University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Kristin Jesse

Introduction to Propensity Score Methods

Mathematics and Statistics
Master's Thesis (30 ECTS credits)

Supervisors: Jaak Sõnajalg, MSc
Krista Fischer, PhD

Tartu 2021

INTRODUCTION TO PROPENSITY SCORE METHODS

Master's thesis

Kristin Jesse

Abstract

Randomised controlled trials (RCTs), while the golden standard of estimating causal effects in clinical studies, are not always possible to conduct due to ethical reasons or other restrictions. Observational studies are an alternative in such cases. However, in such studies, treatment assignment may be subject to systematic biases.

Propensity score (PS) methods are a popular tool to adjust for confounding factors in observational studies. By attempting to mimic RCTs, these methods are quite intuitive. This thesis provides a theoretical overview of the most popular PS methods, and conducts a simulation study to compare PS matching, PS weighting, and conventional covariate adjustment.

CERCS research specialisation: P160 Statistics, operation research, programming, financial and actuarial mathematics.
Key Words: propensity score, observational studies, matching, weighting, covariate adjustment.

SISSEJUHATUS KALDUVUSE MÄÄRA MEETODITESSE

Magistritöö

Kristin Jesse

Lühikokkuvõte

Randomiseeritud uuringud on põhjuslike seoste hindamise kuldstandard kliinilistes uuringutes, kuid neid ei ole alati eetilistel või muudel põhjustel võimalik läbi viia. Vaatlusuuringud on sellisel juhul heaks alternatiiviks, aga ravi määramine võib neis olla süstemaatiline.

Kalduvuse määra (*propensity score*, PS) meetodid on populaarne tööriist segavate tegurite arvessevõtmiseks vaatlusuuringutes. Need meetodid on küllalt intuitiivsed, kuna ideeks on imiteerida randomiseeritud uuringuid. See magistritöö annab teoreetilise ülevaate populaarsematest PS meetoditest ning

viib läbi simulatsiooniuuringu võrdlemaks PS ühildamist, PS kaalumist ja tavapärast tunnustele kohandamist.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.
Märksõnad: kalduvuse määr, vaatlusuuringud, ühildamine, kaalumine, tunnustele kohandamine.

Contents

Introduction

In medical studies, as well as many other fields, it is often of interest how an exposure, also referred to as a treatment or an intervention, affects a certain outcome. To account for possible covariates that affect either the exposure or outcome, or both, different methods can be applied. Such studies, that assess the effect of a treatment on an outcome, can commonly be divided into two: randomised controlled trials (RCTs) and observational studies.

In randomised trials, the study subjects are randomly allocated into the experimental group, that receives the treatment of interest, and the control group, that receives a different treatment or no treatment at all. If the randomisation is properly conducted, it is unlikely that the study groups differ remarkably, on average, in any aspect other than the assigned treatment.

However, it is not always possible or reasonable to conduct RCTs. In this case, observational studies are conducted. Routinely collected register data is one such option to analyse differences in outcomes. However, in observational studies there is no randomisation - treatment assignment may be subject to systematic biases. Propensity scores (PS) have been introduced by Rosenbaum and Rubin (1983) as one option to address this inherent weakness of observational studies.

A propensity score is the probability of a subject being assigned to a particular treatment, given a set of observed covariates. In randomised trials the propensity score is determined by the study design and is known. In observational studies the propensity score is, in general, not known and needs to be estimated from available data - most often using an appropriate logistic model, where treatment status is regressed on available baseline covariates. Other possible methods include random forests (Lee et al. 2010), and neural networks (Setoguchi et al. 2008). In this thesis, only logistic regression will be used.

The purpose of this thesis is to introduce the propensity score and its applications, and to illustrate the similarities and differences between these methods and classic covariate adjustment (logistic regression). A simulation study was conducted to examine these differences. Chapter 1 provides an overview of what the propensity score is and why it is necessary, as well as underlying theorems to show why it works. Chapter 2 covers the assumptions which the propensity score methods rely on. Chapter 3 introduces different

propensity score methods that are commonly used in practice. In Chapter 4, these methods are applied on simulated data.

The original contributions of the author of this thesis are the detailed proofs of the theorems from Rosenbaum and Rubin (1983) in Chapter 1, and the simulation study in Chapter 4. The thesis was written within the Industrial Master's Programme in Quantitative Analysis in collaboration with IQVIA.

1 The Basics of the Propensity Score

In this chapter, the propensity score and its purpose will be introduced. Section 1.2 relies on Rosenbaum and Rubin (1983). The proofs of the theorems are outlined in Rosenbaum and Rubin (1983) and detailed by the author of this thesis.

1.1 Randomised Controlled Trials and Observational Studies

Let us assume we wish to asses the effect of a treatment, also known as exposure or intervention, on a certain outcome. The aim is to compare outcomes of two groups, one of which receives the treatment of interest and the other does not. These are called a treatment group (or experimental group) and a control group, respectively. There may be more than one treatment group or more than one control group to be compared in a study, but in this thesis only one of each will be considered.

The treatment may be a treatment in the colloquial sense, like a drug that a patient is prescribed, or an operation they undergo; or a different kind of exposure, like smoking or having access to higher education. While the latter two would not be referred to as treatments in everyday conversations, here "treatment" refers to any exposure of interest. The outcome may be any event of interest, such as death, recovery from pneumonia, or graduating from high school.

In addition to treatment and outcome, there are other factors to be considered, called confounders. Confounders are any covariates that affect both the outcome and whether the subject received treatment. For example, when studying a drug's effect on recovery from an illness, having a liver disease may mean the person is less likely to be prescribed the drug, but also that the person is more likely to die during treatment and thus not to achieve the outcome of recovery. Since our aim is to assess the true effect of the treatment on the outcome, all such confounders need to be taken into account.

The golden standard of clinical studies is the randomised controlled trial (RCT). In these studies, the subjects are randomly allocated into the treatment or control group. If conducted correctly, this eliminates differences in confounders. Clearly, if the treatment assignment is truly random, then on

average the treatment and control group should not differ remarkably in any other aspect than the treatment which they receive.

However, RCTs are not always the way to go. For example, when studying the effect of smoking during pregnancy on the development of the fetus, it would be highly unethical to conduct a randomised study. In addition to ethical questions, other issues, such as time constraints, may arise. Assessing a certain drug's effect on ten-year morbidity, for example, would clearly require a study that is longer than ten years, which is often not a reasonable length for an RCT. In such cases, observational studies are conducted. For such data, we as investigators have no control over who gets treatment and who does not. One type of observational studies use routinely collected register data, which is the main focus here.

If we wish to study the effect of drug A on an outcome, e.g. 30-day morbidity, just calculating the average effect amongst those who have taken drug A and comparing it to the average amongst those who have not, would most likely give us a skewed picture of the true effect due to aspects that have affected the assignment of treatment. For example, doctors may prefer prescribing drug A to younger patients while using a different approach for older people. Since in general, old people tend to die more often than young people, calculating the average outcome in these groups and claiming this is the true difference in treatment effect would make it seem like drug A reduces 30-day morbidity drastically. Now, if we compared people of similar ages, the picture may be very different.

If we truly wish to know what effect a treatment has on a person, we would need two alternative universes: one where the subject does not receive treatment, and another that is identical in every other way, except that the subject receives treatment. Then we could see which outcome is achieved in either of these scenarios. These scenarios are referred to as potential outcomes or counterfactual outcomes. In reality, we can never compare these situations because a person cannot simultaneously receive and not receive treatment.

Let us formulate this in mathematical terms. Let Z be an indicator for whether a subject received treatment, i.e. $Z = 1$ if the subject received treatment and $Z = 0$ if they did not. While in general, the treatment may also be continuous or have many levels, like the dosage of a drug, here we will only consider a binary treatment.

Let $Y_t,\ t \in \{0, 1\}$, be the counterfactual outcomes, where Y_1 is the outcome

if treatment was received ($Z = 1$) and Y_0 is the outcome if treatment was not received ($Z = 0$). Just like treatment, the outcome may also be a continuous variable or a discrete variable with many levels, but here we will only study binary outcomes, i.e. $Y_t = 1$ if the subject achieves the outcome, and $Y_t = 0$ if the outcome event does not happen to the subject. To reiterate, one subject has two potential outcomes, Y_0 if they do not receive treatment, and Y_1 if they do. These may be equal ($Y_0 = Y_1 = 0$ or $Y_0 = Y_1 = 1$) or different ($Y_0 = 0$ and $Y_1 = 1$, or $Y_0 = 1$ and $Y_1 = 0$), depending on the person. However, we can ever only observe one of these for each subject. We would only know both if we had the aforementioned parallel universes at our command.

Additionally, let $\boldsymbol{X}$ be a vector of observed covariates preceding treatment. Ideally, this would include all confounders that affect the treatment assignment and outcome.

Often in reality, we do not know exactly which confounders are present, and therefore must consult with experts in the relevant field who will have better knowledge of possible causal structures. Sometimes, several different models may need to be considered, analysed and presented, as one can never be completely certain of the underlying causal structures when dealing with observational data.

Causal structures can be illustrated by directed acyclic graphs (DAGs) like in Figure 1. The presence of an arrow pointing from one variable to another indicates that there is a direct causal effect between these variables for at least one individual. The lack of an arrow, on the other hand, means that we know there is no causal effect between those variables for any individual in the population. A path is causal if it consists only of arrows pointing in the same direction; otherwise it is non-causal. (Hernán and Robins 2020)

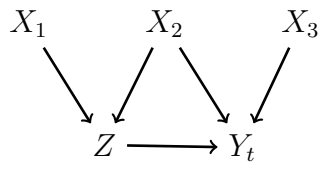

Figure 1: A directed acyclic graph (DAG)

Figure 1 depicts a situation where the treatment Z has a direct causal effect on the outcome Y_t. Of the covariates, X_1 has direct causal effect on the

treatment Z, X_2 has direct causal effects on both the treatment and the outcome, and X_3 has a direct causal effect on the outcome Y_t. While X_1 does not affect the outcome Y_t directly, there is a causal path between them: $X_1 \to Z \to Y_t$. However, there is no causal relationship between Z and X_3, for example, because the paths "collide" at Y_t.

Returning to the example of drug A and its effect on 30-day mortality, we can now express the scenario mathematically. If a patient is prescribed drug A, then for that person $Z = 1$. If a patient is not prescribed this drug, then $Z = 0$. If a patient dies within 30 days of the start of the study, then the outcome $Y = 1$, otherwise $Y = 0$. For simplicity, let us assume that age, denoted by X, is the only confounder.

The average treatment effect (ATE), which we wish to estimate, is the difference between the expected outcome of the population if everyone received treatment and the expected outcome of the population if no one received treatment, i.e.

$$E(Y_1) - E(Y_0), \tag{1}$$

where $E(\cdot)$ is expectation in population. Since we can only observe one of the counterfactual outcomes for each subject based on their treatment status, we can estimate the difference

$$E(Y_1 \mid Z = 1) - E(Y_0 \mid Z = 0), \tag{2}$$

which is usually not equal to the average treatment effect (1).

To illustrate this, let us consider the data in Table 1. Let us say that, in this example, this data set is our entire study population. We can now easily calculate the average treatment effect (1) and the observed difference (2), and see that they are not equal:

$$E(Y_1) - E(Y_0) = \frac{3}{8} - \frac{4}{8} = -0.125$$
$$E(Y_1 \mid Z = 1) - E(Y_0 \mid Z = 0) = \frac{1}{5} - \frac{2}{3} = -0.467$$

Table 1: Example of a possible study population. Here, Z is treatment with drug A, Y_0 and Y_1 are the counterfactual outcomes (30-day mortality if not treated or if treated, respectively), and X is age in full years. Observed outcome is in bold text.

Subject	**Z**	$\mathbf{Y_0}$	$\mathbf{Y_1}$	**X**
Mary	1	0	**0**	19
John	1	1	**0**	25
Will	1	0	**1**	27
Martin	1	0	**0**	35
Tony	1	1	**0**	36
Tina	0	**0**	1	48
Jane	0	**1**	1	60
Wanda	0	**1**	0	77

If we now also pay attention to the age of the patient, we notice that all the younger patients (ages 19 to 36) were treated with drug A and none of the older patients (ages 48 to 77) were treated with the drug. Due to these circumstances, we would severely overestimate the actual effect of drug A on 30-day mortality if we were to use the difference between observed outcome averages as an estimate.

To address this inherent weakness of observational studies, propensity scores have been introduced as one possible option. The following section covers definitions and theorems necessary to understand the concept.

1.2 Definitions and Theorems

This section relies on Rosenbaum and Rubin (1983). The proofs of the theorems are outlined in Rosenbaum and Rubin (1983) and detailed by the author of this thesis. In the following, $\boldsymbol{x}$ is a realisation of the random variable $\boldsymbol{X}$.

Definition 1 (Rosenbaum and Rubin 1983)**.** *The conditional probability of being assigned treatment ($Z = 1$) given the covariates $\boldsymbol{X}$ is called* ***propensity score****, and denoted*

$$ps(\boldsymbol{x}) := P(Z = 1 \mid \boldsymbol{X} = \boldsymbol{x}),$$

where $P(\cdot)$ is the probability function.

In the example in Table 1, the propensity scores would be

$$ps(\boldsymbol{x} \in \{19, 25, 27, 35, 36\}) = 1$$

and

$$ps(\boldsymbol{x} \in \{48, 60, 77\}) = 0.$$

In this small population, for any other age, the propensity score is undefined. Generally, when dealing with larger populations, we will expect the in-between ages (or values of other confounders) also to be present and only to be dealing with a sample instead of the whole population. In that case, the propensity score values can be interpolated naturally, assuming that we know the nature of the relationship between different covariates and the treatment assignment.

Definition 2 (Rosenbaum and Rubin 1983)**.** *We say that function b is a* ***balancing score*** *if the distribution of $\boldsymbol{X}$ given $b(\boldsymbol{X})$ is the same for treated and untreated units, i.e.*

$$P(\boldsymbol{X} = \boldsymbol{x} \mid Z = 0,\ b(\boldsymbol{X}) = b(\boldsymbol{x})) = P(\boldsymbol{X} = \boldsymbol{x} \mid Z = 1,\ b(\boldsymbol{X}) = b(\boldsymbol{x}))$$

for all $\boldsymbol{x}$. In that case, we use the notation $\boldsymbol{X} \perp\!\!\!\perp Z \mid b(\boldsymbol{X})$.

In the example in Table 1, such function b cannot be found. Conditional probability is only defined if the probability of the event we are conditioning on is greater than zero. In the given example, however, at least one of the probabilities $P(Z = 0,\ b(\boldsymbol{X}) = b(\boldsymbol{x}))$ and $P(Z = 1,\ b(\boldsymbol{X}) = b(\boldsymbol{x}))$ is always equal to zero, unless b is a constant function, in which case it gives no additional information and the conditional probabilities are still not equal.

Let us give another example to illustrate what a balancing score is. Consider the data in Table 2, where $b_1(X_1, X_2) = 2X_1 + 3X_2$ and $b_2(X_1, X_2) = X_1 + X_2$. Calculating the necessary distributions is then straightforward. We can see that the distributions $P(\boldsymbol{X} \mid Z = 0,\ b_1)$ and $P(\boldsymbol{X} \mid Z = 1,\ b_1)$ are equal, because

Table 2: Example of a function that is a balancing score (b_1) and a function that is not a balancing score (b_2). Here $b_1(X_1, X_2) = 2X_1 + 3X_2$ and $b_2(X_1, X_2) = X_1 + X_2$.

$\mathbf{Z}$	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{b_1}$	$\mathbf{b_2}$
0	1	1	5	2
0	1	0	2	1
0	0	1	3	1
0	0	1	3	1
1	1	1	5	2
1	1	1	5	2
1	1	0	2	1
1	0	1	3	1

$$
\begin{aligned}
P(\boldsymbol{X} = \boldsymbol{x} \mid Z = 0,\ b_1 = 5) &= P(\boldsymbol{X} = \boldsymbol{x} \mid Z = 1,\ b_1 = 5) \\
&= \begin{cases} 1, & \text{if } \ \boldsymbol{x} = (1,1), \\ 0, & \text{otherwise,} \end{cases} \\
P(\boldsymbol{X} = \boldsymbol{x} \mid Z = 0,\ b_1 = 3) &= P(\boldsymbol{X} = \boldsymbol{x} \mid Z = 1,\ b_1 = 3) \\
&= \begin{cases} 1, & \text{if } \ \boldsymbol{x} = (0,1), \\ 0, & \text{otherwise,} \end{cases} \\
P(\boldsymbol{X} = \boldsymbol{x} \mid Z = 0,\ b_1 = 2) &= P(\boldsymbol{X} = \boldsymbol{x} \mid Z = 1,\ b_1 = 2) \\
&= \begin{cases} 1, & \text{if } \ \boldsymbol{x} = (1,0), \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

For b_2, however, the distributions are not equal, because

$$
\begin{aligned}
P(\boldsymbol{X} = \boldsymbol{x} \mid Z = 0,\ b_2 = 1) &= \begin{cases} \frac{1}{3}, & \text{if } \ \boldsymbol{x} = (1,0), \\ \frac{2}{3}, & \text{if } \ \boldsymbol{x} = (0,1), \\ 0, & \text{otherwise,} \end{cases} \\
P(\boldsymbol{X} = \boldsymbol{x} \mid Z = 1,\ b_2 = 1) &= \begin{cases} \frac{1}{2}, & \text{if } \ \boldsymbol{x} = (1,0) \text{ or } \ \boldsymbol{x} = (0,1), \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

Therefore, b_1 is a balancing score, but b_2 is not.

The following definitions are used as assumptions in the theorems that follow. More about the assumptions can be read in Chapter 2.

Definition 3. *We say that* ***exchangeability*** *holds if, given measured confounders, the potential outcomes are independent of observed exposure, i.e.*

$$(Y_1, Y_0) \perp\!\!\!\perp Z \mid (\boldsymbol{X} = \boldsymbol{x}) \quad \forall \boldsymbol{x}.$$

Definition 4. *We say that* ***positivity*** *holds if the probability of each individual being assigned to the treatment group or control group is non-zero, i.e.*

$$0 < P(Z = 1 \mid \boldsymbol{X} = \boldsymbol{x}) < 1 \quad \forall \boldsymbol{x}.$$

Definition 5 (Rosenbaum and Rubin 1983)**.** *We say that treatment assignment is* ***strongly ignorable*** *if*

$$(Y_1, Y_0) \perp\!\!\!\perp Z \mid (\boldsymbol{X} = \boldsymbol{x}), \quad 0 < P(Z = 1 \mid \boldsymbol{X} = \boldsymbol{x}) < 1 \quad \forall \boldsymbol{x},$$

i.e. both exchangeability and positivity hold.

In the data in Table 1, the treatment assignment is clearly not strongly ignorable, because the positivity condition does not hold.

The following theorem shows the relationship between the propensity score and balancing scores.

Theorem 1 (Rosenbaum and Rubin 1983)**.** *Let b be some function of $\boldsymbol{X}$. Then $b(\boldsymbol{X})$ is a balancing score if and only if there exists a function f such that $ps(\boldsymbol{X}) = f(b(\boldsymbol{X}))$, where ps is the propensity score.*

Proof. Necessity ($\Leftarrow$): Let $ps(\boldsymbol{X}) = f(b(\boldsymbol{X}))$ for some f. We need to show that $\boldsymbol{X} \perp\!\!\!\perp Z \mid b(\boldsymbol{X})$, which is equivalent to

$$\begin{aligned} P(Z = 1 \mid \boldsymbol{X} = \boldsymbol{x}, b(\boldsymbol{X}) = b(\boldsymbol{x})) &= P(Z = 1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) \\ \text{or} \quad P(\boldsymbol{X} = \boldsymbol{x} \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) &= 0 \quad \forall \boldsymbol{x} \end{aligned}$$

by definition of conditional independence. Since $b(\boldsymbol{X})$ is a function of $\boldsymbol{X}$, we have

$$P(Z = 1 \mid \boldsymbol{X} = \boldsymbol{x},\ b(\boldsymbol{X}) = b(\boldsymbol{x})) = P(Z = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = ps(\boldsymbol{x}).$$

Therefore, it is sufficient to show that

$$P(Z = 1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) = ps(\boldsymbol{x}).$$

It holds that

$$\begin{aligned} P(Z = 1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) &= E(P(Z = 1 \mid \boldsymbol{X} = \boldsymbol{x}) \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) \\ &= E(ps(\boldsymbol{x}) \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) && \text{(ps def.)} \\ &= E(f(b(\boldsymbol{x})) \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) && \text{(assum.)} \\ &= f(b(\boldsymbol{x})) && (*) \\ &= ps(\boldsymbol{x}) && \text{(assum.)}, \end{aligned}$$

where (*) holds due to the property of conditional expectation that for a random variable W, $E(f(W) \mid W) = f(W)$.

Sufficiency ($\Rightarrow$): Let b be a balancing score. Suppose, for the sake of contradiction, that there exist $\boldsymbol{x}_1, \boldsymbol{x}_2$ such that $b(\boldsymbol{x}_1) = b(\boldsymbol{x}_2)$, but $ps(\boldsymbol{x}_1) \neq ps(\boldsymbol{x}_2)$, meaning that there is no such function f that $ps(\boldsymbol{X}) = f(b(\boldsymbol{X}))$.

From the discussion in proof of necessity, and assumption that $ps(\boldsymbol{x}_1) \neq ps(\boldsymbol{x}_2)$, we get

$$P(Z = 1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x}_1)) = ps(\boldsymbol{x}_1) \neq ps(\boldsymbol{x}_2) = P(Z = 1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x}_2)).$$

On the other hand, since $b(\boldsymbol{x}_1) = b(\boldsymbol{x}_2)$, it must hold that

$$P(Z = 1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x}_1)) = P(Z = 1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x}_2)).$$

This is a contradiction, and therefore, if b is a balancing score there must exist a function f such that $ps(\boldsymbol{X}) = f(b(\boldsymbol{X}))$.

□

It follows directly from Theorem 1, taking f to be the identity function, that the propensity score itself is also a balancing score.

Theorem 2 (Rosenbaum and Rubin 1983)**.** *If treatment assignment is strongly ignorable given $\boldsymbol{X}$ then it is also strongly ignorable given $b(\boldsymbol{X})$, that is if*

$$(Y_1, Y_0) \perp\!\!\!\perp Z \mid (\boldsymbol{X} = \boldsymbol{x}), \quad 0 < P(Z = 1 \mid \boldsymbol{X} = \boldsymbol{x}) < 1 \quad \forall \boldsymbol{x},$$

then

$$(Y_1, Y_0) \perp\!\!\!\perp Z \mid (b(\boldsymbol{X}) = b(\boldsymbol{x})), \quad 0 < P(Z = 1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) < 1 \quad \forall b(\boldsymbol{x}),$$

where b is a balancing score.

Proof. Since b is a balancing score, then from the proof of Theorem 1

$$P(Z = 1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) = P(Z = 1 \mid \boldsymbol{X} = \boldsymbol{x}),$$

and the inequality $0 < P(Z = 1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) < 1$ follows trivially from $0 < P(Z = 1 \mid \boldsymbol{X} = \boldsymbol{x}) < 1$. Thus the proof of positivity is complete.

To prove the exchangeability, assuming that the counterfactual outcomes (Y_1, Y_0) are independent of treatment Z given covariates $\boldsymbol{X}$, we need to show that $(Y_1, Y_0) \perp\!\!\!\perp Z \mid b(\boldsymbol{X})$ holds; equivalently

$$P(Z = 1 \mid Y_1,\ Y_0,\ b(\boldsymbol{X}) = b(\boldsymbol{x})) = P(Z = 1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x}))$$

or $P((Y_1, Y_0) \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) = 0$.

Again, from proof of Theorem 1, we have $P(Z = 1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) = ps(\boldsymbol{x})$. Therefore if suffices to show that $P(Z = 1 \mid Y_1,\ Y_0,\ b(\boldsymbol{X}) = b(\boldsymbol{x})) = ps(\boldsymbol{x})$.

Indeed, if f is a function such that $ps(\boldsymbol{X}) = f(b(\boldsymbol{X}))$, then

$$\begin{aligned}
P(Z = 1 \mid Y_1, Y_0, b(\boldsymbol{X}) = b(\boldsymbol{x})) = & \\
&= E(P(Z = 1 \mid Y_1, Y_0, \boldsymbol{X} = \boldsymbol{x}) \mid Y_1, Y_0, b(\boldsymbol{X}) = b(\boldsymbol{x})) \\
&= E(P(Z = 1 \mid \boldsymbol{X} = \boldsymbol{x}) \mid Y_1, Y_0, b(\boldsymbol{X}) = b(\boldsymbol{x})) && \text{(assum.)} \\
&= E(ps(\boldsymbol{x}) \mid Y_1, Y_0, b(\boldsymbol{X}) = b(\boldsymbol{x})) && \text{(ps def.)} \\
&= E(f(b(\boldsymbol{x})) \mid Y_1, Y_0, b(\boldsymbol{X}) = b(\boldsymbol{x})) && \text{(Th 1)} \\
&= f(b(\boldsymbol{x})) && (*) \\
&= ps(\boldsymbol{x}), && \text{(assum.)}
\end{aligned}$$

where (*) holds due to the property of conditional expectation that for a random variable W, $E(f(W) \mid W) = f(W)$.

□

Theorem 3 (Rosenbaum and Rubin 1983)**.** *Let treatment assignment be strongly ignorable and b be a balancing score. Then the expected difference in*

observed responses to two treatments at $b(\boldsymbol{x})$ is equal to the average treatment effect at $b(\boldsymbol{x})$, i.e.

$$E(Y_1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x}), Z = 1) - E(Y_0 \mid b(\boldsymbol{X}) = b(\boldsymbol{x}), Z = 0) = \\ = E(Y_1 - Y_0 \mid b(\boldsymbol{X}) = b(\boldsymbol{x})).$$

Proof. Given strongly ignorable treatment assignment, it follows directly from Theorem 2 that

$$\begin{aligned} E(Y_1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x}), Z = 1) - E(Y_0 \mid b(\boldsymbol{X}) = b(\boldsymbol{x}), Z = 0) = \\ = E(Y_1 \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) - E(Y_0 \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) \\ = E(Y_1 - Y_0 \mid b(\boldsymbol{X}) = b(\boldsymbol{x})) \end{aligned}$$

□

In other words, Theorem 3 tells us that under strongly ignorable treatment assignment, units from different treatments with the same value of the balancing score b can act as controls for each other in the sense that the expected difference in their responses equals the average treatment effect. In the following chapters, we use the propensity score as a balancing score.

In general, as discussed in the previous section, if treatment assignment is not strongly ignorable, then comparing a randomly selected treated unit to a randomly selected control unit does not result in average treatment effect, that is

$$E(Y_1 \mid Z = 1) - E(Y_0 \mid Z = 0) \neq E(Y_1) - E(Y_0),$$

because sampling has been done from conditional distribution of Y_t given $Z = t$, not from the marginal distribution of Y_t.

Theorem 3 is a powerful tool in observational studies, as long as one remembers that it relies on the assumption of strongly ignorable treatment assignment. If exchangeability or positivity do not hold, then the balancing property of the propensity score is not guaranteed.

2 Assumptions for Propensity Score Methods

To identify causal effects using propensity score methods, four assumptions need to hold: consistency, exchangeability, positivity, and no misspecification of the used models.

2.1 Consistency

Consistency is the assumption that a subject's potential counterfactual outcome under the treatment received is equal to the outcome observed. Note that this differs from the statistical property of consistency, which is that the bias of an estimator approaches zero when information increases. (Cole and Hernán 2008)

This may seem like a fairly obvious assumption that would always be fulfilled. However, problems may arise if treatments and counterfactual outcomes are not well-defined, or if data set is not sufficiently rich. Let us illustrate this with the following example inspired by Hernán and Robins (2020).

Say we want to observe the effect of obesity Z at age 40 on mortality R by age 50. There are many ways a person could become obese by the age of 40. They could have been obese for ten years or only one. They could be slightly over the limit of the definition of obese, or severely so. Therefore there are many different versions of treatment Z and for it not to be ill-defined we need to specify which version of obesity we are interested in.

Even if we managed to unambiguously define the "obesity" to be studied, there are still several ways a person could get to that point. Say person A has a genetic predisposition to large amounts of fat tissue in their waist and in their coronary arteries. If this person is obese at age 40 and has a myocardial infarction at age 48, then the outcome is $Y_1 = 1$. If that same person A would have neutral genes but poor diet and low activity levels, they can still be obese at 40, but might not die by age 50. In that case the outcome is $Y_1 = 0$. Therefore, even under relatively well-defined treatment, the outcome is ill-defined. Ill-defined counterfactual outcomes, in turn, lead to vague causal questions.

To reiterate, for consistency to hold, the potential counterfactual outcome

under the treatment received must be equal to the outcome observed. If the outcome is defined ambiguously, then there might be several different possible values for the same counterfactual outcome: the previous example illustrates that if $Y_1 = 0$ for an obese person if they had "good" genes but poor diet, and $Y_1 = 1$ for the same person if they had "bad" genes and good diet, then the potential counterfactual outcome has two different values at the same time, and the observed outcome cannot possibly be equal to both of them.

The process of better specifying the treatment and outcomes will sharpen the question of interest. Say that experts now agree that no meaningful vagueness remains in the definitions of treatment and counterfactual outcomes. Even then, we need to make sure that, when using observational data, there are some individuals that received treatment ($Z = 1$) and some that did not ($Z = 0$). Being able to describe a well-defined intervention is not meaningful if we have no data where, for example, the equality $Y_1 = 1$ holds for at least some individuals. This overlaps partially with the positivity assumption described in §2.3.

The characterisation of the treatment versions should be done in cooperation with experts in the study field, but because even experts are fallible, it is best to make the discussions and assumptions as transparent as possible, so that others can refer to and challenge them. (Hernán and Robins 2020)

2.2 Exchangeability

Exchangeability, in its essence, means the assumption of no unmeasured confounders (Cole and Hernán 2008). That is, given measured confounders, the potential outcomes are independent of observed exposure,

$$(Y_0, Y_1) \perp\!\!\!\perp Z \mid (\boldsymbol{X} = \boldsymbol{x}) \quad \forall \boldsymbol{x},$$

like given in Definition 3.

A randomised experiment is expected to result in exchangeability because independent predictors of the outcome will be approximately equally distributed between the treated and the untreated groups. In observational studies, where treatment is not randomly assigned, the reasons for receiving treatment are likely to be associated with some predictors of outcome. Exchangeability will not hold if there exist unmeasured predictors U of the

outcome such that the probability of receiving treatment depends on U within strata of measured covariates $\boldsymbol{X}$. In other words, if we have an unmeasured confounder that for different values of $\boldsymbol{X}$ affects the treatment assignment Z differently, then exchangeability does not hold. (Hernán and Robins 2020)

For the assumption to hold, we need to measure enough joint predictors of exposure and outcome so that the associations between exposure and outcome, that are due to their common causes, disappear. Exchangeability assumptions are not testable in observed data, but there is certain sensitivity analysis that can be applied. (Cole and Hernán 2008)

2.3 Positivity

Positivity assumption (Definition 4) is the condition that the probability of each individual being assigned to each level of treatment is non-zero,

$$0 < P(Z = 1 \mid \boldsymbol{X} = \boldsymbol{x}) < 1 \quad \forall \boldsymbol{x}.$$

Positivity and exchangeability together give the previously defined strongly ignorable treatment assignment in Definition 5.

If a subject cannot possibly be exposed to a treatment at one or more levels of the confounders, then positivity is violated because there is a zero probability of receiving treatment. For example, if liver disease is a contraindication for taking a medication, then when studying the effects of that medication, people with liver disease have a near-zero probability of receiving treatment. One simple solution in that case would be to restrict the inference to a subset where positivity holds, i.e. we exclude people with liver disease and do not claim to draw any conclusions about that sub-population. (Cole and Hernán 2008)

Even if structural zeros are absent, we may encounter zeros by chance because of small sample sizes or high dimensional data. In fact, when modelling continuously distributed covariates, random zeros are essentially a given due to the infinite number of possible values. In such cases, the use of parametric models smooths over the random zeros by borrowing information from individuals with histories similar to those that, by chance, resulted in zeros. (Cole and Hernán 2008)

Weighting methods (covered in Chapter 3) are more sensitive to random zeros than standard regression or stratification methods. For example, inverse

probability weights would be undefined for zero-probabilities. Non-weighted methods like standard regression and stratification implicitly extrapolate to levels of the covariates with lack of positivity. (Cole and Hernán 2008)

Covariates that cause severe non-positivity bias because of a strong association with exposure, may need to be omitted. (Cole and Hernán 2008)

2.4 Correct model specification

To appropriately use the methods described in Chapter 3, it is important to correctly specify the model for treatment assignment, i.e. the propensity score. As we are focusing on estimating the propensity score using logistic regression, the same problems may arise as with any regression model. On one hand, if we leave out important covariates, our estimates could be biased. On the other hand, if we include too many covariates, we might run into over-specification issues, such as inflated standard errors.

To specify the correct propensity score model, statistical methods are usually not enough, and we must consult with experts in the relevant field who will have better knowledge of possible causal structures. Several different models may need to be considered and presented, as we cannot be completely certain of the underlying causal structures in observational data. And even then, there is no guarantee of no misspecification as the approaches may be biased in the same direction. (Hernán and Robins 2020)

3 Propensity Score Methods

The following gives an overview of methods where the propensity score (PS) is used in practice, and why, relying on the assumptions covered in Chapter 2, these methods give the desired results. The presented corollaries also hold for any other balancing score, but only propensity scores are of interest to us.

3.1 Matching

3.1.1 Overview

Since, in general, $E(Y_t \mid Z = t) \neq E(Y_t),\ t \in \{0, 1\}$, then the expected difference between the average outcome of all available treated units and the average outcome of all available control units does not necessarily equal the expected treatment effect.

The goal of matching is, for each treated unit, to find a comparable control unit (or several) based on observed covariates. Ideally, matching would be done exactly on all covariates $\boldsymbol{x}$. In that case the resulting sample distributions of $\boldsymbol{x}$ would be identical for the treated and control units. By Theorem 1, it is sufficient to match exactly on a balancing score b, e.g. propensity score, to obtain the same probability distribution of baseline covariates for the treated and control groups.

Corollary 3.1 follows directly from Theorem 3.

Corollary 3.1 (Rosenbaum and Rubin 1983)**.** *Suppose treatment assignment is strongly ignorable. Further suppose that a value of the propensity score, $ps(\boldsymbol{x})$ is randomly sampled from the population of units, and then one treated unit and one control unit are sampled with this value of $ps(\boldsymbol{x})$. Then the expected difference in response to the two treatments for the units in the matched pair equals the average treatment effect at $ps(\boldsymbol{x})$. Moreover, the mean of matched pair differences obtained by this two-step sampling process is unbiased for the average treatment effect.*

Due to the potentially infinite amount of possible values of the estimated propensity score, or more generally, any balancing score, finding an exact match to a treated unit among control units is often impossible. Thus, a

control unit with a value of the estimated propensity score close enough to that of the sampled treated unit will be chosen. Which difference in value is deemed small enough will be determined for each study separately.

In most studies, one-to-one matching is used, but many-to-one matching or matching using a varying amount of controls to one treated unit is also possible. Different approaches to matching include, for example, matching with or without replacement, and greedy or optimal matching. (Austin 2011)

When matching with replacement, the same control unit can be matched to several different treated units. Then, variance estimation must account for this fact. (Austin 2011)

In greedy matching, first a treated unit is sampled and then the control unit closest in estimated propensity score value will be chosen as a match for it. This process is repeated until all treated units have been matched or until no control unit can be found to match a treated unit. The remaining units in the sample will then be excluded from the following analysis. In optimal matching, matches are made so that the total within-pair difference of the propensity score is minimized. (Austin 2011)

After matched groups have been formed, the treatment effect can be estimated by directly comparing the treated and untreated units in these groups. The reporting of treatment effects can then be done in the same metrics as in randomised controlled trials. Just like in randomised controlled trials, in propensity score matching, in case of large samples, the single covariates are, on average, similarly distributed in different treatment groups. (Austin 2011)

Propensity score matching requires a substantial overlap in the distributions of the propensity score in treatment and control groups. If there is little overlap then a match cannot be found for a lot of units based on their propensity score, and Corollary 3.1 cannot be applied. In this case, inferences could be made for only a small subset of the population.

R software offers a package called `MatchIt` (Ho et al. 2011) for matching purposes.

3.1.2 Example

Let us consider the propensity score distributions in Figure 2. In blue, we have the propensity score distribution of the treatment group, and in red, the propensity score distribution of the control group. There are two differ-

ent scenarios depicted: Figure 2a shows a sample with quite a considerable amount of overlap, while Figure 2b clearly has a large number of treated subjects that cannot be matched to a similar subject of the control group, and an even larger number of control group subjects that will not even be considered for a match with a treated subject. However, the overlap regions of the propensity scores are approximately from 0.2 to 0.75 for both scenarios.

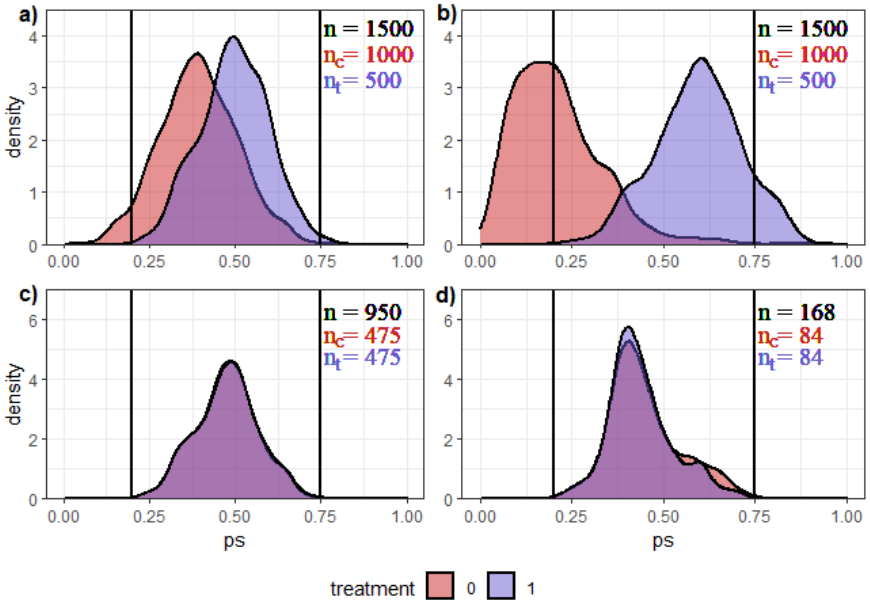

Figure 2: Samples with different overlap in the distributions of propensity scores in treatment and control groups.
a) Unmatched data, sufficient overlap.
b) Unmatched data, insufficient overlap.
c) Matched data corresponding to a).
d) Matched data corresponding to b).

After one-on-one matching without replacement based on the propensity score, we get new distributions for both groups, seen in Figures 2c and 2d. While both scenarios result in seemingly good matches, it is important to note that while in the left-hand side scenario, 95% of the treated people have been matched to corresponding control group units, only about 17% of the treatment group has been matched to control units on the right-hand side, leaving us with only about 11% of the original data, in total. Even in

the sufficient overlap scenario, only slightly over 60% of the total data set remains.

In the right-hand side scenario, certain inferences could still be made, depending on the outcome of interest, but the study question would need to be revised to reflect the actual subset of the population that the remaining data represents.

Such pairs of figures are often used in practice to illustrate how well the groups have been matched, but we must keep in mind that, while a good visual aid, they should not be used without considering how many subjects are actually matched.

Matching has been criticised for discarding a lot of information, even if most treated units find a match, like in Figures 2a and 2c. Additionally, matching on propensity score in particular, has been noted by King and Nielsen (2019) to increase imbalance and bias.

3.2 Stratification

3.2.1 Overview

In the stratification method, units are divided into subclasses or strata based on the observed covariates $\boldsymbol{x}$. The following Corollary is an immediate inference from Theorem 3.

Corollary 3.2 (Rosenbaum and Rubin 1983)**.** *Suppose treatment assignment is strongly ignorable. Suppose further that a group of units is sampled using* $ps(\boldsymbol{x})$ *such that* $ps(\boldsymbol{x})$ *is constant for all units in the group, and at least one unit in the group received each treatment. Then, for these units, the expected difference in treatment means equals the average treatment effect at that value of* $ps(\boldsymbol{x})$*. Moreover, the weighted average of such differences, that is, the directly adjusted difference, is unbiased for the treatment effect, when the weights equal the fraction of the population at* $ps(\boldsymbol{x})$*.*

When classifying directly based on the covariates $\boldsymbol{x}$, the number of subclasses grows fast; even if each covariate only has two possible values, the number of strata would be 2^k, where k is the number of different covariates. Thus, the more covariates we observe, the more subclasses will likely not have both treated and control units in them. Stratifying on propensity score is a good

alternative, given that the assumptions for Corollary 3.2 hold. (Rosenbaum and Rubin 1983)

Although Corollary 3.2 only talks of constant propensity scores, in practice, stratification means dividing the data into a certain small number of subclasses based on the propensity score. Then, in each subclass, the propensity score values for the treated and untreated are roughly similar, and thus the distribution of observed baseline covariates will be roughly similar for the treatment groups as well. In general, to estimate the treatment effect in the entire population, stratum-specific estimates are weighted by the proportion of subjects within that stratum. (Austin 2011)

Rosenbaum and Rubin (1984) showed that such stratification on the propensity score eliminates approximately 90% of the bias due to measured confounders when estimating a linear treatment effect.

3.2.2 Example

In Figure 3a, we see the same propensity score distribution as in Figure 2a. The overlap region is now divided into five strata of equal lengths: [0.2, 0.31), [0.31, 0.42), ..., [0.64, 0.75). Figures 3b-f show the distributions of propensity scores in these strata.

On visual inspection, the PS distributions for control and treatment groups are closer to each other within the strata than in the entire sample. The last subgroup, where propensity scores range from 0.64 to 0.75, has very few observations and the distributions there are not as similar as in the other strata. Different subgroups could be considered to reach even more similar distributions.

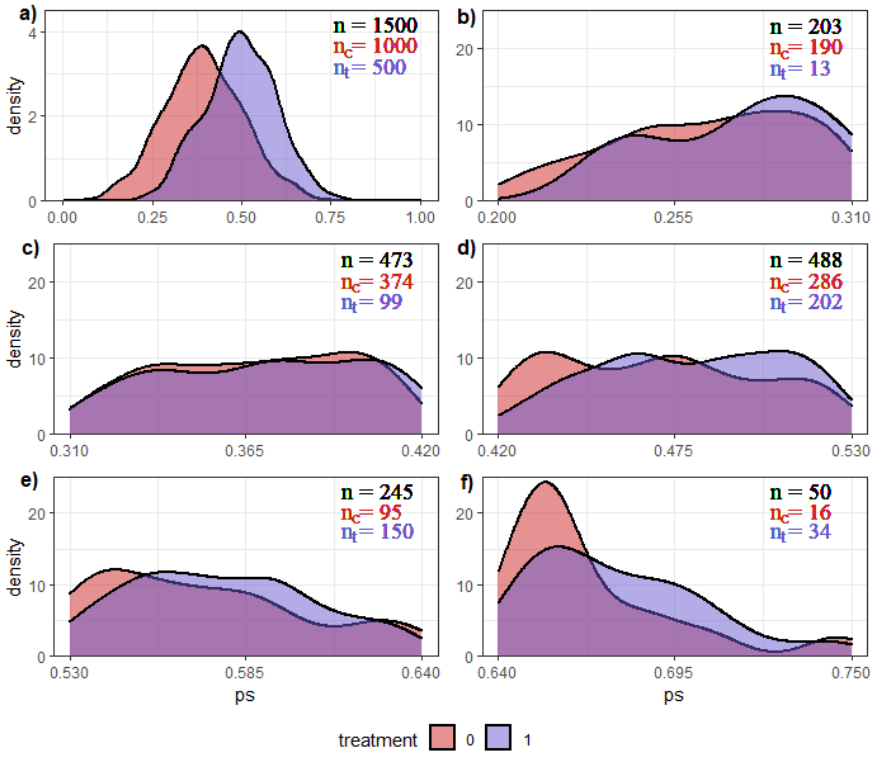

Figure 3: Distributions of propensity scores for treatment and control groups in different strata.
a) Distributions in the entire data set.
b)-f) Distributions in 5 strata of equal length.

3.3 Covariate Adjustment using Propensity Score

In this method, the outcome variable is regressed on the estimated propensity score and an indicator denoting treatment status. Corollary 3.3 follows from Theorem 3.

Corollary 3.3 (Rosenbaum and Rubin 1983)**.** *Suppose treatment assignment is strongly ignorable, so that in particular,* $E(Y_t \mid ps(\boldsymbol{x}), Z = t) = E(Y_t \mid ps(\boldsymbol{x}))$ *for propensity score* ps*. Further suppose that the conditional expectation of* Y *given* $ps(\boldsymbol{x})$ *is linear:*

$$E(Y_t \mid ps(\boldsymbol{x}), Z = t) = \alpha_t + \beta_t ps(\boldsymbol{x}),\ t \in \{0, 1\}.$$

Then the estimator

$$(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)ps(\boldsymbol{x})$$

is conditionally unbiased given $ps(\boldsymbol{x}_i)$ $(i = 1, ..., n)$ *for the treatment effect at* $ps(\boldsymbol{x})$*, namely* $E(Y_1 - Y_0 \mid ps(\boldsymbol{x}))$*, if* $\hat{\alpha}_t$ *and* $\hat{\beta}_t$ *are conditionally unbiased estimators of* α_t *and* β_t*, such as least squares estimators. Moreover,*

$$(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)\hat{ps},$$

where $\hat{ps} = n^{-1} \sum ps(\boldsymbol{x}_i)$*, is unbiased for the average treatment effect if the units in the study are a simple random sample from the population.*

Covariate adjustment using propensity score relies heavily on that the model of the relationship between the propensity score and the outcome is specified correctly. (Austin 2011)

3.4 Inverse Probability of Treatment Weighting

3.4.1 Overview

Propensity score weighting methods use a function of the propensity score to achieve balance in the sample. The populations are reweighted, thus creating a pseudo-population where the treatment assignment and observed covariates are independent. Unlike propensity score matching, weighting keeps most of the units in the analysis, thus offering increased precision in estimates. Several different weighting methods are used, including inverse probability of treatment weighting, fine stratification weighting, standardised mortality ratio weighting, matching weighting, and overlap weighting. (Desai and Franklin 2019)

In this thesis, only inverse probability of treatment weighting (IPTW) will be covered. In IPTW, units are weighted by the inverse probability of receiving the study treatment actually received, i.e.

$$w_i = \frac{z_i}{ps(\boldsymbol{x}_i)} + \frac{1 - z_i}{1 - ps(\boldsymbol{x}_i)} = \begin{cases} \dfrac{1}{ps(\boldsymbol{x}_i)}, & \text{for treated } (z_i = 1), \\ \dfrac{1}{1 - ps(\boldsymbol{x}_i)}, & \text{for controls } (z_i = 0), \end{cases}$$

where z_i and $\boldsymbol{x}_i$ are are the treatment indicator and measured covariates, respectively, for the i-th subject.

After weighting, a subject essentially becomes w subjects in the new, pseudo-population. Since, under the positivity assumption, the propensity score is strictly between zero and one ($0 < ps(\boldsymbol{x}_i) < 1$), then also $0 < 1 - ps(\boldsymbol{x}_i) < 1$, and thus $1 < w_i < \infty$. This means that each subject contributes more than one subject's worth into the pseudo-population after weighting, and the pseudo-population is inevitably larger than the actual population.

This gives unbiased point estimates of average treatment effect, but will most often result in biased standard errors of these point estimates. Thus, stabilised weights are generally preferred, where the weights are calculated as

$$\begin{aligned} sw_i &= \frac{z_i \frac{n_{z=1}}{n}}{ps(\boldsymbol{x}_i)} + \frac{(1 - z_i)\frac{n_{z=0}}{n}}{1 - ps(\boldsymbol{x}_i)} \\ &= \begin{cases} \dfrac{n_{z=1}/n}{ps(\boldsymbol{x}_i)}, & \text{for treated } (z_i = 1), \\ \dfrac{n_{z=0}/n}{1 - ps(\boldsymbol{x}_i)}, & \text{for controls } (z_i = 0), \end{cases} \end{aligned}$$

where z_i and $\boldsymbol{x}_i$ are the treatment indicator and measured covariates for the i-th subject, respectively, $n_{z=1}$ and $n_{z=1}$ are the numbers of treated and control units in the sample, and n is the sample size. This means that instead of simply inverting the treatment probability, we divide the proportion of treated by the subject's propensity score, if the subject is treated, or the proportion of controls by one minus the subject's propensity score, if the subject is a control. (Cole and Hernán 2008)

Extreme weights may occur for subjects that have a very low probability of receiving the treatment they actually received. To prevent variance inflation, weight truncating is often implemented by removing subjects with extreme weights (e.g. smaller than 1st and larger than 99th percentile) from the analysis. The cut-off points are often chosen arbitrarily, but one must keep in mind that while decreasing variance, removing extreme weights might increase bias. (Desai and Franklin 2019)

Variance estimation in regression models requires heteroscedasticity-consistent standard errors, meaning the sample error terms need to be uncorrelated and have constant variance. With weighting, the assumption of constant variance is often not fulfilled, thus resulting in biased variance estimates. Without going into detail, the so-called robust sandwich estimator, also known as

White's (1980) estimator, is used to correct for this flaw. In R, the `sandwich` package (Zeileis et al. 2020; Zeileis 2006) provides the `vcovHC()` command for this purpose.

3.4.2 Example

In Figure 4 on the left, we see another example of propensity score densities for a treatment and a control group. In this case, there are 434 treated subjects and 566 controls in the sample.

After calculating the inverse probability of treatment weights as explained previously, and weighting the data, we get new densities with a much better overlap in the propensity scores. The graph looks exactly the same for both regular and stabilised weights due to the way they are calculated. The only difference is the size of the pseudo-population created by weighting the data. When using regular weights, the size of the pseudo-population would be 1992 in this case, which is almost twice the size of the original data set. With stabilised weights, however, the size remains approximately the same.

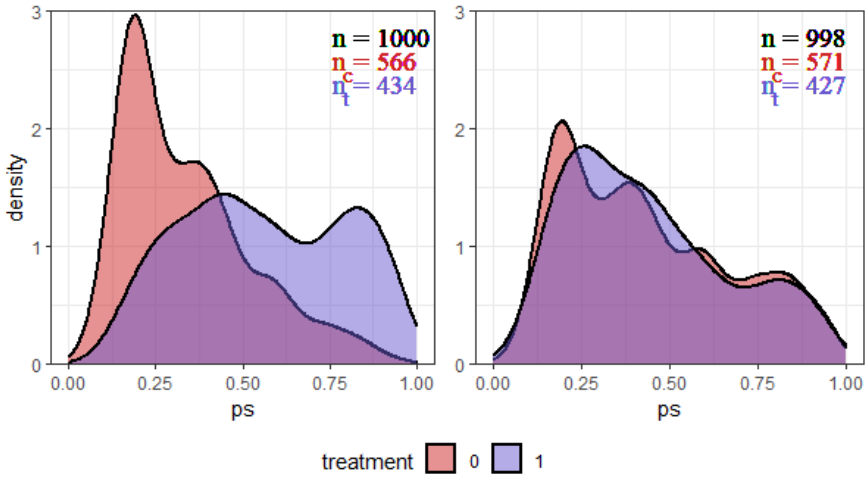

Figure 4: Propensity score distributions in treatment and control group before (left) and after (right) weighting.

3.5 Propensity Score Methods vs. Conventional Covariate Adjustment

One may wonder, why bother with propensity score methods at all. We already have the trustworthy, conventional covariate adjustment, where all relevant covariates are included in a regression model alongside with the treatment when modelling an outcome.

A common concern for covariate adjustment is over-fitting to data when there is a large number of covariates compared to the number of outcome events. As a rule, it is recommended to have at least 10 events per each covariate included in the model. The propensity score reduces the dimensionality of the data, thus also reducing (but not entirely removing) the potential for over-fitting. Propensity score methods also aim to approximate some characteristics of a randomised experiment, making the results easy to comprehend and interpret for practitioners. (Elze et al. 2017)

4 Simulations

The data used in this part are fully simulated using R language (R Core Team 2020) and RStudio software (RStudio Team 2020). Inspiration for a scenario and included variables was obtained from "Seven-day antibiotic courses have similar efficacy to prolonged courses in severe community-acquired pneumonia — a propensity-adjusted analysis" (Choudhury et al. 2011). However, the simulations are only very loosely based on the article and are not expected to give similar results to those presented by Choudhury et al.

The R code for the simulations is available at:
https://github.com/kryzzo/propscore.

4.1 Description of the Baseline Covariates

The population to be studied is all patients admitted to the hospital with severe community-acquired pneumonia in Fakeville, Simulandia.

The treatment of interest is antibiotic courses for 14 days. The control group is people who received antibiotic courses for 7 days. For simplicity we assume that everyone has followed their doctors' orders perfectly.

The outcome to be studied is 30-day mortality within the population, i.e. a patient received the outcome if they died within 30 days of being admitted to the hospital, and did not receive the outcome if they were alive 30 days after the admittance, whether still hospitalised or not.

All baseline covariates, affecting the treatment assignment and/or the outcome, are age, gender, and 5 different comorbidities: congestive cardiac failure, liver disease, diabetes, smoking status, and chronic obstructive pulmonary disease (COPD). They are simulated according to the scheme in Table 3. The data simulated in this manner do not necessarily reflect how such covariates would relate to each other in reality.

For age, first an age group is randomly chosen with the probabilities presented in Table 3, and then an exact age is simulated uniformly within that age group. People over the age of 65 have a much higher probability of having cardiac failure than younger people. All subjects under the age of 18 are non-smokers. COPD is very common in the study population - someone with COPD is more likely to end up in a hospital with severe pneumonia than someone without COPD. Smokers have COPD with twice the proba-

bility of non-smokers (50% vs. 25%). Gender, liver disease and diabetes are independent of the other covariates.

Table 3: Simulation scheme for the baseline covariates.

variable name	**variable description**	**distribution used for simulation**
age	age in years	6% probability to be uniformly in [1, 18) 22% probability to be uniformly in [18, 40) 29% probability to be uniformly in [40, 65) 35% probability to be uniformly in [65, 80) 8% probability to be uniformly in [80, 90)
gender	gender	60% probability to be a man 40% probability to be a woman
smoke	smoking status	0% probability to be a smoker if age <18 20% probability to be a smoker if age $\geq$ 18
cardiac	congestive cardiac failure (CF)	1% probability to have CF if age <65 10% probability to have CF if age $\geq$ 65
COPD	chronic obstructive pulmonary disease	25% probability to have COPD if non-smoker 50% probability to have COPD if smoker
liver	liver disease	5% probability to have disease for all
diab	diabetes mellitus, any type	15% probability to have disease for all

Age is a continuous variable, while all the other covariates are binary. For gender, 0 denotes a woman and 1 denotes a man. For the comorbidities, the variables are indicators: 1 means the patient has the comorbidity, and 0 that they do not. The treatment (variable name "treat") and outcome (variable name "death") will also be denoted with zeros and ones in the same manner.

We are going to consider several different scenarios for simulating treatment and outcome:

1. Randomised trial, where the treatment assignment is independent of all baseline covariates.

2. A scenario where all the baseline covariates are confounders, i.e. all of them affect both treatment and outcome.

3. A more realistic scenario, where some baseline characters are confounders, and some affect only treatment or outcome.

In the following, we describe more precisely how the data sets in these scenarios were simulated, and analyse them.

4.2 Scenario 1: Randomised Trial

4.2.1 Description

Let us start with a simulation of a simple randomised trial, where the treatment assignment is independent of the baseline covariates. Let there be a 40% probability for any study subject to be in the treatment group (antibiotic course of length 14 days) and 60% probability to be in the control group (antibiotic course of length 7 days). The propensity score is thus equal to 0.4, $\quad ps(\boldsymbol{X}) = P(Z = 1 \mid \boldsymbol{X}) = P(Z = 1) = 0.4 \quad \forall \boldsymbol{X}$.

The outcome probability, i.e. the probability to die within 30-days of hospitalisation, is calculated based on a logit-model,

$$p_{out} = \frac{1}{1 + \exp(-m)},$$

where

$$\begin{aligned} m = -3.5 &+ \beta \,\text{treat} + 0.01 \,\text{age} + 0.2 \,\text{cardiac} \\ &+ 0.1 \,\text{COPD} - 0.1 \,\text{diab} + 1 \,\text{smoke}. \end{aligned}$$

Here, β is the expected change in the log odds of the outcome in treatment vs. control group if the other variable values are fixed,

$$\beta = \log(\text{odds}_{Z=1}) - \log(\text{odds}_{Z=0}),$$

$$\text{odds}_{Z=t} = \frac{P(Y = 1 \mid Z = t)}{P(Y = 0 \mid Z = t)} = \frac{P(Y = 1 \mid Z = t)}{1 - P(Y = 1 \mid Z = t)}, \quad t \in \{0, 1\}.$$

For simplicity, we will refer to β as treatment effect throughout this chapter.

Lastly, for each subject, an outcome is randomly generated from a Bernoulli distribution with parameter p_{out}.

We will view two different sub-scenarios: one where treatment has no effect on the outcome ($\beta = 0$) and one where a treated unit is less likely to die within 30 days than a control unit ($\beta = -1$).

4.2.2 Analysis of a Single Data Set

We sampled 1000 individuals from the aforementioned population. There are 387 people in the treatment group and 613 in the control group. A complete

summary of the data can be viewed in Appendix A. Knowing the truth that lies behind the data, we can now estimate the propensity score with logistic regression with all the baseline covariates included, and see if it works the way it is supposed to.

In Figure 5 we see the estimated logit model of the propensity score, i.e.

$$l = \log\left(\frac{ps}{1-ps}\right).$$

The expected model would thus be

$$E(l) = \log\left(\frac{0.4}{1-0.4}\right) \approx -0.405.$$

None of the coefficients in the estimated logit PS model in Figure 5 are statistically significantly different from zero, except for the intercept, which is close to the expected value.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.4793838  0.1982146  -2.419   0.0156 *
gender1     -0.0620243  0.1333605  -0.465   0.6419
age          0.0012450  0.0030614   0.407   0.6842
cardiac1     0.0316846  0.3014423   0.105   0.9163
COPD1       -0.0818905  0.1475286  -0.555   0.5788
liver1      -0.2043553  0.3154339  -0.648   0.5171
diab1        0.1301973  0.1789170   0.728   0.4668
smoke1      -0.0001959  0.1698639  -0.001   0.9991
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5: Estimated logit propensity score model output for a simulated randomised trial where the true PS is 0.4.

The propensity scores are calculated as

$$ps = \frac{1}{1+\exp(-l)}.$$

In Figure 6, we see the propensity score densities for the treatment and control groups. They are overlapping and all very close to 0.4 as expected; the slight differences come only from random sampling.

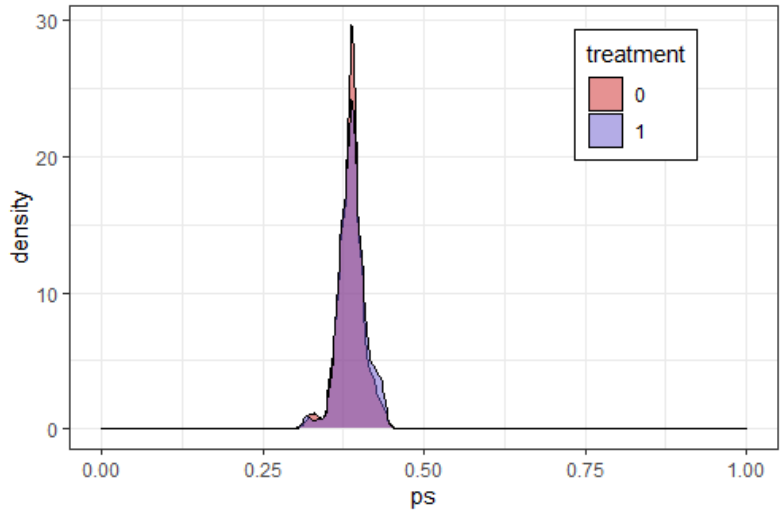

Figure 6: Propensity score distributions for the treated and control units in a simulated randomised trial where the true PS is 0.4

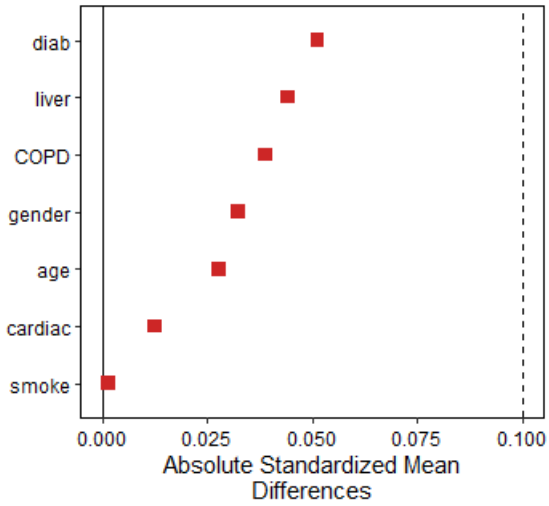

Figure 7: Absolute standardised mean differences between treatment and control group for baseline covariates in a simulated randomised trial.

Since the treatment is generated independently of all the baseline covariates, there should be no imbalances in the covariate distributions between the treatment groups. Of course small imbalances arise from the random sampling. Let us look at the balance plot in Figure 7. It depicts the absolute standardised mean differences in the baseline covariates between treatment and control groups. In practice, variables with an absolute standardised mean difference larger than 0.1 are usually considered imbalanced. Here, we see that no such covariate imbalances are present in our sample, which is also illustrated by the overlapping propensity score distributions in Figure 6.

Although not needed here due to the already balanced covariates, we can also have a look at how matching and weighting based on the propensity score would affect the sample balance.

In PS matching, for each treatment group unit, a control group unit is picked with a similar estimated propensity score. Thus, we create a new data set where we have an equal number of people in each of the two groups. Since in the current data set, there are 387 people in the treatment group, 387 control group subjects are chosen to match them, and therefore 226 people (controls who do not receive a match) are removed from the data set altogether. The changes in the baseline covariate balance and propensity score overlap are minimal, as expected (see Figure 8).

In inverse probability of treatment weighting (IPTW) each unit receives a weight as described in Chapter 3.4. Due to the true propensity score being 0.4, the regular weights should be distributed around $\frac{1}{0.4} = 2.5$ for the treated and $\frac{1}{1-0.4} \approx 1.67$ for the controls. Stabilised weights should have a mean of approximately 1, regardless of the true propensity score. This holds, as can be seen in Figure 9.

Figure 10 shows the balance in propensity score and after weighting. The baseline covariates are near-perfectly balanced here.

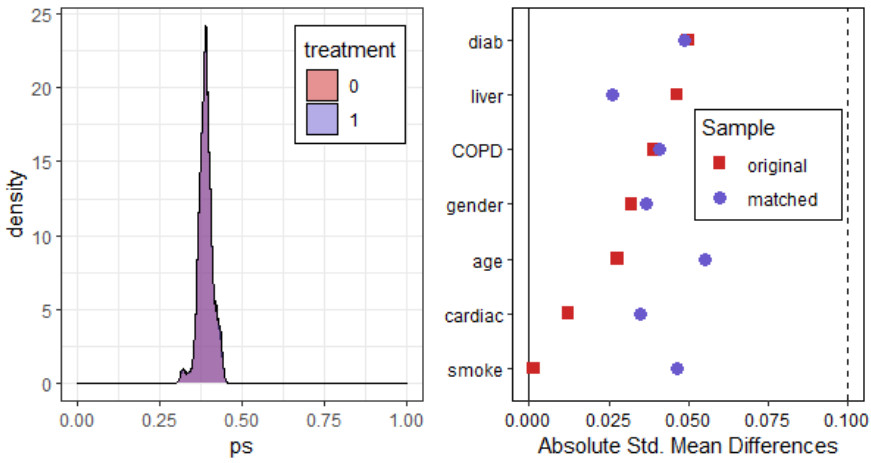

Figure 8: PS distributions (left) and absolute standardised mean differences in baseline covariates (right) between the treatment and control groups after PS matching in a simulated randomised trial.

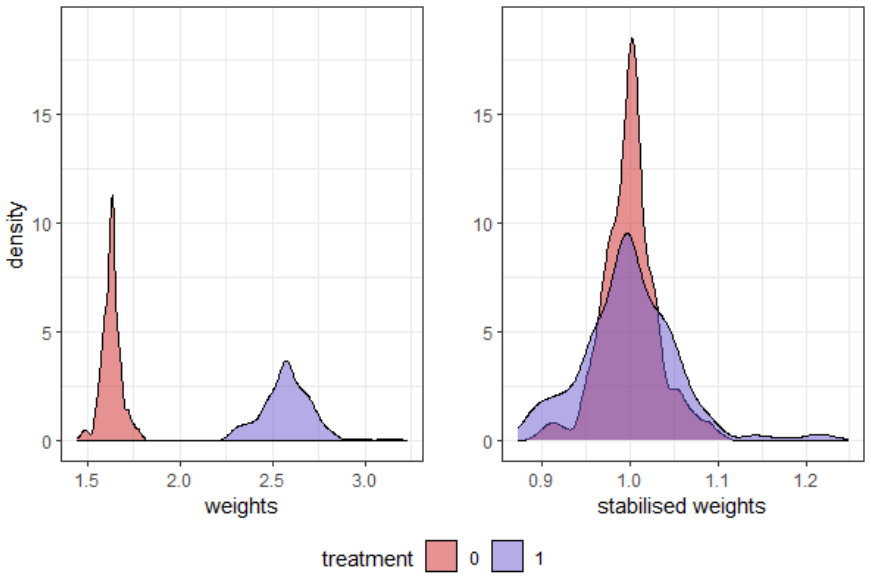

Figure 9: Distributions of weights (left) and stabilised weights (right) in a simulated randomised trial.

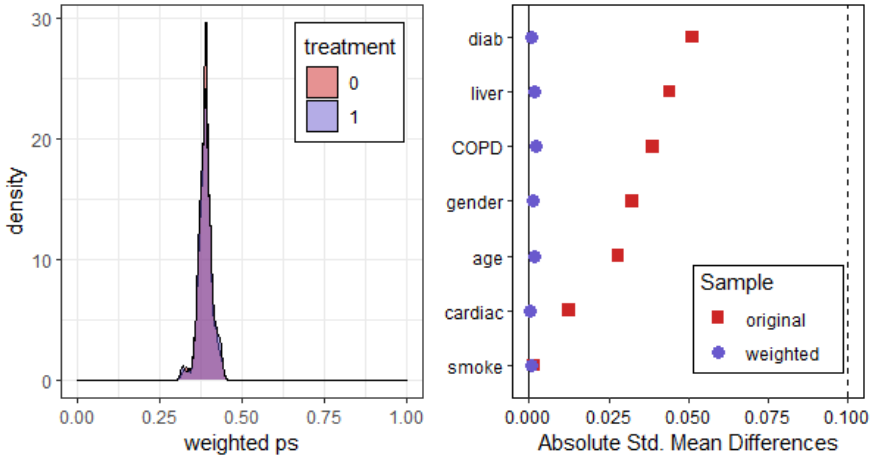

Figure 10: PS distributions (left) and absolute standardised mean differences in baseline covariates (right) between the treatment and control groups after PS weighting in a simulated randomised trial.

As mentioned previously, the outcome was simulated in two different ways: one where $\beta = 0$ and one where $\beta = -1$. The first case means that treatment has no effect on 30-day mortality, and the second case means that for fixed values of all other covariates, the log odds of the treated are one unit smaller than the log odds of the controls.

Table 4: 30-day mortality by treatment

death ($\beta = 0$) **treat**	0	1
0	572	41
1	353	34

death ($\beta = -1$) **treat**	0	1
0	576	37
1	379	8

All the models, for estimating β, here and in the following sections are:

1. logistic regression where treatment is the only included independent variable, all data included (model name in tables: *no adjustment*),

2. logistic regression where treatment and all baseline covariates are included in the model (*all covariates included*),

3. logistic regression on matched data, only treatment included (*matched data*),

4. weighted logistic regression with regular inverse probability of treatment weights, only treatment included (*weights*),

5. weighted logistic regression with stabilised weights, only treatment included (*stabilised weights*),

6. weighted logistic regression with corrected standard error estimate using the sandwich estimator (White 1980) (*corrected standard error for IPTW*).

The last three models will always result in the same point estimate of β, but can have different standard errors of that estimate.

Since we are looking at a (simulated) randomised trial, we can estimate the treatment effect with a simple logistic regression without including any of the baseline covariates in the model. However, we can also see that using covariate adjustment, matching, or weighting does not change the model drastically, as the covariates are balanced between the treatment groups, like demonstrated previously.

For these specific data sets, when the true treatment effect is zero ($\beta = 0$), the estimated treatment effects can be found in Table 5, and when $\beta = -1$, in Table 6. In both cases, the models yield quite similar results, with a slightly wider confidence interval when using matching than in other methods. The estimated coefficients are somewhat different from the true value of β, due to the random sampling, but all the confidence intervals cover the true β.

Complete model outputs can be found in Appendix B.

Table 5: Treatment effect estimates from a simulated randomised trial sample when true $\beta = 0$.

method	**estimated coef. (β)**	**standard error**	**confidence interval (95%)**
no adjustment	0.296	0.242	(-0.178, 0.769)
all covariates included	0.302	0.246	(-0.180, 0.785)
matched data	0.291	0.271	(-0.241, 0.822)
regular weights	0.287	0.239	(-0.181, 0.755)
stabilised weights	0.287	0.242	(-0.187, 0.760)
corrected standard error for IPTW	0.287	0.242	(-0.188, 0.761)

Table 6: Treatment effect estimates from a simulated randomised trial sample when true $\beta = -1$.

method	**estimated coef. (β)**	**standard error**	**confidence interval (95%)**
no adjustment	-1.113	0.396	(-1.888, -0.338)
all covariates included	-1.129	0.401	(-1.915, -0.343)
matched data	-1.000	0.422	(-1.827, -0.173)
regular weights	-1.143	0.371	(-1.870, -0.415)
stabilised weights	-1.143	0.401	(-1.929, -0.356)
corrected standard error for IPTW	-1.143	0.396	(-1.919, -0.366)

4.2.3 Analysis of Repeated Simulations

Now that we have seen an example of one possible sample from the described population, let us repeat this simulation 1000 times to see how much the point estimates and their standard errors vary for each method.

Figures 11 and 12 show violin plots of how these 1000 coefficient estimates and their standard errors, respectively, are distributed for each method when

the true effect is $\beta = 0$. Figures 13 and 14 show similar violin plots when $\beta = -1$.

What we saw in the previously analysed data sets still holds for the 1000 simulations: the point estimates of β are, on average, close to the true value used in the data simulations, and matching gives, on average, less precise estimates (standard errors are higher), for both $\beta = 0$ and $\beta = -1$.

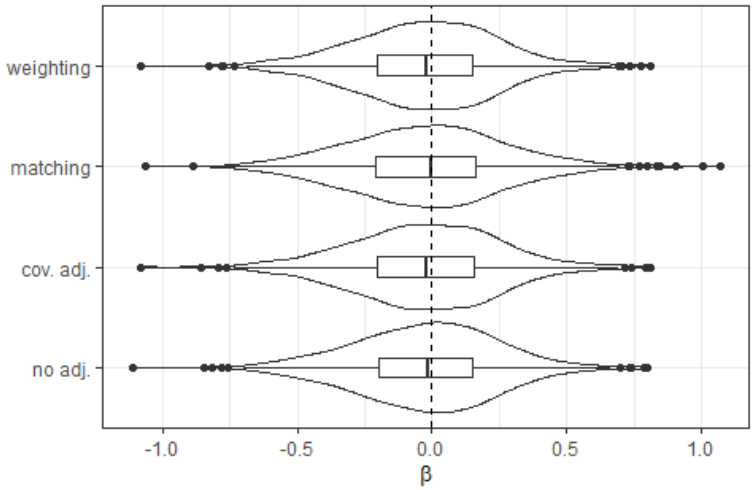

Figure 11: Distribution of point estimates of β for different methods where true $\beta = 0$ in simulated randomised trials.

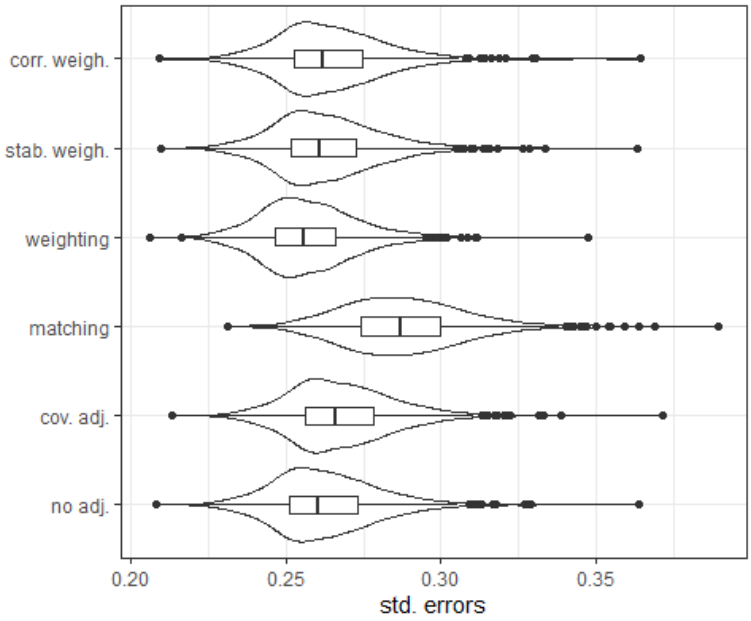

Figure 12: Distribution of standard errors of β estimates for different methods where true $\beta = 0$ in simulated randomised trials.

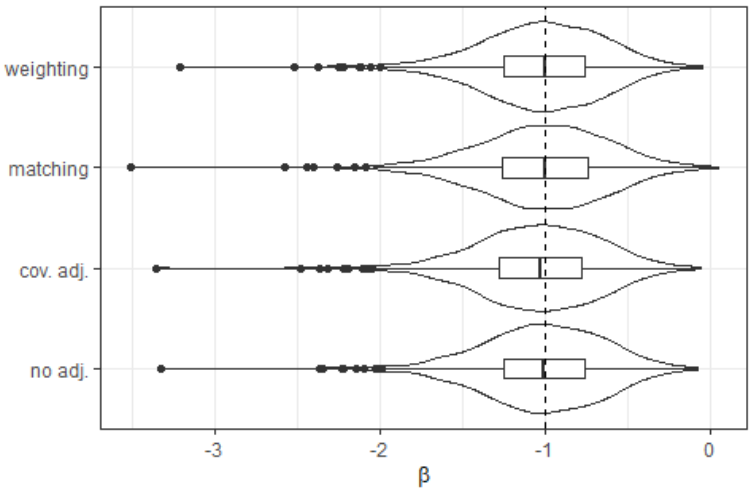

Figure 13: Distribution of point estimates of β for different methods where true $\beta = -1$ in simulated randomised trials.

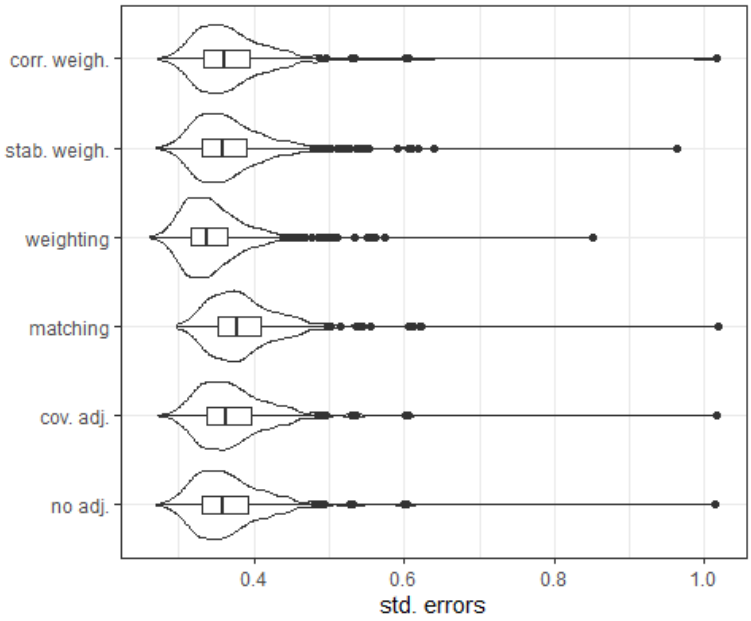

Figure 14: Distribution of standard errors of β estimates for different methods where true $\beta = -1$ in simulated randomised trials.

Lastly, let us have a look at how often the true β lies within the estimated 95% confidence intervals. If the confidence intervals are estimated correctly, then for about 95% of the models, the true β should fall within these bounds. In the case of these simulated randomised trials, this is true for almost all the different models. Regular weighting performs slightly worse than the others, but still gives good enough results. The values for different models are presented in Table 7.

Table 7: Percentage of the 1000 models where the confidence interval (CI) covers the true value of β.

method	**True** β $(\beta = 0)$ **in CI**	**method**	**True** β $(\beta = -1)$ **in CI**
no adjustment	95.1%	no adjustment	95.7%
all covariates included	94.8%	all covariates included	95.7%
matched data	95.2%	matched data	95.6%
regular weights	94.2%	regular weights	94.5%
stabilised weights	94.9%	stabilised weights	95.5%
corrected standard error for IPTW	94.9%	corrected standard error for IPTW	95.5%

4.3 Scenario 2: All Covariates are Confounders

4.3.1 Description

Now that we have seen how the methods behave in a certain simulated randomised trial, let us move on to a slightly more complicated scenario, where all baseline covariates affect both the treatment and the outcome.

Let the treatment logit model be the following:

$$\begin{aligned} m_{tr} = 1 &- 0.02\,\text{age} - 0.2\,\text{gender} - 0.2\,\text{cardiac} - 0.2\,\text{liver} \\ &- 0.2\,\text{COPD} - 0.2\,\text{diab} - 0.2\,\text{smoke} \end{aligned}$$

This means that older people and men (gender = 1) are less likely to be assigned treatment, and each comorbidity the person has, reduces the odds of receiving treatment as well.

After generating the baseline covariates as given in Table 3, the probability of being assigned treatment ($Z = 1$) is calculated for each subject as

$$p_{tr} = \frac{1}{1 + \exp(-m_{tr})}.$$

Then, a treatment is randomly sampled from a Bernoulli distribution with probability p_{tr} for each unit in the sample.

Let the outcome probability be calculated as

$$p_{out} = \frac{1}{1 + \exp(-m_{out})},$$

where

$$\begin{aligned} m_{out} = -2 + \beta\,\text{treat} &+ 0.01\,\text{age} + 0.1\,\text{gender} + 0.1\,\text{cardiac} + 0.1\,\text{liver} \\ &+ 0.1\,\text{COPD} + 0.1\,\text{diab} + 0.1\,\text{smoke}, \end{aligned}$$

i.e. the odds of dying within 30 days of hospitalisation are bigger for older people and men, and each comorbidity raises the odds as well.

Lastly, an outcome is sampled from a Bernoulli distribution with probability p_{out}.

4.3.2 Analysis of a Single Data Set

Again, we sampled 1000 individuals according to the aforementioned scheme. There are 409 treated and 591 control units in this sample. A complete summary of the data is given in Appendix A.

Let us estimate the propensity score with logistic regression. Ideally, we would see the exact value of m_{tr} for each patient, but since we have randomly sampled units, we would expect just something similar. Indeed, this is the case, as shown in Figure 15. Cardiac failure is the only comorbidity that has been estimated to have a rather different coefficient than the true one (-0.98 instead of -0.2), possibly due to the fact that cardiac failure is rare in our sample - only 44 people out of a 1000 have it. We will continue the analysis with this propensity score, as we would in practice, assuming it did not contradict any expert knowledge at hand.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.020984   0.196353   5.200  2.0e-07 ***
gender1     -0.112915   0.135557  -0.833   0.4049
age         -0.023170   0.003077  -7.529  5.1e-14 ***
cardiac1    -0.979387   0.425723  -2.301   0.0214 *
COPD1        0.102282   0.151275   0.676   0.4990
liver1      -0.160141   0.307062  -0.522   0.6020
diab1       -0.196207   0.203538  -0.964   0.3351
smoke1      -0.268593   0.176981  -1.518   0.1291
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 15: Estimated logit propensity score model output.

In Figure 16, the propensity score densities in the treatment and control groups are depicted. Unlike in the randomised trial, here the distributions are quite different from each other, and a simple logistic regression with treatment as its only independent variable would likely not yield correct results.

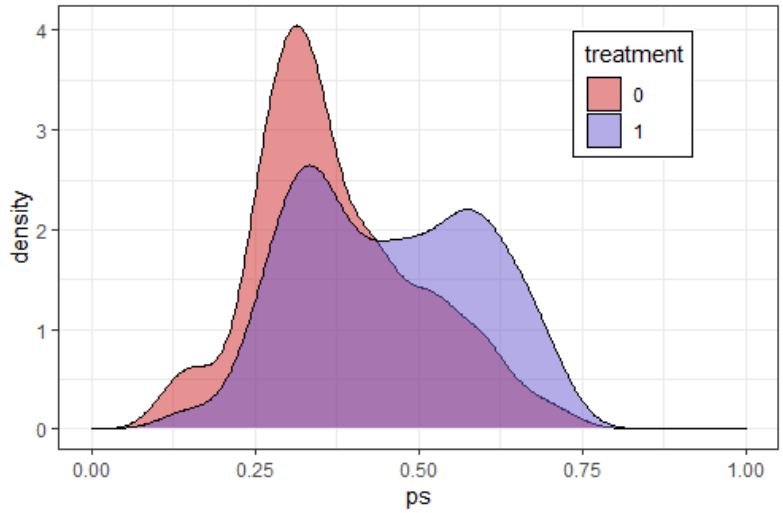

Figure 16: Propensity score distributions for the treated and control units.

We perform one-on-one nearest-neighbour matching on the data, in the hopes of balancing the baseline covariates between treatment groups. This means we match 409 control units to the treated units, and discard 182 people from the original data. However, as can be seen in Figure 17, even in such a matched data set, some imbalance remains. Specifically, the absolute standardised mean difference in age is imbalanced in the original data set, and remains so in the matched data (see Figure 17).

In hopes of a better balance, let us match the data with a smaller caliper - instead of a nearest neighbour, let us look for a match only within a certain range of the treated unit's propensity score. This means that some treated units may be left without a match, if no control units with a similar enough PS exist. In this case we pick 0.1 standard deviances of the PS to be the caliper. This leaves us with a data set of size 710 (355 treated and 355 controls). Figure 18 shows that this gives us a better balance in baseline covariates and the propensity score, at the cost of more than one fourth of the initial data set.

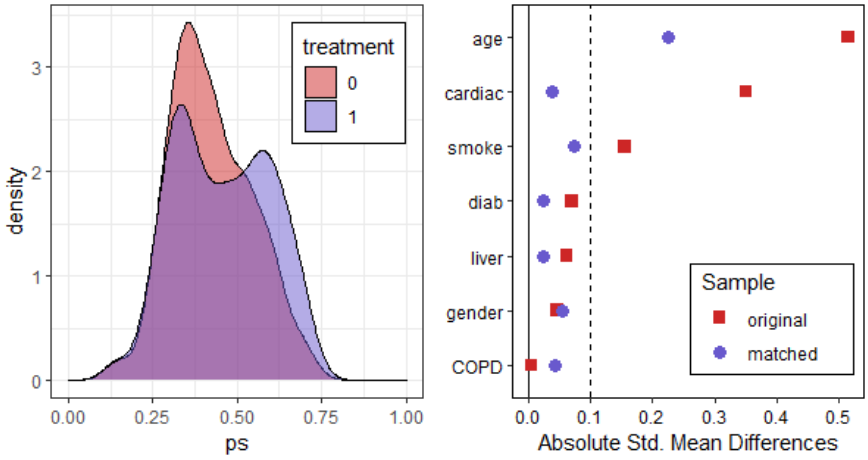

Figure 17: PS distributions (left) and absolute standardised mean differences in baseline covariates (right) between the treatment and control groups after PS matching. Not a good match.

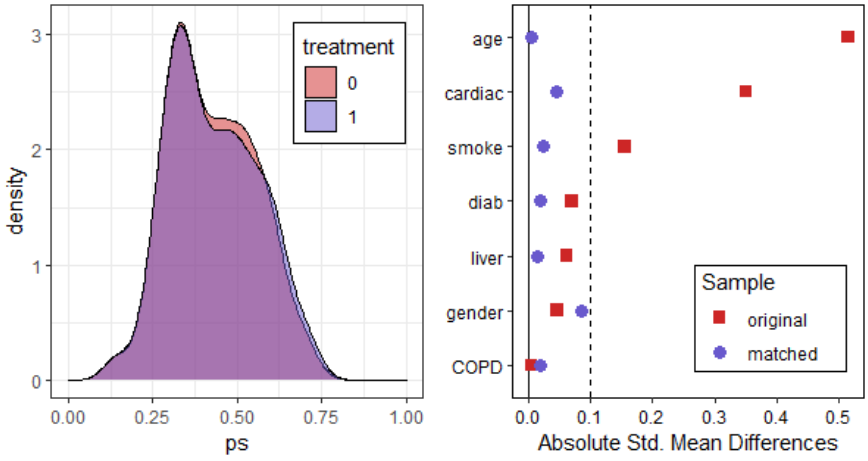

Figure 18: PS distributions (left) and absolute standardised mean differences in baseline covariates (right) between the treatment and control groups after PS matching with a smaller caliper.

For IPTW, the distributions of the weights are shown in Figure 19. The regular weights have a mean value of 2 (2.46 for treated and 1.69 for controls).

The mean values of the stabilised weights are approximately 1 for each of the treatment groups as well as the whole sample.

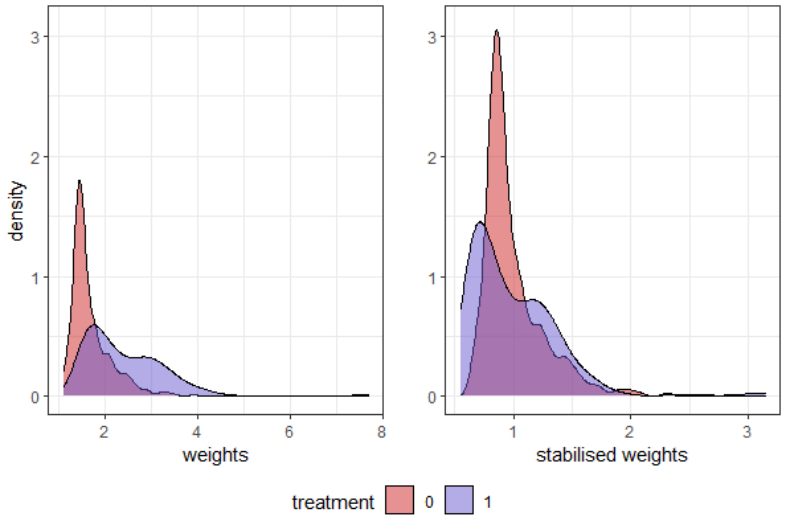

Figure 19: Distributions of weights (left) and stabilised weights (right).

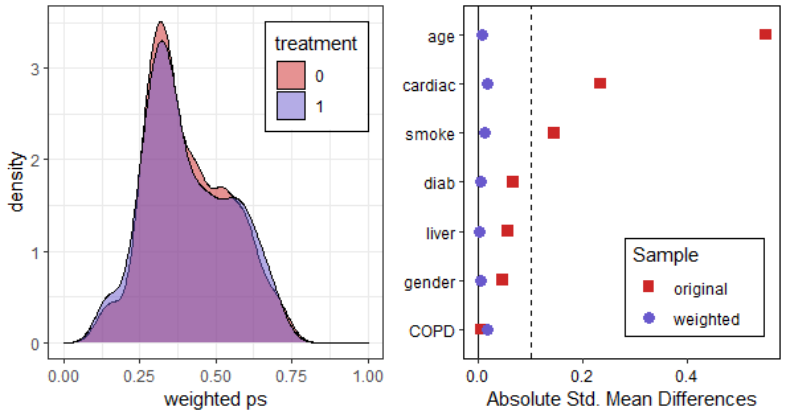

Figure 20: PS distributions (left) and absolute standardised mean differences in baseline covariates (right) between the treatment and control groups after PS weighting.

With weighting, a good balance is achieved in the data, as illustrated by Figure 20.

Again, for 30-day mortality we cover two cases, one of which has $\beta = 0$ and the other $\beta = -1$. Table 8 shows the distribution of the outcome by treatment groups.

Table 8: 30-day mortality by treatment

death ($\beta = 0$) **treat**	0	1
0	460	131
1	328	81

death ($\beta = -1$) **treat**	0	1
0	460	131
1	368	41

Tables 9 and 10 show the estimates of β with different methods in these specific simulated data sets. Complete model outputs are available in Appendix B.

Table 9: Treatment effect estimates when true $\beta = 0$.

method	**estimated coef. (β)**	**standard error**	**confidence interval (95%)**
no adjustment	-0.143	0.159	(-0.454, 0.169)
all covariates included	0.034	0.166	(-0.292, 0.361)
matched data	-0.068	0.184	(-0.428 0.293)
regular weights	0.028	0.155	(-0.275, 0.331)
stabilised weights	0.028	0.157	(-0.280, 0.336)
corrected standard error for IPTW	0.028	0.167	(-0.300, 0.356)

Table 10: Treatment effect estimates when true $\beta = -1$.

method	**estimated coef. (β)**	**standard error**	**confidence interval (95%)**
no adjustment	-0.938	0.192	(-1.315, -0.562)
all covariates included	-0.876	0.198	(-1.264, -0.487)
matched data	-0.971	0.203	(-1.368, -0.574)
regular weights	-0.821	0.181	(-1.174, -0.467)
stabilised weights	-0.821	0.188	(-1.189, -0.452)
corrected standard error for IPTW	-0.821	0.206	(-1.224, -0.418)

All the methods seem to perform relatively well, as the true value of β is in all the confidence intervals. This time, the logistic regression with only treatment as an independent variable gives a different estimate than the other methods. Matched data has a larger standard error than the other methods again. In the following, we study whether these differences are systematic.

4.3.3 Analysis of Repeated Simulations

Let us repeat this simulation 1000 times to see how much the point estimates and their standard errors vary for each method. Figures 21 and 22 show violin plots with the distributions of the estimated treatment effects and standard errors, respectively, when the true value of β is zero. Figures 23 and 24 show the same plots when $\beta = -1$.

It is clear that in both cases, the model with no adjustment for the covariates gives biased estimates of β, while the other methods, on average, work well. Again, in matching we get visibly larger standard errors than in other methods, which can be explained by the smaller sample size. Weighting without any correction, on the other hand, gives smaller standard errors, because the pseudo-sample that we get, with weights that average at 2, is two times larger than the original.

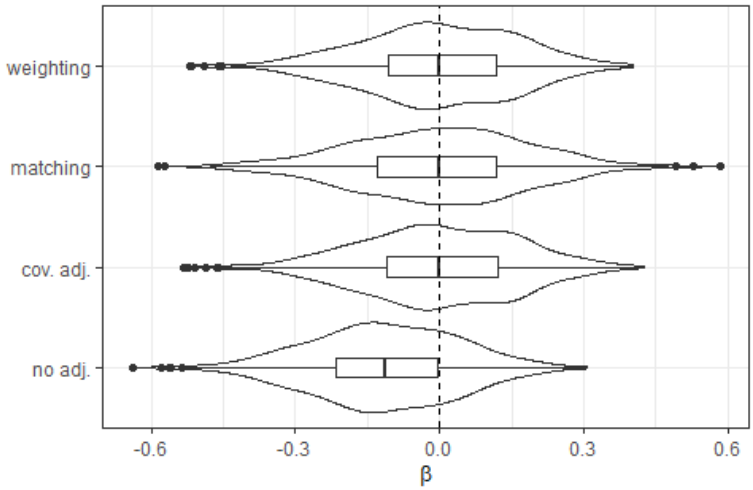

Figure 21: Distribution of point estimates of β for different methods where true $\beta = 0$.

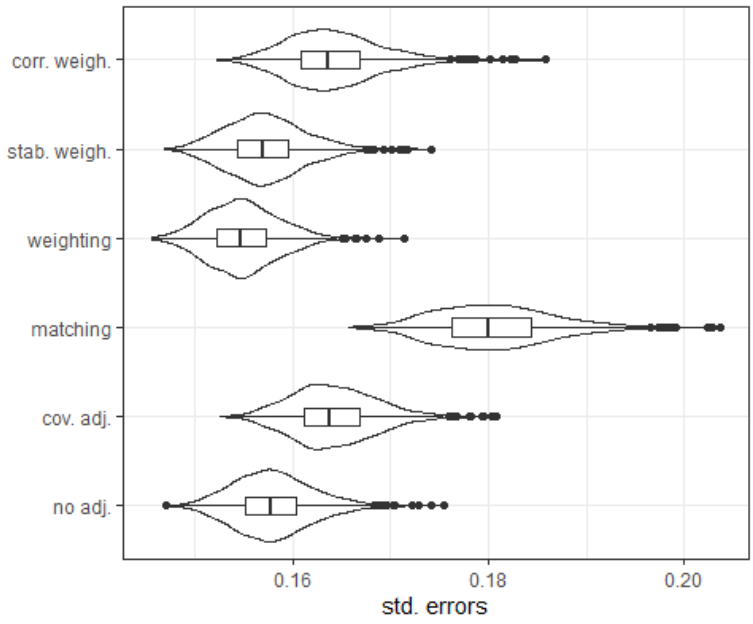

Figure 22: Distribution of standard errors of β estimates for different methods where true $\beta = 0$.

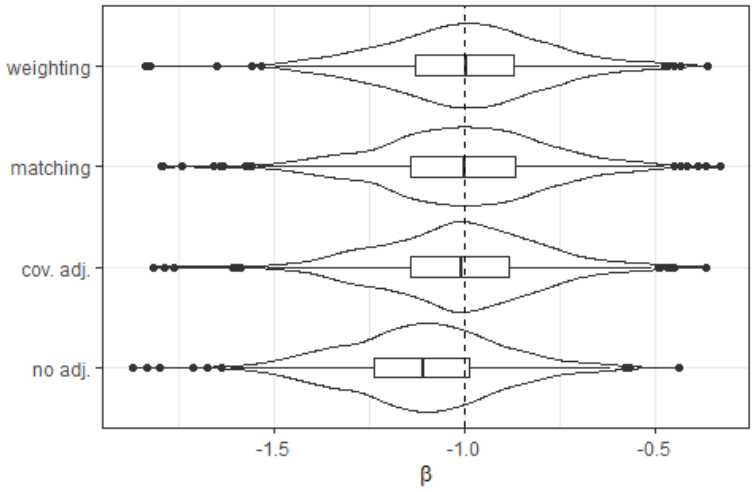

Figure 23: Distribution of point estimates of β for different methods where true $\beta = -1$.

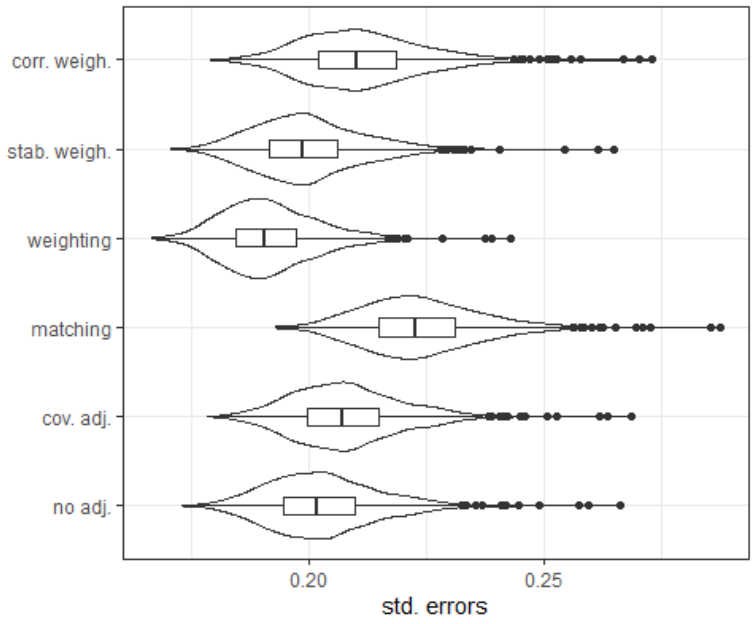

Figure 24: Distribution of standard errors of β estimates for different methods where true $\beta = -1$.

As for the confidence intervals, for all methods (except no adjustment) they cover the true value of β in approximately 95% of the simulations when treatment has no effect on the outcome. When the true β is equal to -1, regular weighting gives too small confidence intervals, that cover the true value only in less than 93% of the cases.

Table 11: Percentage of the 1000 models where the confidence interval (CI) covers the true value of β.

method	**True** β $(\beta = 0)$ **in CI**	**method**	**True** β $(\beta = -1)$ **in CI**
no adjustment	89.4%	no adjustment	92.9%
all covariates included	95.9%	all covariates included	95.1%
matched data	94.9%	matched data	95.6%
regular weights	94.9%	regular weights	92.8%
stabilised weights	95.2%	stabilised weights	94.3%
corrected standard error for IPTW	96.2%	corrected standard error for IPTW	95.1%

In conclusion, for simulated scenario 2, all adjustment methods gave unbiased estimates. Matching gave the largest standard errors, while weighting without any correction resulted in the smallest. Compared to covariate adjustment, weighting, with sandwich estimator-corrected standard errors, gave the most similar results.

4.4 Scenario 3: A More Realistic Case

4.4.1 Description

Now we will consider a more realistic scenario where some baseline covariates are confounders, while some affect only treatment or outcome.

This time, let the treatment logit model be

$$\begin{aligned} m_{tr} = 0.5 - 0.02\,\text{age} + 0.2\,\text{gender} + 1\,\text{liver} \\ + 1\,\text{COPD} - 2\,\text{smoke} \end{aligned}$$

If all other covariates are fixed, then older people, women, and smokers are less likely to be assigned antibiotic courses of 14 days ($Z = 1$), while people with liver disease or COPD are more likely to be prescribed the longer courses.

After generating the baseline covariates as given in Table 3, the probability of being assigned treatment ($Z = 1$) is calculated for each subject as

$$p_{tr} = \frac{1}{1 + \exp(-m_{tr})}.$$

Then, a treatment is randomly sampled from a Bernoulli distribution with probability p_{tr} for each unit in the sample.

The outcome probability will be calculated as

$$p_{out} = \frac{1}{1 + \exp(-m_{out})},$$

where

$$\begin{aligned} m_{out} = -3.5 + \beta\,\text{treat} + 0.02\,\text{age} + 0.2\,\text{cardiac} \\ + 2\,\text{smoke} - 0.1\,\text{diab}, \end{aligned}$$

i.e. the odds of dying within 30 days of hospitalisation are bigger for older people, smokers, and those with cardiac failure, while diabetics are less likely to die.

Lastly, an outcome is sampled from a Bernoulli distribution with probability p_{out}.

4.4.2 Analysis of a Single Data Set

As in the previous scenarios, we sampled 1000 people as described. There are 404 treated and 596 control subjects in the sample. A complete summary of the data set can be seen in Appendix A.

Once again, we estimate the propensity score with logistic regression. The estimated coefficients are given in Figure 25. The estimations look to be close to the true values used in the data simulation, and the ones that are not in the model m_{tr}, are not statistically significantly different from zero in the estimated model.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.407365   0.201705   2.020  0.04342 *
gender1      0.324484   0.145223   2.234  0.02546 *
age         -0.019793   0.003278  -6.038 1.56e-09 ***
cardiac1     0.272879   0.342244   0.797  0.42526
COPD1        1.299663   0.168538   7.711 1.24e-14 ***
liver1      -1.268941   0.427707  -2.967  0.00301 **
diab1        0.068451   0.188448   0.363  0.71643
smoke1      -2.252578   0.262588  -8.578  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 25: Estimated logit propensity score model output.

The propensity score densities in treatment and control group shown in Figure 26 are not similar and, once again, need balancing.

Just like in the previous scenario, simple nearest-neighbour matching, without restricting the distance, does not give a very good match, and the imbalances remain (see Figure 27). Like before, we choose the caliper to be 0.1 standard deviances of the PS to get a better match. When matching like that, only 628 people remain in the new data set, meaning that we have removed 90 people from the treatment group and 282 people from the control group. However, the new data set is nicely balanced, as can be seen in Figure 28.

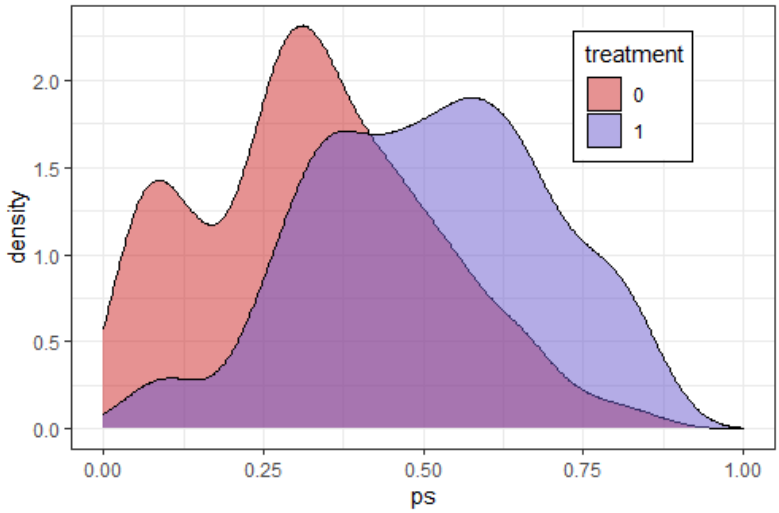

Figure 26: Propensity score distributions for the treated and control units.

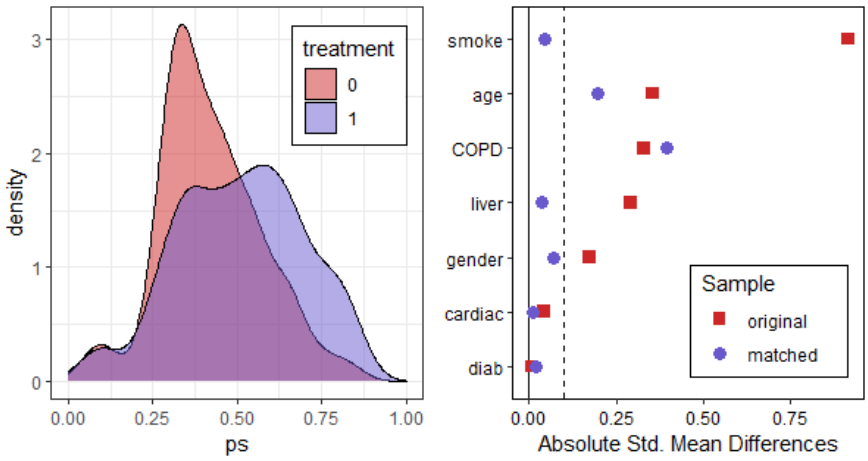

Figure 27: PS distributions (left) and absolute standardised mean differences in baseline covariates (right) between the treatment and control groups after PS matching. Not a good match.

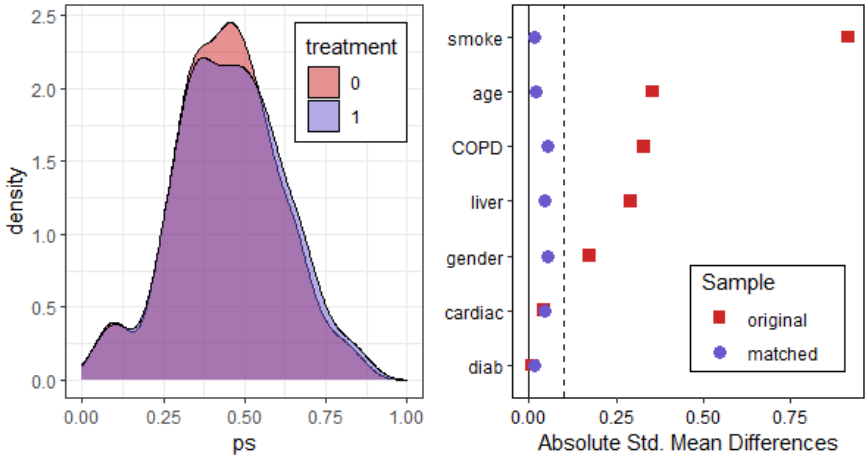

Figure 28: PS distributions (left) and absolute standardised mean differences in baseline covariates (right) between the treatment and control groups after PS matching with a smaller caliper.

When dealing with IPTW, some large weights occur, as can be seen in Figure 29. We will now consider the weighted models for two different cases: one where all data is included (the extreme weights remain), and one where we have trimmed weights, i.e. removed the data points with weights larger than 10 (Figure 30). This means that in addition to the previously listed 6 models, in this case we will estimate 3 more: regular weighting after trimming, stabilised weighting after trimming, as well as one with corrected standard error.

Trimming the data removed 10 observations in this case. Some imbalances remain in the baseline covariates in both trimmed and non-trimmed weighting. In the data set with trimmed weights, smoking is slightly more out of balance than in the data that includes the larger weights (see Figures 31 and 32).

The two different cases of treatment effect give us the outcome distributions by treatment group shown in Table 12.

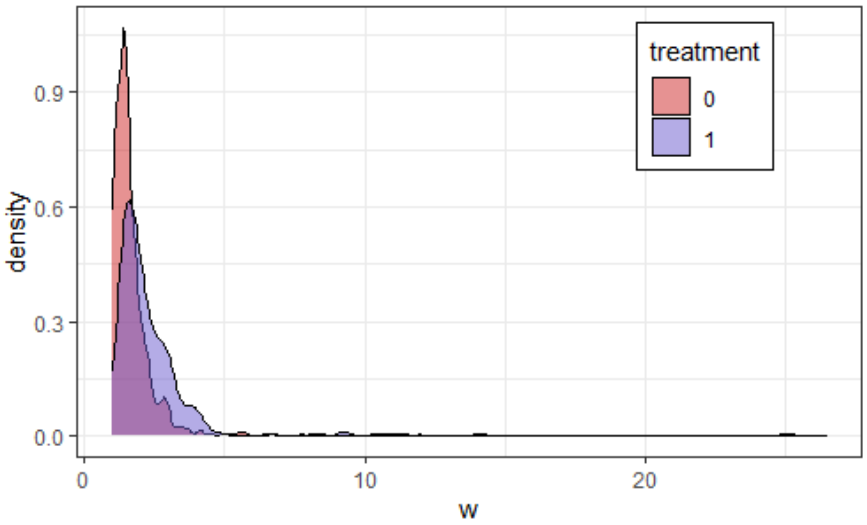

Figure 29: Distribution of weights.

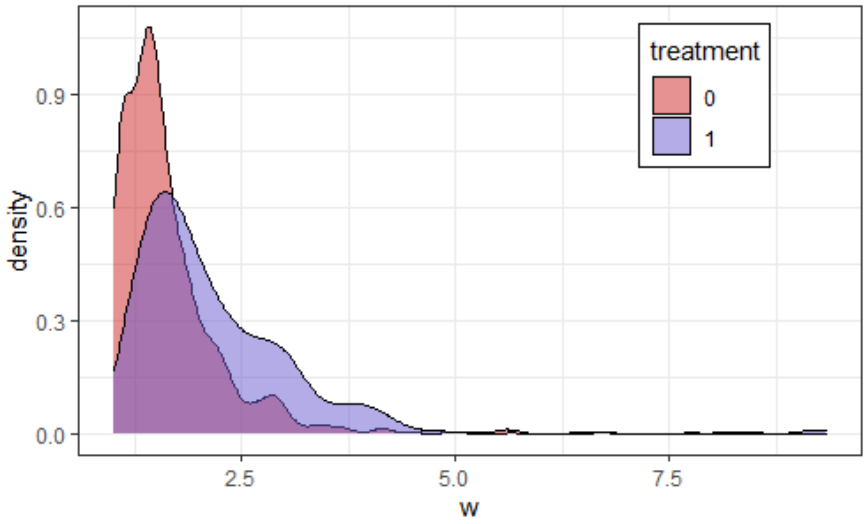

Figure 30: Distribution of weights (trimmed).

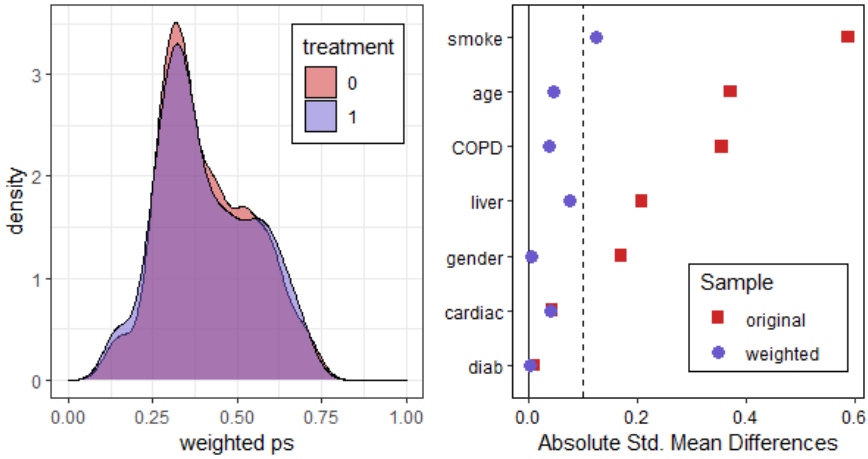

Figure 31: PS distributions (left) and absolute standardised mean differences in baseline covariates (right) between the treatment and control groups after PS weighting.

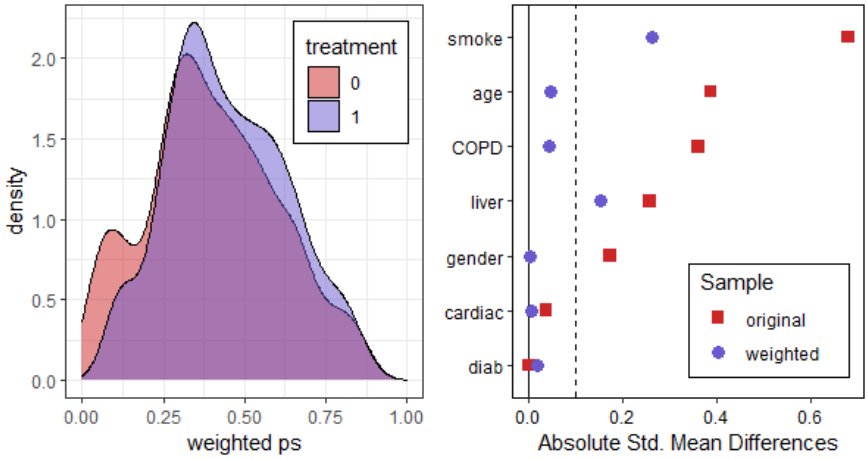

Figure 32: PS distributions (left) and absolute standardised mean differences in baseline covariates (right) between the treatment and control groups after PS weighting (trimmed).

Table 12: 30-day mortality by treatment

death ($\beta = 0$) **treat**	0	1
0	485	111
1	363	41

death ($\beta = -1$) **treat**	0	1
0	489	107
1	387	17

Tables 13 and 14 show the β estimates and their confidence intervals when estimated with different models. Expectedly, the model with no covariate or propensity score adjustment gives an inaccurate estimate for the treatment effect, and the true β is not covered by the confidence intervals. The confidence intervals from the weighting methods without trimming also do not cover the true value if $\beta = -1$. The standard error of the model coefficient from matched data is once again the largest.

Table 13: Treatment effect estimates when true $\beta = 0$.

method	**estimated coef. (β)**	**standard error**	**confidence interval (95%)**
no adjustment	-0.706	0.196	(-1.089, -0.323)
all covariates included	-0.039	0.226	(-0.482, 0.403)
matched data	0.068	0.260	(-0.442, 0.578)
regular weights	-0.262	0.188	(-0.631, 0.106)
stabilised weights	-0.262	0.192	(-0.640, 0.115)
corrected standard error for IPTW	-0.262	0.238	(-0.728, 0.204)
regular weights (trimmed)	-0.287	0.191	(-0.661, 0.087)
stabilised weights (trimmed)	-0.287	0.199	(-0.677, 0.103)
corrected standard error for IPTW (trimmed)	-0.287	0.225	(-0.727, 0.153)

Table 14: Treatment effect estimates when true $\beta = -1$.

method	**estimated coef. (β)**	**standard error**	**confidence interval (95%)**
no adjustment	-1.606	0.270	(-2.134, -1.077)
all covariates included	-1.132	0.289	(-1.698, -0.566)
matched data	-1.049	0.349	(-1.732, -0.366)
regular weights	-1.604	0.279	(-2.150 -1.057)
stabilised weights	-1.604	0.299	(-2.189, -1.018)
corrected standard error for IPTW	-1.604	0.291	(-2.173, -1.034)
regular weights (trimmed)	-1.431	0.270	(-1.959, -0.902)
stabilised weights (trimmed)	-1.431	0.292	(-2.002, -0.859)
corrected standard error for IPTW (trimmed)	-1.431	0.287	(-1.994, -0.867)

4.4.3 Analysis of Repeated Simulations

We repeat the previously described simulation 1000 times to try and identify some patterns. Figures 33 and 35 show violin plots with the distribution of the estimated effects and standard errors, respectively, for when $\beta = 0$. Figures 34 and 36 depict the same for when $\beta = -1$.

While our single simulation showed better results with trimmed weights, the 1000 simulations show that, for data generated in this manner at least, the trimmed weights give us biased estimates. Covariance adjustment, matching, and weighting without trimming, however, give less biased or even unbiased results. The largest standard errors come from matching and corrected weighting estimates.

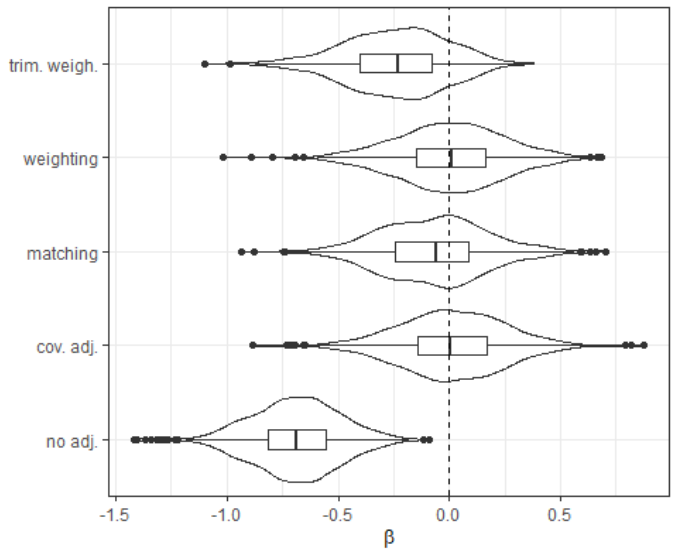

Figure 33: Distribution of point estimates of β for different methods where true $\beta = 0$.

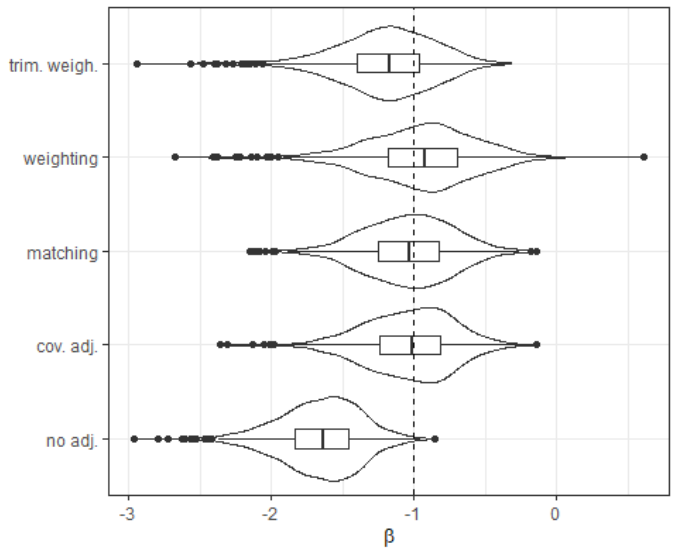

Figure 34: Distribution of point estimates of β for different methods where true $\beta = -1$.

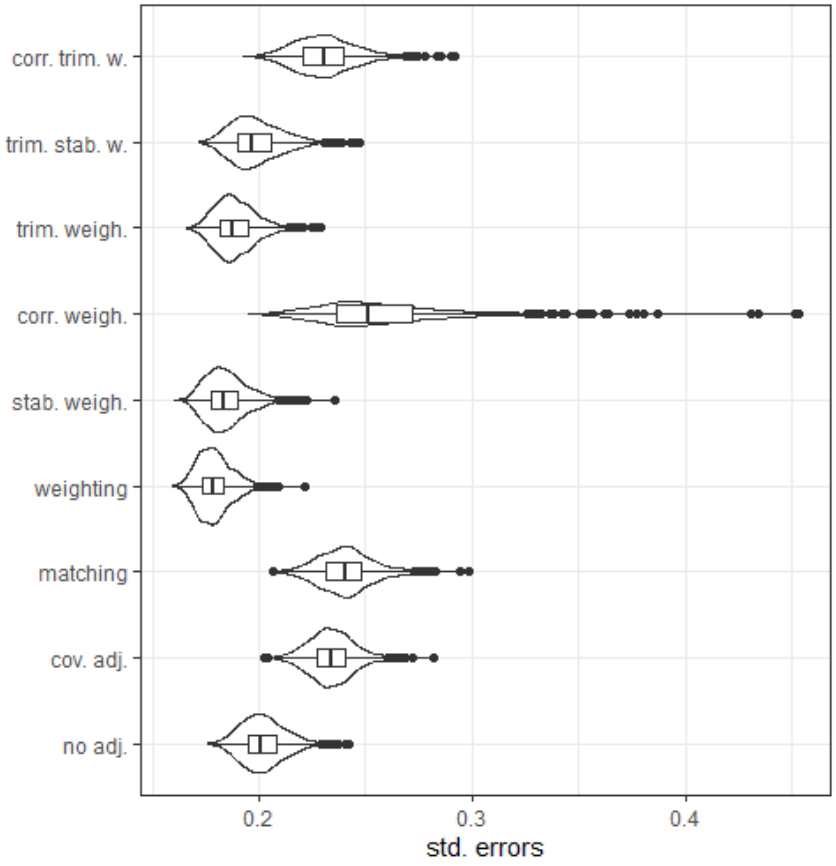

Figure 35: Distribution of standard errors of β estimates for different methods where true $\beta = 0$.

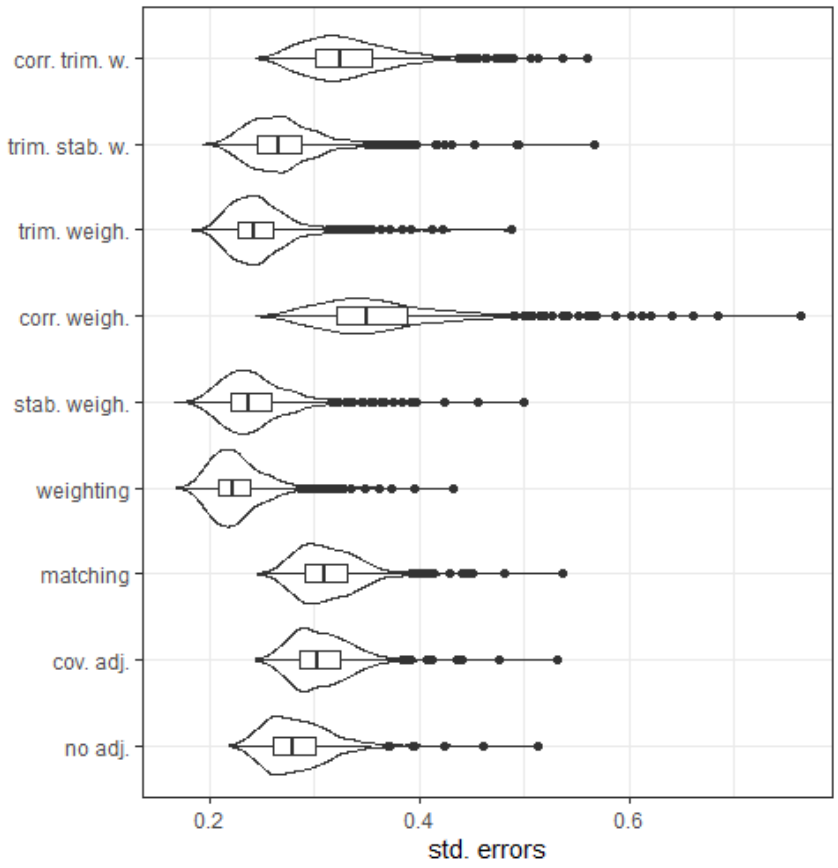

Figure 36: Distribution of standard errors of β estimates for different methods where true $\beta = -1$.

In this scenario, the only methods that give us true 95% confidence intervals are conventional covariate adjustment, matching, and IPTW with corrected standard errors (without trimming), as shown in Table 15.

Table 15: Percentage of the 1000 models where the confidence interval (CI) covers the true value of β.

method	**True** β $(\beta = 0)$ **in CI**
no adjustment	5.9%
all covariates included	95.3%
matched data	94.5%
regular weights	86.6%
stabilised weights	87.4%
corrected standard error for IPTW	96.5%
regular weights (trimmed)	71.8%
stabilised weights (trimmed)	76.0%
corrected standard error for IPTW (trimmed)	84.3%

method	**True** β $(\beta = -1)$ **in CI**
no adjustment	34.1%
all covariates included	95.7%
matched data	95.6%
regular weights	75.8%
stabilised weights	79.8%
corrected standard error for IPTW	94.6%
regular weights (trimmed)	82.8%
stabilised weights (trimmed)	86.3%
corrected standard error for IPTW (trimmed)	93.1%

In conclusion, for simulated scenario 3, covariate adjustment was unbiased and with good standard error estimates (95% confidence interval covered the true β value in 95% of cases). Matching gave the largest standard errors again, while weighting without any correction resulted in the smallest variance. Weight trimming, while making standard errors more similar to those of conventional covariate adjustment, resulted in biased estimation of β, on average.

4.5 Discussion

We simulated data sets of a population of patients admitted to the hospital with severe community-acquired pneumonia. The treatment we are interested in is antibiotic courses of length 14 days. The control group is those with antibiotic courses of exactly 7 days. The outcome of interest is 30-day mortality, i.e. whether the patient dies within 30 days of admittance to the hospital. Included baseline covariates are age, gender, congestive cardiac failure, liver disease, diabetes, chronic obstructive pulmonary disease (COPD), and smoking.

We investigated three scenarios: a randomised trial, a case where all covariates are confounders, and a more realistic case with some covariates as confounders while others only affect treatment or outcome. In all these scenarios we applied propensity score matching and weighting, as well as regular covariate adjustment. All the methods worked in accordance with what was described in Chapter 3.

Since matching removes a part of the available data, then the smaller sample size causes larger standard errors, but the estimated treatment effects are unbiased. Matching, while very intuitive, is often criticised for its tendency to remove a very large part of the data, especially in observational studies where more control data is available than treatment data.

Weighting, when standard errors are corrected using the sandwich estimator, provides unbiased estimates and reasonable standard errors if there are no extreme weights. In case of small treatment probabilities, large weights occur and cause an inflation of standard errors. To counteract this, weights are trimmed by removing observations with very large weights. This, in its turn, causes biased estimations of the treatment effect. This bias-variance trade-off must be taken into account when dealing with extreme weights.

Covariate adjustment provides unbiased and stable estimates. The only issue that may arise, is when there are too few outcome events for the amount of covariates in the model, in which case a regression model cannot be fitted or will be severely over-fitted. In that case propensity score methods can be a good alternative.

Conclusion

Propensity score methods are one way to balance data for causal effect estimations in observational studies, and these methods are becoming increasingly common in the medical field. They include matching, stratification, weighting, and covariate adjustment using propensity score. This thesis covered the theory behind these methods, and applied the matching and weighting in a simulation study. The simulation part also compared these two methods to the conventional covariate adjustment.

All the applied methods worked well in the scenarios implemented here. Matching resulted in larger standard errors of the estimated treatment effects than other methods due to the smaller sample size. If very large weights occur in the weighting methods it can result in an inflation of the standard errors. This can be combated by trimming the weights, i.e. removing observations with extreme weights. However, this in its turn causes biased estimates of the treatment effect, so the bias-variance trade-off needs to be taken into account when using this method. Conventional covariate adjustment was unbiased and stable in all the implemented cases.

While propensity score methods provide a great overview of balance in the baseline covariates and help mimic an RCT-like scenario, covariate adjustment remains a reliable method for causal effect estimation in observational studies.

References

Austin, Peter (2011). "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies". In: *Multivariate Behavioral Research* 46.3, pp. 399–424. DOI: 10.1080/00273171.2011.568786.

Choudhury, Gourab, Pallavi Bedi, Aran Singanayagam, A.R. Akram, James Chalmers, and Adam Hill (Apr. 2011). "Seven-day antibiotic courses have similar efficacy to prolonged courses in severe community-acquired pneumonia — a propensity-adjusted analysis". In: *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 17, pp. 1852–8. DOI: 10.1111/j.1469-0691.2011.03542.x.

Cole, Stephen and Miguel Hernán (Oct. 2008). "Constructing Inverse Probability Weights for Marginal Structural Models". In: *American journal of epidemiology* 168, pp. 656–64. DOI: 10.1093/aje/kwn164.

Desai, Rishi and Jessica Franklin (Oct. 2019). "Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: A primer for practitioners". In: *BMJ* 367. DOI: 10.1136/bmj.l5657.

Elze, Markus, John Gregson, Usman Baber, Elizabeth Williamson, Samantha Sartori, Roxana Mehran, Melissa Nichols, Gregg Stone, and Stuart Pocock (Jan. 2017). "Comparison of Propensity Score Methods and Covariate Adjustment". In: *Journal of the American College of Cardiology* 69, pp. 345–357. DOI: 10.1016/j.jacc.2016.10.060.

Hernán, Miguel and James Robins (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart (2011). "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference". In: *Journal of Statistical Software* 42.8, pp. 1–28. URL: https://www.jstatsoft.org/v42/i08/.

King, Gary and Richard Nielsen (May 2019). "Why Propensity Scores Should Not Be Used for Matching". In: *Political Analysis* 27, pp. 1–20. DOI: 10.1017/pan.2019.11.

Lee, Brian, Justin Lessler, and Elizabeth Stuart (2010). "Improving propensity score weighting using machine learning". In: *Statistics in Medicine* 29.3, pp. 337–346. DOI: 10.1002/sim.3782.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rosenbaum, Paul and Donald Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1, pp. 41–55. DOI: 10.1093/biomet/70.1.41.

Rosenbaum, Paul and Donald Rubin (1984). "Reducing Bias in Observational Studies Using Sub-Classification on the Propensity Score". In: *Journal of the American Statistical Association* 79, pp. 516–524. DOI: 10.2307/2288398.

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA. URL: http://www.rstudio.com/.

Setoguchi, Soko, Sebastian Schneeweiss, Alan Brookhart, Robert Glynn, and Francis Cook (2008). "Evaluating uses of data mining techniques in propensity score estimation: a simulation study". In: *Pharmacoepidemiology and Drug Safety* 17.6, pp. 546–555. DOI: 10.1002/pds.1555.

White, H. (Jan. 1980). "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity". In: *Econometrica* 50.

Zeileis, Achim (2006). "Object-Oriented Computation of Sandwich Estimators". In: *Journal of Statistical Software* 16.9, pp. 1–16. DOI: 10.18637/jss.v016.i09.

Zeileis, Achim, Susanne Köll, and Nathaniel Graham (2020). "Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R". In: *Journal of Statistical Software* 95.1, pp. 1–36. DOI: 10.18637/jss.v095.i01.

A Simulated Data Set Summaries

Table A1: Scenario 1. Baseline covariate summary by treatment group.

	Treated		**Control**	
Variable	n	(%)	n	(%)
N	387		613	
Age				
[1, 18)	20	(5.2)	35	(5.7)
[18, 40)	90	(23.3)	140	(22.8)
[40, 65)	112	(28.9)	200	(32.6)
[65, 80)	134	(34.6)	196	(32.0)
[80, 90)	31	(8.0)	42	(6.9)
Gender				
male	230	(59.4)	374	(61.0)
female	157	(40.6)	239	(39.0)
Cardiac failure	20	(5.2)	30	(4.9)
Liver disease	16	(4.1)	31	(5.1)
COPD	105	(27.1)	177	(28.9)
Diabetes	64	(16.5)	90	(14.7)
Current smoker	73	(18.9)	116	(18.9)

Table A2: Scenario 2. Baseline covariate summary by treatment group.

	Treated		**Control**	
Variable	n	(%)	n	(%)
N	409		591	
Age				
[1, 18)	44	(10.8)	19	(3.2)
[18, 40)	123	(30.1)	106	(17.9)
[40, 65)	112	(27.4)	146	(24.7)
[65, 80)	109	(26.7)	269	(45.5)
[80, 90)	21	(5.1)	51	(8.6)
Gender				
male	228	(55.7)	343	(58.0)
female	181	(44.3)	248	(42.0)
Cardiac failure	7	(1.7)	37	(6.3)
Liver disease	19	(4.6)	35	(5.9)
COPD	119	(29.1)	173	(29.3)
Diabetes	47	(11.5)	81	(13.7)
Current smoker	66	(16.1)	129	(21.8)

Table A3: Scenario 3. Baseline covariate summary by treatment group.

	Treated		**Control**	
Variable	n	(%)	n	(%)
N	404		596	
Age				
[1, 18)	40	(9.9)	17	(2.9)
[18, 40)	118	(29.2)	127	(21.3)
[40, 65)	110	(27.2)	173	(29.0)
[65, 80)	111	(27.5)	219	(36.7)
[80, 90)	25	(6.2)	60	(10.1)
Gender				
male	258	(63.9)	331	(55.5)
female	146	(36.1)	265	(44.5)
Cardiac failure	18	(4.5)	32	(5.3)
Liver disease	8	(2.0)	36	(6.0)
COPD	155	(38.4)	133	(22.3)
Diabetes	70	(17.3)	101	(16.9)
Current smoker	21	(5.2)	152	(25.5)

B Model outputs

B.1 Simulation Scenario 1: Randomised Trial

```
Call:
glm(formula = death ~ treat, family = "binomial", data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4289  -0.4289  -0.3721  -0.3721   2.3258

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6356     0.1617 -16.302   <2e-16 ***
treat1        0.2955     0.2416   1.223    0.221
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 532.77  on 999  degrees of freedom
Residual deviance: 531.29  on 998  degrees of freedom
AIC: 535.29

Number of Fisher Scoring iterations: 5
```

Figure B1: Scenario 1. R output of model without adjustment, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat + gender + age + cardiac + COPD +
    liver + diab + smoke, family = "binomial", data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8464  -0.4138  -0.3441  -0.2761   2.8159

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.947304   0.457089  -8.636  < 2e-16 ***
treat1       0.302491   0.246123   1.229 0.219063
gender1      0.286451   0.259187   1.105 0.269078
age          0.014858   0.006288   2.363 0.018130 *
cardiac1    -1.033815   0.742821  -1.392 0.164001
COPD1        0.058567   0.270061   0.217 0.828313
liver1      -0.515329   0.739565  -0.697 0.485928
diab1        0.348532   0.310476   1.123 0.261620
smoke1       0.988228   0.264041   3.743 0.000182 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 532.77  on 999  degrees of freedom
Residual deviance: 505.01  on 991  degrees of freedom
AIC: 523.01

Number of Fisher Scoring iterations: 6
```

Figure B2: Scenario 1. R output of model with conventional covariate adjustment, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "binomial", data = matched_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4289  -0.4289  -0.3730  -0.3730   2.3239

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6308     0.2031 -12.956   <2e-16 ***
treat1        0.2907     0.2711   1.072    0.284
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 422.09  on 773  degrees of freedom
Residual deviance: 420.93  on 772  degrees of freedom
AIC: 424.93

Number of Fisher Scoring iterations: 5
```

Figure B3: Scenario 1. R output of model with matching, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = sim_data,
    weights = w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7694  -0.6794  -0.4799  -0.4737   3.6992

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.6286     0.1787 -14.712   <2e-16 ***
treat1        0.2868     0.2387   1.202     0.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 2.004147)

    Null deviance: 1090.8  on 999  degrees of freedom
Residual deviance: 1087.9  on 998  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure B4: Scenario 1. R output of model with weighting, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = sim_data,
    weights = sw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4787  -0.4226  -0.3757  -0.3709   2.4205

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.6286     0.1614 -16.291   <2e-16 ***
treat1        0.2868     0.2416   1.187    0.236
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.002043)

    Null deviance: 533.87  on 999  degrees of freedom
Residual deviance: 532.48  on 998  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure B5: Scenario 1. R output of model with stabilised weighting, true $\beta = 0$.

```
z test of coefficients:

            Estimate Std. Error  z value Pr(>|z|)
(Intercept) -2.62859    0.16192 -16.2335   <2e-16 ***
treat1       0.28677    0.24208   1.1846   0.2362
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure B6: Scenario 1. R output of model with weighting corrected with sandwich estimator, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "binomial", data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3529  -0.3529  -0.3529  -0.2044   2.7853

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7452     0.1696 -16.187  < 2e-16 ***
treat1       -1.1129     0.3955  -2.814  0.00489 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 367.04  on 999  degrees of freedom
Residual deviance: 357.37  on 998  degrees of freedom
AIC: 361.37

Number of Fisher Scoring iterations: 6
```

Figure B7: Scenario 1. R output of model without adjustment, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat + gender + age + cardiac + COPD +
    liver + diab + smoke, family = "binomial", data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8413  -0.2954  -0.2413  -0.1644   3.1208

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.716734   0.484233  -5.610 2.02e-08 ***
treat1      -1.128941   0.400941  -2.816  0.00487 **
gender1     -0.614798   0.315029  -1.952  0.05099 .
age          0.002381   0.007636   0.312  0.75523
cardiac1    -0.360675   0.768210  -0.470  0.63871
COPD1       -0.582413   0.379248  -1.536  0.12461
liver1       0.374228   0.636755   0.588  0.55673
diab1       -0.990428   0.610375  -1.623  0.10466
smoke1       1.440640   0.326249   4.416 1.01e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 367.04  on 999  degrees of freedom
Residual deviance: 330.26  on 991  degrees of freedom
AIC: 348.26

Number of Fisher Scoring iterations: 6
```

Figure B8: Scenario 1. R output of model with conventional covariate adjustment, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat, family = "binomial", data = matched_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3340  -0.3340  -0.2044  -0.2044   2.7853

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.8581     0.2244  -12.74   <2e-16 ***
treat1       -1.0000     0.4219   -2.37   0.0178 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 247.39  on 773  degrees of freedom
Residual deviance: 241.12  on 772  degrees of freedom
AIC: 245.12

Number of Fisher Scoring iterations: 6
```

Figure B9: Scenario 1. R output of model with matching, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = sim_data,
    weights = w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4732  -0.4507  -0.4429  -0.3233   4.5339

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.7488     0.1883  -14.60  < 2e-16 ***
treat1       -1.1425     0.3710   -3.08  0.00213 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 2.004148)

    Null deviance: 672.83  on 999  degrees of freedom
Residual deviance: 650.95  on 998  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Figure B10: Scenario 1. R output of model with weighting, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = sim_data,
    weights = sw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3705  -0.3528  -0.3468  -0.2011   2.8205

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.7488     0.1701 -16.165  < 2e-16 ***
treat1       -1.1425     0.4011  -2.848  0.00449 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.002086)

    Null deviance: 364.70  on 999  degrees of freedom
Residual deviance: 354.69  on 998  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Figure B11: Scenario 1. R output of model with stabilised weighting, true $\beta = -1$.

```
z test of coefficients:

            Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -2.74880    0.16985 -16.1833 < 2.2e-16 ***
treat1      -1.14251    0.39604  -2.8848  0.003916 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure B12: Scenario 1. R output of model with weighting corrected with sandwich estimator, true $\beta = -1$.

B.2 Simulation Scenario 2: All Covariates are Confounders

```
Call:
glm(formula = death ~ treat, family = "binomial", data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7079  -0.7079  -0.6644  -0.6644   1.7996

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.25603    0.09903 -12.683   <2e-16 ***
treat1      -0.14254    0.15875  -0.898    0.369
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1033.2  on 999  degrees of freedom
Residual deviance: 1032.4  on 998  degrees of freedom
AIC: 1036.4

Number of Fisher Scoring iterations: 4
```

Figure B13: Scenario 2. R output of model without adjustment, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat + gender + age + cardiac + COPD +
    liver + diab + smoke, family = "binomial", data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0121  -0.7311  -0.6350  -0.4886   2.2012

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.392216   0.287249  -8.328  < 2e-16 ***
treat1       0.034355   0.166434   0.206 0.836464
gender1      0.153576   0.159427   0.963 0.335397
age          0.014998   0.003926   3.821 0.000133 ***
cardiac1    -0.154359   0.376758  -0.410 0.682024
COPD1        0.267611   0.171431   1.561 0.118515
liver1       0.329247   0.319096   1.032 0.302161
diab1        0.105269   0.228319   0.461 0.644755
smoke1       0.104594   0.195511   0.535 0.592665
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1033.2  on 999  degrees of freedom
Residual deviance: 1010.4  on 991  degrees of freedom
AIC: 1028.4

Number of Fisher Scoring iterations: 4
```

Figure B14: Scenario 2. R output of model with conventional covariate adjustment, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "binomial", data = matched_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6993  -0.6993  -0.6785  -0.6785   1.7786

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.28382    0.12878  -9.969   <2e-16 ***
treat1      -0.06763    0.18393  -0.368    0.713
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 732.20  on 709  degrees of freedom
Residual deviance: 732.06  on 708  degrees of freedom
AIC: 736.06

Number of Fisher Scoring iterations: 4
```

Figure B15: Scenario 2. R output of model with matching, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = sim_data,
    weights = w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9241  -0.9996  -0.8687  -0.7898   4.8527

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.31784    0.10983 -11.999   <2e-16 ***
treat1       0.02803    0.15450   0.181    0.856
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 2.006908)

    Null deviance: 2077.4  on 999  degrees of freedom
Residual deviance: 2077.4  on 998  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure B16: Scenario 2. R output of model with weighting, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = sim_data,
    weights = sw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2305  -0.7118  -0.6375  -0.5559   3.1035

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.31784    0.10100 -13.048   <2e-16 ***
treat1       0.02803    0.15693   0.179    0.858
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.002963)

    Null deviance: 1037.1  on 999  degrees of freedom
Residual deviance: 1037.1  on 998  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Figure B17: Scenario 2. R output of model with stabilised weighting, true $\beta = 0$.

```
z test of coefficients:

             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -1.317837   0.102238 -12.8899   <2e-16 ***
treat1       0.028029   0.167164   0.1677   0.8668
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure B18: Scenario 2. R output of model with weighting corrected with sandwich estimator, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "binomial", data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7079  -0.7079  -0.4596  -0.4596   2.1448

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.25603    0.09903 -12.683  < 2e-16 ***
treat1      -0.93848    0.19212  -4.885 1.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 918.09  on 999  degrees of freedom
Residual deviance: 891.63  on 998  degrees of freedom
AIC: 895.63

Number of Fisher Scoring iterations: 4
```

Figure B19: Scenario 2. R output of model without adjustment, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat + gender + age + cardiac + COPD +
    liver + diab + smoke, family = "binomial", data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0065  -0.6923  -0.5127  -0.3983   2.4007

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.964516   0.301818  -6.509 7.57e-11 ***
treat1      -0.875531   0.198275  -4.416 1.01e-05 ***
gender1      0.324145   0.176827   1.833   0.0668 .
age          0.008784   0.004215   2.084   0.0372 *
cardiac1    -1.094293   0.540076  -2.026   0.0427 *
COPD1        0.089471   0.191522   0.467   0.6404
liver1       0.452943   0.340087   1.332   0.1829
diab1        0.167487   0.244105   0.686   0.4926
smoke1      -0.200904   0.224302  -0.896   0.3704
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 918.09  on 999  degrees of freedom
Residual deviance: 876.45  on 991  degrees of freedom
AIC: 894.45

Number of Fisher Scoring iterations: 4
```

Figure B20: Scenario 2. R output of model with conventional covariate adjustment, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat, family = "binomial", data = matched_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7183  -0.7183  -0.4596  -0.4596   2.1448

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2231     0.1180 -10.368  < 2e-16 ***
treat1       -0.9714     0.2025  -4.796 1.62e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 729.56  on 817  degrees of freedom
Residual deviance: 704.88  on 816  degrees of freedom
AIC: 708.88

Number of Fisher Scoring iterations: 4
```

Figure B21: Scenario 2. R output of model with matching, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = sim_data,
    weights = w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3803  -0.8776  -0.8104  -0.6191   5.8442

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.2882     0.1089 -11.828  < 2e-16 ***
treat1       -0.8208     0.1805  -4.548 6.08e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 2.007102)

    Null deviance: 1774.8  on 999  degrees of freedom
Residual deviance: 1731.2  on 998  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Figure B22: Scenario 2. R output of model with weighting, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = sim_data,
    weights = sw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.6604  -0.5667  -0.3960   3.7376

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.2882     0.1001 -12.863  < 2e-16 ***
treat1       -0.8208     0.1880  -4.367 1.39e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.002977)

    Null deviance: 918.59  on 999  degrees of freedom
Residual deviance: 897.82  on 998  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Figure B23: Scenario 2. R output of model with stabilised weighting, true $\beta = -1$.

```
z test of coefficients:

            Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -1.28816    0.10181 -12.6529 < 2.2e-16 ***
treat1      -0.82075    0.20568  -3.9904 6.596e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure B24: Scenario 2. R output of model with weighting corrected with sandwich estimator, true $\beta = -1$.

B.3 Simulation Scenario 3: A More Realistic Case

```
Call:
glm(formula = death ~ treat, family = "binomial", data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6420  -0.6420  -0.4626  -0.4626   2.1391

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4746     0.1052 -14.015  < 2e-16 ***
treat1       -0.7062     0.1955  -3.613 0.000303 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 852.33  on 999  degrees of freedom
Residual deviance: 838.32  on 998  degrees of freedom
AIC: 842.32

Number of Fisher Scoring iterations: 4
```

Figure B25: Scenario 3. R output of model without adjustment, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat + gender + age + cardiac + COPD +
    liver + diab + smoke, family = "binomial", data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4320  -0.5189  -0.4125  -0.3215   2.5392

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.175975   0.350816  -9.053  < 2e-16 ***
treat1      -0.039475   0.225835  -0.175 0.861242
gender1     -0.049429   0.194831  -0.254 0.799727
age          0.018201   0.004852   3.751 0.000176 ***
cardiac1     0.287709   0.380389   0.756 0.449436
COPD1       -0.168772   0.217456  -0.776 0.437677
liver1      -1.122896   0.631370  -1.779 0.075321 .
diab1       -0.180851   0.264439  -0.684 0.494036
smoke1       2.031530   0.221033   9.191  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 852.33  on 999  degrees of freedom
Residual deviance: 726.79  on 991  degrees of freedom
AIC: 744.79

Number of Fisher Scoring iterations: 5
```

Figure B26: Scenario 3. R output of model with conventional covariate adjustment, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "binomial", data = matched_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4788  -0.4788  -0.4637  -0.4637   2.1371

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.17617    0.18652  -11.67   <2e-16 ***
treat1       0.06774    0.26033    0.26    0.795
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 422.19  on 627  degrees of freedom
Residual deviance: 422.12  on 626  degrees of freedom
AIC: 426.12

Number of Fisher Scoring iterations: 4
```

Figure B27: Scenario 3. R output of model with matching, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = sim_data,
    weights = w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5650  -0.7690  -0.6755  -0.5928   6.7456

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7585     0.1279 -13.752   <2e-16 ***
treat1       -0.2624     0.1879  -1.396    0.163
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 2.041331)

    Null deviance: 1587.2  on 999  degrees of freedom
Residual deviance: 1583.2  on 998  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure B28: Scenario 3. R output of model with weighting, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = sim_data,
    weights = sw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6303  -0.5580  -0.4937  -0.3954   4.2876

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7585     0.1169 -15.047   <2e-16 ***
treat1       -0.2624     0.1924  -1.363    0.173
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.016269)

    Null deviance: 801.15  on 999  degrees of freedom
Residual deviance: 799.23  on 998  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Figure B29: Scenario 3. R output of model with stabilised weighting, true $\beta = 0$.

```
z test of coefficients:

            Estimate Std. Error  z value Pr(>|z|)
(Intercept) -1.75849    0.10987 -16.0046   <2e-16 ***
treat1      -0.26237    0.23777  -1.1034   0.2698
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure B30: Scenario 3. R output of model with weighting corrected with sandwich estimator, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = trim_data,
    weights = w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5712  -0.7637  -0.6727  -0.5912   6.3624

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7585     0.1234 -14.253   <2e-16 ***
treat1       -0.2871     0.1910  -1.503    0.133
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.900252)

    Null deviance: 1463.0  on 989  degrees of freedom
Residual deviance: 1458.7  on 988  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Figure B31: Scenario 3. R output of model with weighting (trimmed), true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = trim_data,
    weights = sw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2130  -0.5523  -0.4911  -0.3908   4.0440

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7585     0.1137 -15.472   <2e-16 ***
treat1       -0.2871     0.1988  -1.444    0.149
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 0.9611689)

    Null deviance: 750.97  on 989  degrees of freedom
Residual deviance: 748.92  on 988  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Figure B32: Scenario 3. R output of model with stabilised weighting (trimmed), true $\beta = 0$.

```
z test of coefficients:

            Estimate Std. Error  z value Pr(>|z|)
(Intercept) -1.75849    0.10988 -16.0044   <2e-16 ***
treat1      -0.28714    0.22452  -1.2789   0.2009
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure B33: Scenario 3. R output of model with weighting (trimmed) corrected with sandwich estimator, true $\beta = 0$.

```
Call:
glm(formula = death ~ treat, family = "binomial", data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6291  -0.6291  -0.2932  -0.2932   2.5172

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5195     0.1067 -14.237  < 2e-16 ***
treat1       -1.6057     0.2698  -5.952 2.65e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 749.64  on 999  degrees of freedom
Residual deviance: 702.04  on 998  degrees of freedom
AIC: 706.04

Number of Fisher Scoring iterations: 5
```

Figure B34: Scenario 3. R output of model without adjustment, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat + gender + age + cardiac + COPD +
    liver + diab + smoke, family = "binomial", data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2244  -0.5185  -0.3523  -0.2478   2.7021

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.489943   0.351758  -7.079 1.46e-12 ***
treat1      -1.132228   0.288819  -3.920 8.85e-05 ***
gender1     -0.095397   0.208278  -0.458   0.6469
age          0.009156   0.005078   1.803   0.0714 .
cardiac1     0.379377   0.398173   0.953   0.3407
COPD1       -0.055804   0.238964  -0.234   0.8154
liver1       0.317275   0.425555   0.746   0.4559
diab1       -0.668864   0.322642  -2.073   0.0382 *
smoke1       1.579292   0.228012   6.926 4.32e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 749.64  on 999  degrees of freedom
Residual deviance: 640.22  on 991  degrees of freedom
AIC: 658.22

Number of Fisher Scoring iterations: 6
```

Figure B35: Scenario 3. R output of model with conventional covariate adjustment, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat, family = "binomial", data = matched_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4637  -0.4637  -0.2792  -0.2792   2.5552

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.1762     0.1865 -11.666   <2e-16 ***
treat1       -1.0493     0.3485  -3.011   0.0026 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 318.78  on 627  degrees of freedom
Residual deviance: 308.66  on 626  degrees of freedom
AIC: 312.66

Number of Fisher Scoring iterations: 6
```

Figure B36: Scenario 3. R output of model with matching, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = sim_data,
    weights = w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5609  -0.6953  -0.5706  -0.3350   4.9729

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7727     0.1285 -13.794  < 2e-16 ***
treat1       -1.6035     0.2790  -5.747 1.21e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 2.041387)

    Null deviance: 1212.5  on 999  degrees of freedom
Residual deviance: 1127.5  on 998  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure B37: Scenario 3. R output of model with weighting, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = sim_data,
    weights = sw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2051  -0.5334  -0.4399  -0.2129   3.1608

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7727     0.1175 -15.092  < 2e-16 ***
treat1       -1.6035     0.2990  -5.363 1.01e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.016267)

    Null deviance: 653.50  on 999  degrees of freedom
Residual deviance: 613.92  on 998  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Figure B38: Scenario 3. R output of model with stabilised weighting, true $\beta = -1$.

```
z test of coefficients:

            Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -1.77272    0.11150 -15.8992 < 2.2e-16 ***
treat1      -1.60347    0.29076  -5.5148 3.491e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure B39: Scenario 3. R output of model with weighting corrected with sandwich estimator, true $\beta = -1$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = trim_data,
    weights = w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5609  -0.6910  -0.5716  -0.3637   4.8498

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7727     0.1240 -14.297  < 2e-16 ***
treat1       -1.4305     0.2699  -5.301 1.42e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.900143)

    Null deviance: 1181.8  on 989  degrees of freedom
Residual deviance: 1115.8  on 988  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure B40: Scenario 3. R output of model with weighting (trimmed), true $\beta = -1$.

```
Call:
glm(formula = death ~ treat, family = "quasibinomial", data = trim_data,
    weights = sw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2051  -0.5327  -0.4387  -0.2312   3.0826

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7727     0.1142 -15.519  < 2e-16 ***
treat1       -1.4305     0.2915  -4.907 1.08e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 0.9611603)

    Null deviance: 639.6  on 989  degrees of freedom
Residual deviance: 609.2  on 988  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Figure B41: Scenario 3. R output of model with stabilised weighting (trimmed), true $\beta = -1$.

```
z test of coefficients:

            Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -1.77272    0.11150 -15.8990 < 2.2e-16 ***
treat1      -1.43055    0.28739  -4.9776 6.436e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure B42: Scenario 3. R output of model with weighting (trimmed) corrected with sandwich estimator, true $\beta = -1$.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kristin Jesse,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „*Introduction to Propensity Score Methods*“, mille juhendajad on Jaak Sõnajalg ja Krista Fischer, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kristin Jesse
25.05.2021