UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Flexible estimation of causal effects with observational data

Auteur: **Tom Van Zeebroeck**
Promoteur: **Eugen Pircalabelu**
Lecteur: **Christian Hafner**
Année académique 2023–2024
Master [120] en statistique, orientation générale

Acknowledgments

I am grateful to my supervisor, Eugen Pircalabelu, for his guidance, his insightful feedback and more generally the time that he dedicated to reviewing my thesis.

My appreciation also extends to Christian Hafner for taking the time to read this work.

Lastly, I extend my thanks to my family and friends for their unwavering support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

An important question in social or medical sciences is that of evaluating the **causal** effect of a binary treatment or policy on a well defined outcome. Researchers want to answer questions such as

- What is the effect of a job training program on subsequent earnings?
- What is the effect of giving a cash bonus to unemployed on unemployment duration?
- What is the effect of a given treatment against a disease?

In order to answer these questions, researchers typically collect data on a sample from the population and use it to compare treated and untreated units. They then leverage statistical methods to infer the effect of the treatment in the population of interest from that sample.

The gold standard to collect such a sample is to conduct a randomized controlled trial (RCT) in which subjects are randomly assigned to a treatment status (receiving the treatment/policy or not). Randomization is used to ensure that units in treated and untreated samples are comparable. The outcome of treated and untreated units can then reasonably be compared. However, conducting an RCT is not always feasible due to among others, ethical concerns or financial constraints. For example, it would not be ethical to use an RCT to evaluate the effect of smoking on physical condition given the evidence on the risks of smoking. When RCTs can not be conducted, researchers will typically rely on observational data.

In an observational setting, nothing guarantees that treated and untreated units are comparable. For example, if participation to a job training program is voluntary, participants may be more motivated than non participants. At the same time, motivation may be associated with higher wages. Hence, motivated participants would have earned higher wages even without the program. In this context, observing higher wages for the treated individuals from the observational data does not allow us to conclude on the effect of the program.

Therefore, drawing conclusions about the effect of a treatment based on observational data necessitates making additional assumptions about how the data were generated.

These assumptions, in turn, compel researchers to employ more sophisticated statistical analysis methods. Popular methods in this context rely on procedures where a model is assumed for the process whereby the data were generated. In social or medical sciences building an accurate model describing how the data have been generated may be challenging. At the same time, relying on imperfect models may have bad consequences on the conclusions drawn by the analyst.

This has encouraged researchers to develop methods that relax these modeling assumptions. Some of these methods involve Machine Learning (ML) algorithms. However, ML algorithms are designed for prediction tasks which makes it difficult to understand the asymptotic properties of estimators involving these algorithms. Hopefully, some progress has recently been made in understanding the asymptotic properties of such estimators that rely on ML algorithms. This thesis discusses these recent developments in comparison with the more traditional approaches to the statistical estimation of treatment effects.

The thesis is structured as follows. In chapter 2, we provide a formal definition of causal effects. This will be done based on the potential outcomes framework. We will use this framework to formalize the identification problem that arises when relying on observational data to infer causal effects. We will then see how additional assumptions can be introduced to circumvent this identification problem. Subsequently, we explore how treatment effects can be **estimated** from a sample when the assumptions discussed in chapter 2 hold. We focus on the large sample properties of these estimators. We start by reviewing standard parametric approaches to the estimation of treatment effects in chapter 3. Then, in chapter 4, we discuss estimators that rely on traditional non-parametric approaches and in chapter 5, we discuss how flexible Machine learning methods can be accommodated in the estimation of treatment effects. In chapter 6, we build a Monte Carlo simulation experiment to evaluate the final sample properties of the presented estimators. Finally, in chapter 7, we illustrate the methods on a real application.

Part I

Introduction to causal inference

Chapter 2

Preliminaries and basis of causal inference

Before delving into the estimation of causal effects from a sample, it is essential to clearly define what causal effects are. We will start this chapter by introducing the Potential Outcomes (PO) framework first developed by Splawa-Neyman (1923) in the context of randomized experiments and extended to observational studies by Rubin (1974). This framework will enable us to formalize the definition of causal effects as well as to clarify the identification problem that arises when using observational data to determine causal effects. It will also be useful to discuss assumptions that can be leveraged to circumvent this identification problem.

2.1 The potential outcomes framework

Let W be a random variable describing the treatment assignment. In the PO framework, the causal effect of a binary treatment for a unit of a population of interest is defined as a contrast between two states of the world. A state of the world in which the unit receives the treatment ($W = 1$) and a state of the world in which the unit does not receive the treatment ($W = 0$). Each of these two states of the world is characterized by a potential outcome. Let $Y(w)$ be the random variable describing the potential outcome under treatment value w for $w \in \{0, 1\}$. It is the outcome that would be observed under treatment value w for a unit in the population. In the example of a job training program, $Y(0)$ describes the outcome that would be observed in the absence of training and $Y(1)$ describes the outcome that would be observed under participation in the program.

The causal effect for a unit, τ_{ind}, is defined as a contrast between the two potential outcomes for that unit. For example, the difference between the two individual potential outcomes,

$$\tau_{ind} = Y(1) - Y(0).$$

In practice, however, **only one of the 2 potential outcomes can ever be observed** for a given individual. This has been described as the fundamental problem of causal inference by Holland (1986). If a unit has treatment assignment $W = 1$, only

$Y(1)$ can be observed for that unit and if the unit has treatment assignment $W = 0$, only $Y(0)$ can be observed and measured. The unobservable potential outcome is called the counterfactual. This can be formalized using a switching equation

$$Y = WY(1) + (1 - W)Y(0) = \begin{cases} Y(1) & \text{if } W = 1 \\ Y(0) & \text{if } W = 0 \end{cases} \tag{2.1}$$

where Y is a random variable describing the outcome that can be observed.[1]

Example 1

We present a simple example of the evaluation of a job training program (the treatment) on earnings (the outcome) to elucidate the ongoing discussion and illustrate the concepts that will be presented in the next paragraphs.
Let the treatment assignment W be characterized by $W \sim Bern(\pi)$ where we fix $\pi = 0.5$. Let $Y(0)$ be the random variable that measures the potential outcome in the absence of treatment and $Y(1)$ be the random variable that measures the potential outcome under treatment.
We define

$$\begin{cases} Y(0)|W = 0 & \text{earnings under no treatment given that a unit is untreated} \\ Y(0)|W = 1 & \text{potential earnings under no treatment given that a unit is treated} \\ Y(1)|W = 0 & \text{potential earnings under treatment given that a unit is untreated} \\ Y(1)|W = 1 & \text{earnings under treatment given that a unit is treated} \end{cases}$$

Note that only $Y(0)|W = 0$ and $Y(1)|W = 1$ can ever be observed. At this stage, nothing guarantees that the marginal distributions of $Y(0)$ and $Y(1)$ are the same for treated and untreated units. Indeed, units participating in a training program may have different potential outcomes in the absence of treatment than people who do not participate. In particular, let

$$\begin{cases} Y(0)|W = 0 \sim \mathcal{N}(\mu_0^0, \sigma^2) \\ Y(0)|W = 1 \sim \mathcal{N}(\mu_0^1, \sigma^2) \\ Y(1)|W = 0 \sim \mathcal{N}(\mu_1^0, \sigma^2) \\ Y(1)|W = 1 \sim \mathcal{N}(\mu_1^1, \sigma^2) \end{cases}$$

where we fix $\mu_0^0 = 1$, $\mu_0^1 = 2$, $\mu_1^0 = 1.5$ and $\mu_1^1 = 3$. For simplicity, we assume that the variance parameter is the same for all 4 distributions.

The impossibility to observe the counterfactual makes it impossible to contrast individual potential outcomes. To overcome this limitation and enable the comparison of potential outcomes, researchers must turn to analyses involving multiple units. To facilitate this comparison across various units, the PO framework introduces a simplifying assumption, the Stable Unit Treatment Value Assumption (SUTVA)

Assumption 2.1. *SUTVA (Imbens and Rubin, 2015)*
The potential outcomes for any unit do not vary with the treatments assigned to other

[1] In this thesis, we refer to the random variable Y as the outcome that *can be* observed or the observable potential outcome to avoid any confusion with the outcome that is observed from a sample to which we refer as a realization.

units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

SUTVA has 2 components.

First, there should be **no interference** between units. The fact that a unit receives the treatment should not influence the outcome for other units. The problem with interference is best illustrated in the following example from Imbens and Rubin (2015). Two individuals suffering from headache are in the same room. Individual 2's headache is caused by individual 1's constant complaining about his headache. If individual 1 takes aspirin and this reduces his headache, individual 2's condition will improve. In this case, 4 potential outcomes exist for individual 2: the potential outcome under treatment if individual 1 is treated, the potential outcome under treatment if individual 1 is not treated, the potential outcome under no treatment if individual 1 is treated and the potential outcome under no treatment if individual 1 is not treated. This leads to 6 possible contrasts between potential outcomes for individual 2 and hence, 6 definitions of causal effects.
The no interference assumption restricts the number of potential outcomes for an individual to 2 (the potential outcome under treatment and the potential outcome in the absence of treatment) and the number of contrasts to 1. Interference would increase the number of potential outcomes per unit making it harder to define causal effects.

Second, the **no hidden variations of treatments** component requires that each level of the potential outcomes corresponds to a unique version of the treatment. In the aspirin example presented above, this assumption would be violated if the aspirin tablets available for the 2 individuals were different. One of the tablets could for example be outdated and hence less effective than the other one.

2.2 Causal effects

The potential outcomes framework can now be used to formally define causal estimands in the context of comparing multiple units. These estimands are defined at the population level.

A popular causal estimand is the **Average Treatment Effect** (ATE) which is defined as the expectation of the individual treatment effects in the population

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

In the illustration presented above, using the law of total expectation, the expected potential outcomes can be computed as

$$\begin{cases} \mathbb{E}[Y(0)] = \mathbb{E}[Y(0)|W=0] \times Pr(W=0) + \mathbb{E}[Y(0)|W=1] \times Pr(W=1) = 1.5 \\ \mathbb{E}[Y(1)] = \mathbb{E}[Y(1)|W=0] \times Pr(W=0) + \mathbb{E}[Y(1)|W=1] \times Pr(W=1) = 2.25 \end{cases}$$

yielding an ATE of 0.75.

Studying the average treatment effect over the entire population may not always be relevant. For example, participation to some job training programs happen on a voluntary basis. If participation is intended to remain voluntary, it makes little sense

to study the effect of the program on non participating individuals. This motivates the definition of another popular causal estimand, the **average treatment effect for the treated** (ATT) which is defined as

$$\tau_t = \mathbb{E}[Y(1) - Y(0)|W = 1].$$

It describes the average of the individual causal effects for individuals who are treated. In example 1 above, the ATT is $3 - 2 = 1$.

Similarly, one can define the **average treatment effect for the untreated** (ATU) as

$$\tau_u = \mathbb{E}[Y(1) - Y(0)|W = 0]$$

which in our example is $1.5 - 1 = 0.5$.

Alternatively, researchers can also be interested in the average effect of the treatment for sub-populations sharing some common characteristics. Let $\mathbf{X}$ be the $p \times 1$ random vector of pre-treatment observed covariates $\mathbf{X}^t = \begin{pmatrix} X_1, & \ldots, & X_p \end{pmatrix}$. The conditional average treatment effect (CATE) is defined as

$$\tau(\mathbf{X}) = \mathbb{E}[Y(1) - Y(0)|\mathbf{X} = \mathbf{x}].$$

CATE is a very fine definition of treatment effects in the sense that it considers a different treatment effect for all possible values in the support of $\mathbf{X}$. In practice, researchers are often interested in aggregation levels between an average effect and this extremely fine definition of treatment effect. This motivates the definition of group level average treatment effects (GATE) (Knaus et al., 2021a)

$$GATE = \mathbb{E}[Y(1) - Y(0)|\mathbf{G} = \mathbf{g}]$$

where $\mathbf{G}$ is a subset of $\mathbf{X}$.

In the remainder of this thesis we will **focus on the ATE**. In the next section, we discuss identification of these causal effects.

2.3 Identification

The ATE can not be identified without further assumptions

An estimand is said to be **identifiable** if it can be uniquely determined from the distribution of the observed data (Hernán and Robins, 2020). In practice, the distribution of the 2 potential outcomes $Y(0)$ and $Y(1)$ cannot be observed and only the distribution of the observable outcome Y can be observed. Hence, without additional assumptions, the ATE is not identifiable.

This can be illustrated by thinking about how we would determine the ATE based on data that can be observed. Relying on data that can be observed, a natural way to determine the ATE would be to compute the difference between observed potential outcomes, i.e, the difference between the outcome for treated and untreated units

$$\mathbb{E}[Y(1)|W = 1] - \mathbb{E}[Y(0)|W = 0] = \mathbb{E}[Y|W = 1] - \mathbb{E}[Y|W = 0]$$

which holds by equation (2.1).

Using the law of total expectation, this contrast can be rewritten as

$$\mathbb{E}[Y(1)|W=1] - \mathbb{E}[Y(0)|W=0] = \underbrace{\mathbb{E}[Y(1)-Y(0)]}_{\text{ATE}} + B$$

where

$$B = \underbrace{\mathbb{E}[Y(0)|W=1] - \mathbb{E}[Y(0)|W=0]}_{\text{Selection bias}} + \underbrace{Pr(W=0)(\tau_t - \tau_u)}_{\text{Heterogeneity}}$$

Details on how this decomposition is obtained can be found in Appendix A.1.

The expected difference in observed outcomes is equal to the ATE plus an additional term which is not equal to 0 without additional assumptions. This additional term is composed of the **selection bias** and a term reflecting **heterogeneity** in the effect of the treatment.

Selection bias refers to the fact that treated and untreated units have different potential outcomes in the absence of treatment. For example, this would be the case in a job training program if participants have higher wages than non participants in the absence of the program. If participants have higher wages in the absence of the program, the expected difference in observable outcomes will overrate the ATE.

The heterogeneity bias arises due to differences in how the treated and untreated units benefit from the treatment. For job training programs for example, a usual concern relates to *cream skimming*. The fact that program examinators select participants based on how much they think participants will benefit from the program. If examinators favor participants for whom they anticipate a positive effect of the treatment, the effect of the treatment might well be greater for participants than for non participants. The difference in expected observable outcomes would then overstate the ATE.

In the example presented above, the selection bias is $2-1=1$ and the heterogeneity bias is $0.5 \times (1-0.5) = 0.25$. Hence, the difference in expectations for the observable outcome is $3-1=2$ which is an overstatement of the ATE by 1.25.

This discussion highlights the unidentifiability of the ATE. Indeed, as such, the difference in expected observable outcomes is compatible with different values for the ATE depending on the selection and heterogeneity bias. For example, it is compatible with a value for the ATE greater than the difference in expected observable outcomes if there is a negative selection bias and no heterogeneity bias. At the same time it is compatible with a value for the ATE smaller than the difference in expected observable outcomes if the selection bias is negative and there is no heterogeneity bias.

Introducing identifying assumptions

To identify the ATE, we need to introduce identifying assumptions, i.e, assumptions that allow identification of the ATE. These assumptions will constrain the way in which units are assigned to the treatment. Imbens and Rubin (2015) consider 3 classes of assignment mechanisms. Classical randomized experiments, regular assignment mechanisms and non regular assignment mechanisms.

Classical randomized experiments are characterized by the randomization of individuals to the treatment. Randomization ensures that

$$W \perp\!\!\!\perp Y(0), Y(1).$$

Under this assumption, the ATE can directly be identified from the difference in expected observed outcomes in treated and untreated units. Indeed,

$$\begin{aligned}\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] &= \mathbb{E}[Y(1)|W=1] - \mathbb{E}[Y(0)|W=0] \\ &= \mathbb{E}[Y|W=1] - \mathbb{E}[Y|W=0]\end{aligned}$$

where the first equality holds due to independence and the second due to the switching equation. The final expression depends only on data that can be observed. This explains why randomized controlled trials (RCTs) are considered as the gold standard in the program evaluation literature. However, RCTs are not always feasible. When researchers need to rely on observational data, other identifying assumptions have to be made.

The **regular assignment mechanism** is characterized by two identifying assumptions. Under these two assumptions, identifiability of the ATE can be established. The first assumption, states that treatment assignment W is independent of potential outcomes conditional on a $p \times 1$ random vector of pre-treatment covariates $\mathbf{X}$ [2]. This assumption is often termed the unconfoundedness or conditional independence assumption,

Assumption 2.2. $W \perp\!\!\!\perp Y(1), Y(0)|\mathbf{X}$.

The second assumption states that the probability (at the population level) to be assigned to the treatment, also called the propensity score, is positive for each level of the covariates. This assumption is termed overlap.

Assumption 2.3. $0 < Pr(W = 1|\mathbf{X}) < 1$.

While assumption 2.3 may seem restrictive, in practice, it can be relaxed by excluding units that have no chance of being treated from the population of interest.

The most challenging assumption is clearly assumption 2.2. It requires that all covariates that are associated to both the treatment and potential outcomes are measured and conditioned upon. For example, in a job training program, one could argue that conditional on motivation, the treatment and the potential outcomes are independent. But in practice, it is challenging to have perfect knowledge of all variables related to treatment and potential outcomes on which one should condition.
In economics this assumption is also controversial within the framework of economic theory. Economic theory, which assumes that agents are rational and optimize their behaviour to make choices that benefit them the most. A fact that is sometimes viewed as incompatible with the idea that it would be possible to explain individual's treatment decisions based on a set of observed covariates (Imbens and Rubin, 2015).

[2] As in most of the literature (see for example Knaus et al., 2021a), we refer to $\mathbf{X}$ as the union of the pre-treatment covariates that are used to define $\tau(\mathbf{X})$ and the pre-treatment covariates that are used to ensure unconfoundedness. In practice, however, they do not have to overlap.

The third class of assignment mechanisms, **non regular assignment mechanisms**, contains all situations in which randomization or unconfoundedness are not realistic. In these cases, other identifying assumptions have to be introduced. One example, of such an identifying assumption is assuming the availability of an instrument (a variable correlated to treatment assignment but unrelated to the outcome). Note that such an assumption allows identification of a refined version of the ATE, namely the local average treatment effects (LATE).

In this thesis, we will focus mainly **on the ATE under the regular assignment mechanism**. In the next section, we elaborate on how assumptions (2.2) and (2.3) allow identification of the ATE.

Establishing identification under the regular assignment mechanism

Identification of the ATE under assumptions 2.2 and 2.3 can be demonstrated in 3 ways. The resulting representations of the ATE have important implications for the construction of estimators for the ATE.

Firstly, identification of the ATE can be determined by considering the **conditional (on covariates) expectations of the outcome** for treated and untreated individuals. Note that

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0)|\mathbf{X}] &= \mathbb{E}[Y(1)|\mathbf{X}] - \mathbb{E}[Y(0)|\mathbf{X}] \\ &= \mathbb{E}[Y(1)|W = 1, \mathbf{X}] - \mathbb{E}[Y(0)|W = 0, \mathbf{X}] \\ &= \mathbb{E}[Y|W = 1, \mathbf{X}] - \mathbb{E}[Y|W = 0, \mathbf{X}]\end{aligned}$$

where the last expression can be uniquely determined from observed data. The first equality follows from linearity of expectations, the second equality follows from unconfoundedness and the third equality holds by overlap. To understand the importance of overlap, note that the conditional probability density function of the outcome can be written as $f(y|w = 1, \mathbf{X}) = \frac{f(y,x,w=1)}{f(x,w=1)}$ which would not exist for a 0 propensity score[3]. The ATE can then be obtained from the above expression as

$$\tau = \mathbb{E}\Big[\mathbb{E}[Y(1) - Y(0)|\mathbf{X}]\Big] = \mathbb{E}\Big[\mathbb{E}[Y|W = 1, \mathbf{X}] - \mathbb{E}[Y|W = 0, \mathbf{X}]\Big]$$

where the outer expectation is taken over the covariates. This suggests the possibility of introducing estimation methods that rely on estimation of the conditional expectations functions $\mathbb{E}[Y|W = 1, \mathbf{X}]$ and $\mathbb{E}[Y|W = 0, \mathbf{X}]$. An important complement to this representation of the ATE is the following result by Rosenbaum and Rubin (1983)

$$W \perp\!\!\!\perp Y(0), Y(1)\big|\mathbf{X} \implies W \perp\!\!\!\perp Y(0), Y(1)\big|\pi(\mathbf{X}) \tag{2.2}$$

where $\pi(\mathbf{X}) \equiv \mathbb{E}[W|\mathbf{X}]$ is the propensity score. A proof of this result can be found in Appendix A.2. This result states that under unconfoundedness, it is sufficient to condition on the propensity score to achieve independence between the treatment assignment and potential outcomes. By following the same argument that we used when conditioning on $\mathbf{X}$, the ATE can be identified by conditioning on the propensity score $\pi(\mathbf{X})$ only.

[3]More details on the necessity of overlap for identifying the ATE and how it can be relaxed when considering the ATT are provided in Appendix A.3

Secondly, identification of the ATE under the regular assignment mechanism can be shown by **considering the Inverse Probability Weighting** (IPW) representation of the ATE. The ATE can be rewritten as a weighted average of the observed outcome. Consider the following representation of the potential outcome under treatment $\mathbb{E}[Y(1)]$ for which we have

$$\begin{aligned}
\mathbb{E}[Y(1)] &= \mathbb{E}\big[\mathbb{E}[Y(1)|\mathbf{X}]\big] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[W|\mathbf{X}]\mathbb{E}[Y(1)|\mathbf{X}]}{\mathbb{E}[W|\mathbf{X}]}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[WY(1)|\mathbf{X}]}{\mathbb{E}[W|\mathbf{X}]}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{WY(1)}{\pi(\mathbf{X})}\middle|\mathbf{X}\right]\right] \\
&= \mathbb{E}\left[\frac{WY(1)}{\pi(\mathbf{X})}\right] \\
&= \mathbb{E}\left[\frac{WY}{\pi(\mathbf{X})}\right]
\end{aligned} \tag{2.3}$$

where the first equality follows from the law of iterated expectations (LIE). The second equality is obtained by multiplying and dividing by the propensity score. The third equality holds by unconfoundedness. The fourth equality holds because the propensity score is a function of $\mathbf{X}$. The fifth equality is obtained by LIE and the last equality follows from equation (2.1) and the binary structure of W.
The same argument applies to $\mathbb{E}[Y(0)]$ which can be expressed as $\mathbb{E}\left[\frac{(1-W)Y}{1-\pi(\mathbf{X})}\right]$. This justifies the following representation of the ATE

$$\tau = \mathbb{E}\big[\frac{WY}{\pi(\mathbf{X})}\big] - \mathbb{E}\big[\frac{(1-W)Y}{1-\pi(\mathbf{X})}\big].$$

which can be uniquely determined from data that can be observed.

Thirdly, identification of the ATE under the regular assignment mechanism can be justified by considering its **Augmented Inverse Probability Weighted** (AIPW) representation. Note that the potential outcome under treatment can be represented as

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[\mathbb{E}[Y|W=1,\mathbf{X}] + \frac{W(Y-\mathbb{E}[Y|W=1,\mathbf{X}])}{\pi(\mathbf{X})}\right]. \tag{2.4}$$

Under unconfoundedness, the first term recovers the expectation of the potential outcome under treatment by the LIE while the second term can be shown to be null under

unconfoundedness. Indeed,

$$\begin{aligned}
\mathbb{E}\left[\frac{W(Y-\mathbb{E}[Y|W=1,\mathbf{X}])}{\pi(\mathbf{X})}\right] &= \mathbb{E}\left[\frac{WY}{\pi(\mathbf{X})}\right] - E\left[\frac{W\mathbb{E}[Y|W=1,\mathbf{X}]}{\pi(\mathbf{X})}\right] \\
&= \mathbb{E}[Y(1)] - \mathbb{E}\left[\mathbb{E}\left[\frac{W\mathbb{E}[Y|W=1,\mathbf{X}]}{\pi(\mathbf{X})}\middle|\mathbf{X}\right]\right] \\
&= \mathbb{E}[Y(1)] - \mathbb{E}\left[\frac{\mathbb{E}[W|\mathbf{X}]\mathbb{E}[Y|W=1,\mathbf{X}]}{\pi(\mathbf{X})}\right] \\
&= \mathbb{E}[Y(1)] - \mathbb{E}[Y(1)]
\end{aligned}$$

where the first equality follows from the IPW representation of the ATE. The second equality holds by LIE. The third equality holds because $\pi(\mathbf{X})$ and $E[Y|W=1,\mathbf{X}]$ are functions of $\mathbf{X}$ and the last equality holds because the propensity score is the conditional expectation of the treatment. An analog representation can be derived for the potential outcome in the absence of treatment. This motivates the following representation of the ATE

$$\begin{aligned}
\tau = \mathbb{E}\Big[&\mathbb{E}[Y|W=1,\mathbf{X}] - \mathbb{E}[Y|W=0,\mathbf{X}] \\
&+ \frac{W(Y-\mathbb{E}[Y|W=1,\mathbf{X}])}{\pi(\mathbf{X})} - \frac{(1-W)(Y-\mathbb{E}[Y|W=0,\mathbf{X}])}{1-\pi(\mathbf{X})}\Big]
\end{aligned} \tag{2.5}$$

which can be uniquely determined from observed data. This representation seems unhelpful at first because the last term is equal to 0. However, we will see in chapters 3 and 5 that estimators motivated by this representation of the ATE share very interesting properties.

These different representations of the ATE motivate different estimation strategies. In particular, they highlight the central role that estimation of the outcome regression and propensity score functions will play in the construction of estimators for the ATE.

Identification of the other causal estimands described in section 2.2 follows directly using the same arguments that we developed in this section. For the remainder of the thesis, we introduce the following notations. Let $\mathbb{E}[Y(w)] \equiv \mu_w$, $\mathbb{E}[Y(w)|\mathbf{X}=\mathbf{x}] \equiv \mu_w(\mathbf{x})$ and $\mathbb{E}[Y|\mathbf{X}=\mathbf{x},W=w] \equiv g_w(\mathbf{x})$ with $w \in \{0,1\}$. We also define $\mathbb{E}[\mathbf{X}] = \mu_{\mathbf{X}}$. Note also that under unconfoundedness and overlap, $\mu_w(\mathbf{x})$ and $g_w(\mathbf{x})$ are equivalent.

2.4 Representation of causal effects in terms of a parameter in semi-parametric models

It is sometimes useful to represent causal quantities as parameters in semi-parametric models under the unconfoundedness and overlap assumptions (see for example Chernozhukov et al., 2018 or Hines et al., 2022). An example of such a semi-parametric model that can be used to represent the ATE is the partially linear model.
If we assume that there is no heterogeneity in the effect of the treatment (the effect of the treatment does not change with $\mathbf{X}$) and that unconfoundedness holds, the ATE

can be represented by the parameter θ in the model

$$\mathbb{E}[Y|\mathbf{X}, W = w] = \theta w + g_0(\mathbf{X}). \tag{2.6}$$

To see this, it is useful to decompose the potential outcomes as

$$\begin{cases} Y(0) &= \mu_0 + s_0 \quad \text{where} \quad \mathbb{E}[s_0] = 0 \\ Y(1) &= \mu_1 + s_1 \quad \text{where} \quad \mathbb{E}[s_1] = 0. \end{cases} \tag{2.7}$$

Then, note that

$$\begin{aligned} \mathbb{E}[Y|\mathbf{X}, W] &= \mathbb{E}[WY(1)|\mathbf{X}, W] + \mathbb{E}[(1-W)Y(0)|\mathbf{X}, W] \\ &= \mathbb{E}[WY(1)|\mathbf{X}, W] + \mathbb{E}[Y(0)|\mathbf{X}, W] - \mathbb{E}[WY(0)|\mathbf{X}, W] \\ &= \mathbb{E}[Y(0)|\mathbf{X}, W] + \mathbb{E}[WY(1) - WY(0)|\mathbf{X}, W] \\ &= \mathbb{E}[Y(0)|\mathbf{X}] + \mathbb{E}[W|\mathbf{X}, W]\left(\mathbb{E}[Y(1)|\mathbf{X}, W] - \mathbb{E}[Y(0)|\mathbf{X}, W]\right) \\ &= \mathbb{E}[Y(0)|\mathbf{X}] + W\left(\mathbb{E}[Y(1)|\mathbf{X}] - \mathbb{E}[Y(0)|\mathbf{X}]\right) \\ &= \mathbb{E}[Y(0)|\mathbf{X}] + W\left(\mu_1 + \mathbb{E}[s_1|\mathbf{X}] - \mu_0 - \mathbb{E}[s_0|\mathbf{X}]\right) \end{aligned} \tag{2.8}$$

where the first line holds by equation (2.1). The second and third lines hold by the linearity of expectations. The fourth and fifth equalities hold by unconfoundedness and the last equality holds by decomposition (2.7). If we assume that there is no heterogeneity in the effect of the treatment, then

$$\mathbb{E}[s_1|\mathbf{X}] = \mathbb{E}[s_0|\mathbf{X}]$$

and by unconfoundedness, we recover the partially linear model described in equation (2.6). Note that if we add a functional form assumption for $g_0(\mathbf{X})$, it is straightforward to see that τ can be consistently estimated by OLS. We come back to this point in the next chapter.

Note also that in the presence of heterogeneity it is clear from equation (2.8), that the ATE can not be separated in an additive way. Then θ in the partially linear model does not capture the ATE anymore. Chernozhukov et al. (2018) refer to this case as the interactive regression model.

Part II

Estimation of causal effects

Chapter 3

Parametric estimation

The objective of this chapter is to review traditional parametric estimators $\hat{\tau}$ for the ATE **under the regular assignment mechanism**. The previous chapter highlighted the importance of the outcome regression functions $g_1(\mathbf{X})$, $g_0(\mathbf{X})$ and of the propensity score $\pi(\mathbf{X})$ for identification of the ATE. We saw that the ATE can be determined as the expectation of the difference between $g_1(\mathbf{X})$ and $g_0(\mathbf{X})$ or as an expectation of the outcomes weighted by the propensity score. Estimation strategies will therefore generally take the form of sample averages where the regression functions or the propensity score functions have to be estimated in a first step. For example, one could build estimators $\hat{g}_0(\mathbf{X})$ and $\hat{g}_1(\mathbf{X})$ for $g_0(\mathbf{X})$ and $g_1(\mathbf{X})$ using a sample of size N with $i = 1, \ldots, N$ and estimate the ATE as

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \big(\hat{g}_1(\mathbf{X}_i) - \hat{g}_0(\mathbf{X}_i) \big). \tag{3.1}$$

The outcome regression and propensity score functions will not be of primary interest but have to be estimated in a first step. We will sometimes refer to these functions as nuisance functions.

One possibility for constructing estimators for the nuisance functions is to assume a parametric model for these functions and rely on conventional parametric estimation strategies. However, the properties of the resulting estimators rely crucially on the validity of the assumed parametric model. Assuming a model that is not accurate will have implications for the resulting estimator's properties. We will illustrate this in the next section in the context of Ordinary Least Squares (OLS) estimation of the outcome regression functions.

For the remainder of this chapter as well as for chapters 4 and 5, assumptions (2.1), (2.2) and (2.3) are assumed to hold. In addition, we assume that there is a sample of size N available where observations $\mathbf{O}_i \equiv (Y_i, W_i, \mathbf{X}_i)_{i=1}^N$ are independent draws from a joint distribution $\mathbb{F}$

$$(Y, W, \mathbf{X}) \sim \mathbb{F}$$

where

$$Y = WY(1) + (1 - W)Y(0).$$

While discussing the different estimators, emphasis will be placed on their large sample properties. In particular, we will discuss **consistency** and the **asymptotic distribution** of the presented estimators.

An estimator $\hat{\tau}$ for τ, is said to be consistent if it converges in probability to τ. We shall also discuss the rate of convergence at which the estimator approaches the true value of the estimand. The convergence rate of an estimator can be characterized by the greatest possible constant by which it can be multiplied while remaining bounded in probability. For example, if $\sqrt{N}\hat{\tau} = O_p(1)$, then $\hat{\tau} = O_p(N^{-1/2})$ and $\hat{\tau}$ is said to be $N^{-1/2}$ consistent. This is referred to as the parametric convergence rate (Hines et al., 2022) and is often the best convergence rate that can be achieved for an estimator.

Estimators will also be discussed on the grounds of their asymptotic distribution. Knowledge of the asymptotic distribution of an estimator is valuable for making inference about the estimand of interest. In particular, it provides the researcher with information that can be used to build asymptotically valid confidence intervals and hypothesis tests.

In this section, we discuss estimators for the ATE that rely on models for the estimation of the nuisance functions. In practice, researchers still rely a lot on model based approaches to estimate the nuisance functions. For example Ordinary Least Squares (OLS) remains a popular choice to estimate $g_0(\mathbf{X})$ and $g_1(\mathbf{X})$ (Imbens and Rubin, 2015; Hines et al., 2022). In the next subsection, we discuss the OLS estimator for the ATE and its sensitivity to modeling assumptions. Then, we briefly discuss model-based IPW and AIPW estimators.

3.1 An OLS estimator for the ATE

The estimator and its properties

In order for OLS based estimators to consistently estimate the ATE under the regular assignment mechanism, underlying linear models have to be assumed by the researcher for $g_0(\mathbf{X})$ and $g_1(\mathbf{X})$,

Assumption 3.1. $\begin{cases} g_0(\mathbf{X}) = \alpha_0 + \mathbf{X}^t\boldsymbol{\beta_0} \\ g_1(\mathbf{X}) = \alpha_1 + \mathbf{X}^t\boldsymbol{\beta_1} \end{cases}$

where $\boldsymbol{\beta}_w$ is a $p \times 1$ vector of parameters and α_w is a scalar parameter. The expected potential outcomes can now be defined as $\mu_w = \alpha_w + \mu_{\mathbf{X}}{}^t\boldsymbol{\beta}_w$ with $w \in \{0, 1\}$.

Let $\hat{\alpha}_0$, $\hat{\boldsymbol{\beta}}_0$ and $\hat{\alpha}_1$, $\hat{\boldsymbol{\beta}}_1$ be the corresponding OLS estimators obtained from regressions of Y_i on $\begin{pmatrix} 1 & \mathbf{X}_i^t \end{pmatrix}$ on untreated and treated samples respectively. Then, if the researcher has access to the vector of expected values of the covariates, an estimator for the expected potential outcomes could be

$$\tilde{\mu}_w = \mathbb{E}_{\mathbf{X}}[\hat{\alpha}_w + \mathbf{X}^t\hat{\boldsymbol{\beta}}_w] = \hat{\alpha}_w + \mu_{\mathbf{X}}{}^t\hat{\boldsymbol{\beta}}_w \qquad w \in \{0, 1\}$$

where the subscript $\mathbf{X}$ on the expectation indicates that the expectation is taken over the distribution of $\mathbf{X}$ keeping $\hat{\alpha}_w$ and $\hat{\beta}_w$ fixed. This yields the following estimator for the ATE

$$\tilde{\tau}_{reg} = \tilde{\mu}_1 - \tilde{\mu}_0.$$

In practice, $\mu_{\mathbf{X}}$ is rarely known by the researcher. The potential outcomes can then be estimated by considering the vector of empirical averages $\bar{\mathbf{X}}$ instead of $\mu_{\mathbf{X}}$. Indeed,

$$\hat{\mu}_w = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\alpha}_w + \mathbf{X}_i^t\hat{\boldsymbol{\beta}}_w\right) = \hat{\alpha}_w + \bar{\mathbf{X}}^t\hat{\boldsymbol{\beta}}_w \qquad w \in \{0,1\}.$$

This yields the following estimator for the ATE

$$\hat{\tau}_{reg} = \hat{\mu}_1 - \hat{\mu}_0.$$

Remark 1. Under assumptions (2.2), (2.3) and (3.1), $g_0(\mathbf{X})$ and $g_1(\mathbf{X})$ can be rewritten as

$$g_w(\mathbf{X}) = \mu_w + \dot{\mathbf{X}}^t\boldsymbol{\beta}_w \qquad w \in \{0,1\} \tag{3.2}$$

where $\dot{\mathbf{X}} = (\mathbf{X} - \mu_{\mathbf{X}})$. And by the properties of OLS, $\tilde{\mu}_0$ and $\tilde{\mu}_1$ can be obtained directly as the OLS estimators from regressions of Y_i on $\begin{pmatrix}1 & \dot{\mathbf{X}}_i^t\end{pmatrix}$ on untreated and treated samples respectively. Equivalently, $\hat{\mu}_0$ and $\hat{\mu}_1$ can be obtained as the OLS estimators of the intercept in the regressions of Y_i on $\begin{pmatrix}1 & \ddot{\mathbf{X}}_i^t\end{pmatrix}$ on untreated and treated samples respectively where $\ddot{\mathbf{X}}_i = (\mathbf{X}_i - \bar{\mathbf{X}})$.

More details on remark 1 can be found in Appendix B.1. The fact that the estimator reduces to an intercept in a regression model facilitates the discussion of the asymptotic properties of the resulting estimator for the ATE.

Theorem 1. *(Imbens and Wooldridge, 2009). Under the usual OLS assumptions, the following asymptotic distributions hold*

$$\sqrt{N}(\tilde{\tau}_{reg} - \tau) \xrightarrow{d} \mathcal{N}(0, v_0 + v_1) \tag{3.3}$$

$$\sqrt{N}(\hat{\tau}_{reg} - \tau) \xrightarrow{d} \mathcal{N}(0, v_0 + v_1 + v_T) \tag{3.4}$$

where $v_w = N\mathbb{E}\left[(\tilde{\mu}_w - \mu_w)^2\right]$*, is the usual scaled variance of OLS intercepts with* $w \in \{0,1\}$*,* $v_T = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^t\boldsymbol{\Omega}(\mathbf{X})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$ *and* $\boldsymbol{\Omega}(\mathbf{X})$ *is the variance covariance matrix of the random vector* $\mathbf{X}$*.*

A proof of theorem 1 can be found in Appendix B.2.

Note that using $\bar{\mathbf{X}}$ instead of $\mu_{\mathbf{X}}$ has implications for the asymptotic variance of the obtained estimator. The variance is larger because one needs to account for the uncertainty in the estimation of $\bar{\mathbf{X}}$. In practice, this assumption is often ignored by researchers (Imbens and Wooldridge, 2009).

Sensitivity of the OLS estimator to the modeling assumption

Consistency of $\hat{\tau}_{reg}$ relies crucially on assumption (3.1), i.e, the ability of the researcher to postulate a correct model for $g_1(\mathbf{X})$ and $g_0(\mathbf{X})$. In this subsection, we discuss the impact on consistency of using the OLS estimator when assumption (3.1) fails.

Since $\hat{\tau}_{reg} = \hat{\mu}_1 - \hat{\mu}_0$, we only discuss consistency of $\hat{\mu}_1$, the estimator for $\mathbb{E}[Y(1)]$ but the same argument holds for $\hat{\mu}_0$. Consider the following decomposition of $\hat{\mu}_1$

$$\begin{aligned}
\hat{\mu}_1 &= \frac{1}{N}\sum_{i=1}^{N} \hat{\alpha}_1 + \mathbf{X}_i^t \hat{\boldsymbol{\beta}}_1 \\
&= \frac{1}{N}\sum_{i=1}^{N} W_i(\hat{\alpha}_1 + \mathbf{X}_i^t \hat{\boldsymbol{\beta}}_1) + \frac{1}{N}\sum_{i=1}^{N}(1-W_i)(\hat{\alpha}_1 + \mathbf{X}_i^t \hat{\boldsymbol{\beta}}_1) \\
&= \frac{N_1}{N}\hat{\alpha}_1 + \frac{N_1}{N}\frac{1}{N_1}\sum_{i=1}^{N} W_i \mathbf{X}_i^t \hat{\boldsymbol{\beta}}_1 + \frac{N_0}{N}\hat{\alpha}_1 + \frac{N_0}{N}\frac{1}{N_0}\sum_{i=1}^{N} W_i \mathbf{X}_i^t \hat{\boldsymbol{\beta}}_1 \\
&= \frac{N_1}{N}\underbrace{\left(\hat{\alpha}_1 + \bar{\mathbf{X}}_1^t \hat{\boldsymbol{\beta}}_1\right)}_{\equiv \mathbb{E}[\widehat{Y(1)|W=1}]} + \frac{N_0}{N}\underbrace{\left(\hat{\alpha}_1 + \bar{\mathbf{X}}_0^t \hat{\boldsymbol{\beta}}_1\right)}_{\equiv \mathbb{E}[\widehat{Y(1)|W=0}]}
\end{aligned} \tag{3.5}$$

where N_w is the number of observations in the sample with treatment assignment $W = w$, $\hat{\boldsymbol{\beta}}_1$ is the OLS estimator obtained from the regression described in the previous section and $\bar{\mathbf{X}}_w$ is the empirical mean of the covariates for the sample with treatment value $W = w$ with $w \in \{0, 1\}$. The second equality holds by the switching equation. Note that we introduced $\mathbb{E}[\widehat{Y(1)|W=1}]$ and $\mathbb{E}[\widehat{Y(1)|W=0}]$ as estimators for the potential outcome under treatment for treated and untreated units respectively.

Now, by the law of total expectations,

$$\mathbb{E}[Y(1)] = Pr(W=1) \times \mathbb{E}[Y(1)|W=1] + Pr(W=0) \times \mathbb{E}[Y(1)|W=0] \tag{3.6}$$

and by the law of large numbers, $\frac{N_0}{N}$ and $\frac{N_1}{N}$ are consistent for $Pr(W=0)$ and $Pr(W=1)$. Hence, consistency of $\hat{\mu}_1$ depends only on the consistency of $\mathbb{E}[\widehat{Y(1)|W=0}]$ and $\mathbb{E}[\widehat{Y(1)|W=1}]$.

Let us first study the probability limit of $\mathbb{E}[\widehat{Y(1)|W=1}]$, the estimated potential outcome under treatment for untreated units

$$\begin{aligned}
\operatorname{plim}\left(\mathbb{E}[\widehat{Y(1)|W=1}]\right) &= \operatorname{plim}\left(\hat{\alpha}_1 + \bar{\mathbf{X}}_1^t \hat{\boldsymbol{\beta}}_1\right) \\
&= \operatorname{plim}\left(\bar{Y}_1 - \bar{\mathbf{X}}_1^t \hat{\boldsymbol{\beta}}_1 + \bar{\mathbf{X}}_1^t \hat{\boldsymbol{\beta}}_1\right) \\
&= \operatorname{plim}\left(\bar{Y}_1\right) \\
&= \mathbb{E}[Y(1)|W=1]
\end{aligned} \tag{3.7}$$

where $\bar{Y}_1$ is the empirical mean of Y for treated units. The second equality holds by realizing that $\hat{\alpha}_1 = \bar{Y}_1 - \bar{\mathbf{X}}_1^t \hat{\boldsymbol{\beta}}_1$ by the properties of OLS[1]. Equation (3.7) implies that the first term on the right hand side of (3.5) is consistent for the first term on the right hand side of equation (3.6) even if the model is misspecified. This makes sense since the potential outcome under treatment is observed for treated units.

[1] We refer the reader to Appendix B.1 for more details about this property.

Let us consider the probability limit of $\mathbb{E}[\widehat{Y(1)|W=0}]$, which we can rewrite as

$$\begin{aligned}\operatorname{plim}\Big(\mathbb{E}[\widehat{Y(1)|W=0}]\Big) &= \operatorname{plim}\Big(\hat{\alpha}_1 + \bar{\mathbf{X}}_0^t\hat{\boldsymbol{\beta}}_1\Big)\\ &= \operatorname{plim}\Big(\bar{Y}_1 + (\bar{\mathbf{X}}_0 - \bar{\mathbf{X}}_1)^t\hat{\boldsymbol{\beta}}_1\Big)\\ &= \operatorname{plim}\big(\bar{Y}_1\big) + \operatorname{plim}\big((\bar{\mathbf{X}}_0 - \bar{\mathbf{X}}_1)^t\hat{\boldsymbol{\beta}}_1\big)\\ &= \mathbb{E}[Y(1)|W=1] + \Big[\mathbb{E}[\mathbf{X}|W=0] - \mathbb{E}[\mathbf{X}|W=1]\Big]^t \operatorname{plim}(\hat{\boldsymbol{\beta}}_1).\end{aligned} \tag{3.8}$$

The second equality is obtained by the same argument as that in (3.7) and the last equality holds by the law of large numbers.
Using the law of iterated expectations, one can rewrite the last equality of equation (3.8) as

$$\begin{aligned}\operatorname{plim}\big(\mathbb{E}[\widehat{Y(1)|W=0}]\big) = \mathbb{E}[Y(1)|W=0] &- \Big[\mathbb{E}\big[\mathbb{E}[Y(1)|\mathbf{X},W=1]\big] - \mathbb{E}\big[\mathbb{E}[Y(1)|\mathbf{X},W=0]\big]\Big]\\ &+ \Big[\mathbb{E}[\mathbf{X}|W=0] - \mathbb{E}[\mathbf{X}|W=1]\Big]^t \operatorname{plim}(\boldsymbol{\beta}_1)\end{aligned}$$

where the first line is obtained by adding and subtracting $\mathbb{E}[Y(1)|W=0]$. Then, we have that

$$\begin{aligned}\operatorname{plim}(\mathbb{E}[\widehat{Y(1)|W=0}]) - \mathbb{E}[Y(1)|W=0] = &-\Big[\mathbb{E}\big[\mathbb{E}[Y(1)|\mathbf{X},W=1]\big] - \mathbb{E}\big[\mathbb{E}[Y(1)|\mathbf{X},W=0]\big]\Big]\\ &+ \Big[\mathbb{E}[\mathbf{X}|W=0] - \mathbb{E}[\mathbf{X}|W=1]\Big]^t \operatorname{plim}(\hat{\boldsymbol{\beta}}_1).\end{aligned} \tag{3.9}$$

The estimator $\mathbb{E}[\widehat{Y(1)|W=0}]$ would be consistent for $\mathbb{E}[Y(1)|W=0]$ if $\operatorname{plim}(\mathbb{E}[\widehat{Y(1)|W=0}]) - \mathbb{E}[Y(1)|W=0] = 0$. From equation (3.9) it appears that this will not be the case except in two scenarios.

First, if assumption (3.1) is correct, then, $\operatorname{plim}(\hat{\boldsymbol{\beta}}_1) = \boldsymbol{\beta}_1$ and the two terms on the right hand side of equation (3.9) are equivalent. Hence, their difference is 0 and $\mathbb{E}[\widehat{Y(1)|W=0}]$ is consistent.

Second, if the distribution of the covariates is the same for treated and untreated units, the two terms on the right hand side of equation (3.9) are equal to 0.

This implies that, if either model (3.1) is correctly specified or if the distribution of the covariates is identical between treated and untreated units, $\mathbb{E}[\widehat{Y(1)|W=0}]$ is consistent (implying that $\hat{\mu}_1$ is consistent for $\mathbb{E}[Y(1)]$). However, if the model is misspecified and the treated and untreated units have different distributions for the covariates, the estimator of the potential outcome under treatment for non treated units is not consistent (implying non consistency of $\hat{\mu}_1$ for $\mathbb{E}[Y(1)]$). We illustrate this in example 2 below.

Example 2: Impact of using a misspecified model

Let the treatment assignment W be characterized by $W \sim Bern(\pi)$ where we fix $\pi = 0.5$. Let X be a random variable characterizing a unique confounder and suppose that

$$X \sim \begin{cases} \mathcal{N}(\mu_X^1, \sigma_X^2) & \text{if} \quad W = 1 \\ \mathcal{N}(\mu_X^0, \sigma_X^2) & \text{if} \quad W = 0 \end{cases}$$

where we fix $\sigma_X^2 = 1$, $\mu_X^1 = 1.5$ and $\mu_X^0 = 0$.

Finally, suppose that $Y(1) = \mu_1(X) + \epsilon$ where $\mu_1(X)$ is the conditional expectation of the potential outcome under treatment and $\epsilon \sim \mathcal{N}(0, 0.5)$.

We assume that a researcher is interested in the expected potential outcome under treatment μ_1 but does not have access to the true data generating process. He would like to estimate μ_1 using an i.i.d sample of size $N = 10000$. To do so, he assumes that $\mu_1(X) = \alpha + X\beta$ where α and β are scalar parameters. He estimates μ_1 as $\hat{\mu}_1 = \hat{\alpha} + \hat{\beta}\bar{X}$ where $\hat{\alpha}$ and $\hat{\beta}$ are the OLS estimators of α and β obtained from the sample.

We consider the obtained estimator $\hat{\mu}_1$ under two different specifications for $\mu_1(X)$, details of which are given in the table (3.1).

In case 1, the model specified by the researcher is correct and the estimator is very close to the true value of the expected potential outcome. Case 1 is illustrated in figure (3.1b) where treated units are represented by a dot, untreated units are represented by a cross and the line represents the estimated regression line.

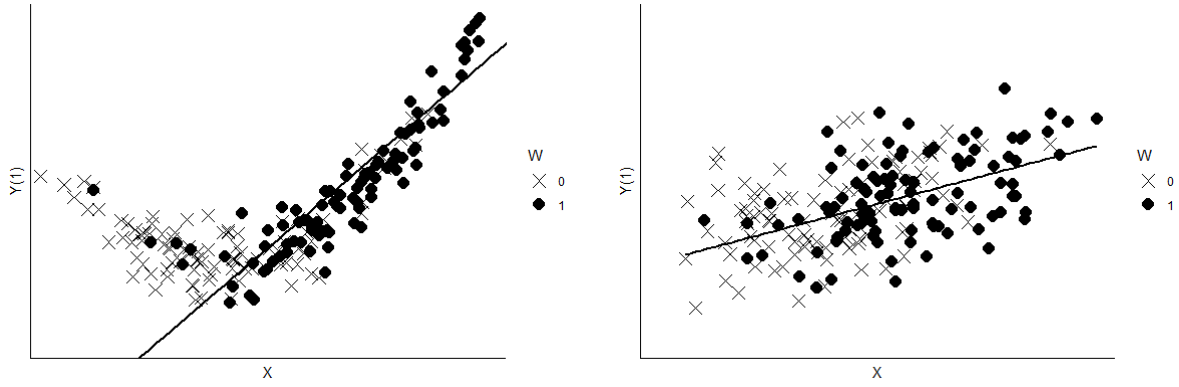
In case 2 however, the model assumed by the researcher is misspecified and the estimator is far from the target. Case 2 is illustrated in figure 3.1a. We immediately see that the potential outcome under treatment for untreated units will be underestimated because the regression coefficients are estimated on treated units and the estimates are not accurate in the region where untreated units lie. In that region, the estimated model is a very bad approximation to the true model. Hence, unless the assumed model is a good approximation to the true model globally, the OLS estimator of the expected potential outcome under treatment can be severly biased.

Table 3.1: Sensitivity of the OLS estimator to misspecification

	$\mu_1(X)$	μ_1	$\hat{\mu}_1$
Case 1	$0.2X$	0.15	0.1527
Case 2	$0.2X + 0.5X^2$	1.21	0.64

The same reasoning holds for the expected potential outcome in the absence of treatment where the estimation of the potential outcome in the absence of treatment for treated units will in general be unsatisfactory unless the postulated model is correct.

Figure 3.1: Sensitivity of the OLS estimator to model misspecification



(a) Misspecified model

(b) Correct model

3.2 Model-based IPW estimation

In the previous chapter, we introduced the Inverse Probability Weighting representation of the ATE described in equation (2.3). This representation motivates the construction of an IPW estimator where $\pi(\mathbf{X})$ in equation (2.3) is replaced by an estimator $\hat{\pi}(\mathbf{X})$

$$\hat{\tau}_{IPW1} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i Y_i}{\hat{\pi}(\mathbf{X}_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)Y_i}{1-\hat{\pi}(\mathbf{X}_i)}.$$

A common alternative is to use a normalized version where the weights add up to 1, namely,

$$\hat{\tau}_{IPW2} = \Big(\sum_{i=1}^{N}\frac{W_i}{\hat{\pi}(\mathbf{X}_i)}\Big)^{-1}\sum_{i=1}^{N}\frac{W_i Y_i}{\hat{\pi}(\mathbf{X}_i)} - \sum_{i=1}^{N}\Big(\frac{1-W_i}{1-\hat{\pi}(\mathbf{X}_i)}\Big)^{-1}\sum_{i=1}^{N}\frac{(1-W_i)Y_i}{1-\hat{\pi}(\mathbf{X}_i)}. \tag{3.10}$$

One possibility to estimate the propensity score, is to assume an underlying parametric model. An example of such a model is the logit model of the form

$$\pi(\mathbf{X}) = \frac{\exp(\mathbf{X}^t\mathbf{\Phi})}{1+\exp(\mathbf{X}^t\mathbf{\Phi})},$$

where $\mathbf{\Phi}$ is a $p \times 1$ vector of parameters. The parameters of the assumed model can then be estimated using maximum likelihood (MLE).

When the assumed model is correctly specified, the estimator can be shown to be consistent and asymptotically normal. We refer to Lunceford and Davidian (2004) for a detailed treatment of the asymptotic properties of such IPW estimators. Just as the OLS estimator described in the previous section, consistency of these estimators relys on the correct specification of the model. For a discussion on the impact of model misspecification on the large sample properties of IPW based estimators, we refer the interested reader to Waernbaum and Pazzagli (2023).

3.3 Model-based AIPW estimation

The AIPW representation of the ATE motivates the use of estimators where both the outcome regression and propensity score functions have to be estimated. An AIPW based estimator can then be obtained as the sample equivalent to the AIPW representation of the ATE presented in the previous chapter, namely

$$\hat{\tau}_{AIPW} = \frac{1}{N}\sum_{i=1}^{N}\Bigg[\hat{g}_1(\mathbf{X}_i) - \hat{g}_0(\mathbf{X}_i) + \frac{W_i(Y_i - \hat{g}_1(\mathbf{X}_i))}{\pi(\mathbf{X}_i)} - \frac{(1-W_i)(Y_i - \hat{g}_0(\mathbf{X}_i))}{1-\pi(\mathbf{X}_i)}\Bigg]. \tag{3.11}$$

This estimator is doubly robust in the sense that it is consistent if either the outcome regression or the propensity score model is correctly specified.

Proposition 1. *(Tsiatis, 2006). Under assumptions (2.1), (2.2) and (2.3), $\hat{\tau}_{AIPW}$ is a consistent estimator for τ if either the model for the propensity score $\pi(\mathbf{X})$ or the model for the outcome regression $\mathbb{E}[Y|W,\mathbf{X}]$ is correctly specified.*

A proof of Proposition 1 can be found in Appendix B.3. This is an interesting property as it gives researchers two chances for specifying a correct model. However, when both models are misspecified, nothing guarantees consistency of the estimator. We come back to the asymptotic properties of AIPW estimators at a more general level in the next chapter. The next example illustrates the ongoing discussion.

Example 3: AIPW as a protection against misspecification

Let $\mathbf{X}$ be the random vector satisfying

$$\mathbf{X} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right).$$

Let the treatment assignment W be characterized by

$$W \sim Bern(\pi(\mathbf{X}))$$

where

$$\pi(\mathbf{X}) = \frac{\exp(0.2\times X_1 + 0.1\times X_1\times X_2 + 0.1\times X_2^2)}{1+\exp(0.2\times X_1 + 0.1\times X_1\times X_2 + 0.1\times X_2^2)}.$$

Finally, suppose that there is no heterogeneity in the effect of the treatment such that the ATE can be represented by τ in the following model (by equation 2.6)

$$Y = \tau\times W + g_0(\mathbf{X}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0,1)$$

where we fix $\tau = 1$ and $g_0(\mathbf{X}) = 0.2\times X_1 + 0.5\times X_1\times X_2 + 0.5\times X_2^2$.

Assume that a researcher is interested in estimating τ from an i.i.d sample of size $N = 10000$. He considers $\hat{\tau}_{reg}$, $\hat{\tau}_{IPW}$ and $\hat{\tau}_{AIPW}$ as possible estimators. For the IPW and AIPW estimators, $\pi(\mathbf{X})$ is estimated using a logistic regression estimator and $g_0(\mathbf{X})$ is estimated using OLS.

The obtained estimates are displayed under 3 different scenarios in table 3.2.

In scenario 1, the researcher uses a correct specification for $g_0(\mathbf{X})$ but includes only main effects (no interactions or higher order effects) in the logistic regression. In scenario 2, the propensity score function is correctly specified but the researcher includes only main terms for estimating $g_0(\mathbf{X})$. In scenario 3, both functions are misspecified by the researcher.

Noting that the target is 1, the OLS and IPW estimators are close to τ in scenarios 1 and 2 respectively. In these two cases, the AIPW estimator works well. However, in the last case, all estimators are biased.

Table 3.2: Comparison of OLS, IPW and AIPW under 3 misspecification scenarios

	$\hat{\tau}_{reg}$	$\hat{\tau}_{IPW}$	$\hat{\tau}_{AIPW}$
Scenario 1	1.002	1.26	1.003
Scenario 2	1.26	1.002	1.002
Scenario 3	1.24	1.25	1.25

Chapter 4

Traditional non-parametric estimation

In the previous chapter, we presented estimators for the ATE that relied on models for estimation of the nuisance functions. We saw that relying on modeling assumptions can have bad consequences on the properties of the resulting estimators for the ATE when the model is misspecified. This is of particular concern in the social or medical sciences where specifying a correct model can be challenging (Hines et al., 2022). Therefore, researchers have developed methods that relax these assumptions. In this section, we discuss how traditional non-parametric techniques such as Nearest Neighbor (NN) or Kernel regression have been leveraged to estimate the ATE.

4.1 Nearest neighbor estimation of the outcome regression functions

In this section, we describe an estimator discussed in Abadie and Imbens (2006) that is based on nearest neighbors estimation of the outcome regression functions.

The estimation procedure

Abadie and Imbens (2006) consider the following small adjustment[1] to equation (3.1) to construct an estimator for the ATE

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\big(\hat{Y}_i(1) - \hat{Y}_i(0)\big) \tag{4.1}$$

with

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0 \\ \hat{g}_1(\mathbf{X}_i) & \text{if } W_i = 1 \end{cases}$$

[1]Note that for some estimators, equations (3.1) and (4.1) are equivalent. For example for OLS, $\frac{1}{N_1}\sum_{i=1}^{N} W_i(\hat{\alpha}_1 + \mathbf{X}_i^t\hat{\boldsymbol{\beta}}_1) = \hat{\alpha}_1 + \bar{\mathbf{X}}_1^t\hat{\boldsymbol{\beta}}_1 = \bar{Y}_1$ where the last equality comes from the fact that the empirical mean of the estimated residuals is 0 by construction in OLS estimation.

and

$$\hat{Y}_i(1) = \begin{cases} \hat{g}_0(\mathbf{X}_i) & \text{if } W_i = 0 \\ Y_i & \text{if } W_i = 1 \end{cases}$$

where $\hat{g}_0(\mathbf{X}_i)$ and $\hat{g}_1(\mathbf{X}_i)$ are obtained from Nearest Neighbor (NN) regressions with fixed number of neighbors on untreated and treated samples respectively. The NN regression estimator for $g_w(\mathbf{X}_i)$ is defined as

$$\hat{g}_w(\mathbf{X}_i) = \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j \qquad \text{for} \quad w \in \{0, 1\} \tag{4.2}$$

where $\mathcal{J}_M(i)$ is the set of M *closest* units to unit i in the sample of observations with treatment value different from W_i (since we only impute the counterfactual).

To define the set $\mathcal{J}_M(i)$, we first need to define a distance metric. As we consider a multidimensional covariate space, the distance can be determined by a vector norm $||x||_V = \sqrt{x\mathbf{V}x^t}$ where $\mathbf{V}$ is a positive definite symmetric matrix. For example, if $\mathbf{V}$ is the inverse of the sample covariance matrix of $\mathbf{X}$ that we define as

$$\frac{1}{N} \sum_{i=1}^{N} (\mathbf{X}_i - \bar{\mathbf{X}}_i)(\mathbf{X}_i - \bar{\mathbf{X}}_i)^t,$$

then, the distance metric is the standard Mahalanobis distance. Now, let $j_m(i)$ be the m^{th} closest unit to unit i, where $j_m(i)$ satisfies $W_{j_m(i)} = 1 - W_i$ and

$$m = \sum_{l:W_l = 1 - W_i} \mathbb{1}\left(||\mathbf{X}_l - \mathbf{X}_i|| \leq ||\mathbf{X}_{j_m(i)} - \mathbf{X}_i||\right)$$

with $\mathbb{1}$ the indicator function. The set of M closest units to unit i in the sample with treatment assignment different from that of unit i can now be formally defined as $\mathcal{J}_M(i) = \{j_1(i), \ldots, j_M(i)\}$.

The estimation procedure is illustrated in table 4.1 for an hypothetical example with $N = 10$, $M = 2$ and a single continuous covariate where $K_M(i)$ refers to the number of times that unit i is used as a neighbor.

Table 4.1: Example of the simple Nearest Neighbors estimation approach

i	Y_i	X_i	W_i	$\mathcal{J}_M(i)$	$K_M(i)$	$\hat{Y}_i(0)$	$\hat{Y}_i(1)$
1	6	5.5	0	{8,7}	3	6	2
2	0	6	0	{7,8}	3	0	2
3	4	7	0	{9,10}	2	4	4.5
4	2	8.5	0	{10,9}	1	2	4.5
5	5	4	0	{6,8}	1	5	4.25
6	3	3	1	{5,1}	1	5.5	3
7	1	6	1	{2,1}	2	3	1
8	3	5.5	1	{1,2}	3	3	3
9	5	7	1	{3,2}	2	2	5
10	4	7.5	1	{3,4}	2	3	4

Consistency and asymptotic distribution of the estimator

To better understand the asymptotic properties of the estimator in (4.1), Abadie and Imbens (2006) decompose it as

$$\hat{\tau}_{ma} - \tau = (\overline{\tau(\mathbf{X})} - \tau) + E_M + B_M \tag{4.3}$$

where

$$\begin{aligned}
\overline{\tau(\mathbf{X})} &= \frac{1}{N}\sum_{i=1}^{N}\big(\mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i)\big) \\
E_M &= \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\big(1 + K_M(i)\big)\big(Y_i - \mu_{W_i}(\mathbf{X}_i)\big) \\
B_M &= \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\frac{1}{M}\sum_{m=1}^{M}\big(\mu_{1-W_i}(\mathbf{X}_i) - \mu_{1-W_i}(\mathbf{X}_{j_m(i)})\big).
\end{aligned}$$

The asymptotic properties of the NN estimator can be discussed by considering each term separately.

The first term, $\overline{\tau(\mathbf{X})}$, is a sample average of the difference between the true conditional expectation of $Y(1)$ and the true conditional expectation of $Y(0)$. Since $\overline{\tau(\mathbf{X})}$ is a sample average of a deterministic function, $\overline{\tau(\mathbf{X})} \xrightarrow{p} E[\tau(\mathbf{X})]$ by the law of large numbers and the continuous mapping theorem. Note that $E[\tau(\mathbf{X})] = \tau$ by the law of iterated expectations. Hence,

$$\overline{\tau(\mathbf{X})} - \tau \xrightarrow{p} 0.$$

By the central limit theorem, we also have that

$$\sqrt{N}(\overline{\tau(\mathbf{X})} - \tau) \xrightarrow{d} \mathcal{N}\big(0, V^{\tau(\mathbf{X})}\big) \tag{4.4}$$

where $V^{\tau(\mathbf{X})} = E[(\overline{\tau(\mathbf{X})} - \tau)^2]$.

The second term, E_M, is a weighted average of the residuals (deviations of the outcome from its conditional mean). Abadie and Imbens (2006) show that under regularity conditions

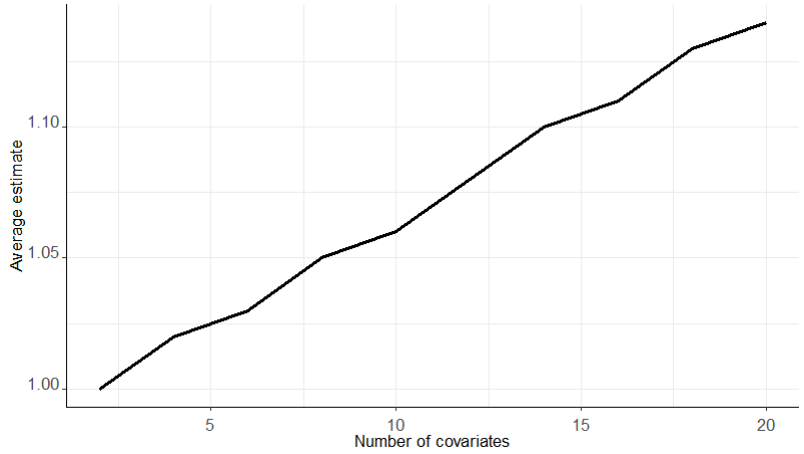
$$\sqrt{N}E_M \xrightarrow{d} \mathcal{N}\big(0, V^E\big) \tag{4.5}$$

where $V^E = \frac{1}{N}\sum_{i=1}^{N}\left(1 + \frac{K_M(i)}{M}\right)^2 \sigma^2(\mathbf{X}_i, W_i)$ with $\sigma^2(\mathbf{X}_i, W_i) = V(Y|\mathbf{X}_i, W_i)$.

The asymptotic behaviour of the last term, B_M, is harder to understand. It depends on the sum of $\mu_0(\mathbf{X}_i) - \mu_0(\mathbf{X}_{j_m(i)})$ for treated units and the sum of $\mu_1(\mathbf{X}_i) - \mu_1(\mathbf{X}_{j_m(i)})$ for untreated units. These terms capture differences in expected outcomes between units and their neighbors. When neighbors are imperfect matches, this term will be different from 0. The asymptotic behaviour of B_M is summarized in the following proposition.

Proposition 2. *(Abadie and Imbens, 2006). Under unconfoundedness, overlap and additional regularity conditions,* $B_M = O_p(N^{-\frac{1}{p}})$.

Figure 4.1: Sensitivity of the NN estimator to the number of covariates



Note: The average estimates for every p are computed based on 50 repetitions of the DGP of example 3 with N = 5000. In example 3, only 2 covariates were considered. Additional covariates are included as noise covariates in the sense that they do not enter the nuisance functions.

The proof of Proposition 2 and more details about the regularity conditions can be found in Abadie and Imbens (2006). Proposition 2 is important because it implies that the convergence rate of B_M deteriorates with the number of covariates p and that it is in general not equal to $\sqrt{N}$. The intuition behind this important result is that when the number of covariates increases, it becomes harder to find neighbors that are close to unit i. Then, units that are further away from unit i are considered and the differences in conditional expectations between unit i and its neighbors increase.

Let us now consider the implications of (4.4), (4.5) and proposition 2 for the asymptotic properties of $\hat{\tau}_{ma}$. First, note that $\hat{\tau}_{ma}$ is consistent for τ.

Proposition 3. *(Abadie and Imbens, 2006). Under unconfoundedness, overlap and regularity conditions,*

$$\hat{\tau}_{ma} \xrightarrow{p} \tau.$$

Proposition 3 follows directly from (4.4), (4.5) and proposition 2. Note that while the Nearest Neighbor estimator is consistent, it will only be $\sqrt{N}$ consistent as long as $p \leq 2$ because for $p > 2$, the B_M term dominates the asymptotic behaviour of the matching estimator.

This is illustrated on example 3 in figure (4.1). The ATE was estimated using the Nearest Neighbors estimator for different number of covariates on 50 repetitions of the data generating process of example 3 where the true ATE was 1. When $p = 2$, the empirical bias of the estimator is close to 0. However, increasing the number of noise variables induces a strong bias in the studied estimator.

A general asymptotic distribution for the NN estimator can not be obtained as there is no central limit theorem available for B_M. However, it useful to consider the distribution of the estimator when B_M can be ignored.

Theorem 2. *(Abadie and Imbens, 2006). Under unconfoundedness, overlap and some*

additional regularity conditions,

$$\sqrt{N}(\hat{\tau}_{ma} - B_M - \tau) \xrightarrow{p} \mathcal{N}(0, V^E + V^{\tau(\mathbf{X})}). \tag{4.6}$$

This follows directly from the results presented above and the fact that $\overline{\tau(\mathbf{X})}$ and E_M are asymptotically independent. Note that when $p = 1$, B_M can be ignored when studying the asymptotic properties of $\hat{\tau}_{ma}$. In that case, the asymptotic distribution described above is that of $\sqrt{N}(\hat{\tau}_{ma} - \tau)$.

Abadie and Imbens (2006) propose the following consistent estimator for $V^E + V^{\tau(\mathbf{X})}$ that can be used for inferential purposes

$$\begin{aligned}\hat{V} = &\frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i(1) - \hat{Y}_i(0) - \hat{\tau}_M)^2 \\ &+ \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{K_M(i)}{M}\right)^2 + \left(\frac{2M-1}{M}\right)^2\left(\frac{K_M(i)}{M}\right)\right]\hat{\sigma}^2(\mathbf{X}_i, W_i)\end{aligned} \tag{4.7}$$

where $\hat{\sigma}^2(\mathbf{X}_i, W_i)$ is an estimator of the conditional variance of Y which is based on a new NN procedure to estimate the conditional expectation of Y. It is computed as

$$\hat{\sigma}^2(\mathbf{X}_i, W_i) = \frac{J}{J+1}\left(Y_i - \frac{1}{J}\sum_{l \in \mathcal{L}_J(i)} Y_l\right)^2 \tag{4.8}$$

where $\mathcal{L}_J(i)$ is the set of J *closest* units to unit i *among the individuals in the sample with same treatment value as unit i* that are used in the new NN procedure.

Bias correction

As such, theorem 2 is of limited practical importance because researchers are often interested in cases with $p > 1$. In order to improve the convergence order of B_M such that it can be ignored asymptotically, Abadie and Imbens (2011) propose a bias corrected version of the NN regression estimator presented above. In particular, they propose to correct the outcomes of the neighbors to take into account the difference between unit i and its neighbors. They perform this by adjusting the outcome of the considered neighbors with the difference in predicted values from an estimated regression. Concretely, let

$$\tilde{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0 \\ \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j + \tilde{g}_0(\mathbf{X}_i) - \tilde{g}_0(\mathbf{X}_{j_m(i)}) & \text{if } W_i = 1 \end{cases}$$

and

$$\tilde{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j + \tilde{g}_1(\mathbf{X}_i) - \tilde{g}_1(\mathbf{X}_{j_m(i)}) & \text{if } W_i = 0 \\ Y_i & \text{if } W_i = 1 \end{cases}$$

where $\tilde{g}_w(\mathbf{X})$ is an estimator for $g_w(\mathbf{X})$. Then, the bias corrected Nearest Neighbor estimator is

$$\tilde{\tau}_{ma} = \frac{1}{N}\sum_{i=1}^{N}(\tilde{Y}_i(1) - \tilde{Y}_i(0)).$$

One can decompose the bias adjusted Nearest Neighbor estimator as

$$\begin{aligned}\tilde{\tau}_{ma} &= \hat{\tau}_{ma} + \frac{1}{N}\sum_{i=1}^{N} W_i \frac{1}{M}\sum_{m=1}^{M}\left\{\tilde{g}_0(\mathbf{X}_i) - \tilde{g}_0(\mathbf{X})_{j_m(i)}\right\} - \frac{1}{N}\sum_{i=1}^{N}(1-W_i)\frac{1}{M}\sum_{m=1}^{M}\left\{\tilde{g}_1(\mathbf{X}_i) - \tilde{g}_1(\mathbf{X})_{j_m(i)}\right\} \\ &= \hat{\tau}_{ma} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}(2W_i-1)\frac{1}{M}\sum_{m=1}^{M}\left\{\tilde{g}_{1-W_i}(\mathbf{X}_i) - \tilde{g}_{1-W_i}(\mathbf{X})_{j_m(i)}\right\}}_{\hat{B}_M}.\end{aligned}$$

This decomposition clarifies how $\hat{B}_M$ serves as an estimator for the bias. If $\hat{B}_M$ is a consistent estimator for B_M it can be ignored asymptotically.

Abadie and Imbens (2011) show that under regularity conditions, when a non-parametric series estimator is used to estimate the bias, $\tilde{\tau}_{ma}$ has the asymptotic distribution derived in Theorem 2. They also show that using an OLS estimator for the bias works well in simulations. However, this will likely depend on the accuracy of the assumptions necessary for OLS to work well for the data at hand.

4.2 Other aspects of traditional non-parametric estimation of the nuisance functions

A drawback of using a Nearest Neighbors approach is that the same weights are given to all neighbors. In practice, neighbors close to unit i are likely to be more representative for that unit. It would be desirable to take this fact into account in the estimation procedure. This can be done by using Kernel regression estimators instead.
In the multivariate case, the kernel regression estimator of the outcome regression function $g_w(x)$ is

$$\hat{\mu}_0(x) = \sum_{i:W_i=0}^{N} Y_i\lambda_i \qquad \text{where } \lambda_i = \frac{k(H^{-1}(X_i - x))}{\sum_{i:W_i=w_i} k(H^{-1}(X_i - x))}$$

where H^{-1} is a matrix of bandwidths and k is a multivariate kernel. The large sample properties of such estimators have been studied by Heckman et al. (1998). As in the case of the NN estimator described above, a general problem with such estimators is that their rate of convergence is very sensitive to the number of covariates (Heckman et al., 1998).

In order to address this curse of dimensionality, the conventional approach has been to make use of the result described in equation (2.2). Instead of matching on the entire covariate space, one could match on a one dimensional function of it, namely the propensity score $\pi(\mathbf{X})$. However, in practice the propensity score is unknown and has to be estimated as well. If the propensity score is estimated using parametric methods, the problem of model misspecification appears again and if the propensity score is estimated using non-parametric methods, the curse of dimensionality appears as well (Heckman et al., 1998).

Finally, it is also possible to build IPW-type non-parametric estimators where the propensity score is estimated non-parametrically. Hirano et al. (2003) study the prop-

erties of such an estimator when the propensity score is estimated using a logistic sieve estimator.

Chapter 5

Debiased Machine Learning

In this chapter, we consider an estimator for the ATE that accomodates generic Machine-Learning (ML) algorithms to estimate the nuisance functions. For example, one could consider an estimator of the form described in equation 3.1 where $\hat{g}_0(x)$ and $\hat{g}_1(x)$ are obtained using ML algorithms. We will refer to such estimators as so called *plug-in* estimators. We refer to ML techniques in a broad sense to describe standard ML methods such as random forests or Neural Networks as well as parametric estimation procedures with variable selection such as the LASSO. Importantly, ML methods are designed for prediction rather than precise point estimation. This induces a regularization bias in the estimation of the nuisance functions and makes it hard to understand the asymptotic properties of the resulting plug-in estimator.

We will present a generic procedure called Debiased Machine Learning (DML) discussed in Chernozhukov et al. (2018) that can be used in this context. At a general level, DML is useful to obtain estimators with desirable asymptotic properties for estimands that rely on the estimation of nuisance functions in a first step when ML methods are used to estimate these nuisance functions. In the next two sections, we take a step back to introduce the theoretical background underlying the method at a general level building on Hines et al. (2022) and Kennedy (2022). In the third section of this chapter, we discuss how the general method can be used in a causal inference context to construct an estimator for the special case of the ATE.

5.1 Preliminaries

In this section, we present some important results from Von Mises calculus that will be useful to make progress in understanding the asymptotic behaviour of plug-in estimators when ML techniques are used for estimation of the nuisance functions.

Framework and Notations

We will assume that there is a sample of observations $\mathbf{Z}_i^N$ that are identically and independently drawn from a distribution $\mathbb{P}$ assumed to lie in a set of distributions $\mathcal{P}$. No restriction is placed on the set $\mathcal{P}$ in the sense that it contains all probability distributions contained in the sample space as in Kennedy (2022) or Hines et al. (2022).

The objective is to use this sample to estimate a one dimensional estimand which can be expressed as a functional of the true distribution $\mathbb{P}$. Let $\Psi : \mathcal{P} \mapsto \mathbb{R}$ be a functional taking as argument a distribution function contained in $\mathcal{P}$. The estimand, is then defined as $\Psi(\mathbb{P})$.

Let us also introduce some notation from Kennedy (2022) that will be used in the remainder of this section. In particular, $\mathbb{P}_n\big[f(\mathbf{Z})\big]$ will be used to refer to empirical averages of functions of $\mathbf{Z}$. Similarly, we let $\mathbb{P}\big[f(\mathbf{Z})\big] = \int f(\mathbf{z})p(\mathbf{z})\,d\mathbf{z}$ where $p(\mathbf{z})$ refers to the density of $\mathbf{Z}$.

Von Mises calculus

In order to discuss the asymptotic behaviour of plug-in estimators we will rely on some important concepts from Von Mises calculus. We start by defining a parametric submodel.

Definition 1 (Kennedy, 2022)**.** A parametric submodel is a smooth parametric model $\mathcal{P}_t = \{\mathbb{P}_t : t \in \mathbb{R}\}$ that satisfies (i) $\mathcal{P}_t \subset \mathcal{P}$, and (ii) $\mathbb{P}_{t=0} = \mathbb{P}$.

Let $\tilde{\mathbb{P}}$ be an arbitrary distribution contained in $\mathcal{P}$. Then, a possible parametric submodel $\mathbb{P}_t$ is

$$\mathbb{P}_t = (1-t)\mathbb{P} + t\tilde{\mathbb{P}}, \qquad t \in [0,1].$$

Note that if $t = 0$, $\mathbb{P}_t = \mathbb{P}$ and if $t = 1$, then $\mathbb{P}_t = \tilde{\mathbb{P}}$. Parametric submodels are useful to characterize the sensitivity of $\Psi(\mathbb{P})$ to changes in the distribution in the direction of $\tilde{\mathbb{P}}$. In particular, it can be used to formalize the notion of a functional derivative of the form

$$\lim_{t\to 0}\left(\frac{\Psi(\mathbb{P}_t) - \Psi(\mathbb{P})}{t}\right) = \left.\frac{\partial \Psi(\mathbb{P}_t)}{\partial t}\right|_{t=0}.$$

If such a derivative exists, it is refered to as a Von Mises or Gateaux derivative and the estimand is said to be pathwise differentiable. The Von Mises derivative can be computed using the **influence function (IF)** of the estimand where an influence function is defined below.

Definition 2 (Fisher and Kennedy, 2021)**.** For a given functional Ψ, the influence function for Ψ is the function ϕ satisfying

$$\left.\frac{\partial \Psi(G + t(Q-G))}{\partial t}\right|_{t=0} = \int \phi(\mathbf{z}, G)\{q(\mathbf{z}) - g(\mathbf{z})\}\,d\mathbf{z}$$

where $\int \phi(\mathbf{z}, G)g(\mathbf{z})\,d\mathbf{z} = 0$ for any two distributions G and Q with densities g and q.

When the Von Mises derivative exists, the estimand admits a distributional equivalent to the Taylor expansion called the Von Mises expansion. The Von Mises expansion about the point $t = 1$ in the parametric submodel is

$$\Psi(\mathbb{P}) = \Psi(\tilde{\mathbb{P}}) - \left.\frac{\partial \Psi(\mathbb{P}_t)}{\partial t}\right|_{t=1} + R(\mathbb{P}, \tilde{\mathbb{P}}) \tag{5.1}$$

where $R(\mathbb{P}, \tilde{\mathbb{P}})$ is a remainder term. Note that,

$$\begin{aligned}\left.\frac{\partial \Psi(\mathbb{P}_t)}{\partial t}\right|_{t=1} &= \left.\frac{\partial \Psi(\mathbb{P} + t(\mathbb{P} - \tilde{\mathbb{P}}))}{\partial t}\right|_{t=1} \\ &= \int \phi(\mathbf{z}, \tilde{\mathbb{P}})\{\tilde{p}(\mathbf{z}) - p(\mathbf{z})\}\, d\mathbf{z} \\ &= -\int \phi(\mathbf{z}, \tilde{\mathbb{P}}) p(\mathbf{z})\, d\mathbf{z} \\ &= -\mathbb{P}[\phi(\mathbf{Z}, \tilde{\mathbb{P}})]\end{aligned} \tag{5.2}$$

where the second equality holds by definition 2 [1] and the third equality holds by the zero mean property of influence functions. Hence, the Von Mises expansion can be rewritten as

$$\Psi(\mathbb{P}) = \Psi(\tilde{\mathbb{P}}) + \mathbb{P}[\phi(\mathbf{Z}, \tilde{\mathbb{P}})] + R(\mathbb{P}, \tilde{\mathbb{P}}). \tag{5.3}$$

In the next section, we use these results to discuss the asymptotic properties of plug-in estimators when ML algorithms are used to estimate the nuisance functions. We will see that these estimators are subject to a so called plug-in bias and we will discuss methods that can be used to correct these estimators for this plug-in bias.

5.2 The construction of debiased estimators

A possible approach to estimate $\Psi(\mathbb{P})$ could be to *plug-in* an estimator $\hat{\mathbb{P}}_n$ of the true distribution $\mathbb{P}$. Such estimators are refered to as *plug-in* estimators. In the context of estimating the ATE where $\mathbf{Z}$ would be replaced by the random vector $\mathbf{O} = (Y, \mathbf{X}, W)^t$, $\hat{\mathbb{P}}_n$ could be any distribution for $(Y, \mathbf{X}, W)$ such that the marginal distribution of $\mathbf{X}$ is estimated by its empirical distribution and the conditional expectations $g_w(\mathbf{X})$ are estimated using a ML estimation procedure.

Instead of considering an arbitrary distribution $\tilde{\mathbb{P}}$ in $\mathcal{P}$ as in the previous section, we now consider the estimated distribution $\hat{\mathbb{P}}_n$. The results described in the previous section can now be used to investigate the difference between $\Psi(\hat{\mathbb{P}}_n)$ and $\Psi(\mathbb{P})$. In particular, expansion (5.3) can be rewritten as

$$\begin{aligned}\Psi(\hat{\mathbb{P}}_n) - \Psi(\mathbb{P}) &= -\mathbb{P}[\phi(\mathbf{Z}, \hat{\mathbb{P}}_n)] - R(\mathbb{P}, \hat{\mathbb{P}}_n) \\ &= (\mathbb{P}_n - \mathbb{P})[\phi(\mathbf{Z}, \mathbb{P})] - \mathbb{P}_n[\phi(\mathbf{Z}, \hat{\mathbb{P}}_n)] + (\mathbb{P}_n - \mathbb{P})[\phi(\mathbf{Z}, \hat{\mathbb{P}}_n) - \phi(\mathbf{Z}, \mathbb{P})] - R(\mathbb{P}, \hat{\mathbb{P}}_n) \\ &= \underbrace{\frac{1}{N}\sum_{i=1}^{N} \phi(\mathbf{Z}_i, \mathbb{P})}_{S} - \underbrace{\frac{1}{N}\sum_{i=1}^{N} \phi(\mathbf{Z}_i, \hat{\mathbb{P}}_n)}_{B} + \underbrace{(\mathbb{P}_n - \mathbb{P})[\phi(\mathbf{Z}, \hat{\mathbb{P}}_n) - \phi(\mathbf{Z}, \mathbb{P})]}_{T_1} - \underbrace{R(\mathbb{P}, \hat{\mathbb{P}}_n)}_{T_2}\end{aligned} \tag{5.4}$$

where the second equality is obtained by adding and subtracting $(\mathbb{P}_n - \mathbb{P})[\phi(\mathbf{Z}, \mathbb{P})]$ and $\mathbb{P}_n[\phi(\mathbf{Z}, \hat{\mathbb{P}}_n)]$ and the last equality holds by the zero mean property of influence

[1] To see this, it is useful to note that one can redefine $a = 1 - t$ such that $\left.\frac{\partial \Psi(\mathbb{P} + t(\mathbb{P} - \tilde{\mathbb{P}}))}{\partial t}\right|_{t=1} = \left.\frac{\partial \Psi(\tilde{\mathbb{P}} + a(\tilde{\mathbb{P}} - \mathbb{P}))}{\partial a}\right|_{a=0}$ where the application of definition 2 follows naturally.

functions. We can use equation (5.4) to investigate the asymptotic properties of plug-in estimators.

If terms B, T_1 and T_2 can be shown to converge to 0 and be of order at least $\sqrt{N}$, one can write

$$\sqrt{N}\big(\Psi(\hat{\mathbb{P}}_n) - \Psi(\mathbb{P})\big) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \phi(\mathbf{Z}_i, \mathbb{P}) + o_p(1).$$

Such estimators are called asymptotically linear estimators (Newey, 1994) and their asymptotic distribution is equivalent to that of S which is a sample average of a deterministic function of the data whose asymptotic distribution can be understood by the CLT. We have that

$$S \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{P}[\phi(\mathbf{Z}, \mathbb{P})^2]}{N}\right) \tag{5.5}$$

where the 0 expectation follows from the 0 expectation property of the influence function.

A very attractive property of such estimators relates to their variance. In parametric statistics, efficient estimators are estimators whose variance reaches the Cramer-Rao lower bound. This bound can be computed as the variance of the score function.
In non-parametric statistics, the influence function can be viewed as an equivalent to the score function and the lower bound for the variance of a non parametric estimator is the variance of the influence function of the estimand (Newey, 1994, Kennedy, 2022). Importantly, this implies that asymptotically linear estimators are asymptotically efficient.

Given the attractive properties of such estimators, it would be desirable to be able to control the asymptotic behaviour of B, T_1 and T_2 such that one obtains an asymptotically linear estimator when ML methods are involved in the estimation of $\mathbb{P}$.

Plug-in bias and the debiased estimator

When Machine learning algorithms are involved in the estimation of $\mathbb{P}$, understanding the asymptotic behavior of term B is complicated. This is because the asymptotic behavior of $\hat{\mathbb{P}}_n$ is hard to understand when it involves ML algorithms. In particular, ML estimators often suffer from regularization bias and the convergence of this term is too slow. This has motivated the elaboration of strategies to *adjust the plug-in estimators* such that term B can be ignored in the expansion. The most straightforward way to do so is to directly adjust the plug-in estimator by rewriting the above expansion as

$$\begin{aligned}\Psi(\hat{\mathbb{P}}_n) + \frac{1}{N}\sum_{i=1}^{N} \phi(\mathbf{Z}_i, \hat{\mathbb{P}}_n) - \Psi(\mathbb{P}) = \frac{1}{N}\sum_{i=1}^{N} \phi(\mathbf{Z}_i, \mathbb{P}) \\ + (\mathbb{P}_n - \mathbb{P})[\phi(\mathbf{Z}, \hat{\mathbb{P}}_n) - \phi(\mathbf{Z}, \mathbb{P})] - R(\mathbb{P}, \hat{\mathbb{P}}_n),\end{aligned} \tag{5.6}$$

where $\Psi(\hat{\mathbb{P}}_n) + \frac{1}{N}\sum_{i=1}^{N} \phi(\mathbf{Z}_i, \hat{\mathbb{P}}_n)$ can be obtained from the sample and is refered to as the one step estimator of $\Psi(\mathbb{P})$. This implies that the asymptotic properties of the one step estimator can be investigated by considering only terms S, T_1 and T_2.

Other strategies to remove term B include the construction of estimating equations[2] or Targeted Maximum Likelihood.

The empirical process term T_1

Term T_1 is the so called empirical process term. We would like to show that this term is at least of order $\sqrt{N}$ such that it can be ignored in the above expansion.

In the process of building the one-step estimator, the sample is used twice for two related estimation procedures. Once for building the estimator $\hat{\mathbb{P}}_n$ and once to compute $\mathbb{P}_n\left[\phi(\mathbf{Z}, \hat{\mathbb{P}}_n)\right]$. This induces a risk of overfitting (Kennedy, 2022) which is likely to reduce the convergence rate of the empirical process term. One way to avoid overfitting is to use sample splitting where $\hat{\mathbb{P}}_n$ and $\mathbb{P}_n\left[\phi(\mathbf{Z}, \hat{\mathbb{P}}_N)\right]$ are computed on independent subsamples. In the next paragraphs, we discuss why and how sample splitting can be used to control the asymptotic behaviour of the empirical process term.

A prerequisite for controlling the asymptotic behavior of the empirical process term through sample splitting is that $\phi(\mathbf{Z}, \hat{\mathbb{P}}_n)$ converges to $\phi(\mathbf{Z}, \mathbb{P})$ in L_2 norm.

Assumption 5.1 (Kennedy, 2022)**.**

$$||\phi(\mathbf{z}, \hat{\mathbb{P}}_n) - \phi(\mathbf{z}, \mathbb{P})||^2 \equiv \int \{\phi(\mathbf{z}, \hat{\mathbb{P}}_n) - \phi(\mathbf{z}, \mathbb{P})\}^2 \, d\mathbf{z} = o_p(1).$$

The idea of sample splitting is to split the sample $\mathbf{Z}_i^N$ into K disjoint folds where K is fixed in advance. Let $\hat{\mathbb{P}}_{-k}$ be an estimator for $\mathbb{P}$ obtained from all observations except those from fold k. We define $\hat{\Psi}$, the sample splitting version of the one step estimator described above as

$$\hat{\Psi} = \sum_{k=1}^{K} \Big(\frac{N_k}{N}\Big) \hat{\Psi}_k$$

where N_k is the number of observations in fold $k = 1, \cdots, K$ and

$$\hat{\Psi}_k = \Psi\big(\hat{\mathbb{P}}_{-k}\big) + \mathbb{P}_n^k\big[\phi(\mathbf{Z}, \hat{\mathbb{P}}_{-k})\big]$$

is the one step estimator in fold k based on estimated of $\hat{\mathbb{P}}_{-k}$ on all observations except those in fold k. Note that $\mathbb{P}_n^k$ denotes the empirical measure over fold k. For each k, one can now write

$$\hat{\Psi}_k - \Psi = \mathbb{P}_n^k \phi(\mathbf{Z}_i, \mathbb{P}) + (\mathbb{P}_n^k - \mathbb{P})[\phi(\mathbf{Z}_i, \hat{\mathbb{P}}_{-k}) - \phi(\mathbf{Z}_i, \mathbb{P})] - R(\mathbb{P}, \hat{\mathbb{P}}_{-k}).$$

The difference between $\hat{\Psi}$ and Ψ can be written as

$$\hat{\Psi} - \Psi = S + \sum_{k=1}^{K} \Big(\frac{N_k}{N}\Big)\big(T_{1k} + T_{2k}\big)$$

[2]This approach is for example advocated in Chernozhukov et al. (2018) and gives the same results as the one step approach when the influence function is linear in the estimand of interest. This is the case for the ATE as will become clear in the next section.

The order of $\sum_{k=1}^{K}\left(\frac{N_k}{N}\right)\left(T_{1k}+T_{2k}\right)$ will be the same as that of $\max_k(T_{1k}+T_{2k})$ where the asymptotic behaviour of T_{1k} and T_{2k} can be analyzed separately.

The following lemma from Kennedy (2022) can be used to show that when sample splitting is used, T_{1k} (and hence T_1) is $o_p\left(\frac{1}{\sqrt{N}}\right)$.

Lemma 1. *(Kennedy, 2022). Let $\hat{f}(z)$ be a function estimated from a sample $\mathbf{Z}^D = (\mathbf{Z}_{N+1}, \cdots, \mathbf{Z}_D)^t$. Let $\mathbb{P}_n$ denote the empirical measure over $(\mathbf{Z}_1, \cdots, \mathbf{Z}_N)^t$ independent of $\mathbf{Z}^D$. Then, $(\mathbb{P}_n - \mathbb{P})(\hat{f} - f) = O_p\left(\frac{||\hat{f}-f||}{\sqrt{N}}\right)$.*

A proof of this lemma can be found in Kennedy (2022). By a direct application of this lemma

$$\frac{(\mathbb{P}_n^k - \mathbb{P})(\phi(\mathbf{Z}, \hat{\mathbb{P}}_{-k}) - \phi(\mathbf{Z}, \mathbb{P}))}{\sqrt{N}} = O_p(||\phi(\mathbf{Z}, \hat{\mathbb{P}}_{-k}) - \phi(\mathbf{Z}, \mathbb{P})||).$$

Since $||\phi(\mathbf{Z}, \hat{\mathbb{P}}) - \phi(\mathbf{Z}, \mathbb{P})|| = o_p(1)$ by assumption, it follows that $(\mathbb{P}_n^k - \mathbb{P})(\phi(\mathbf{Z}, \hat{\mathbb{P}}_{-k}) - \phi(\mathbf{Z}, \mathbb{P})) = o_p\left(\frac{1}{\sqrt{N}}\right)$.

Remainder term

The asymptotic behaviour of the remainder term T_2 depends on the considered estimand as well as on the estimation methods involved in the estimation of $\mathbb{P}$. Hence, we postpone the discussion of the asymptotic behaviour of the remainder term to the next section where we introduce the one step estimator for the specific case of estimating the ATE.

5.3 The DML estimator of the ATE

In the previous section, we explained how debiased estimators can be built for pathwise differentiable estimands. A very attractive feature of this approach is that it applies to a broad range of estimands (among which many causal estimands). In this section, we derive the debiased estimator for the ATE. As it appears clear from the previous section, deriving the debiased estimator for the ATE relies crucially on showing that the ATE is pathwise differentiable and finding its influence function.

Deriving the influence function of the ATE

We derive the influence function for the ATE following the point mass contamination approach developed in Ichimura and Newey (2015) and applied to the case of the ATE by Hines et al. (2022).

From definition 2, it appears that the influence function can be isolated by considering a specific point $\tilde{\mathbf{z}}$ in the support of $\mathbb{P}$ rather than the entire distribution $\tilde{\mathbb{P}}$. Indeed, one can then rewrite the Gateaux derivative as

$$\phi(\mathbf{z}, \mathbb{P}) = \left.\frac{\partial \Psi(\mathbb{P} + t(\delta_{\tilde{\mathbf{z}}}(\mathbf{z}) - \mathbb{P}))}{\partial t}\right|_{t=0} \tag{5.7}$$

where $\delta_{\tilde{\mathbf{z}}}$ denotes the Dirac delta function which integrates to 1 and is equal to 0 everywhere except at $\tilde{\mathbf{z}}$. Since Gateaux derivatives have the same properties as usual derivatives (Hines et al., 2022), one can compute the efficient influence function using the usual properties of derivatives.

In the next paragraphs, we derive the influence function for the ATE. Under unconfoundedness and overlap, the ATE can be written as

$$\begin{aligned}\Psi(\mathbb{F}) &= \mathbb{E}\Big[\mathbb{E}\left[Y(1)|\mathbf{X}, W=1\right] - \mathbb{E}[Y(0)|\mathbf{X}, W=0]\Big] \\ &= \int\int yf(y|\mathbf{x}, w=1)f(x)\,dy\,d\mathbf{x} - \int\int yf(y|\mathbf{x}, w=0)f(x)\,dy\,d\mathbf{x}\end{aligned} \tag{5.8}$$

where $f(y|\mathbf{x}, w=1)$ denotes the conditional probability density function of Y and $f(\mathbf{x})$ denotes the probability density function of $\mathbf{X}$.
Then,

$$\begin{aligned}\Psi(\mathbb{F}_t) &= \int\int yf_t(y|\mathbf{x}, w=1)f_t(x)dyd\mathbf{x} - \int\int yf_t(y|\mathbf{x}, w=0)f_t(x)\,dy\,d\mathbf{x} \\ &= \int\int y\frac{f_t(y, \mathbf{x}, w=1)f_t(\mathbf{x})}{f_t(\mathbf{x}, w=1)}\,dy\,d\mathbf{x} - \int\int y\frac{f_t(y, \mathbf{x}, w=0)f_t(\mathbf{x})}{f_t(\mathbf{x}, w=0)}\,dy\,d\mathbf{x}\end{aligned} \tag{5.9}$$

where $\mathbb{F}_t$ is the parametric submodel and $f_t(.)$ are the respective density functions of the submodel.

We would like to evaluate the following Gateaux derivative

$$\phi(\mathbf{o}, \mathbb{F}) = \left.\frac{\partial\Psi(\mathbb{F} + t(\delta_{\tilde{\mathbf{x}},\tilde{y},\tilde{w}}(\mathbf{x}, y, w) - \mathbb{F}))}{\partial t}\right|_{t=0}$$

with $\mathbf{o}$ as the realization of $\mathbf{O}$. Since the ATE has the form of a difference between two expectations, we start by evaluating the derivative for the expected potential outcome under treatment (first term on the right hand side of equation 5.9) that we denote $\Psi_1(\mathbb{F})$

$$\begin{aligned}\left.\frac{\partial\Psi_1(\mathbb{F}_t)}{\partial t}\right|_{t=0} &= \int\int y\Bigg\{\frac{f(\mathbf{x})}{f(w=1,\mathbf{x})}\frac{d}{dt}f_t(y, w=1, \mathbf{x})\Big|_{t=0} \\ &\qquad + \frac{f(y, w=1, \mathbf{x})}{f(w=1,\mathbf{x})}\frac{d}{dt}f_t(\mathbf{x})\Big|_{t=0} - \frac{f(y, w=1, \mathbf{x})}{f(w=1,\mathbf{x})^2}\frac{d}{dt}f_t(w=1, \mathbf{x})\Big|_{t=0}\Bigg\}dyd\mathbf{x} \\ &= \int\int \frac{f(y, w=1, \mathbf{x})f(\mathbf{x})}{f(w=1,\mathbf{x})}\left(\frac{\delta_{\tilde{y},\tilde{\mathbf{x}},\tilde{w}}(y, w=1, \mathbf{x})}{f(y, w=1, \mathbf{x})} - \frac{\delta_{\tilde{\mathbf{x}},\tilde{w}}(w=1, \mathbf{x})}{f(w=1,\mathbf{x})} + \frac{\delta_{\tilde{x}}(\mathbf{x})}{f(\mathbf{x})} - 1\right)dyd\mathbf{x}.\end{aligned} \tag{5.10}$$

The first equality is obtained by the chain rule. The second equality is obtained by taking the derivatives of the respective densities with respect to t and simplifying. When evaluating the integral in the last expression, one obtains

$$\phi_1(\mathbf{o}, \mathbb{F}) = \frac{\delta_{\tilde{w}}(w=1)}{\pi(\tilde{\mathbf{x}})}[\tilde{y} - g_1(\tilde{\mathbf{x}})] + g_1(\tilde{\mathbf{x}}) - \Psi_1(\mathbb{F}), \tag{5.11}$$

which is the influence function for the expected potential outcome under treatment. The same reasoning can be used to derive the influence function for the expected potential outcome in the absence of treatment

$$\phi_0(\mathbf{o}, \mathbb{F}) = \frac{\delta_{\tilde{w}}(w=0)}{1-\pi(\tilde{\mathbf{x}})}[\tilde{y} - g_0(\tilde{\mathbf{x}})] + g_0(\tilde{\mathbf{x}}) - \Psi_0(\mathbb{F}). \tag{5.12}$$

Combining these two expressions yields the influence function for the ATE as

$$\phi(\mathbf{O}, \mathbb{F}) = \frac{W}{\pi(\mathbf{X})}[Y - g_1(\mathbf{X})] - \frac{1-W}{1-\pi(\mathbf{X})}[Y - g_0(\mathbf{X})] + (g_1(\mathbf{X}) - g_0(\mathbf{X})) - \Psi(\mathbb{F}).$$

The debiased estimator for the ATE

Using the influence function for the ATE derived above, the one step estimator defined in the previous section is obtained as

$$\begin{aligned}
\hat{\tau}_{os} &= \Psi(\hat{\mathbb{F}}_n) + \frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{O}_i, \hat{\mathbb{F}}_n) \\
&= \Psi(\hat{\mathbb{F}}_n) + \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{W_i}{\pi(\mathbf{X}_i)}\big[Y_i - g_1(\mathbf{X}_i)\big] - \frac{1-W_i}{1-\pi(\mathbf{X}_i)}\big[Y_i - g_0(\mathbf{X}_i)\big] + g_1(\mathbf{X}_i) - g_0(\mathbf{X}_i) - \Psi(\hat{\mathbb{F}}_n)\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{W_i}{\pi(\mathbf{X}_i)}\big[Y_i - g_1(\mathbf{X}_i)\big] - \frac{1-W_i}{1-\pi(\mathbf{X}_i)}\big[Y_i - g_0(\mathbf{X}_i)\big] + g_1(\mathbf{X}_i) - g_0(\mathbf{X}_i)\right\}.
\end{aligned} \tag{5.13}$$

Note that we recover the empirical analog of the Augmented Probability Weighted (AIPW) representation of the ATE discussed in chapters 2 and 3.

The asymptotic properties of the estimator

From the discussion in the previous section, asymptotic normality and efficiency will follow if one can show $\sqrt{N}$ convergence of the empirical process and remainder terms.

The asymptotic behaviour of the empirical process term, can be controlled by applying the sample splitting procedure described in the previous section. Concretely, the sample can be split into K subsamples. For each subsample k, $g_w(\mathbf{X})$ and $\pi(\mathbf{X})$ are estimated using all observations except those in subsample k and $\hat{\tau}_{os}$ is then computed in sample k as

$$\hat{\tau}_{DML} = \sum_{k=1}^{K}\Big(\frac{N_k}{N}\Big)\hat{\tau}_{os}^{k}$$

where N_k is the number of observations in fold k and

$$\hat{\tau}_{os}^{k} = \frac{1}{N_k}\sum_{i=1}^{N_k}\left\{\frac{W_i}{\hat{\pi}^{-k}(\mathbf{X}_i)}\big[Y_i - \hat{g}_1^{-k}(\mathbf{X}_i)\big] - \frac{1-W_i}{1-\hat{\pi}(\mathbf{X}_i)^{-k}}\big[Y_i - \hat{g}_0^{-k}(\mathbf{X}_i)\big] + \hat{g}_1^{-k}(\mathbf{X}_i) - \hat{g}_0^{-k}(\mathbf{X}_i)\right\}$$

where $\hat{\pi}(\mathbf{X})^{-k}$ and $\hat{g}_0^{-k}(\mathbf{X})$ are obtained via Machine Learning algorithms using all observations except those in subsample k. The obtained estimator is equivalent to the Debiased Machine Learning (DML) estimator described in Chernozhukov et al. (2018).

We still need to discuss consistency of the remainder term. We first describe the asymptotic behaviour of the remainder term for the average potential outcome under treatment. From equation (5.3), the remainder for the average potential outcome under treatment can be computed as

$$\begin{aligned}
R_1(\mathbb{F}, \hat{\mathbb{F}}_n) &= \Psi_1(\mathbb{F}) - \Psi_1(\hat{\mathbb{F}}_n) - \mathbb{E}[\phi_1(\mathbf{O}, \hat{\mathbb{F}}_n)] \\
&= \Psi_1(\mathbb{F}) - \Psi_1(\hat{\mathbb{F}}_n) - \mathbb{E}\Bigg\{ \frac{W}{\pi(\mathbf{X}, \hat{\mathbb{F}}_n)} \big[Y - g_1(\mathbf{X}, \hat{\mathbb{F}}_n)\big] + g_1(\mathbf{X}, \hat{\mathbb{F}}_n) - \Psi_1(\hat{\mathbb{F}}_n) \Bigg\} \\
&= -\mathbb{E}\Bigg\{ \frac{WY}{\pi(\mathbf{X}, \hat{\mathbb{F}}_n)} - \frac{g_1(\mathbf{X}, \hat{\mathbb{F}}_n)}{\pi(\mathbf{X}, \hat{\mathbb{F}}_n)} + g_1(\mathbf{X}, \hat{\mathbb{F}}_n) - \Psi_1(\mathbb{F}) \Bigg\} \\
&= -\mathbb{E}\Bigg\{ \frac{E\big[W|\mathbf{X}\big] E\big[Y(1)|\mathbf{X}\big]}{\pi(\mathbf{X}, \hat{\mathbb{F}}_n)} - \frac{g_1(\mathbf{X}, \hat{\mathbb{F}}_n)}{\pi(\mathbf{X}, \hat{\mathbb{F}}_n)} + g_1(\mathbf{X}, \hat{\mathbb{F}}_n) - \Psi_1(\mathbb{F}) \Bigg\} \\
&= -\mathbb{E}\Bigg\{ \left[\frac{\pi(\mathbf{X}, \mathbb{F})}{\pi(\mathbf{X}, \hat{\mathbb{F}}_n)} - 1 \right] \Big[g_1(\mathbf{X}, \mathbb{F}) - g_1(\mathbf{X}, \hat{\mathbb{F}}_n) \Big] \Bigg\}
\end{aligned} \tag{5.14}$$

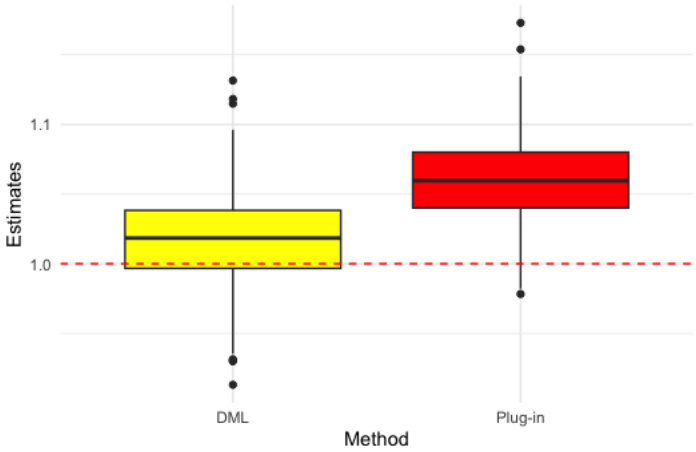
where the last equality follows from the law of iterated expectations and unconfoundedness. Note that $\pi(\mathbf{X}, \mathbb{F})$ refers to the true propensity score while $\pi(\mathbf{X}, \hat{\mathbb{F}}_n)$ refers to its estimated version using a data-adaptive methods. The last equality in equation (5.14) can be bounded to the following expression by a direct application of Cauchy Schwartz inequality

$$\mathbb{E}\Bigg\{ \left[\frac{\pi(\mathbf{X}, \mathbb{F})}{\pi(\mathbf{X}, \hat{\mathbb{F}}_n)} - 1 \right]^2 \Bigg\}^{\frac{1}{2}} \mathbb{E}\Bigg\{ \Big[g_1(\mathbf{X}, \mathbb{F}) - g_1(\mathbf{X}, \hat{\mathbb{F}}_n) \Big]^2 \Bigg\}^{\frac{1}{2}}.$$

The convergence rate of this term is determined by the convergence rate of $g_1(\mathbf{X}, \hat{\mathbb{F}}_n)$ and $\pi(\mathbf{X}, \hat{\mathbb{F}}_n)$. More precisely, it is determined by the product of the convergence rates of the two. This property is termed rate double robustness (Hines et al., 2022) because the convergence rate of the resulting estimator is greater than that of the estimators for the nuisance functions. For example, if both $g_1(\mathbf{X}, \hat{\mathbb{F}}_n)$ and $\pi(\mathbf{X}, \hat{\mathbb{F}}_n)$ are of order $N^{\frac{1}{4}}$, the remainder is of order $N^{\frac{1}{2}}$ and so is the final estimator for the ATE.

Hence, if one can show that a given data-adaptive estimator achieves such convergence rates, the DML estimator is asymptotically efficient with asymptotic variance equal to the variance of the efficient influence function. A discussion on such convergence results exist for most ML algorithms (see for example Chernozhukov et al. (2017) for some references).

Figure 5.1: Bias adjustment due to DML



Note: The boxplots represent the distribution of the DML (yellow) and plug-in (red) estimators on 500 replications of the DGP used in example 3. Each sample was generated with $p = 10$ and $n = 5000$ and the nuisance functions for both estimators were computed using random forests with the same hyperparameters. (A detailed description of Random Forests can be found in the next section)

The bias reduction associated to the use of the DML procedure is illustrated in figure 5.1 on the DGP of example 3. The estimated bias is clearly smaller for the DML estimator than for the Plug-in estimator.

Finally, motivated by equation (5.5) a consistent estimator for the variance of the estimator can be obtained as

$$\hat{\sigma}^2_{DML} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \left[\hat{\tau}^k_{os}\right]^2 . \tag{5.15}$$

Part III

Finite sample results

Chapter 6

Monte Carlo simulations

In this chapter we perform a Monte Carlo simulation study to evaluate the finite sample performance of the debiased ML estimator for the ATE described in the previous chapter. We consider different scenarios for the data generating process and investigate the performance of the DML estimator across these different scenarios. The considered scenarios will involve challenging nuisance functions where simple parametric methods are not expected to work well. The performance of DML estimators will be compared to that of the traditional estimation methods presented in chapters 3 and 4. In addition, we will also benchmark the performance of the DML estimator to Bayesian Additive Regression Trees (BART), a Bayesian approach to the estimation of causal effects involving random forests that has demonstrated very good performance in the recent past (see for example Dorie et al., 2019).

This chapter is organized as follows. The next section discusses the practical implementation of the considered estimators. Then, we present the simulation design as well as the considered scenarios. Finally, the results of the simulation study are discussed.

6.1 Estimators under consideration

In this section, we clarify some practical aspects of the considered estimators. All estimators discussed below are implemented using R (R Core Team, 2021).

6.1.1 Traditional estimation approaches

Since the DGP will be designed to evaluate the performance of the estimators when the nuisance functions are complex and non-linear, we do not consider parametric model based estimation strategies except **OLS** that we use as a benchmark. While it is not expected to work well on the considered DGPs it allows to assess the importance and potential advantages of the more flexible methods in such a context. If not explicitly stated otherwise, we consider a relatively naive implementation of the OLS estimator and only include main effects (no interactions or higher order terms) in the specification.

We also consider the **nearest neighbor** regression estimator as well as its bias corrected version discussed in chapter 4. The nearest neighbor regression estimator is

implemented using the `Matching` package of Sekhon (2008). This package is also used to compute the **bias corrected** version of the estimator where the bias correction is performed using a simple linear regression. The asymptotic variance is estimated based on equation (4.7).

6.1.2 Debiased Machine Learning

DML estimators are implemented using the `DoubleML` package of Bach et al. (2021). DML estimators rely on ML estimation of the nuisance functions. In our applications we will consider Gradient Boosting and Random forests to estimate the nuisance functions. The algorithms are tuned and fitted using the `mlr3` package of Lang et al. (2019). We briefly present Random forests and Gradient Boosting in the next paragraphs.

All DML estimators are constructed using the sample splitting procedure described in the previous chapter. While there are no clear recommendations regarding the number of splits to consider, choosing a number of folds around 4 has been shown to work well in practice (Chernozhukov et al., 2018, McConnell and Lindner, 2019). An estimator for the variance of the estimator is obtained from equation (5.15).

Random forest

The objective of random forests is to use covariates to accurately predict the value of an outcome for a new sample. Random forest is an aggregation of classification trees (for discrete outcomes) or regression trees (for continuous outcomes). A single regression tree is a decision rule partitioning the covariate space into different subspaces. Starting from the entire sample, a split of the sample involving a single splitting covariate and a splitting point on that covariate is considered. The split (splitting covariate and splitting point) is chosen to minimize the sample mean squared error defined (for an outcome Y) as

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \tag{6.1}$$

where $\bar{Y}$ is the empirical mean. From the newly created subsamples, new splits are then considered using the same optimization rule.

Considering the final partition of the sample, predictions for a new observation are obtained as the average outcome in the subsample to which the observation to be predicted belongs. A single regression tree in an ensemble of H regression trees is defined as

$$\hat{g}_h(\mathbf{X}) = r_c \mathbb{1}(\mathbf{X} \in \mathcal{A}_c) \qquad \text{for} \qquad 1 \leq c \leq C$$

where $\mathcal{A}_c$ denotes subsample c with C the total number of subsamples and r_c is the empirical mean in subsample c.

Note that adding additional splits will always reduce the mean squared error in sample. However, very deep trees are subject to overfitting and will not perform well to predict new data. Overfitting can be adressed by introducing some regularization, i.e, choosing hyper parameters that constrain the complexity of the trained trees. For example, one can constrain the maximum number of subspaces that a tree can contain.

Single regression trees tend to produce very discontinuous prediction rules. This motivates the use of random forests which consists in an aggregation of single regression trees. The aggregation procedure is characterized by two important features. First, each tree is trained using a bootstrapped (sampling with replacement from the original sample) sub sample. This procedure is called Bagging. Second, for each potential tree, only a subset of m covariates is chosen from the total number of available covariates p. The procedure is summarized in the following algorithm from Efron and Hastie (2021).

Algorithm 1 Random forest (Efron and Hastie, 2021)

Fix $m < p$ and the number of trees to H.
for `h in 1:H` **do**
(a) Create a bootstrap version of the training data by randomly sampling the N rows with replacement N times.
(b) Grow a tree $\hat{g}_{rt}(\mathbf{X})$ using the bootstrap sample, sampling m of the p covariates at random prior to making each split.
end for
Compute the random forest fit at any prediction point $\mathbf{X}_0$ as the average

$$\hat{g}_{rf}(\mathbf{X}_0) = \frac{1}{H}\sum_{h=1}^{H} \hat{g}_h(\mathbf{X}_0)$$

Recall from the previous chapter that the desirable asymptotic properties of DML estimators are conditional on the convergence behaviour of the considered Machine Learning methods. Biau (2012) shows that the convergence rate of a simplified version of the random forest algorithm depends on the number of *strong covariates*, i.e, covariates that are related to the outcome rather than on the number of total features. The idea behind this result is that irrelevant covariates are likely to be ignored in the process of building the regression tree. This result suggests some robustness of the random forest estimator to the presence of a large number of noise variables in the data. This is an important difference with the Nearest Neighbor or Kernel regression approaches discussed in chapter 3. We refer to Wager and Walther (2015) and Biau (2012) for an in depth discussion of the asymptotic properties of random forest estimators.

Gradient Boosting

The Gradient Boosting algorithm is also a combination of regression trees. However it differs in the way in which the trees are constructed and aggregated. Instead of fitting regression trees on the outcome as for random forests, individual trees are fitted on the residuals of the predictor obtained in the previous step (except for the first tree which is fitted on the outcome) with the aim of minimizing the residual sum of squares. At each step, the predictor is updated considering a fraction of the newly built tree. The idea of fitting trees on the residuals rather than the true outcome is to give more weight to observations that were not accurately predicted by the predictors in the previous step. The procedure is described in the following algorithm from Efron and Hastie (2021).

Algorithm 2 Gradient Boosting (Efron and Hastie, 2021)

Fix the number of steps to H, a shrinkage factor ϵ, a maximum tree depth (maximum number of splits) d. Set the initial fit $\hat{G}_0 \equiv 0$ and the initial vector of residuals to $\mathbf{R} = \mathbf{Y}$.

for `h in 1:H` **do**

(a) Fit a regression tree $\tilde{g}_h$ to the data $(\mathbf{X}, \mathbf{R}_{h-1})$ with maximum depth d where $\mathbf{R}_{h-1}$ are the residuals from the previous step. The splits are chosen to minimize the residual sum of squares

$$\sum_{i=1}^{N} \left(R_{h-1,i} - \hat{G}_h(\mathbf{X}_i) \right)^2.$$

(b) Update the fitted model with a shrunken version of $\tilde{g}_h$ as $\hat{G}_h = \hat{G}_{h-1} + \epsilon \tilde{g}_h$.

(c) Update the residuals $R_i = R_{h-1,i} - \epsilon \tilde{g}_h$ for $i = 1, \ldots, N$

end for

Return the final predictor $\hat{G}_H(\mathbf{X}_i)$.

While there are no clear results on the convergence rate for boosted trees algorithms, an informal discussion on these convergence rates can be found in Yang et al. (2020).

In our simulation study, the hyper parameters (d and ϵ) are tuned to minimize the mean squared error (or classification error for the propensity score) using cross validation.

6.1.3 An additional competitor: Bayesian Additive Regression Trees (BART)

BART was developped by Chipman et al. (2010) and is discussed in the context of causal inference by Hill (2011) and Hahn et al. (2020). BART assumes a model of the form

$$Y = g(\mathbf{X}, W) + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

where $g(\mathbf{X}, W) = E[Y|\mathbf{X}, W]$ is assumed to be an aggregation of H trees

$$g(\mathbf{X}, W) = g_1(\mathbf{X}, W, T_1, M_1) + \ldots + g_H(\mathbf{X}, W, T_H, M_H). \tag{6.2}$$

In each individual tree h, T_h refers to the set of decision rules defining the tree and $M_h = \left(r_{h1}, \ \ldots, \ r_{hc}, \ \ldots, \ r_{hC}\right)^t$ refers to the vector of final means (predictions) in each subspace c determined by the tree where C refers to the number of subspaces specified by the tree.

The parameters of the model are (T_h, M_h) and σ. The idea of BART is to use a traditional Bayesian approach to the estimation of this model. First, a prior joint distribution is defined for the parameters of the model. The prior is then updated using Monte Carlo Markov Chains to build a joint posterior distribution. Bayesian inference can then be performed by sampling from the obtained posterior distribution.

The joint distribution for the parameters can be expressed as

$$Pr\big((T_1, M_1), (T_2, M_2), \ldots, (T_h, M_H), \sigma\big) = Pr(T_1, T_2, \cdots, T_H) Pr(M_1, M_2, \ldots, M_H | T_1, T_2, \ldots, T_H) Pr(\sigma). \tag{6.3}$$

In order to simplify this prior joint distribution, the following simplifications are made by introducing some independence assumptions

$$\begin{aligned} Pr(T_1, T_2, \ldots, T_H) &= \Pi_{h=1}^{H} Pr(T_h) \\ Pr(M_1, M_2, \ldots, M_H | T_1, T_2, \ldots, T_H) &= \Pi_{h=1}^{H} Pr(M_h | T_h) \\ Pr(M_h | T_h) &= \Pi_{h=1}^{H} Pr(r_{hc} | T_h). \end{aligned} \tag{6.4}$$

A prior distribution has to be specified for the set of decisions within a tree (T_h), the within subspace means (r_{hc}) and the variance of the error (σ^2).

The prior for the trees is specified as a uniform distribution for the choice of the split on a particular feature. The probability that the tree stops at a certain split is

$$\alpha(1 + d)^{-\beta} \tag{6.5}$$

where α and β are pre specified parameters and d is the number of subspaces already defined at the stage where a given split is evaluated. Note that this prior strongly favors small trees over large trees as the probability of ending the process of growing the tree becomes very small after a small number of splits. Similar to the discussion regarding random forests and boosted trees, the idea behind this *regularization* prior is to avoid overfitting.

Regarding the within subspace averages, for a given subspace, the average is chosen from a normal distribution

$$r_{hc} \sim \mathcal{N}(0, \sigma_\mu^2) \tag{6.6}$$

where $\sigma_\mu = \frac{0.5}{k\sqrt{m}}$ with k chosen arbitrary. The 0 mean arises because the outcome is first scaled to have mean 0 and lie between -0.5 and 0.5. Finally, the prior for σ^2 is selected from a χ^2 distribution.

Monte Carlo Markov Chains (MCMC) algorithms can be used to obtain the joint posterior distribution for the parameters of the model. This happens by exploring new possibilities for the parameters governing the prior distribution and accepting the changes with a probability that depends on a Maximum likelihood based criterion.

One can then obtain draws of $g(\mathbf{X}, W)$ from this joint posterior. For each individual draw l, one can use $g(\mathbf{X}, W)^l$ and the empirical distribution of $\mathbf{X}$ to compute

$$\frac{1}{N} \sum_{i=1}^{N} (g(\mathbf{X}_i, W = 1)^l - g(\mathbf{X}_i, W = 0)^l).$$

This in turn gives us draws from the joint posterior for the ATE. An estimator for the ATE can then be obtained as the average over the draws and bayesian credible intervals can be computed using the quantiles of interest from the draws.

In our empirical applications, we use a modified version of this basic BART that was proposed by Hahn et al. (2020). They consider two modifications of the basic version presented above. First, the model is slightly changed to model the effect of the treatment explicitly as

$$Y = g(\mathbf{X}) + \tau(\mathbf{X}) \times W + \epsilon \qquad\qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

This modification allows to separate the specification of a prior for the conditional mean and for the effect of the treatment. In practice, the prior for the treatment effect is usually chosen to favor even smaller trees than that for the conditional mean. Second, they suggest to include an estimate of the propensity score among the considered covariates. We implement this approach using the `bartCause` package of Dorie and Hill (2020).

6.2 Simulation design

Different data generating processes will be considered for evaluating the performances of the DML estimator. These DGP will differ in various dimensions. We start this section by presenting some general features of the process whereby we will generate the joint distribution $\mathbb{F}$ for $(W, Y, \mathbf{X})^t$ under the regular assignment mechanism. All the considered DGPs will respect the assumptions of the regular assignment mechanism. In other words, the DGP will be such that the effect of the treatment is confounded (in the sense that a difference in expectations would not identify the effect of the treatment) but assumptions (2.2) and (2.3) hold.

In all the considered joint distributions, the $p \times 1$ random vector of covariates $\mathbf{X}$ will be assumed to follow a multivariate normal distribution

$$\mathbf{X} \sim \mathcal{N}(\mu_{\mathbf{X}}, \Sigma)$$

where $\mu_{\mathbf{X}}^t = \big(0, \quad \dots \quad, 0\big)$ and Σ is a $p \times p$ variance covariance matrix with elements $\Sigma_{ij} = 0.5^{|j-i|}$.

The treatment indicator W is assumed to follow a Bernoulli distribution

$$W \sim \text{Bern}\big(\pi(\mathbf{X})\big)$$

where $\pi(\mathbf{X})$ is the propensity score function with $0 < \pi(\mathbf{X}) < 1$.

The outcome variable Y is generated as follows

$$Y = g(W, \mathbf{X}) + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, 1)$. Recall that by equation (2.8), $g(W, \mathbf{X})$ can be decomposed as

$$g(W, \mathbf{X}) = g_0(\mathbf{X}) + W\tau(\mathbf{X})$$

where by the law of iterated expectations, the ATE is the expectation of $\tau(\mathbf{X})$.

The exact specification of the nuisance functions $g_0(W, \mathbf{X})$ and $\pi(\mathbf{X})$ is key to the determination of the DGP. Two different designs will be considered with respect to the

nuisance functions. For each design, the propensity score is obtained from a logistic function as

$$\pi(\mathbf{X}) = \frac{\exp\big(m(\mathbf{X})\big)}{1 + \exp\big(m(\mathbf{X})\big)}$$

where $m(\mathbf{X})$ has to be specified.

For each of these two designs, we will let the following features vary and study how they affect the performance of the considered estimators.

- **The number of covariates,** p. The higher the number of covariates, the harder it should be for the estimators to recover the ATE.
- **Level of confounding** γ. The greater the level of confounding, the harder it should be to estimate the effect of the treatment.
- **The number of observations,** N. With more observations, we expect the bias to decrease and the estimators to become more precise.
- **Heterogeneity in the effect of the treatment,** $\tau(\mathbf{X})$. Adding heterogeneity should make it harder to estimate the effect of the treatment.

Design 1

The first design follows an approach similar to that of Dorie et al. (2019) where $g_0(\mathbf{X})$ and $m(\mathbf{X})$ are constructed as generative additive functions in which each covariate is first passed into a transformation function. The final nuisance function is then obtained by adding up these transformed covariates. We consider a similar set of potential transformation functions as in Dorie et al. (2019). This approach allows to test the robustness of the method to different types of functional forms for the propensity score and outcome regression functions. The exact specifications for $g_0(\mathbf{X})$ and $m(\mathbf{X})$ are

$$m(\mathbf{X}) = -0.4 + 0.1X_1 + 0.1X_2^2 + 0.1X_1X_2 + 0.1\big(\mathbb{1}(X_3 > 0) + \mathbb{1}(X_3 > 1)\big) + 0.1\left(\frac{(X_4 + X_5)^2}{5}\right)$$

$$g_0(\mathbf{X}) = \gamma\left(0.2X_1 + 0.5X_2^2 + 0.5X_1X_2 + 2 \times \big(\mathbb{1}(X_3 > 0) + \mathbb{1}(X_3 > 1)\big) + 0.1\exp(X_4 + X_5))\right)$$

where X_j, $j = 1, ..., 5$ is the j^{th} covariate in $\mathbf{X}$. The parameter γ captures the degree of confounding. The following scenarios will be considered for design 1

$$N \in \{500, 1000, 5000\} \qquad p \in \{5, 50\}$$
$$\gamma \in \{0.5, 1\} \qquad \tau(\mathbf{X}) \in \{1, 1.25 + 2 \times m(\mathbf{X})\}.$$

With these specifications for the propensity score and outcome regression functions, the average of the outcome and the probability to be treated increases in the 5 first covariates. Hence, there is a positive selection bias. The selection bias, heterogeneity bias and the resulting bias from the difference in expectations can be found in table 6.1. Due to the non-linearities included in the nuisance functions, the simple OLS estimator described in the previous section is not expected to perform well. However, it should still capture part of the relationship between the outcome and the covariates and hence perform better than a simple difference in means estimator. Note that only the 5 first covariates are confounders. When considering scenarios with $p > 5$, the

Table 6.1: Summary of true the causal effects (ATE), selection (S.B), heterogeneity bias (H.B) and Simple Difference in Expectations (SDE) for all scenarios

		Design 1				Design 2			
		p = 5		p = 50		p=5		p=50	
		$\gamma = 0.5$	$\gamma = 1$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 0.5$	$\gamma = 1$
τ	ATE	1	1	1	1	1	1	1	1
	SDE	1.23	1.44	1.23	1.44	1.2	1.41	1.21	1.43
	S.B	0.23	0.44	0.23	0.44	0.2	0.41	0.21	0.43
	H.B	0	0	0	0	0	0	0	0
$\tau(\mathbf{X})$	ATE	1	1	1	1	1.03	1.03	1	1
	SDE	1.31	1.53	1.31	1.53	1.33	1.54	1.33	1.54
	S.B	0.23	0.45	0.23	0.45	0.2	0.41	0.22	0.43
	H.B	0.08	0.08	0.08	0.08	0.09	0.09	0.1	0.1

additional covariates will be noise features. They will be unrelated to the treatment and the outcome.

Design 2

The second design will rely on nuisance functions that are very far from being linear. The design is summarized below.

$$m(\mathbf{X}) = -0.2 + \cos(\mathbf{X}^t\boldsymbol{\beta})$$
$$g_0(\mathbf{X}) = \gamma \times (\cos(\mathbf{X}^t\boldsymbol{\beta}))$$

where $\boldsymbol{\beta}^t = \left(1, \ \frac{1}{2}, \ \ldots, \ \frac{1}{p}\right)$ and the constant in $m(\mathbf{X})$ is chosen to balance the treatment assignment. Trigonometric functions are often used to assess the capacity of estimators to fit complex functions in the context of predictions (Friedman, 1991) or in a context of causal inference (McConnell and Lindner, 2019; Knaus et al., 2021b).

The following scenarios will be considered for design 2

$$N \in \{500, 1000, 5000\} \qquad p \in \{5, 50\}$$
$$\gamma \in \{0.5, 1\} \qquad \tau(\mathbf{X}) \in \{1, (2 \times \pi(\mathbf{X}))\}$$

Note that the heterogeneity depends directly on the propensity score. Units with greater probability of treatment benefit more from the treatment.

In table 6.1, we summarize the causal effects (ATE), difference in expectation (SDE) between treated and untreated units, selection (S.B) and heterogeneity biases (H.B) for the different scenarios[1].

6.3 Performance evaluation

We will draw B samples of size N for each considered DGP. For each sample in these B repetitions, we estimate τ using the methods presented above. The finite sample

[1]When there is heterogeneity in the effect of the treatment, the true ATE was obtained from a simulation over a very large sample (size $N = 1000000$).

properties of a given estimator $\hat{\tau}$ for τ will be evaluated based on the following criteria.

1. **Bias**. The empirical bias measures the average difference between the estimator obtained in one of the B samples and the true value of the estimator

$$\widehat{Bias}(\hat{\tau}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_b - \tau.$$

2. **Variance**. The empirical variance measures the precision in the estimation of τ as

$$\widehat{Var}(\hat{\tau}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\tau}_b - \bar{\hat{\tau}}_b)^2.$$

where $\bar{\hat{\tau}}_b$ is the average of the estimator over the B repetitions.

3. **Mean Squared Error**. The empirical mean squared error combines the empirical bias and variance. It is obtained as the sum of the square of the empirical bias and the empirical variance

$$\widehat{MSE}(\hat{\tau}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\tau}_b - \tau)^2.$$

4. **Coverage and length**. Confidence intervals for the estimators can be constructed based on the normal asymptotic approximations presented in the previous chapter. Confidence intervals are constructed as

$$IC(\alpha) = \left[\hat{\tau} - \hat{\sigma} \times q_{\frac{\alpha}{2}} : \hat{\tau} + \hat{\sigma} \times q_{1-\frac{\alpha}{2}}\right].$$

where $\hat{\sigma}^2$ is an estimator for the variance of the estimator that we discussed in the previous chapters, while $q_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ quantile of the standard normal distribution. We fix $\alpha = 0.05$. Coverage is then defined as the proportion of times that τ is contained in the computed confidence interval.

$$\widehat{cov} = \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}(\tau \in IC_b).$$

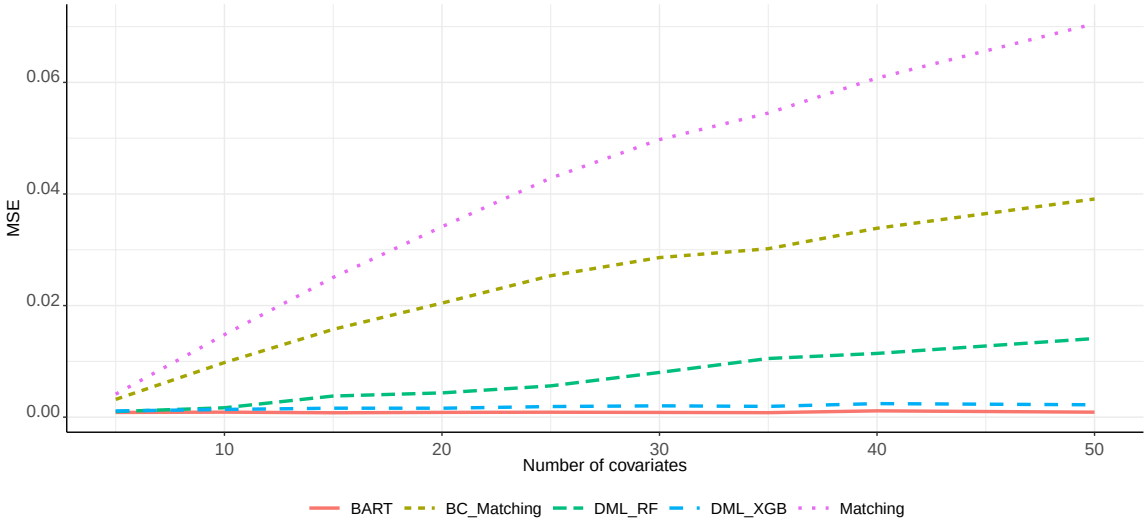
We will also compute the average length of the confidence intervals as

$$\frac{1}{B}\sum_{b=1}^{B}\big((\hat{\tau}_b - \hat{\sigma}_b \times q_{\frac{\alpha}{2}}) - (\hat{\tau}_b + \hat{\sigma}_b \times q_{1-\frac{\alpha}{2}})\big).$$

6.4 Results

We evaluated the performance of the estimators on $B = 200$ samples for each of the 48 scenarios described above. For each scenario, we computed the empirical bias, the empirical MSE, the empirical variance, the average length of confidence intervals and their coverage. The results are presented in tables 6.2 (design 1 without heterogeneity), 6.3 (design 1 with heterogeneity), 6.4 (design 2 without heterogeneity) and 6.5

Figure 6.1: Performances of the estimators in function of p



Note: For each level of p, MSE is computed using 200 repetitions of design 1 without heterogeneity, holding $\gamma = 1$ and $N = 5000$ fixed.

(design 2 with heterogeneity). From these tables, we can make the following interesting observations.

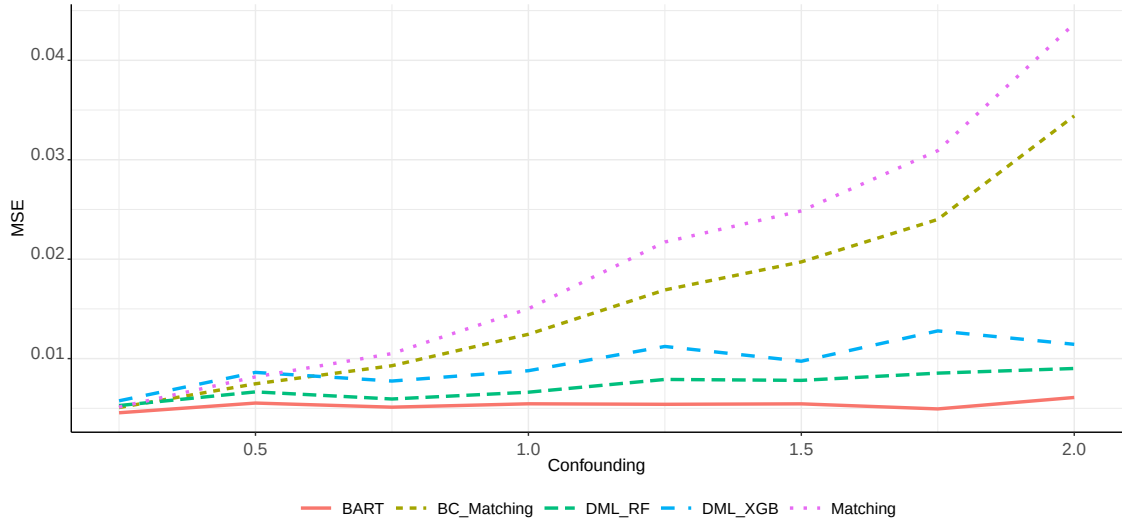
First, as expected, the OLS estimator is heavily biased in every scenario. This results in very high MSE and low coverage compared to the non-parametric approaches. In design 1, the OLS estimator seems to reduce part of the selection bias (the empirical bias is smaller than the selection bias even in the no heterogeneity case) but remains heavily biased. In the second design where the nuisance functions are highly non-linear, the OLS estimator does not even reduce the selection bias.

In terms of empirical MSE, BART tends to perform best in most of the scenarios. This is consistent with some previous results on comparable simulation approaches (Dorie et al., 2019). Note however that BART might benefit from assumptions that are satisfied in these specific simulation designs such as the normal specification for the errors (see Hill, 2011). At the same time, the DML estimation approaches (DMLRF, DMLXG) largely outperform the traditional non-parametric Nearest Neighbor (matching) approaches in all scenarios.

The performance gains of Machine Learning approaches over the more traditional non-parametric approaches are particularly large when the number of potential confounders increases. When the number of confounders increases, the MSE of ML approaches is systematically smaller than that of the traditional non-parametric approaches. The DML estimator based on boosting seems more robust to the presence of many potential confounders than the DML estimator based on Random Forest. This is consistent with the theoretical results underlying each of these ML algorithms (Yang et al., 2020). This fact is illustrated in a more detailed way in figure 6.1 where the MSE is estimated for all estimators by letting the number of covariates vary in a more systematic way.

Moreover, the ML based approaches tend to be less sensitive to the intensity of

Figure 6.2: Performances of the estimators as a function of γ



Note: For each level of γ, the MSE is computed using 200 repetitions of design 1 without heterogeneity, holding $p = 5$ and $N = 1000$ fixed.

confounding in settings with a small number of potential confounders. Indeed, the MSE is less affected by γ for ML-based approaches than for tradtional non-parametric approaches in this context as can be seen in figure 6.2.

While the OLS estimator is largely affected by the inclusion of heterogeneity in the effect of the treatement, this is not the case for the non-parametric approaches except in design 2 when there are a lot of potential confounders.

In terms of inference, when the bias is sufficiently small, all methods tend to have coverage close to the 95% target with comparable average length (at least for large sample sizes). Note however that BART intervals tend to have coverage above the 95% target. Note also that with smaller sample sizes, the boosting-based DML approach tends to produce less precise estimates.

To summarize, ML-based approaches appear to offer an attractive alternative to more traditional estimation approaches to estimate causal effects. These improvements associated to the ML-based approaches are dependent on the generating process. It is also interesting to note that there is some heterogeneity in the performance of the DML-based approaches.

Table 6.2: Design 1 without heterogeneity

			N = 500					N=1000					N=5000				
			Bias	MSE	sd	length	cov	bias	MSE	sd	length	cov	bias	MSE	sd	length	cov
$p = 5$	$\gamma = 0.5$	LR	.16	.04	.11	.43	.69	.16	.03	.08	.31	.51	.16	.03	.03	.14	.02
		DML_RF	.02	.01	.11	.45	.95	.01	.01	.08	.30	.93	.01	.00	.03	.13	.94
		DML_XG	.00	.02	.15	.59	.95	.02	.01	.09	.34	.94	.01	.00	.03	.12	.90
		Matching	.05	.01	.11	.41	.91	.05	.01	.07	.29	.89	.03	.00	.03	.13	.83
		BCMatching	.05	.01	.10	.41	.93	.04	.01	.07	.29	.91	.03	.00	.03	.13	.87
		BART	.01	.01	.11	.44	1	.01	.01	.08	.31	.96	.00	.00	.04	.14	.98
	$\gamma = 1$	LR	.31	.12	.16	.61	.50	.32	.11	.11	.43	.18	.32	.10	.05	.19	.00
		DML_RF	.03	.01	.13	.52	.97	.02	.01	.09	.34	.98	.01	.00	.03	.13	.98
		DML_XG	.01	.02	.16	.64	.97	.02	.01	.09	.36	.96	.01	.00	.03	.12	.94
		Matching	.10	.02	.13	.50	.92	.09	.02	.09	.34	.84	.06	.00	.04	.14	.67
		BCMatching	.09	.02	.12	.48	.92	.07	.01	.08	.33	.87	.05	.00	.04	.14	.73
		BART	.02	.01	.12	.46	.96	.01	.01	.08	.32	.98	.00	.00	.04	.14	.97
$p = 50$	$\gamma = 0.5$	LR	.17	.04	.12	.45	.72	.17	.03	.08	.31	.43	.16	.03	.03	.14	.02
		DML_RF	.13	.03	.11	.44	.79	.10	.02	.08	.29	.72	.06	.00	.03	.12	.50
		DML_XG	.05	.04	.18	.72	.95	.04	.01	.11	.42	.97	.03	.00	.03	.13	.87
		Matching	.16	.05	.14	.55	.78	.16	.04	.10	.39	.60	.13	.02	.05	.18	.19
		BCMatching	.12	.03	.14	.56	.84	.12	.02	.10	.38	.75	.10	.01	.04	.17	.39
		BART	.02	.01	.12	.46	.97	.02	.01	.08	.32	.96	.01	.00	.04	.14	.99
	$\gamma = 1$	LR	.30	.11	.16	.64	.56	.32	.11	.11	.44	.19	.32	.10	.05	.19	.00
		DML_RF	.21	.06	.15	.58	.71	.19	.04	.10	.38	.46	.11	.01	.04	.15	.09
		DML_XG	.06	.05	.22	.85	.95	.07	.02	.11	.45	.91	.04	.00	.03	.13	.82
		Matching	.27	.12	.21	.80	.72	.30	.11	.14	.57	.41	.26	.07	.06	.25	.01
		BCMatching	.21	.08	.19	.75	.81	.23	.06	.13	.50	.60	.19	.04	.06	.22	.07
		BART	.01	.01	.12	.48	.97	.03	.01	.09	.33	.99	.01	.00	.04	.14	.98

Table 6.3: Design 1 with heterogeneity

			N = 500					N=1000					N=5000				
			Bias	MSE	sd	length	cov	bias	MSE	sd	length	cov	bias	MSE	sd	length	cov
$p = 5$	$\gamma = 0.5$	LR	.24	.07	.13	.51	.55	.22	.06	.09	.36	.38	.22	.05	.04	.16	.01
		DML_RF	.03	.01	.12	.48	.97	.01	.01	.08	.32	.96	.01	.00	.03	.13	.94
		DML_XG	.02	.03	.16	.63	.94	.01	.01	.09	.35	.95	.02	.00	.03	.13	.91
		Matching	.04	.01	.12	.47	.95	.03	.01	.08	.32	.97	.02	.00	.04	.14	.89
		BCMatching	.03	.01	.12	.46	.96	.02	.01	.08	.32	.98	.02	.00	.04	.14	.91
		BART	-.01	.01	.12	.46	.99	-.01	.00	.08	.32	.98	.00	.00	.04	.14	.97
	$\gamma = 1$	LR	.40	.19	.18	.71	.41	.37	.16	.13	.50	.16	.38	.15	.06	.22	.00
		DML_RF	.04	.02	.14	.56	.95	.03	.01	.09	.36	.96	.01	.00	.04	.14	.95
		DML_XG	.03	.06	.17	.67	.91	.02	.01	.10	.39	.95	.02	.00	.03	.13	.89
		Matching	.10	.03	.14	.56	.89	.07	.01	.10	.38	.92	.05	.00	.04	.16	.77
		BCMatching	.08	.03	.14	.54	.90	.06	.01	.09	.37	.95	.04	.00	.04	.15	.83
		BART	-.01	.01	.12	.47	.98	.00	.01	.08	.33	.96	.00	.00	.04	.14	.97
$p = 50$	$\gamma = 0.5$	LR	.24	.08	.14	.54	.62	.23	.06	.09	.37	.33	.22	.05	.04	.16	.00
		DML_RF	.18	.04	.13	.50	.72	.14	.03	.08	.33	.62	.08	.01	.03	.13	.34
		DML_XG	.04	.05	.20	.78	.94	.04	.01	.11	.44	.95	.03	.00	.03	.13	.85
		Matching	.16	.05	.16	.64	.85	.16	.04	.11	.45	.69	.12	.02	.05	.20	.35
		BCMatching	.13	.04	.16	.65	.86	.12	.03	.11	.43	.82	.08	.01	.05	.19	.65
		BART	.00	.01	.12	.48	.97	.00	.01	.08	.33	.97	.00	.00	.04	.14	.98
	$\gamma = 1$	LR	.36	.16	.19	.73	.52	.38	.16	.13	.51	.16	.38	.15	.06	.22	.00
		DML_RF	.25	.09	.17	.66	.71	.23	.06	.11	.43	.41	.14	.02	.04	.16	.06
		DML_XG	.07	.05	.22	.85	.94	.07	.02	.13	.49	.94	.04	.00	.04	.14	.75
		Matching	.27	.13	.23	.90	.77	.30	.11	.16	.64	.53	.25	.07	.07	.28	.07
		BCMatching	.22	.09	.22	.86	.85	.22	.06	.15	.57	.72	.17	.03	.06	.25	.19
		BART	-.02	.01	.13	.51	.97	.01	.00	.09	.35	.99	.00	.00	.04	.15	.99

Table 6.4: Design 2 without heterogeneity

			N = 500					N=1000					N=5000				
			Bias	MSE	sd	length	cov	bias	MSE	sd	length	cov	bias	MSE	sd	length	cov
$p = 5$	$\gamma = 0.5$	LR	.20	.05	.09	.37	.44	.20	.05	.07	.26	.18	.20	.04	.03	.12	.00
		DML_RF	.00	.01	.12	.46	.97	.00	.01	.08	.31	.93	.00	.00	.03	.13	.98
		DML_XG	.07	.03	.14	.55	.94	.05	.01	.08	.32	.92	.02	.00	.03	.12	.91
		Matching	.04	.01	.10	.40	.95	.03	.01	.07	.28	.92	.02	.00	.03	.12	.90
		BCMatching	.04	.01	.10	.40	.95	.03	.01	.07	.28	.92	.02	.00	.03	.12	.90
		BART	.02	.01	.12	.46	.99	.02	.01	.08	.32	.97	.01	.00	.04	.14	.99
	$\gamma = 1$	LR	.40	.17	.11	.42	.03	.40	.17	.07	.29	.00	.41	.17	.03	.13	.00
		DML_RF	.02	.01	.12	.48	.96	.01	.01	.08	.32	.93	.01	.00	.03	.13	.99
		DML_XG	.06	.02	.15	.58	.91	.05	.01	.09	.34	.91	.03	.00	.03	.13	.81
		Matching	.09	.02	.10	.41	.87	.07	.01	.07	.28	.85	.05	.00	.03	.13	.73
		BCMatching	.09	.02	.10	.41	.88	.08	.01	.07	.29	.84	.05	.00	.03	.13	.74
		BART	.05	.01	.12	.47	.96	.03	.01	.08	.33	.95	.01	.00	.04	.14	.99
$p = 50$	$\gamma = 0.5$	LR	.22	.06	.10	.39	.37	.22	.05	.07	.27	.11	.22	.05	.03	.12	.00
		DML_RF	.19	.05	.10	.38	.47	.18	.04	.07	.26	.27	.14	.02	.03	.12	.01
		DML_XG	.26	.10	.18	.69	.68	.19	.05	.10	.41	.57	.13	.02	.03	.12	.04
		Matching	.20	.05	.11	.45	.63	.18	.04	.08	.32	.41	.17	.03	.04	.15	.01
		BCMatching	.20	.05	.13	.49	.66	.18	.04	.09	.33	.42	.17	.03	.04	.15	.00
		BART	.13	.03	.12	.46	.85	.14	.02	.08	.32	.63	.10	.01	.04	.15	.23
	$\gamma = 1$	LR	.44	.21	.11	.44	.05	.43	.19	.08	.30	.00	.43	.19	.03	.13	.00
		DML_RF	.37	.15	.11	.43	.07	.34	.12	.07	.29	.02	.27	.07	.03	.13	.00
		DML_XG	.39	.19	.19	.74	.41	.28	.09	.11	.44	.24	.22	.05	.03	.13	.00
		Matching	.37	.16	.13	.50	.19	.36	.14	.09	.36	.03	.34	.12	.04	.16	.00
		BCMatching	.40	.18	.14	.55	.18	.37	.14	.09	.37	.03	.34	.12	.04	.16	.00
		BART	.28	.09	.13	.51	.41	.26	.07	.09	.36	.17	.17	.03	.04	.16	.00

Table 6.5: Design 2 with heterogeneity

			N = 500					N=1000					N=5000				
			Bias	MSE	sd	length	cov	bias	MSE	sd	length	cov	bias	MSE	sd	length	cov
$p = 5$	$\gamma = 0.5$	LR	.29	.09	.10	.39	.22	.29	.09	.07	.27	.02	.30	.09	.03	.12	.00
		DML_RF	.00	.01	.12	.47	.96	.01	.01	.08	.32	.95	.00	.00	.03	.13	.98
		DML_XG	.06	.02	.15	.57	.96	.04	.01	.09	.34	.90	.03	.00	.03	.13	.88
		Matching	.07	.02	.10	.41	.91	.06	.01	.07	.29	.87	.04	.00	.03	.13	.82
		BCMatching	.08	.02	.10	.41	.92	.06	.01	.07	.29	.87	.04	.00	.03	.13	.82
		BART	.05	.01	.12	.47	.97	.03	.01	.08	.33	.94	.01	.00	.04	.14	.99
	$\gamma = 1$	LR	.49	.25	.11	.44	.01	.50	.25	.08	.31	.00	.50	.25	.04	.14	.00
		DML_RF	.02	.01	.13	.50	.98	.01	.01	.08	.33	.93	.01	.00	.03	.13	.96
		DML_XG	.04	.03	.16	.62	.93	.05	.01	.09	.35	.90	.04	.00	.03	.13	.81
		Matching	.13	.03	.11	.42	.80	.10	.02	.07	.29	.73	.06	.00	.03	.13	.53
		BCMatching	.13	.03	.11	.42	.80	.10	.02	.07	.29	.71	.06	.00	.03	.13	.53
		BART	.07	.01	.12	.48	.96	.04	.01	.09	.34	.94	.02	.00	.04	.15	.98
$p = 50$	$\gamma = 0.5$	LR	.32	.12	.10	.41	.11	.32	.11	.07	.28	.01	.32	.10	.03	.12	.00
		DML_RF	.28	.09	.10	.40	.20	.26	.07	.07	.28	.05	.20	.04	.03	.12	.00
		DML_XG	.32	.13	.17	.66	.48	.25	.07	.11	.42	.29	.18	.03	.03	.13	.00
		Matching	.31	.11	.12	.47	.26	.29	.09	.09	.34	.08	.28	.08	.04	.16	.00
		BCMatching	.31	.11	.13	.52	.35	.29	.09	.09	.35	.04	.28	.08	.04	.16	.00
		BART	.21	.06	.12	.48	.60	.21	.05	.09	.34	.31	.14	.02	.04	.15	.03
	$\gamma = 1$	LR	.54	.31	.12	.47	.01	.53	.29	.08	.32	.00	.54	.29	.04	.14	.00
		DML_RF	.46	.22	.12	.45	.03	.42	.18	.08	.31	.00	.33	.11	.03	.13	.00
		DML_XG	.45	.24	.20	.77	.34	.34	.13	.13	.51	.26	.27	.07	.04	.14	.00
		Matching	.49	.26	.14	.53	.05	.47	.23	.10	.38	.00	.45	.20	.04	.17	.00
		BCMatching	.51	.28	.15	.59	.06	.48	.24	.10	.40	.01	.45	.20	.04	.17	.00
		BART	.35	.14	.14	.54	.24	.32	.11	.10	.38	.06	.21	.04	.04	.16	.00

Chapter 7

Real data applications

A challenge when considering real data to illustrate causal inference methods under unconfoundedness is that unconfoundedness is an untestable assumption. In practice, justifying this assumption is an important part of empirical research and has to be motivated by expert domain knowledge. The objective of this section is not to offer a thorough justification of unconfoundedness for a new empirical application. Instead, we will illustrate the methods presented in this thesis on example datasets that are well established in the causal inference litterature on estimation of causal effects under unconfoundedness.

7.1 The effect of a retirement program on savings

As a first application, we reconsider an analysis of a retirement program initially conducted by Poterba et al. (1995) and reevaluated among others by Abadie (2003) and Chernozhukov et al. (2018). The objective of Poterba et al. (1995) is to study the effect of the 401(k) retirement plan introduced in 1978 in the US on subsequent saving behavior.

The 401(k) retirement plan is a fiscally favorable pension plan offered at the employer level. The objective of this program is to encourage people to save for their retirement. An important question is whether the 401(k) plan effectively increases retirement savings or whether it simply reallocates savings that would otherwise occur through other mechanisms. The outcome Y of interest in this case is total net savings and the treatment W is the 401(k) retirement plan. Each individual has a potential outcome $Y(0)$ if it does not participate in the retirement plan and a potential outcome $Y(1)$ under participation.

In this context, the ideal situation for a researcher would be one in which individuals are randomly assigned to participate in the 401(k) program. The researcher could then compare average savings after implementation of the program between participants and non-participants. In particular, following the arguments given in the first chapter, the difference in average savings in the two groups would yield an unbiased estimator of the ATE of the program.

Unfortunately, this is not how the 401(k) retirement savings program was imple-

Table 7.1: Description of a subset of the variables contained in the data

Variable name	Description
nettfa	Net financial assets (in $)
a401	savings in 401(k) program (in $)
e401	employer offers 401(k) program
age	age of the unit
income	household income (in $)
marr	marital status
fsize	family size
educ	years of education
p401	participation into 401(k)

mented. Hence, the only available option for researchers is to rely on observational data. In this context, Poterba et al. (1995) used data from the Survey on Income and Program Participation (SIPP). The SIPP data contains detailed data on income, employment, saving patterns and government program participation collected on a yearly basis for a representative sample of American households. While the initial study considered the waves of 1984, 1987 and 1991, we will restrict ourselves to the 1991 wave studied in Abadie (2003) and Chernozhukov et al. (2018). The considered sample consists of 9915 households reference persons and spouses aged between 25 and 64. A subset of the available variables are described in table 7.1. The main outcome variable Y will be net financial assets and the treatment W is participation in the 401(k) program. Descriptive statistics for a subset of the variables can be found in table 7.2.

A naive idea would be to estimate the average treatment effect of the program as the difference in net total financial assets between those who participate in the program and those who do not. This would yield an estimate of 27372 (see table 7.2). However, since participation to the treatment is voluntary, participants self select into the treatment and nothing guarantees that both groups are comparable. For example, from table 7.2 it appears that participants have higher incomes and a higher probability to be married than non-participants. This raises the concern that participants may have a different saving behavior than non-participants in the absence of treatment. In particular, it is reasonable to believe that participants may have stronger preferences for saving and would have saved more than non-participants even without the program. If that is the case, the potential outcome in the absence of treatment is larger for treated individuals than for untreated individuals. Then, as we saw in chapter 2, the ATE can not be identified by the difference in expectations between the treated and untreated units. This implies that the naive estimation approach is subject to a selection bias.

A possible approach to recover identification of the ATE in such a situation is to introduce identifying assumptions such as unconfoundedness. One could argue that conditional on a vector of observable covariates, treatment is independent from potential outcomes. Then, one can use the estimation approaches discussed in this thesis to obtain an unbiased estimate of the ATE of 401(k) participation. Estimates and associated standard errors obtained from the estimators discussed in chapters 3, 4 and 5 are given in panel A of table 7.3. The specification considered for the OLS estimator follows Abadie (2003) and includes income, age, marital status and family size. The

Table 7.2: Descriptive statistics for the 401(k) data

	Sample	Participation status		Eligibility to 401(k)	
		Treated	Non Treated	Eligible	Non eligible
Treatment 1					
401(k) participation	0.26			0.7	0
	(0.44)			(0.46)	(0)
Treatment 2					
401(k) eligibility	0.37	1	0.15		
	(0.48)	(0)	(0.36)		
Outcome					
Total net financial assets	18052	38262	10890	30347	10788
	(63522)	(79088)	(55257)	(74800)	(54518)
Selected covariates					
Household income	37200	49366	32890	46862	31494
	(24774)	(27208)	(22315)	(25958)	(22151)
Age	41	42	41	41	41
	(10)	(10)	(11)	(10)	(11)
Marital Status	0.6	0.7	0.57	0.67	0.56
	(0.49)	(0.46)	(0.49)	(0.47)	(0.5)
Education	13.2	13.9	13	13.8	12.9
	(2.8)	(2.6)	(2.8)	(2.6)	(2.9)

Notes: The statistics are different than those from Abadie (2003) because they remove very low and very high earners from the sample.

estimators for which data adaptive methods are used for estimation of the nuisance functions include a larger set of covariates chosen based on Chernozhukov et al. (2018). The hyperparameters of the boosting algorithm are chosen via cross validation. The sample splitting procedure for the DML estimators is repeated 20 times to stabilize the estimates as suggested by Chernozhukov et al. (2018). The number of considered matches for matching estimators is set to 2.

First, note that these estimates are smaller than the naive estimate based on the simple difference in means between groups. This suggests that the naive estimator is indeed subject to a selection bias in the present case.
Second, the estimate based on OLS is close to the estimates obtained from the more flexible estimation methods presented in this thesis. This suggests that for these data, there is little gain from using more flexible methods. Note also that matching estimators tend to have larger standard errors than the DML estimators.

Unfortunately, the unconfoundedness assumption is untestable. Hence, while selection bias is certainly attenuated in the estimates presented in panel A of table 7.3 it remains unclear whether this bias has been totally eliminated.

To increase the credibility of unconfoundedness in this setting, Poterba et al. (1995)

Table 7.3: Estimates and standard errors

	OLS	Matching	BC matching	DML RF	DML Boost	BART
A. Participation						
	12430	12093	10662	11275	11660	12707
	(1366)	(1603)	(1597)	(1226)	(1180)	(1322)
B. Eligibility						
	5236	8483	6973	8039	8254	8718
	(1253)	(1394)	(1392)	(1199)	(1149)	(1282)

suggested to redefine the treatment as *beeing eligible to the 401(k) program (treatment 2)* rather than effective participation. Since, only employees from firms offering the 401(k) program are eligible to the plan, eligibility is determined by the employer. If employee decisions to join a firm can be considered to be independent of individual saving behavior, the comparison of eligible and non-eligible groups could be used to infer the effect of eligibility. There are reasons to believe that this may not be true. For example, it is likely that income determines an employee's choice of employer as well as its saving behavior. At the same time, employers offering higher wages may be more inclined to offer saving plans. However, Poterba et al. (1995) argue that conditional on a few covariates (mainly income), the decision to join a firm is independent of savings behaviour. They provide evidence for this by showing that conditional on income, saving behaviour of eligible and non-eligible employees were equivalent at the beginning of the program. Note that if participation has a positive effect on savings, the ATE of eligibility is expected to be smaller than the effect of participation because part of the eligible subjects do not participate (about 15%) as can be seen from table 7.1.

The estimate obtained from the naive estimator is now reduced to 19559. Estimates and standard errors obtained from the estimation procedures described in this thesis are given in panel B of table 7.3. The different estimations are performed using the same specifications as for the evaluation of the effect of participation.

The estimates are broadly similar but much smaller than the naive estimates. This suggests that considering eligibility rather than participation is insufficient on its own to eliminate selection bias. The simple OLS estimator yields smaller results than the more flexible estimators. While we cannot compare the obtained results with those of the initial paper because they compute treatment effects within income brackets, the obtained results are in the same range as those obtained by Poterba et al. (1995).

7.2 Effect of a job training program

A drawback of the previous analysis was that it was not possible to assess the performance of the different methods because the true ATE was unknown. This is a general limitation of causal inference with observational data stemming from the fact that identifying assumptions (in this case unconfoundedness) are essentially untestable. In this context, LaLonde (1986) suggested to compare results from observational studies

with those obtained from Randomized experiments where estimates are known to be unbiased.

The National Supported Work demonstration (NSW) program offered unemployed a paid (subsidized) job with the objective of helping them moving to employment. Only disadvantaged unemployed (ex-drug addicts, ex-criminal offenders, high school dropouts, ...) were considered for participation into the program. An important feature of the program is that eligible applicants were randomly assigned to receive the program. In this context, an unbiased estimate of the Average Treatment Effect of the program on job related outcomes can be obtained by contrasting outcomes from treated and untreated applicants. The outcome considered by LaLonde (1986) is earnings at the end of the program (in 1978). The data that we will use in this thesis are a subsample of male applicants to the NSW program that was made publicly available by Dehejia and Wahba (1999). Descriptive statistics and main variables contained in the dataset are described in table 7.4. Note that the treated and untreated groups are very similar in terms of observed characteristics which was expected because they were randomized to receive the program. The ATE estimated as the difference in earnings at the end of the program between treated and untreated units is 886$.

To evaluate the performances of program evaluation methods under unconfoundedness, LaLonde (1986) proposed to recreate the conditions of an observational study. This observational data can then be used to benchmark the results obtained in an observational context with those from the randomized experiment. To do so, he gathered a random sample of the population from the Panel Study on Income Dynamics (PSID). This is a sample of untreated units whose earnings in 1978 could be compared to those of the treated units in the NSW experiment. Descriptive statistics on the PSID control group can be found in table 7.4. In this setting, a naive estimate of the treatment effect obtained as the difference in average earnings between the treated and PSID group is -15577$.

The PSID control group is very different than the treated group of the NSW experiment. For example, the individuals in the PSID sample are much younger, less likely to be married and more likely to be high school drop outs. This raises concern on the comparability of the treated and control groups. Indeed, treated individuals were specifically selected into treatment because they were disadvantaged workers. Hence, we would expect their potential outcome in the absence of treatment to be smaller than that of treated units. This suggests that the simple difference in means estimate is subject to a selection bias.

To recover identification of the ATE, researchers can introduce identifying assumptions such as unconfoundedness. We could argue that conditional on a vector of covariates, treatment is independent of potential outcomes. Then, we can use the methods discussed in this thesis to obtain unbiased estimates of the ATE. Estimates and associated standard errors obtained by the methods discussed above are presented in panel A of table 7.5. The results are rather disappointing. Indeed, while the obtained estimates are slightly closer to the true effect of the program compared to the naive estimate, they are still very far from the target obtained from the randomized experiment. Note that the bias corrected matching and BART estimators are closer (but still very far) from the target. The estimated standard errors of the DML estimators are smaller

Table 7.4: Descriptive statistics for Lalonde data

	NSW experiment		PSID controls	
	Treated	Non Treated	Full sample	Unemployed
Outcome				
Earnings in 1978	5976	5090	21554	9995
	(6924)	(5718)	(15555)	(11184)
Selected covariates				
Age	24.6	24.4	34.9	36
	(6.7)	(6.6)	(10)	(12)
Education	10.3	10.2	12.1	10.7
	(1.81)	(1.61)	(3.08)	(3.17)
Marital Status	0.17	0.16	0.87	0.73
	(0.37)	(0.36)	(0.34)	(0.44)
Black	0.8	0.8	0.25	0.39
	(0.4)	(0.4)	(0.43)	(0.48)
Hispanic	0.09	0.11	0.03	0.068
	(0.29)	(0.32)	(0.17)	(0.25)
Earnings 75'	3066	3026	19063	7569
	(4874)	(5201)	(13596)	(9041)
High school dropout	0.73	0.81	0.3	0.49
	(0.44)	(0.39)	(0.46)	(0.5)
N	297	425	2490	253

than those of the other estimators.

An important concern with the above results relates to the overlap assumption. Indeed, the two samples are very different and most observations from the PSID control group would not meet the admission criteria such that their propensity score should be 0. A detailed comparison of (estimated) propensity scores between treated and untreated units can for example be found in Dehejia and Wahba (1999). To adress this issue, LaLonde (1986) proposed to restrict the control sample to a subsample more representative of units meeting the inclusion criteria. One such subsample considers only the unemployed from the PSID control group. Descriptive statistics for this restricted sample can be found in the last column of table 7.4. Estimates and associated standard errors obtained by the methods discussed above are presented in panel A of table 7.5. The obtained estimates are much closer to the target but still relatively far.

In a later reassessment, Dehejia and Wahba (1999) proposed a few modifications to the initial study. First, they proposed to refine the experimental sample to those units for which information on the earnings in 1974 was available. This was aimed at in-

Table 7.5: Estimates and standard errors

	OLS	Matching	BC matching	DML RF	DML Boost	BART
A. Full sample						
	-11960	-12394	-6797	-10597	-10931	-6923
	(2426)	(2987)	(2530)	(798)	(1030)	(2394)
B. Unemployed						
	-1496	-1469	-686	-946	-797	-859
	(988)	(1135)	(1103)	(969)	(1615)	(1107)
B. ATT						
	161	525	1218	1188	1483	320
	(869)	(1173)	(1189)	(893)	(1179)	(1946)

Notes: The OLS estimate differ from the original results from LaLonde (1986). This is because we allow for heterogeneity in the effect of the treatment while Lalonde assumed a partially linear model. We made this choice for two reasons. First, because assuming no heterogeneity in the effect of the treatment seems to be a strong assumption in this context. Second, to ensure comparability with other estimators who account for heterogeneity. The parameters for the data-adaptive methods are chosen in a similar way as in the previous case study.

creasing the credibility of the unconfoundedness assumption. On this new experimental sample, the estimated ATE that serves as the new target is 1794$.

Second, they considered the average treatment effect on the treated (ATT) rather than the ATE. While in the experimental data, the ATT and the ATE should be equivalent, this should not necessary be the case for the observational data. In the context of a job training program targeted to a specific group, it makes little sense to study the impact of the program for the entire population. Furthermore, considering the ATT solves part of the overlap problem because the treatment effect is evaluated in the support of treated units only. Hence, we need a weaker version of the overlap assumption requiring only that the propensity score is bounded away from 1. Details on why this result holds can be found in Appendix A.3. Modifying the estimators presented in chapters 3, 4 and 5 to estimate the ATT is straightforward. For example, for nearest neighbors estimators, one would only consider nearest neighbors for treated units (among the untreated units). For the DML estimators, one would consider the efficient influence function for the ATT and construct the estimator accordingly. Estimates and associated standard errors obtained with the methods presented above tailored to the estimation of the ATT are presented in table 7.5. The newly obtained estimators are much closer to the target in this situation confirming the results obtained by Dehejia and Wahba (1999). Note that the bias corrected matching estimator and the Debiased ML estimators seem to perform best in this case.

Chapter 8

Conclusion

In chapter 2 of this thesis, the potential outcomes framework was presented. This framework allowed us to provide a clear definition of causal estimands and to clarify the assumptions that are needed to identify these estimands. We presented 3 ways in which popular estimands could be identified under unconfoundedness and overlap and saw that these identification results rely on the determination of so-called nuisance functions. This suggested empirical estimators where the nuisance functions have to be estimated in a first step.

In practice, researchers typically rely on parametric models to estimate these nuisance functions. In chapter 3, we saw that relying on such models can have negative consequences on the properties of estimators for causal estimands when the postulated models are incorrect. To mitigate the problem, researchers have considered traditional non-parametric approaches such as Nearest Neighbor or kernel regressions to estimate the nuisance functions. Under certain conditions, the resulting estimators for the causal estimands can be shown to have desirable asymptotic properties. In chapter 4, we discussed in detail an estimator based on Nearest Neighbor regression of the nuisance function. While this estimator relaxes the modeling assumptions, it requires strong assumptions in terms of the complexity of the nuisance function.

In this context, researchers have considered the use of Machine Learning algorithms to estimate the nuisance functions. However, understanding the asymptotic properties of estimators for causal estimands when ML is used in a first step is hard. Indeed, these ML algorithms have been designed for prediction rather than for precise point estimation. In chapter 5, we presented a general theory that facilitates the understanding of the asymptotic properties of these ML based estimators. We saw that the tradtional approach of building an estimator by plugging in the estimated nuisance functions does not lead to an estimator with desirable asymptotic properties. But we saw how such estimators can be corrected to build so-called Debiased Machine Learning (DML) estimators with desirable asymptotic properties when ML methods are used to estimate the nuisance functions. We then applied this theory to the construction of an estimator for the ATE.

In chapter 6, we studied the final sample properties of the estimators by performing a Monte Carlo simulation study. The data generating processes were designed to create a

challenging environment for model-based approaches. We studied various adaptations to the baseline designs by varying the number of covariates, the intensity of confounding, the sample size and the intensity of heterogeneity in the effect of the treatment. Unsurprisingly, the modeling approaches were heavily biased. While all non-parametric estimators performed better than their simple parametric counterparts, the DML estimator was better than the traditional non-parametric approaches, especially when the number of considered covariates increases.

Finally, in chapter 7, we applied the methods discussed in chapters 3, 4 and 5 to real data. First, we considered the evaluation of the effect of a retirement saving program on net savings. The different estimators produced comparable estimates suggesting that for that dataset, simple modeling approaches were sufficient to obtain the ATE. Second, we reconsidered the LaLonde experiment where the results obtained from an RCT could be used as a benchmark for the true causal effect. For estimation of the ATE, all methods performed poorly and there was no gain from using the more flexible approaches reinforcing the concern with the identifying assumptions. Following the litterature, we redefined the causal estimand and the studied sample of interest to soften the identifying assumptions. In this setting, the estimates produced by the DML estimators are closer to the target than their model-based counterparts.

In most of the thesis we considered the ATE under the regular assignment mechanism. However, most of the aspects covered in the thesis are valid for causal estimands at a more general level. In particular, using the theory developed in chapter 5, DML estimators can be constructed for most of the causal estimands discussed in the first chapter. Extensions to estimation of the ATT and ATU are relatively straightforward based on chapter 5. Derivation of the efficient influence function and resulting DML estimator for the instrumental variable based Local Average Treatment Effect can be found in Kennedy (2022) or Chernozhukov et al. (2018).

In chapter 6, we observed that estimates can vary according to the method used and the data generating process. An important question in this context relates to which estimator should be used on which dataset. In this context, an interesting approach to validate causal inference methods is proposed in Parikh et al. (2022) and Athey et al. (2021). They propose to use deep generative algorithms to generate synthetic data that mimic the empirical distribution of the available sample. The analyst can then specify the form of causal effects and use a simulation experiment to assess the performance of different estimators on data that closely resembles the true data generating process.

Appendix A

Additional computations in the Introduction

A.1 Bias in the simple difference in expectations

$$\mathbb{E}[Y(1)|W=1] - \mathbb{E}[Y(0)|W=0] = \underbrace{\mathbb{E}[Y(1) - Y(0)]}_{\text{ATE}} + B.$$

Equivalently,

$$B = \mathbb{E}[Y(1)|W=1] - \mathbb{E}[Y(0)|W=0] - ATE.$$

Using the law of total expectation, we can rewrite the ATE as

$$\begin{aligned} ATE = {} & Pr(W=0)\mathbb{E}[Y(1)|W=0] + Pr(W=1)\mathbb{E}[Y(1)|W=1] \\ & - Pr(W=0)\mathbb{E}[Y(0)|W=0] - Pr(W=1)\mathbb{E}[Y(0)|W=1]. \end{aligned}$$

Hence, B can be rewritten as

$$\begin{aligned} B = {} & \mathbb{E}[Y(1)|W=1] - \mathbb{E}[Y(0)|W=0] \\ & - Pr(W=0)\mathbb{E}[Y(1)|W=0] - Pr(W=1)\mathbb{E}[Y(1)|W=1] \\ & + Pr(W=0)\mathbb{E}[Y(0)|W=0] + Pr(W=1)\mathbb{E}[Y(0)|W=1]. \end{aligned}$$

We then add and subtract $\mathbb{E}[Y(0)|W=1]$ and $\mathbb{E}[Y(0)|W=0]$ to obtain

$$\begin{aligned} B = {} & \mathbb{E}[Y(1)|W=1] - \mathbb{E}[Y(0)|W=0] \\ & - Pr(W=0)\mathbb{E}[Y(1)|W=0] - Pr(W=1)\mathbb{E}[Y(1)|W=1] \\ & + Pr(W=0)\mathbb{E}[Y(0)|W=0] + Pr(W=1)\mathbb{E}[Y(0)|W=1] \\ & + \mathbb{E}[Y(0)|W=1] - \mathbb{E}[Y(0)|W=1] + \mathbb{E}[Y(0)|W=0] - \mathbb{E}[Y(0)|W=0] \end{aligned}$$

which can be rearranged as

$$\begin{aligned}
B &= \mathbb{E}[Y(0)|W=1] - \mathbb{E}[Y(0)|W=0] \\
&\quad + (1 - Pr(W=1))\mathbb{E}[Y(1)|W=1] - (1 - Pr(W=1))\mathbb{E}[Y(0)|W=1] \\
&\quad - (1 - Pr(W=0) - 1)\mathbb{E}[Y(0)|W=0] - Pr(W=0)\mathbb{E}[Y(1)|W=0] \\
&= \mathbb{E}[Y(0)|W=1] - \mathbb{E}[Y(0)|W=0] \\
&\quad + Pr(W=0)\{\mathbb{E}[Y(1)|W=1] - \mathbb{E}[Y(0)|W=1]\} \\
&\quad + Pr(W=0)\{\mathbb{E}[Y(0)|W=0] - \mathbb{E}[Y(1)|W=0]\}.
\end{aligned}$$

The terms in brackets in the last two lines correspond to the ATT and ATU respectively. Hence, we recover the desired expression

$$\begin{aligned}
B &= \mathbb{E}[Y(0)|W=1] - \mathbb{E}[Y(0)|W=0] \\
&\quad + Pr(W=0)\{\underbrace{\big(\mathbb{E}[Y(1)|W=1] - \mathbb{E}[Y(0)|W=1]\big)}_{ATT} - \underbrace{\big(\mathbb{E}[Y(0)|W=0] - \mathbb{E}[Y(1)|W=0]\big)}_{ATU}\}.
\end{aligned}$$

A.2 Unconfoundedness under the propensity score - result 2.2

We would like to show that

$$W \perp\!\!\!\perp Y(0), Y(1) \big| \mathbf{X} \implies W \perp\!\!\!\perp Y(0), Y(1) \big| \pi(\mathbf{X}).$$

Proving this result is equivalent to proving that if unconfoundedness holds (left hand side term above), we have that

$$Pr(W=1|Y(1), Y(0), \pi(\mathbf{X})) = Pr(W=1|\pi(\mathbf{X})).$$

This follows from the fact that equality of the conditional and marginal distributions implies independence. Since W is a binary variable, its distribution is entirely determined by its expectation. Hence, equality of the conditional and unconditional expectation of W implies independence.

We will show that the equality holds by considering both sides separately.

Consider first the right hand side which can be rewritten as

$$Pr(W=1|\pi(\mathbf{X})) = \mathbb{E}\Big[\mathbb{E}[W|\mathbf{X}, \pi(\mathbf{X})]\Big|\pi(\mathbf{X})\Big] = \mathbb{E}\Big[\mathbb{E}[W|\mathbf{X}]\Big|\pi(\mathbf{X})\Big] = Pr(\pi(\mathbf{X})\big|\pi(\mathbf{X})) = \pi(\mathbf{X})$$

where the second equality is a direct application of the law of iterated expectations and the third equality holds because the propensity score is a function of $\mathbf{X}$.

Now, consider the left hand side

$$\begin{aligned}
Pr(W=1|Y(0), Y(1), \pi(\mathbf{X})) &= \mathbb{E}[W|Y(1), Y(0), \pi(\mathbf{X})] \\
&= \mathbb{E}\Big[[\mathbb{E}[W|Y(1), Y(0), \mathbf{X}, \pi(\mathbf{X})]\Big|Y(0), Y(1), \pi(\mathbf{X})]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}[W|\mathbf{X}, \pi(\mathbf{X})]\Big|Y(0), Y(1), \pi(\mathbf{X})\Big] \\
&= \mathbb{E}\Big[\mathbb{E}[W|\mathbf{X}]\Big|Y(0), Y(1), \pi(\mathbf{X})\Big] \\
&= \mathbb{E}[\pi(\mathbf{X})|\pi(\mathbf{X}), Y(0), Y(1)] = \pi(\mathbf{X}).
\end{aligned}$$

The second equality is a direct application of the general LIE. The third equality uses the assumption of the theorem (unconfoundedness). The fourth equality holds because the propensity score is a function of $\mathbf{X}$.

A.3 The importance of overlap for identification

In this section, we provide some precisions on the importance of overlap for the identification of the ATE. Remember that overlap states that

$$0 < \pi(\mathbf{X}) < 1$$

As we argued in chapter 1, under unconfoundedness and overlap, the ATE can be written as

$$\begin{aligned}\Psi(\mathbb{F}) &= \mathbb{E}\Big[\mathbb{E}\left[Y|\mathbf{X}, W=1\right] - \mathbb{E}[Y|\mathbf{X}, W=0]\Big] \\ &= \int\int yf(y|\mathbf{x}, w=1)f(\mathbf{x})\,dy\,d\mathbf{x} - \int\int yf(y|\mathbf{x}, w=0)f(\mathbf{x})\,dy\,d\mathbf{x} \\ &= \int\int y\frac{f(y, \mathbf{x}, w=1)f(\mathbf{x})}{f(\mathbf{x}, w=1)}\,dy\,d\mathbf{x} - \int\int y\frac{f(y, \mathbf{x}, w=0)f(\mathbf{x})}{f(\mathbf{x}, w=0)}\,dy\,d\mathbf{x}\end{aligned} \tag{A.1}$$

Note that,

$$\pi(\mathbf{X}) = Pr(W=1|\mathbf{X}) = \frac{f(\mathbf{x}, w=1)}{f(\mathbf{x})} = 0 \implies f(\mathbf{x}, w=1) = 0$$

Hence, the first term on the right of equation (A.1) is not defined for a null propensity score.
Similarly,

$$\pi(\mathbf{X}) = 1 \implies Pr(W=0|\mathbf{X}) = 0 \implies \frac{f(\mathbf{x}, w=0)}{f(\mathbf{x})} = 0 \implies f(\mathbf{x}, w=0) = 0$$

Hence, the second term on the right of equation (A.1) is not defined for a propensity score of 1.

Interestingly, the overlap assumption can be relaxed to the following assumption when considering identification of the ATT

$$\pi(\mathbf{X}) < 1.$$

This can be immediately seen by noting that one can write (under unconfoundedness and the weaker overlap assumption) the ATT as

$$\begin{aligned}\Psi(\mathbb{F}) &= \mathbb{E}_{\mathbf{X}, W=1}\Big[\mathbb{E}\left[Y|\mathbf{X}, W=1\right] - \mathbb{E}[Y|\mathbf{X}, W=0]\Big] \\ &= \int\int yf(y|\mathbf{x}, w=1)f(\mathbf{x}, w=1)\,dy\,d\mathbf{x} - \int\int yf(y|\mathbf{x}, w=0)f(\mathbf{x}, w=1)\,dy\,d\mathbf{x} \\ &= \int\int yf(y, \mathbf{x}, w=1)\,dy\,d\mathbf{x} - \int\int y\frac{f(y, \mathbf{x}, w=0)f(\mathbf{x}, w=1)}{f(\mathbf{x}, w=0)}\,dy\,d\mathbf{x}\end{aligned}$$

where the index $\mathbf{X}, W=1$ on the expectation refers to the fact that the expectation is taken over the support of $\mathbf{X}$ for treated units.

Appendix B

Properties of the OLS estimator for the ATE

B.1 Remark 1

1. It is useful to realize that under assumptions (2.2), (2.3) and (3.1), we can rewrite $g_0(\mathbf{X})$ and $g_1(\mathbf{X})$ in terms of linear models containing the expected potential outcomes as intercepts

$$g_w(\mathbf{X}) = \alpha_w + \mathbb{E}[\mathbf{X}]^t\boldsymbol{\beta}_w + (\mathbf{X} - \mathbb{E}[\mathbf{X}])^t\boldsymbol{\beta}_w = \mu_w + \dot{\mathbf{X}}^t\boldsymbol{\beta}_w \qquad w \in \{0,1\} \quad \text{(B.1)}$$

   where $\dot{\mathbf{X}} = (\mathbf{X} - \mu_{\mathbf{X}})$.

   The first equality is obtained by adding and subtracting $\mathbb{E}[\mathbf{X}]^t\boldsymbol{\beta}_w$. The second equality is obtained by realizing that

$$\mu_w = \mathbb{E}[Y(w)] = \mathbb{E}\Big[\mathbb{E}[Y(w)|\mathbf{X}]\Big] = \mathbb{E}[g_w(\mathbf{X})] = \alpha_w + \mathbb{E}[\mathbf{X}]^t\boldsymbol{\beta_w} \qquad w \in \{0,1\}$$

   where the first equality holds by iterated expectations.

2. Let $\tilde{\mu}_1$, $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\mu}_0$, $\tilde{\boldsymbol{\beta}}_0$ be the intercepts and coefficients in a regression of Y_i on $\begin{pmatrix}1 & \dot{\mathbf{X}}_i^t\end{pmatrix}$ for treated and untreated samples respectively. We want to show that $\tilde{\mu}_1 = \hat{\alpha}_1 + \mu_{\mathbf{X}}{}^t\hat{\boldsymbol{\beta}}_1$ and that $\tilde{\mu}_0 = \hat{\alpha}_0 + \mu_{\mathbf{X}}{}^t\hat{\boldsymbol{\beta}}_0$.

   First, note that $\tilde{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1$ because demeaning the covariates does not change the estimated slope coefficients. For ease of exposition, this can be seen from the well known OLS expressions for the slope coefficient in the simple linear regression context. Indeed,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N W_i(X_i - \bar{X}_1)\sum_{i=1}^N W_i(Y_i - \bar{Y}_1)}{\sum_{i=1}^N W_i(X_i - \bar{X}_1)^2}$$

   and

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^N W_i(X_i - \mathbb{E}[X] - \bar{X}_1 + \mathbb{E}[X])\sum_{i=1}^N W_i(Y_i - \bar{Y}_1)}{\sum_{i=1}^N W_i(X_i - \mathbb{E}[X] - \bar{X}_1 + \bar{X})^2}.$$

These two expressions are equivalent.

Now, note that the intercepts, can be rewritten as

$$\hat{\alpha}_1 = \bar{Y}_1 - \bar{\mathbf{X}}_1^t \hat{\beta}_1 \tag{B.2}$$

where $\bar{\mathbf{X}}_1 = \frac{1}{N_1}\sum_{i=1}^N W_i \mathbf{X}_i$ and $\bar{Y}_1 = \frac{1}{N_1}\sum_{i=1}^N W_i Y_i$ with $N_1 = \sum_{i=1}^N W_i$. Equivalently,

$$\tilde{\mu}_1 = \bar{Y}_1 - \bar{\dot{X}}_1^t \tilde{\beta}_1 \tag{B.3}$$

where $\bar{\dot{X}}_1 = \frac{1}{N_1}\sum_{i=1}^N W_i(\mathbf{X}_i - \mu_{\mathbf{X}}) = \frac{1}{N_1}\sum_{i=1}^N W_i \mathbf{X}_i - \mu_{\mathbf{X}} = \bar{\mathbf{X}}_1 - \mu_{\mathbf{X}}$.

Solving equation B.2 for $\bar{Y}_1$ and plugging it into equation B.3 gives

$$\begin{aligned} \tilde{\mu}_1 &= \hat{\alpha_1} + \bar{\mathbf{X}}_1^t \hat{\beta}_1 - \bar{\dot{X}}_1^t \tilde{\beta}_1 \\ &= \hat{\alpha_1} + \bar{\mathbf{X}}_1^t \hat{\beta}_1 - (\bar{\mathbf{X}}_1 - \mu_{\mathbf{X}})^t \tilde{\beta}_1 \\ &= \hat{\alpha_1} + {\mu_{\mathbf{X}}}^t \hat{\beta}_1 \end{aligned} \tag{B.4}$$

where we replaced $\tilde{\beta}_1$ by $\hat{\beta}_1$ in the last line because the two are equivalent.

A similar reasoning can be used to show that $\hat{\mu}_w = \hat{\alpha}_w + \bar{\mathbf{X}}^t \hat{\beta_w}$.

B.2 Proof of theorem 1

Proof. Asymptotic distribution of $\tilde{\tau}_{reg}$.

We consider the following models for $\mathbb{E}[Y|\mathbf{X}, W=0]$ and $\mathbb{E}[Y|\mathbf{X}, W=1]$

$$\begin{cases} \mathbb{E}[Y|\mathbf{X}, W=0] = \mu_0 + \dot{\mathbf{X}}^t \boldsymbol{\beta}_0 = \dot{\mathbf{R}}^t \boldsymbol{\delta}_0 \\ \mathbb{E}[Y|\mathbf{X}, W=1] = \mu_1 + \dot{\mathbf{X}}^t \boldsymbol{\beta}_1 = \dot{\mathbf{R}}^t \boldsymbol{\delta}_1 \end{cases} \tag{B.5}$$

where $\dot{\mathbf{R}}^t = \begin{pmatrix} 1 & \dot{\mathbf{X}}^t \end{pmatrix}$ and $\boldsymbol{\delta}_w^t = \begin{pmatrix} \mu_w & \boldsymbol{\beta}_w \end{pmatrix}$. Let $\tilde{\boldsymbol{\delta}}_0^t = \begin{pmatrix} \tilde{\mu}_0 & \tilde{\boldsymbol{\beta}}_0^t \end{pmatrix}$ and $\tilde{\boldsymbol{\delta}}_1^t = \begin{pmatrix} \tilde{\mu}_1 & \tilde{\boldsymbol{\beta}}_1^t \end{pmatrix}$ be the OLS estimators obtained from a regression of Y_i on $(1,\ \dot{\mathbf{X}}_i^t)$ on treated and untreated samples respectively. The proposed estimator for the ATE is

$$\tilde{\tau}_{reg} = \tilde{\mu}_1 - \tilde{\mu}_0$$

which is the first element of the random vector $\tilde{\boldsymbol{\delta}}_1 - \tilde{\boldsymbol{\delta}}_0$. We start by considering the asymptotic distribution of $\tilde{\boldsymbol{\delta}}_1$ and $\tilde{\boldsymbol{\delta}}_0$ separately before investigating the asymptotic distribution of their difference.

We rewrite the model for treated units as

$$Y = \dot{\mathbf{R}}^t \boldsymbol{\delta}_1 + \epsilon(1) \qquad \text{for} \qquad W = 1 \tag{B.6}$$

where $\mathbb{E}[W\epsilon(1)] = 0$ and $\mathbb{E}[W\epsilon(1)|\dot{\mathbf{R}}] = 0$. We also make the classical OLS assumption that $\operatorname{rank} \mathbb{E}[W\mathbf{R}^t\mathbf{R}] = p + 1$.

The OLS estimator for $\boldsymbol{\delta}_1$ can be written as

$$\begin{aligned}
\tilde{\boldsymbol{\delta}}_1 &= \left(\frac{1}{N}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i \dot{\mathbf{R}}_i^t\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i Y_i\right) \\
&= \left(\frac{1}{N}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i \dot{\mathbf{R}}_i^t\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i \left(\dot{\mathbf{R}}_i^t \boldsymbol{\delta}_1 + \epsilon_i(1)\right)\right) \\
&= \left(\frac{1}{N}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i \dot{\mathbf{R}}_i^t\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i \dot{\mathbf{R}}_i^t \boldsymbol{\delta}_1 + \frac{1}{N}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i^t \epsilon_i(1)\right) \\
&= \boldsymbol{\delta}_1 + \left(\frac{1}{N}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i \dot{\mathbf{R}}_i^t\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i \epsilon_i(1)\right).
\end{aligned} \tag{B.7}$$

This implies that

$$\sqrt{N}(\tilde{\boldsymbol{\delta}}_1 - \boldsymbol{\delta}_1) = \left(\frac{1}{N}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i \dot{\mathbf{R}}_i^t\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i \epsilon_i(1)\right).$$

By the Law of Large Numbers (LLN) and the Continuous Mapping Theorem (CMT),

$$\left(\frac{1}{N}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i \dot{\mathbf{R}}_i^t\right)^{-1} \xrightarrow{p} \left(\mathbb{E}\left[W_i \dot{\mathbf{R}}_i \dot{\mathbf{R}}_i^t\right]\right)^{-1} \equiv \mathbf{A}_1^{-1}.$$

By the CLT,

$$\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i \epsilon_i(1)\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[W_i \dot{\mathbf{R}}_i \dot{\mathbf{R}}_i^t \epsilon_i(1)^2\right]\right)$$

because

$$\begin{aligned}
\mathbb{E}\left[W_i \dot{\mathbf{R}}_i \epsilon_i(1)\right] &= \mathbb{E}\left[\mathbb{E}\left[W_i \dot{\mathbf{R}}_i \epsilon_i(1) | \dot{\mathbf{R}}_i\right]\right] \\
&= \mathbb{E}\left[\dot{\mathbf{R}}_i \underbrace{\mathbb{E}\left[W_i \epsilon_i(1) | \dot{\mathbf{R}}_i\right]}_{=0}\right] \\
&= 0
\end{aligned} \tag{B.8}$$

where the first line is an application of the law of iterated expectations.
And

$$Var\left[W_i \dot{\mathbf{R}}_i \epsilon_i(1)\right] = \mathbb{E}\left[W_i \dot{\mathbf{R}}_i \dot{\mathbf{R}}_i^t \epsilon_i(1)^2\right] - \underbrace{\mathbb{E}\left[W_i \dot{\mathbf{R}}_i \epsilon_i(1)\right]}_{=0} \underbrace{\mathbb{E}\left[W_i \dot{\mathbf{R}}_i \epsilon_i(1)\right]^t}_{=0} \tag{B.9}$$

Hence,

$$\sqrt{N}(\tilde{\boldsymbol{\delta}}_1 - \boldsymbol{\delta}_1) = \mathbf{Z}_1 + o_p(1)$$

where $\mathbf{Z}_1 = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{A}_1^{-1} W_i \dot{\mathbf{R}}_i \epsilon_i(1)$. And

$$\boldsymbol{Z_1} \sim \mathcal{N}\left(0, \mathbf{V_1}\right) \tag{B.10}$$

where $\mathbf{V_1} = \mathbf{A}_1^{-1}\mathbb{E}[W_i\dot{\mathbf{R}}_i\dot{\mathbf{R}}_i^t\epsilon_i(1)^2]\mathbf{A}_1^{-1}$.

The same argument can be used to show that

$$\sqrt{N}(\tilde{\boldsymbol{\delta}}_0 - \boldsymbol{\delta}_0) = \mathbf{Z}_0 + o_p(1) \tag{B.11}$$

where $\mathbf{Z}_0 = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{A_0}^{-1}W_i\mathbf{R}_i\epsilon_i(0)$, and

$$\mathbf{Z}_0 \sim \mathcal{N}(0, \mathbf{V}_0) \tag{B.12}$$

where $\mathbf{V}_0 = \mathbf{A}_0^{-1}\mathbb{E}[(1-W_i)\dot{\mathbf{R}}_i\dot{\mathbf{R}}_i^t\epsilon_i(0)^2]\mathbf{A}_0^{-1}$.

Finally, we are interested in the first element of

$$\sqrt{N}\left((\tilde{\boldsymbol{\delta}}_1 - \tilde{\boldsymbol{\delta}}_0) - (\boldsymbol{\delta}_1 - \boldsymbol{\delta}_0)\right) = \sqrt{N}(\tilde{\boldsymbol{\delta}}_1 - \boldsymbol{\delta}_1) - \sqrt{N}(\tilde{\boldsymbol{\delta}}_0 - \boldsymbol{\delta}_0) = \mathbf{Z}_1 - \mathbf{Z}_0 + o_p(1) \tag{B.13}$$

$\mathbf{Z}_1$ and $\mathbf{Z}_0$ are normally distributed random variables. Null covariance between the 2 would imply that their difference follows a normal distribution with variance being the sum of the variances. The covariance between the two is

$$\begin{aligned} Cov\left[\mathbf{Z}_1, \mathbf{Z}_0\right] &= Cov\left[\mathbf{A}^{-1}(1)W_i\dot{\mathbf{R}}_i\epsilon_i(1), \mathbf{A}_0^{-1}(1-W_i)\dot{\mathbf{R}}_i\epsilon_i(0)\right] \\ &= \underbrace{\mathbb{E}\left[\mathbf{A}_1^{-1}W_i\dot{\mathbf{R}}_i\epsilon_i(1)\mathbf{A}_0^{-1}(1-W_i)\dot{\mathbf{R}}_i\epsilon_i(0)\right]}_{=0} \\ &\quad - \underbrace{\mathbb{E}\left[\mathbf{A}_1^{-1}W_i\dot{\mathbf{R}}_i\epsilon_i(1)\right]}_{=0}\underbrace{\mathbb{E}\left[\mathbf{A}_0^{-1}(1-W_i)\dot{\mathbf{R}}_i\epsilon_i(0)\right]}_{=0} \\ &= 0 \end{aligned} \tag{B.14}$$

where the first equality holds because we have i.i.d data and the second because $W_i(1-W_i) = 0$. Hence,

$$\sqrt{N}\left((\tilde{\boldsymbol{\delta}}_1 - \tilde{\boldsymbol{\delta}}_0) - (\boldsymbol{\delta}_1 - \boldsymbol{\delta}_0)\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_0 + \mathbf{V}_1) \tag{B.15}$$

And the asymptotic variance of $\sqrt{N}(\tilde{\tau} - \tau)$ is the first element of the $\mathbf{V}_0 + \mathbf{V}_1$ variance covariance matrix. □

Proof. Asymptotic distribution of $\hat{\tau}_{reg}$.

We start by discussing the asymptotic distribution of $\hat{\mu}_1$. By least squares mechanics, $\hat{\mu}_1$ can be expressed in terms of $\tilde{\mu}_1$. Indeed, using a similar argument to that in remark 1

$$\hat{\mu}_1 = \tilde{\mu}_1 + (\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}])^t\hat{\boldsymbol{\beta}}_1 \tag{B.16}$$

Then,

$$\sqrt{N}(\hat{\mu}_1 - \mu_1) = \sqrt{N}(\tilde{\mu}_1 - \mu_1) + \sqrt{N}(\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}])^t\hat{\boldsymbol{\beta}}_1 \tag{B.17}$$

Following the same reasoning,

$$\sqrt{N}(\hat{\mu}_0 - \mu_0) = \sqrt{N}(\tilde{\mu}_0 - \mu_0) + \sqrt{N}(\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}])^t\hat{\boldsymbol{\beta}}_0 \tag{B.18}$$

which implies that

$$\sqrt{N}(\hat{\tau}_0 - \tau_0) = \sqrt{N}(\tilde{\mu}_1 - \mu_1) - \sqrt{N}(\tilde{\mu}_0 - \mu_0) + \sqrt{N}(\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}])^t(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0) \quad \text{(B.19)}$$

Note that

$$(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)\sqrt{N}(\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}]) = \sqrt{N}(\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}])^t(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + o_p(1) \quad \text{(B.20)}$$

because $(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0) \xrightarrow{p} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$. Then, by the CLT,

$$\sqrt{N}(\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}]) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\mathbf{X}_i - \mathbb{E}[\mathbf{X}]) \xrightarrow{d} \mathcal{N}\Big(0, \boldsymbol{\Omega}(\mathbf{X})\Big) \quad \text{(B.21)}$$

where $\boldsymbol{\Omega}(\mathbf{X}) = Var[\mathbf{X} - \mathbb{E}[\mathbf{X}]]$, the variance covariance matrix of $\mathbf{X}$ and the 0 expectation follows because $\mathbb{E}[\mathbf{X} - \mathbb{E}[\mathbf{X}]] = 0$. Hence,

$$\sqrt{N}(\hat{\tau}_0 - \tau_0) = \sqrt{N}(\tilde{\mu}_1 - \mu_1) - \sqrt{N}(\tilde{\mu}_0 - \mu_0) + \sqrt{N}(\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}])^t(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + o_p(1) \quad \text{(B.22)}$$

This implies that, $\sqrt{N}(\hat{\tau}_0 - \tau_0)$ can be expressed as the sum of 3 normally distributed random variables and a term that can be ignored asymptotically. If the 3 normally distributed random variables are pairwise independent, the asymptotic distribution of the sum will be a normal with a variance that is the sum of the 3. Pairwise independence between the first two terms was shown for the case of $\tilde{\tau}_{reg}$.

Let's investigate the independence between $\sqrt{N}(\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}])$ and $(\tilde{\mu}_1 - \mu_1)$ which is a component of $\sqrt{N}(\tilde{\boldsymbol{\delta}}_1 - \boldsymbol{\delta}_1)$. Hence, we will show independence between $\sqrt{N}(\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}])$ and $\sqrt{N}(\tilde{\boldsymbol{\delta}}_1 - \boldsymbol{\delta}_1)$. Since these are both normally distributed, it suffices to show that the cross covariance between the two is 0.

$$\begin{aligned}
Cov\Big(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} A_1^{-1} W_i \dot{\mathbf{R}}_i \epsilon_i(1), \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\mathbf{X}_i - \mathbb{E}[\mathbf{X}])\Big) &= Cov\Big(\boldsymbol{A}_{\mathbf{1}}^{-1} W_i \dot{\mathbf{R}}_i \epsilon_i(1), (\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i])\Big) \\
&= \mathbb{E}\Big(\boldsymbol{A}_{\mathbf{1}}^{-1} W_i \dot{\mathbf{R}}_i \epsilon_i(1) (\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i])^t\Big) \\
&= \mathbb{E}\Big(\boldsymbol{A}_{\mathbf{1}}^{-1} W_i \dot{\mathbf{R}}_i \epsilon_i(1) \mathbf{X}_i^t\Big) \\
&= \mathbb{E}\Big(\mathbb{E}\big(\boldsymbol{A}_{\mathbf{1}}^{-1} W_i \dot{\mathbf{R}}_i \epsilon_i(1) \mathbf{X}_i^t \big| \dot{\mathbf{R}}_i\big)\Big) \\
&= \boldsymbol{A}_{\mathbf{1}}^{-1} \mathbb{E}\Big(\dot{\mathbf{R}}_i \mathbf{X}_i^t \underbrace{\mathbb{E}\big(W_i \epsilon_i(1) | \dot{\mathbf{R}}_i\big)}_{=0}\Big) \\
&= 0
\end{aligned} \quad \text{(B.23)}$$

The same argument can be used for independence between $\sqrt{N}(\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}])$ and $\sqrt{N}(\tilde{\mu}_0 - \mu_0)$.
Hence,

$$\sqrt{N}(\hat{\tau}_0 - \tau_0) = T + o_p(1) \quad \text{(B.24)}$$

with $T \sim \mathcal{N}(0, v_0 + v_1 + v_t)$ where v_0 is the first element of $\boldsymbol{V}_0$, v_1 is the first element of $\boldsymbol{V}_{\mathbf{1}}$ and $v_t = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^t \boldsymbol{\Omega}(\mathbf{X})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$.

□

Note that for v_0 and v_1, we can use the usual estimators for the OLS variance. A consistent estimator for the additional term v_T is provided in Imbens and Wooldridge (2009) as

$$\hat{v}_T = (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_{\mathbf{0}})^t \frac{1}{N} \sum_{i=1}^{N} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^t (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_{\mathbf{0}})$$

B.3 Double Robustness property of the AIPW estimator

Proof. (Adapted from Tsiatis (2006)) We would like to investigate consistency of the AIPW estimator described in chapter 2 when estimation of the nuisance functions is based on parametric models. The propensity score and outcome regression functions are assumed to depend on unknown parameter vectors ψ and ξ respectively. In this context, let $\pi(\mathbf{X}, \psi)$ be the propensity score and $g_w(\mathbf{X}, \xi)$ be the outcome regression function. Let $\pi(\mathbf{X}, \hat{\psi})$ be an estimator of $\pi(\mathbf{X}, \psi)$ obtained using a parametric estimation technique such as a logistic regression estimated by maximum likelihood where $\hat{\psi} \xrightarrow{p} \psi^*$. Similarly, let $g_w(\mathbf{X}, \hat{\xi})$ be an estimator of $g_w(\mathbf{X}, \xi)$ obtained using a parametric estimation technique such as a linear regression estimated by OLS where $\hat{\xi} \xrightarrow{p} \xi^*$. If $\psi^* = \psi$ the propensity score model is correctly specified and if $\xi^* = \xi$ the outcome regression model is correctly specified. We will show that the AIPW estimator described above is consistent for the ATE if only one of the two models is correctly specified.

The AIPW estimator described in equation (3.11) can be rewritten as[1]

$$\underbrace{\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{W_i Y_i}{\pi(\mathbf{X}_i,\hat{\psi})} - \frac{[W_i - \pi(\mathbf{X}_i,\hat{\psi})]g_1(\mathbf{X}_i,\hat{\xi})}{\pi(\mathbf{X}_i,\hat{\psi})}\right\}}_{\hat{\mu}_1} - \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{(1-W_i)Y_i}{1-\pi(\mathbf{X}_i,\hat{\psi})} - \frac{[W_i - \pi(\mathbf{X}_i,\hat{\psi})]g_0(\mathbf{X}_i,\hat{\xi})}{1-\pi(\mathbf{X}_i,\hat{\psi})}\right\}}_{\hat{\mu}_0}.$$

If $\hat{\mu}_1$ and $\hat{\mu}_0$ are consistent for $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ respectively, the AIPW estimator is consistent. Let us first consider the probability limit of $\hat{\mu}_1$ such that one can investigate the conditions under which $\hat{\mu}_1$ is consistent for $\mathbb{E}[Y(1)]$. We can rewrite

$$\begin{aligned}
\hat{\mu}_1 &= \mathbb{E}\left\{\frac{WY}{\pi(\mathbf{X},\psi^*)} - \frac{[W-\pi(\mathbf{X},\psi^*)]g_1(\mathbf{X},\xi^*)}{\pi(\mathbf{X},\psi^*)}\right\} + o_p(1) \\
&= \mathbb{E}\left\{\frac{WY(1)}{\pi(\mathbf{X},\psi^*)} - \frac{[W-\pi(\mathbf{X},\psi^*)]g_1(\mathbf{X},\xi^*)}{\pi(\mathbf{X},\psi^*)}\right\} + o_p(1) \\
&= \mathbb{E}\left\{\frac{WY(1)}{\pi(\mathbf{X},\psi^*)} - \frac{\pi(\mathbf{X},\psi^*)Y(1)}{\pi(\mathbf{X},\psi^*)} + Y(1) - \frac{[W-\pi(\mathbf{X},\psi^*)]g_1(\mathbf{X},\xi^*)}{\pi(\mathbf{X},\psi^*)}\right\} + o_p(1) \\
&= \mathbb{E}\left\{\frac{Y(1)[W-\pi(\mathbf{X},\psi^*)]}{\pi(\mathbf{X},\psi^*)} + Y(1) - \frac{[W-\pi(\mathbf{X},\psi^*)]g_1(\mathbf{X},\xi^*)}{\pi(\mathbf{X},\psi^*)}\right\} + o_p(1) \\
&= \mathbb{E}[Y(1)] + \mathbb{E}\left\{\frac{\left[Y(1) - g_1(\mathbf{X},\xi^*)\right]\left[W-\pi(\mathbf{X},\psi^*)\right]}{\pi(\mathbf{X},\psi^*)}\right\} + o_p(1). \qquad \text{(B.25)}
\end{aligned}$$

[1] by multiplying and dividing the first two terms in (3.11) by $\pi(\mathbf{X}, \hat{\psi})$.

The first equality holds by the law of large numbers because $\hat{\mu}_1$ is a sample average and the second holds by the switching equation. We can now investigate consistency of $\hat{\mu}_1$ in two different scenarios.

First, consider the case in which the propensity score model is correctly specified, i.e, $\psi^* = \psi$. Then, one can rewrite $\hat{\mu}_1$ as

$$\begin{aligned}
\hat{\mu}_1 &= \mathbb{E}[Y(1)] + \mathbb{E}\left\{ \frac{W\big[Y(1) - g_1(\mathbf{X}, \xi^*)\big]}{\pi(\mathbf{X}, \psi)} - \frac{\pi(\mathbf{X}, \psi)\big[Y(1) - g_1(\mathbf{X}, \xi^*)\big]}{\pi(\mathbf{X}, \psi)} \right\} + o_p(1) \\
&= \mathbb{E}[Y(1)] + \mathbb{E}\left[\mathbb{E}\left\{ \frac{W\big[Y(1) - g_1(\mathbf{X}, \xi^*)\big]}{\pi(\mathbf{X}, \psi)} \middle| \mathbf{X} \right\} \right] - \mathbb{E}\left\{ \big[Y(1) - g_1(\mathbf{X}, \xi^*)\big] \right\} + o_p(1) \\
&= \mathbb{E}[Y(1)] + \mathbb{E}\left[\frac{\mathbb{E}[W|\mathbf{X}]\big[\mathbb{E}[Y(1)|\mathbf{X}] - g_1(\mathbf{X}, \xi^*)\big]}{\pi(\mathbf{X}, \psi)} \right] - \mathbb{E}\left\{ \big[Y(1) - g_1(\mathbf{X}, \xi^*)\big] \right\} + o_p(1) \\
&= \mathbb{E}[Y(1)] + \mathbb{E}\Big[\mathbb{E}[Y(1)|\mathbf{X}] - g_1(\mathbf{X}, \xi^*) \Big] - \mathbb{E}\left\{ \big[Y(1) - g_1(\mathbf{X}, \xi^*)\big] \right\} + o_p(1) \\
&= \mathbb{E}[Y(1)] + o_p(1).
\end{aligned}$$

The second equality holds by the law of iterated expectations, the third equality holds by the properties of conditional expectations and the fourth equality holds by definition of the propensity score. Hence, even if the outcome regression model is misspecified, $\hat{\mu}_1$ is consistent for $\mathbb{E}[Y(1)]$ as long as the propensity score model is correctly specified.

Second, consider the case when the outcome regression model is correctly specified, i.e, $\xi^* = \xi$. Under that assumption, equation (B.25) can be rewritten as

$$\begin{aligned}
\hat{\mu}_1 &= \mathbb{E}[Y(1)] + \mathbb{E}\left\{ \frac{W\big[Y(1) - g_1(\mathbf{X}, \xi)\big]}{\pi(\mathbf{X}, \psi^*)} - \frac{\pi(\mathbf{X}, \psi^*)\big[Y(1) - g_1(\mathbf{X}, \xi)\big]}{\pi(\mathbf{X}, \psi^*)} \right\} + o_p(1) \\
&= \mathbb{E}[Y(1)] + \mathbb{E}\left[\mathbb{E}\left\{ \frac{W\big[Y(1) - g_1(\mathbf{X}, \xi)\big]}{\pi(\mathbf{X}, \psi^*)} \middle| W, X \right\} \right] + o_p(1) \\
&= \mathbb{E}[Y(1)] + \mathbb{E}\left[\frac{W\big[\mathbb{E}[Y(1)|W, \mathbf{X}] - g_1(\mathbf{X}, \xi)\big]}{\pi(\mathbf{X}, \psi^*)} \middle| W, X \right\} \right] + o_p(1) \\
&= \mathbb{E}[Y(1)] + o_p(1).
\end{aligned}$$

The second equality holds by the law of iterated expectations and because by unconfoundedness $\mathbb{E}[g_1(\mathbf{X}, \xi)] = E[Y(1)]$. The third equality holds by the properties of conditional expectations. The last equality holds because under unconfoundedness, $\mathbb{E}[Y(1)|W, \mathbf{X}] = \mathbb{E}[Y|W = 1, \mathbf{X}] = g_1(\mathbf{X}, \xi)$. Hence, $\hat{\mu}_1$ is consistent if only the outcome regression is consistent.
Similar arguments can be used to show that $\hat{\mu}_0$ is consistent for $\mathbb{E}[Y(0)]$. Hence, the AIPW estimator is consistent for the ATE if either the propensity score or the outcome regression model is correctly specified. □

Bibliography

Splawa-Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 465–472.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, *66*(5), 688–701.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604–620.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 1349–1382.

Poterba, J. M., Venti, S. F., & Wise, D. A. (1995). Do 401 (k) contributions crowd out other personal saving? *Journal of Public Economics*, *58*(1), 1–32.

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *The review of economic studies*, *65*(2), 261–294.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062.

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, *113*(2), 231–263.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in medicine*, *23*(19), 2937–2960.

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*(1), 235–267.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data* (Vol. 4). Springer.

Sekhon, J. S. (2008). Multivariate and propensity score matching software with automated balance optimization: The matching package for r. *Journal of Statistical Software*.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, *47*(1), 5–86.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 266–298.

Abadie, A., & Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, *29*(1), 1–11.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*(1), 1063–1095.

Ichimura, H., & Newey, W. (2015). The Influence Function of Semiparametric Estimators. *arXiv preprint arXiv:1508.01378*.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Wager, S., & Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, *107*(5), 261–65.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, C1–C68.

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, *34(1)*, 43–68.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). Mlr3: A modern object-oriented machine learning framework in r. *Journal of Open Source Software*, *4*(44), 1903.

McConnell, K. J., & Lindner, S. (2019). Estimating treatment effects with machine learning. *Health services research*, *54*(6), 1273–1282.

Dorie, V., & Hill, J. (2020). Package 'bartcause'.

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, *15*(3), 965–1056.

Hernán, M., & Robins, J. (2020). *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC.

Yang, J.-C., Chuang, H.-C., & Kuan, C.-M. (2020). Double machine learning with gradient boosting and its application to the big n audit quality effect. *Journal of Econometrics*, *216*(1), 268–283.

Athey, S., Imbens, G. W., Metzger, J., & Munro, E. (2021). Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*, 105076.

Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (2021). Doubleml–an object-oriented implementation of double machine learning in r. *arXiv preprint arXiv:2103.09603*.

Efron, B., & Hastie, T. (2021). *Computer age statistical inference, student edition: Algorithms, evidence, and data science* (Vol. 6). Cambridge University Press.

Fisher, A., & Kennedy, E. H. (2021). Visually communicating and teaching intuition for influence functions. *The American Statistician*, *75*(2), 162–172.

Knaus, M. C., Lechner, M., & Strittmatter, A. (2021a). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence [Publisher: Oxford University Press]. *The Econometrics Journal*, *24*(1), 134–161.

Knaus, M. C., Lechner, M., & Strittmatter, A. (2021b). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, *24*(1), 134–161.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Hines, O., Dukes, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*, *76*(3), 292–304.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*.

Parikh, H., Varjao, C., Xu, L., & Tchetgen, E. T. (2022). Validating causal inference methods. *International conference on machine learning*, 17346–17358.

Waernbaum, I., & Pazzagli, L. (2023). Model misspecification and bias for inverse probability weighting estimators of average causal effects. *Biometrical Journal*, *65*(2), 2100118.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des Sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/sc