikosti vzorca v metodah nagnjenja in pri tem preučiti, s pomočjo simulacijske študije, kako dobro delujejo metode nagnjenja z malimi vzorci in kaj so najmanjši možni vzorci, ki še omogočajo uspešno odstranitev pristranskosti iz študijske zasnove.

[36] »Uspešna implementacija« pomeni, da smo uspeli uravnotežiti študijsko zasnovo, glede na opazovane sospremenljivke, tako uspešno, da je možna preostala neuravnoteženost (residual balance) brezpomembna.

Poleg tega, doktorska disertacija podaja še dva izvirna prispevka: (i) natančna definicija metod nagnjenja, s pomočjo katere pripomoremo k razjasnitvi literature s področja metod nagnjenja, na podlagi katere so metode tako pogosto napačno uporabljene; (ii) razširitev uporabne vrednosti metod nagnjenja tudi na opazovalne študije, kjer narava podatkov ne omogoča ocenjevanja vzročnih učinkov, zato lahko ocenjujemo le pogojne asociacije, glede na sospremenljivke za katere raziskovalec meni, da so bistvenega pomena.

Doktorska disertacija ob koncu poda tudi dve aplikativni študiji. Namen prve aplikativne študije je preveriti, kako aplikativni so rezultati naše simulacijske študije v praksi. Namen druge aplikativne študije je predstaviti realne podatke, kjer ni mogoče zanesljivo oceniti vzročnih učinkov, zato ocenimo le pogojne asociacije glede na izbor sospremenljivk, ki so v kontekstu študije v našem največjem interesu.

3. Metode nagnjenja

Metode nagnjenja so bile primarno razvite z namenom ocenjevanja vzročnih učinkov v opazovalnih študijah. V doktorski disertaciji razširimo uporabno vrednost metod tudi na opazovalne študije, kjer raziskovalna vprašanja morda so vzročna, vendar pa narava podatkov ne omogoča ocene vzročnih učinkov. V teh primerih so lahko metode nagnjenja uporabljene z namenom ocenjevanja pogojnih asociacij, glede na sospremenljivke za katere raziskovalec meni, da so bistvenega pomena. Da bi lahko enostavneje razumeli naravo podatkov, ki ne omogoča ocenjevanje vzročnih učinkov, podamo najprej definicijo o tem kaj je vzročno, in glede na to definicijo potem izpeljemo definicijo za pogojne asociacije.

Ko ocenjujemo vzročne učinke obravnav (treatment) v. neobravnav (control) moramo biti zmožni definirati (1) intervencijo, ki bi lahko bila dodeljena vsem »obravnavanim« enotam in jih tako spremeniti v »neobravnavane« enote (npr., namesto zdravila damo placebo), in (2) podobno intervencijo, ki bi lahko bila dodeljena vsem »neobravnavanim« enotam in jih tako spremenila v »obravnavane« enote (npr., namesto placebo damo zdravilo). Vse te resnične ali hipotetične verzije (1) in (2) morajo voditi v iste možne izide obravnavanih enot, $Y(1)$, in neobravnavanih enot $Y(0)$, zato da sta izpolnjeni naslednji dve predpostavki : (i) obravnava, dodeljena eni enoti, ne sme vplivati na izid druge enote, ne glede na to, če je drugi enoti obravnava bila dodeljena ali ne (Cox 1958); in (ii) za vsako enoto obstaja samo en tip obravnave oziroma neobravnave[37]. Vse meritve, ki so narejene oziroma vsaj določene preden je intervencija (1) ali (2) dodeljena posamezni enoti, predstavljajo osnovne sospremenljivke (baseline covariates), in vse meritve, ki so narejene po dodelitvi intervencije, predstavljajo spremenljivko izida (outcome variable).

Ko nam narava opazovalnih podatkov ne dopušča prepričljivo formulirati intervencije, kot smo jo definirali zgoraj, ne moremo ocenjevati vzročnih učinkov obravnave v. neobravnave. V takšnih primerih lahko zato ocenjujemo le pogojne asociacije med dihotomko Z in drugo spremenljivko Y, pogojno na sospremenljivke X, ki so izbrane s strani raziskovalca kot posebej zanimive. Primer takšnih podatkov je ocenjevanje učinka statusa manjšin na vključenost v izobraževalni sistem. Na primer, če imamo dve skupini študentov (npr., kavkazijskega in afriškega izvora) je pri tem nemogoče spremeniti raso kavkazijcev v raso afričanov in vice versa, zato lahko v tem primeru ocenjujemo samo pogojne asociacije.

Kot že rečeno, so metode nagnjenja osnovane na podlagi pristopa na osnovi načrta zasnove (design-based approach). Njihova uporaba sestoji iz dveh pomembnih delov: (i) faza načrta, kjer poskrbimo za statistično uravnoteženo študijsko zasnovo, glede na opazovane sospremenljivke (brez uporabe informacij o izidu). Ta faza vključuje orodja za uravnoteženje študijske zasnove in orodja za ocenjevanje uravnoteženosti študijske zasnove; in (ii) faza analize, kjer uporabimo informacijo o izidu z namenom ocene vzročnih učinkov ali pogojnih asociacij. V fazi analize pa

[37] Te predpostavke so del SUTVA (Stable Unit Treatment Value Assumption) (Rubin 1990)

lahko izvedemo tudi dodatna statistična uravnavanja, in v primeru ocenjevanja vzorčnih učinkov, izvedemo tudi analizo občutljivosti (sensitivity analysis) dobljenih ocen. Analiza občutljivosti pa ni potrebna v primeru uporabe metod nagnjenja z namenom ocenjevanja pogojnih asociacij.

Glavni cilj faze načrta, pri ocenjevanju vzročnih učinkov v opazovalnih študijah, je uspešna odprava pristranskosti iz študijske zasnove. Študijske zasnove, kjer je pristranskost prisotna, imenujemo neuravnotežene zasnove (unbalanced designs) zaradi neuravnoteženosti sospremenljivk med skupinama, katerima je dodeljena različna obravnava (npr., skupina enot, katerim je obravnava dodeljena ($W=1$) – obravnavana skupina, in skupina enot, katerim obravnava ni dodeljena ($W=0$) – kontrolna skupina). Raven pristranskosti je tako lahko ponazorjena z ravnjo neuravnoteženosti sospremenljivk med obravnavano in kontrolno skupino.

Na drugi strani, je glavni cilj faze zasnove, pri ocenjevanju pogojnih asociacij v opazovalnih študijah, učinkovito kontrolirati sospremenljivke, ki jih raziskovalec izbere kot posebej zanimive. Pogojevati na X pomeni poiskati enote z $Z=1$ in enote z $Z=0$, ki imajo identično vrednost X oziroma enote z $Z=1$, ki imajo enako distribucijo X kot enote z $Z=0$. Bolj kot je porazdelitev X v $Z=1$ enotah, v povprečju, podobna porazdelitvi X v $Z=0$ enotah, bolj uspešno kontroliramo X v tej študiji primerljivosti. V tem smislu uravnotežimo študijsko zasnovo glede na X, ker pa pri takšnih študijah ne moremo govoriti o mehanizem dodeljevanja, neuravnoteženost študijske zasnove, glede na X, ne sme biti razumljena kot pristranskosti.

3.1 Faza načrta

Faza načrta vsebuje orodja za uravnoteženje študijske zasnove in orodja za ocenjevanje uspešnosti uravnoteženja študijske zasnove. Orodja za uravnoteženje vsebujejo tehnike in metode za odpravo pristranskosti v študijski zasnovi pri ocenjevanju vzročnih učinkov, oziroma uravnovešenje študijske zasnove, glede na posebej zanimive sospremenljivke, pri ocenjevanju pogojnih asociacij. V obeh primerih je cilj doseči uravnoteženo študijsko zasnovo, glede na opazovane sospremenljivke. Orodja za ocenjevanje procesa uravnavanja študijske zasnove (balance assessment tools), ki ocenjujejo uspešnost procesa uravnavanja študijske zasnove (kako uspešno sta obravnavana in kontrolna skupina uravnoteženi, glede na opazovane sospremenljivke) pa morajo biti uporabljena pred in med fazo načrta.

Glavni element orodij uravnoteženja študijske zasnove je stopnja nagnjenja (propensity score) (Rosenbaum in Rubin 1983), ki je balansirana stopnja (balancing score) in tako pomembna komponenta v procesu uravnoteženja študijske zasnove (odprave pristranskosti pri ocenjevanju vzročnih učinkov, oziroma odprave neuravnoteženosti sospremenljivk, med dvema skupinama, pri ocenjevanju pogojnih asociacij). Stopnja nagnjenja, $e(X)$, je definirana kot pogojna verjetnost, da je enota obravnavana $W=1$, glede na opazovane sospremenljivke, X:

$$e(X) = pr(W = 1 \mid X),$$

kar pomeni, da sta W in X pogojno neodvisna glede na $e(X)$ (Rosenbaum in Rubin 1983a). Nagnjenje je tako funkcija opazovanih sospremenljivk. Njen glavni cilj je v uravnoteženju obravnavane in kontrolne skupine glede na sospremenljivke. Tako enote v obravnavani in kontrolni skupini, ki imajo približno enako vrednost nagnjenja, rezultirajo v približno podobnih porazdelitvah njunih osnovnih sospremenljivk (Rosenbaum in Rubin 1985). Ko sta obe skupini (obravnavana in kontrolna) uravnoteženi, glede na osnovne sospremenljivke, je vzročni učinek razlika med izidi obeh skupin (enako velja v primeru ocenjevanja pogojnih asociacij).

Ocene stopenj nagnjenja so verjetnosti za razvrstitev obravnave, pogojno na opazovane sospremenljivke. Tako na stopnje nagnjenja v glavnem vplivata dva dejavnika: izbira sospremenljivk za oceno modela nagnjenja in izbrana metoda za ocenjevanje stopnje nagnjenja. Pri ocenjevanju vzročnih učinkov Rubin in Thomas (1996) predlagata, da bi naj v model nagnjenja vključili vse sospremenljivke, povezane z izidom, četudi morda niso močno povezane z obravnavo. Rubin (1997) nadalje predlaga, da tudi če je sospremenljivka samo šibko povezana z izidom in je hkrati povezana z obravnavo, mora biti vključena v model, saj bi se v nasprotnem primeru pristranskost povečala močneje, kot bi bila izguba učinkovitosti (efficiency), zaradi njene vključitve. Izguba učinkovitosti se zmanjša, ko vključimo spremenljivko, ki je povezana z obravnavo in nepovezana z izidom. V primeru ocenjevanja pogojnih asociacij, raziskovalec vključi v model nagnjenja sospremenljivke, za katere meni, da so bistvenega pomena za uspešno uravnoteženje študijske zasnove. V skladu s tem je raziskovalec dolžan ustrezno zagovarjati izbor sospremenljivk.

Na podlagi izbranih sospremenljivk ocenimo stopnje nagnjenja, katerih glavni namen je v efektivni kontroli sospremenljivk. Stopnje nagnjenja so lahko ocenjene s pomočjo diskriminantne analize ali logistične regresije, pod pogojem, da sospremenljivke ne vsebujejo manjkajočih podatkov. V zadnjem času so se začele uporabljati tudi nekatere druge metode, kot so razvrstitveno drevo (classification tree) (Westreichab, in drugi 2010) in odborne metode (ensemble methods), kot so okrepljena regresija (boosted regression) (Mccaffrey, in drugi 2004) in slučajni gozd (random forest) (Siroky 2009).

Na podlagi ocenjenih stopenj nagnjenja, uporabimo tako imenovane metode uravnavanja (adjustment methods), kot so usklajevanje (matching), subklasifikacija/stratifikacija (subclassification/stratification), ali uteževanje. Vloga teh metod je v uravnoteženju obravnavane in kontrolne skupine glede na sospremenljivke.

Preden so možni izidi obeh skupin lahko primerjani in tako ocenjen vzročni učinek ali pogojna asociacija, je potrebno diagnosticirati kvaliteto uravnoteženosti študijske zasnove. Pogosto uporabljena tehnika, za takšno diagnostiko, je standardizirana razlika v povprečnih vrednostih sospremenljivk med obravnavano in kontrolno skupino (Rosenbaum in Rubin 1985). Tri preostale kvantitativne tehnike so: razlika v povprečnih vrednostih stopenj nagnjenja med obema skupinama, razmerje varianc stopnje nagnjenja obeh skupin, in razmerje ostankov varianc sospremenljivk po uravnavanju (Rubin 2001). Poleg kvantitativnih tehnik pa je priporočljivo uporabiti tudi grafična orodja, kot so Q-Q diagrami ali grafikoni kvantilov (Ho in drugi 2007), ali grafična orodja, ki vključujejo tudi kvantitativno informacijo (Love 2002; Steiner in drugi 2011).

3.2 Faza analize

Faza analize vključuje oceno vzročnih učinkov ali pogojnih asociacij, dodatna uravnoteženja sospremenljivk z namenom odstranitve ostankov neuravnoteženosti sospremenljivk med skupinama (t.j., neuravnoteženost po dokončanju faze načrta), kot tudi analizo občutljivosti ocen vzročnih učinkov. Cilj analize občutljivosti je odgovoriti na vprašanje, kako bi se pridobljene ocene vzročnih učinkov lahko spremenile ob prisotnosti skrite pristranskosti (hidden bias), in kako obsežna bi morala biti skrita pristranskost, da bi spremenila naše zaključke glede ocen vzročnih učinkov. Analiza občutljivosti ni potrebna v primeru ocenjevanja pogojnih asociacij, ker v teh primerih ne govorimo o pristranskosti, ampak le o neuravnoteženi študijski zasnovi.

4. Vloga velikosti vzorca

Glavni problem majhnih vzorcev na splošno, v statističnem sklepanju, so velike standardne napake. Tako je jasno, da je statistično sklepanje na podlagi majhnih vzorcev manj natančno, kot v primeru velikih vzorcev. To pa ni edini problem s katerim se soočamo, ko ocenjujemo vzročne učinke ali pogojne asociacije z metodami nagnjenja v primeru majhnih vzorcev.

Da bi lahko odpravili neuravnoteženost v sospremenljivkah med dvema skupinama, moramo najprej oceniti stopnje nagnjenja. Cilj ocenjenih stopenj nagnjenja je v tem, da le-te predstavljajo uravnotežene stopnje (balancing scores). Manjši kot je vzorec, manj natančne so ocene stopenj nagnjenja, in ta natančnost se verjetno zmanjšuje, ko se število opazovanih sospremenljivk povečuje. Tako lahko uravnoteženje študijske zasnove z manj natančnimi stopnjami nagnjenja rezultira v večjem ostanku neuravnoteženosti (neuravnoteženost v sospremenljivkah med skupinama po izvedbi faze načrta metod nagnjenja).

Poleg tega so lahko opazovalne študije z majhnimi vzorci problematične tudi zaradi pomanjkanja prekrivanja (overlap) med porazdelitvijo sospremenljivk v obravnavani in kontrolni skupini, kot tudi v doseganju področja skupne podpore (common support). Slabo prekrivanje ali šibko doseganje področja skupne podpore, v primerih majhnih vzorcev, oteži proces uravnoteženja sospremenljivk med obravnavano in kontrolno skupino.

Pri pregledovanju publikacij metod nagnjenja z majhnimi vzorci, je bil naš interes predvsem v publikacijah, ki raziskujejo, kako dobro lahko uravnotežimo študijsko zasnovo, glede na opazovane sospremenljivke, in katera metoda uravnavanja je najprimernejša v primeru majhnih vzorcev. Tri simulacijske študije, ki vključujejo majhne in velike vzorce (Rubin in Thomas 1996; Zhao 2004; Luellen 2007) potrjujejo, da je uspešna implementacija metod nagnjenja (t.j., pridobitev primerljivih skupin) odvisna od velikosti vzorca. Tako te raziskave kažejo na to, da se metode nagnjenja pri malih vzorcih obnašajo drugače pri velikih.

Glede na teoretične lastnosti metod nagnjenja in glede na rezultate predhodnih raziskav o vlogi velikosti vzorca pri metodah nagnjenja smo sklenili, da je metoda usklajevanja večinoma najbolj primerna metoda uravnavanja v primeru majhnih vzorcev. Hkrati je metoda usklajevanja tudi najbolj pogosto uporabljena metoda uravnavanja pri ocenjevanju vzročnih učinkov z metodami nagnjenja.

Da bi lahko raziskali, kako dobro delujejo metode nagnjenja z malimi vzorci in kaj so najmanjši možni vzorci, ki še omogočajo uspešno odstranitev pristranskosti iz študijske zasnove, smo izvedli vrsto simulacijskih študij za raziskovanje različnih scenarijev.

5. Simulacijska študija

Niz simulacijskih študij vključuje majhne in srednje velike vzorce, različno število opazovanih sospremenljivk in različne stopnje začetne pristranskosti v študijskem načrtu (t.j., raven neuravnoteženosti študijske zasnove). Raziskujemo tudi različna razmerja med velikostjo vzorca kontrolne in obravnavane skupine.

Simulacijska študija metod nagnjenja je izvedena tako z ocenjenimi stopnjami nagnjenja, kot tudi s pravimi stopnjami nagnjenja. Na tak način preučujemo razlike v rezultatih za scenarij, ki smo ga deležni v realnem svetu z opazovalnimi podatki, in za scenarij, ki bi ga bili deležni v primeru slučajnih poskusov. Hkrati pa preučujemo tudi razlike v rezultatih študije nagnjenja, ko uravnavanje študijskega načrta izvajamo z dvema različnima algoritmoma usklajevanja (t.j., optimalni in požrešni (greedy) algoritem).

Analiza simulacijskih podatkov je narejena deskriptivno in z analizo variance. Rezultati kažejo, da je v primeru majhnih obravnavanih vzorcev mogoče uravnotežiti študijsko zasnovo enako uspešno, kot v primeru srednje velikih obravnavanih vzorcev. Vendar pa je uspeh študij nagnjenja (propensity score study) z majhnimi obravnavanimi vzorci (kako uspešno lahko uravnotežimo študijsko zasnovo), predvsem odvisen od velikosti skupine kontrolnih enot. Majhni obravnavni vzorci zahtevajo veliko večje razmerje med skupinama (t.j., med skupino kontrolnih enot in skupino obravnavanih enot) kot srednje veliki obravnavani vzorci. Hkrati, je pri majhnih obravnavanih vzorcih, zahtevano razmerje med skupinama, predvsem odvisno od velikosti skupine obravnavanih enot, števila opazovanih sospremenljivk in velikostjo začetne pristranskosti, oziroma ravni neuravnoteženosti študijske zasnove.

Manjši kot je vzorec skupine obravnavanih enot, več opazovanih sospremenljivk kot imamo in bolj kot je študijska zasnova neuravnotežena, glede na opazovane sospremenljivke, večja so zahtevana razmerja med skupinama (t.j., med kontrolno in obravnavano skupino). Po drugi strani pa ima število opazovanih sospremenljivk, pri srednje velikih vzorcih, le malenkosten vpliv na zahtevano razmerje skupin.

Razlog, da število opazovanih sospremenljivk igra tako pomembno vlogo pri majhnih vzorcih obravnavane skupine, je v procesu ocenjevanja stopenj nagnjenja. Več kot imamo opazovanih sospremenljivk in manjši kot je vzorec obravnavane skupine, težje je natančno oceniti stopnje nagnjenja.

Kakorkoli, natančna ocena stopenj nagnjenja ni edini problem majhnih vzorcev obravnavanih skupin. Pri zelo majhnih obravnavanih vzorcih $n_t \leq 50$ in majhnem razmerju med skupinama $R = 2,$ je logistična regresija, za ocenjevanje stopenj nagnjenja, v primeru z desetimi opazovanimi sospremenljivkami rezultirala v ekstremnih vrednostih 0 in 1. Tako za te primere ocenjujemo, da je predpostavka o strogi neodvisnosti kršena[38]. V takšnih primerih morda ni modro nadaljevati s študijo nagnjenja in ocenitvijo vzročnih učinkov ali pogojnih asociacij.

Glede na dobljene rezultate sklepamo, da je »uspeh« logistične regresije (logistična regresija ne rezultira v ekstremnih vrednostih 0 in 1) primarno odvisen od velikosti obravnavane skupine in števila opazovanih sospremenljivk. S 30 opazovanimi sospremenljivkami in s celotnim vzorcem $n = 104$, ter razmerjem med skupinama $R = 12$ (t.j., $n_t = 8$ in $n_c = 96$), logistična regresija rezultira v ekstremnih vrednostih, 0 in 1, pri vseh stopnjah začetne neuravnoteženosti (pristranskosti) študijske zasnove, ki jih preučujemo, in hkrati pri vseh simulacijskih ponovitvah.

[38] Predpostavka o strogi neodvisnosti ima dva kriterija: (i) v model nagnjenja za oceno stopenj nagnjenja morajo biti vključene vse sospremenljivke, povezane z obravnavo in izidom; in (ii) stopnja nagnjenja, e(X), mora biti na intervalu med 0 in 1. V primeru ocenjevanja vzročnih učinkov je potrebno »zadovoljiti« oba kriterija, medtem ko v primeru ocenjevanja pogojnih asociacij prvi kriterij odpade.

Na drugi strani, s 30 opazovanimi sospremenljivkami in s celotnim vzorcem $n = 120$ in razmerjem med skupinama $R = 3$ (t.j., $n_t = 30$ in $n_c = 90$), logistična regresija ni bila »uspešna« v samo desetih odstotkih simulacijskih ponovitev, vendar to le za najvišjo stopnjo začetne neuravnoteženosti študijske zasnove $B^2 = 1.5$. Te ugotovitve so podprte tudi z analizo variance, kjer rezultati kažejo, da je velikost obravnavane skupine najbolj vpliven faktor (Tabela 5.21).

Rezultati simulacijskih študij kažejo, da so majhni vzorci obravnavanih skupin (tako majhni kot $n_t = 8$) enako uspešni pri uravnoteženju študijske zasnove, kot srednje veliki vzorci (t.j., n_t of 200 or 500), če je le razmerje med skupinama pri majhnih vzorcih zadosti veliko in je predpostavki o strogi pogojnosti neodvisnosti zadoščeno. Vendar pa so ocene vzročnih učinkov ali pogojnih asociacij z majhnimi vzorci obravnavane skupine pričakovano veliko manj natančne v primerjavi s srednje velikimi vzorci. Pomanjkanje natančnosti v ocenah vzorčnih učinkov ali pogojnih asociacij pri majhnih obravnavanih vzorcih je v veliko večjih standardnih napakah, v primerjavi s standardnimi napakami ocen vzročnih učinkov srednje velikih vzorcev obravnavane skupine. Standardne napake ocen vzročnih učinkov z majhnimi vzorci so tako lahko več kot desetkrat večje v primerjavi s standardnimi napakami ocen vzročnih učinkov s srednje velikimi vzorci. Čeprav je to pričakovano dognanje, pa rezultati hkrati kažejo na to, da velikost standardnih napak z majhnimi obravnavanimi vzorci ni odvisna samo od velikosti obravnavane skupine, temveč tudi od števila opazovanih sospremenljivk. Več kot imamo opazovanih sospremenljivk, večje so standardne napake ocen vzorčnih učinkov, čeprav se z naraščanjem števila opazovanih sospremenljivk zahtevano razmerje med kontrolno in obravnavano skupino povečuje in je zato celoten vzorec vedno večji.

Poleg tega, rezultati simulacijske študije kažejo, da izbira algoritma za izvedbo metode usklajevanja ne igra bistvene vloge, čeprav je ta vloga večja pri majhnih vzorcih, kot pri srednje velikih. Hkrati ima izbira algoritma usklajevanja le malo opazen učinek na zahtevano razmerje med skupinama (t.p., optimalen algoritem bo za uravnoteženje študijske zasnove pri majhnih vzorcih v povprečju zahteval malenkostno manjše razmerje med skupinama, kot požrešni (greedy) algoritem). Hkrati pa imajo vzročni učinki ocenjeni z majhnimi obravnavanimi vzorci pri uporabi optimalnega algoritma v povprečju manjše standardne napake.

Rezultati dveh dodatnih simulacijski študiji kažejo, na to da: (i) korelacijska struktura, med opazovanimi sospremenljivkami in spremenljivko izida, ne igra nobene vloge v študijah stopenj nagnjenja, ko je zadoščeno predpostavki pogojne neodvisnosti; in (ii) vrsta spremenljivke izida (t.j., zvezna ali dihotomna) ne vpliva na proces uravnoteženja študijske zasnove v fazi načrta metod nagnjenja (t.p., začetna neuravnoteženost študijske zasnove je lahko uspešno odpravljena ne glede na tip spremenljivke izida). Do razlik prihaja le v fazi analize metod nagnjenja, kjer podatki z dihotomno spremenljivko izida zahtevajo drugačen pristop pri odpravljanju ostankov neuravnoteženosti, kot podatki z zvezno spremenljivko izida.

6. Aplikaciji

V doktorski disertaciji predstavimo tudi dve aplikativni študiji. Prva aplikativna študija uporabi realne opazovalne podatke (Lalonde 1983) za katere so ocene vzročnih učinkov slučajnega poskusa znane. Na podlagi deskriptivne statistike rezultatov naše simulacijske študije, glede minimalno zahtevanih razmerij med skupinama pri majhnih obravnavanih vzorcih, izvedemo simulacijo na Lalondovih podatkih. Cilj takšne simulacijske študije je preveriti, kako aplikativni so rezultati naše teoretične simulacijske študije v praksi.

Rezultati Lalondove simulacijske študije potrjujejo uporabnost rezultatov teoretične simulacijske študije, glede zahtevanih razmerij med skupinama pri določeni velikosti obravnavanega vzorca, ravni začetne neuravnoteženosti študijske zasnove in števila opazovanih sospremenljivk, vendar le v primeru, ko zadovoljimo predpostavko o strogi pogojni neodvisnosti.

Druga aplikativna študija predstavlja opazovalne podatke, kjer so raziskovalna vprašanja morda vzročna, vendar pa zaradi narave podatkov ne moremo nepristransko oceniti vzročnih učinkov, zato ocenjujemo le pogojne asociacije glede na izbor sospremenljivk, ki so v kontekstu študije v našem največjem interesu.

Pogojne asociacije so v preteklosti bile ocenjevane predvsem s pristopom na osnovi modela, kjer z uporabo različnih regresijskih metod hkrati odstranimo neuravnoteženost v sospremenljivkah med skupinama in ocenimo pogojne asociacije. Ta aplikacija razkriva, kako je lahko v primeru majhnih vzorcev in velike ravni neuravnoteženosti sospremenljivk, ter majhnih razmerij med skupinama, ocenjevanje pogojnih asociacij s pristopom na osnovi modela varljivo, saj takšno ocenjevanje temelji v veliki meri na ekstrapolacijah.

Uporaba metod nagnjenja nam v takšnih primerih signalizira že v fazi načrta, da kombinacija majhnih vzorcev in majhnih razmerij med skupinama, ter večje neuravnoteženosti sospremenljivk med skupinama, ne omogoča uspešno balansirati študijske zasnove. Ocenjevanje pogojnih asociacij v takšnem primeru ne bo podalo zaupanja vrednih rezultatov.

7. Zaključek

Doktorska disertacija najprej definira metode nagnjenja v smislu pravilne implementacije metod v opazovalnih študijah pri ocenjevanju vzročnih učinkov. Hkrati razširimo uporabnost metod tudi na ocenjevanje pogojnih asociacij in z aplikacijo pokažemo, kako je pristop na osnovi zasnove (temelj metod nagnjenja) veliko bolj zanesljiv način ocenjevanja pogojnih asociacij v primeru majhnih vzorcev, kot uporaba pristopa na osnovi modela (regresijske metode).

Raziskovanje uspešne implementacije metod nagnjenja z majhnimi vzorci, pri ocenjevanju vzročnih učinkov ali pogojnih asociacij, kaže na to, da se metode nagnjenja, v primeru majhnih vzorcev, obnašajo drugače, kot v primeru velikih vzorcev. V skladu s tem so zahteve pri uporabi metod nagnjenja z majhnimi vzorci drugačne, kot v primeru velikih vzorcev.

Metoda usklajevanja se je izkazala kot najprimernejša metoda uravnavanja v primeru majhnih vzorcev, vendar pa za uravnoteženje sospremenljivk med skupinama majhni obravnavani vzorci zahtevajo veliko večje razmerje med skupinama (kontrolna skupina mora biti veliko večja od obravnavane), kot srednje veliki obravnavani vzorci. Ob pogoju, da je razmerje med skupinama v primeru majhnih obravnavanih vzorcev dovolj veliko, so le-ti sposobni enako učinkovito uravnotežiti skupini, glede na opazovane sospremenljivke.

V prihodnjih raziskavah bi bilo smiselno razširiti raziskovanje majhnih vzorcev tudi s vključitvijo mešanih tipov sospremenljivk[39] (zvezne in diskretne), raziskati drugo pomembno cenilko usklajevanja (usklajevanje z Mahalanobis razdaljo na originalnih opazovanih sospremenljivkah – brez uporabe stopenj nagnjenja), ter uporabiti različne vrste izidov (npr.: ne-paralelni izid, nelinearni izid[40]). Hkrati bi bilo v prihodnje smiselno raziskati tudi področje izbire sospremenljivk za ocenjevanje stopenj nagnjenja z majhnimi vzorci, in kako pravilno oceniti standardne napake vzročnih učinkov v primeru majhnih vzorcev.

[39] Vse sospremenljivke v simulacijski študiji so zvezne.

[40] »Ne-paralelni izid« - funkciji izida obravnavane in kontrolne skupine sta različni. »Nelinearni izid« - funkcija izida ima nelinearno obliko.

8. Slovarček

O metodah nagnjenja se v slovenski statistični literaturi še ni pisalo, zato je bilo potrebno posloveniti kar nekaj statistične terminologije, ki se uporablja na tem statističnem področju. Slovenjenje izraza »propensity score« je precej problematično, zato smo se odločili za izmenično uporabo terminov stopnja nagnjenja oziroma nagnjenje, ko govorimo o »propensity score«.

Tabela 8.1: Slovarček

Angleški izraz	**Slovenski izraz**
adjustment methods	metode uravnavanja
assignment mechanism	mehanizem dodeljevanja
baseline covariates	osnovne sospremenljivke
broken experimental design	kršena zasnova poskusa
common support	skupne podpore
covariate	sospremenljivka
design based approach	pristop na osnovi načrta zasnove
matching	usklajevanje
model-based approach	pristop na osnovi modela zasnove
outcome variable	spremenljivka izida
potential outcomes	možni izidi
potential outcomes framework	okvir možnih izidov
propensity score	stopnja nagnjenja, nagnjenje
propensity score methods	metode nagnjenja
Propensity score study	Študija nagnjenja
Rubin Causal Model - RCM	Rubinov model vzročnosti - RMV
sensitivity analysis	analiza občutljivosti
strong ignorability assumption	predpostavka o strogi pogojni neodvisnosti
treatment	obravnava
treatment condition	obravnavano stanje
treatment effect	učinek obravnave
unbalanced design	neuravnotežena zasnova