Methods for Improving Covariate Balance in Observational Studies

Philip Fowler

Department of Statistics, USBE
Umeå 2017
A straight line may be the shortest distance between two points,
but it is by no means the most interesting.

- Doctor Who, The Time Warrior
List of Papers

The thesis is based on the following papers:


Paper II is reprinted with the kind permission of Elsevier. Paper III is reprinted with the kind permission of the editorial board of Observational Studies.
Abstract

This thesis contributes to the field of causal inference, where the main interest is to estimate the effect of a treatment on some outcome. At its core, causal inference is an exercise in controlling for imbalance (differences) in covariate distributions between the treated and the controls, as such imbalances otherwise can bias estimates of causal effects. Imbalance on observed covariates can be handled through matching, where treated and controls with similar covariate distributions are extracted from a data set and then used to estimate the effect of a treatment.

The first paper of this thesis describes and investigates a matching design, where a data-driven algorithm is used to discretise a covariate before matching. The paper also gives sufficient conditions for if, and how, a covariate can be discretised without introducing bias.

Balance is needed for unobserved covariates too, but is more difficult to achieve and verify. Unobserved covariates are sometimes replaced with correlated counterparts, usually referred to as proxy variables. However, just replacing an unobserved covariate with a correlated one does not guarantee an elimination of, or even reduction of, bias. In the second paper we formalise proxy variables in a causal inference framework and give sufficient conditions for when they lead to nonparametric identification of causal effects.

The third and fourth papers both concern estimating the effect an enhanced cooperation between the Swedish Social Insurance Agency and the Public Employment Service has on reducing sick leave. The third paper is a study protocol, where the matching design used to estimate this effect is described. The matching was then also carried out in the study protocol, before the outcome for the treated was available, ensuring that the matching design was not influenced by any estimated causal effects. The third paper also presents a potential proxy variable for unobserved covariates, that is used as part of the matching. The fourth paper then carries out the analysis described in the third paper, and uses an instrumental variable approach to test for unobserved confounding not captured by the supposed proxy variable.

KEYWORDS: causal effect; coarsening; discretisation; proxy variables; register study; swedish social insurance agency; unobserved variables
Populärvetenskaplig sammanfattning


Anledningen till att man vill slumpa vem som blir behandlad är för att det annars kan bli så att de som blir behandlade systematiskt skiljer sig åt från de som inte blir det. Exempelvis skulle det kunna vara så att de som blir behandlade i regel är yngre än de som inte blir det. Om åldern även påverkar utfallet, uppstår problem med att avgöra om skillnaden i utfall mellan de behandlade och de obehandlade beror på behandlingen eller på åldersfördelningen mellan grupperna. Ett sätt att hantera detta problem är att hitta behandlade och obehandlade som har samma åldersfördelningar, en process som kallas för matching. Problemet är dock att det i regel är fler variabler än bara ålder som man vill matcha för. Ju fler variabler man vill matcha för desto svårare blir det att hitta jämförbara grupper.

Den här avhandlingen handlar om metoder för att öka jämförbarheten mellan två grupper när behandlingen ej har kunnat slumpas. Avhandlingen tar även upp scenariot där vissa av dessa variabler inte ens har mätts. Ett stärkt samarbete mellan Försäkringskassan och Arbetsförmedlingen utvärderas också med hjälp av metoder som behandlas i denna avhandling.
Preface

While visiting my father at work one day when I was younger, I told him that I would never want his kind of job. After all, his work day seemed to consist of sitting in front of a computer and entering numbers into Excel. Flash forward to today and I can safely say that I have succeeded in not following in his footsteps. R is, after all, a very different piece of software than Excel.

I didn’t plan on becoming a statistician, then again, few of us do. My plans upon graduating from high-school were in fact very clear; I was going to become a history teacher for high-school students. History lead me to economic history, which in turn lead me to economics. While studying economics, I decided that taking a course or two of statistics would be a good idea. However, it turns out that in the Markov chain of life, studying statistics is an absorbing state, so here I am today!

Now that I am at the end of my PhD studies, there are many people I would like to thank. First and foremost I want to thank my supervisor Xavier de Luna. Thank you for your guidance, knowledge and generosity with your time. I could not have asked for a better supervisor! I have much appreciated our discussions and your feedback on things I have written, rewritten and re-rewritten. Your ability to put things into perspective has also been very helpful, as have your reminders to be more concise in my writing. I also want to thank my co-supervisor Ingeborg Waernbaum for opening my eyes to the field of causal inference to begin with, and whose feedback, help and support has been very valuable these past years. Thank you, Ingeborg!

I would also like to extend my thanks to my co-authors Per Johansson, Petra Ornstein, Sofia Bill and Peje Bengtsson for their valuable contributions and knowledge. It has been a joy to collaborate with all of you.

To all my fellow PhD students, thank you for making these years extra memorable. Special thanks to Anita, who (according to an anonymous source named Anita) is much better than I am in every way. Thank you for all the laughter and strange very reasonable conversations that we have had throughout the years. You’re a great friend.

To all my other colleagues, too many to mention by name, thank you for providing such a great working environment and for taking the time to answer my many questions about statistics, probability theory,
teaching and legislation.

As always, my family deserves a special thank you. Dad, thanks for the many walks we have taken during lunch breaks and for your never-ending support. Thank you also for your offers to come and talk about local anesthesia during statistics classes I’ve been teaching, even though I still haven’t quite figured out what it has to do with the course material. Mom, thank you for always having my back, for our lunches together and for reminding me that I need to have some time off work. Denise, thank you for sending me random messages, for bringing me food and for continuing the family legacy of being a nerd.

To Renhoren, my favourite band and musical home, thanks a bunch for going on tour during the last year of my PhD programme, when I didn’t have time to join along. Jokes aside, you have been and continue to be very important to me. As the saying goes, Renhoren is more reliable than the weather! My thanks to all of you, not least Erik Sturesson for being awesome at just the right moment.

To all my friends outside of work, thank you for your understanding that I haven’t had that much time for y’all lately. Well, except for Petter whose understanding I take for granted, seeing how he’s also going through the same academic process. But to the rest of you, thank you! Also, thank you all for the messages you’ve sent me, lunches together, board game evenings and invitations to various nerdy events.

Finally, to Erica, my love, thank you for being so patient, supportive and understanding through all these years, especially during the times I had to work very late in the evenings (it happened once or twice) or during the weekends. Through all ups and downs of my PhD student years you have always been there for me and given me your support. I love you.

Umeå, September 2017
Philip Fowler
1 Introduction

One of the first things pointed out about causality while studying statistics, is the old saying that correlation does not imply causation. In other words, just because two things are correlated does not mean that one causes the other. Ice-cream sales are likely correlated to the number of people swimming at beaches, but that does not mean that ice-cream causes people to swim nor that swimming causes people to eat ice-cream! The two are likely to be correlated simply because outdoor temperature affects both ice-cream sales and the willingness of people to go swimming.

The ambition of making causal statements is an old one, at least dating back to Democritus of Abdera (ca 460-370 BC), with his claim "(I would) rather discover one cause than gain the kingdom of Persia." (Freeman, 1948). While the aim of finding causes of effects is intriguing, this thesis is concerned with the more humble goal of wanting to estimate the effects of causes. The difference between "cause of an effect" and "effect of a cause" might seem small at first glance, but is from a scientific point of view rather large. The former aims at answering questions such as "why do some people develop lung cancer?", while the latter focuses on more targeted questions such as "does smoking lead to an increased risk of lung cancer?". Indeed, the difference is that of understanding a causal mechanism on one hand and, on the other hand, wondering what an intervention of some sort would lead to. It is from this point of view much easier to answer "If we do A, what will happen to B?" than it is to answer "Why does B happen?". Holland (1986) gives a more thorough discussion of this distinction.

One way of estimating causal effects is randomly to divide individuals into two groups, where one of the groups is exposed to some sort of treatment while the other is not. The difference in outcome between these groups could then be used to estimate the effect of the treatment since the only thing affecting whether or not an individual is treated is the randomisation mechanism. In the example with ice-cream and swimming, we could throughout the course of a year randomly select people into two groups, one that is given ice-cream and one that is not. We could then see if the ice-cream group tends to swim more than the group without ice-cream. Since both groups would experience the same weather and since no factor outside the randomisation affected if an
individual was assigned to the ice-cream group or not, simple comparisons of differences in outcomes between the two groups would lead to an unbiased estimate of the effect ice-cream has on the willingness to go swimming.

Even though randomised experiments in many ways are the gold standard for estimating causal effects, there are cases where they are not possible to do. For instance, if one was to examine the effect of smoking on the risk of developing some disease, a randomised experiment would require that some individuals were randomly selected to smoke. With all we know about the dangerous properties of smoking, it would be highly unethical to force people to smoke, and thus a randomised experiment cannot be performed. Further reasons why randomised experiments are not always plausible are monetary or time constraints, experimental conditions not mimicking the real world closely enough for valid conclusions to be drawn, or difficulties in ensuring compliance to the assigned treatment (as would be the case in the smoking example). When randomisation is not possible, researchers have to rely on observational studies.

Despite the problems associated with observational studies, estimating causal effects can be possible under the right conditions, which will be outlined in Section 2. The key to observational studies lies in the control of factors that affect both the outcome and the chance of being treated. As opposed to randomised experiments, it is usually the case in observational studies that those treated differ in some covariate distributions from those not treated. For example, smokers and non-smokers might differ in their alcohol consumptions. If that is the case, we say that our data is imbalanced with regards to alcohol consumption. Causal inference is to a large extent an exercise in handling such pre-treatment imbalances, both on observed and unobserved covariates, and we thus need methods for improving covariate balance in observational studies.

The remainder of this thesis is structured in the following way. Section 2 reviews the potential outcomes framework that greatly facilitates discussion of causal effects. Directed acyclic graphs are also discussed. Section 3 covers the topic of matching estimators, one of many ways of estimating causal effects from observational data. Propensity score matching, a very common type of matching technique, is discussed in
Section 4. Section 5 gives an introduction to coarsening (discretisation) of covariates, while Section 6 is dedicated to the topic of using proxy variables in place of unobserved confounders. Section 7 covers instrumental variables and discusses how they can be used to examine a proxy variable assumption. Section 8 is about the Swedish Social Insurance Agency (Försäkringskassan in Swedish), since two of the four papers of this thesis analyses data from said agency. Section 9 summarises the papers of this thesis, which are discussed briefly in Section 10 together with suggestions for future research.

2 The Potential Outcomes Framework

In order to properly discuss estimation of causal effects, we first need to define what causal effects are. In non-mathematical terms, a causal effect is a comparison of what the outcome would be for an individual, if he or she would be treated, as opposed to if he or she would not. For instance, if a medicine has a positive causal effect on recovering from some disease, we would expect that a patient who is administered the medicine would recover more quickly than if he or she was not. The potential outcomes framework, also known as the Neyman-Rubin causal model, is a mathematical formalisation of these ideas. It was first introduced by Jerzy Neyman in the context of randomised experiments in 1923 [Neyman 1923] and by Donald Rubin in the 1970’s for observational studies [Rubin 1974]. In this section we briefly review this framework.

Suppose that for each individual \(i = 1, \ldots, n\), we measure a binary variable of treatment, \(T_i\), such that \(T_i = 1\) if the individual is treated and \(T_i = 0\) if he or she is not. Each individual has two potential outcomes, \(Y_{1i}\) and \(Y_{0i}\), denoting the outcome he or she would have if treated and untreated, respectively. For instance, if studying the effect that participation in a labour market programme has on future employment, \(Y_{1i}\) would then be the future employment status of individual \(i\), if he or she entered into the programme, while \(Y_{0i}\) would be the future employment status if he or she did not. The effect the treatment would have on individual \(i\) is thus a comparison between \(Y_{1i}\) and \(Y_{0i}\), say \(Y_{1i} - Y_{0i}\). If both \(Y_{1i}\) and \(Y_{0i}\) would be known to us, the causal effect would per definition also be known. However, for each individual we can either observe
\(Y_{1i} \) or \(Y_{0i}\). They cannot both be observed since the treatment either is administered or it is not, i.e., we only observe \(Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i}\). This is the fundamental problem of causal inference (Holland 1986). Because of this, individual level causal effects are not observed. A researcher must instead focus on aggregated causal effects such as the average causal effect, \(\tau = E(Y_{1i} - Y_{0i})\), or the average causal effect on the treated, \(\tau^t = E(Y_{1i} - Y_{0i} \mid T_i = 1)\). The former parameter is of interest if the treatment is to be administered to the entire population, while the latter is more suitable in scenarios where applying the treatment to everybody in the population is not reasonable. In the labour market programme example, those who actively enroll in such a programme might differ from those who do not, for instance if those who do not enroll are individuals unlikely to return to the workforce even if they would have enrolled in the programme. The effect that the programme under investigation has for the average unemployed person is then arguably not as interesting as the effect it has on those who are likely to participate, making \(\tau^t\) the more suitable parameter. For a further discussion of these parameters, see Imbens and Wooldridge (2009). As we will discuss in Section 2.2, identification of \(\tau\) and \(\tau^t\) is possible under certain assumptions.

### 2.1 Directed Acyclic Graphs

In casual inference, a directed acyclic graph (DAG) is a graphical representation of how variables are causally connected. We use them in Section 2.2, 6.2 and 7 to illustrate assumptions made in this thesis. In this section we give an introduction to DAGs.

Arrows in a DAG indicate causality, e.g., an arrow pointing from a variable \(A\) to another variable \(B\), states that \(A\) is a cause of \(B\). An example of a DAG is shown in Figure 1 below.

```
A

B  C

Figure 1: An example of a DAG.
```

The DAG in Figure 1 states that the distribution of \(B\) varies with...
A, since $A$ is a cause of $B$. As $A$ also is a cause of $C$, we say that the association between $B$ and $C$ is confounded by $A$. This confounding means that $B$ and $C$ are dependent due to their common cause, $A$. If we were to estimate the causal effect that $B$ has on $C$, we would get biased estimates unless we control for $A$.

A variable $D$ is said to be a collider for the association between two variables $C$ and $E$ if there are arrows from both $C$ and $E$ pointing into $D$. For instance, in the DAG $C \rightarrow D \leftarrow E$, $D$ is a collider since the arrows from $C$ and $E$ "collide" at $D$. Conditioning on a collider creates a dependency between the variables that are causes of it. It is thus possible for two variables to be independent of each other, but not being independent conditional on a third variable. As an example of this, consider the eye colour of the two biological parents of a child. It is not unreasonable to assume that the parents’ eye colours are independent of each other. However, the eye colour of a child is dependent on the eye colour of both his or her parents. This makes the child’s eye colour a collider for his or her parents’ eye colours. If we know that the child has brown eyes and that parent 1 has blue eyes, the likelihood of parent 2 having brown eyes is higher than if we knew parent 1 has brown eyes. Expressed differently, the eye colours of parents, while independent of each other, become conditionally dependent given the eye colour of their children.

In a DAG, a path between two variables is a set of arrows connecting them directly or indirectly via other variables. As an example, consider the DAG in Figure 2 below.

$$
A \rightarrow B \quad C \rightarrow D \leftarrow E
$$

Figure 2: An illustration of paths in a DAG.

There is no path from $A$ to $E$ in Figure 2, but there is a path between $A$ and $B$ as well as between $C$ and $E$. A path is said to be open if there are no colliders along the path. Paths that are not open are said to be closed. Conditioning on a non-collider variable along an open path closes the path. Conditioning on a collider creates paths between all variables that are causes of the collider. To get unbiased estimates of causal effects, we must condition on variables such that no open paths between the treatment and the potential outcomes remain. Consider Figure 3, where $Y_{ji}$, for $j = 0, 1$, are the potential outcomes for individual $i$. 


In Figure 3, the path $T_i \leftarrow X_1 \rightarrow X_2 \leftarrow X_3 \rightarrow Y_{ji}$ is closed since $X_2$ is a collider, while the path $T_i \leftarrow X_4 \rightarrow Y_{ji}$ is open. By conditioning on $X_4$, all (the only) open paths between $T_i$ and $Y_{ji}$ are closed, meaning that $\tau$ and $\tau^t$ are identified. However, by conditioning on $X_2$, a collider, we open up a path between $X_1$ and $X_3$, meaning that we have an open path from $T_i$ to $Y_{ji}$ going through $X_1$ and $X_3$. Unless we also condition on $X_1$ and/or $X_3$, the path will remain open, thus causing bias in our estimates of $\tau$ and $\tau^t$. For a more thorough introduction to DAGs, see e.g. Morgan and Winship (2014).

### 2.2 Causal Inference Assumptions

The fact that $\tau^t$ and $\tau$ can be identified, despite $Y_{1i}$ and $Y_{0i}$ not being jointly observed, is arguably one of the most important results in the causal inference literature. Identification is, however, not possible without some assumptions. Using the independence notation of Dawid (1979), where $A \perp \perp B \mid C$ denotes that $A$ is conditionally independent of $B$ given $C$, we first make the following assumption:

**Assumption 1 Unconfoundedness**

\[ i) \quad Y_{0i} \perp T_i \mid X_i, \]
\[ ii) \quad Y_{1i} \perp T_i \mid X_i, \]

where $X_i$ is a set of pretreatment (observed) covariates. Assumption 1 states that we have no unobserved confounders, i.e., that we have controlled for enough covariates to ensure that any systematic difference in outcome between the treated and the controls is due to the treatment and not due to background characteristics that differ between the two groups. For instance, suppose we wish to estimate the effect of some surgical procedure on recovery from a disease, but do not have access to
information about the age of the patients. If those who undergo surgery tend to be younger than those who do not, then the effect that the surgical procedure has on recovery cannot be separated from the effect of age. In such a scenario, Assumption 1 does not hold.

Letting $U_i$ denote an unobserved covariate, Figure 4 shows two DAGs that illustrate cases where Assumption 1 holds and does not hold respectively.

\[ \begin{align*}
X_i 
\downarrow 
T_i 
\downarrow 
Y_{ji}
\end{align*} \]

(a) Assumption 1 holds.

\[ \begin{align*}
X_i 
\downarrow 
U_i 
\downarrow 
T_i 
\downarrow 
Y_{ji}
\end{align*} \]

(b) Assumption 1 does not hold.

Figure 4: A simple illustration of Assumption 1.

Assumption 1 is strong and untestable, so the plausibility of it holding must be determined by experts in the research field and preferably combined with a sensitivity analysis of the robustness of the study’s results to deviations from this assumption. If a large number of relevant covariates are available to the researcher, the assumption seems more plausible than if only a few are. That said, there are dangers in conditioning on the wrong or too many covariates, see e.g. de Luna et al. (2011) as well as the discussion in a series of letters between Shrier (2008), Rubin (2008, 2009), Pearl (2009) and Sjölander (2009). In a randomised experiment Assumption 1 holds by design since the only thing affecting whether or not an individual is treated is the randomisation mechanism. This would in a DAG be illustrated by having no arrow leading into $T_i$.

**Assumption 2**  
Common support

1. \( \Pr(T_i = 0 \mid X_i) > 0 \),
2. \( \Pr(T_i = 1 \mid X_i) > 0 \).

Assumption 2 states that all individuals have a non-zero probability of becoming controls and also a non-zero probability of becoming treated. In other words, we assume that there are no values of $X_i$ that automatically lead to an individual becoming treated (or not treated).
If this does not hold, say if individuals under a certain age never can be treated, then estimation of causal effects would rely heavily on model extrapolation. Interpretation of the estimated causal parameter would also be more difficult, since it is not clear what a comparison with something that could never happen would say. Again, this assumption holds by design in a randomised experiment since the treatment assignment mechanism is under the experimenter’s control.

In addition to these two assumptions we need an assumption referred to as the Stable Unit Treatment Value Assumption (Rubin, 1980, 1986).

Assumption 3

**Stable Unit Treatment Value Assumption**

The potential outcomes of one individual does not depend on the treatment assignment of other individuals.

Assumption 3 is needed since the definition of potential outcomes in Section 2 otherwise are not sufficient for defining causal effects.

There are many cases where Assumption 3 is reasonable. Whether or not some surgical procedure will help cure a patient of his or her disease is not affected by whether or not another patient undertakes surgery. However, there are also many cases where Assumption 3 is not reasonable, for instance when it comes to vaccination. The more people that are vaccinated, the more protected the unvaccinated become. In that case, \( Y_{0i} \) then depends on the treatment assignment of other individuals, which violates Assumption 3. As opposed to Assumptions 1 and 2, the Stable Unit Treatment Value Assumption does not necessarily hold by design in experimental studies.

If Assumptions 1-3 hold, \( \tau \) and \( \tau^t \) are identified. In fact, for identification of \( \tau^t \), only Assumptions 1), 2) and 3) are needed.

### 3 Matching

There are many ways of estimating causal effects from observational data or, expressed differently, many ways to condition on \( X_i \). One could fit a suitable regression model and use the estimated coefficients from the model to draw inference about \( \tau^t \) and \( \tau \). For instance, suppose that \( Y_{1i} = \alpha_1 + \beta_1 X_i + \varepsilon_{1i} \) and \( Y_{0i} = \alpha_0 + \beta_0 X_i + \varepsilon_{0i} \), where \( \varepsilon_{1i} \) and \( \varepsilon_{0i} \) are independent error terms. One can then show that linear regression of \( Y_i \) on \( X_i \), \( T_i \) and \( X_i T_i \) gives parameter estimates that can be combined
to yield consistent estimates of $\alpha_1$, $\alpha_0$, $\beta_1$ and $\beta_0$. This means that $\tau^t$ and $\tau$ can be estimated consistently since, in this example, $\tau = \alpha_1 - \alpha_0 + (\beta_1 - \beta_0)E(X_i)$ and $\tau^t = \alpha_1 - \alpha_0 + (\beta_1 - \beta_0)E(X_i | T_i = 1)$. However, misspecification of the regression models for $Y_{1i}$ and $Y_{0i}$ can yield biased estimates of $\tau^t$ and $\tau$. Because of this problem, estimation techniques that are more robust to model misspecification have emerged and among them we find matching.

The basic idea of matching is rather straightforward. Instead of specifying a model of $Y_{0i}$ for the treated, we instead find untreated individuals with similar values of $X_i$ as the treated and use their outcomes in place of $Y_{0i}$. Similarly, matching can be used to find surrogates of $Y_{1i}$ for the controls. In other words, matching builds on the principle that people who share the same characteristics, should on average respond similarly to being treated/not treated. In essence, this is what Assumption 1 states.

Suppose we wish to estimate $\tau^t$ and have a set of $n_1$ treated individuals and $n_0 > n_1$ controls. For each treated individual we then find a control with the same values of $X_i$ as the treated. Unmatched controls are discarded. Once this has been done for each treated individual, we can compare the average outcome in the treatment group with that of the matched control group. Given that Assumptions 1-3 hold, a good estimate of $\tau^t$ can then be obtained.

It is often the case that exact matches are not available for some or all of the treated. For instance, if the dimensionality of $X_i$ is high or if each covariate has many levels, exact matching can become infeasible. Furthermore, if any of the covariates are continuous, exact matching goes from difficult to impossible. In those scenarios, the matching problem becomes finding matches that are as similar as possible on $X_i$ so that the remaining differences in $X_i$ are small enough to not bias the results of the study to any large extent.

Then there is the issue of how many matches are to be chosen for each individual. While the most intuitive approach perhaps is to do 1-to-1 matching, i.e., finding the closest match for each unit, $K$-to-1 matching can also be done. In that case, each individual is matched to the $K$ closest observations from the opposite treatment group. The outcomes of those $K$ matches are then averaged to form an estimate of the missing potential outcome of the individual they were matched to.
The choice of $K$ has a bias-variance trade-off. The larger the $K$, the higher the bias typically becomes, since the second closest match will likely be more dissimilar with regard to $X_i$ than the closest match is. However, by using more than one match, variance is reduced.

Matching estimators of causal effects can be formalised more precisely in mathematical terms. Suppose we take a sample of size $n_1$ treated and $n_0$ controls, making up a total sample size of $n = n_1 + n_0$. Using notation similar to that of Abadie and Imbens (2006), let $J_K(i)$ be the set of indices for the $K$ closest matches to individual $i$. An individual $i' \neq i$ is thus matched to $i$ if $i' \in J_K(i)$. Also, let

$$
\hat{Y}_{0i} = \begin{cases} 
\frac{1}{K} \sum_{i' \in J_K(i)} Y_{i'} & \text{if } T_i = 1, \\
Y_i & \text{if } T_i = 0,
\end{cases}
$$

and

$$
\hat{Y}_{1i} = \begin{cases} 
Y_i & \text{if } T_i = 1, \\
\frac{1}{K} \sum_{i' \in J_K(i)} Y_{i'} & \text{if } T_i = 0,
\end{cases}
$$

be the estimated potential outcomes for individual $i$. A matching estimator of $\tau$ and $\tau^t$ can then be formalised as:

$$
\hat{\tau} = \frac{1}{n} \sum_i \left( \hat{Y}_{1i} - \hat{Y}_{0i} \right)
$$

and

$$
\hat{\tau}^t = \frac{1}{n_1} \sum_{i: T_i = 1} \left( Y_i - \hat{Y}_{0i} \right)
$$

respectively.

Matching can be done with or without replacement, i.e. allowing/not allowing each unit to be used as a match for more than one unit from the opposite treatment group. The main benefit of matching with replacement is that of bias reduction, since not allowing units to be used as matches more than once can by definition mean that units will have to settle for worse matches than they otherwise would have to. Furthermore, the order of which units are matched does not matter when matching with replacement, unlike when matching without. All matching done in this thesis is with replacement.
3.1 Balance

The aim of matching is to match treated and controls in such a way that the distribution of $X_i$ among the matched treated is as similar as possible to that among the matched controls (recall Assumption I). Two groups are said to be balanced with regards to $X_i$ if they have similar distributions of $X_i$. Visualisation and comparisons of high-dimensional distributions are problematic, so usually low-dimensional summaries such as variance ratios and differences in means are presented. Due to covariates often being measured on very different scales, the differences in means are usually standardised by dividing the difference with some pre-matching standard deviation, e.g., the standard deviation among the treated (Stuart 2010) or the average of the standard deviation of the treated and controls (Rosenbaum and Rubin 1985). Love plots, named after one of the authors of Ahmed et al. (2006), are a visual way to display these standardised mean differences. In a Love plot, the absolute standardised mean difference pre- and post-matching are displayed for each covariate in order to visualise if matching was successful or not in improving the covariate balance. An example of a Love plot is shown in Figure 5, where matching is done to improve balance on five covariates $X_1, \ldots, X_5$.

In Figure 5, blue squares and red circles represent pre- and post-matching absolute standardised mean difference respectively. For instance, the $X_3$ covariate had, pre-matching, an absolute standardised mean difference of about 0.19, which matching reduced to around 0.05. With the exception of $X_4$, matching improved the balance for all covariates.

While statistical tests of differences in means post matching could be carried out (and often are, see Ali et al. 2015), it can be argued that such tests do not contribute to the analysis since balance is a within-sample property that should be improved as much as possible, regardless of if some test statistic is significant or not. Imai et al. (2008) has a more thorough discussion on the matter, where they show that randomly discarding observations can lead to a significant t-test becoming insignificant despite no systematic improvement of balance having occurred.

In this thesis we use the absolute standardised mean difference as a measure of balance, though other measures also exist in the literature,
Figure 5: An example of a Love plot. Matching improved balance on all covariates except $X_4$.

see Austin (2009); Stuart et al. (2013); Ali et al. (2015).

4 Propensity Score Matching

Finding good matches gets more and more difficult as the number of covariates grows, an example of what often is referred to as the curse of dimensionality. The paradox of matching is therefore that the higher the dimensionality of $X_i$, the more plausible Assumption 1 might be, but the more difficult it becomes to find close matches. Furthermore, Abadie and Imbens (2006) showed that matching on the Euclidean norm of $X_i$ does not result in root-$n$ consistent estimates of $\tau$ unless $X_i$ contains at most one continuous covariate.

In a landmark paper, Rosenbaum and Rubin (1983) proved that if conditioning on $X_i$ leads to unbiased inference, then so does conditioning on the probability of receiving treatment given $X_i$. This probability is commonly referred to as the propensity score and denoted $e(X_i) = Pr(T_i = 1 \mid X_i)$. Since the propensity score is one-dimensional regardless of the dimensionality of $X_i$, matching on $e(X_i)$ instead of
on $X_i$ directly, greatly simplifies the matching problem. That said, this requires some estimate of $e(X_i)$, which at first glance might indicate that we are back at the model selection problems that matching tries to avoid. However, the goal of matching is not to specify correctly some model for the propensity score (although it certainly helps!), but to find a model such that the matched treated and controls have as similar distributions of $X_i$ as possible. It is possible to have an incorrectly specified propensity score model and still achieve unbiased estimates of causal effects ([Waernbaum, 2010]). If an incorrectly specified propensity score model still leads to matched pairs with very similar distributions of $X_i$, then the misspecification is likely of little importance. Furthermore, how well the matching managed to create two matched groups with similar distributions of $X_i$ can be checked from the data, as discussed in Section 3.1.

4.1 Lasso Estimation of Propensity Scores

Propensity scores are often estimated with a logit or probit model. However, in Paper III we use a logistic lasso ([Tibshirani, 1996]) to estimate them. For ease of notation, we here give a brief introduction to the lasso in the linear regression case. Extensions to other regression models, e.g. logistic regression, are straightforward. A more thorough introduction to the lasso and similar shrinkage techniques can be found in [Hastie et al., 2009].

Suppose that an outcome $Y_i$ can be described by the following regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon_i,$$

where $\varepsilon_i$ is an error term independent of $X_{1i}, \ldots, X_{pi}$. Letting $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$, ordinary least squares (OLS) searches for the solution that minimizes the mean squared error of prediction,

$$\hat{\beta}^{OLS} = \arg \min_{\beta} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{k=1}^{p} \beta_k x_{ki} \right)^2 \right\}.$$

The lasso solution, on the other hand, involves finding the OLS solution subject to the restriction that the sum of the absolute values of the
coefficients does not exceed some predetermined constant. Equivalently, the lasso solution solves:

$$\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{k=1}^{p} \beta_k x_{ki} \right)^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right\},$$

for some $\lambda \geq 0$. When $\lambda = 0$, the lasso solution is equivalent to the OLS solution and when $\lambda = \infty$ the lasso sets all coefficients to 0. As $\lambda$ is increased, the coefficients are all shrunk towards zero (though not by necessity monotonously) which reduces the variance of the parameter estimates at the cost of introducing bias in the estimation of $\beta$. This variance reduction is especially useful when confounders are highly correlated since parameter estimates then can fluctuate wildly between samples. Unlike the closely related ridge regression (see e.g. [Hoerl and Kennard, 1970; Hastie et al., 2009]), the lasso can set coefficients to exactly zero even when $\lambda < \infty$, meaning that it can be used for covariate selection.

In Paper III, we use the logistic lasso to create hundreds of different propensity score models, simply by varying the penalty term $\lambda$. Typically in lasso regression, $\lambda$ is chosen through cross validation. However, as the aim of matching is to minimise differences in post-matching distributions of $X_i$, and not in prediction of $T_i$ per se, we approach the matter from a different point of view. For each of the hundreds of estimated propensity scores, we perform matching and check the resulting balance. The choice of $\lambda$ that leads to the best balance between the treated and control groups is then selected.

5 Coarsening

Coarsening – also known under many other names, such as discretisation or partitioning – is the process of taking a continuous variable and transforming it into a discrete one. More formally, a coarsening function $C_m(X)$ partitions a continuous variable $X$ into $m$ strata such that $C_m(x) = c$ if $\delta_{c-1} < x \leq \delta_c$ for some $\delta_0 < \delta_1 < \ldots < \delta_m$. Typically $\delta_0$ and $\delta_m$ are $-\infty$ and $\infty$ respectively, so that the entire domain of $X$ is covered.

Coarsening can have negative consequences for inference, including bias, reduced power and spurious correlations, see e.g. [Maxwell and De-
so it should not be done haphazardly. However, coarsening covariates also has its merits. Visualisation of distributions is often done through histograms, which by their very nature coarsen covariates. Furthermore, coarsening can be done in order to estimate causal effects within subgroups of a continuous variable, for instance to investigate causal effect heterogeneity between age groups. In matching, coarsening one covariate could lead to better balance on other, perhaps more important, covariates. Coarsening can also be done in order to put bounds on post matching imbalance, a matching strategy discussed in Section 5.1.

5.1 Coarsened Exact Matching

presented a matching strategy where continuous variables are coarsened and where exact matching is then performed on the no longer continuous variables. This approach, called Coarsened Exact Matching (CEM), is appealing due to its simplicity and its ability to put bounds on how far apart two matched observations are allowed to be on any given covariate. Furthermore, since the coarsening of one covariate can be set without affecting that of other covariates, researchers can choose a finer coarsening on covariates believed to be more important to control for more exactly, while letting less important covariates being coarsened more, i.e. having wider partitions. This means that balance can be improved on one covariate (by coarsening it less) without changing the bounds on how far apart observations can be on the other covariates. This property, coined the Monotonic Balance Imbounding (MIB) property , is not shared with all matching techniques. In particular, propensity score matching puts no limits on how far apart two observations can be on any covariate, since the distance it is minimising is that between the estimated propensity scores and not between $X_i$ directly. As a consequence, respecification of a propensity score model to improve balance on one of the covariates can (and often will) worsen the balance on others.

Despite its appealing properties, CEM has a number of drawbacks. To begin with, since CEM only cares about balancing covariates up to the coarsening set by the researchers before matching, increasing the sample size will not generally lead to better balance on the covariates, unless the coarsening becomes finer and finer as the sample size grows.
Moreover, while sidestepping the problem of estimating a propensity score, CEM requires the researcher to determine how to coarsen the covariates. That is, CEM requires a choice of cutpoints $\delta_1, \ldots, \delta_m$ for each of the continuous variables, as well as a choice of how many cutpoints to use, $m$. These cutpoint choices can have a great impact on the resulting estimates of $\tau$ or $\tau^t$. Coarsen too much (too wide intervals) and the matched pairs might have very different distributions of the covariates in question, coarsen too little (too fine intervals) and finding matches becomes exceedingly difficult.

As the dimensionality of $X_i$ grows, one would expect fewer and fewer observations falling in the same multidimensional stratum, making CEM less plausible unless the covariates are highly coarsened. Iacus et al. (2011) point out that not being able to find matches for a given coarsening can indicate that the data is not suitable for causal inference. The lack of overlap between the treated and control groups’ covariate distributions (recall Assumption 2) would then lead to high model dependence.

While CEM bounds the maximum difference in $X_i$ for each matched pair, this is done at the expense of missing matches that are close to each other but happen to fall into different strata of the coarsening. This is illustrated in Figure 6 where a univariate variable $X_k$ is coarsened into three strata. One treated observation, illustrated by the black circle in Stratum 2, is there to be matched to one of two controls, illustrated by black and gray triangles respectively. The closest observation, with regards to $X_k$, is the gray triangle in Stratum 3, but in CEM that would not be a valid match and thus the black triangle of Stratum 2 is instead selected.

An alternative would be to search for the closest match within some pre-specified distance from the observation, not doing any coarsening. This is known as caliper matching and would allow the gray triangle as a match to the black circle in Figure 6. Observations matched this way are not guaranteed to fall into some predetermined strata, which, as we

![Figure 6](image-url)
will discuss in Section 5.2 can be a drawback from a theoretical point of view.

An alternative approach to CEM could be to coarsen only a few of the covariates and then combine exact matching on those coarsened covariates with inexact matching on the remaining covariates, e.g. via propensity score matching (Iacus et al., 2012). This could either be done to put a bound on how different matches can be on certain key covariates or to ensure that matching respects some natural cutoffs, e.g. if legislation is different for different age groups. This type of matching would however lose the MIB property of CEM.

5.2 How to Coarsen

How we coarsen a variable should reflect the aim of the coarsening. When we coarsen a variable to display its distribution in a histogram, we want the resulting discretisation to properly represent the shape of the underlying distribution. If coarsening is done to estimate causal effects separately within different levels of a variable, e.g. within age groups, the coarsening should be chosen in such a way that the resulting subgroups are meaningful. In causal inference, the main goal is to draw conclusions about the effect a treatment has on an outcome of some sort. Thus coarsening, if done at all, should have as its aim to reduce the bias or MSE of the estimated causal effect. In matching, this could be achieved if coarsening one covariate improves the post-matching balance for other covariates or if exact matching on the coarsened covariate balances the covariate better. Coarsening techniques used for histogram creation, such as those of Sturges (1926) and Freedman and Diaconis (1981), do not have this purpose in mind, and can result in unnecessary bias in the estimation of \( \tau \) and \( \tau^t \) if used in causal inference.

Since coarsening can lead to bias, it is of interest to know if there are cases where a coarsened covariate can be used in place of its uncoarsened counterpart and still yield consistent estimates of \( \tau \) and \( \tau^t \). As we show in Paper I, the answer is yes. If some covariates affect the treatment assignment and/or the potential outcomes in a discrete way, then those covariates only need to be controlled for up to that discretisation. Any further adjustment is not needed from a bias point of view. This would be the case if age affects some outcome continuously, but that an individual’s chance of receiving treatment only depends on if he or she
is above some age threshold, say 18 years. It would then be sufficient to
condition on an indicator variable for if the individual is older than 18
years or not, despite age affecting the outcome continuously. However,
even if such a coarsening would exist, it is typically unknown to the
researcher in real world applications. Therefore, Paper I also presents
an algorithm that can be used to search for such a coarsening.

6 Proxy Variables

The assumptions needed for identification of causal effects, outlined in
Section 2.2, are very strong. To assume that no unobserved confounders
exist is often not reasonable, especially if the number of observed co-
variates is low. Sometimes the researcher might have access to variables
believed to be related to one or more unobserved confounders. For in-
stance, suppose that intelligence is thought to be a confounder of some
causal effect of interest. Since intelligence is an abstract concept and
thus unmeasurable, one might be tempted to instead use IQ in its place
in the analysis. Such ”replacement variables” are often referred to as
proxy variables in the literature. This practice of using proxy variables
in place of unobserved confounders is not unproblematic and, as Frost
(1979) showed, simply replacing a covariate with a correlated counter-
part does not by necessity lead to lower bias than just excluding the
covariate altogether. It is thus of interest to know which circumstances
lead to $\tau^t$ and $\tau$ being identified conditional on proxy variables instead
of on the unobserved confounders.

6.1 Parametric Setting

In a linear regression setting, Wooldridge (2010, pp. 67-72) gives con-
ditions that lead to identification of $\tau^t$ and $\tau$ when a proxy variable is
substituted for an unobserved confounder. Suppose that we can write
$Y_{1i}$ and $Y_{0i}$ as linear functions of $X_i$ and $U_i$ in the following way:

\begin{align*}
Y_{1i} &= \alpha_1 + \beta_1 X_i + \gamma U_i + \varepsilon_{1i}, \quad (1) \\
Y_{0i} &= \alpha_0 + \beta_0 X_i + \gamma U_i + \varepsilon_{0i}, \quad (2)
\end{align*}
where \(\varepsilon_0i\) and \(\varepsilon_1i\) are error terms, independent of \(T_i, X_i, U_i\) and each other. The observed outcome \(Y_i = T_iY_{1i} + (1 - T_i)Y_{0i}\) then equals

\[
Y_i = \alpha_0 + (\alpha_1 - \alpha_0)T_i + \beta_0X_i + (\beta_1 - \beta_0)X_iT_i + \gamma U_i + \varepsilon_{Y_i},
\]  

(3)

where \(\varepsilon_{Y_i} = \varepsilon_0i + (\varepsilon_1i - \varepsilon_0i)T_i\). Since the error terms \(\varepsilon_0i\) and \(\varepsilon_1i\) are independent of \(T_i, X_i, U_i\) it follows that \(\varepsilon_{Y_i}\) is uncorrelated with \(T_i, X_i, U_i\). Suppose further that the association between \(U_i, P_i\) and \(X_i\) is such that

\[
U_i = E(U_i \mid P_i, X_i) + \varepsilon_{Ui},
\]

(4)

where \(E(U_i \mid P_i, X_i) = \theta + \psi X_i + \phi P_i\) and where \(\varepsilon_{Ui}\) is an independent error term. Then substituting (4) into (3) gives, after some algebra,

\[
Y_i = \alpha_0 + \gamma \theta + (\alpha_1 - \alpha_0)T_i + (\gamma \psi + \beta_0)X_i + (\beta_1 - \beta_0)X_iT_i + \gamma \phi P_i + \varepsilon_{\tilde{Y}_i},
\]

where \(\varepsilon_{\tilde{Y}_i} = \gamma \varepsilon_{Ui} + \varepsilon_{Y_i}\) and is uncorrelated with \(T_i, X_i\). Noting that here \(\tau = \alpha_1 - \alpha_0 + (\beta_1 - \beta_0)E(X_i)\), we see that \(\tau\) is identified through linear regression of \(Y_i\) on \(T_i, X_i, P_i\) and \(X_iT_i\). Likewise \(\tau^t\) can also be identified since \(\tau^t = \alpha_1 - \alpha_0 + (\beta_1 - \beta_0)E(X_i \mid T_i = 1)\).

### 6.2 Nonparametric Setting

The conditions in [Wooldridge (2010)] lead to identification of \(\tau\) and \(\tau^t\) if one is able to specify linear models such as (1) and (2). However, they fail to do so if \(U_i\) affects \(Y_{1i}\) and \(Y_{0i}\) in different ways, i.e., if \(\gamma\) in (1) and (2) was instead \(\gamma_1\) and \(\gamma_0 \neq \gamma_1\) respectively. Furthermore, if we are not willing to make parametric assumptions then the conditions are also insufficient. That said, the solution for the parametric case gives some insight into where a nonparametric solution can be found. Note that because \(P_i\) and \(X_i\) contain all information about \(U_i\) that is relevant for unbiased prediction of \(Y_i\), we achieve identification of \(\tau\) and \(\tau^t\). A similar reasoning can be made for the nonparametric case. Consider the DAG presented in Figure 7.

In Figure 7 the proxy variable, \(P_i\), (together with \(X_i\)) blocks all causal paths between the unobserved confounder \(U_i\) and the potential outcomes \(Y_{1i}\) and \(Y_{0i}\). As a consequence of this, all paths between \(T_i\) and the potential outcomes are blocked conditional on \(X_i\) and \(P_i\), and thus \(\tau\) and \(\tau^t\) are identified. We can also note that identification is also achieved if \(P_i\) instead blocks all paths from \(U_i\) to \(T_i\), as shown in Figure 8.
Figure 7: A DAG representation where $P_i$ is a proxy for an unobserved confounder $U_i$.

Figure 8: Another type of proxy variable, where conditioning on $P_i$ and $X_i$ instead closes all paths between $U_i$ and $T_i$.

Again, all paths between $T_i$ and the potential outcomes are blocked conditional on $X_i$ and $P_i$, leading to $\tau$ and $\tau^t$ being identified. The difference between these two types of proxy variables thus lies in that conditional on $X_i$ and $P_i$, the first type makes $U_i$ independent of $Y_{ji}$, whereas the second makes $U_i$ independent of $T_i$.

In Paper II we formalise these two types of proxy variables and prove that they lead to nonparametric identification of $\tau$ and $\tau^t$. As a technical aside, it is worth noting that the proofs do not rely upon the existence of any compatible DAG. So while DAGs such as those of Figure 7 and 8 are helpful for aiding understanding, they are not required for the existence of proxy variables. The reader interested in the distinction between independence in a graph and independence in a probability distribution is referred to Pearl (1988, Chapter 3).
7 Instrumental Variables to Test Proxy Properties

An instrumental variable is a variable that is a cause of the treatment but that does not affect the potential outcomes other than through its influence on the treatment assignment. Due to the latter, conditioning on an instrumental variable is not needed for identification of causal effects if Assumption 1 holds. However, instrumental variables can be used to examine if a (believed) proxy variable indeed makes treatment assignment independent of the unobserved confounders, conditional on the proxy and the other observed covariates.

Recall from Section 2.1 that conditioning on a collider can introduce dependence between two variables that otherwise were independent of each other. This fact lets us use instrumental variables to investigate if a believed proxy stops all unobserved confounding. Consider the DAG in Figure 9 below. There, \( Z_i \) is an instrument as it affects the treatment but not the potential outcomes \( Y_{ji} \). In this case \( P_i \) and \( X_i \) blocks all paths from \( T_i \) to \( Y_{ji} \), so \( P_i \) is a proxy variable for \( U_i \).

![Figure 9: An illustration of an instrumental variable \( Z_i \).](image)

Suppose now instead that \( P_i \) did not stop all influence from \( U_i \) on \( T_i \), i.e., that it is not a proxy variable as defined in Paper II. This is illustrated in Figure 10.

![Figure 10: \( P_i \) fails to block all direct influence that \( U_i \) has on \( T_i \).](image)

In Figure 10 we note that \( T_i \) is a collider for the association between \( Z_i \) and \( U_i \). As pointed out in Section 2.1, this means that \( Z_i \) and \( U_i \) become
conditionally dependent given $T_i$. Since $Y_{ji}$ in turn is dependent on $U_i$, this would mean that $Z_i$ and $Y_{ji}$ become dependent when conditioning on $T_i$ and that this dependency is not removed by also conditioning on $P_i$ and $X_i$. This is not the case if $P_i$ is a proxy for $U_i$, as in Figure 9. Since $Y_{0i} = Y_i$ conditional on $T_i = 0$, we can investigate if $Z_i$ and $Y_i$ are dependent given $(T_i = 0, P_i, X_i)$. If they are, we then have evidence against the assumption that $P_i$ is a proxy variable for $U_i$. We use this fact in Paper IV to examine the supposed proxy property of one of the variables.

8 The Swedish Social Insurance Agency

In this section we give a short description of the Swedish Social Insurance Agency and describe a joint cooperation studied in Papers III and IV of this thesis.

The Swedish Social Insurance Agency (SIA) is, as the name implies, the agency tasked with handling most kinds of social insurance in Sweden, including sick leave benefits, parental leave and disability benefits. In 2016, the total costs of insurance from the SIA totalled 227 billion Swedish Kronor, or about 5% of Sweden’s GDP (Försäkringskassan, 2017). Since being a resident of Sweden for longer than a year, or living in Sweden with the intent of staying for over a year, gives access to some basic insurance, a large proportion of the population is covered. The extent of the insurance depends on, e.g., employment status and income. For individuals with employment, the first 14 days of sick leave are paid by the employer, excluding the first work day. After those 14 days, the individual can receive sick leave benefits from the SIA. For the unemployed, sick leave benefits from SIA can instead start from day two (Försäkringskassan, 2017).

To qualify for sick leave benefits, an individual needs a certificate from his or her medical doctor. Once such a certificate is obtained, caseworkers at the SIA then determine the eligibility of the individual to the sick leave benefits.

In 2011, the Swedish government tasked the SIA and the Public Employment Service to enhance their cooperation (Socialdepartementet, 2011). As part of this enhanced cooperation, joint assessment meetings were implemented at the SIA. In such a meeting, an individual on
sick leave meets with representatives from both the SIA and the Public Employment Service in order to investigate his or her working ability. These meetings can then lead to further interventions from the Public Employment Service, with the aim of helping the individual return to work.

One measure of sick leave is the total extent of the sick leave (TESL). The TESL measures the proportion of an individual’s work ability he or she receives sick leave benefits for. To clarify the meaning of TESL, consider two individuals, A and B. Individual A works half-time but is on paid sick leave from all of that. Individual B usually works full-time and is on sick leave for 50% of it. Then individual A has a TESL of 1 (since he or she is on sick leave for all of his or her work supply) while individual B has a TESL of 0.5 (since he or she is on paid sick leave for half of his or her work supply). So despite the two individuals receiving sick leave benefits for the same amount of hours, their TESL differs. Thus TESL is a relative measure, not an absolute one, of how much sick leave an individual is on.

In Paper III and IV, we investigate the effect of being called to a joint assessment meeting on TESL. Paper III is a study protocol, describing the matching design and presenting pre- and post-matching balance. This was done without access to the outcome data for the treated, ensuring that the matching design was not influenced by any estimated causal effects. Paper IV then carries out the analysis.

9 Summary of Papers

9.1 Paper I

Coarsening is the process of taking a continuous covariate and partitioning it into strata so that the resulting covariate is discrete. This practice of coarsening covariates is not unproblematic and can lead to biased estimation of causal effects. However, if coarsening some covariates leads to improved balance on other more important covariates, this could lower the bias. Paper I defines sufficient coarsening, i.e., such coarsening that the treatment assignment is independent of the uncoarsened covariate, conditional on its coarsened counterpart and the other covariates. We show that average causal effects are identified despite some of the covariates being coarsened, given that the coarsening per-
formed is sufficient coarsening. This means that, if such a coarsening is known, covariates can be coarsened without introducing bias. However, sufficient coarsening functions are typically unknown to the researcher. We therefore present a data-driven algorithm to search for such a coarsening. The algorithm is evaluated through a simulation study, where exact matching on the coarsened covariate is combined with propensity score matching on the remaining covariates. We present the bias and RMSE in estimation of the average causal effect.

9.2 Paper II

Differences in covariate distributions between treated and controls need to be adjusted for in order to draw causal inference about the effect a treatment has on some outcome. Such adjustment needs to be made for confounders, whether or not they are observed. While observed confounders can be handled through, e.g., matching, unobserved covariates are more problematic since they per definition cannot be directly controlled for. One approach is to replace unobserved covariates with correlated counterparts, often called proxy variables. However, this replacement of covariates does not necessarily solve the problem and can even increase the bias. Paper II defines two types of proxy variables that lead to nonparametric identification of average causal effects despite unobserved confounders being present. Examples of when such proxies could be obtained in practice are also given. The proxy variables defined within the article are contrasted with the proxy variable defined by Wooldridge (2010).

9.3 Paper III

The Swedish Social Insurance Agency and the Public Employment Service were in 2011 tasked to enhance their cooperation. As a consequence, joint assessment meetings were implemented, where individuals on sick leave meet with representatives from the two agencies. It is of interest to know if this enhanced cooperation is successful in reducing the total extent of sick leave. In Paper III we present a study protocol to investigate this through a matching design. In order to ensure that no estimated causal effect influenced the matching design, the outcome for the treated was not available at the time of writing the paper. Paper III
describes the matching design, where propensity score matching, based on the logistic lasso, is combined with exact matching on important covariates. One of these important covariates is a prognosis of the expected duration of the sick leave. We argue that this prognosis is a proxy variable for unobserved confounders.

9.4 Paper IV

In Paper IV we perform the analysis described in Paper III, now with access to all outcome data. We estimate the effect of being called to a joint assessment meeting on future total extent of the sick leave for those called. The results indicate that the joint collaboration is detrimental from a total extent of sick leave point of view. Subgroup analyses are performed and give similar pictures as the main analysis. An instrumental variable approach is used to investigate the plausibility of the proxy variable assumption connected to the prognosis covariate.

10 Final Remarks and Further Research

This thesis is concerned with improving balance on both observed and unobserved covariates in observational studies. We showed in Paper I that, under certain conditions, we can coarsen covariates and still get unbiased estimates of causal effects. However, the coarsening algorithm presented was developed for coarsening one variable at a time, and a multivariate coarsening approach would be useful. Furthermore, other algorithms for searching out coarsening would be of interest, the main crux of the problem lying in the need for good conditional independence tests. Coarsening based upon discrete associations with the outcome, is also of interest to evaluate and contrast to contrast with coarsening based upon association with the treatment.

In Paper IV we point out that it is plausible that the treated are deemed in more need of work rehabilitation interventions than their matched counterparts. The aim of the matching, especially on the prognosis variable, was to make the matched controls similar to the treated in all aspects, including unobserved confounders. The instrumental variable approach used in Paper IV gives some support for the claim of no unobserved confounding remaining. However, a sensitivity analysis
of how much unobserved confounding is required for the results of our study to become insignificant, would shed further light on the question.

References


