DISS. ETH NO. 26810

Essays on Causal Inference in Economics: Methods and Applications

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

ELIAS MOOR

M.Sc. in Economics, Ludwig-Maximilians-Universität München

born on 03.01.1989

citizen of Vordemwald AG, Switzerland

accepted on the recommendation of

Prof. Dr. Antoine Bommier (ETH Zurich), examiner
Prof. Dr. Massimo Filippini (ETH Zurich), co-examiner
Prof. Dr. Michael Lechner (University of St. Gallen), co-examiner

2020

Acknowledgements

Writing this dissertation would not have been possible without the help and support of many people. First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Antoine Bommier. I greatly appreciated his professional guidance and advice at all stages of this dissertation. He supported me in every possible way and gave me the opportunity to pursue my own research ideas. I am very grateful for the time at his chair.

I would further like to thank Prof. Massimo Filippini and Prof. Michael Lechner for serving as co-supervisors and evaluating my research. Moreover, I would like to thank Prof. Hans Gersbach for chairing the examination committee.

I wish to express my sincere gratitude to Markus Hersche, with whom I had the privilege to work on the second and third chapter of this dissertation. Working with him has shown me how exciting and enriching collaborative research can be. I am incredibly grateful for his support throughout the PhD, from facilitating the beginning to encouragement towards the end. I really appreciate that he was always available to discuss and give feedback.

During my time as PhD student at the Chair of Integrative Risk Management and Economics, I was fortunate to meet and benefit from many colleagues. I would like to thank Amélie, Arnaud, Aurore, Bruno, Claudio, Daniel Ha., Daniel He., François, Hélène, Irina, Jean-Philippe, Jere, Marie-Charlotte, Sabine, and Wei. Moreover, I would like to thank the people I was fortunate to meet at ZUE, especially Adrian, Afsoon, Ewelina, Fabio, Florian, Martin, Moritz, Oliver, Philippe, Samuel, and Tobias, who greatly enriched my time.

I would like to express my special thanks to Nicola, for her unconditional support during the PhD, and for the love and joy she brings to my life. Our relationship is invaluable to me. Our discussions were a tremendous help throughout my studies.

I am extremely grateful to my parents, Renate and Werner, and to my siblings, Raphael and Salome. They supported me in every possible way and were always there for me. A special thanks goes to my longtime friends Beat, Jimmy, Nicholas, Noah, Raphael, and Samuel. I really appreciate our friendship and the time we spend together. I would also like to thank my friends from the Marzili Highway, and all the friends I got to know during my studies in Bern and Munich.

Abstract

Chapter 1 discusses estimation of average treatment effects under the assumption of unconfoundedness. I study regression estimators, matching on the propensity score, inverse probability weighting, and hybrid methods such as bias-corrected matching and doubly robust estimators. These estimators require the estimation of the conditional outcome means, the propensity score, or both. In empirical applications, these functions are often estimated with ordinary least squares or logit. In this chapter I additionally consider machine learning methods to estimate these functions. To analyze the treatment effect estimators, I conduct two Monte Carlo simulation studies in which the true treatment effects are known. I find that machine learning based estimators are in many cases more accurate in terms of treatment effect root-mean-square error than estimators relying on ordinary least squares or logit. The differences are more pronounced when the underlying relationships are nonlinear and nonadditive, or when selection into treatment is strong.

Chapter 2 discusses identification and estimation of causal intensive margin effects.[1] The causal intensive margin effect is defined as the treatment effect on the outcome of individuals with a positive outcome irrespective of whether they are treated or not, and is of interest for outcomes with corner solutions. The main issue is to deal with a potential selection problem that arises when conditioning on positive outcomes. We propose using difference-in-difference methods - conditional on positive outcomes - to estimate causal intensive margin effects. We derive sufficient conditions under which the difference-in-difference estimator identifies the causal intensive margin effect. In contrast to standard difference-in-difference methods, two monotonicity assumptions are additionally required to identify the causal intensive margin effect. We apply the methodology to estimate the causal intensive margin effect of reaching the full retirement age on working hours.

Chapter 3 estimates the labor supply response when the spouse reaches the full retirement age.[2] We exploit the age difference within couples and changes in pension legislation in Switzerland to identify the causal effect. In contrast to the majority of previous contributions, we estimate the effect not only on labor market participation (*extensive margin*), but also on working hours (*intensive margin*). We find that the labor force participation of women decreases, on average, by approximately 3 percentage points in response to the spouse reaching the full retirement age. We find no evidence that men adjust their labor force participation when their wives reach the full retirement age. At the intensive margin, we find only small and mostly non-significant effects for both men and women, although older workers use working hours to adjust their labor supply. We argue that the response can be explained by *complementarity in leisure* and *liquidity* effects.

[1] Chapter 2 is joint work with Markus Hersche. Both authors contributed equally to this chapter.

[2] Chapter 3 is joint work with Markus Hersche. Both authors contributed equally to this chapter.

Zusammenfassung

Kapitel 1 behandelt die Schätzung durchschnittlicher kausaler Effekte (Behandlungseffekte) unter der Annahme, dass alle Störfaktoren beobachtbar sind. Die Analyse umfasst Regressionsschätzer, Matching-Methoden basierend auf der bedingten Behandlungswahrscheinlichkeit, Schätzer basierend auf inverser Wahrscheinlichkeitsgewichtung, sowie hybride Schätzer. Diese Schätzer erfordern die Schätzung der bedingten Ergebnismittelwerte und/oder der bedingten Behandlungswahrscheinlichkeit. In empirischen Anwendungen werden diese Funktionen oft mit der Methode der kleinsten Quadrate bzw. mit logistischer Regression geschätzt. In diesem Kapitel betrachte ich zusätzlich Methoden des maschinellen Lernens zur Schätzung dieser Funktionen. Um die Schätzer der Behandlungseffekte zu analysieren, führe ich zwei Monte-Carlo-Simulationen durch, in denen die wahren Behandlungseffekte bekannt sind. Die Analyse zeigt, dass Schätzer, die auf Methoden des maschinellen Lernens basieren, den durchschnittlichen Behandlungseffekt in vielen Fällen genauer schätzen als Schätzer, die auf der Methode der kleinsten Quadrate bzw. der logistischen Regression basieren. Die Unterschiede sind grösser, wenn die zugrundeliegenden Beziehungen nicht linear und additiv sind oder wenn die Behandlungsselektion stark ist.

Kapitel 2 befasst sich mit der Identifikation und der Schätzung von kausalen Mengenentscheidungseffekten.[3] Der kausale Mengenentscheidungseffekt wird definiert als der Behandlungseffekt - beispielsweise der Effekt einer politischen Massnahme - auf das Ergebnis von Personen, welche ein positives Ergebnis aufweisen, unabhängig davon, ob sie behandelt wurden oder nicht. Dieser Effekt ist insbesondere bei Ergebnissen mit Randlösungen von Interesse. Das Hauptproblem besteht darin, ein potenzielles Selektionsproblem zu lösen, welches bei der Konditionierung auf positive Ergebnisse entsteht. Wir schlagen vor, Differenz-von-Differenzen (DvD) Methoden - konditioniert auf Individuen mit positivem Ergebnis - anzuwenden, um den kausalen Mengenentscheidungseffekt zu schätzen. Wir leiten hinreichende Bedingungen her, unter welchen die DvD Methode den kausalen Mengenentscheidungseffekt identifiziert. Im Vergleich zu herkömmlichen DvD Methoden werden zusätzlich zwei Monotonieannahmen benötigt, um den kausalen Mengenentscheidungseffekt zu identifizieren. Wir wenden die Methode an, um den kausalen Mengenentscheidungseffekt des Erreichens des ordentlichen Rentenalters auf die Arbeitsstunden zu schätzen.

Kapitel 3 schätzt die Veränderung des Arbeitsangebotes, wenn der Partner das ordentliche Rentenalter erreicht.[4] Wir nutzen Altersunterschiede von Paaren sowie eine Reform des Frauenrentenalters, um den kausalen Effekt zu identifizieren. Im Gegensatz zu den meisten bisherigen Studien schätzen wir den Effekt nicht nur auf die Arbeitsmarktbeteiligung, sondern auch auf die Mengenentscheidung (Anzahl Arbeitsstunden). Die Analyse zeigt, dass die Arbeitsmarktbeteiligung von Frauen im Durchschnitt um drei Prozentpunkte abnimmt, sobald der Partner das ordentliche Rentenalter erreicht. Wir finden keine Hinweise, dass Männer ihre Arbeitsmarktbeteiligung anpassen, wenn die Partnerin das Rentenalter erreicht.

[3] Kapitel 2 wurde in Zusammenarbeit mit Markus Hersche verfasst.
[4] Kapitel 3 wurde in Zusammenarbeit mit Markus Hersche verfasst.

Bezüglich der Anzahl Arbeitsstunden finden wir nur geringe und nicht signifikante Effekte, sowohl für Frauen als auch für Männer. Wir argumentieren, dass die Effekte sowohl durch Komplementarität in der Freizeit, als auch durch Liquiditätseffekte erklärt werden können.

Contents

Appendices

Introduction

Understanding causal relationships is at the core of empirical economics and other empirical social sciences. In many cases, the focus is on understanding the causal effect of one variable on an outcome of interest.[5] Empirical economists are for example interested in the causal effect of a minimum wage introduction on employment, the causal effect of a deductible on health care utilization, or the causal effect of attending a job training program on subsequent earnings. Knowledge of causal effects is essential for evidence-based policy design and evaluation. A job training program might, for example, target decreasing the unemployment duration, or increasing subsequent earnings. However, such a job training program is also costly. A natural starting point of an evaluation is to assess whether the job training program, on average, decreases unemployment duration or increases subsequent earnings of its participants.

In order to discuss causal effects, the notion of a causal effect needs to be formally defined. Two frameworks are widely used when addressing causality: the potential outcome framework and the directed acyclic graph (DAG) approach.[6] In the potential outcome framework - which is used in most parts of this thesis - we start with the definition of the treatment variable. The treatment variable denotes the variable that can be manipulated - for example participation in a job training program. For each possible level of the treatment variable, a corresponding potential outcome indicates the value the outcome variable would take if the individual were to receive this level of treatment. With a binary treatment, each individual thus has two potential outcomes:[7] the potential outcome in the case of treatment, and the potential outcome in the case of no treatment.[8] The potential outcome framework then defines the causal effect as the difference between the two potential outcomes.[9] However, we observe only one of the potential outcomes for each individual. This is the "fundamental problem of causal inference", as Holland (1986) puts it. If an individual was treated, we observe the potential outcome in the case of treatment. If an individual was not treated, we observe the potential outcome in the case of no treatment. This implies that the causal effect, at the level of an individual, is never actually observed.

[5] That is, the interest is in so-called *effects of causes* (Holland, 1986). A different objective would be to study the *causes of effects*.

[6] For a comparison of the two frameworks from an economics perspective, see Imbens (2019).

[7] Instead of an individual, the subject could also be a firm, a market, or a country.

[8] With a binary treatment, I refer to the two levels of the treatment variable as *treated* or *treatment*, and *untreated*, *no treatment*, *not treated*, or *control*.

[9] See Section 1.3 for an introduction to the potential outcome framework.

However, under certain conditions, it is possible to estimate, for example, the average of the causal effect. This requires a sample of both treated and untreated individuals, with information on at least the outcome and the treatment variable. Depending on the setting and the methods applied, we require additional variables, such as personal characteristics or past outcomes. Learning causal effects from a sample includes three important steps: identification, estimation, and inference. Identification is the task of demonstrating that the causal parameter - e.g. the average causal effect - is identified in the population. A causal parameter such as the average causal effect is identified in the population if we are able to rewrite the generally unobserved average causal effect - using identifying assumptions - into a quantity that is observed in the population. Estimation is the computational part that takes the data as input and produces an estimate of the causal parameter as output. Finally, inference is the part that generalizes the results from the sample to the population.[10] This requires taking into account the uncertainty in the estimation. In the potential outcome framework, there are two main sources of uncertainty. First, we only observe one potential outcome for each individual, never both. We do not know what the outcome of a given individual would be if treatment were assigned differently. Second, we are observing only a sample from the population. In another sample, the estimate of the average causal effect would likely differ (Imbens, 2004). Inference involves estimating a confidence interval for the causal parameter. Using the confidence interval, we can then conduct hypothesis tests; for example, whether the average causal effect is statistically significantly different from zero.

A key focus in the potential outcome framework is on the treatment assignment mechanism. If treatment is randomly assigned, for example in a randomized experiment, identification and estimation of average causal effects is rather straightforward (Imbens & Rubin, 2015).[11] A simple and unbiased estimator for the average causal effect is given by the difference in mean outcomes of treated and untreated observations. In practice however, running a randomized experiment is often not feasible, for financial, ethical, or other reasons (Athey & Imbens, 2017). For example, if we are interested in the causal effect of college attendance on earnings, it is unimaginable to run an experiment in which some students would be randomly assigned to attend college (treatment group), while others would be prohibited from attending college (control group). In these cases, we have to rely on observational data to estimate the causal effect of a treatment.

However, estimating causal effects is more difficult with observational data. A major problem is that the relationship between treatment and outcome is potentially confounded by other variables. This means that there are some variables, called confounders, that have an effect both on whether the individual is treated or not and on the outcome. In the job training example, education is a potential confounder. Individuals with a higher education

[10] This is the case if one is interested in the population average treatment effect. It is also possible to conduct inference for the sample average treatment effect.

[11] Randomized experiments have their own drawbacks. For example the usually high costs, the problem of individuals assigned to the treatment group refusing to participate in the treatment, or conversely the problem of individuals assigned to the control group trying to obtain alternative (similar) treatments (Smith & Todd, 2005).

might be more aware of the benefits of a job training program and would therefore be more likely to enroll for such training. Besides that, individuals with a higher education tend to have higher earnings. A difference in earnings between job training participants and non-participants could be the result of differences in such confounding variables, and not because of a causal effect of treatment. Thus, ignoring confounders could lead to a biased estimate of the causal effect of attending job training on earnings. To overcome this problem, the literature on treatment effects and program evaluation relies on different methods to isolate the causal effect. The typical toolbox includes methods that rely on the unconfoundedness assumption, difference-in-difference, synthetic control, instrumental variable approaches, and regression discontinuity designs.[12] These methods differ both in terms of identifying assumptions and data requirements.

Two of the aforementioned methods are at the core of this dissertation: methods that rely on the unconfoundedness assumption and difference-in-difference methods. The former methods basically assume that selection into treatment is based on observable characteristics.[13] Difference-in-difference methods, by contrast, allow for some selection into treatment based on unobservable characteristics, at the cost of making an assumption about time trends. Although these methods are well-studied and widely applied in practice, there is room for potential improvements and extensions.

Consider the methods that rely on unconfoundedness. In practice, commonly used estimators are regression or matching on the propensity score. In many cases, ordinary least squares (OLS) regression is used to estimate the conditional outcome means, and logit or probit to estimate the propensity score.[14] In this context, there are at least two potential improvements. First, instead of using either an estimator that relies only on the conditional outcome means, or only on the propensity score, it could be beneficial to use hybrid estimators. Hybrid estimators use both the conditional outcome means and the propensity score. The idea of hybrid methods is to provide additional robustness. A question that arises in this context is whether there is any cost - for example in terms of variance of these estimators - for providing additional robustness. Second, instead of using OLS, logit, or probit, it could be beneficial to employ more flexible methods to estimate the conditional outcome means and the propensity score. This could be of interest in settings with many covariates, or if the underlying relationships are unknown and possibly nonlinear and/or nonadditive.[15] Again, a question arises as to whether there is any cost for allowing flexible functions.

Difference-in-difference methods are commonly applied to estimate the total effect of a treatment.[16] Using data from pre- and post-treatment periods, difference-in-difference

[12] For an overview see for example Athey and Imbens (2017), Abadie and Cattaneo (2018), Imbens and Wooldridge (2009), and Angrist and Pischke (2009).

[13] See footnote 8 in Chapter 1 for a more rigorous definition.

[14] The conditional outcome means denote the conditional expectation of the outcome given treatment and covariates. The propensity score denotes the conditional probability of treatment given covariates.

[15] Following Diamond and Sekhon (2013), nonlinearity refers to the presence of quadratic terms (or higher order polynomials) in the true underlying model, while nonadditivity refers to the presence of interaction terms in the true underlying model.

[16] Difference-in-difference usually identifies the average treatment effect on the treated (ATT).

is able to deal with the selection problem arising from unobserved confounding. When the outcome of interest is non-negative and has a mass point at zero, we are sometimes interested in a decomposition of the total effect into an extensive and an intensive margin effect. To estimate the intensive margin effect, the estimation sample is usually restricted to individuals with a positive outcome. However, this creates an additional selection problem. A possible extension of the standard difference-in-difference estimator is therefore to analyze the case in which the estimation sample is restricted to individuals with a positive outcome.

This dissertation addresses these potential improvements and extensions in Chapters 1 and 2. Chapter 3 presents an application of causal effect estimation in empirical economics. The thesis thus consists of both methodological and applied chapters on the identification and estimation of causal effects with observational data.

Chapter 1 focuses on methods that rely on the unconfoundedness assumption. Examples of these methods include regression estimators, matching estimators, inverse probability weighting, and doubly robust estimators. These methods all rely on estimating either the propensity score, the conditional outcome means, or both. There are thus two central decisions: which treatment effect estimator to use (regression, matching, doubly robust, etc.), and how to estimate the conditional outcome means and the propensity score. The goal of Chapter 1 is to shed light on these two decisions. I focus on three research questions. First, do hybrid methods estimate treatment effects more accurately than estimators that rely either only on the propensity score or only on the conditional outcome means? Second, does machine learning based estimation of the propensity score and/or the conditional outcome means improve treatment effect estimation, compared to logit or OLS based estimation? Third, how does the accuracy of the estimators depend on changes in the degree of linearity and additivity of the underlying functions, or on changes in the strength of selection into treatment? To answer these questions, I conduct two Monte Carlo simulation studies in which the true treatment effects are known. Moreover, I conduct a within-study comparison in the spirit of LaLonde (1986). The main contribution of this chapter is to provide empirical insights on the two described potential improvements of treatment effect estimation under unconfoundedness. First, I provide insights on the use of hybrid estimators as alternative to estimators that rely only on the propensity score or only on the conditional outcome means. Second, I provide insights on the use of machine learning methods to estimate the propensity score and/or the conditional outcome means. In simulation studies or within-study comparisons, a central question is to what extent the findings can be generalized. The design of the simulation studies affect the results. If, for example, the underlying relationships are all linear and additive, flexible machine learning methods are likely to be of limited use. The goal of this chapter is to demonstrate that there are indeed cases where hybrid methods or machine learning based estimation are beneficial.

Chapter 2 is concerned with a methodological extension of the difference-in-difference estimator.[17] The idea is to use difference-in-difference to estimate the causal intensive mar-

[17] Chapter 2 is joint work with Markus Hersche. Both authors contributed equally to this chapter.

gin effect. The causal intensive margin effect is defined as the treatment effect on the outcome of individuals with a positive outcome irrespective of whether they are treated or not. Even if treatment is randomly assigned, a mean comparison of treatment and control groups with positive outcomes does not identify the causal intensive margin effect without additional assumptions (Angrist, 2001). The main issue is a potential selection problem that arises when conditioning on positive outcomes. Difference-in-difference methods were developed to address selection problems. Using data from pre- and post-treatment periods, difference-in-difference allows for some selection on unobservables. The main contribution is to derive sufficient conditions under which the difference-in-difference estimator - conditional on positive outcomes - identifies the causal intensive margin effect. In contrast to standard difference-in-difference methods, two monotonicity assumptions are additionally required to identify the causal intensive margin effect. These monotonicity assumptions are indeed rather strong and thus limit the range of potential applications. Chapter 2 provides the methodological foundation for an estimator used in Chapter 3.

Chapter 3 comprises an application of causal effect estimation in empirical economics.[18] As many developed countries are forced to reform their pension systems, there is a need for a detailed understanding of the labor supply behavior of older workers. The full retirement age represents one of the main policy instruments for the government. A large body of literature has focused on the estimation of the *direct effect*, that is, the labor supply response of individuals directly affected by pension reforms (Mastrobuoni, 2009). However, the majority of older workers are married, and several studies indicate that older couples coordinate their exit from the labor force (Hospido & Zamarro, 2014). As a result, changes in incentives of one member of the couple may have spillover effects on the labor supply of the spouse (*indirect effect*). This chapter estimates the causal effect of the spouse reaching the full retirement age on labor supply. In contrast to the majority of previous contributions, we estimate the effect not only on labor market participation (*extensive margin*), but also on working hours (*intensive margin*).

[18] Chapter 3 is joint work with Markus Hersche. Both authors contributed equally to this chapter.

References

Abadie, A., & Cattaneo, M. D. (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, *10*, 465–503.

Angrist, J. D. (2001). Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors. *Journal of Business & Economic Statistics*, *19*(1), 2–28.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.

Athey, S., & Imbens, G. (2017). The State of Applied Econometrics - Causality and Policy Evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Diamond, A., & Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics & Statistics*, *95*(3), 932–945.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 968.

Hospido, L., & Zamarro, G. (2014). Retirement Patterns of Couples in Europe. *IZA Journal of European Labor Studies*, *3*(12).

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity : A Review. *The Review of Economics and Statistics*, *86*(1), 4–29.

Imbens, G. W. (2019). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *NBER Working Paper Series, No. 26104*.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

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

Mastrobuoni, G. (2009). Labor Supply Effects of the Recent Social Security Benefit Cuts: Empirical Estimates Using Cohort Discontinuities. *Journal of Public Economics*, *93*(11-12), 1224–1233.

Smith, J. A., & Todd, P. E. (2005). Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? *Journal of Econometrics*, *125*, 305–353.

Chapter 1

Machine Learning Based Estimation of Average Treatment Effects under Unconfoundedness

1.1 Introduction

Estimating causal effects by means of observational data can be a challenging task. The main difficulty is that the relationship between treatment and outcome is potentially confounded. That is, there exist some variables that have an effect on both, treatment and outcome. As a consequence, differences in outcomes between the treatment group and the control group can arise for two reasons. First, because of the causal effect of treatment on the outcome, and second, because of differences in the confounding factors. If one is interested in the causal effect, it is therefore the second reason that causes problems.

The idea behind methods that rely on the unconfoundedness assumption is to eliminate outcome differences that arise due to differences in the confounding factors. The central assumption is that all confounding factors are observed.[1] Examples of such methods include regression estimators, matching estimators, inverse probability weighting, and hybrid methods such as doubly robust estimators. These methods rely on estimating either the conditional outcome means, the propensity score, or both.[2] In practice, these functions are often estimated with conventional methods such as ordinary least squares (OLS) or logit. The researcher then has to decide which variables to include in the model, and how to include them. In many cases, the specifications are linear and additive, without quadratic or interaction terms.[3] However, it is possible that the true underlying relationships are not linear and additive. If the models are misspecified, this may lead to a biased estimate of the causal effect. As an alternative to OLS and logit, one might consider estimating the conditional outcome means and the propensity score by machine learning (ML) methods.

[1] See footnote 8 for a more rigorous definition.

[2] For a definition of the conditional outcome means and the propensity score, see Section 1.4.

[3] The definition of linearity and additivity follows Diamond and Sekhon (2013), see footnote 15.

In recent years, ML methods have gained interest both in industry and in academia. One of the reasons is the success of ML methods in problems concerned with prediction. In this type of problems, the goal is to predict an outcome on the basis of a set of covariates as accurately as possible. Although ML methods have been predominantly used for such prediction problems, there is a growing literature on these methods being employed for estimating causal effects.[4] However, as Goller, Lechner, Moczall, and Wolff (2019) note, it is not entirely clear that the successes achieved with prediction problems will also apply in settings where one is interested in causal effects. A crucial difference is that the outcome of interest is observed in prediction problems. It is thus possible to compare different models on the basis of performance measures such as the root-mean-square error (RMSE). This is not possible when estimating causal effects, because the true causal effect - at the individual level - is not observed.

In the context of estimating treatment effects under the assumption of unconfoundedness, estimation involves two main decisions: first, the decision on which treatment effect estimator to use (regression, matching, doubly robust, etc.); and second, how to estimate the conditional outcome means and/or the propensity score. These two decisions form the basis of my research questions.

This chapter aims to answer the following research questions. First, does estimating the propensity score and/or the conditional outcome means with machine learning methods increase accuracy of treatment effect estimation compared to OLS or logit based estimation (within-estimator comparison)?[5] Second, do hybrid methods estimate treatment effects more accurately than estimators that rely either only on the propensity score or only on the conditional outcome means (between-estimator comparison)? Third, how does the accuracy of the estimators depend on changes in a) the degree of linearity and additivity in the relationships between treatment and covariates and between outcome and covariates, and b) the strength of selection into treatment?

To answer these questions I conduct two simulation studies. The advantage of the simulation studies is that the true causal effect is known. It is therefore possible to compare the performance of different treatment effect estimators in terms of RMSE, bias, and variance. The first simulation study is based on a simulation of Diamond and Sekhon (2013), while the second is based on the LaLonde (1986) dataset and follows the simulation design of Busso, DiNardo, and McCrary (2014). The two simulation studies differ substantially. The first is based on a stylized data generating process (DGP) and characterized by a constant and additive treatment effect, good overlap between treated and untreated observations, and a treated fraction of approximately 50%. The second is based on a real dataset. It allows for heterogeneous treatment effects, exhibits bad overlap between treated and untreated observations, and includes only approximately 17% treated observations. In the first simulation

[4] For an overview, see for example Athey and Imbens (2017), Mullainathan and Spiess (2017), Athey (2018), and Athey and Imbens (2019).

[5] Accuracy of treatment effect estimation refers to the RMSE of the treatment effect estimators, see Section 1.6.5.

study, I additionally consider different misspecification scenarios. Moreover, I conduct a small within-study comparison in the spirit of LaLonde (1986).

I find that machine learning based estimators often estimate the treatment effect more accurately than estimators relying on OLS and/or logit. The differences between machine learning based estimators and estimators relying on OLS and/or logit are small when the underlying relationships are linear and additive, or when selection into treatment is weak. However, when the underlying relationships become nonlinear and nonadditive, or when the strength of selection into treatment increases, the performance of machine learning based estimators can be substantially better than OLS and/or logit based estimators. Moreover, I find that hybrid estimators do not generally outperform estimators that rely either only on the propensity score or only on the conditional outcome means. However, hybrid estimators are in most cases among the estimators with the lowest RMSE. In many cases, hybrid estimators exhibit the lowest bias, sometimes at the cost of an increased variance. It might therefore be advantageous to employ hybrid methods to guard against misspecification, especially if one is more concerned about bias than variance.

The remainder of the chapter is structured as follows. Section 1.2 presents a short review of the literature. Section 1.3 describes the setting and the assumptions. Section 1.4 outlines the structure of the estimation procedure. Section 1.5 presents an overview of the methods used to estimate the conditional outcome means and the propensity score, while Section 1.6 discusses the treatment effect estimators. The two simulation studies are presented in Section 1.7 and 1.8. Section 1.9 presents the within-study comparison. Section 1.10 presents results from a supplementary analysis, analyzing how the estimator performance depends on cross-fitting and repeated sample splitting. Finally, Section 1.11 concludes the chapter.

1.2 Literature Review

This chapter relates to the literature that evaluates estimators for average treatment effects under the assumption of unconfoundedness. The challenge in such an evaluation is that the true treatment effect is generally unobserved. Consequently, in order to make an evaluation, we need either to create an artificial setting in which the true treatment effect is observed, or use a setting in which a reliable proxy for the true treatment effect is available. These two possibilities represent the two main strands of the literature to which this chapter is related. The first strand is the literature that uses Monte Carlo simulation studies to assess the performance of the estimators. The second strand is the literature on within-study comparisons, where the true treatment effect is proxied by an estimate from a randomized experiment. Compared to simulation studies, within-study comparisons have the disadvantage that there is no guarantee that the unconfoundedness assumption holds. For this reason, within-study comparisons are sometimes used to ”test” the unconfoundedness assumption, and not to provide evidence on the performance of estimators.

1.2.1 Monte Carlo Simulation Studies

The simulation study literature contains at least four different approaches for the design of the simulation. The first approach uses what Advani, Kitagawa, and Słoczyński (2019) call stylized DGPs. Stylized DGPs are often characterized by a) covariates that are drawn from normal or Bernoulli distributions, b) parametrically specified associations between the variables, and c) a good overlap between treated and untreated observations (Frölich, 2004; Lunceford & Davidian, 2004; Zhao, 2004). Such settings provide a high level of control over the DGP, at the cost of less realistic DGPs and therefore lower external validity. The simulation in Section 1.7 is based on a stylized DGP that was previously used by Setoguchi, Schneeweiss, Brookhart, Glynn, and Cook (2008), Lee, Lessler, and Stuart (2010), Diamond and Sekhon (2013), Pirracchio, Petersen, and Van Der Laan (2015), and Cannas and Arpino (2019).

To increase the external validity of the simulations, the simulation literature has moved towards using empirical Monte Carlo studies. In this type of simulations, the goal is to generate datasets with distributions that are as similar as possible to real datasets. Advani et al. (2019) distinguish between structured empirical Monte Carlo studies and placebo empirical Monte Carlo studies. Structured empirical Monte Carlo studies fit parametric distributions to a real dataset. New samples are then generated from the fitted distributions (Abadie & Imbens, 2011; Busso et al., 2014; Diamond & Sekhon, 2013). The simulation in Section 1.8 is based on the structured Monte Carlo simulation design of Busso et al. (2014). Placebo empirical Monte Carlo studies directly draw control observations from a real dataset and simulate a placebo treatment. The placebo treatment can be based on a propensity score estimated in the full dataset - as for example in Huber, Lechner, and Wunsch (2013), and Goller et al. (2019) - or by a matching approach, as for example in Frölich, Huber, and Wiesenfarth (2017). Most recently, Athey, Imbens, Metzger, and Munro (2019) propose a fourth approach which utilizes Generative Adversarial Networks (GANs). The idea of GANs is to adjust a *generator* neural network that generates simulated data, until a *discriminator* neural network is no longer able to distinguish between simulated and real data. The usefulness of empirical Monte Carlo studies to rank treatment effect estimators is questioned by Advani et al. (2019). They suggest the use of a series of estimators and a comparison of the range of treatment effect estimates.

1.2.2 Within-Study Comparisons

Second, this chapter is related to the literature on within-study comparisons. This literature combines experimental data with nonexperimental data to conduct an evaluation of treatment effect estimators. There are two different approaches within this literature. The first approach relies on the assumption that the randomized experiment gives an unbiased estimate of the true treatment effect. Subsequently, the experimental control group is substituted with a nonexperimental comparison group to mimic a situation in which no experiment is available. It is then analyzed whether the treatment effect estimators are able to recover the true treatment effect from the experiment. This procedure was applied in the landmark study by LaLonde (1986), as well as in subsequent contributions (Dehejia & Wahba, 1999, 2002). One drawback of this approach is that the experimentally estimated treatment effect can be quite noisy. The second approach aims for a more direct analysis of the selection bias. Instead of substituting the experimental control group with a nonexperimental comparison group, the experimental treatment group is substituted with a nonexperimental comparison group.[6] The benefit of this approach is that the true treatment effect is known to be zero, since no observation was actually treated. This approach was adopted, for example, by Heckman, Ichimura, and Todd (1997), Heckman, Ichimura, Smith, and Todd (1998), and Smith and Todd (2005), and is used in the within-study comparison in Section 1.9.

1.3 Setting and Assumptions

I consider the Rubin Causal Model (RCM) with an outcome Y and a binary treatment D. Each individual i is characterized by two potential outcomes. The potential outcome in the case of treatment is denoted by Y_i^1, the potential outcome in the case of no treatment is denoted by Y_i^0. Since we observe one of the two potential outcomes only, the individual treatment effect, defined as $\tau_i = Y_i^1 - Y_i^0$, is never observed. In this chapter I am interested in the *average treatment effect* (ATE), defined as

$$\tau_{ATE} = E(Y_i^1 - Y_i^0) \, ,$$

and the *average treatment effect on the treated* (ATT),[7] defined as

$$\tau_{ATT} = E(Y_i^1 - Y_i^0 | D_i = 1) \, .$$

In the following, I refer to the ATE and the ATT as *treatment effects*. In addition, each individual is characterized by a vector of covariates X_i. For a comprehensive introduction to the RCM, see Imbens and Wooldridge (2009) or Imbens and Rubin (2015).

In the setting I consider, identification of the ATE and ATT relies on the following assumptions: unconfoundedness, overlap, stable unit treatment value assumption (SUTVA), and no effect of treatment on covariates.

The unconfoundedness assumption formally states that treatment is independent of potential outcomes, conditional on covariates $X_i = x$, i.e.

$$(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i \mid X_i = x \;\text{ for all } x \text{ in the support of } X_i. \tag{1.1}$$

[6] The treatment indicator is then replaced to indicate membership to the experimental control group.

[7] More specifically, the interest is in the population versions - i.e. the population average treatment effect, and the population average treatment effect on the treated. Therefore, the ATE captures the expected causal effect for an individual chosen at random from the population. The ATT captures the expected causal effect for an individual chosen at random from the subpopulation who received treatment.

The assumption requires that all confounders - i.e. all variables affecting both treatment *and* outcome - are observed.[8] The assumption implies that treatment is as good as randomly distributed conditional on $X_i = x$.[9] For identification of the ATT, the unconfoundedness assumption can be relaxed to $Y_i^0 \perp\!\!\!\perp D_i | X_i = x$. The unconfoundedness assumption (or similar versions) is sometimes also called conditional independence assumption, selection on observables, or exogeneity.

The overlap assumption concerns the joint distribution of treatment and covariates. Formally, the assumption states that the conditional probability of treatment (propensity score) is strictly between zero and one, i.e.

$$0 < p(x) < 1 \text{ for all } x \text{ in the support of } X_i, \tag{1.2}$$

where $p(x) = P(D_i = 1 | X_i = x)$ is the propensity score. The overlap assumption requires that, for all x in the support of X_i, both treated and untreated observations are available. For identification of the ATT, the overlap assumption can be relaxed to $p(x) < 1$. The combination of unconfoundedness and overlap is also called *strong ignorability* (Rosenbaum & Rubin, 1983).

Furthermore, the Stable Unit Treatment Value Assumption (SUTVA) is assumed,

$$Y_i = Y_i^1 D_i + Y_i^0 (1 - D_i) \, . \tag{1.3}$$

The SUTVA states that the observed outcome of any individual i depends only on the treatment status of individual i, but not on any other treatment status of individual $j \neq i$. Hence, spill-over and general equilibrium effects are assumed to be absent.

Lastly, it is assumed that there is no effect of treatment on covariates,

$$X_i^1 = X_i^0 = X_i \, . \tag{1.4}$$

Conditioning on covariates which are themselves affected by the treatment would either remove part of the total causal effect we are interested in, or introduce a collider bias, depending on the direction of the causal relationship between the covariate and the outcome.[10]

[8] This description of the unconfoundedness assumption is only correct in conjunction with the assumption that there is no effect of treatment on covariates, see assumption (1.4). Furthermore, as Huber (2019) notes, the unconfoundedness assumption is also satisfied if - conditional on the observed covariates - the effects from unobserved confounders on treatment and outcome are ”blocked”. The concept of ”blocking”, also known as ”d-separation”, stems from the directed acyclic graph (DAG) approach to causality - see e.g. Pearl (2009), or Imbens (2019) for a comparison of the DAG approach to the potential outcomes approach.

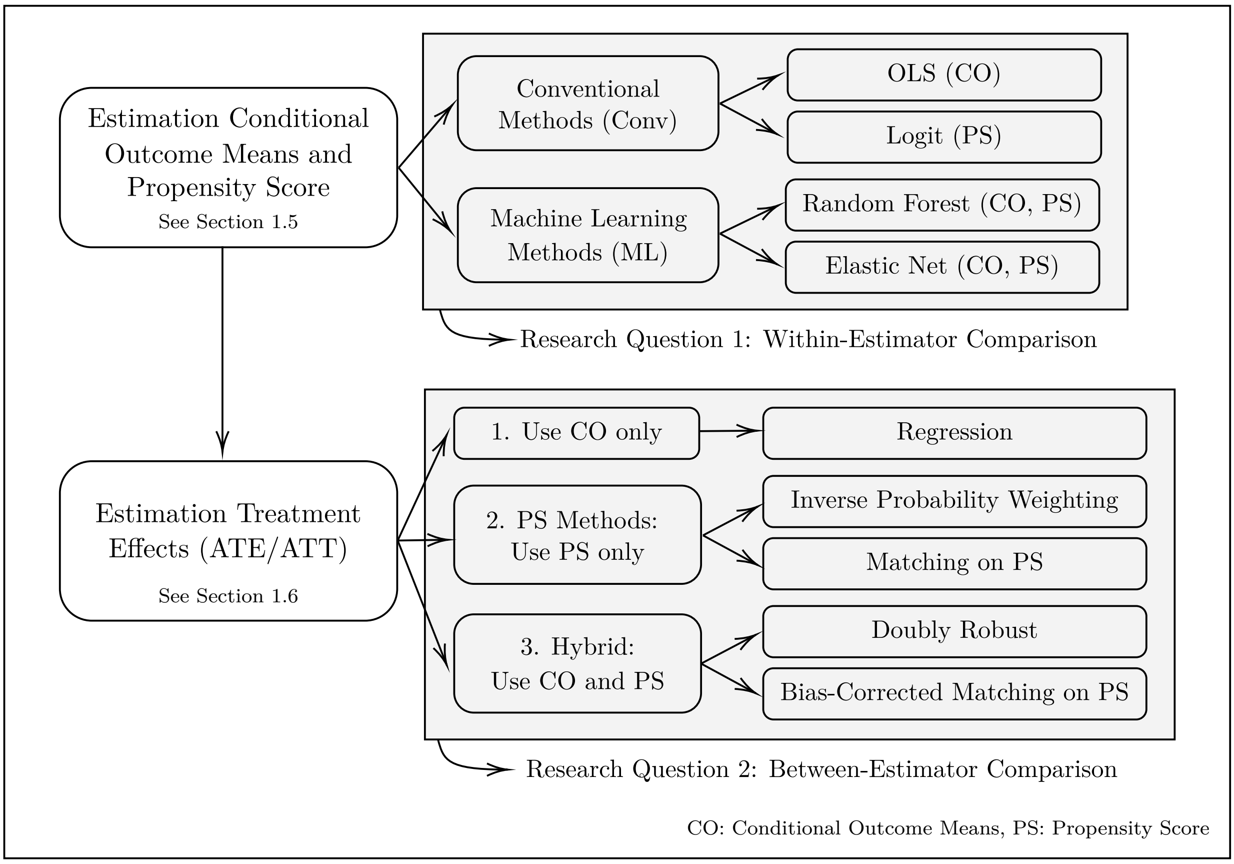
[9] A sufficient weaker version of the unconfoundedness assumptions is mean independence, i.e. $E(Y_i^1 | D_i, X_i) = E(Y_i^1 | X_i)$ and $E(Y_i^0 | D_i, X_i) = E(Y_i^0 | X_i)$. As Imbens (2004) points out, it might be hard to find a case where the weaker mean independence assumption is fulfilled, but the stronger unconfoundedness assumption is violated.

[10] A collider bias occurs when we condition on a collider variable. A collider variable is a variable that is affected by both the treatment and the outcome. Conditioning on a collider creates an association between treatment and outcome, even if the true treatment effect is zero.

1.4 Structure of the Estimation Procedure

This section presents an overview of the estimation procedure. Figure 1.1 illustrates the different parts, which are briefly described below and explained in more detail in Sections 1.5 and 1.6.

Figure 1.1: Structure of the Estimation Procedure


Estimation Conditional Outcome Means and Propensity Score
See Section 1.5
Conventional Methods (Conv)
OLS (CO)
Logit (PS)
Machine Learning Methods (ML)
Random Forest (CO, PS)
Elastic Net (CO, PS)
Research Question 1: Within-Estimator Comparison
Estimation Treatment Effects (ATE/ATT)
See Section 1.6
1. Use CO only
Regression
2. PS Methods: Use PS only
Inverse Probability Weighting
Matching on PS
3. Hybrid: Use CO and PS
Doubly Robust
Bias-Corrected Matching on PS
Research Question 2: Between-Estimator Comparison
CO: Conditional Outcome Means, PS: Propensity Score

As described in Section 1.3, I am interested in estimating the ATE (simulation Section 1.7) or the ATT (simulation Section 1.8). Estimation of these treatment effects follows a two-step procedure. In the first step, the conditional outcome means and/or the propensity score are estimated. The conditional outcome means are defined as

$$m_1(x) = E(Y_i|D_i = 1, X_i = x) \text{ and } m_0(x) = E(Y_i|D_i = 0, X_i = x)\,.$$

The propensity score is defined as

$$p(x) = P(D_i = 1|X_i = x)\,.$$

In this chapter, I estimate these functions both with the conventional methods OLS (conditional outcome means) and logit (propensity score), as well as with the machine learning methods random forest and elastic net (see boxes at the top of Figure 1.1). These methods are described in Section 1.5. To answer the first research question, I investigate whether machine learning based estimation of treatment effects is more accurate than estimation based on conventional methods.

In the second step, the fitted values of the conditional outcome means and/or the propensity score are plugged into the treatment effect estimators. I consider five different treatment effect estimators (see boxes at the bottom of Figure 1.1). These estimators can be classified into three categories. The *Regression* estimator in the first category uses only the conditional outcome means to estimate the treatment effects. The propensity score methods in the second category use only the propensity score to estimate the treatment effects. This category includes the *Inverse Probability Weighting (IPW)* and the *Matching on PS* estimators. The third category consists of hybrid estimators using both the conditional outcome means as well as the propensity score to estimate the treatment effects. This category includes the *Doubly Robust* and the *Bias-Corrected (BC) Matching on PS* estimators. The treatment effect estimators are described in Section 1.6. To answer the second research question, I investigate whether hybrid methods estimate treatment effects more accurately than estimators that rely either only on the propensity score or only on the conditional outcome means.

1.5 Estimation Conditional Outcome Means and Propensity Score

In this section I describe the methods used to estimate the conditional outcome means and the propensity score. The fitted values of these functions are plugged into the treatment effect estimators described in Section 1.6. In the last part of this section, I describe certain additional estimation issues.

1.5.1 Conventional Methods

The conventional based estimators use OLS to estimate the conditional outcome means and logit to estimate the propensity score. Ordinary least squares and logit are widely used methods in empirical economics. For reasons of space, I do not describe these methods here. A comprehensive overview of OLS and logit can be found, for example, in Cameron and Trivedi (2005).

1.5.2 Machine Learning Methods

In this section I present the basics of random forests and elastic net. There are several reasons for using these two methods. Both machine learning methods are very intuitive, easily implemented using standard software packages, and do not require extensive hyperparameter tuning (see section 1.5.3). As Goller et al. (2019) note, the two methods follow a different approach to approximate the unknown propensity score or conditional outcome means. While random forests approximate the unknown function locally, elastic net aims to approximate the unknown function globally.[11] For a more in-depth overview, see for example Hastie, Tibshirani, and Friedman (2001), James, Witten, Hastie, and Tibshirani (2013), and Efron and Hastie (2016).

Random Forests

A random forest is a collection of many regression or classification trees (Breiman, 2001). A regression or classification tree is a recursive partitioning of the covariate space into exhaustive and mutually exclusive subgroups.[12] In each subgroup of the partition, a simple model is fitted to obtain the predicted outcome for observations with covariates corresponding to this subgroup. Often, the simple model is just a constant, and the predicted outcome for observations with covariates corresponding to subgroup k is given by the group average of observations in group k.

To obtain the recursive partitioning of the covariate space, the algorithm starts with the unpartitioned covariate space. Only binary splits are considered. For the first split, the algorithm aims to find a splitting variable v and a splitting point s which reduce a given loss function the most. For example, the regression tree algorithm used in this chapter minimizes

$$\min_{v,s} \left[\sum_{x_i \in R_1(v,s)} (y_i - c_1)^2 + \sum_{x_i \in R_2(v,s)} (y_i - c_2)^2 \right], \tag{1.5}$$

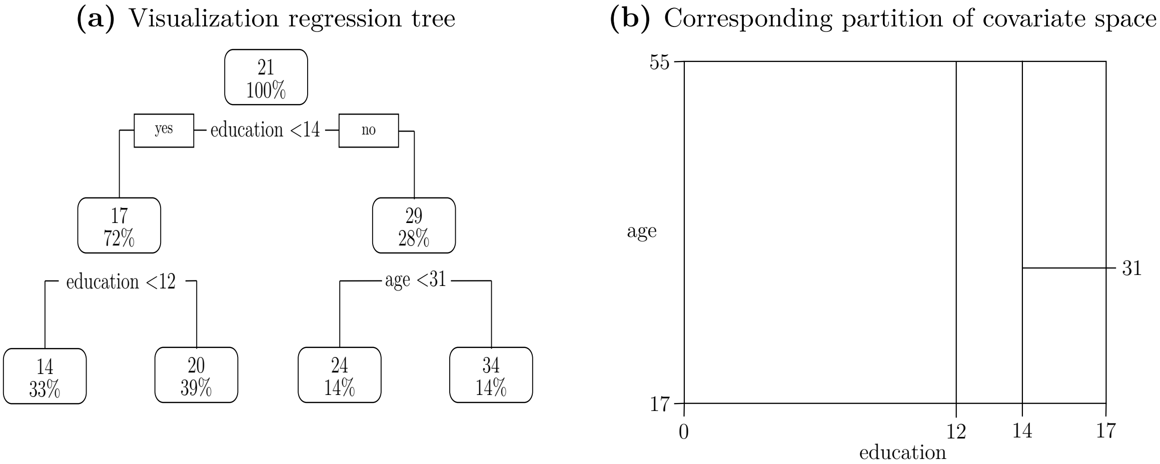
where $R_1(v,s) = \{X | X_v \leq s\}$ and $R_2(v,s) = \{X | X_v > s\}$ are the subgroups defined by the binary covariate split, $c_g = \frac{1}{n_g} \sum_{x_i \in R_g(v,s)} y_i$ for $g \in \{1,2\}$ is the average outcome in subgroup $R_g(v,s)$, and n_g is the number of observations in subgroup $R_g(v,s)$.

After each split, the algorithm continues to find a new splitting variable v and a splitting point s in the subgroups defined by the previous split. Figure 1.2 presents an illustrative example of a regression tree (left panel) and the corresponding partitioning of the two-dimensional covariate space (right panel). The regression tree was fit on the LaLonde (1986) data with earnings in 1978 as outcome (in 1000), and age and education as covariates.

[11] Instead of elastic net, Goller et al. (2019) consider Lasso.

[12] If the outcome is continuous, the tree is called a regression tree. If the outcome is categorical, the tree is called a classification tree.

Figure 1.2: Illustration of a regression tree


(a) Visualization regression tree

(b) Corresponding partition of covariate space

The recursive partitioning is repeated until some stopping criterion is reached, for example if fewer than five observations were to end up in a subgroup. Due to the recursive partitioning, regression trees automatically model interactions between covariates.

A random forest is a large collection of individual trees. Usually, the individual trees are grown deeply. This means that the recursive splitting is applied many times. The resulting individual tree has lower bias, but higher variance than a shallow tree with only a few splits. The final prediction is obtained by averaging the predictions of all individual trees. To decrease the variance of the random forest estimator, the individual trees are decorrelated. To decorrelate the individual trees, two forms of randomness are introduced in a random forest. First, for each new tree, a bootstrap sample is drawn from the original data, and only the bootstrap sample is used to grow the tree. Second, at each split, only a randomly drawn subset of the covariates is considered to split upon. The size of the randomly drawn subset is sometimes regarded as a tuning parameter.

Elastic Net

Elastic net is a penalized regression method (Zou & Hastie, 2005). Penalized regression means that the objective function includes a penalty term on the coefficients. As a result, the estimated coefficients are shrunken towards zero, and some coefficients are directly set to zero (variable selection). To estimate the conditional outcome means, I use the linear regression version. In this case, the elastic net estimator is the solution to

$$\min_{\beta_0,\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i'\beta)^2 + \lambda\Big[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\Big] \ . \tag{1.6}$$

To estimate the propensity score, I use the logistic regression version. In this case, the elastic net estimator is the solution to

$$\min_{\beta_0,\beta} -\left\{\frac{1}{N}\sum_{i=1}^{N}[y_i(\beta_0 + x_i'\beta) - \log(1+\exp(\beta_0 + x_i'\beta))]\right\} + \lambda\Big[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\Big] \ , \tag{1.7}$$

where the propensity score is modelled as $p(x) = \frac{1}{1+\exp[-(\beta_0+x'\beta)]}$. The coefficients are estimated with penalized maximum likelihood.

For $\alpha = 1$, the estimator is called least absolute shrinkage and selection operator (Lasso) and penalizes the absolute values of the coefficients.[13] For $\alpha = 0$, the estimator is called ridge regression and penalizes the squared values of the coefficients. For $0 < \alpha < 1$, the estimator is called elastic net and employs a combination of Lasso and ridge penalization. Due to the Lasso penalization, some coefficients are set directly to zero (Efron & Hastie, 2016, p. 305).[14] The strength of the penalization is determined by the penalty parameter λ. The larger λ, the stronger the penalization and the more the coefficients are shrunken towards zero. For $\lambda \to \infty$, all coefficients are set to zero. For $\lambda = 0$, the OLS solution occurs.

Compared to OLS, elastic net allows for the bias-variance trade-off.[15] Due to the penalization, elastic net is a biased estimator. The bias-variance trade-off is controlled by the penalty parameter λ. The larger λ, the higher the bias and the smaller the variance. The overall goal is to achieve a smaller (mean squared) prediction error because the variance is reduced more than the introduced (squared) bias.

1.5.3 Practical Estimation Issues

For the estimation of $m_1(x)$, $m_0(x)$, and $p(x)$, I apply cross-fitting as proposed in Chernozhukov et al. (2018). In the default specification, I randomly split the data into five folds. I set aside one fold and estimate the functions $m_1(x)$, $m_0(x)$, and $p(x)$ using the remaining four folds. Then, using the fitted functions, I predict $\widehat{m_1}(X_i)$, $\widehat{m_0}(X_i)$, and $\widehat{p}(X_i)$ for all observations in the remaining fold. This procedure is repeated until each fold has been left out once and thus until each observation has predictions $\widehat{m_1}(X_i)$, $\widehat{m_0}(X_i)$, and $\widehat{p}(X_i)$. In Section 1.10, I investigate the effect of the number of cross-fitting folds on the performance of the treatment effect estimators.

Many machine learning methods rely on tuning parameters. Tuning parameters have to be chosen prior to the estimation, and influence the bias and the variance of the estimator. For the random forest, this is for example the number of randomly chosen variables to consider for a split. For elastic net, this is the penalty term λ and the mixing parameter α.

In practice, the tuning parameters are often chosen using cross-validation. The goal of cross-validation is to obtain an estimate of the out-of-sample performance of an estimator - i.e. to estimate how well the estimator is able to predict unseen data. First, the researcher chooses a grid of possible values for the tuning parameter. Then, for each possible value (or combination) of the tuning parameter(s), the out-of-sample performance is estimated via cross-validation. Similar to cross-fitting, the data is randomly divided into K folds. Then,

[13] The covariates are standardized prior to the estimation. Otherwise, asymmetric penalization is imposed, which is usually not desired.

[14] Generally, a combination of Lasso and ridge regression does not set as many coefficients to zero as Lasso only.

[15] Ordinary least squares is an unbiased estimator and thus does not allow for the bias-variance trade-off.

$K - 1$ folds are used to fit the function, followed by predicting the outcomes in the left-out fold. The procedure is repeated until each observation was once in the left-out fold and thus until each observation has an out-of-sample predicted outcome. Next, a performance measure such as the RMSE is calculated. The whole process is repeated for a new value in the grid of tuning parameters. Finally, the tuning parameter corresponding to the best out-of-sample performance is selected as *final* tuning parameter.[16]

In this chapter, I use 10-fold cross-validation to choose the penalty parameter λ for elastic net. For computational reasons, I do not use cross-validation to choose the mixing parameter α, but instead set $\alpha = 0.5$. Moreover, for random forests, I do not apply cross-validation to choose the tuning parameters. Instead, I follow the software package default specifications.[17]

1.6 Estimation Treatment Effects (ATE/ATT)

In this section I describe the five treatment effects estimators. All estimators described in this section use predicted values of the propensity score and/or the conditional outcome means, for which estimation is described in Section 1.5. Throughout this section, I assume a random sample of N observations of which N_T are treated.

1.6.1 Regression Estimator

Regression estimators (see Hahn (1998), Heckman et al. (1997), Heckman, Ichimura, and Todd (1998), Imbens, Whitney, and Ridder (2006)) estimate the ATE as the difference in the predicted values of the conditional outcome means, averaged over the sample.[18] Hence, the ATE is estimated as

$$\hat{\tau}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} \left[\widehat{m_1}(X_i) - \widehat{m_0}(X_i) \right] . \tag{1.9}$$

Similarly, the ATT is estimated as the difference between the observed outcomes of the treated and their predicted values of the conditional outcome mean fitted in the untreated sample. Hence, for the ATT, we do not need to estimate the conditional outcome mean

[16] There are alternatives in choosing the *final* tuning parameter, e.g. the one-standard error rule, which chooses the least complex model whose error is within one standard error of the best model.

[17] In the *R* package *randomForest*, the parameter defining the number of randomly chosen variables to consider for a split, for example, is called *mtry* and its default is $p/3$ for regression problems and $\sqrt{p}$ for classification problems, where p is the number of variables.

[18] An alternative version would plug in the observed outcome, i.e.

$$\hat{\tau}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} \left[D_i Y_i + (1 - D_i)\widehat{m_1}(X_i) - D_i \widehat{m_0}(X_i) - (1 - D_i) Y_i \right] . \tag{1.8}$$

If the average of the observed outcome for the treated and the control is equal to the average predicted outcome for the treated and the untreated, the two versions are equivalent.

fitted in the treated sample. The ATT is estimated as

$$\hat{\tau}_{ATT} = \frac{1}{N_T}\sum_{i=1}^{N} D_i\Big[Y_i - \widehat{m_0}(X_i)\Big]\ . \tag{1.10}$$

It is important to note that the regression estimator using OLS to estimate $m_1(x)$ and $m_0(x)$ is generally different from the OLS regression of Y on D and X.[19] This implies that an OLS regression of Y on D and X does not estimate the ATE or ATT except for special cases (see Angrist and Pischke (2009) and Abadie and Cattaneo (2018)). A valid alternative to estimating $m_1(x)$ and $m_0(x)$ separately using OLS is to estimate a single OLS regression of Y on a constant, D, X, and $D(X - \bar{X})$, where $\bar{X}$ is the sample average of X.

1.6.2 Propensity Score Methods

Inverse Probability Weighting

Inverse probability weighting (see Horvitz and Thompson (1952), Robins, Rotnitzky, and Zhao (1994), Hirano, Imbens, and Ridder (2003)) relies on reweighting the outcome such that treated individuals with a high propensity score receive a smaller weight than treated individuals with a low propensity score. Similarly, untreated individuals with a low propensity score receive a smaller weight than untreated individuals with a high propensity score. The idea of reweighting is again to make the treated and untreated groups comparable in terms of their covariate distributions.
The ATE is estimated by the average difference in reweighted outcomes

$$\hat{\tau}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_iY_i}{\widehat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\widehat{p}(X_i)}\right]. \tag{1.11}$$

Similarly, the ATT is estimated as

$$\hat{\tau}_{ATT} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_iY_i}{c} - \frac{(1-D_i)Y_i\widehat{p}(X_i)}{(1-\widehat{p}(X_i))c}\right]. \tag{1.12}$$

where c is the fraction of treated, i.e. $c = \frac{1}{N}\sum_{i=1}^{N} D_i$. A potential problem with *IPW* is that the weights can become very large when the estimated propensity scores are very close to zero or one (Imbens & Wooldridge, 2009). As a result, the variance of the estimator

[19] In a setting where X is saturated - i.e. with J dummy variables d_j, where $d_{ij} = \mathbb{1}\{X_i = x_j\}$ - the coefficient of D in the OLS regression of Y on D and X equals

$$\tau_{OLS} = \sum_{j=1}^{J}(E[Y_i|D_i = 1, X_i = x_j] - E[Y_i|D_i = 0, X_i = x_j])\omega_j\ ,$$

with weights $\omega_j = \frac{Var(D_i|X_i=x_j)P(X_i=x_j)}{\sum_l^J Var(D_i|X_i=x_l)P(X_i=x_l)}$ (Abadie & Cattaneo, 2018; Angrist & Pischke, 2009). The ATE estimator can be written in a similar form, but with weights $\omega_k = P(X_i = x_k)$, and the ATT estimator with weights $\omega_k = P(X_i = x_k|D_i = 1)$. As Angrist and Pischke (2009) and Abadie and Cattaneo (2018) point out, OLS of Y on D and X estimates a variance-weighted ATE.

increases. For this reason, I apply a trimming rule (see Section 1.6.4).

Matching on Propensity Score

Rosenbaum and Rubin (1983) have demonstrated that unconfoundedness implies independence of treatment and potential outcomes conditional on the propensity score, that is $(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i | p(X_i)$. As a result, there is no confounding problem for observations with the same propensity score. The idea of matching on the propensity score (see Heckman, Ichimura, and Todd (1998), Dehejia and Wahba (2002), Abadie and Imbens (2016)) is to estimate the missing potential outcome by averaging the outcomes of the nearest neighbors in the opposite treatment group. The nearest neighbors are the observations with the smallest absolute difference in the propensity score. In this chapter, I apply one-to-one nearest neighbor matching with replacement. One-to-one means that only one nearest neighbor is considered for each individual. Matching with replacement implies that a given observation can be matched to more than one observation of the opposite treatment group. The result of Rosenbaum and Rubin (1983) is based on the true propensity score $p(x)$. In practice, the true propensity score is usually not observed.[20] The true propensity score is then replaced by the estimated propensity score.

Let $\ell(i)$ denote the nearest neighbor of individual i in the opposite treatment group. Formally, $\ell(i)$ equals integer $j \in \{1, \ldots, N\}$, if $D_j \neq D_i$, and

$$|\widehat{p}(X_j) - \widehat{p}(X_i)| = \min_{k: D_k \neq D_i} |\widehat{p}(X_k) - \widehat{p}(X_i)|, \tag{1.13}$$

where $\hat{p}(\cdot)$ is the estimated propensity score.[21]

The missing potential outcome of each observation is imputed by the outcome of the nearest neighbor. Then, the ATE is estimated as the average difference between the observed outcome and the estimated missing potential outcome,

$$\hat{\tau}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} \Big[D_i(Y_i - Y_{\ell(i)}) + (1 - D_i)(Y_{\ell(i)} - Y_i) \Big] \, . \tag{1.14}$$

Similarly, the ATT is estimated as the average difference between the observed outcome and the estimated missing potential outcome among the treated,

$$\hat{\tau}_{ATT} = \frac{1}{N_T} \sum_{i=1}^{N} D_i \Big[Y_i - Y_{\ell(i)} \Big] \, . \tag{1.15}$$

[20] An exception where the true propensity score is known are experiments, where treatment assignment is either completely randomized or based on observed characteristics. However Abadie and Imbens (2016) have shown that it is beneficial to use the estimated propensity score even when the true propensity score is known.

[21] Notation and definition based on Imbens and Wooldridge (2009).

1.6.3 Hybrid Methods

Finally, I consider hybrid methods that combine the aforementioned estimators. The goal of hybrid methods is to make the estimators more robust.

Weighting and Regression (Doubly Robust)

Doubly robust estimation (see Robins et al. (1994), Robins and Rotnitzky (1995), Bang and Robins (2005)) combines weighting and regression. The estimator is consistent if either the conditional outcome means or the propensity score is correctly specified.
Chernozhukov et al. (2018) discuss the use of machine learning methods for this estimator in a high-dimensional setting.
The ATE is estimated as

$$\hat{\tau}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left[\widehat{m_1}(X_i) - \widehat{m_0}(X_i) + \frac{D_i(Y_i - \widehat{m_1}(X_i))}{\widehat{p}(X_i)} - \frac{(1-D_i)(Y_i - \widehat{m_0}(X_i))}{1-\widehat{p}(X_i)}\right]. \quad (1.16)$$

Similarly, the ATT is estimated as

$$\hat{\tau}_{ATT} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_i(Y_i - \widehat{m_0}(X_i))}{c} - \frac{(1-D_i)(Y_i - \widehat{m_0}(X_i))\widehat{p}(X_i)}{(1-\widehat{p}(X_i))c}\right]. \quad (1.17)$$

where c is the fraction of treated, i.e. $c = \frac{1}{N}\sum_{i=1}^{N} D_i$.

Matching and Regression (Bias-Corrected Matching on PS)

Bias-corrected matching combines matching and regression. The idea of bias-corrected matching was introduced by Abadie and Imbens (2011), after Abadie and Imbens (2006) had demonstrated that matching can be biased.[22] The findings of Abadie and Imbens (2006, 2011) relate primarily to matching on more than one continuous covariate. As Imbens (2004) points out, matching on the propensity score matches only on a single variable. As a result, the bias vanishes asymptotically.

The idea of bias-corrected matching is to adjust the imputed potential outcome by an estimate of this bias. As an estimate of this bias, the above authors suggest the difference in the conditional outcome means.

Therefore, the missing potential outcome of each observation is imputed by the outcome of the nearest neighbor, adjusted by the difference in the conditional outcome means. Then, the ATE is estimated as the average difference between the observed outcome and the estimated missing potential outcome,

[22] The problem is that the imputed missing potential outcome $Y_{\ell(i)}$ is unbiased for $m_0(X_{\ell(i)})$ and $m_1(X_{\ell(i)})$, but not for $m_0(X_i)$ and $m_1(X_i)$.

$$\hat{\tau}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\Big[D_i(Y_i - [Y_{\ell(i)} + \widehat{m_0}(X_i) - \widehat{m_0}(X_{\ell(i)})]) + (1-D_i)([Y_{\ell(i)} + \widehat{m_1}(X_i) - \widehat{m_1}(X_{\ell(i)})] - Y_i)\Big]. \tag{1.18}$$

Similarly, the ATT is estimated as the average difference between the observed outcome and the estimated missing potential outcome among the treated,

$$\hat{\tau}_{ATT} = \frac{1}{N_T}\sum_{i=1}^{N} D_i(Y_i - [Y_{\ell(i)} + \widehat{m_0}(X_i) - \widehat{m_0}(X_{\ell(i)}]) . \tag{1.19}$$

1.6.4 Trimming

The overlap assumption - also called common support assumption - is a central assumption for the treatment effect estimators discussed in this chapter (see Section 1.3). In practice, there is often *limited overlap* or even *no overlap* in some region of the covariate space (Lechner & Strittmatter, 2019). That is, for some x in the support of X_i, we have only a few or even no observations in either the treated or the control group. As Crump, Hotz, Imbens, and Mitnik (2009) note, this can increase the bias and variance of the treatment effect estimators. Given the result of Rosenbaum and Rubin (1983), one way to analyze overlap is to compare the histograms of propensity scores for the treated and untreated. The overlap assumption is violated if there are values of the propensity score for which only treated or only control individuals exist. In the population, the overlap assumption is fulfilled when $0 < p(x) < 1$ (Lechner & Strittmatter, 2019). As Lechner and Strittmatter (2019) point out, this does not guarantee, however, that the overlap assumption is satisfied in the sample.

A related problem occurs when the propensity score of treated and control observations is close to 0 or 1. These observations receive a relatively large weight in the estimation, which can also increase the variance of the estimators.

Lechner and Strittmatter (2019) discuss several approaches to dealing with overlap problems. In this chapter, I apply two trimming rules. First, I discard observations with an estimated propensity score larger (smaller) than the maximum (minimum) estimated propensity score among the control (treated) group. Second, I discard observations with an estimated propensity score larger (smaller) than 0.99 (0.01).[23] Generally, this procedure changes the (sub-)population for which the treatment effect is estimated.

1.6.5 Performance Measures

To measure the performance of the treatment effect estimators, I use primarily the root-mean-square error (RMSE). In addition, I consider the absolute bias (|Bias|) and the standard deviation (SD) of the estimators. These performance measures are related to each other

[23] For the ATT estimation, I only discard observations with an estimated propensity score larger than these thresholds, but not those lower than the described thresholds.

in the following way

$$\text{RMSE} = \sqrt{\text{Bias}^2 + \text{SD}^2}. \tag{1.20}$$

Hence, the MSE (squared RMSE) can be decomposed into the squared bias and the variance of the estimator.

The RMSE of an estimator e is defined as

$$\text{RMSE}_e = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\tau}_{e,r} - \tau)^2}\,, \tag{1.21}$$

where R is the number of simulation replications, $\hat{\tau}_{e,r}$ is the estimated treatment effect of estimator e in simulation replication r, and τ is the true treatment effect (ATE or ATT). The term *accuracy* of an estimator refers to the RMSE of an estimator.

The absolute bias of an estimator e is defined as

$$|\text{Bias}|_e = \left|\frac{1}{R}\sum_{r=1}^{R}(\hat{\tau}_{e,r} - \tau)\right|. \tag{1.22}$$

The standard deviation of an estimator e is defined as

$$\text{SD}_e = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\tau}_{e,r} - \overline{\hat{\tau}_e})^2}\,, \tag{1.23}$$

where $\overline{\hat{\tau}_e} = \frac{1}{R}\sum_{r=1}^{R}\hat{\tau}_{e,r}$ is the average of the estimated treatment effects.

1.6.6 Software

Throughout the analysis, I use the software *R* (R Core Team, 2019). To fit the random forests, I use the *randomForest* package (Liaw & Wiener, 2002). For elastic net, I use the *glmnet* package (Friedman, Hastie, & Tibshirani, 2010). The treatment effect estimators based on matching employ the *Matching* package (Sekhon, 2011). The other treatment effect estimators are self-implemented.

1.7 Stylized Simulation Study

The DGP of the first simulation study is based on the simulation design of Diamond and Sekhon (2013), which was also used in Setoguchi et al. (2008), Lee et al. (2010), Pirracchio et al. (2015) and Cannas and Arpino (2019). From an applied perspective, the design of this simulation study is rather unrealistic. The covariates are either standard normal or Bernoulli distributed and independent of each other. The relationship between treatment and covariates is specified such that the resulting distribution of propensity scores exhibits a good overlap of control and treated observations.

1.7.1 Data Generating Process

In the default specification, a simulated dataset consists of 1000 observations with ten covariates, six of them are dummy variables and four continuous variables. The covariates are distributed as follows:

$$\begin{gathered}X_1 \sim Ber(0.5), X_3 \sim Ber(0.5),\ X_5 \sim Ber(0.5),\\ X_6 \sim Ber(0.5), X_8 \sim Ber(0.5),\ X_9 \sim Ber(0.5),\\ X_2 \sim N(0,1),\ X_4 \sim N(0,1),\ X_7 \sim N(0,1),\ X_{10} \sim N(0,1).\end{gathered}$$

For the within-estimator comparison as well as for the between-estimator comparison, I consider a default scenario in which the relationships between treatment and covariates and between outcome and covariates are moderately nonlinear and nonadditive.[24] The nonlinearity and nonadditivity is introduced by adding quadratic terms and interaction terms. In Section 1.7.3, I analyze changes in the data generating process. First, I analyze the effect of changes in the degree of linearity and additivity on the performance of the estimators. Second, I analyze the effect of changes in the strength of selection into treatment.

The relationship between treatment and covariates includes seven main effects, three quadratic terms, and ten interaction terms.[25] The specification is given by:

$$\begin{aligned}D^* =&\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 +\\ &\beta_2 X_2^2 + \beta_4 X_4^2 + \beta_7 X_7^2 + 0.5\beta_1 X_1 X_3 + 0.7\beta_2 X_2 X_4 + 0.5\beta_3 X_3 X_5 +\\ &0.7\beta_4 X_4 X_6 + 0.5\beta_5 X_5 X_7 + 0.5\beta_1 X_1 X_6 + 0.7\beta_2 X_2 X_3 + 0.5\beta_3 X_3 X_4 +\\ &0.5\beta_4 X_4 X_5 + 0.5\beta_5 X_5 X_6\ ,\end{aligned} \tag{1.24}$$

with $\beta_0 = 0, \beta_1 = 0.8, \beta_2 = -0.25, \beta_3 = 0.6, \beta_4 = -0.4, \beta_5 = -0.8, \beta_6 = -0.5, \beta_7 = 0.7.$[26]

The observed treatment indicator D is drawn from a Bernoulli distribution with probability equal to the propensity score:

$$D \sim Ber\left(\frac{1}{1 + exp(-D^*)}\right), \tag{1.25}$$

where the propensity score is given by the logistic transformation on D^*.

Similarly, the observed outcome is a function of the treatment indicator, seven main effects, three quadratic terms, ten interaction terms, and a noise term. Here the simulation design differs from Diamond and Sekhon (2013) in two ways. First, they include only the seven main effects and therefore consider a linear and additive relationship between outcome and covariates. Second, they do not include a noise term in the outcome specification. The

[24] The definition of nonlinearity and nonadditivity follows Diamond and Sekhon (2013), see footnote 15.

[25] The covariates are divided into three types. The covariates X_1, X_2, X_3, and X_4 have an effect on both treatment and outcome (confounders). The covariates X_5, X_6, and X_7 have an effect only on treatment, while covariates X_8, X_9, and X_{10} have an effect only on the outcome.

[26] This is scenario G in Diamond and Sekhon (2013).

specification is given by:

$$
\begin{aligned}
Y = \tau D &+ \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_8 X_8 + \alpha_9 X_9 + \alpha_{10} X_{10} + \\
&\alpha_2 X_2^2 + \alpha_4 X_4^2 + \alpha_{10} X_{10}^2 + 0.5\alpha_1 X_1 X_3 + 0.7\alpha_2 X_2 X_4 + 0.5\alpha_3 X_3 X_4 + \\
&0.7\alpha_4 X_4 X_8 + 0.5\alpha_8 X_8 X_{10} + 0.5\alpha_1 X_1 X_{10} + 0.7\alpha_2 X_2 X_3 + 0.5\alpha_3 X_3 X_9 + \\
&0.5\alpha_4 X_4 X_{10} + 0.5\alpha_9 X_9 X_{10} + \epsilon \, ,
\end{aligned}
\tag{1.26}
$$

with $\tau = -0.4$ being the constant treatment effect, $\alpha_0 = -3.85, \alpha_1 = 0.3, \alpha_2 = -0.36, \alpha_3 = -0.73, \alpha_4 = -0.2, \alpha_8 = 0.71, \alpha_9 = -0.19, \alpha_{10} = 0.26$, and ϵ being a noise term from a normal distribution with mean 0 and variance σ_ϵ^2. In an attempt to approximately balance the noise in the treatment and outcome specification, σ_ϵ^2 is chosen such that the pseudo R^2 of the treatment indicator is approximately equal to the R^2 of the outcome specification.[27] Since the treatment effect is constant, there is no difference between the ATE and the ATT. In the estimation, I use the treatment effect estimators for the ATE.

1.7.2 Misspecification Scenarios

In theory, the regression estimator is consistent if the conditional outcome means are correctly specified. Likewise, the propensity score methods are consistent if the propensity score is correctly specified. The hybrid estimators are consistent if either the conditional outcome means or the propensity score is correctly specified.

In this simulation I depart from the situation in which the conditional outcome means and the propensity score are always correctly specified. I consider four scenarios. The scenarios differ in terms of misspecification of the conditional outcome means and the propensity score. As described in Section 1.7.1, the DGP is specified such that the *true* conditional outcome means are linear-regression-type functions, including the main effects of the covariates, quadratic terms, and interaction terms. Similarly, the *true* propensity score is a logit-type function, including the main effects of the covariates, quadratic terms, and interaction terms. Hence, a correctly specified conditional outcome mean is estimated with OLS on the *correct* set of variables. The *correct* set of variables includes the main effects, quadratic terms, and interaction terms that were used to generate the data. A correctly specified propensity score is estimated with logit on the *correct* set of variables. Misspecification is introduced by omitting the quadratic and interaction terms in the estimation, or by using machine learning methods instead of OLS and logit.

In misspecification scenario I, both the propensity score and the conditional outcome means are correctly specified. This means that a) the propensity score $p(x)$ is estimated using logit on the *correct* set of variables, and b) the conditional outcome means $m_1(x)$ and $m_0(x)$ are estimated using OLS on the *correct* set of variables. As no machine learning based estimation is involved, this scenario is only of limited interest in this context, and the results are presented only in the Appendix.

[27] See Appendix 1.A.1

In misspecification scenario II, only the propensity score is correctly specified, and the conditional outcome means are misspecified. This means that the propensity score is estimated as in misspecification scenario I. The conditional outcome means are estimated using either a) OLS on the set of main variables, i.e. omitting the quadratic and interaction terms, or b) random forests and elastic net. For random forests, only the main variables are included. For elastic net, the set of main variables and all possible quadratic and two-way interaction terms are included.

In misspecification scenario III, only the conditional outcome means are correctly specified, and the propensity score is misspecified. This means that the conditional outcome means are estimated as in misspecification scenario I. The propensity score is estimated using either a) logit on the set of main variables, i.e. omitting the quadratic and interaction terms, or b) random forests and elastic net. For random forests, only the main variables are included. For elastic net, the set of main variables and all possible quadratic and two-way interaction terms are included.

In misspecification scenario IV, both the propensity score and the conditional outcome means functions are misspecified. The misspecified functions are estimated as in misspecification scenarios II and III.

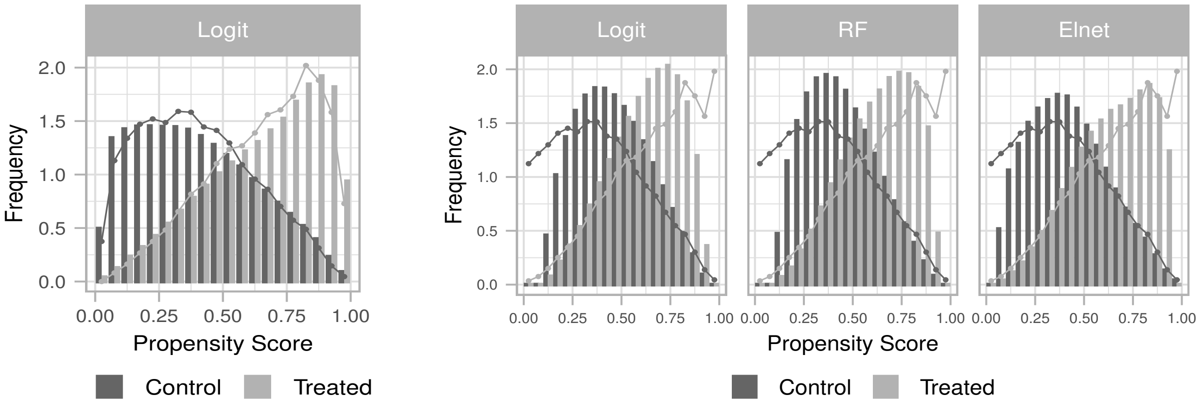
1.7.3 Results

Within-Estimator Comparison

Before I elaborate on the results of the different misspecification scenarios, I demonstrate that this simulation design actually generates propensity score distributions with fairly good overlap. Figure 1.3 presents the median histogram of the trimmed propensity scores. The left panel plots the estimated (bars) and true (lines) propensity scores for misspecification scenarios I/II, i.e. for a correctly specified propensity score. It can be seen that the distribution of estimated propensity scores closely follows the distribution of the true propensity scores. The right panel plots the estimated and true propensity scores for misspecification scenarios III/IV, i.e. for a misspecified propensity score. In this case there is a considerable difference between estimated propensity scores and true propensity scores. Logit and random forest, in particular, overestimate the fraction of propensity scores close to 0.5, and underestimate the fraction of propensity scores close to 0 and 1.[28]

[28] The difference in the histogram of true propensity scores between the left and right panels of Figure 1.3 is due to the trimming (see Section 1.6.4).

Figure 1.3: Distribution of Estimated Propensity Scores



(a) MS I/II: PS correctly specified

(b) MS III/IV: PS misspecified

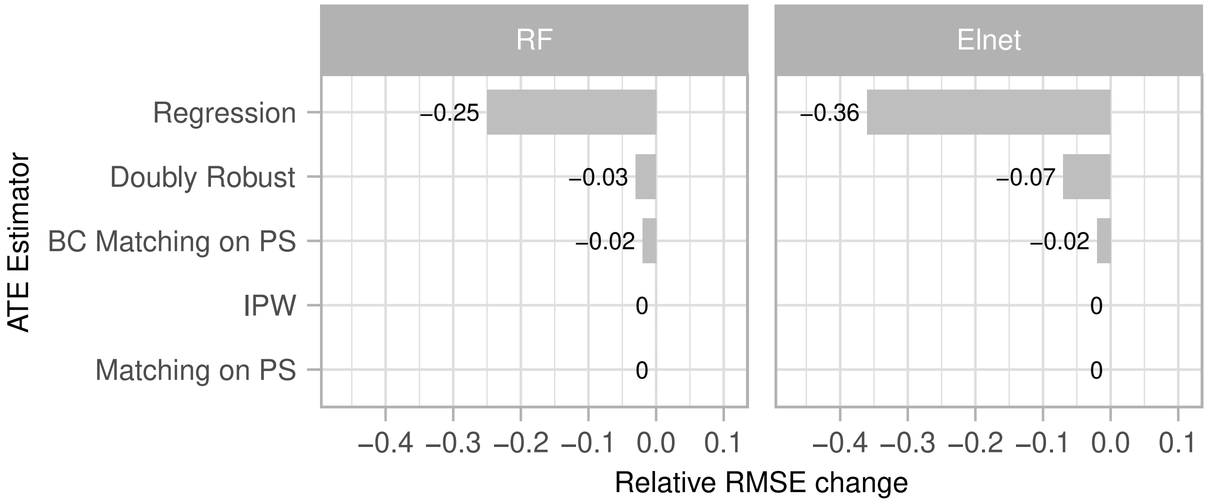
Note: The left and right panels represent misspecification scenarios I/II and III/IV. The graph indicates the median (over 5000 simulation replications) histogram of estimated propensity scores by logistic regression (Logit), random forests (RF), and elastic net (Elnet). The darker gray bars (line) represent the distribution of the estimated (true) propensity scores of control observations, the lighter gray bars (line) represent the distribution of the estimated (true) propensity scores of treated observations.

This section compares - for each treatment effect estimator - conventional based estimation (OLS and logit) to machine learning based estimation (random forest and elastic net). Since there is no difference between conventional and machine learning based estimation in misspecification scenario I (propensity score and conditional outcome means correctly specified), the results for this misspecification scenario are not presented.

Misspecification Scenario II: Propensity score correctly specified, conditional outcome means misspecified

The results for the misspecification scenario in which the propensity score is correctly specified, but the conditional outcome means are misspecified, are presented in Figure 1.4.

Figure 1.4: Within-Estimator Comparison for Misspecification Scenario II



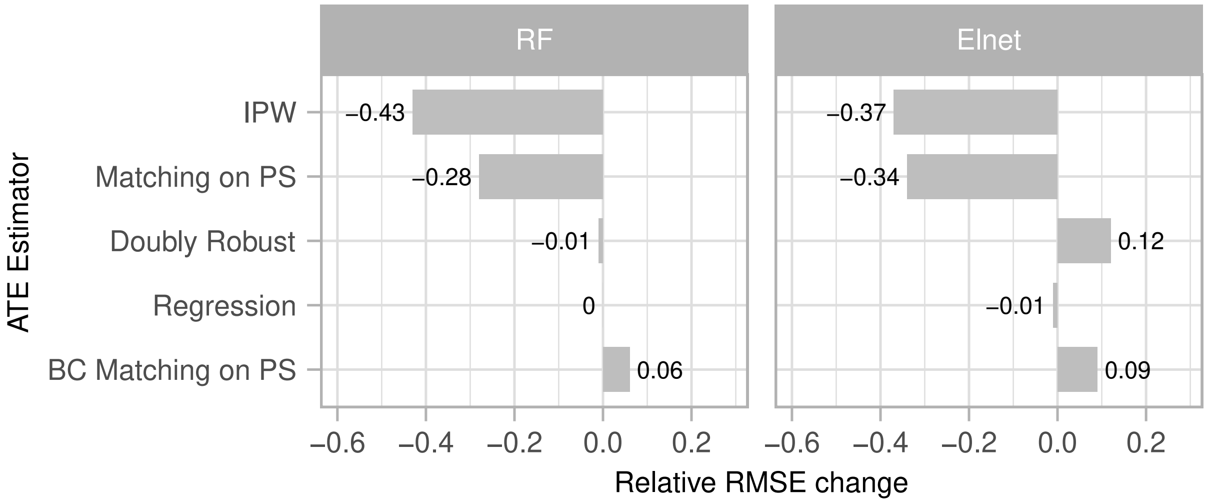
Note: Simulation based on 5000 replications. The bars display the relative change in RMSE of machine learning based estimation of the ATE (random forest in the left panel, elastic net in the right panel), compared to conventional based estimation of the ATE (OLS/Logit). A negative value indicates that the RMSE of the machine learning based estimator was lower, i.e. that the treatment effect was estimated more accurately with machine learning methods.

Since *IPW* and *Matching on PS* use only the (correctly specified) propensity score, there is no difference between conventional and machine learning based estimators. For *BC Matching on PS*, *Regression*, and *Doubly Robust*, there are improvements in the RMSE when the conditional outcome means are estimated with ML methods. The relative improvements are moderate (2%-7%) for the hybrid methods *BC Matching on PS* and *Doubly Robust*, and substantial (25% and 36%) for the *Regression* estimator. To sum up, I find no improvements, or only minor ones, for methods employing the correctly specified propensity score, but sizable improvements for methods employing only the misspecified conditional outcome means.

Misspecification Scenario III: Conditional outcome means correctly specified, propensity score misspecified

The results for the misspecification scenario in which the conditional outcome means are correctly specified, but the propensity score is misspecified, are presented in Figure 1.5.

Figure 1.5: Within-Estimator Comparison for Misspecification Scenario III



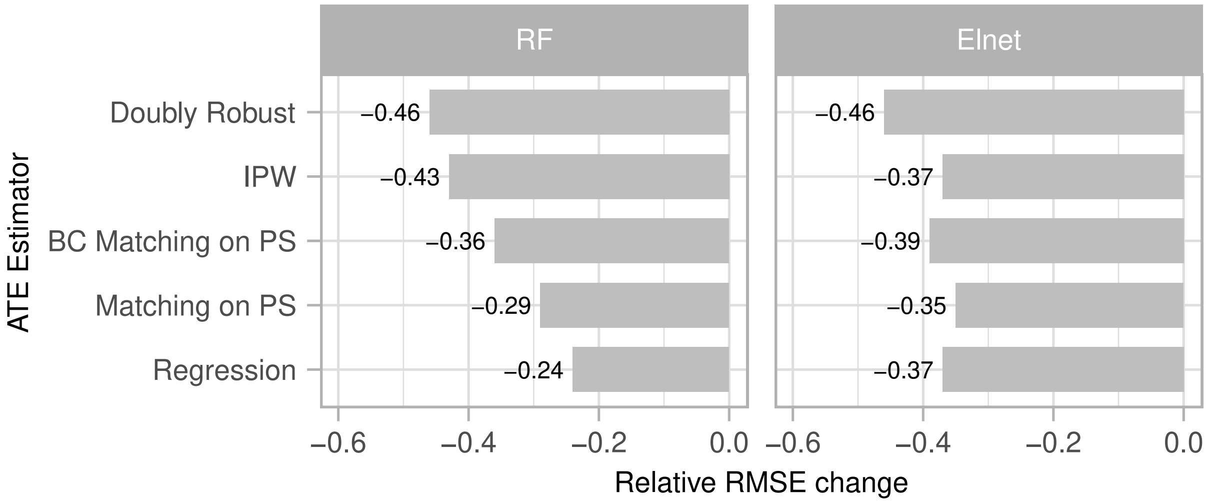
Note: Simulation based on 5000 replications. The bars display the relative change in RMSE of machine learning based estimation of the ATE (random forest in the left panel, elastic net in the right panel), compared to conventional based estimation of the ATE (OLS/Logit). A negative value indicates that the RMSE of the machine learning based estimator was lower, i.e. that the treatment effect was estimated more accurately with machine learning methods.

In this misspecification scenario, *Regression* uses only the (correctly specified) conditional outcome means. As a result, there is no difference between conventional and machine learning based estimators. For *IPW* and *Matching on PS*, there are large improvements in the RMSE when the propensity score is estimated with ML methods. The relative improvements range between 28% and 43%. The hybrid estimators *Doubly Robust* and *BC Matching on PS* estimate the ATE up to 12% less accurately when the propensity score is estimated with machine learning methods. As demonstrated in the between-estimator comparison in the next section, this is due to an increase in the variance of these estimators. To sum up, I find sizable improvements for methods employing only the misspecified propensity score. Estimators employing the correctly specified conditional outcome means are either unaffected by design (*Regression*), or perform somewhat worse (*Doubly Robust* and *BC Matching on PS*).

Misspecification Scenario IV: Both propensity score and conditional outcome means misspecified

The results for the misspecification scenario in which both the propensity score and the conditional outcome means are misspecified are presented in Figure 1.6.

Figure 1.6: Within-Estimator Comparison for Misspecification Scenario IV



Note: Simulation based on 5000 replications. The bars display the relative change in RMSE of machine learning based estimation of the ATE (random forest in the left panel, elastic net in the right panel), compared to conventional based estimation of the ATE (OLS/Logit). A negative value indicates that the RMSE of the machine learning based estimator was lower, i.e. that the treatment effect was estimated more accurately with machine learning methods.

The RMSE improvements of *IPW* and *Matching on PS* are the same as they are in misspecification scenario III. This is because *IPW* and *Matching on PS* rely only on the propensity score, which is misspecified in the same way in both misspecification scenarios. Similarly, since *Regression* relies only on the conditional outcome means, the RMSE improvement of *Regression* is the same as in misspecification scenario II. For the hybrid estimators *BC Matching on PS* and *Doubly Robust*, I find substantial improvements in RMSE when the propensity score and the conditional outcome means are estimated with machine learning methods compared to conventional methods. The relative improvements range from 36% to 46%. Overall, I find that all treatment effect estimators have a lower RMSE when the propensity score and/or the conditional outcome means are estimated with machine learning methods.

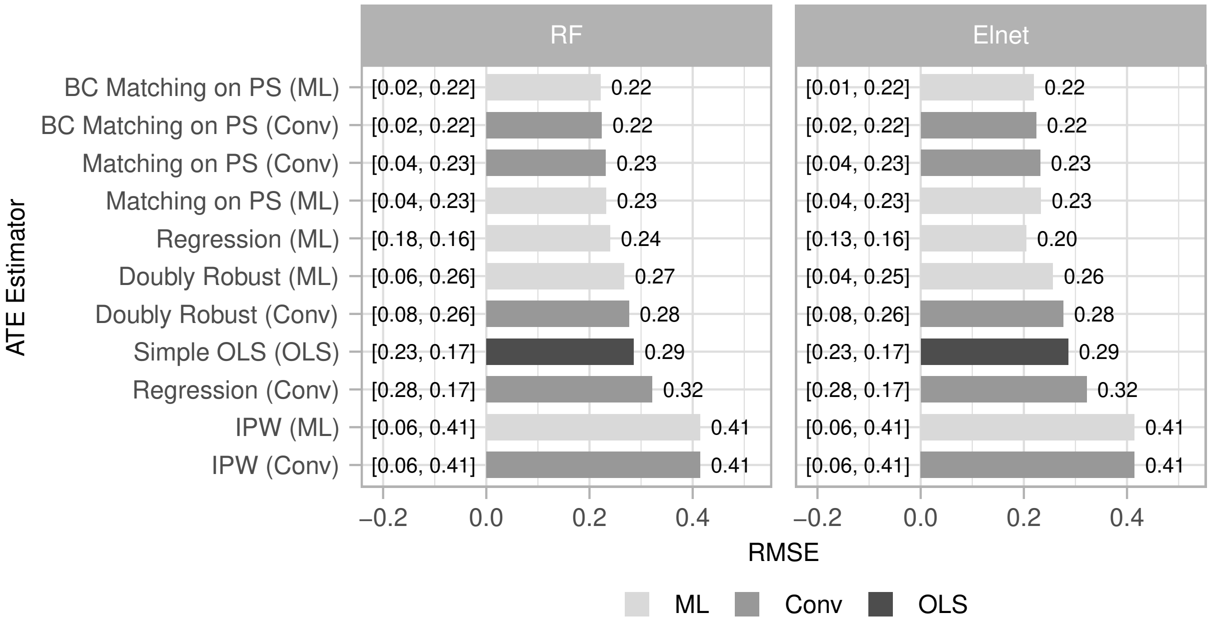
Between-Estimator Comparison

In the between-estimator comparison, I compare the performance of all estimators (conventional and machine learning based). Moreover, I include the *Simple OLS* estimator regressing the outcome on the treatment indicator and the ten covariates. As discussed in Section 1.6, the *Simple OLS* estimator is generally not a valid estimator for the ATE, but often applied in practice. For this reason it is valuable to compare the ATE estimators to the *Simple OLS* estimator. Note that in all misspecification scenarios the *Simple OLS* estimator includes only the treatment indicator and the ten main effects, but no quadratic and interaction terms. For reasons of space, the results for misspecification scenario I are presented in Appendix 1.A.

Misspecification Scenario II: Propensity score correctly specified, conditional outcome misspecified

The results for the misspecification scenario in which the propensity score is correctly specified, but the conditional outcome means are misspecified are presented in Figure 1.7.

Figure 1.7: Between-Estimator Comparison for Misspecification Scenario II



Note: The bars indicate the RMSE of the ATE estimator over 5000 simulation replications. The numbers in brackets to the left of the bars indicate the absolute bias and the standard deviation of the estimator, i.e. [|Bias|, SD]. The light gray bars represent estimators that use machine learning methods (random forests and elastic net) to estimate the conditional outcome means. The dark gray bars represent estimators that use conventional methods (OLS) to estimate the conditional outcome means. The black bar represents the simple OLS estimator. The results for random forest (RF) are presented in the left panel, and those for elastic net in the right panel.

With respect to the RMSE, various estimators perform almost equally well. *BC Matching on PS*, *Matching on PS*, *Doubly Robust*, *Regression*, and *Simple OLS* are in the same range of RMSE. Only the *IPW* estimators perform somewhat worse.

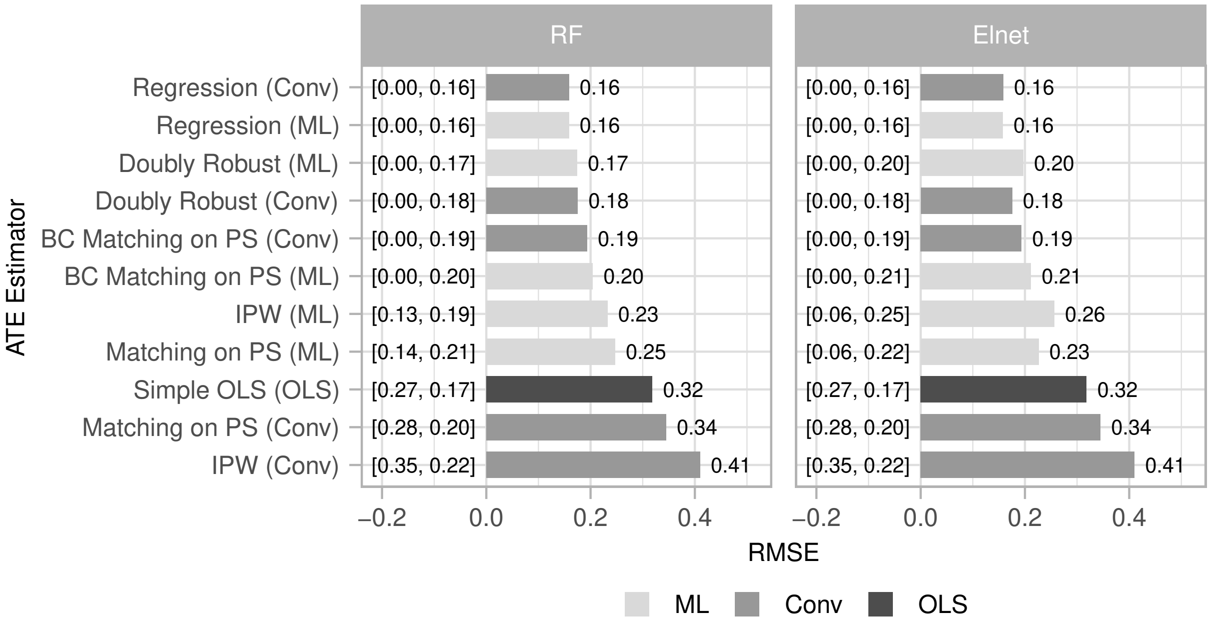
In terms of absolute bias, the estimators differ considerably. Estimators employing the correctly specified propensity score have a low bias (0.01 to 0.08). This illustrates that these methods are able to remove the bias. By contrast, the estimators employing only the misspecified conditional outcome means (*Regression* and *Simple OLS*) have a higher bias (0.13 to 0.28).

An examination of the standard deviation of the estimators reveals that, *Regression* (0.16 and 0.17) and *Simple OLS* (0.17) have the lowest standard deviation. The estimators based on matching as well as the *Doubly Robust* estimators have a higher standard deviation, ranging from 0.22 to 0.26. The *IPW* estimators exhibit by far the largest variance (SD: 0.41). Overall, this misspecification scenario does not provide evidence that hybrid methods are superior to estimators that rely only on the propensity score or only on the conditional outcome means.

Misspecification Scenario III: Conditional outcome correctly specified, propensity score misspecified

The results for the misspecification scenario in which the conditional outcome means are correctly specified, but the propensity score is misspecified are presented in Figure 1.8.

Figure 1.8: Between-Estimator Comparison for Misspecification Scenario III



Note: The bars indicate the RMSE of the ATE estimator over 5000 simulation replications. The numbers in brackets to the left of the bars indicate the absolute bias and the standard deviation of the estimator, i.e. [|Bias|, SD]. The light gray bars represent estimators that use machine learning methods (random forests and elastic net) to estimate the propensity score. The dark gray bars represent estimators that use conventional methods (logit) to estimate the propensity score. The black bar represents the simple OLS estimator. The results for random forest (RF) are presented in the left panel, and those for elastic net in the right panel.

In terms of RMSE, I find that the estimators employing the correctly specified conditional outcome means perform almost equally well. The RMSE ranges from 0.16 for *Regression* to 0.21 for the elastic net based *BC Matching on PS* estimator. The performance of machine learning based *IPW* and *Matching on PS* is slightly worse. The conventional based *IPW* estimator again has the highest RMSE.

With regard to the absolute bias, the simulation demonstrates that the estimators using the correctly specified conditional outcome means are able to remove the bias completely. Estimators employing only the misspecified propensity score are biased. The bias ranges from 0.06 for elastic net based *IPW* and *Matching on PS*, to 0.35 for conventional based *IPW*.

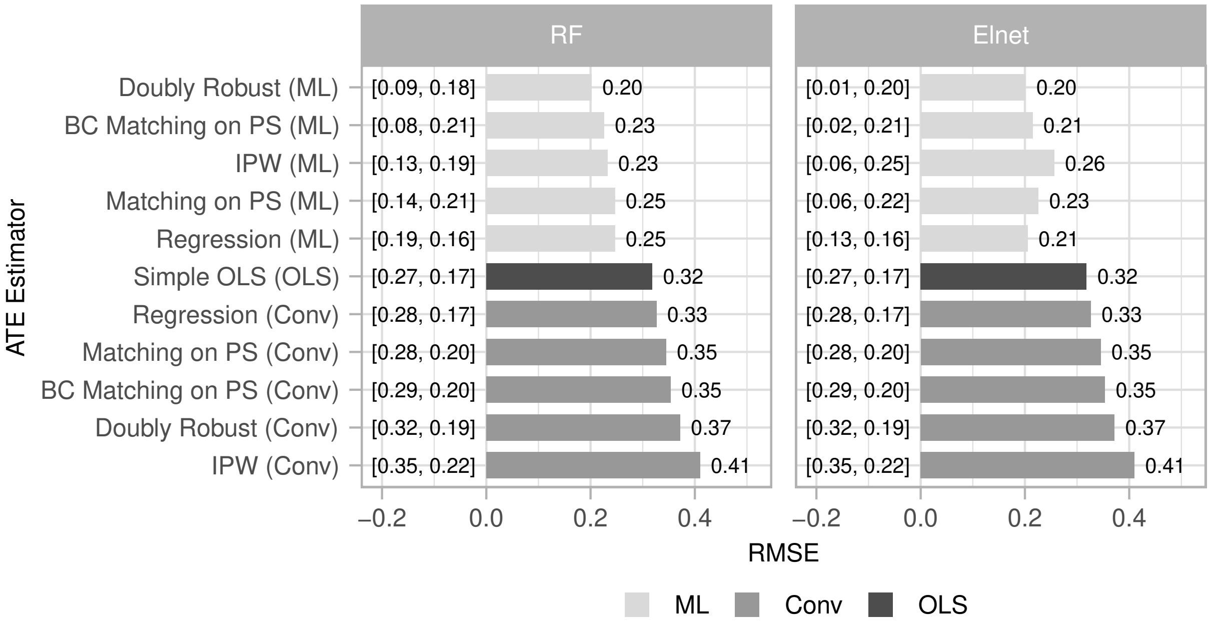
The standard deviations are all rather similar, ranging from 0.16 for *Regression* to 0.25 for the elastic net based *IPW* estimator.

Overall, the simulation results for this misspecification scenario again do not provide evidence that hybrid methods are superior. However, in conjunction with the results from misspecification scenario II, it can be observed that in both scenarios the hybrid methods are among the estimators with the lowest RMSE, or have only a slightly higher RMSE than the estimator with the lowest RMSE. This slight increase is due to an increase in the variance of the estimator.

Misspecification Scenario IV: Both propensity score and conditional outcome model misspecified

The results for the misspecification scenario in which both the propensity score and the conditional outcome means are misspecified are presented in Figure 1.9.

Figure 1.9: Between-Estimator Comparison for Misspecification Scenario IV



Note: The bars indicate the RMSE of the ATE estimator over 5000 simulation replications. The numbers in brackets to the left of the bars indicate the absolute bias and the standard deviation of the estimator, i.e. [|Bias|, SD]. The light gray bars represent estimators that use machine learning methods (random forests and elastic net) to estimate the conditional outcome means and the propensity score. The dark gray bars represent estimators that use conventional methods (OLS and logit) to estimate the conditional outcome means and the propensity score. The black bar represents the simple OLS estimator. The results for random forest (RF) are presented in the left panel, and those for elastic net in the right panel.

With respect to RMSE, the machine learning based *Doubly Robust* estimators perform best (both 0.20). The performance of the other machine learning based estimators is very similar, ranging from 0.21 to 0.26. The conventional based estimators perform worse, especially the hybrid estimators and *IPW*. Interestingly, the conventional based estimators perform even worse than the *Simple OLS* estimator.

In terms of absolute bias, I find that the machine learning based hybrid estimators and the elastic net based propensity score methods have the smallest bias, ranging from 0.01 to 0.09. This provides evidence that these methods are able to approximate the true underlying confounding relationships, even though they are misspecified. The random forest based propensity score methods, machine learning based *Regression*, as well as the conventional based estimators exhibit a higher bias.

As for misspecification scenario III, the standard deviations of the estimators are all very similar, with the lowest standard deviation observed for the machine learning based *Regression* estimators (0.16).

In general, I again find no evidence that hybrid estimators outperform the estimators that rely only on the propensity score or only on the conditional outcome means. Although the machine learning based hybrid methods have lower bias in this scenario, this is partly offset by an increase in the variance (compared to e.g. *Regression*).

The between-estimator comparison of misspecification scenario II, III, and IV demonstrates that hybrid estimators are often among the estimators with the lowest RMSE. If not, the difference to the estimator with the lowest RMSE is small. This provides evidence that using the hybrid methods might be advantageous to guard against misspecification, sometimes at the cost of an increase in the variance.

Analysis of Changes in the Data Generating Process

In this section I analyze the effect of changes in the DGP on the performance of the treatment effect estimators. I consider changes in the degree of linearity and additivity, as well as changes in the strength of selection into treatment.

Linearity and Additivity

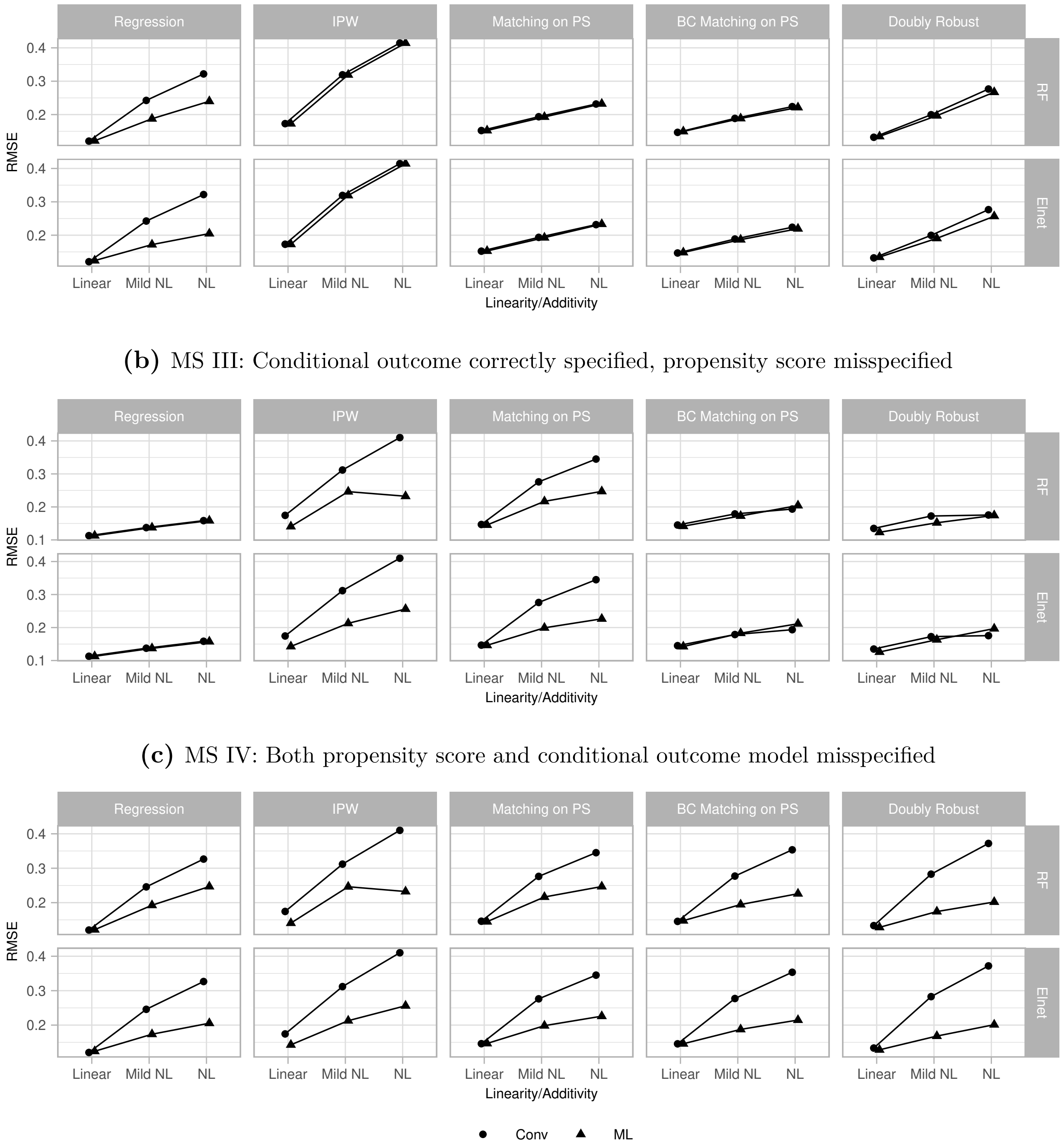
First, I analyze the effects of changes in the degree of linearity and additivity in the relationships between treatment and covariates and between outcome and covariates. As described in Section 1.7.1, the default specification for both treatment and outcome includes seven main effects, three quadratic terms, and ten interaction terms. In the following, the default specification is abbreviated as *NL*. Similar to Diamond and Sekhon (2013), I compare specification *NL* to a specification where treatment and outcome are linear and additive functions of the main effects only, without quadratic or interaction terms.[29] This specification is abbreviated as *Linear*. Moreover, I include an intermediate specification where treatment and outcome are functions of the main effects, one quadratic term, and four interaction terms, abbreviated as *Mild NL*.[30] The results are presented in Figure 1.10.

The analysis provides several insights. First - and not surprisingly - the overall performance of the treatment effect estimators does not improve when the underlying relationships change from linear and additive to nonlinear and nonadditive. Second, the RMSE of the treatment effect estimators are all rather similar in the *Linear* specification. Moreover, the differences in RMSE between machine learning based and conventional based estimators are small. Third, when the underlying relationships change from linear and additive to nonlinear and nonadditive, the performance of the estimators depends heavily on misspecification. In

[29] Strictly speaking, the relationship between treatment and covariates is never linear, since the logistic transformation is applied and D is a binary variable. In this context, the linearity and additivity applies to the underlying function for D^*.

[30] See Appendix 1.A.1 for details on the specifications.

Figure 1.10: Analysis of Changes in Linearity and Additivity



(a) MS II: Propensity score correctly specified, conditional outcome misspecified

(b) MS III: Conditional outcome correctly specified, propensity score misspecified

(c) MS IV: Both propensity score and conditional outcome model misspecified

Note: The top/middle/bottom panels represent misspecification scenarios II/III/IV. In each panel, the columns represent the treatment effect estimators, the rows indicate whether random forest (RF) or elastic net (Elnet) was used to estimate the propensity score and/or the conditional outcome means. In each subgraph, the x-axis plots three degree of linearity and additivity. *Linear* corresponds to a DGP where both the outcome specification and the treatment specification are linear and additive, i.e. include only the main effects. To improve readability, the points are slightly offset horizontally. *Mild NL* and *NL* correspond to DGPs where both the outcome specification and the treatment specification are nonlinear and nonadditive. *Mild NL* includes the main effects, one quadratic term and four interaction terms. *NL* includes the main effects, three quadratic terms, and ten interaction terms. The shape of the points represents the conventional based estimators (circle) and the machine learning based estimators (triangle).

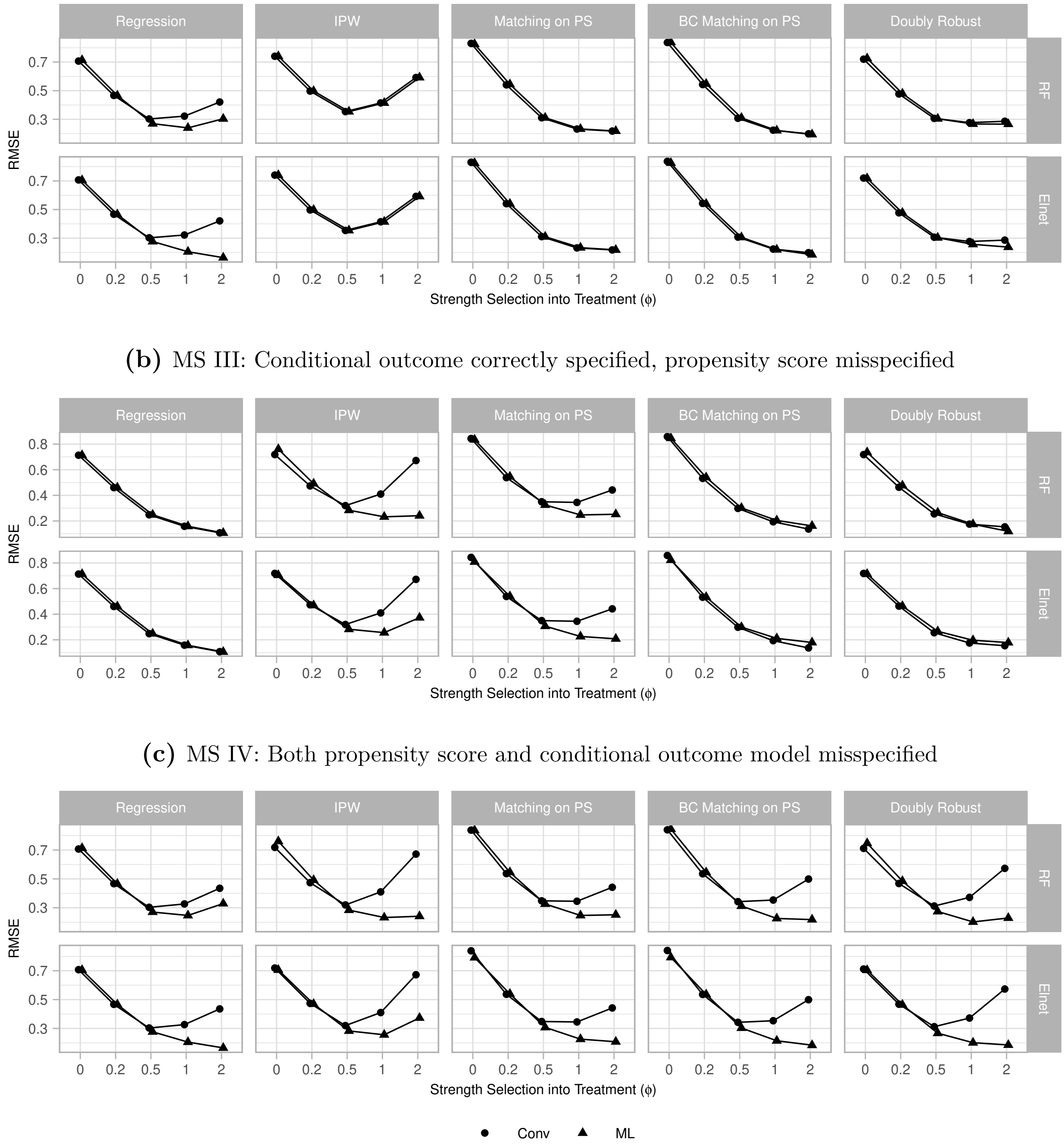
the top panel (MS II), there is a stark difference between estimators employing the correctly specified propensity score (*IPW*, *Matching on PS*, *BC Matching on PS*, *Doubly Robust*), and the estimator employing only the misspecified conditional outcome means (*Regression*). The performance of machine learning based *Regression* deteriorates less than the performance of conventional based *Regression*. For the other estimators there is almost no difference between machine learning based and conventional based estimation. A similar pattern is observed for MS III and MS IV. In the middle panel (MS III), the difference emerges between estimators that employ the correctly specified conditional outcome means (*Regression*, *BC Matching on PS*, *Doubly Robust*), and those that employ only the misspecified propensity score (*IPW*, *Matching on PS*). In the bottom panel (MS IV), all estimators rely on misspecified propensity scores and/or conditional outcome means. Again, I find that the performance of machine learning based estimators deteriorates less - as the underlying relationships become nonlinear and nonadditive - than the performance of conventional based estimators. In summary, this analysis identifies large differences between machine learning based estimators and conventional based estimators when these estimators employ only misspecified propensity scores and/or conditional outcome means in a nonlinear and nonadditive setting. This indicates that machine learning based estimators are better able to approximate the nonlinear and nonadditive underlying relationships.

Strength Selection Into Treatment

The strength of selection into treatment captures the relative weight of the covariates in determining whether an observation is treated or not. As it can be seen in equation (1.24), D^* is only a function of the covariates. However, the observed treatment indicator is a Bernoulli draw, and hence a function of both D^* and a random component. To analyze changes in the strength of selection into treatment, I follow Frölich (2004) and Huber et al. (2013), and change the relative weight of the covariates by multiplying D^* by a parameter ϕ. The adjusted propensity score is then given by $\frac{1}{1+exp(-\phi D^*)}$, where $\phi \in \{0, 0.2, 0.5, 1, 2\}$. In the default specification, ϕ is set to 1. Decreasing this parameter towards zero decreases the variance of D^*. After the logistic transformation, the propensity scores are less spread out, and thus the relative weight of the covariates is smaller. Hence, the effect of the covariates on the propensity score and thus on the treatment indicator, is less pronounced. The specification with $\phi = 0$ represents the extreme case of random treatment assignment, with propensity scores all equal to 0.5. The results are presented in Figure 1.11.

As the strength of selection into treatment increases, the performance of the treatment effect estimators either decreases monotonically or follows a U-shaped pattern. This can be explained by different bias and variance patterns of the estimators. Figures 1.19 and 1.20 in Appendix 1.A.2 plot the absolute bias and the standard deviation of the estimators. The bias remains constant for the estimators employing the correctly specified propensity score and/or conditional outcome means. However, the bias increases for estimators employing only the misspecified propensity score and/or conditional outcome means. The strongest bias increase is observed for conventional based estimators. By contrast, the standard deviation of

Figure 1.11: Analysis of Changes in the Strength of Selection Into Treatment



(a) MS II: Propensity score correctly specified, conditional outcome misspecified

(b) MS III: Conditional outcome correctly specified, propensity score misspecified

(c) MS IV: Both propensity score and conditional outcome model misspecified

Note: The top/middle/bottom panels represent misspecification scenarios II/III/IV. In each panel, the columns represent the treatment effect estimators, the rows indicate whether random forest (RF) or elastic net (Elnet) was used to estimate the propensity score and/or the conditional outcome means. In each subgraph, the x-axis plots ϕ, the strength of selection into treatment. The larger ϕ, the stronger the selection into treatment. With $\phi = 0$, selection into treatment is random. The default specification corresponds to $\phi = 1$. The shape of the points represents the conventional based estimators (circle) and the machine learning based estimators (triangle).

the estimators decreases as selection into treatment increases, except for the *IPW* estimator. Hence, this analysis illustrates that machine learning based estimators perform as well as conventional based estimators when selection into treatment is weak (or random), or when the estimators employ the correctly specified propensity score and/or conditional outcome means. However, when selection into treatment is more pronounced, and the estimators use a misspecified propensity score and/or conditional outcome means, the machine learning based estimators outperform the conventional based estimators.

1.8 Empirical Simulation Study: LaLonde Data

The DGP of the second simulation study is based on the LaLonde (1986) dataset and closely follows the simulation design of Busso et al. (2014). The simulation study in Section 1.7 is characterized by certain unrealistic features, such as that the covariates are drawn from either a standard normal distribution or a Bernoulli distribution with a probability of 0.5. Moreover, the covariates are all independent of each other, and the relationship between treatment and covariates is specified such that approximately half of the sample is treated and half is control. Such characteristics are often not present in empirical settings. In order to apply the estimators in a more realistic setting, I conduct an empirical simulation study.

The idea of the empirical simulation study in this section is to base the DGP on characteristics of a real dataset. The procedure is to fit parametric distributions to a real dataset and subsequently generate a population from the fitted distributions. In the simulation, I repeatedly draw samples from the population. The real dataset I consider is the Dehejia and Wahba (1999) subsample of the NSW/PSID1 dataset considered in LaLonde (1986). The dataset is described in more detail in Section 1.9. Following Busso et al. (2014), I further restrict the sample to African Americans. In the following I refer to this dataset as the LaLonde dataset. The objective is to estimate the ATT of a job training program on earnings.

The LaLonde dataset consists of 780 individuals, 156 of whom were treated and 624 not treated, with eight covariates. This already marks an important difference in relation to the simulation in Section 1.7. The number of treated observations is considerably smaller than the number of control observations in the LaLonde dataset. As is demonstrated later, this has an effect on the distributions of the propensity scores. Overlap is limited in this dataset, which makes treatment effect estimation generally more challenging.

1.8.1 Data Generating Process

The population consists of 1 million observations with eight covariates, four of which are dummy variables and four continuous variables. In the simulation I repeatedly draw samples of 1000 observations. The population is generated in order to calculate the population ATT, i.e. the true causal effect of interest.[31] Following Busso et al. (2014), the covariates are generated as follows:

1. Draw four dummies for *married*, *no degree*, unemployed in the year 1974 (*u74*), and unemployed in the year 1975 (*u75*) from four Bernoulli distributions with probabilities equal to the sample means of *married*, *no degree*, *u74*, and *u75* in the LaLonde data. Each possible combination of these four dummies represents a *group*.

2. For each of the 16 groups defined by the four dummies, draw the variables *age*, *education*, earnings in 1974 (*re74*, in 1000), and earnings in 1975 (*re75*, in 1000) from a group-specific multivariate normal distribution.[32] The means and covariance matrix of the group-specific normal distribution are equal to the sample means and covariances of *age*, *education*, *re74*, and *re75* in the group-specific subset of the LaLonde data. The variables *re74* and *re75* are restricted to be in the interval defined by the group-specific minimum and maximum of *re74* and *re75* in the LaLonde data. Draws outside the minimum or maximum are set to the minimum or maximum. The variables *age* and *education* are rounded to integers.

3. To model the relationship between treatment and the covariates, a logit model is fitted to the LaLonde data. Following Busso et al. (2014), I include the eight main variables as well as squared *re74*, squared *re75*, and interactions between *u74* and *u74*, and between *re74* and *re75* in the logit model. I refer to the coefficients of the fitted logit model as $\widehat{\beta}^L$. Then, the population relationship between treatment and the covariates is based on
$$D^* = Z'\widehat{\beta}^L - u \;, \tag{1.27}$$
where Z is the set of variables consisting of the main effects as well as the quadratic and interaction terms, and u is a noise term from a logistic distribution with location 0 and scale 1.

4. The observed treatment indicator D is constructed as:
$$D = \mathbb{1}\{D^* > 0\} \;,$$
where $\mathbb{1}\{\cdot\}$ is the indicator function.

5. To model the relationship between the outcome and the covariates, two separate OLS regressions are fitted to the LaLonde data. The set of variables included in the OLS regressions are the same as for the logit model in step 3. The first OLS regression is fitted in the subsample of treated observations, and the coefficients are referred to as

[31] Unlike in the simulation of Section 1.7, the treatment effect of this simulation is allowed to be heterogeneous.

[32] The group-specific multivariate distribution ensures that the variables are dependent, e.g. unemployment status and education are correlated.

$\hat{\gamma}_1^L$. The second OLS regression is fitted in the subsample of untreated observations, and the coefficients are referred to as $\hat{\gamma}_0^L$.

6. Calculate the potential outcomes Y^1 and Y^0 as

$$Y^1 = Z'\hat{\gamma}_1^L + \epsilon^1 \text{ , and} \tag{1.28}$$

$$Y^0 = Z'\hat{\gamma}_0^L + \epsilon^0 \text{ ,} \tag{1.29}$$

   where Z is again the set of variables consisting of the main effects as well as the quadratic and interaction terms, and ϵ^1 and ϵ^0 are noise terms from two normal distributions with mean 0 and variances $\sigma^2_{\epsilon_1}$ and $\sigma^2_{\epsilon_0}$. The variances $\sigma^2_{\epsilon_1}$ and $\sigma^2_{\epsilon_0}$ correspond to the means of the squared residuals from the two OLS regressions fitted to the LaLonde data.

7. Finally, the observed outcome Y is given by

$$Y = DY^1 + (1 - D)Y^0 \text{ .} \tag{1.30}$$

Compared to the simulation in Section 1.7, this DGP does not impose a constant treatment effect. Since two OLS regressions are fitted in the treated and untreated subsample, this DGP allows for heterogeneity in treatment effects.

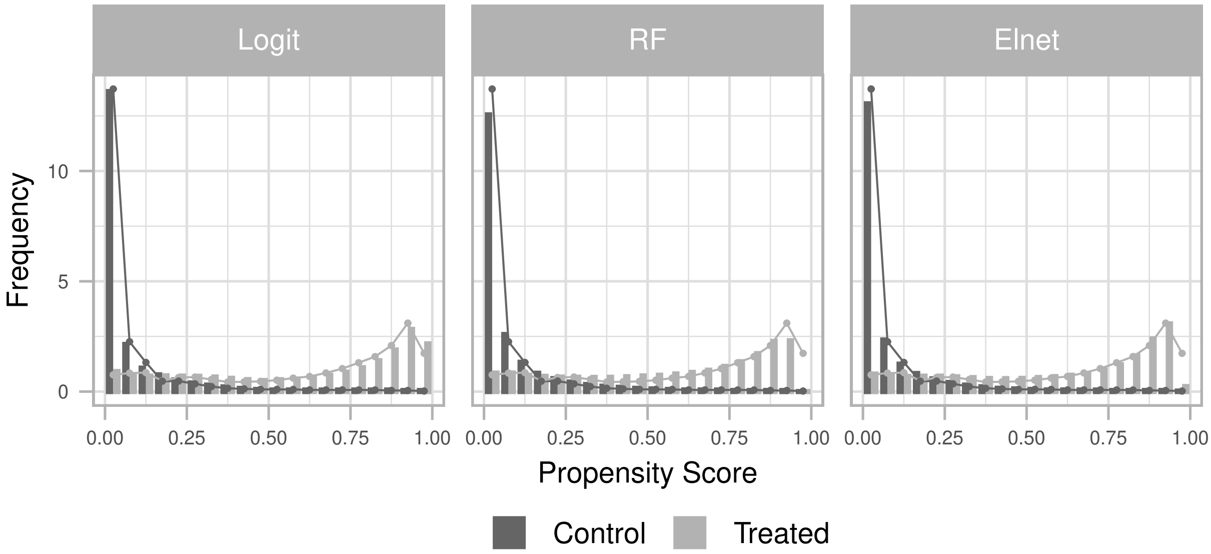
1.8.2 Results

Within-Estimator Comparison

As described at the beginning of Section 1.8, the LaLonde dataset consists of substantially more control observations than treated observations. Since the simulated datasets are based on functions fitted in the LaLonde dataset, this characteristic is also present in the simulated datasets. As Figure 1.12 illustrates, the estimated propensity score distributions have limited overlap. The mass of the distribution of the control group propensity scores is close to zero. As a result, there are few observations in the control group with an estimated propensity score greater than 0.1.[33] This indicates that overlap is limited in this setting.

[33] Each bin in the histogram contains a propensity score range of 0.05. Hence, the two leftmost bins indicate the frequency of observation with propensity scores 0 to 0.1.

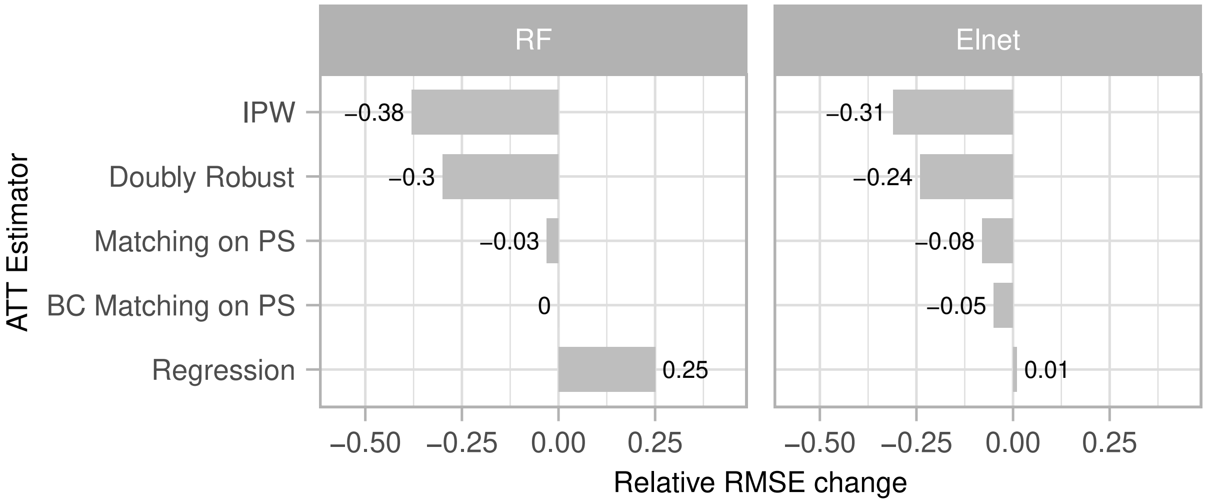
Figure 1.12: Distribution of Estimated Propensity Scores



Note: The graph indicates the median (over 5000 simulation replications) histogram of estimated propensity scores by logistic regression (Logit), random forests (RF), and elastic net (Elnet). The darker gray bars (line) represent the distribution of the estimated (true) propensity scores of control observations, while the lighter gray bars (line) represent the distribution of the estimated (true) propensity scores of treated observations.

Analogous to Section 1.7, for each treatment effect estimator I first compare conventional based estimation (OLS and logit) to machine learning based estimation (random forest and elastic net). The results are presented in Figure 1.13.

Figure 1.13: Within-Estimator Comparison Empirical Simulation Study



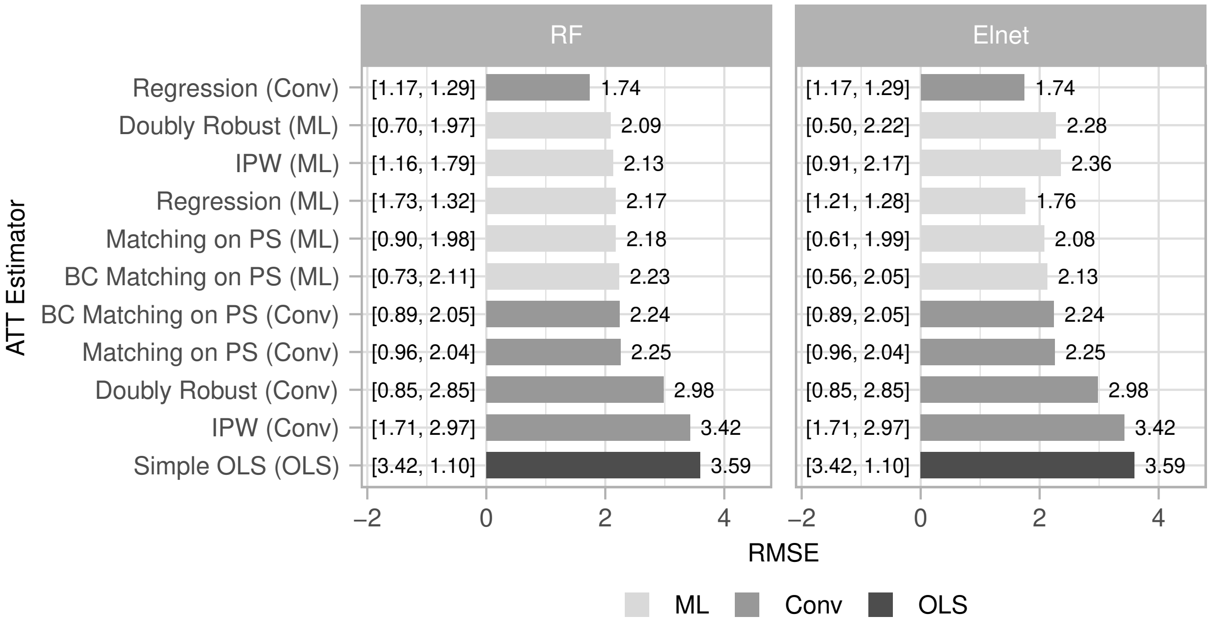
Note: Simulation based on 5000 replications. The bars indicate the relative change in RMSE of machine learning based estimation of the ATT (random forest in the left panel, elastic net in the right panel), compared to conventional based estimation of the ATT (OLS/Logit). A negative value indicates that the RMSE of the machine learning based estimator was lower, i.e. that the treatment effect was estimated more accurately with machine learning methods.

The estimators based on weighting (*IPW* and *Doubly Robust*) estimate the ATT much more accurately when the propensity score - and in case of the *Doubly Robust* estimator also the conditional outcome means - is estimated with machine learning methods. The largest improvements are observed for *IPW* (38% RMSE improvement with random forest, 31% RMSE improvement with elastic net). The RMSE of the *Doubly Robust* estimator decreases by 30% (RF) and 24% (Elnet). For the estimators based on matching (*Matching on PS* and *BC Matching on PS*), the RMSE improvements are much smaller. The RMSE of the *Matching on PS* estimator decreases by 3% (RF) and 8% (Elnet), the RMSE of the *BC Matching on PS* estimator by 5% (Elnet), while no RMSE change is observed for the random forest based version. The *Regression* estimator performs worse when the conditional outcome is estimated with machine learning methods. The increase in RMSE is substantial in the case of random forest (25%), and minor in the case of elastic net (1%). In summary, the within-estimator comparison for this dataset provides evidence that, with the exception of the *Regression* estimators, machine learning based estimators of the ATT are more accurate than conventional based estimators.

Between-Estimator Comparison

As in the first simulation, the between-estimator comparison indicates the performance of all treatment effect estimators (conventional and machine learning based). Again, I include the *Simple OLS* estimator regressing the outcome on the treatment indicator and the eight covariates as a benchmark. The results are presented in Figure 1.14.

Figure 1.14: Between-Estimator Comparison Empirical Simulation Study



Note: The bars indicate the RMSE of the ATT estimator over 5000 simulation replications. The numbers in brackets to the left of the bars indicate the absolute bias and the standard deviation of the estimator, i.e. [|Bias|, SD]. The light gray bars represent estimators that use machine learning methods (random forests and elastic net) to estimate the conditional outcome means and the propensity score. The dark gray bars represent estimators that use conventional methods (OLS and logit) to estimate the conditional outcome means and the propensity score. The black bar represents the simple OLS estimator. The results for random forest (RF) are presented in the left panel, and those for elastic net in the right panel.

In terms of RMSE, OLS based *Regression* (RMSE: 1.74) and elastic net based *Regression* (RMSE: 1.76) estimate the ATT most accurately. Thereafter, the performance of a large set of estimators is very similar. The RMSE of random forest based *Regression*, machine learning based *IPW* and *Doubly Robust*, and both versions of the *Matching on PS* and *BC Matching on PS* estimators range between 2.08 and 2.36. The conventional based versions of *IPW* (RMSE: 3.42) and *Doubly Robust* (RMSE: 2.98) perform worse. The ATT is estimated least accurately by *Simple OLS* (RMSE: 3.59).

Considering the bias of the treatment effect estimators, the machine learning based hybrid estimators *Doubly Robust* and *BC Matching on PS* exhibit the smallest absolute bias. Both the conventional and the machine learning based *Regression* estimators have substantially higher bias than the hybrid estimators. *Matching on PS* achieves relatively low bias, especially the elastic net based version. The *Simple OLS* estimator exhibits by far the largest bias (3.42).

In terms of standard deviation of the estimators, the *Simple OLS* estimator (1.10) and the *Regression* estimators have the smallest standard deviation. By contrast, the matching based estimators and *Doubly Robust* demonstrate relatively high variance. The conventional based *IPW* estimator has the highest variance.

Overall, I find that the good RMSE performance of OLS based *Regression* is due primarily to its relatively low variance. The hybrid estimators exhibit low bias, but have relatively high variance. As a result, I do not find evidence that hybrid estimators are able to estimate the ATT more accurately in terms of RMSE.

1.9 Within-Study Comparison: LaLonde Data

1.9.1 Within-Study Comparisons

In an influential contribution, LaLonde (1986) introduced the concept of *within-study comparisons*, which consist of two steps. In the first step, an experimental dataset is used to estimate the causal effect of treatment. Since treatment is randomly assigned in the experiment, there are no confounders. Therefore, an unbiased estimate of the causal effect is given by the difference in the mean outcomes of the treatment and control group. In a second step, either the experimental control group or the experimental treatment group is replaced by a nonexperimental comparison group. This should mimic a situation in which no experiment is available and confounding is a problem. The idea is then to analyze whether non-experimental econometric estimators are able to recover the ”true” causal effect from the experiment with the non-experimental data. LaLonde (1986) finds that results are sensitive to both econometric specification and subgroup used in analysis. In many cases, his non-experimental econometric estimators were not able to recover the ”true” causal effect.

1.9.2 Data

The analysis of LaLonde (1986) was based on the National Supported Work (NSW) Demonstration - a US job training program for disadvantaged workers. The goal of the program was to give participants work experience and assistance in a protected environment. For an in-depth description, see e.g. LaLonde (1986), Smith and Todd (2005), and Calónico and Smith (2017). The four target groups were 1) women receiving *Aid to Families with Dependent Children* (AFDC), 2) former drug addicts, 3) former offenders, and 4) high school dropouts. The main criteria for eligibility were that the person was unemployed at the time of being selected for the job training and had not been employed for more than three months in the preceding six months.[34] The eligible applicants were randomly assigned to either the treatment group (receiving job training) or the control group (receiving no job training). For both treatment and control group, data on earnings and other demographic variables was collected before and after the treatment.[35] The outcome of interest was earnings in the post-training year, which was 1978 for men and 1979 for women. The observed covariates were age, education, pre-treatment earnings in 1974 and 1975, a categorical variable for ethnicity,

[34] Further criteria for the group receiving AFDC were that they had been receiving AFDC for at least 30 out of the 36 preceding months and that they had no child younger than 6 years old.

[35] Sample attrition was a problem and possibly led to biased estimates of the program impact, see LaLonde (1986).

and dummies for being married and high school dropouts.[36] The job training program was voluntary, and only a small fraction of eligible people participated in the program.

LaLonde (1986) created nonexperimental comparison groups from Westat's Matched Current Population Survey - Social Security Administration File (CPS-SSA) and the Panel Study of Income Dynamics (PSID). The CPS-SSA and PSID are stratified random samples of the US population.[37] There are several problems with both the experimental and the non-experimental datasets of LaLonde (1986). First, the experimental dataset is rather small, both in terms of the number of observations and the number of covariates. The experimental dataset of male participants consists of 297 treated and 425 control observations, with eight covariates.[38] The limited number of observations affects the precision of the estimated causal effect. In the experimental dataset, LaLonde (1986) estimates a causal effect of $886. However, the standard error is $476 and thus relatively large. Another issue is that the individuals in the non-experimental comparison groups were often from a different local labor market (geographic mismatch), and earnings were measured in a different way to the experimental dataset (Smith & Todd, 2005).

1.9.3 Implementation

In this section I conduct a within-study comparison. I follow the implementation applied in Heckman et al. (1997), Heckman, Ichimura, Smith, and Todd (1998), and Smith and Todd (2005). These authors note that another way to evaluate whether non-experimental estimators are able to remove the selection bias is to compare the experimental control group with the non-experimental comparison groups. Therefore, the treatment indicator indicates whether the individual belongs to the experimental control group or the non-experimental comparison group. As both groups did not receive job training, the true causal effect is known to be zero. Thus, I first create a combined dataset consisting of the experimental control group and the PSID comparison group.[39] I then follow Advani et al. (2019) and apply a resampling procedure. I draw 5000 bootstrap samples from the original dataset. For each sample I draw 260 treated and 2490 control observations with replacement.

This dataset raises two fundamental problems related to the identifying assumptions. First, overlap of treated and control observations is very limited. This is illustrated in Figure 1.15. The vast majority of control group observations have an estimated propensity

[36] Earnings in 1974 are only available for the Dehejia and Wahba (1999) subgroup. This subgroup is used in the analysis.

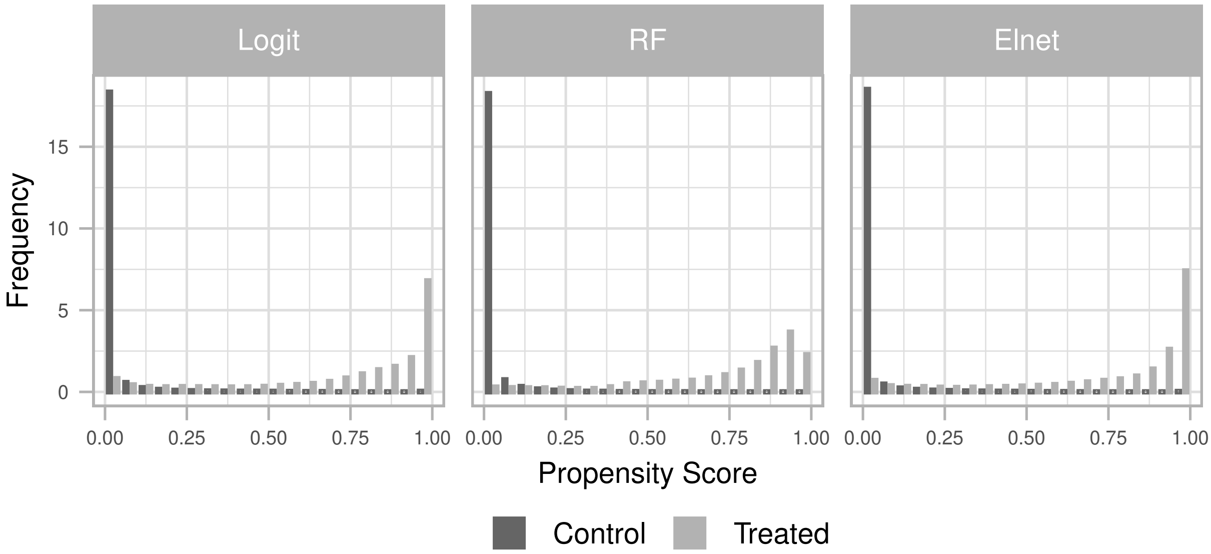
[37] Individuals older than 55 were excluded. In addition, individuals with a nominal own income of more than $20'000 (family income more than $30'000) were excluded from the CPS, and individuals reporting as being retired in 1975 were excluded from the PSID.

[38] The literature following LaLonde (1986) focused primarily on the subsample of male participants. One reason for this is that LaLonde (1986) concludes that "[...] econometric procedures are more likely to replicate the experimental results in the case of female rather than male participants." In other words, LaLonde (1986) finds that the selection problem is more pronounced for male participants. Recently, Calónico and Smith (2017) reanalyzed the female subsample.

[39] As in section 1.8, I consider the Dehejia and Wahba (1999) subsample of the NSW/PSID1 dataset considered in LaLonde (1986).

score smaller than 0.05. Only a few control group observations have an estimated propensity score above 0.75, where the mass of the distribution of treated observations is. Second, unlike in the two simulation studies, we do not know whether the unconfoundedness assumption holds. In fact, given the limited number of covariates and pre-treatment information, it is rather unlikely that unconfoundedness holds. For example, it might be required to include a more detailed employment history. As these two problems challenge the core assumptions of the treatment effect estimators, the results should be interpreted with caution.

Figure 1.15: Distribution of Estimated Propensity Scores

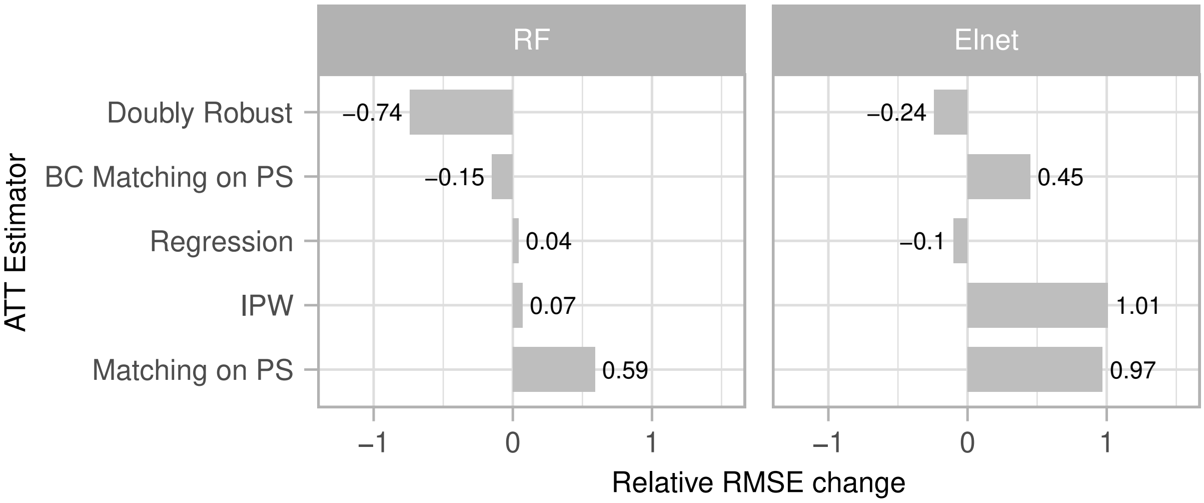


Note: The graph indicates the median (over 5000 simulation replications) histogram of estimated propensity scores by logistic regression (Logit), random forests (RF), and elastic net. The lines represent the distribution of the true propensity score. The darker gray represents the distribution of the estimated propensity scores of control observations, while the lighter gray represents the distribution of treated observations.

1.9.4 Results

Similar to the two simulation studies, I conduct a within-estimator comparison as well as a between-estimator comparison. The results for the within-estimator comparison are presented in Figure 1.16.

Figure 1.16: Within-Estimator Comparison LaLonde Data

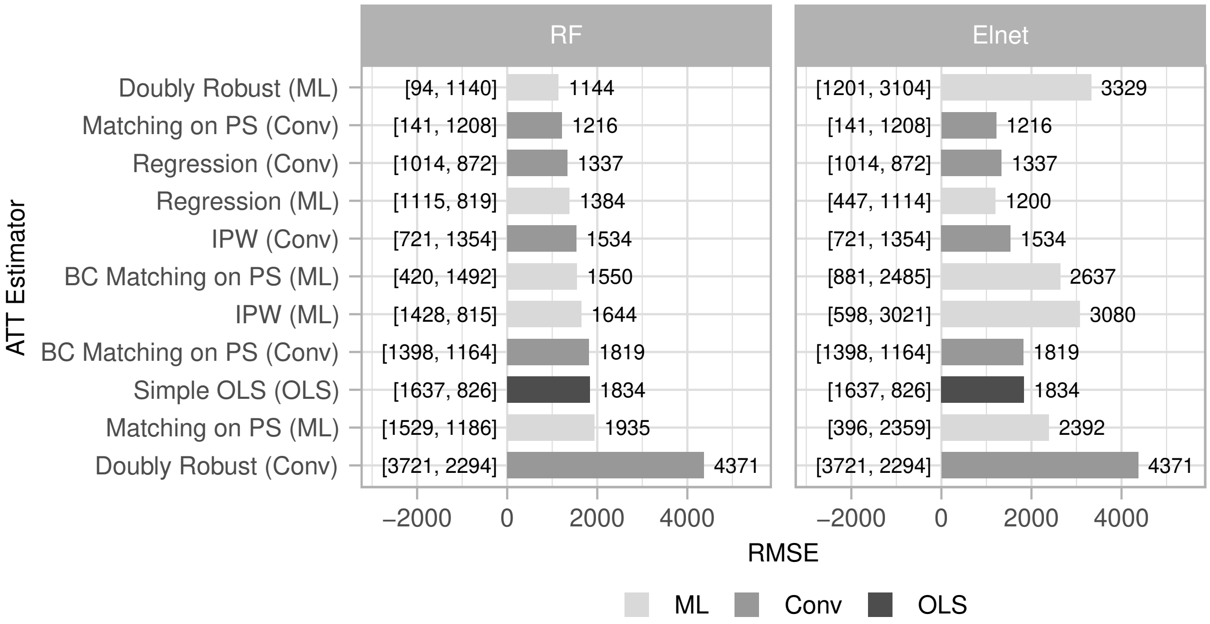


Note: Results based on 5000 bootstrap replications. The bars indicate the relative change in RMSE of machine learning based estimation of the ATT (random forest in the left panel, elastic net in the right panel), compared to conventional based estimation of the ATT (OLS/logit). A negative value indicates that the RMSE of the machine learning based estimator was lower, i.e. that the treatment effect was estimated more accurately with machine learning methods.

The results are mixed. For the *Doubly Robust* estimator, I find substantial improvements in RMSE when the propensity score and the conditional outcome means are estimated with random forests (74% improvement) or with elastic net (24% improvement). By contrast, I find that machine learning based *IPW* and *Matching on PS* perform considerably worse than their conventional counterpart. The elastic net based *IPW* and *Matching on PS* estimators, in particular, perform poorly (a 101 % and 97% deterioration, respectively). *BC Matching on PS* performs better (15% improvement) when the propensity score and the conditional outcome means are estimated with random forests, but worse (45% deterioration) with elastic net.

The results for the between-estimator comparison are presented in Figure 1.17.

Figure 1.17: Between-Estimator Comparison LaLonde Data



Note: The bars indicate the RMSE of the ATT estimator over 5000 bootstrap replications. The numbers in brackets to the left of the bars indicate the absolute bias and the standard deviation of the estimator, i.e. [|Bias|, SD]. The light gray bars represent estimators that use machine learning methods (random forests and elastic net) to estimate the conditional outcome means and the propensity score. The dark gray bars represent estimators that use conventional methods (OLS and logit) to estimate the conditional outcome means and the propensity score. The black bar represents the simple OLS estimator. The results for random forest (RF) are presented in the left panel, and those for elastic net in the right panel.

In terms of RMSE, random forest based *Doubly Robust* estimates the ATT most accurately (RMSE: 1144). This is in stark contrast to the conventional based *Doubly Robust* estimator, whose RMSE is almost four times larger (RMSE: 4371). Furthermore, the elastic net based estimators employing the propensity score (*Doubly Robust*, *BC Matching on PS*, *Matching on PS*, *IPW*) perform poorly.

With respect to the bias of the treatment effect estimators, the random forest based *Doubly Robust* estimator exhibits the lowest bias (94). The bias of the conventional based *Matching on PS* estimator is also reasonably low (141). The conventional based *Doubly Robust* estimator has the largest bias (3721).

In terms of standard deviation of the estimators, the random forest based *IPW* estimator (815), the random forest based *Regression* estimator (819), and the *Simple OLS* estimator (826) have the smallest standard deviation. By contrast, the elastic net based estimators employing the propensity score have the highest variance.

Overall, in line with the two simulation studies, I do not find evidence that hybrid estimators outperform the estimators that rely only on the propensity score or only on the conditional outcome means. Although the best performance is achieved by the random forest based *Doubly Robust* estimator, the performance of the elastic net based hybrid estimators is worrisome.

1.10 Supplementary Analysis

In this section I analyze the effect of two estimation choices a researcher has to make when applying the treatment effect estimators discussed in this chapter, especially when using the machine learning based estimators. I analyze the effects of cross-fitting and repeated sample splitting.

1.10.1 Cross-Fitting

Cross-fitting is described in Section 1.5.3. In the default specification, the number of cross-fitting folds is equal to five. I analyze the effect of changing the number of cross-fitting folds on the performance of the estimators. Figure 1.21 in Appendix 1.A presents the results for the simulation of Section 1.7.

I find that the RMSE differences between two-fold, five-fold, and ten-fold cross-fitting are small. Often, the performance is worst with only two cross-fitting folds. This is most pronounced for the *IPW* estimator. The benefits of increasing the number of cross-fitting folds from five to ten are modest. Since the associated increase in computational cost is sizable, this simulation provides evidence that five cross-fitting folds are sufficient.

1.10.2 Repeated Sample Splitting

As described in 1.5.3, cross-fitting randomly splits the sample into different folds. Since this split is random, the results in a finite sample depend on the split. The goal of repeated sample splitting is to decrease the dependency on a single random split. In small samples, this spit can be unrepresentative of the whole sample and thus influence the results. In the default specification, I do not consider repeated sample splitting for computational reasons. Figure 1.22 in Appendix 1.A presents the results for the default specification, as well as for 10 and 25 repeated sample splits. That is, in each of the 5000 simulation replications, I take the median over 10 and 25 estimated treatment effects, respectively.[40]

Even with only 10 repeated sample splits, there are certain RMSE improvements of repeated sample splitting. The largest RMSE improvements are observable for *BC Matching on PS* and *Matching on PS*, especially their machine learning based versions. I find no changes for the *Regression* estimator and only minor improvements for the *Doubly Robust* estimator. The difference between 10 and 25 repeated sample splits is small.

1.11 Conclusion

This chapter has analyzed the performance of treatment effect estimators assuming unconfoundedness. I examined the regression estimator, matching on the propensity score, inverse

[40] For computational reasons, I only examine 10 and 25 repeated sample splits. In an empirical example, Chernozhukov et al. (2018) use 100 repeated sample splits.

probability weighting, bias-corrected matching on the propensity score, and the doubly robust estimator.

To answer the first research question, I analyzed whether estimating the propensity score and/or the conditional outcome means with machine learning methods increases the accuracy of treatment effect estimation compared to estimating these functions with OLS and/or logit. The two simulation studies provide evidence that in many cases the machine learning based estimators are more accurate than the estimators relying on OLS and/or logit. The differences are most pronounced when both the propensity score and the conditional outcome means are misspecified. The results from the within-study comparison are mixed, and should be interpreted with caution since the identifying assumptions are potentially violated.

To answer the second research question, I analyzed whether hybrid methods estimate the treatment effects more accurately than estimators that rely either only on the propensity score or only on the conditional outcome means. I do not find evidence that hybrid estimators generally outperform the other estimators. However, hybrid estimators are often among the estimators with the lowest RMSE. In many cases, hybrid estimators exhibit the lowest bias, sometimes at the cost of an increased variance. This could be interpreted as the cost of allowing misspecification in either the propensity score or the conditional outcome means. It might therefore be advantageous to use hybrid methods to guard against misspecification, especially if one is more concerned about bias than variance.

To answer the third research question, I analyzed how the accuracy of the estimators depends on changes in a) the degree of linearity and additivity in the relationships between treatment and covariates and between outcome and covariates, and b) the strength of selection into treatment. I find that in cases where the underlying relationships are linear and additive, or when selection into treatment is weak, the accuracy of the treatment effect estimators is similar. Moreover, the differences between machine learning based and conventional based estimators are small. However, when the underlying relationships become nonlinear and nonadditive, or when selection into treatment is more pronounced, and the estimators use a misspecified propensity score and/or conditional outcome means, the machine learning based estimators outperform the conventional based estimators.

In a supplementary analysis, I find that the RMSE differences between two-fold, five-fold, and ten-fold cross-fitting are small. Often, the performance is worst with only two cross-fitting folds. In addition, I find that repeated sample splitting improves treatment effect estimation in some cases.

This chapter has several limitations. As discussed in Section 1.2, there are numerous ways to conduct a simulation study. Every decision a researcher takes when designing a simulation study could potentially have an effect on the findings. In the simulation in Section 1.7, for example, the decision that the default specification includes nonlinearities undoubtedly has an effect on the results. Moreover, the way the nonlinearities are introduced potentially affects the results. Even though the simulation in Section 1.8 tries to imitate a real dataset, it is not clear whether the findings can be generalized to a real empirical application. Whether it is beneficial to use machine learning methods to control for covariates will depend on the unknown underlying relationships. If the underlying relationships are linear and additive, linear and additive models will most likely work well, and flexible machine learning methods will be of limited use. Furthermore, the scope of the study is limited. I only apply five treatment effect estimators. Several other interesting estimators have been proposed - for example approximate residual balancing (Athey, Imbens, & Wager, 2018). Moreover, I implemented rather basic versions of these five estimators. Various alternatives for these estimators were suggested, for example normalizing the *IPW* weights so that they add up to one, using more than one nearest neighbor in *Matching on PS*, or *Matching on PS* without replacement. Further, I only consider the machine learning methods random forests and elastic net. It would be very interesting to analyze the accuracy of treatment effect estimators using other machine learning methods, such as boosting or neural networks. It could also be interesting to use ensemble methods that combine different machine learning methods.

It is important to stress that the treatment effect estimators discussed in this chapter rely on the unconfoundedness assumption. If unconfoundedness does not hold, neither machine learning based estimators nor hybrid estimators will solve this problem. Given that the unconfoundedness assumption is credible, it might be recommendable to estimate the treatment effect with more than one estimator, and compare the estimated treatment effects. This analysis suggests the use of the hybrid estimators *Doubly Robust* and *BC Matching on PS*, especially if one is more concerned about bias than variance. The simulation results further indicate that the *Simple OLS* estimator - i.e. regressing the outcome on the treatment indicator and the covariates - often performs poorly. It is therefore not recommended to rely only on the *Simple OLS* estimator.

1.12 Bibliography

Abadie, A., & Cattaneo, M. D. (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, *10*, 465–503.

Abadie, A., & Imbens, G. W. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, *74*(1), 235–267.

Abadie, A., & Imbens, G. W. (2011). Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics*, *29*(1), 1–11.

Abadie, A., & Imbens, G. W. (2016). Matching on the Estimated Propensity Score. *Econometrica*, *84*(2), 781–807.

Advani, A., Kitagawa, T., & Słoczyński, T. (2019). Mostly Harmless Simulations? Using Monte Carlo Studies for Estimator Selection. *Journal of Applied Econometrics*, *34*(6), 893–910.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics.* Princeton University Press.

Athey, S. (2018). The Impact of Machine Learning on Economics. In *The economics of artificial intelligence: An agenda* (pp. 507–547).

Athey, S., & Imbens, G. (2017). The State of Applied Econometrics - Causality and Policy Evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Athey, S., & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, *11*, 685–725.

Athey, S., Imbens, G. W., Metzger, J., & Munro, E. (2019). Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations. *Stanford GSB Working Paper*, *No. 3824*.

Athey, S., Imbens, G. W., & Wager, S. (2018). Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *80*(4), 597–623.

Bang, H., & Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, *61*, 962–972.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Busso, M., DiNardo, J., & McCrary, J. (2014). New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators. *The Review of Economics and Statistics*, *96*(5), 885–897.

Calónico, S., & Smith, J. (2017). The Women of the National Supported Work Demonstration. *Journal of Labor Economics*, *35*(S1), S65-S97.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications.* Cambridge University Press.

Cannas, M., & Arpino, B. (2019, 7). A Comparison of Machine Learning Algorithms and Covariate Balance Measures for Propensity Score Matching and Weighting. *Biometrical Journal*, *61*(4), 1049–1072.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., &

Robins, J. (2018). Double/debiased Machine Learning for Treatment and Structural Parameters. *Econometrics Journal*, *21*(1), C1-C68.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing With Limited Overlap in Estimation of Average Treatment Effects. *Biometrika*, *96*(1), 187–199.

Dehejia, R. H., & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluation the Evaluation of Training Programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics*, *84*(1), 151–161.

Diamond, A., & Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics & Statistics*, *95*(3), 932–945.

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Frölich, M. (2004). Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators. *The Review of Economics and Statistics*, *86*(1), 77–90.

Frölich, M., Huber, M., & Wiesenfarth, M. (2017). The Finite Sample Performance of Semi- and Non-Parametric Estimators for Treatment Effects and Policy Evaluation. *Computational Statistics & Data Analysis*, *115*, 91–102.

Goller, D., Lechner, M., Moczall, A., & Wolff, J. (2019). Does the Estimation of the Propensity Score by Machine Learning Improve Matching Estimation? The Case of Germany's Programmes for Long Term Unemployed. *IZA Discussion Papers*, *12526*.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average. *Econometrica*, *66*(2), 315–331.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, *66*(5), 1017–1098.

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching As An Economic Evaluation Estimator. *Review of Economic Studies*, *65*(2), 261–294.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, *64*(4), 605–654.

Hirano, B. Y. K., Imbens, G. W., & Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, *71*(4), 1161–1189.

Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, *47*(260), 663–685.

Huber, M. (2019). An Introduction to Flexible Methods for Policy Evaluation. *FSES Working Papers*, *504*.

Huber, M., Lechner, M., & Wunsch, C. (2013). The Performance of Estimators Based on the Propensity Score. *Journal of Econometrics*, *175*, 1–21.

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity : A Review. *The Review of Economics and Statistics*, *86*(1), 4–29.

Imbens, G. W. (2019). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *NBER Working Paper Series*, *No. 26104*.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences.* Cambridge University Press.

Imbens, G. W., Whitney, N., & Ridder, G. (2006). *Mean-squared-error Calculations for Average Treatment Effects.* Retrieved from `https://scholar.harvard.edu/files/imbens/files/mean-squared-error_calculations_for_average_treatment_effects.pdf`

Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer.

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

Lechner, M., & Strittmatter, A. (2019). Practical Procedures to Deal With Common Support Problems in Matching Estimation. *Econometric Reviews*, *38*(2), 193–207.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving Propensity Score Weighting Using Machine Learning. *Statistics in Medicine*, *29*(3), 337–346.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Lunceford, J. K., & Davidian, M. (2004). Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study. *Statistics in Medicine*, *23*(19), 2937–2960.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Pearl, J. (2009). *Causality.* Cambridge University Press.

Pirracchio, R., Petersen, M. L., & Van Der Laan, M. (2015). Practice of Epidemiology Improving Propensity Score Estimators' Robustness to Model Misspecification Using Super Learner. *American Journal of Epidemiology*, *181*(2), 108–119.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression

Models with Missing Data. *Journal of the American Statistical Association*, *90*, 122–129.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, *89*(427), 846–866.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*(1), 41–55.

Sekhon, J. S. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. *Journal of Statistical Software*, *42*(7), 1–52.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study. *Pharmacoepidemiology and Drug Safety*, *17*, 546–555.

Smith, J. A., & Todd, P. E. (2005). Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? *Journal of Econometrics*, *125*, 305–353.

Zhao, Z. (2004). Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence. *The Review of Economics and Statistics*, *86*(1), 91–107.

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*(2), 301–320.

Appendix 1.A Appendix to Stylized Simulation Study

1.A.1 Data Generating Process

The analysis of different degrees of linearity and additivity in Section 1.7.3 is based on the following specifications.

$$\begin{aligned}
D^*_L =& \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\
D^*_{mNL} =& D^*_L + \beta_2 X_2^2 + 0.5\beta_1 X_1 X_3 + 0.7\beta_2 X_2 X_4 + 0.5\beta_4 X_4 X_5 + 0.5\beta_5 X_5 X_6 \\
D^*_{NL} =& D^*_L + \beta_2 X_2^2 + \beta_4 X_4^2 + \beta_7 X_7^2 + \\
& 0.5\beta_1 X_1 X_3 + 0.7\beta_2 X_2 X_4 + 0.5\beta_3 X_3 X_5 + 0.7\beta_4 X_4 X_6 + 0.5\beta_5 X_5 X_7 + \\
& 0.5\beta_1 X_1 X_6 + 0.7\beta_2 X_2 X_3 + 0.5\beta_3 X_3 X_4 + 0.5\beta_4 X_4 X_5 + 0.5\beta_5 X_5 X_6 \ ,
\end{aligned}$$

where D^*_L is the linear and additive specification, D^*_{mNL} the mild nonlinear and nonadditive specification, and D^*_{NL} is the nonlinear and nonadditive specification. The three specifications correspond to the scenarios A, E, and G in Diamond and Sekhon (2013). The β coefficients are the same as in Section 1.7.3. The same applies to the outcome specifications:

$$\begin{aligned}
Y^*_L =& \tau D + \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_8 X_8 + \alpha_9 X_9 + \alpha_{10} X_{10} \\
Y^*_{mNL} =& Y^*_L + \alpha_2 X_2^2 + 0.5\alpha_1 X_1 X_3 + 0.7\alpha_2 X_2 X_4 + 0.5\alpha_4 X_4 X_9 + 0.5\alpha_8 X_8 X_{10} \\
Y^*_{NL} =& Y^*_L + \alpha_2 X_2^2 + \alpha_4 X_4^2 + \alpha_{10} X_{10}^2 + \\
& 0.5\alpha_1 X_1 X_3 + 0.7\alpha_2 X_2 X_4 + 0.5\alpha_3 X_3 X_4 + 0.7\alpha_4 X_4 X_8 + 0.5\alpha_8 X_8 X_{10} + \\
& 0.5\alpha_1 X_1 X_{10} + 0.7\alpha_2 X_2 X_3 + 0.5\alpha_3 X_3 X_9 + 0.5\alpha_4 X_4 X_{10} + 0.5\alpha_9 X_9 X_{10} \ ,
\end{aligned}$$

where Y^* denotes the outcome without the effect of treatment and without the error term.

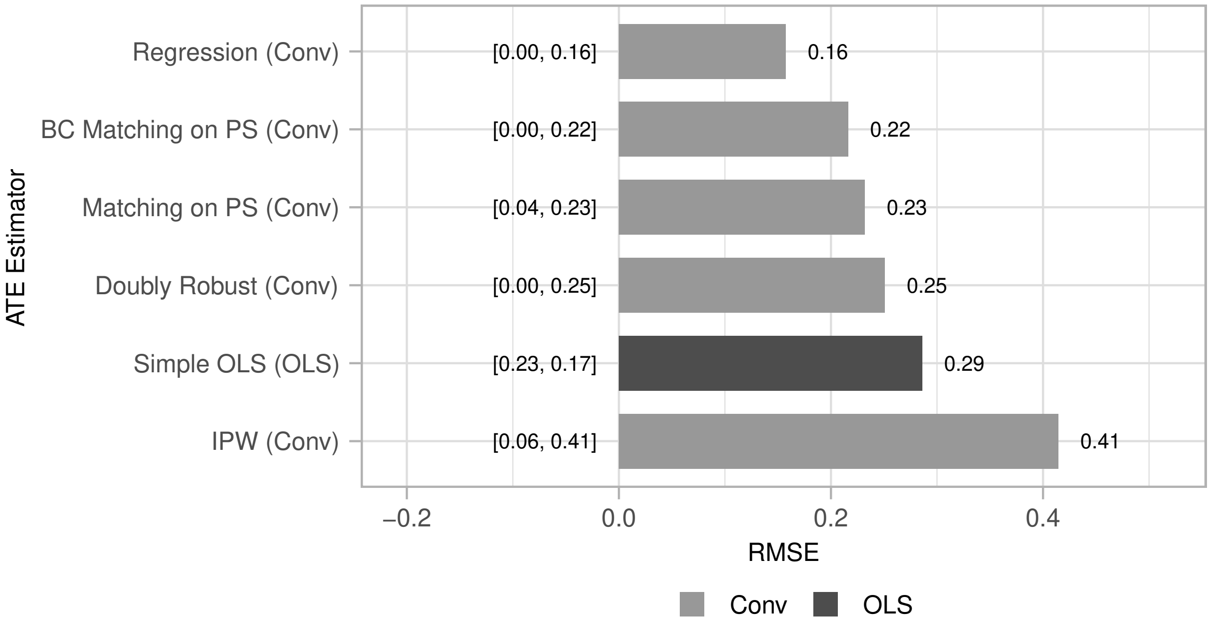
In order to determine σ^2_ϵ, I first calculate the McFadden pseudo R^2 of the treatment specification as $R^2_{pseudo} = 1 - \frac{\log(L_{full})}{\log(L_{null})}$, where L_{full} is the Likelihood of the logistic regression model, including all variables used in the propensity score specification, and L_{null} is the likelihood of the null model (logistic regression) including only an intercept. I then calculate the variance of the outcome without noise term, denoted by $Var(Y^*)$. Then,

$$\begin{aligned}
R^2 &= \frac{Var(Y^*)}{Var(Y)} \\
R^2 &= \frac{Var(Y^*)}{Var(Y^*) + \sigma^2_\epsilon} \\
\sigma^2_\epsilon &= \frac{1 - R^2}{R^2} V(Y^*)
\end{aligned}$$

Substituting R^2 for R^2_{pseudo} yields $\sigma^2_\epsilon = \frac{1-R^2_{pseudo}}{R^2_{pseudo}} V(Y^*)$. In 1000 simulations of D and Y (independent from the simulations in Section 1.7.3), the mean of the R^2_{pseudo} was 0.263 (SD: 0.022), the mean of σ^2_ϵ was 4.628 (SD: 0.605), and the mean of the R^2 was 0.274 (SD: 0.032).

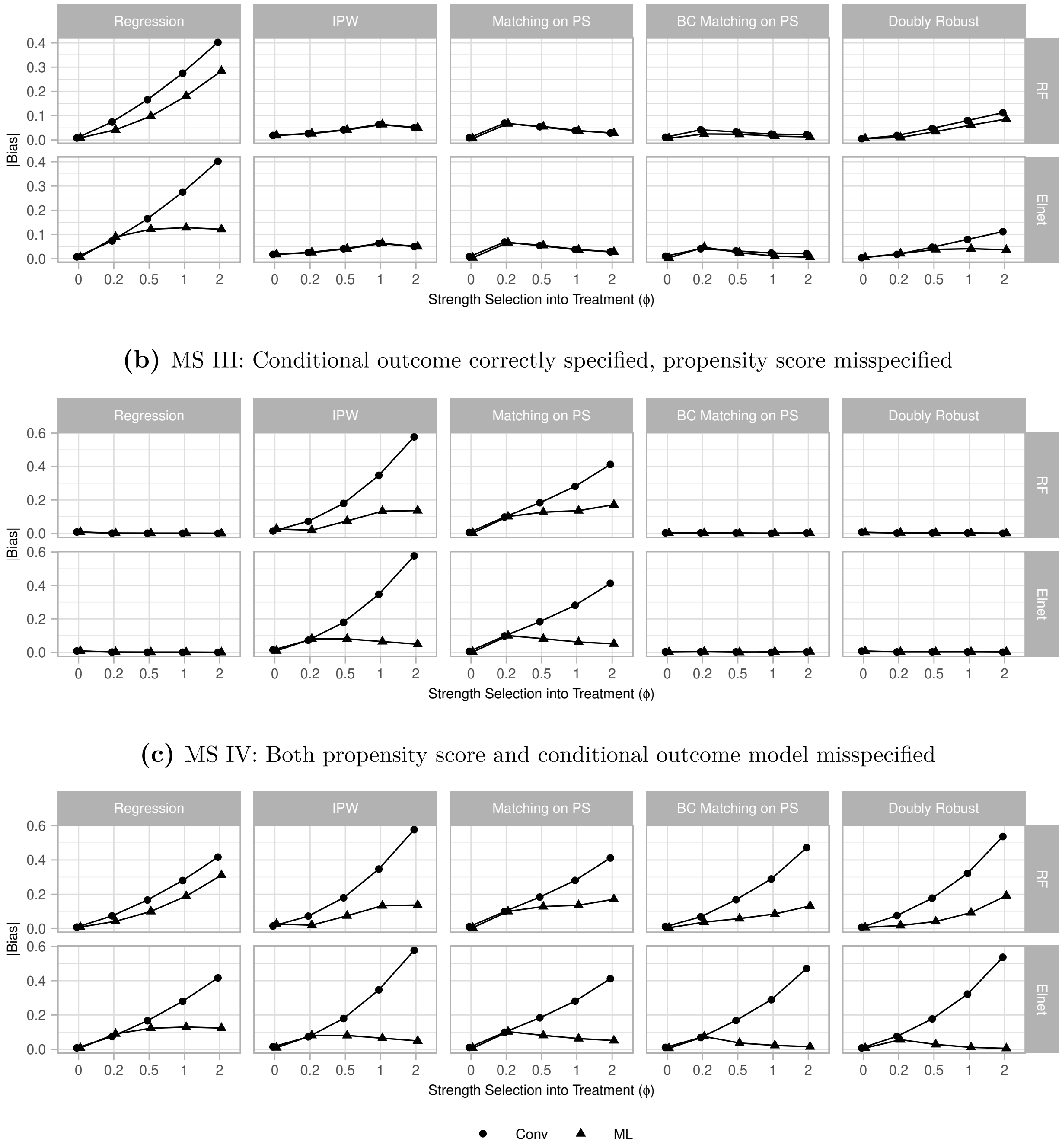
1.A.2 Additional Results

Figure 1.18: Between-Estimator Comparison for Misspecification Scenario I



Note: The bars indicate the RMSE of the ATE estimator over 5000 simulation replications. The dark gray bars represent estimators that use conventional methods to estimate the conditional outcome means (OLS) and the propensity score (logit). The black bar represents the OLS estimator regressing the outcome on the treatment indicator and the ten covariates.

Figure 1.19: Analysis of Changes in the Strength of Selection Into Treatment: Bias



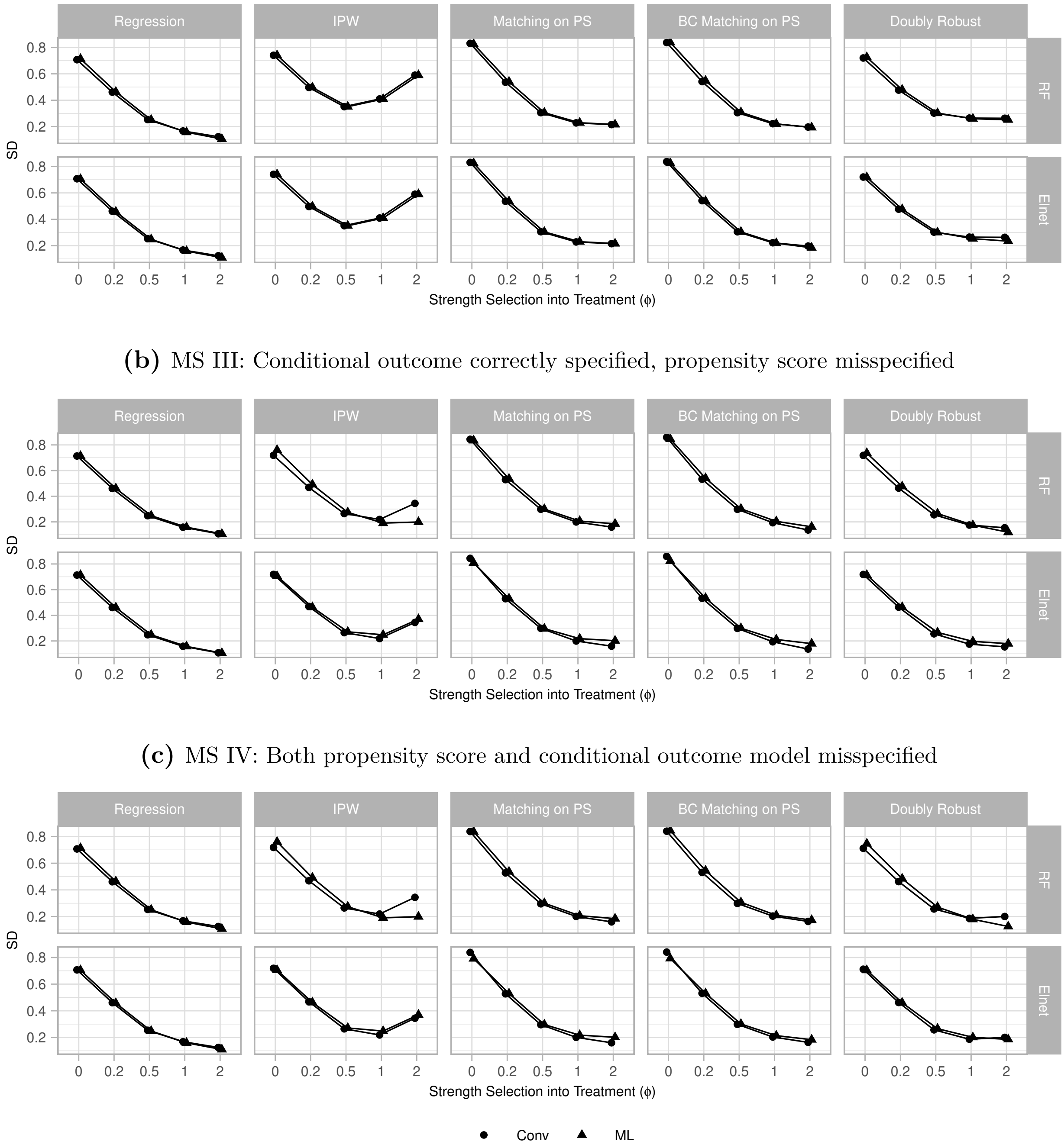
(a) MS II: Propensity score correctly specified, conditional outcome misspecified

(b) MS III: Conditional outcome correctly specified, propensity score misspecified

(c) MS IV: Both propensity score and conditional outcome model misspecified

Note: The top/middle/bottom panels represent misspecification scenarios II/III/IV. In each panel, the columns represent the treatment effect estimators, the rows indicate whether random forest (RF) or elastic net (Elnet) was used to estimate the propensity score and/or the conditional outcome means. In each subgraph, the x-axis plots ϕ, the strength of selection into treatment. The larger ϕ, the stronger is the selection into treatment. With $\phi = 0$, selection into treatment is random. The default specification corresponds to $\phi = 1$. The shape of the points represents the conventional based estimators (circle) and the machine learning based estimators (triangle).

Figure 1.20: Analysis of Changes in the Strength of Selection Into Treatment: SD



(a) MS II: Propensity score correctly specified, conditional outcome misspecified

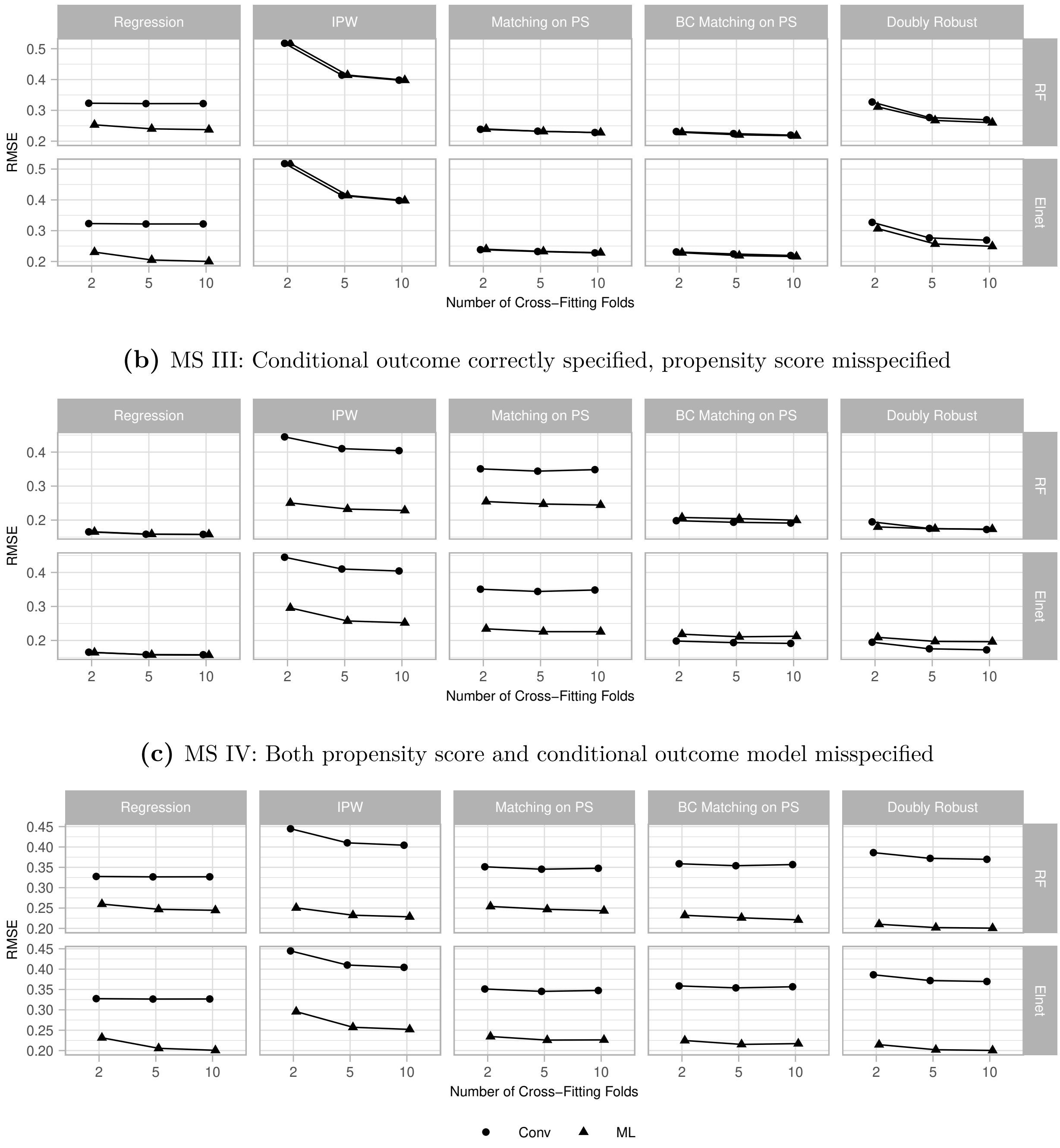
(b) MS III: Conditional outcome correctly specified, propensity score misspecified

(c) MS IV: Both propensity score and conditional outcome model misspecified

Note: The top/middle/bottom panels represent misspecification scenarios II/III/IV. In each panel, the columns represent the treatment effect estimators, the rows indicate whether random forest (RF) or elastic net (Elnet) was used to estimate the propensity score and/or the conditional outcome means. In each subgraph, the x-axis plots ϕ, the strength of selection into treatment. The larger ϕ, the stronger is the selection into treatment. With $\phi = 0$, selection into treatment is random. The default specification corresponds to $\phi = 1$. The shape of the points represents the conventional based estimators (circle) and the machine learning based estimators (triangle).

Figure 1.21: Analysis of Cross-Fitting

(a) MS II: Propensity score correctly specified, conditional outcome misspecified



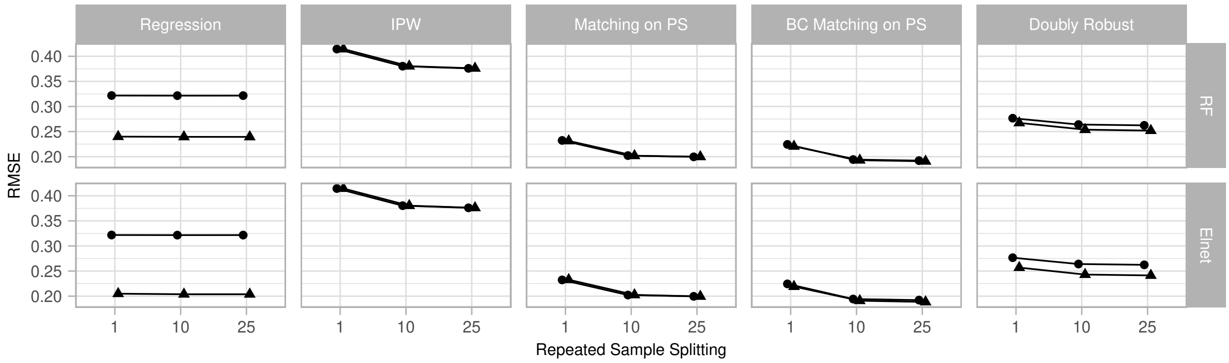
(b) MS III: Conditional outcome correctly specified, propensity score misspecified

(c) MS IV: Both propensity score and conditional outcome model misspecified

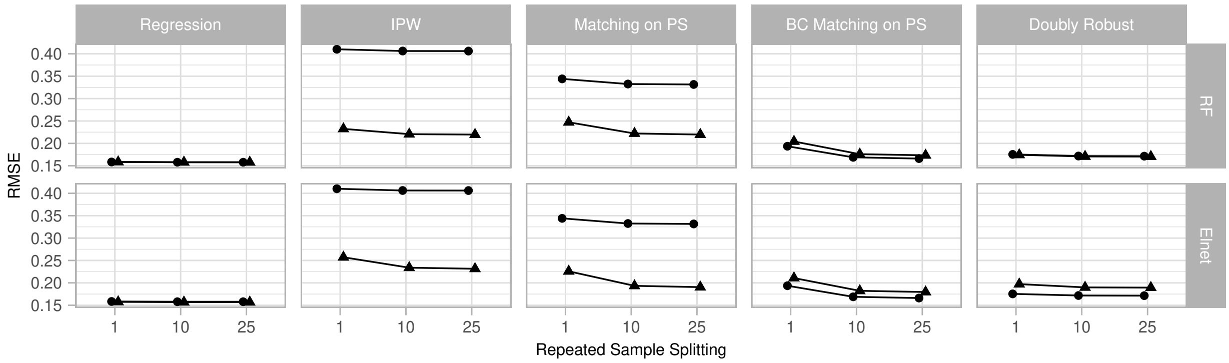
Note: The top/middle/bottom panels represent misspecification scenarios II/III/IV. In each panel, the columns represent the treatment effect estimators, the rows indicate whether random forest (RF) or elastic net (Elnet) was used to estimate the propensity score and/or the conditional outcome means. In each subgraph, the x-axis plots the number of cross-fit folds. The shape of the points represents the conventional based estimators (circle) and the machine learning based estimators (triangle).

Figure 1.22: Analysis of Repeated Sample Splitting

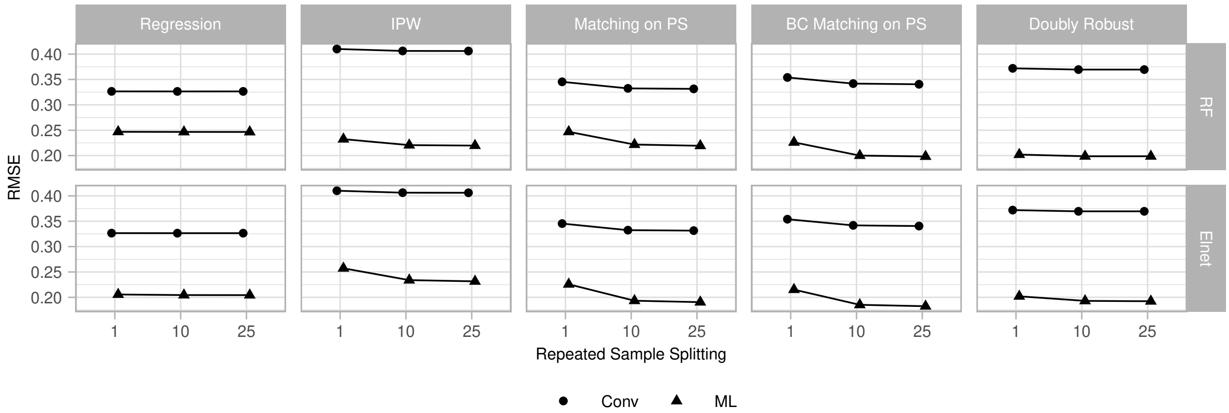
(a) MS II: Propensity score correctly specified, conditional outcome misspecified



(b) MS III: Conditional outcome correctly specified, propensity score misspecified



(c) MS IV: Both propensity score and conditional outcome model misspecified



Note: The top/middle/bottom panels represent misspecification scenarios II/III/IV. In each panel, the columns represent the treatment effect estimators, the rows indicate whether random forest (RF) or elastic net (Elnet) was used to estimate the propensity score and/or the conditional outcome means. In each subgraph, the x-axis plots the number of repeated sample splits. If the number of repeated sample splits is equal to 1, no repeated sample splitting was applied. The shape of the points represents the conventional based estimators (circle) and the machine learning based estimators (triangle).

Chapter 2

Identification and Estimation of Causal Intensive Margin Effects by Difference-in-Difference Methods[1]

2.1 Introduction

A decomposition of a binary treatment into extensive and intensive margin effects is of special interest when studying outcomes with a corner solution at zero.[2] Outcomes with corner solutions include working hours, health expenditures, and trade volumes. The average effect of a treatment on an outcome with a corner solution at zero can be decomposed into 1) the average change in the outcome of those with a positive outcome irrespective of treatment, plus 2) the average outcome of those with a positive outcome in case of treatment and a zero outcome in case of no treatment, minus 3) the average outcome of those with a zero outcome in case of treatment and a positive outcome in case of no treatment (Lee, 2012, 2017; Staub, 2014). Part 1) represents the weighted causal intensive margin effect. The sum of 2) and 3) captures the weighted causal extensive margin effect.[3]

Take as an example the effect of the introduction of a partial retirement policy on labor supply. Suppose that in the status quo, individuals must withdraw the full pension at a given age, but are allowed to continue working. Under the partial retirement policy, individuals have the choice between a partial and a full pension, and are allowed to continue working. The total effect of such a policy on labor supply might be zero, suggesting that the policy has been ineffective. The zero result, however, could be explained by a positive extensive margin effect that was offset by a negative intensive margin effect. Older workers who

[1] This chapter is joint work with Markus Hersche. An earlier version was published as working paper "Identification of Causal Intensive Margin Effects by Difference-in-Difference Methods", *CER-ETH Economics Working Paper Series, 11/2018*, see Hersche and Moor (2018). This chapter is also part of the doctoral thesis "Theoretical and Empirical Essays on Labor Supply of the Elderly", *Diss. ETH No. 25377*, see Hersche (2018).

[2] Corner solutions at alternative thresholds are possible as well. For simplicity and illustration, we consider the case where the threshold is at zero.

[3] The weights are given by the relative size of the group in the population.

would have retired in the absence of a partial retirement policy, now decide to stay in the labor market. Likewise, individuals who would have worked full-time in the absence of a partial retirement policy, now decide to work part-time. In such cases the total effect masks interesting subeffects at the extensive and intensive margin.

Even if treatment is randomly assigned, estimating intensive margin effects is challenging. A mean comparison of treatment and control groups with positive outcomes does not identify the causal intensive margin effect without additional assumptions (Angrist, 2001). In the labor supply example, the sample of individuals with positive working hours consists of two groups: 1) the group of individuals with positive working hours irrespective of whether they are treated or not, i.e. irrespective of whether they have the possibility to withdraw a partial pension; and 2) the group of individuals with positive working hours only because they are treated, i.e. only because they have the possibility to withdraw a partial pension, who would not work if they could only withdraw the full pension.[4] For the causal intensive margin effect, we are only interested in the first group. Group membership is however not observed in the data, because we observe either the outcome in case of treatment or the outcome in case of no treatment. The unobserved characteristics of the two groups are likely to be different. Individuals in the first group might be more motivated than individuals in the second group. Therefore, average working hours in the first group are likely higher than in the second group. As a result, a difference in the means of treated and untreated - conditional on positive working hours - could be the result of differences in these unobserved characteristics, and not because of a causal effect of treatment.

This constitutes a selection problem. In a general setting without random treatment, we are thus faced with two selection problems. The first selection problem is the standard selection problem in observational studies. In the presence of confounding variables, a mean comparison of treated and untreated individuals does not identify the causal effect. The second selection problem arises because we condition on positive outcomes. Difference-in-difference methods were developed to deal with the first selection problem. Using data from pre- and post-treatment periods, difference-in-difference allows for some selection on unobservables. This comes at the cost of making an assumption about outcome trends over time. It seems reasonable to extend the difference-in-difference methodology to include the second selection problem as well.

In this chapter, we discuss difference-in-difference methods to estimate the causal intensive margin effect. In contrast to standard difference-in-difference estimators, we condition the sample on individuals with positive outcomes.[5] We derive sufficient conditions under which the causal intensive margin effect is identified. Compared to standard difference-in-difference methods, two monotonicity assumptions are additionally required to identify the

[4] Here we neglect a possible third group, the group of individuals with positive working hours only because they are not treated, i.e. only because they *do not* have the possibility to withdraw a partial pension, and who would not work if they had the choice between a partial and a full pension. In this example this case seems rather unlikely.

[5] We refer to the term *standard difference-in-difference* to denote difference-in-difference methods that do not condition on positive outcomes.

causal intensive margin effect. We apply the difference-in-difference methodology to estimate the causal intensive margin effect of reaching the full retirement age on working hours. Moreover, we discuss how the identifying assumptions can be motivated in practice.

The main contribution of this chapter is to extend the literature on identification and estimation of intensive margin effects by borrowing well established difference-in-difference methods from the policy evaluation literature. The intensive margin effect is of interest in cases where the total effect masks relevant subeffects, e.g. when the extensive and intensive margin effect have different signs. The difference-in-difference estimator on positive outcomes represents an alternative to estimating the intensive margin effect with models for outcomes with corner solutions or selection models.

This chapter is thus related to the literature on models for outcomes with corner solutions, e.g. Tobit (McDonald & Moffitt, 1980; Tobin, 1958) or two-part models (Cragg, 1971; Duan, Manning, Morris, & Newhouse, 1983), and selection models (Heckman, 1979). Moreover, this chapter is connected to the literature employing principal stratification (Frangakis & Rubin, 2002) to study causal extensive and intensive margin treatment effects for variables with nonnegative outcomes (Lee, 2012, 2017; Staub, 2014). This literature decomposes the average treatment effect into a population-weighted sum of treatment effects of participants and switchers.[6] Studying outcomes with a corner solution at zero, Staub (2014) derives nonparametric bounds for the treatment effects of participants and switchers. He further discusses point identification of causal intensive and extensive margin effects in censored regression, selection, and two-part models. Lee (2012, 2017) analyzes total, extensive, and intensive margin effects in general sample selection models, with the corner solution outcome as a special case. Lee (2012) analyzes nonparametric methods to estimate extensive and intensive margin effects, whereas Lee (2017) discusses point identification of intensive and extensive margin effects in semiparametric linear models. The idea of principal stratification is also used in instrumental variable approaches (Angrist, Imbens, & Rubin, 1996), and in mediation analysis (Deuchert, Huber, & Schelker, 2019). In instrumental variable approaches, the stratification is based on the treatment variable (always-takers, compliers, defiers, never-takers), whereas in mediation analysis, the stratification in based on the mediator. In our context, the stratification is based on the outcome variable. More generally, this chapter often draws upon Lechner (2010), who provides a survey on difference-in-difference methods from a potential outcomes perspective.

The remainder of this chapter is organized as follows. Section 2.2 introduces the notation and describes the conventional as well as the causal decomposition of a treatment effect. Identification of the causal intensive margin effect is described in Section 2.3. Section 2.4 discusses estimation and inference. An empirical application is presented in Section 2.5. The last section concludes.

[6] Participants represent individuals with a positive outcome irrespective of treatment. Switchers represent individuals with a positive outcome in case of treatment and a zero outcome in case of no treatment, as well as individuals with a zero outcome in case of treatment and a positive outcome in case of no treatment.

2.2 Notation and Decomposition of a Treatment Effect

2.2.1 Notation

We consider the standard potential outcome framework with a non-negative outcome Y and a binary treatment D (Rubin, 1974), extended to two periods (Lechner, 2010). We observe individuals in the pre-treatment period $t-1$, and in the post-treatment period t; that is we observe $Y_{i,t-1}$ and $Y_{i,t}$. In each period, each individual i has two potential outcomes. The potential outcomes in case of treatment ($D_i = 1$) are denoted by $Y^1_{i,t}$ and $Y^1_{i,t-1}$, and in case of no treatment ($D_i = 0$) by $Y^0_{i,t}$ and $Y^0_{i,t-1}$.[7] In each period, we only observe one of the two potential outcomes. Moreover, each individual is characterized by a vector of observed covariates X_i, assumed to be constant over time. The starting point of the decompositions described in Sections 2.2.2 and 2.2.3 is the average treatment effect on the treated (ATT), defined as

$$ATT_t = E(Y^1_{i,t} - Y^0_{i,t} | D_i = 1) \, . \tag{2.1}$$

The ATT measures the expected treatment effect for a treated observation.

2.2.2 Conventional Decomposition

As described in Section 2.1, the estimation of causal intensive margin effects entails two selection problems. The first selection problem arises from confounding variables, the second selection problem arises from conditioning on observations with positive outcomes. To illustrate the second selection problem, we consider in this subsection the case of random treatment assignment, and thus eliminate the first selection problem. This illustration closely follows Staub (2014). Random treatment assignment implies that treatment is independent of the potential outcomes, i.e. $(Y^1_{i,t}, Y^0_{i,t}) \perp\!\!\!\perp D_i$. Hence, the ATT at time t is identified by the difference in mean outcomes of treated and untreated:[8]

$$\begin{aligned} ATT_t &= E(Y^1_{i,t} | D_i = 1) - E(Y^0_{i,t} | D_i = 1) & (2.2) \\ &= E(Y_{i,t} | D_i = 1) - E(Y_{i,t} | D_i = 0) & (2.3) \end{aligned}$$

A non-negative outcome (with a point mass at zero) is often decomposed into an extensive and an intensive part as $E(Y_{i,t}) = E(Y_{i,t} | Y_{i,t} > 0) P(Y_{i,t} > 0)$. Similar to Staub (2014), the

[7] Note that the treatment indicator is not indexed with a time index, i.e. $D_i = 1$ for individuals treated between $t-1$ and t, and $D_i = 0$ for individuals not treated between $t-1$ and t.

[8] In the case of random treatment assignment, the average treatment effect on the treated is equal to the average treatment effect, defined as $ATE_t = E(Y^1_{i,t} - Y^0_{i,t})$.

difference in mean outcomes can then be rewritten as

$$\begin{align}
ATT_t =& E(Y_{i,t}|D_i = 1) - E(Y_{i,t}|D_i = 0) \tag{2.4}\\
=& E(Y_{i,t}|Y_{i,t} > 0, D_i = 1)P(Y_{i,t} > 0|D_i = 1) \tag{2.5}\\
& - E(Y_{i,t}|Y_{i,t} > 0, D_i = 0)P(Y_{i,t} > 0|D_i = 0) \tag{2.6}\\
=& \Big[P(Y_{i,t} > 0|D_i = 1) - P(Y_{i,t} > 0|D_i = 0)\Big]E(Y_{i,t}|Y_{i,t} > 0, D_i = 1) \tag{2.7}\\
& + \Big[E(Y_{i,t}|Y_{i,t} > 0, D_i = 1) - E(Y_{i,t}|Y_{i,t} > 0, D_i = 0)\Big]P(Y_{i,t} > 0|D_i = 0)\,. \tag{2.8}
\end{align}$$

The terms in (2.7) represent the extensive margin effect, the terms in (2.8) the intensive margin effect. Under random treatment, the terms in (2.7) and (2.8) can be rewritten as

$$\begin{align}
ATT_t =& \Big[P(Y^1_{i,t} > 0) - P(Y^0_{i,t} > 0)\Big]E(Y^1_{i,t}|Y^1_{i,t} > 0) \tag{2.9}\\
& + \Big[E(Y^1_{i,t}|Y^1_{i,t} > 0) - E(Y^0_{i,t}|Y^0_{i,t} > 0)\Big]P(Y^0_{i,t} > 0)\,. \tag{2.10}
\end{align}$$

The difference in (2.9) is a causal comparison and captures the causal effect of treatment on the probability of having a positive outcome. However, the difference in (2.10) does generally not have a causal interpretation, because we compare two possibly different subgroups of the population. The subgroup with a positive outcome in case of treatment ($Y^1_{i,t} > 0$) and the subgroup with a positive outcome in case of no treatment ($Y^0_{i,t} > 0$). Hence, conditioning on positive outcomes induces a selection problem. As a result, the difference in mean outcomes of treated and untreated - conditional on positive outcomes - does not identify the causal intensive margin effect (without additional assumptions, see Appendix 2.A). In the next section, we use a decomposition in which both the extensive and the intensive part have a causal interpretation.

2.2.3 Causal Decomposition

Following Lee (2012) and Staub (2014), we define four exhaustive and mutually exclusive subgroups based on the joint distribution of potential outcomes in period t:

Table 2.1: Subgroups Based on the Joint Distribution of Potential Outcomes in Period t

	$Y^0_{i,t} = 0$	$Y^0_{i,t} > 0$
$Y^1_{i,t} = 0$	Nonparticipants	Switchers 2
$Y^1_{i,t} > 0$	Switchers 1	Participants

Based on this definition, we decompose the average treatment effect on the treated (ATT) at time t as follows:

$$
\begin{aligned}
ATT_t &= E(Y^1_{i,t} - Y^0_{i,t}|D_i = 1) && (2.11)\\
&= E(Y^1_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} = 0, D_i = 1)P(Y^1_{i,t} > 0, Y^0_{i,t} = 0|D_i = 1) && (2.12)\\
&- E(Y^0_{i,t}|Y^1_{i,t} = 0, Y^0_{i,t} > 0, D_i = 1)P(Y^1_{i,t} = 0, Y^0_{i,t} > 0|D_i = 1) && (2.13)\\
&+ E(Y^1_{i,t} - Y^0_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} > 0, D_i = 1)P(Y^1_{i,t} > 0, Y^0_{i,t} > 0|D_i = 1) && (2.14)
\end{aligned}
$$

The terms in (2.12) and (2.13) represent the weighted causal extensive margin effect. The term in (2.12) describes the effect of treatment on the outcome of individuals with positive outcome in case of treatment and zero outcome in case of no treatment (Switchers 1), weighted by the fraction of Switchers 1. The term in (2.13) describes the effect of treatment on the outcome of individuals with zero outcome in case of treatment and positive outcome in case of no treatment (Switchers 2), weighted by the fraction of Switchers 2. The contribution of individuals with zero outcome in the cases of treatment and no treatment (Nonparticipants) is zero and therefore dropped.

The term in (2.14) represents the weighted causal intensive margin effect. It captures the effect of treatment on the outcome of individuals having a positive outcome irrespective of treatment status (Participants), weighted by the fraction of Participants.

In this decomposition, both the extensive margin effect and the intensive margin effect have a causal interpretation.

2.3 Identification

In this chapter we are interested in the intensive margin effect. Hence, we focus on the first term in (2.14). In this section, we discuss identification of the intensive margin average treatment effect on the treated (IMATT),

$$
\begin{aligned}
IMATT_t &= E(Y^1_{i,t} - Y^0_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} > 0, D_i = 1) && (2.15)\\
&= E\left[\underbrace{E(Y^1_{i,t} - Y^0_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} > 0, D_i = 1, X_i = x)}_{\gamma_t(x)}\middle| Y^1_{i,t} > 0, Y^0_{i,t} > 0, D_i = 1\right]. && (2.16)
\end{aligned}
$$

We will first derive sufficient conditions under which $\gamma_t(x)$, i.e. the conditional-on-X version of the intensive margin average treatment effect on the treated, is identified. In a second step, we state conditions under which the conditional-on-X version can be aggregated to $E(Y^1_{i,t} - Y^0_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} > 0, D_i = 1)$.

2.3.1 Difference-in-Difference on Positive Outcomes

Difference-in-difference on positive outcomes is given by the difference of the time differences between treated and untreated observations

$$\begin{aligned}\gamma_t^{DiD}(x) =& E(Y_{i,t} - Y_{i,t-1}|Y_{i,t} > 0, Y_{i,t-1} > 0, D_i = 1, X_i = x) \\ & - E(Y_{i,t} - Y_{i,t-1}|Y_{i,t} > 0, Y_{i,t-1} > 0, D_i = 0, X_i = x).\end{aligned} \tag{2.17}$$

The following sufficient conditions identify the intensive margin average treatment effect on the treated.

Proposition 1 (Identification Difference-in-Difference on Positive Outcomes) *Sufficient conditions to identify the intensive margin average treatment effect on the treated using difference-in-difference on positive outcomes are*

1. *stable unit treatment value assumption (SUTVA),*
2. *no pre-treatment effect,*
3. *common trend in positive outcomes,*
4. *no effect of treatment on covariates,*
5. *overlap,*
6. *treatment monotonicity at the extensive margin, and*
7. *time monotonicity at the extensive margin.*

Assumptions 1-5 are also required in similar form in standard difference-in-difference. Assumptions 6 and 7 are specific to difference-in-difference on positive outcomes. These assumptions are additionally required to eliminate the selection problem arising from conditioning on individuals with positive outcomes. In the following we describe the assumptions in more detail.

Assumption 1 (SUTVA) *The stable unit treatment value assumption is given by*

$$\begin{aligned}Y_{i,t} &= (1 - D_i)Y_{i,t}^0 + D_i Y_{i,t}^1 \quad \forall i, \quad \textit{and} \\ Y_{i,t-1} &= (1 - D_i)Y_{i,t-1}^0 + D_i Y_{i,t-1}^1 \quad \forall i,\end{aligned}$$

where $D_i \in \{0, 1\}$ denotes treatment status.

The *SUTVA* assumption ensures that we actually observe the potential outcomes in the treatment and control groups. The *SUTVA* assumption implies that the observed outcome of individual i only depends on the potential outcomes and the treatment status D_i, but not on the treatment status D_j of any other individual j. Thus, *SUTVA* rules out general equilibrium effects and spill-over effects.

Assumption 2 (No pre-treatment effect) *The no pre-treatment effect assumption is given by*

$$E(Y^1_{i,t-1} - Y^0_{i,t-1} | Y_{i,t} > 0, Y_{i,t-1} > 0, D_i = 1, X_i = x) = 0 \textit{ for all } x \textit{ in the support of } X_i.$$

The *no pre-treatment effect* assumption requires that the treatment effect in the pre-treatment period is zero. Hence in expectation, individuals do not change their behavior in period $t-1$ because they will be treated between period $t-1$ and t.[9]

Assumption 3 (Common trend in positive outcomes) *The common trend in positive outcomes assumption is given by*

$$\begin{aligned} &E(Y^0_{i,t} - Y^0_{i,t-1} | Y_{i,t} > 0, Y_{i,t-1} > 0, D_i = 1, X_i = x) \\ &= E(Y^0_{i,t} - Y^0_{i,t-1} | Y_{i,t} > 0, Y_{i,t-1} > 0, D_i = 0, X_i = x) \textit{ for all } x \textit{ in the support of } X_i. \end{aligned}$$

The *common trend in positive outcomes* assumption represents the key assumption for identification. The *common trend in positive outcomes* assumption is closely related to the standard common trend assumption, except that we require the common trend to hold in the subsample of individuals with a positive outcome in period t and $t-1$.[10] The *common trend in positive outcomes* assumption requires that the treated and the control group would experience the same time trend in case of no treatment.[11] As Lechner (2010) points out, the common trend assumption can be rewritten as a ”constant bias” assumption. That is, the bias arising from unobserved confounders is assumed to be constant over time.

Assumption 4 (No effect of treatment on covariates) *The no effect of treatment on covariates assumption is given by*

$$X^1_i = X^0_i = X_i \quad \forall i \; .$$

The *no effect of treatment on covariates* assumption is required to ensure that conditioning on X does not condition away parts of the causal effect we are interested in, or introduce a collider bias.[12]

Assumption 5 (Overlap) *The overlap assumption is given by*

$$P(D_i = 1 | Y_{i,t} > 0, Y_{i,t-1} > 0, X_i = x) < 1 \textit{ for all } x \textit{ in the support of } X_i.$$

[9] Using *SUTVA*, the *no pre-treatment effect* assumption can be rewritten to $E(Y^1_{i,t-1} - Y^0_{i,t-1} | X_i = x, Y^1_{i,t} > 0, Y^1_{i,t-1} > 0, D_i = 1) = 0$, which is the version used in the proof of identification.

[10] In the standard difference-in-differences, the common trend assumption is given by $E(Y^0_{i,t} - Y^0_{i,t-1} | D_i = 1) = E(Y^0_{i,t} - Y^0_{i,t-1} | D_i = 0)$.

[11] Using *SUTVA*, the *common trend in positive outcomes* assumption can be rewritten to $E(Y^0_{i,t} - Y^0_{i,t-1} | Y^1_{i,t} > 0, Y^1_{i,t-1} > 0, D_i = 1, X_i = x) = E(Y^0_{i,t} - Y^0_{i,t-1} | Y^0_{i,t} > 0, Y^0_{i,t-1} > 0, D_i = 0, X_i = x)$, which is the version used in the proof of identification.

[12] See footnote 10 in Chapter 1.

The *overlap* assumption requires that for all x in the support of X_i, there exist not only treated individuals in the subsample with positive outcomes in period t and $t-1$.

Assumption 6 (Treatment monotonicity at the extensive margin) *The treatment monotonicity at the extensive margin assumption is given by*

$$\begin{aligned} Y^1_{i,t} > 0 \;\; &\Rightarrow \;\; Y^0_{i,t} > 0 \quad \forall i, \;\; \textit{or} \\ Y^0_{i,t} > 0 \;\; &\Rightarrow \;\; Y^1_{i,t} > 0 \quad \forall i. \end{aligned}$$

The assumption of *treatment monotonicity at the extensive margin* states that a positive outcome in case of treatment implies a positive outcome in case of no treatment or vice versa. Therefore, the treatment response is monotone with respect to the extensive margin decision. Note that this assumption only restricts the sign of the extensive margin effect. Thus, given the potential outcome in case of treatment is positive, the potential outcome in case of no treatment is allowed to be higher or lower than the potential outcome in case of treatment. The assumption only requires that the potential outcome in case of no treatment is positive.

Assumption 7 (Time monotonicity at the extensive margin) *The time monotonicity at the extensive margin assumption is given by*

$$\begin{aligned} Y^0_{i,t} > 0 \;\; &\Rightarrow \;\; Y^0_{i,t-1} > 0 \quad \forall i, \;\; \textit{and} \\ Y^1_{i,t} > 0 \;\; &\Rightarrow \;\; Y^1_{i,t-1} > 0 \quad \forall i. \end{aligned}$$

The assumption of *time monotonicity at the extensive margin* states that a positive outcome in period t implies a positive outcome in period $t-1$, both in case of treatment and no treatment. Thus, we assume that there are no individuals with a positive outcome in period t who have a zero outcome in period $t-1$. This assumption again only restricts the sign of the extensive margin effect. Given the potential outcome in period t is positive, the potential outcome in period $t-1$ is allowed to be higher or lower than the potential outcome in period t.

Proof Assuming *SUTVA*, equation (2.18) can be rewritten to

$$\begin{aligned} \gamma^{DiD}_t(x) =& E(Y^1_{i,t} - Y^1_{i,t-1} | Y^1_{i,t} > 0, Y^1_{i,t-1} > 0, D_i = 1, X_i = x) \\ & - E(Y^0_{i,t} - Y^0_{i,t-1} | Y^0_{i,t} > 0, Y^0_{i,t-1} > 0, D_i = 0, X_i = x)\,. \end{aligned} \tag{2.18}$$

Adding and subtracting $E(Y^0_{i,t-1} | Y^1_{i,t} > 0, Y^1_{i,t-1} > 0, D_i = 1, X_i = x)$ and

$E(Y^0_{i,t}|Y^1_{i,t} > 0, Y^1_{i,t-1} > 0, D_i = 1, X_i = x)$ to equation (2.18) and rearranging yields

$$\begin{aligned}
\gamma^{DiD}_t(x) = {} & E(Y^1_{i,t} - Y^0_{i,t}|Y^1_{i,t} > 0, Y^1_{i,t-1} > 0, D_i = 1, X_i = x) && (2.19)\\
& + E(Y^0_{i,t-1} - Y^1_{i,t-1}|Y^1_{i,t} > 0, Y^1_{i,t-1} > 0, D_i = 1, X_i = x) && (2.20)\\
& + E(Y^0_{i,t} - Y^0_{i,t-1}|Y^1_{i,t} > 0, Y^1_{i,t-1} > 0, D_i = 1, X_i = x) && (2.21)\\
& + E(Y^0_{i,t-1} - Y^0_{i,t}|Y^0_{i,t} > 0, Y^0_{i,t-1} > 0, D_i = 0, X_i = x)\,. && (2.22)
\end{aligned}$$

Assuming *SUTVA* and *common trend in positive outcomes*, the sum of the two terms in (2.21) and (2.22) equals 0. Moreover, under the *no pre-treatment effect* assumption, the sum of the term in (2.20) is equal to zero. Assuming *time* and *treatment monotonicity at the extensive margin*, the term in (2.19) can be rewritten to $E(Y^1_{i,t} - Y^0_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} > 0, D_i = 1, X_i = x)$.[13] This identifies the conditional-on-X version of the intensive margin average treatment effect on the treated.

The *overlap* assumption then guarantees that all conditional-on-X versions of the IMATT exist. Based on (2.15), the conditional-on-X versions are aggregated with respect to the distribution of X in the subsample with $Y^1_{i,t} > 0$, $Y^0_{i,t} > 0$, and $D_i = 1$. Assuming *time* and *treatment monotonicity at the extensive margin*, this subsample is identical to the subsample with $Y^1_{i,t} > 0$, $Y^1_{i,t-1} > 0$, and $D_i = 1$.[14] By *SUTVA*, this subsample is again identical to the subsample with $Y_{i,t} > 0$, $Y_{i,t-1} > 0$, and $D_i = 1$, which is an observed subsample.

An obvious alternative to difference-in-difference is the simple difference estimator, given by

$$\gamma^D_t(x) = E(Y_{i,t}|Y_{i,t} > 0, D_i = 1, X_i = x) - E(Y_{i,t}|Y_{i,t} > 0, D_i = 0, X_i = x)\,. \qquad (2.23)$$

In Appendix 2.A, we state sufficient conditions under which the simple difference estimator identifies the conditional-on-X intensive margin average treatment effect on the treated.

2.3.2 Special Case: Random Treatment

When treatment is randomly assigned, we do not need to condition on X to identify the causal effect. If we do not condition on X, we do not require the *overlap* and the *no effect of treatment on covariates* assumptions. The other assumptions are still required to identify the intensive margin average treatment effect on the treated. A further implication of random treatment is that we can also identify the ATE, since the ATT equals the ATE under random treatment.

[13] By the *time monotonicity at the extensive margin*, the conditioning set can be reduced to $E(Y^1_{i,t} - Y^0_{i,t}|Y^1_{i,t} > 0, D_i = 1, X_i = x)$. Using the *treatment monotonicity at the extensive margin*, the conditioning set can again be expanded to $E(Y^1_{i,t} - Y^0_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} > 0, D_i = 1, X_i = x)$.

[14] Analogous to the rewriting of (2.19), see footnote 13.

2.4 Estimation and Inference

Difference-in-difference estimation requires estimating different conditional expectations. Here we adopt a split sample approach. Let $\Delta Y_{i,t} = Y_{i,t} - Y_{i,t-1}$. For the difference-in-difference on positive outcomes estimator, we first estimate

$$m_1(x) = E(\Delta Y_{i,t} | Y_{i,t} > 0, Y_{i,t-1} > 0, D_i = 1, X_i = x)\ , \text{ and} \tag{2.24}$$

$$m_0(x) = E(\Delta Y_{i,t} | Y_{i,t} > 0, Y_{i,t-1} > 0, D_i = 0, X_i = x)\ , \tag{2.25}$$

using ordinary least squares. That is, we regress $\Delta Y_{i,t}$ on X_i separately in the treated sample and in the untreated sample, restricted to the observations with positive outcomes in period t and $t-1$. Since we condition the sample on observations with a positive outcome in period t and $t-1$, we require panel data.[15] Using the fitted functions $\widehat{m_1}(x)$ and $\widehat{m_0}(x)$, we then calculate fitted values $\widehat{m_1}(X_i)$ and $\widehat{m_0}(X_i)$. The intensive margin average treatment effect on the treated is then estimated as

$$\widehat{IMATT}_t^{DID} = \frac{1}{N_T} \sum_{\substack{i:Y_{i,t}>0,\\ Y_{i,t-1}>0,\\ D_i=1}} \left[\widehat{m_1}(X_i) - \widehat{m_0}(X_i)\right]\ , \tag{2.26}$$

where N_T is the number of treated observations with positive outcome in period t and $t-1$.[16]

To conduct inference, we employ a nonparametric quantile bootstrap (Efron & Tibshirani, 1993). From the sample of observations with positive outcomes in period t and $t-1$, we repeatedly draw a bootstrap sample of the same sample size. In the bootstrap sample, we estimate the IMATT as described above. This gives a distribution of bootstrap estimated IMATTs: $\widehat{IMATT}_t^1, \ldots, \widehat{IMATT}_t^B$, where B is the number of bootstrap replications. We then construct a bootstrap estimated confidence interval as

$$[q^*_{\alpha/2}, q^*_{1-\alpha/2}]\ , \tag{2.27}$$

where $q^*_{1-\alpha/2}$ is the $(1-\alpha/2)$-percentile of the distribution of bootstrap estimated IMATTs.

2.5 Empirical Application: Causal Effect of Reaching the Full Retirement Age on Working Hours

We apply the difference-in-difference methodology to estimate the causal intensive margin effect of reaching the full retirement age on working hours of women. We exploit a pension reform in Switzerland taking place in 2004. In this pension reform, the full retirement age

[15] This marks a difference to the standard difference-in-difference estimator, for which it is also possible to use repeated cross-sections.

[16] Alternatively, one could replace $\widehat{m_1}(X_i)$ in (2.26) with $\Delta Y_{i,t}$, such that it would not be necessary to estimate $\widehat{m_1}(x)$.

(FRA) of women was increased from age 63 to age 64.[17] This implies that women with year of birth 1941 or earlier reach FRA at age 63, while women with year of birth 1942 or later reach FRA at age 64. We use data from the Swiss Labor Force Survey (SLFS) from 2002-2009. The outcome of interest is working hours, denoted by $Y_{i,t}$.[18] We restrict the sample to women aged 63. Therefore, treatment $D_i = 1$ for women who have reached FRA (year of birth 1941 or earlier), and $D_i = 0$ for women who have not reached FRA (year of birth 1942 or later). Since the reform affects individuals only based on their year of birth, we assume that confounding is not a problem. For illustrative purposes, we include a categorical education variable and a dummy for being a Swiss citizen. We consider two estimation samples. The first sample consists of women with year of birth 1941 or 1942. That is, women exactly at the threshold of the pension reform. This sample is cleaner in terms of identification, but the small number of observations decreases the power. For this reason we consider a second estimation sample, which includes women with year of birth 1939 to 1946. This sample includes more observations, but might pose a threat to identification if there is a time trend in working hours.

2.5.1 Discussion: Assumptions

With the exception of *time monotonicity at the extensive margin* and *overlap*, we cannot directly test the identifying assumptions. Instead, we propose alternative tests that can be used to motivate the identifying assumptions and discuss whether the assumptions are likely to be fulfilled in the context of our empirical application.[19]

SUTVA: This assumption cannot be tested. There is evidence for spillover effect within couples (see Chapter 3). That is, the labor supply of one individual depends on whether the spouse has reached FRA. We are aware that this might pose a threat for identification, but assume that the spillover effects are negligible.

No pre-treatment effect: This assumption rules out that people adjust their working hours in anticipation of reaching FRA in the next period. We cannot directly test this assumption. We motivate the assumption by comparing the mean working hours in period $t-1$, conditional on having positive working hours in period t and $t-1$. The mean in the control group is 23.8 hours, in the treatment group 24.6 hours. A simple Welch two sample t-test does not reject the null hypothesis of equal means (p-value: 0.54). This indicates that the assumption is fulfilled. Moreover, if there is a pre-treatment effect, this effect will likely have the same sign as the treatment effect. As a result, the estimated treatment effect could be interpreted as a lower bound.

Common trend in positive outcomes: This assumption requires that the treatment group

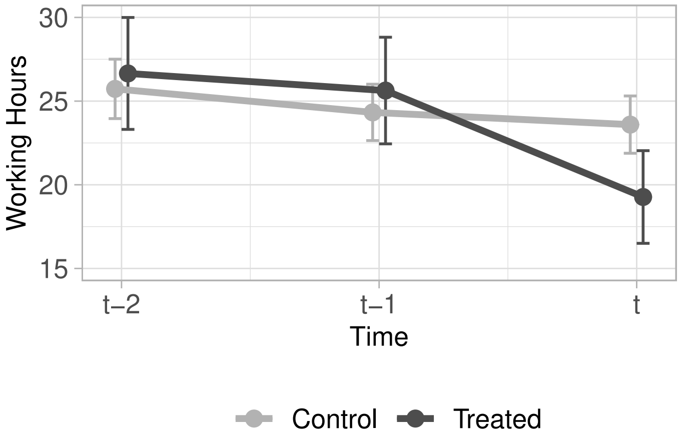
[17] The FRA denotes the age at which a first pillar pension can be claimed without a deduction. There was a second pension reform in 2001, increasing the FRA of women from 62 to 63, which is not considered in this application.

[18] We use contracted hours for wage employed and usual hours for self employed as our measure of working hours.

[19] The (alternative) tests are always based on the larger estimation sample, i.e. the sample including women with year of birth 1939 to 1946.

would experience the same time trend in working hours in case of no treatment as the control group. We cannot directly test this assumption, but we motivate the assumption by examining the pre-treatment trends of the control and treatment group. In Figure 2.1, we plot the mean working hours of women with positive hours in period t, $t-1$ and $t-2$. We observe that the trends between period $t-2$ and $t-1$ are roughly parallel, indicating that the assumption is fulfilled.

Figure 2.1: Assessment of Common Trend Assumption



Note: Dots indicate the mean working hours of women aged 63 with positive working hours in period t, t-1 and t-2. Bars indicate the 95% normal approximation confidence interval for the mean. Year of birth between 1939 and 1946. Treated is the group which reaches FRA in period t (women with year of birth 1941 or earlier), Control is the group which does not reach FRA in period t (women with year of birth 1942 or later).

No effect of treatment on covariates: In the empirical application, we include a categorical education variable and a dummy for being a Swiss citizen. It seems unlikely that reaching FRA has an effect on these variables.

Overlap: This assumption can be tested. In each covariate cell, we calculate the fraction of treated observations. The results are presented in Table 2.3 in Appendix 2.B. We observe that there is no covariate cell with only treated observations. Therefore, the assumption is fulfilled.

Treatment monotonicity at the extensive margin: This assumption rules out that people start to work because they reach FRA. There are indeed incentives to take up a job after reaching FRA. For example, part of the earnings are exempted from social security contributions. This increases the net wage. On the other hand, it seems plausible that reaching FRA either has no effect or drives people out of the labor market.

Time monotonicity at the extensive margin: This assumption can be tested. In the treated and control subsample, we calculate the fraction of individuals with positive working hours in period t, conditional on not working in period $t-1$. In the sample of women with year of birth 1939-46, 5% of the treated and 7.4% of the control sample state that they returned to work after having not worked in the period before. This poses a threat to our identification. However, the overall pattern in the age range 60-70 is that people rather leave the labor force as they become older.

2.5.2 Estimation Results

The results of the difference-in-difference estimation are presented in Table 2.2. In the estimation sample including only women with year of birth 1941-42 (left column), the estimated intensive margin average treatment effect on the treated is -5.00. That is, reaching FRA reduces the working hours of women with positive working hours irrespective of whether they have reached FRA or not on average by 5 hours. The bootstrap estimated 95% confidence interval does not include zero, indicating that the effect is statistically significantly different from zero. In the sample including women with year of birth 1939-46 (right column), the estimated intensive margin ATT is -4.22. Again, the bootstrap estimated 95% confidence interval does not include zero. This analysis provides evidence that women react at the intensive margin when reaching FRA.

Table 2.2: Results Difference-in-Difference on Positive Outcomes

	Outcome: Working Hours	
	Sample 1941-42	Sample 1939-46
IMATT: FRA reached	-5.00	-4.22
95% C.I.	[-8.47, -1.58]	[-6.58, -1.95]
Obs. (treat/control)	63/87	156/405

Note: Confidence interval based on 1000 bootstrap replications. Sample includes women aged 63 with positive working hours in period t and t-1. Women with year of birth 1942 or later have FRA 64 (Control), women with year of birth 1941 or earlier have FRA 63 (Treated). The left column presents the results for women with year of the birth 1941-1942, and the right column those for women with year of birth 1939-1946.

2.6 Conclusion

This chapter extends the literature on the identification and estimation of causal intensive margin effects. The intensive margin effect is of interest when subeffects are masked by the total effect. This is the case, for example, when the extensive and intensive margin effect have different signs. We use difference-in-difference methods to identify the causal intensive margin effect. We derive sufficient conditions under which the difference-in-difference estimator on positive outcomes identifies the causal intensive margin effect. We demonstrate that the difference-in-difference estimator on positive outcomes, compared to the standard difference-in-difference estimator, additionally requires *time* and *treatment monotonicity at* *the extensive margin*. Our proposed difference-in-difference estimator represents an alternative to models for outcomes with corner solutions or selection models. We apply the methodology to estimate the causal intensive margin effect of reaching the full retirement age on working hours.

2.7 Bibliography

Angrist, J. D. (2001). Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors. *Journal of Business & Economic Statistics*, *19*(1), 2–28.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, *91*, 444–455.

Cragg, J. (1971). Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica*, *39*(5), 829–844.

Deuchert, E., Huber, M., & Schelker, M. (2019). Direct and Indirect Effects Based on Difference-in-Differences With an Application to Political Preferences Following the Vietnam Draft Lottery. *Journal of Business and Economic Statistics*, *37*(4), 710–720.

Duan, N., Manning, W. G., Morris, C. N., & Newhouse, J. P. (1983). A Comparison of Alternative Models for the Demand for Medical Care. *Journal of Business & Economic Statistics*, *1*(2), 115–126.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall/CRC.

Frangakis, C. E., & Rubin, D. B. (2002). Principal Stratification in Causal Inference. *Biometrics*, *58*(1), 21–29.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*(1), 153–161.

Hersche, M. (2018). Theoretical and Empirical Essays on Labor Supply of the Elderly. (Diss. ETH No. 25377).

Hersche, M., & Moor, E. (2018). Identification of Causal Intensive Margin Effects by Difference-in-Difference Methods. *CER-ETH Economics Working Paper Series*(18/302).

Lechner, M. (2010). The Estimation of Causal Effects by Difference-in-Difference Methods. *Foundations and Trends in Econometrics*, *4*(3), 165–224.

Lee, M.-j. (2012). Treatment Effects in Sample Selection Models and Their Nonparametric Estimation. *Journal of Econometrics*, *167*(2), 317–329.

Lee, M.-j. (2017). Extensive and Intensive Margin Effects in Sample Selection Models : Racial Effects on Wages. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *180*, 817–839.

McDonald, J. F., & Moffitt, R. A. (1980). The Uses of Tobit Analysis. *The Review of Economics and Statistics*, *62*(2), 318–321.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Staub, K. (2014). A Causal Interpretation of Extensive and Intensive Margin Effects in Generalized Tobit Models. *The Review of Economics and Statistics*, *96*(2), 371–375.

Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, *26*(1), 24–36.

Appendix 2.A Identification Simple Difference Estimator on Positive Outcomes

The *simple difference estimator on positive outcomes* is given by

$$\gamma_t^D(x) = E(Y_{i,t}|Y_{i,t} > 0, D_i = 1, X_i = x) - E(Y_{i,t}|Y_{i,t} > 0, D_i = 0, X_i = x). \tag{2.28}$$

The following sufficient conditions identify the conditional-on-X intensive margin average treatment effect on the treated.

Proposition 2 (Identification simple difference estimator on positive outcomes) *Sufficient conditions to identify the causal intensive margin effect using the simple difference estimator on positive outcomes are:*

1. *SUTVA (assumption 1),*
2. *no effect of treatment on covariates (assumption 4),*
3. *overlap (assumption 5),*
4. *unconfoundedness (assumption 8), and*
5. *no Switchers (assumption 9),*

Or

5. *conditional mean independence (assumption 10).*

Assumption 8 (Unconfoundedness) *The unconfoundedness assumption is given by*

$$(Y^1_{i,t}, Y^0_{i,t}) \perp\!\!\!\perp D_i \mid X_i \; .$$

The *unconfoundedness* assumption requires that treatment is independent of the potential outcomes, conditional on covariates X_i.

Assumption 9 (No Switchers) *The assumption of no Switchers is given by*

$$Y^1_{i,t} > 0 \quad \Leftrightarrow \quad Y^0_{i,t} > 0 \quad \forall i.$$

The assumption of *no Switchers* states that the potential outcome in case of treatment is positive if and only if the potential outcome in case of no treatment is positive. It therefore excludes the possibility that individuals have a positive outcome in case of treatment and a zero outcome in case of no treatment (Switchers 1), or vice versa (Switchers 2).

Assumption 10 (Conditional mean independence) *The conditional mean independence assumption is given by*

$$E(Y^1_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} = 0, D_i = 1, X_i = x) = E(Y^1_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} > 0, D_i = 1, X_i = x), \quad \textit{and}$$
$$E(Y^0_{i,t}|Y^1_{i,t} = 0, Y^0_{i,t} > 0, D_i = 1, X_i = x) = E(Y^0_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} > 0, D_i = 1, X_i = x).$$

The assumption of *conditional mean independence* states that the expected potential outcome in case of treatment of Switchers 1 is equal to the expected potential outcome in case of treatment of Participants. Furthermore, the expected potential outcome in case of no treatment of Switchers 2 is equal to the expected potential outcome in case of no treatment of Participants.

Proof Under *SUTVA* and *unconfoundedness*, and by the law of iterated expectations, equation (2.28) can be rewritten as

$$\begin{aligned}
\gamma^D_t(x) = \Big[&pE(Y^1_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} > 0, D_i = 1, X_i = x) && (2.29)\\
&+(1-p)E(Y^1_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} = 0, D_i = 1, X_i = x)\Big] && (2.30)\\
-\Big[&qE(Y^0_{i,t}|Y^1_{i,t} > 0, Y^0_{i,t} > 0, D_i = 1, X_i = x) && (2.31)\\
&+(1-q)E(Y^0_{i,t}|Y^1_{i,t} = 0, Y^0_{i,t} > 0, D_i = 1, X_i = x)\Big],
\end{aligned}$$

where $p \equiv P(Y^0_{i,t} > 0|Y^1_{i,t} > 0, D_i = 1, X_i = x)$ and $q \equiv P(Y^1_{i,t} > 0|Y^0_{i,t} > 0, D_i = 1, X_i = x)$. This term is equal to the causal intensive margin of interest in equation (2.15) if a) $p = q = 1$ (assumption of *no Switchers*), or if b) the corresponding expectations in the brackets are identical, i.e. the expected potential outcome in case of treatment of Switchers 1 is equal to the expected potential outcome of Participants, and the expected potential outcome in case of no treatment of Switchers 2 is equal to the expected potential outcome of Participants (assumption of *conditional mean independence*).

Appendix 2.B Overlap

Table 2.3: Analysis of Overlap

Secondary Education	Higher Education	Swiss citizen	Fraction Treated
0	0	0	0.091
0	0	1	0.308
0	1	0	0.350
0	1	1	0.262
1	0	0	0.154
1	0	1	0.298

Note: This table displays the fraction of treated observations (last column) for all possible covariate cells.

Chapter 3

Labor or Leisure? Labor Supply of Older Couples and the Role of Full Retirement Age[1]

3.1 Introduction

Declining fertility rates and increasing life expectancy force many developed countries to reform their pension systems. Designing policy reforms requires a detailed understanding of the labor supply behavior of older workers. For this reason, a large body of literature has estimated labor supply responses of individuals directly affected by pension reforms (*direct effect*), see e.g. Börsch-Supan and Schnabel (1999), or Mastrobuoni (2009).

The approach focusing on the *direct effect* abstracts from the fact that a large proportion of older workers are married.[2] Several studies find that older couples coordinate their exit from the labor force, see e.g. Gustman and Steinmeier (2004) or Hospido and Zamarro (2014). As a result, changes in incentives of one member of the couple may have spillover effects on the labor supply of the spouse (*indirect effect*). In contrast to the evidence on the *direct effect*, existing studies on *indirect effects* find ambiguous results, depending on the particular country and reform under consideration.

Previous studies examine the *indirect effect* on the participation decision (*extensive margin*). Changes at the *extensive margin*, however, do not fully capture the change in total labor supply. Individuals may adjust their working hours to change their labor supply (*intensive margin*). The prevalence of gradual retirement indicates that older workers use working hours to adjust their labor supply, see Kantarci and Van Soest (2008) for a summary of evidence in Europe and the US.

[1] This chapter is joint work with Markus Hersche. An earlier version was published as working paper “Labor or Leisure? Labor Supply of Older Couples and the Role of Full Retirement Age”, *Netspar Working Paper Series, 03/2018*, see Hersche and Moor (2018). This chapter is also part of the doctoral thesis ”Theoretical and Empirical Essays on Labor Supply of the Elderly”, *Diss. ETH No. 25377*, see Hersche (2018).

[2] According to census data from the Swiss Federal Office for Statistics, 75% of men aged between 55 and 70, and 64% of women in Switzerland were married in 2010.

In this chapter, we estimate the effect of having a spouse at or above the full retirement age (FRA) on labor supply at *extensive margin* and *intensive margin* in Switzerland. FRA represents the age at which first pillar pensions can be claimed without deductions. The full retirement age is of interest in two ways. First, full retirement age in the first pillar represents the main policy instrument for the government. Knowledge on the expected spousal reaction will provide information on spillover effects of future pension reforms. Second, changes in hours and hazard rates of retirement peak at FRA. Therefore, the estimate on spousal labor supply reaction will be informative on the relationship between spousal retirement and own labor supply.

We find that the labor force participation of women decreases on average by approximately 3 percentage points in response to the spouse reaching FRA. By contrast, men do not react at the extensive margin. At the intensive margin, we find only small and mostly non-significant effects for both men and women. We argue that the response can be explained by *complementarity in leisure* and *liquidity effects*. For women, the absence of a substantial intensive margin reaction can be explained by the presence of *fixed costs of work*.

We use two sources of variation to identify the effect. First, we exploit variation in age difference within couples. Second, we use a pension reform which increased the FRA of women from 62 to 64. In our analysis, the treatment group consists of individuals with a spouse who has reached FRA. The control group consists of individuals whose spouse has not yet reached FRA. The key identifying assumption of our approach is that, after controlling for confounders, a difference in labor supply arises only from the difference in the FRA status of the spouse.

In contrast to Chapters 1 and 2, we do not explicitly use the potential outcome framework to describe causal effects. Rather, we directly model realized outcomes, and interpret model parameters as causal effects. At the extensive margin, we estimate the causal effect using a probit model. At the intensive margin, we employ a Tobit, a two-part, and a difference-in-difference estimator on positive outcomes. The approach we take in this chapter has several drawbacks. For example, since the causal effect is a fixed model parameter, this approach does not allow for treatment effect heterogeneity. Moreover, it is not possible to define the causal effect without referring to a specific model (Imbens & Wooldridge, 2009).

The analysis is based on data drawn from the Swiss Labor Force Survey for the time period between 1991 and 2009. In contrast to administrative Social Security data sets, this data set provides information on working hours. Furthermore, it includes rich information on labor supply and a large set of sociodemographic variables relating to the interviewed individual and the spouse. In this survey, individuals are interviewed every year for up to five consecutive years. We exploit the panel structure of the survey to estimate the intensive margin effect with the difference-in-difference estimator on positive outcomes.

This chapter relates to the literature on labor supply of older couples pioneered by Hurd (1990), Zweimüller, Winter-Ebmer, and Falkinger (1996), and Blau (1998). In this literature, two effects are frequently studied. First, the literature studies the effect of spouse B retiring on the labor supply of spouse A. Second, studies examine the effect of spouse B reaching FRA on the labor supply of spouse A. In this chapter, we contribute to the latter literature. This chapter is closely related to contributions by Cribb, Emmerson, and Tetlow (2013), Selin (2017), Stancanelli (2017) and Lalive and Parrotta (2017). These contributions use social security reforms or pension legislation to identify the causal effect of the spouse reaching FRA. Cribb et al. (2013) analyze the spillover effects of an increase in female FRA in the UK. They find positive spillover effects on the labor supply of men. Selin (2017) exploits an occupational pension reform in Sweden which primarily affected female workers. He finds no evidence for a response of men married to affected women. Stancanelli (2017) analyzes the 1994 French pension reform which increased the contribution length required to receive the maximum pension. She finds that the reform decreased the retirement probability of men whose spouse was affected by the reform by approximately 1 percentage point. By contrast, she finds no evidence that women react when their husband was affected by the reform. Lalive and Parrotta (2017) use Swiss census data from 1990 and 2000 to estimate the effect of pension eligibility on labor supply in a couple. They find that labor force participation of women drops by 2 to 3 percentage points when their spouses reach pension eligibility age. They find no significant effects for men. In contrast to the aforementioned contributions examining the effect on the participation decision, we additionally investigate the causal effect of the spouse reaching FRA on hours worked (intensive margin).

The remainder of this chapter is organized as follows. Section 3.2 describes the Swiss pension and tax system, outlining the financial incentives faced by older couples. Section 3.3 outlines mechanisms that can explain the labor supply reaction when the spouse reaches FRA. Sections 3.4 and 3.5 describe the data and general labor supply patterns. Section 3.6 presents the empirical approach. Results are presented in Section 3.7 and discussed in Section 3.8. The last section concludes.

3.2 Incentives of Older Couples in Switzerland

In this section, we describe the financial incentives faced by older workers in Switzerland. In particular, we focus on the description of incentives for older married couples.

3.2.1 Pension System

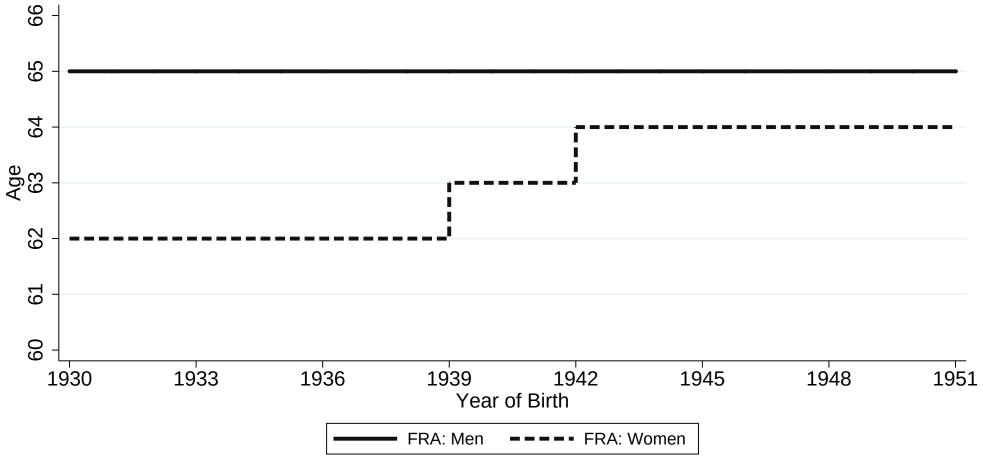
The Swiss pension system consists of three pillars. The old age and survivor insurance (OASI) represents the first pillar. The OASI is a pay-as-you-go insurance, with a strong redistributive motive. The OASI is financed by payroll taxes and government transfers. Its main purpose is to cover basic living costs. Individual pension entitlements are a function of contribution years and average earnings.[3] Individuals who contributed each year from age 20 to FRA are entitled to a full pension. The FRA is defined as the age at which a first pillar pension can be claimed without deductions. For each missing contribution year, benefits are

[3] Average earnings depend on the lifetime earnings, as well as on educational and care credits.

reduced by at least 2.3%. Depending on average earnings, the monthly full pension in 2005 ranged from a minimum of 1075 CHF to a maximum of 2150 CHF.[4] The sum of the two individual pensions within a couple is capped at 150% of a maximum individual pension. It is not possible to borrow against future first pillar entitlements.

Individuals reaching FRA can claim pensions and continue working. No earnings test applies for pensions from the first pillar. In addition, workers are able to postpone claiming pensions from the first pillar. The pension increases from 5.2% for a one-year delay to 31.5% for a five-year delay. Individuals working past FRA continue paying payroll taxes, with an allowance of 16'800 CHF. These contributions do not increase future pension entitlements.

Figure 3.1: Full Retirement Age of Men and Women by Year of Birth



In the time period under consideration, the OASI was reformed once in 1997. Most prominently, the FRA for women was increased in two steps from 62 to 64. In 2001, the FRA was increased from 62 to 63. In 2004, the FRA was increased in a second step from 63 to 64. Furthermore, the possibility of claiming early retirement benefits from the first pillar was introduced. Figure 3.1 depicts the evolution of the FRA for men and women by year of birth.

The second pillar is an occupational pension scheme. The objective is to ensure the continuation of the living standard held prior to retirement. Contributions to the occupational pension system are age-dependent and compulsory for wage employed above a given threshold.[5] In general, the regulated retirement age of occupational pension schemes coincides with the FRA of the first pillar.[6] However, pension funds are free to set more generous regulations. Upon reaching the regulated retirement age, the retiree can choose between a lifelong monthly annuity, a lump-sum transfer of the accumulated capital, or a combination of both. The share of married men insured in the second pillar amounts to approximately 70%. By contrast, approximately 40% of women are insured in the second pillar.[7] This can

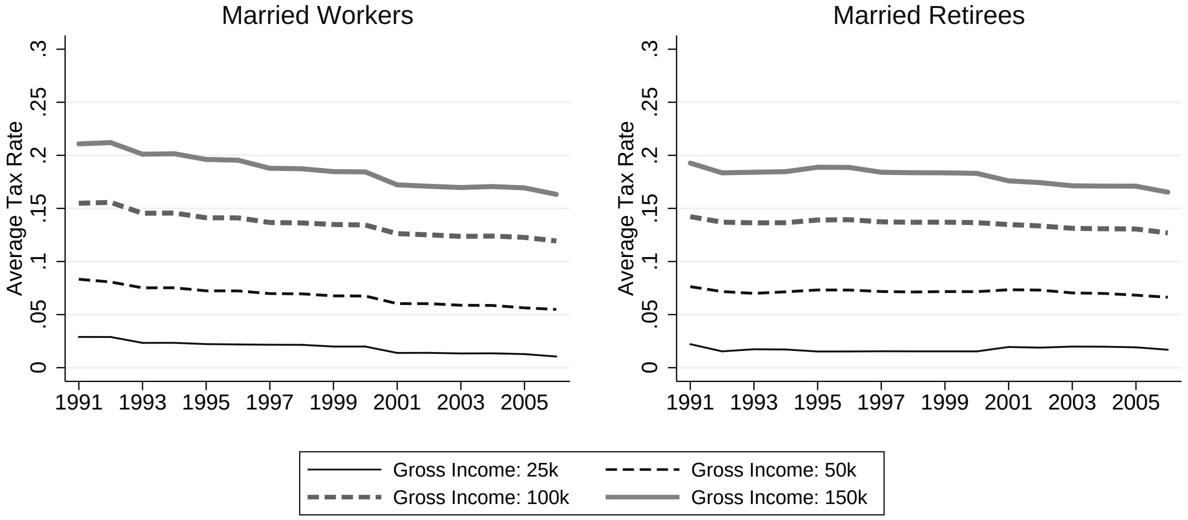
[4] As a comparison, the monthly median labor income in Switzerland amounted to 5250 CHF (2005).

[5] Only the amount exceeding the threshold is insured. Threshold 1991: yearly earnings 19'200 CHF, 2009: yearly earnings 20'520 CHF.

[6] The regulated retirement age of occupational pension schemes is the age at which occupational pensions can be claimed without deductions.

[7] See Figure 3.9 in Appendix 3.D.

Figure 3.2: Average Tax Rates of Married Couples by Gross Labor Income



Note: Average combined tax rates (federal, cantonal, community level) in cantonal capitals by year and gross labor/pension income for a married couple. Standard deductions without verification requirements for wage earners are applied. For retirees, standard deductions without verification requirements for the case where both individuals have reached FRA are applied. Data source: Own calculations, tax rates from Federal Tax Administration.

be explained by the fact that women have lower labor force participation rates and are more likely to work part-time. The third pillar consists of voluntary, tax-favored savings.

3.2.2 Income Taxes

In contrast to the majority of OECD countries, Switzerland has a system where the income of married couples is taxed based on the concept of family taxation. Income from both partners is aggregated and taxed as a single unit. Tax rates for unmarried individuals and married couples are different. Income is taxed at community, cantonal, and federal level. Cantons have fiscal sovereignty, and are therefore free to set tax rates and establish deductions.

The income tax schedule in Switzerland is progressive by law. Gross labor income is subject to a set of deductions. The left graph in Figure 3.2 displays the average tax rates for a given *gross labor income* before social security deductions for the time period 1991-2006. The average tax rates decreased slightly for all income brackets in the time period under consideration. Pension income is not exempted from income taxation. First, second, and third pillar pensions are generally taxed at the same rate as labor income. Similar to labor income, retirees are eligible for a set of tax deductions.

The right graph in Figure 3.2 displays the average tax rates for a given *gross pension income*.[8] The average tax rates for retirees decreased moderately in the time period under consideration. The differences over time between the average tax rates are however small.[9]

Occupational pension funds in Switzerland aim at a replacement rate of 50%-60%. Com-

[8] Before social security deductions.

[9] In 2007, the Federal tax administration changed the statistical procedure of reporting average tax rates for retirees. Therefore, we do not report tax rates after 2007.

bined with pension income from first pillar and third pillar, this results in a total replacement rate of 70%-80%, see Bütler (2009) for a discussion. Tax rates therefore generally decrease when an individual reaches FRA.

3.3 Mechanisms

Table 3.1 presents a non-exclusive list of mechanisms explaining a change in labor supply of individual A when spouse B reaches FRA. The labor supply reaction of A depends on whether the spouse B reduces labor supply when reaching FRA. Therefore, we divide the analysis into two parts: the case in which B reduces labor supply (left column), and the case in which B does not reduce labor supply (right column). The latter case includes the situation where B retired before FRA. Furthermore, it includes the case where B continues working without changing working hours. Although B does not reduce labor supply, a full pension can be claimed when reaching FRA. The resulting change in income can have an impact on A's labor supply.

Table 3.1: Mechanisms and Expected Sign of Labor Supply Reaction to Spouse Reaching FRA

Mechanism	Expected sign of labor supply reaction of individual A when spouse B reaches FRA and …	
	B reduces labor supply	B does not reduce labor supply
1. Complementarity in leisure	Negative	Zero
2. Liquidity Effect	Positive	Negative
3. Joint Taxation	Positive	Negative
4. Housework	Positive	Zero

First, we expect couples to enjoy their leisure time more when it is spent together. Previous studies find evidence that *complementarities in leisure* are a strong driver of labor market decisions in older couples, see e.g. Coile (2004) or Banks, Blundell, and Casanova Rivas (2010). Therefore, if B reduces labor supply, *complementarities in leisure* lead, ceteris paribus (cet. par.), to a decrease in labor supply of individual A. If B does not reduce labor supply, the expected effect is zero.

Second, *liquidity effects* can occur as soon as spouse B reaches FRA and claims a pension.[10] In the case where B reaches FRA, claims a pension, and reduces labor supply, there is a drop in household income, since replacement rates are below one.[11] Hence A, cet. par.,

[10] According to Bundesamt für Sozialversicherungen (2009), the proportion not claiming a first pillar pension at FRA amounted to less than 1% (2009). We disregard this possibility in the discussion of mechanisms.

[11] For both liquidity and tax mechanism, we assume that a reduction in labor income is not fully replaced by pension income. Therefore, we exclude the case that household income increases when labor supply

increases labor supply to compensate the loss in household income. In the case where B reaches FRA and claims a pension, but does not reduce labor supply, the pension will, cet. par., increase household income. Hence A, cet. par., decreases labor supply.

Third, due to the system of progressive *joint taxation*, the marginal tax rate of A increases (decreases) if total household income increases (decreases). Evidence from other countries suggests that labor supply of older workers increases with decreasing tax rates (Alpert & Powell, 2014; Laun, 2017). If B reduces labor supply, cet. par., household income decreases.[12] This decrease leads to a lower marginal tax rate for individual A, and therefore to an expected increase in labor supply. If B does not reduce labor supply but claims a pension, the household income increases. Therefore, the marginal tax rate for A increases, and we expect a negative effect on the labor supply of individual A.

Fourth, there is evidence that retirement increases hours of *housework* (Ciani, 2016; Stancanelli & Van Soest, 2012). If individual B reduces labor supply and increases housework, individual A may decrease housework. In this case, individual A may be willing to increase labor supply. If B does not reduce labor supply, we expect no effect on labor supply of individual A.

3.4 Data

For the analysis, we use data drawn from the Swiss Labor Force Survey (SLFS)[13] for the time period between 1991 and 2009. The SLFS is a rotating yearly panel of individuals above the age of 15. The survey is administered by the Federal Statistical Office (FSO). Participation in the survey is voluntary and individuals are interviewed for up to five consecutive years. For the sample of individuals aged between 58 and 70, 30% participated in one interview, 19% in two interviews, 14% in three interviews, 9% in four interviews, and 28% in five interviews. In the time period under consideration, the survey was carried out by telephone in the second quarter (April-June) of each year. The number of respondents aged between 58 and 70 increased from 2233 in 1991 to 8825 in 2009.

The survey provides extensive information on sociodemographic variables, labor supply status, earnings and household income of the respondent. The survey provides a variable for the year of birth of the respondent, but not the birth date. The spouse of the respondent is not directly interviewed. The respondent provides answers to questions on labor supply behavior and age of the spouse. There is information on the age of the spouse, but not on the year of birth. We impute the year of birth using the year of the interview and the age of the spouse.[14] There is sparse information on the health of the respondent and no information

decreases.

[12] See footnote 11.

[13] In German: Schweizerische Arbeitskräfteerhebung (SAKE).

[14] Given the year of interview and the age, the exact year of birth of the spouse is not identified (only a range of two years is identified). Since treatment classification is based on year of birth, there are spouses in the sample for whom we are not able to identify whether they have reached FRA or not. For example, the year of birth of a female spouse aged 61 and observed in 2000 could be either 1939 (FRA 63) or 1938 (FRA

on the health of the spouse.

For working individuals, the SLFS provides a set of variables describing the amount of time spent at work. The set includes *usual hours*, *contracted hours* and *actual hours in the previous week*. In this chapter, we use *contracted hours per week* as a measure of labor supply for wage employed. The underlying survey question is: "How many hours do you work according to your written or verbal contract per week?".[15] For self-employed, our measure of labor supply is *usual working hours per week* and the underlying survey question is: "How many hours do you usually work per week?".[16] We classify an individual as being in the labor force if the individual reports positive weekly working hours. In the sensitivity analysis, we check whether the results differ when using *actual working hours in the previous week* as an alternative measure for labor supply. We do not use *actual working hours in the previous week* in the main analysis since this variable contains more observations with zero working hours who are nevertheless working, e.g. they were sick or on holidays in the previous week. We consider this variable to be noisier and less reliable with respect to defining labor market participation.

Based on a set of questions, the SLFS classifies each respondent as being either employed, apprentice, unemployed, or non-participating. The group of non-participants includes disabled individuals, retirees and others. Non-participants are not asked about their working hours. We set the hours of unemployed and non-participants to zero.

In order to examine possible labor market frictions, we use *desired hours* as measure of *desired labor supply*. The underlying question is: "How many hours per week would you like to work?".[17]

3.5 Labor Supply Patterns

Before presenting the causal analysis, we examine the labor force patterns of older married individuals in Switzerland. Figure 3.3 displays the labor force participation rates by age and labor market status of the spouse for the time period 1991-2012.[18]

As illustrated in Figure 3.3a, labor force participation rates of married men with a working spouse are between 5 and 25 percentage points higher than the rates of men with a non-working spouse. The difference in participation rates between the two groups increases with age. Furthermore, male participation rates remain high - at above 80% - until the age of 60.

62), depending on whether the birthday is before or after the day of the interview in that year. Mastrobuoni (2009) deal with this issue by assuming that all birth dates within a year are equally likely. In contrast to his approach, we do not include observations with uncertain treatment status.

[15] German: "Wieviele Stunden pro Woche schaffen Sie gemäss mündlichem oder schriftlichem Arbeitsvertrag?". The corresponding SLFS variable is EK01.

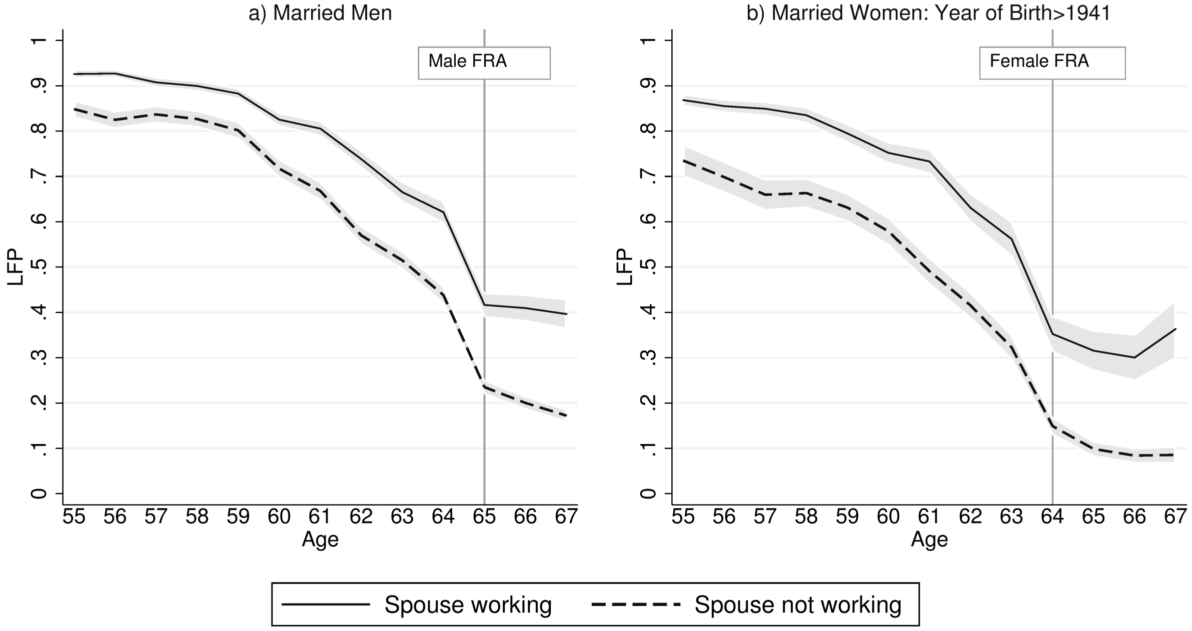
[16] German: "Wieviele Stunden schaffen Sie normalerweise pro Woche?". The corresponding SLFS variable is EK01.

[17] German: Wieviele Stunden in der Woche würden Sie gerne schaffen? The corresponding SLFS variable is EK07. In contrast to similar surveys in other countries, individuals are not asked to assume a constant hourly wage rate when answering the question for desired hours.

[18] Unlike our estimation samples (1991-2009), we use data until 2012 in order to have a larger sample size for working individuals past FRA. The pattern is very similar for the time period 1991-2009.

Female labor force participation rates are set out in Figure 3.3b. Again, the labor force participation rates are substantially higher for women with a working spouse than for women with a non-working spouse. In contrast to men, female labor force participation rates start to drop before the age of 60.

Figure 3.3: Labor Force Participation Rates by Labor Market Status of the Spouse



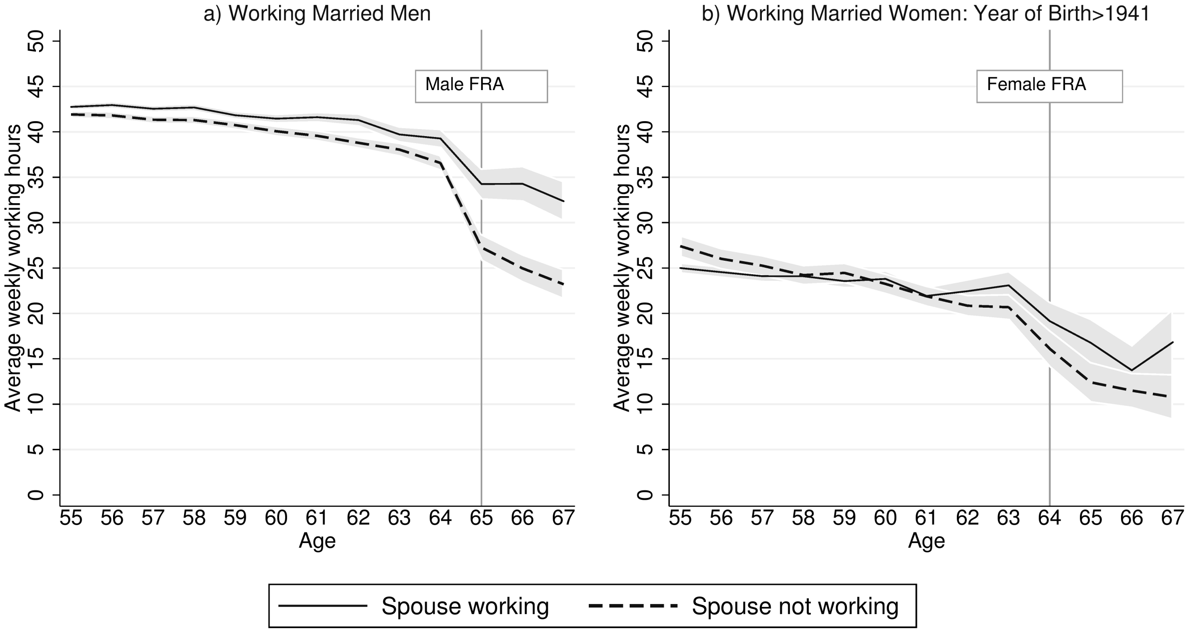
Note: Labor force participation rates (LFP) by age and labor market status of the spouse. Average values for period 1991-2012. Single and widowed individuals excluded. For women, only cohorts born after 1941 (FRA 64) are considered. Shaded gray area represents the 95% confidence interval for the mean estimate. Data source: Own calculations based on SLFS data, FSO.

Figure 3.4 displays the average weekly working hours by age and labor market status of the spouse of individuals participating in the labor market. Men work an average of approximately 40 hours per week before FRA is reached. In Switzerland, working 40 hours corresponds to a full-time employment.[19] There is a drop in hours worked at FRA. On average, men whose wives are not in the labor market work fewer hours at all ages. The difference is increasing with age.

Until the age of 57, the average working hours of women with a working husband are lower compared to women with a non-working husband. There is no difference in working hours between the ages of 58 and 61. Beyond the age of 62, women with a working husband work, on average, more hours than women with a non-working husband. Since the confidence intervals for the mean estimates overlap in most cases, these differences should be interpreted with caution.

[19] The legal maximum weekly working time in Switzerland is set at 45 hours for industrial, administrative, commercial, technical and sales jobs. All other sectors have a maximum of 50 hours. Working time regulation did not change in the period under consideration (1991-2012).

Figure 3.4: Average Weekly Working Hours by Labor Market Status of the Spouse



Note: Estimated average weekly working hours conditional on positive hours by age and labor market status of the spouse. Contracted hours for wage employed and usual hours for self-employed are used. Average values are for the period 1991-2012. Single and widowed individuals are excluded. For women, only cohorts born after 1941 (FRA 64) are considered. Shaded gray area represents the 95% confidence interval for the mean estimate. Data source: Own calculations based on SLFS data, FSO.

3.6 Empirical Approach

We estimate the causal effect of having a spouse B at or above FRA (treatment) on the labor supply of individual A. The total labor supply effect induced by having a spouse at or above FRA can be decomposed into 1) the average change in working hours of those working irrespective of treatment, plus 2) the average hours worked of those working in the case of treatment, and not working in the case of no treatment, minus 3) the average hours worked of those not working in the case of treatment, and working in the case of no treatment (Angrist, 2001; Staub, 2014). We refer to individuals working irrespective of treatment as *Participants*, to individuals working only in the case of treatment (or only in the case of no treatment) as *Switchers*.

We are interested in two causal effects. First, the causal effect of treatment on the probability of working (extensive margin). Second, the causal effect of treatment on working hours of individuals having positive hours irrespective of treatment, i.e. individuals working irrespective of whether their spouse has reached FRA or not (intensive margin).

3.6.1 Extensive Margin

Let h_{it} denote weekly working hours of individual i in interview year t. We estimate the extensive margin effect using a probit model of the form

$$P(h_{it} > 0 | T_{it}, \mathbf{X_{it}}) = \mathbf{\Phi}(\alpha_0 + \alpha_1 T_{it} + \mathbf{X_{it}} \boldsymbol{\alpha_2}), \tag{3.1}$$

where $\mathbf{\Phi}(\cdot)$ denotes the cumulative normal distribution. The treatment variable T_{it} is defined as

$$T_{it} = \begin{cases} 1 & \text{if the spouse of individual i is at or above FRA in period t,} \\ 0 & \text{otherwise.} \end{cases}$$

The matrix of controls $\mathbf{X}$ includes age dummies, age of the spouse, age of the spouse squared, education dummies, dummies for the year of the interview, a dummy whether the household size is larger than two, and a dummy whether the respondent is a Swiss citizen.

We are interested in the average partial effect of T_{it}, which measures the causal effect of having a spouse at or above the FRA on the probability of working. Our identifying assumption is that after controlling for covariates $\mathbf{X}$, respondents whose spouse has not yet reached FRA (control group) do not differ from respondents whose spouse has reached FRA (treatment group) with respect to observable and unobservable characteristics. Therefore, the remaining differences in labor supply participation rates between the treatment and the control group can be attributed to having a spouse at or above FRA.

We use two sources of variation to identify the effect. First, we exploit variation in age difference within couples. The variation in age difference is depicted in Figure 3.8 in Appendix 3.A. Second, we use a pension reform which increased the FRA for women in two steps. In 2001, the FRA was increased from 62 to 63. In 2004, the FRA was increased in a second step from 63 to 64, see Section 3.2.

A threat to our identification strategy are unobserved confounders affecting both; FRA status of spouse B, and the labor supply of individual A. Potential unobserved confounders include the health of spouse B, the birth of grand children, unobserved preferences, and changes in tax rates over time.

The health of spouse B potentially affects the decision whether and how much individual A works. Moreover, the health status of spouse B is associated with the age of spouse B, and therefore also with whether the spouse has reached FRA. On average, the older a spouse is, the lower the health status. We control for age of the spouse with a linear and a quadratic term. Figure 3.10 in Appendix 3.E provides evidence that the specification with a linear and a quadratic age term is sufficient to capture the effects of the age of spouse B on labor supply of individual A. If this approximation is not sufficient to capture the effect of spousal health, this poses a threat to our identification strategy. However, if reaching FRA affects the health of spouse B directly, we would not want to control for the health of spouse B. This case is part of the causal effect we want to measure.

The case of the birth of grandchildren is very similar. Grandchildren potentially affect the decision whether and how much individual A works. Moreover, having grandchildren is associated with the age of spouse B, and therefore also with whether the spouse has reached FRA. On average, the older a spouse is, the older the children. The older the children, the more likely are the children to have children themselves. Again, we assume that controlling for the age of individual A using dummies, and the age of spouse B with a linear and a quadratic term is sufficient to capture unobserved effects from grandchildren.

Finally, unobserved preferences of individual A that affect the labor supply of A may also be associated with the age of spouse B, see e.g. Bloemen and Stancanelli (2015). For example, individuals with preferences for a younger spouse could be willing to work more or work longer. We assume that controlling for the age of spouse B with a linear and a quadratic term is sufficient to capture effects from these unobserved preferences.

Another concern could be changes in tax rates over time. By including year dummies we capture changes in financial incentives induced by changes in tax rates over time. Moreover, we find that the changes in tax rates over time are small, see Figure 3.2.

3.6.2 Intensive Margin

At the intensive margin, we are interested in the causal effect of having a spouse at or above FRA on working hours of individuals working irrespective of whether their spouse has reached FRA or not. We employ a Tobit, a two-part, and a difference-in-difference estimator. For the intensive margin part of the two-part model as well as the difference-in-difference estimator, the estimation sample is restricted to individuals with positive working hours. For this reason, we need additional assumptions to identify the causal effect. The main issue is that a potential selection problem arises when conditioning on individuals with positive working hours (Angrist, 2001; Staub, 2014).[20] Alternatively, a Heckman selection model could be applied (Heckman, 1979). Although identification in the Heckman selection model relies on the functional form, an exclusion restriction is often required in practice to ensure identification (Cameron & Trivedi, 2009). Since we did not find a convincing exclusion restriction, we do not estimate the selection model.

Tobit Model

For the Tobit specification following Tobin (1958), we estimate the following equation using maximum likelihood

$$h_{it} = \max(0, \beta_1 \mathrm{T}_{it} + \mathbf{X_{it}} \boldsymbol{\beta_2} + u_{it}) \tag{3.2}$$

where we assume that $u_{it} \sim \text{Normal}(0, \sigma^2)$.[21] Treatment status T_i and the matrix of controls $\mathbf{X}$ are defined analogously to the extensive margin.

[20] See Chapter 2. The group of individuals with positive working hours consists of two potentially different subgroups. First, the subgroup with positive working hours irrespective of whether their spouse has reached FRA or not; and second, the subgroup with positive working hours only because their spouse has reached FRA (or only because their spouse has not yet reached FRA), with zero working hours otherwise. For the intensive margin effect, we are interested in the first group only. However we do not observe group membership. If the two groups differ in unobserved characteristics, a selection bias occurs.

[21] The usual motivation for the Tobit model is to assume a latent variable $y^* = X'\beta + \epsilon$ with $\epsilon \sim \text{Normal}(0, \sigma^2)$, and an observation rule such that the observed variable is equal to the latent variable if the latent variable is positive, and equal to zero if the latent variable is zero or negative (Cameron & Trivedi, 2009). In our case of working hours, a latent variable for working hours, allowed to be negative, is not very intuitive. Therefore we do not adopt the notion of a latent variable.

If the Tobit model is correctly specificed, β_1 represents the causal intensive margin effect, see Staub (2014). The coefficient β_1 captures the effect of having a spouse B at or above FRA on working hours of individual A with positive working hours irrespective of whether their spouse has reached FRA or not. In the Tobit model, extensive and intensive margin effects are closely linked due to the functional form. As a result, extensive and intensive margin effects are restricted to have the same sign.

Two-part Model

Following Cragg (1971), the two-part model specifies separate mechanisms for the participation decision (extensive margin), and the hours decision for individuals with positive hours (intensive margin). Since we are interested in the intensive margin effect, we focus on the second part, the hours decision for individuals with positive working hours. In the sample of individuals with positive working hours, we estimate an OLS regression of the form

$$\log h_{it} = \beta_1 \mathrm{T}_{it} + \mathbf{X_{it}}\boldsymbol{\beta_2} + \epsilon_{it}, \quad \text{for } h_{it} > 0, \tag{3.3}$$

where treatment status T_i and the matrix of controls $\mathbf{X}$ are defined analogously to the extensive margin. The log specification ensures that predicted hours are positive.

We are interested in the average partial effect of T_{it} on h_{it} (not on $\log h_{it}$) for individuals with positive hours. To achieve this, we apply the smearing retransformation proposed by Duan (1983). In the two-part model, the partial effect of T_{it} on h_{it} has a causal interpretation if treatment T_i has no effect on the participation decision (assumption of no Switchers) or if the assumption of conditional mean independence holds (see Angrist (2001), Staub (2014) and Appendix 2.A of Chapter 2).[22] If individuals leave the labor market due to their spouse reaching FRA (no Switchers assumption violated) and if individuals reacting at the extensive margin have different average hours than Participants (conditional mean independence assumption violated), the estimated intensive margin effect is biased.

Difference-in-Difference on Positive Outcomes

The difference-in-difference estimator on positive outcomes represents an alternative to the two-part model. The identification of the causal effect in the difference-in-difference estimator is based on the findings of Chapter 2. The estimation, however, differs for reasons of consistency within this chapter. In this chapter we employ models for realized outcomes. Hence, we estimate - in the sample of individuals with positive working hours in two subsequent periods - an OLS regression of the form

$$\Delta h_{it} = \beta_1 \Delta \mathrm{T}_{it} + \Delta \mathbf{X_{it}}\boldsymbol{\beta_2} + \nu_{it} \qquad \text{for } h_{it} > 0, h_{it-1} > 0, \tag{3.4}$$

[22] Conditional mean independence assumes that the mean of individuals who are working irrespective of treatment status is equal to the mean of individuals working only in case of treatment or only in case of no treatment.

where $\Delta z_{it} = z_{it} - z_{i,t-1}$ for $z_{it} \in \{h_{it}, T_{it}, \mathbf{X}_{it}\}$.[23] Treatment status T_i is defined analogously to the extensive margin. The matrix of controls $\mathbf{X}$ includes age dummies, age of the spouse, age of the spouse squared, dummies for the year of the interview, and a dummy whether the household size is larger than two. Dummies for education and whether the respondent is a Swiss citizen are constant and therefore dropped.

We are interested in β_1, the causal intensive margin effect. The coefficient β_1 captures the causal effect of having a spouse at or above FRA on working hours for individuals with positive working hours irrespective of whether their spouse has reached FRA or not.

Compared to the intensive margin part of the two-part estimator, the difference-in-difference estimator on positive outcomes does not assume that treatment has no effect on the participation decision or that conditional mean independence holds. As we demonstrate in Chapter 2, the central assumptions of the difference-in-difference estimator on positive outcomes are *no pre-treatment effect*, *treatment* and *time monotonicity at the extensive margin*, and *common trend in positive outcomes.*

No pre-treatment effect requires that in expectation, having a spouse who reaches FRA in period t does not affect the respondent's labor supply in period $t-1$. If this assumption is violated, it is likely that individuals adjusted their labor supply in the same direction in period $t-1$ as they do in period t. In this case, the estimated causal effect is biased towards zero, and would therefore represent a lower bound of the true causal effect (in absolute terms). *Treatment monotonicity at the extensive margin* excludes the possibility that individuals work when their spouse has reached FRA, but would not work if their spouse had not yet reached FRA. Similarly, *Time monotonicity at the extensive margin* excludes unretirement, i.e. the possibility that individuals work in period t, but do not work in period $t-1$. *Common trend in positive outcomes* assumes a common trend between individuals with positive hours in case their spouse has reached FRA and individuals with positive hours in case their spouse has not yet reached FRA. The common trend assumption can be motivated in the data using pretreatment observations. We compare the change in hours from the penultimate period $(t-2)$ to the previous period $(t-1)$ for the treatment group (spouse reaches FRA in period t), and the control group (spouse does not reach FRA in period t). We find no evidence for a difference in trends relating to working hours between the two groups, see Table 3.5 in Appendix 3.A.

In Appendix 3.B, we present summary statistics for the samples used to estimate the probit and tobit model, the two-part model, and the difference-in-difference estimator. The differences between the means of treatment and control group are largest for *age* and *age spouse*. For differences in *age*, we control using age dummies. For differences in *age spouse*,

[23] Note that this setting differs from the conventional difference-in-difference setting with two groups and two periods, in which both groups are untreated in the first period and one group is treated in the second period. We are in a setting with multiple time periods and multiple groups. Individuals can be treated at different points in time. This version of the difference-in-difference estimator can be written as $h_{it} = \delta_i + \delta_t + \beta_1 T_{it} + \mathbf{X_{it}}\boldsymbol{\beta_2} + \epsilon_{it}$, where δ_i and δ_t are fixed effects for each individual i and each time period t, and T_{it} is the indicator for being in a treatment group after treatment occurred. This is equivalent to a fixed effects model. Taking first differences yields $\Delta h_{it} = \{\delta_t - \delta_{t-1}\} + \beta_1 \Delta \mathrm{T}_{it} + \Delta \mathbf{X_{it}}\boldsymbol{\beta_2} + \nu_{it}$, which is the equation we estimate, except that the time fixed effects are included in the matrix of controls $\mathbf{X}$.

we control using a linear and a quadratic term of age of the spouse.

3.7 Results

The results section is divided into extensive and intensive margin. Both parts start with simple graphical evidence. Subsequently, we present estimates of the causal effect of having a spouse B at or above the FRA on labor supply of individual A.

3.7.1 Extensive Margin

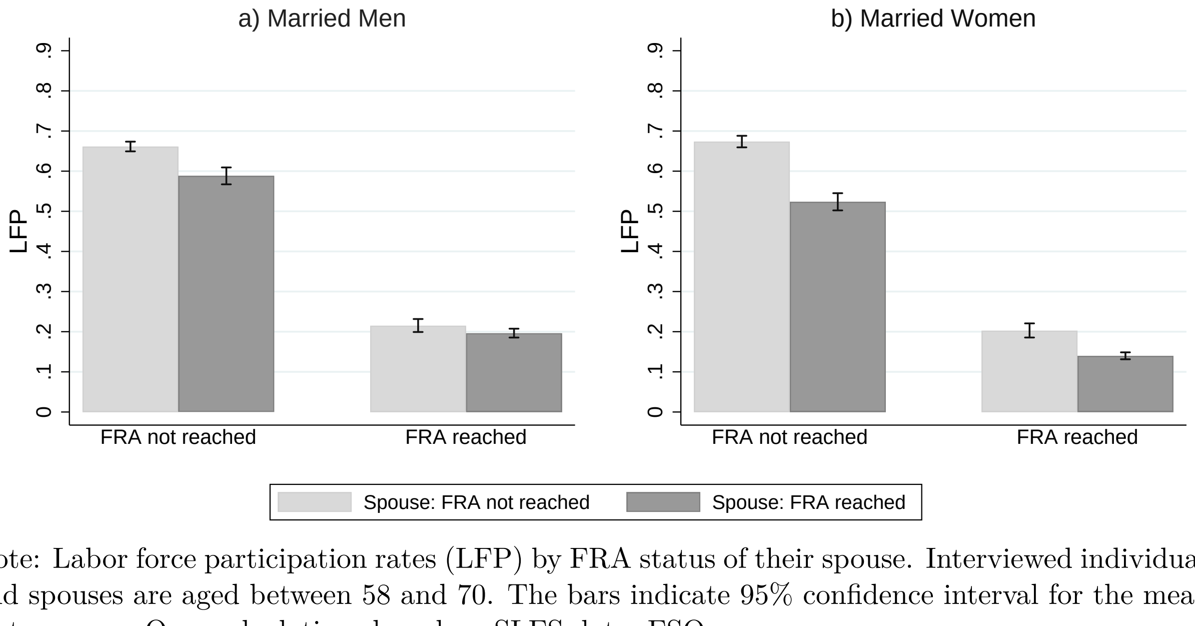
Graphical Evidence

Figure 3.5 presents the labor force participation (LFP) rates of married men (left panel) and married women (right panel), depending on whether they have reached their own FRA or not. The light-gray bars on the left indicate the LFP rates of respondents with a spouse who has not yet reached FRA; the dark-gray bars on the right indicate the LFP rates of respondents with a spouse who has reached FRA. Not surprisingly, reaching their own FRA is associated with a decrease in LFP rates for both men and women.

Men who have not reached their own FRA have a slightly higher LFP rate in case their spouse has not reached FRA. Men who have reached their own FRA have similar LFP rates regardless of whether their spouse has reached FRA or not.

A more distinct pattern can be observed among women in Figure 3.5b. Women who have not reached their own FRA have a higher LFP rate in case their spouse has not reached FRA. Similarly, women who have reached their own FRA have also a higher LFP rate in case their spouse has not yet reached FRA.

Figure 3.5: Labor Force Participation Rates by FRA Status of Spouse



Note: Labor force participation rates (LFP) by FRA status of their spouse. Interviewed individuals and spouses are aged between 58 and 70. The bars indicate 95% confidence interval for the mean. Data source: Own calculations based on SLFS data, FSO.

Estimation Results

In the graphical analysis, we controlled for age of the respondent, but not for other potential confounders. By estimating the probit model described in Section 3.6.1, we can both control for potential confounders and increase precision of the estimate of interest. Besides controlling for age and age of the spouse, we also control for the year of the interview, education, household size, and whether the respondent is Swiss. The results are presented in Table 3.2.

Table 3.2: Estimation Extensive Margin

	Dep. variable: Indicator $\mathbb{1}$(working hours > 0)			
	Men		Women	
	APE	SE(APE)	APE	SE(APE)
Spouse FRA reached	0.012	(0.014)	-0.034^{***}	(0.013)
FRA reached			-0.133^{***}	(0.016)
Age dummies	Yes		Yes	
Year dummies	Yes		Yes	
Age spouse	Yes		Yes	
Age spouse squared	Yes		Yes	
Education dummies	Yes		Yes	
Household size > 2	Yes		Yes	
Swiss citizenship	Yes		Yes	
Observations	15683		14643	

Note: Results Probit estimation. Average partial effects (APE) reported. Interviewed individuals and spouses are aged between 58 and 70. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

In the case of men, we find no evidence that having a spouse at or above FRA has an effect on LFP. This finding is line with graphical results from Figure 3.5. By contrast, women react when their spouses reach FRA. On average, women whose spouses have reached FRA are 3.4 percentage points less likely to be in the labor force than women whose spouses have not yet reached FRA.

Since there is variation in the FRA of women, we are also able to identify the effect of women reaching their own FRA.[24] Women who have reached their own FRA are 13.3 percentage points less likely to work compared to women who have not yet reached their own FRA. Therefore, the *direct* effect of -13.3 percentage points is approximately four times larger than the *indirect* effect of -3.4 percentage points.

These results are in line with Lalive and Parrotta (2017). Using another data source - census data from Switzerland - and a double regression discontinuity design, they find the

[24] If there was no variation in the FRA, reaching the own FRA would be multicollinear with the age dummies.

same gender asymmetry as we do. The LFP of women decreases by approximately 2 to 3 percentage points when their spouses reach FRA. Moreover, the LFP of women decreases by approximately 12 percentage points when they reach their own FRA. They also find no significant indirect effect for men.

Anticipation and Delay Effects

Individuals with a spouse approaching FRA potentially anticipate treatment and therefore change behavior already before their spouse has actually reached FRA. If this anticipation effect has the same sign as the causal effect, the estimated causal effect is biased towards zero. We examine anticipation by including a dummy in the estimation equation which is equal to 1 if the spouse is one year younger than his/her FRA and zero otherwise. Table 3.11 in Appendix 3.C presents the results. We do not find evidence for anticipation. For both men and women, the estimated effect is small and not significant.

On the other hand, individuals might not adjust their labor supply immediately when their spouse reaches FRA, but with delay. If these delayed effects have the same sign as the causal effect, the estimated causal effect represents again a lower bound (in absolute terms) of the total causal effect. We examine the delay effect by modifying our dummy variables of interest. First, we include a dummy that equals one when the spouse's age equals his/her FRA. Second, we include a dummy that equals one when the spouse is one or more years older than his/her FRA. In this specification, the first dummy captures the immediate effect, i.e. the effect of having a spouse directly at FRA relative to having a spouse who has not yet reached FRA. The second dummy captures the sum of immediate and delayed effect, i.e. the effect of having a spouse who is one or more years older than FRA relative to having a spouse who has not yet reached FRA.[25] Table 3.12 in Appendix 3.C presents the results. We find evidence that women respond with delay. The immediate effect of having a spouse directly at FRA is -2.7 percentage points, the combined effect of immediate and delayed effects is -4.7 percentage points.

3.7.2 Intensive Margin

Graphical Evidence

Figure 3.6 illustrates the average working hours - conditional on positive working hours - of married men (left panel) and married women (right panel), depending on whether they have reached their own FRA or not. The light-gray bars on the left indicate the average working hours of respondents with a spouse who has not yet reached FRA; the dark-gray bars on the right indicate the average working hours of respondents with a spouse who has reached FRA. We observe that the intensive margin is a margin at which men and women adjust their labor supply. Reaching their own FRA is associated with a decrease in working hours

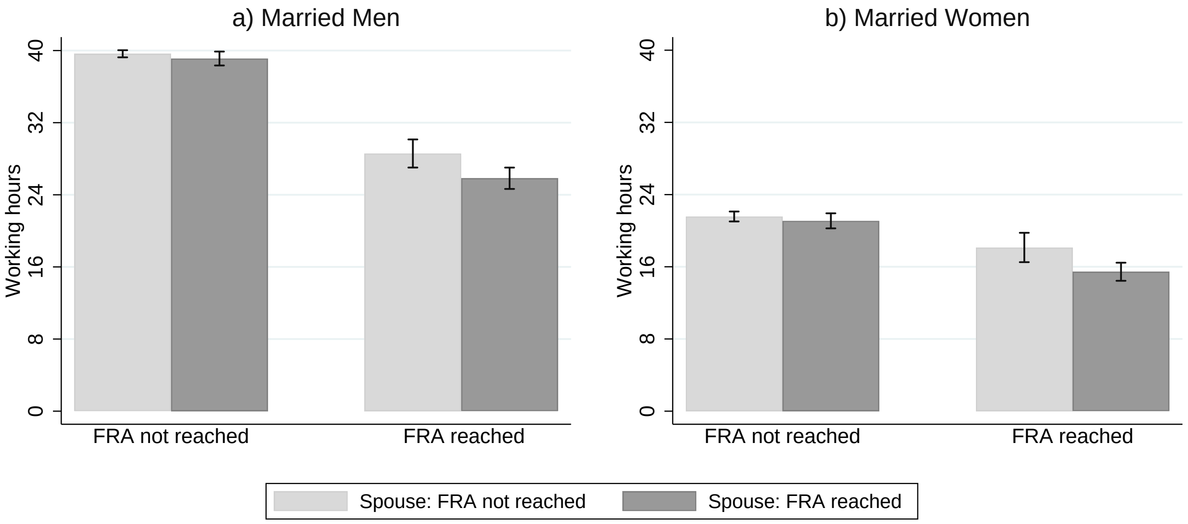
[25] Note that the sum of immediate and delayed effect is not equal to the effect presented in Table 3.2. The effect presented in Table 3.2 corresponds to a weighted average of immediate and delayed effect.

for both men and women.

Men who have not reached their own FRA have similar average working hours regardless of whether their spouse has reached FRA or not. Men who have reached their own FRA have slightly higher average working hours in case their spouse has not reached FRA.

We observe a similar pattern for women. The average working hours of women who have not reached their own FRA are similar, regardless of whether their spouse has reached FRA or not. Women who have reached their own FRA have slightly higher average working hours in case their spouse has not reached FRA.

Figure 3.6: Weekly Working Hours by FRA Status of Spouse



Note: Weekly working hours of individuals with positive working hours, by FRA status of their spouse. Interviewed individuals and spouses are aged between 58 and 70. Error bars indicate 95% confidence interval for the mean. Data source: Own calculations based on SLFS data, FSO.

Estimation Results

We employ a Tobit, a two-part, and a difference-in-difference estimator. The results for men are presented in Table 3.3, the results for women in Table 3.4. We do not find evidence that men adjust their working hours when their spouses reach FRA. For all models the estimated effect is not statistically significant. The estimate of the Tobit model is positive, whereas the estimates of the two-part and difference-in-difference estimator are negative.

In combination with the graphical evidence presented in Figure 3.6a, we conclude that the intensive margin is a margin at which men adjust their labor supply, but there is no evidence of spillover effects from their wives.

Table 3.3: Estimation Intensive Margin Men

	Dependent variable: Working hours		
	Tobit	Two-part	Diff-in-diff
Spouse FRA reached	0.624	−0.786	−0.810
	(1.318)	(0.687)	(0.715)
Age dummies	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes
Age spouse	Yes	Yes	Yes
Age spouse squared	Yes	Yes	Yes
Education dummies	Yes	Yes	No
Household size > 2	Yes	Yes	Yes
Swiss citizenship	Yes	Yes	No
Observations	15683	6724	3995

Note: Results intensive margin effect for men with Tobit, two-part, and difference-in-difference estimator. Interviewed individuals and spouses are aged between 58 and 70. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

As for women, we find only weak evidence of a small negative indirect effect. In the Tobit model, the effect of having a spouse at or above FRA on working hours is approximately -2.8 hours and statistically significant at the 1% level.[26] In the two-part model, the estimated indirect effect on working hours is approximately -0.7 hours and statistically significant at the 10% level. The estimated indirect effect using the difference-in-difference estimator is negative but not statistically significant.

In the Tobit model, extensive and intensive margin effects are restricted to have the same sign. Since we find a negative effect at the extensive margin for women (see Table 3.2), the intensive margin effect is restricted to be negative as well. Therefore, the finding of a negative intensive margin effect could be driven only by the extensive margin, without a *true* intensive margin effect. Since we are concerned with this restriction, the Tobit estimate should be interpreted with caution.

The two-part model allows for separate mechanisms determining extensive and intensive margin. The two-part estimate has a causal interpretation if treatment has no effect on the participation decision or if the assumption of conditional mean independence holds (see Section 3.6.2). Since we find a negative effect at the extensive margin for women (see Table 3.2), the assumption that treatment has no effect on the participation decision is likely violated. The conditional mean independence assumption is also restrictive. It is possible that unobserved characteristics of individuals who are working irrespective of treatment status are different from unobserved characteristics of individuals working only in case of

[26] This is the causal effect on working hours of women with positive hours irrespective of whether their spouse has reached FRA or not, see Section 3.6.2.

treatment or only in case of no treatment. If individuals reacting at the extensive margin have lower average working hours than *Participants*, the estimated intensive margin effect is biased towards zero.[27]

Compared to the Tobit and two-part model, the difference-in-difference estimator relies on fewer observations. This is because we need positive outcomes in two subsequent periods. Moreover, since we take first differences, the number of observations is further reduced. As a result, the standard error of the intensive margin effect, cet. par., increases.

We conclude that there is only weak evidence for a small indirect effect. If there is an effect and women adjust their working hours when their spouse reaches FRA, the magnitude of the average effect is likely to be small.

Due to the variation in FRA of women, we can also identify the direct effect of women reaching their own FRA. Depending on the model, the direct effect varies between -2.2 and -10.7 hours. The effects are statistically significant at least at the 10% level. Again, the Tobit assumptions are likely too restrictive. The estimates from the two-part and the difference-in-difference estimator are more credible.

Table 3.4: Estimation Intensive Margin Women

	Dependent variable: Working hours		
	Tobit	Two-part	Diff-in-diff
Spouse FRA reached	−2.821***	−0.707*	−1.252
	(0.987)	(0.419)	(0.763)
FRA reached	−10.723***	−3.469***	−2.178*
	(1.361)	(0.593)	(1.162)
Age dummies	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes
Age spouse	Yes	Yes	Yes
Age spouse squared	Yes	Yes	Yes
Education dummies	Yes	Yes	No
Household size > 2	Yes	Yes	Yes
Swiss citizenship	Yes	Yes	No
Observations	14643	5191	2993

Note: Results intensive margin effect for women with Tobit, two-part, and difference-in-difference estimator. Interviewed individuals and spouses are aged between 58 and 70. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

[27] This statement is true under the assumption that the true intensive margin effect is negative and that there are no individuals working only in case their spouse has reached FRA, and not working in case their spouse has not reached FRA. If the true intensive margin effect is positive, the estimated intensive margin effect is biased away from zero. If there are individuals working only in case their spouse has reached FRA, and not working in case their spouse has not reached FRA, the direction of the bias can be positive or negative.

3.7.3 Robustness

Extensive Margin

Instead of using a probit model, we estimate a linear probability model. The estimation results are presented in Table 3.13 in Appendix 3.E. The results are very similar in both sign and magnitude. In the men sample, the effect is not statistically significant. In the women sample, the estimated indirect effect is approximately 3.9 percentage points and statistically significant at the 1% level.

Instead of using contracted working hours as the dependent variable, we use actual working hours of the previous week as an alternative measure for labor supply. The estimation results are presented in Table 3.14 in Appendix 3.E. The results are very similar in both sign and magnitude. We find no evidence that men adjust their labor force participation rate when their wives reach FRA. By contrast, the LFP rate of women decreases by 2.7 percentage points when their spouses reach FRA.[28]

As discussed in Section 3.6, we control for the age of the spouse with a linear and a quadratic term in the main specification. In a robustness check, we test whether the results are robust to this specification. We exclude the age of the spouse as a control variable, but restrict the age of the spouse to be between two years before and two years after reaching FRA. The results are presented in Table 3.17, Appendix 3.E. For men, the effect of the spouse reaching FRA is, once again, small and not statistically significant. For women, the effect is -4.4 percentage points and significant.

We conclude that the extensive margin results are not sensitive with respect to the model for the binary outcome, with respect to the definition of labor supply, or with respect to the specification with which the age of the spouse is included in the model.

Intensive Margin

Instead of using contracted working hours, we use actual working hours of the previous week. The results are presented in Tables 3.15 and 3.16 in Appendix 3.E. For men, the estimated intensive margin effect are not statistically significant. As for women, the estimated intensive margin effects are smaller in absolute terms compared to the main specification. The estimates of the indirect effect from the two-part and the difference-in-difference estimator are not statistically significant. Moreover, the results are not sensitive to the specification with which the age of the spouse is included in the model (Tables 3.18 and 3.19 in Appendix 3.E). These robustness checks provide further evidence that - if there is an intensive margin effect for women - the effect is small in magnitude.

[28] Differences between having positive contracted working hours and having positive working hours in the previous week arise for example when a respondent was sick or on holidays in the previous week. Hence there are natural situations in which contracted working hours are positive and actual working hours in the previous week are zero. Vice versa, a situation where actual working hours in the previous week are positive and contracted working hours are zero is unlikely. Overall, this means that actual working hours in the previous week contains more zeros than contracted working hours. This explains why the treatment effect magnitude is slightly smaller when using actual working hours in the previous week.

3.8 Discussion

As indicated in Table 3.1, the sign of the expected labor supply reaction of individual A to spouse B reaching FRA depends on whether B reduces labor supply. In our estimation sample, approximately 32% of men and 22% of women reduce their labor supply by 8 or more hours when reaching FRA. Of those who do not reduce their labor supply, approximately 76% of men and 67% of women are already retired, while 24% of men and 33% of women are still working.

In the case of women, we observe a negative labor supply reaction. If the effect is driven primarily by women whose husbands reduce labor supply at FRA, complementarities in leisure must be sufficiently large to outweigh liquidity, joint taxation, and housework effects. The negative labor supply reaction, however, can also be explained by liquidity and joint taxation effects of women whose husbands do not reduce labor supply at FRA. In the case of men, we do not find evidence of a labor supply reaction. This does not rule out that men have preferences for joint leisure time, since liquidity, joint taxation, and housework effects possibly outweigh complementarity in leisure effects.

Difference Between Men and Women

There is heterogeneity in complementarity in leisure, liquidity, joint taxation and housework effects, which may explain part of the asymmetric reaction of men and women. The change in labor supply of men reducing their workload when reaching their *own FRA* is larger than the reaction of women. Considering only individuals who reduce labor supply at FRA by 8 or more hours, we find that men reduce weekly working hours on average by 33 hours (extensive and intensive reaction combined) whereas women decrease their weekly working hours by 23 when reaching their *own FRA*. This difference can explain part of the asymmetry in the indirect effect since, cet. par., the complementarity effect is stronger the larger the labor supply reaction of the spouse.

The *liquidity* and *joint taxation* effect depend on the labor supply reaction of the spouse. We consider first the case where spouse B reduces labor supply at own FRA. In the analysis above, we found that men react more strongly to their own FRA. Assuming men and women achieve the same replacement rate, the drop in household income is larger when the husband reaches his FRA. Therefore, tax and liquidity effects are positive and larger for women than for men. For this reason, tax and liquidity effects partially offset the asymmetry stemming from differences in complementarity in leisure effects. In the case where spouse B does not reduce labor supply at own FRA, we do not find evidence for asymmetries with respect to *liquidity* and *joint taxation effects*. On the basis of questions on early retirement in the SLFS, we find that only 0.93% of men and 0.85% of women answered that taxes were the main determinant of their early retirement decision. These results suggest that tax considerations are only of secondary importance when deciding when to retire.

Changes in relation to housework upon retirement are similar for men and women. On the basis of questions on housework in the SLFS, we find that men increase the amount of housework they do by approximately 40 minutes a day, whereas this increase is approximately 60 minutes in the case of women.

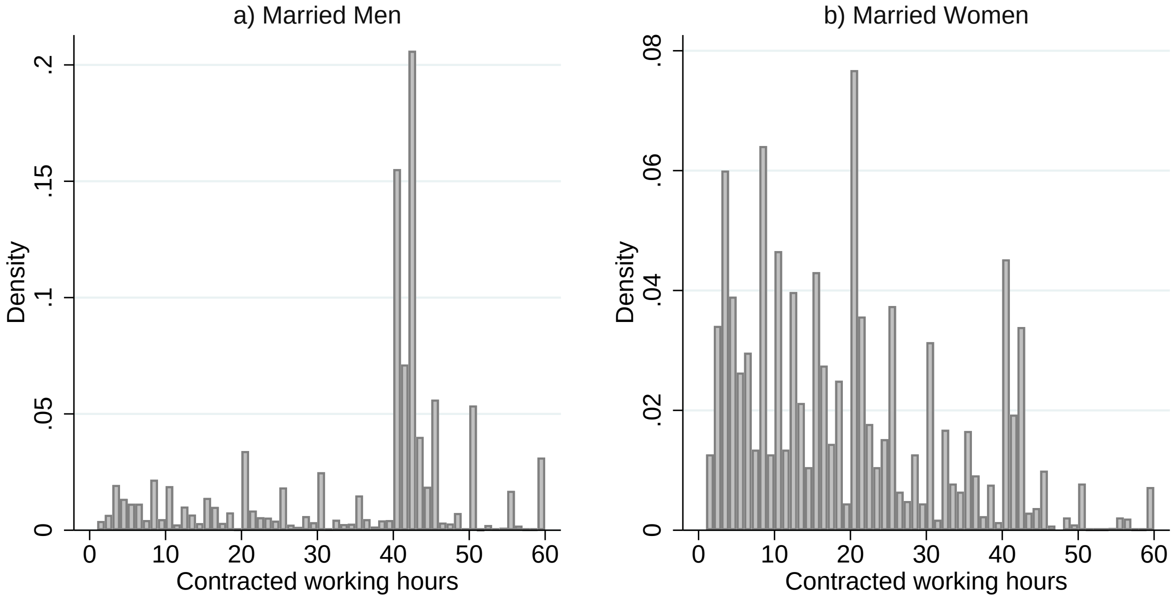
To sum up, our analysis provides evidence that *complementarity in leisure effects* are an important mechanism for the indirect effect. *Liquidity effects* can play a role when reacting to the spouse reaching FRA. Finally, we cannot exclude the possibility that the asymmetric reaction is driven by gender differences in relation to preferences.

Margin of Reaction

We find that women react at the extensive margin, but there is no evidence of a substantial intensive margin reaction. We would like to point out several potential explanations.

First, women may want to reduce their working hours, but are prevented from doing so by hours constraints set by firms. We examine this mechanism by analyzing desired working hours. Instead of using contracted working hours as dependent variable, we use desired hours. The results are presented in Tables 3.9 and 3.10 in Appendix 3.C. We do not find evidence that women would like to reduce their working hours in response to their husbands reaching FRA. Second, social norms in relation to working hours may discourage women from adjusting their working hours when their husbands reach FRA. The distribution of weekly working hours for men and women is presented in Figure 3.7. The graph suggests that social norms are less pronounced for women than for men. Third, *fixed costs of work* imply that individuals are not willing to work below a minimum number of hours. A large proportion of women work 21 hours per week or less, see Figure 3.7. This group is likely to react at the extensive margin as a result of the presence of fixed costs of work. Fourth, *complementarities in leisure* may be discontinuous at zero working hours. For example, it may be necessary for both partners to be out of the labor force if they want to change residence for retirement, or travel for an extended period.

Figure 3.7: Distribution of Weekly Working Hours of Men and Women



Note: Distribution of weekly working hours of married men and women aged 58-70. Only respondents who work between 1 and 60 hours per week are shown. Own calculations based on SLFS.

3.9 Conclusion

In this chapter, we estimated labor supply responses to the spouse reaching FRA. We find that labor force participation of women drops by approximately 3 percentage points when their spouses reach FRA. By contrast, the LFP of men does not respond to their spouses reaching FRA. At the intensive margin, we find no evidence for substantial effects. The estimated effects are small in magnitude and mostly non-significant. This is despite the fact that older workers use working hours to adjust their labor supply when they reach FRA.

We identify four different mechanisms that could explain the effect on labor supply of having a spouse at or above FRA: complementarities in leisure, joint taxation, liquidity, and housework effects. Since we find a negative indirect effect for women, we argue that complementarities in leisure and liquidity effects are important mechanisms for the indirect effect. We explain the absence of a substantial intensive margin reaction in the case of women on the basis of the presence of fixed costs of work.

3.10 Bibliography

Alpert, A., & Powell, D. (2014). Estimating Intensive and Extensive Tax Responsiveness: Do Older Workers Respond to Income Taxes? *RAND Working Paper Series*, *WR-987-1*.

Angrist, J. D. (2001). Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors. *Journal of Business & Economic Statistics*, *19*(1), 2–28.

Banks, J., Blundell, R., & Casanova Rivas, M. (2010). The Dynamics of Retirement Behavior in Couples: Reduced-Form Evidence from England and the US. *Working Paper*.

Blau, D. M. (1998). Labor Force Dynamics of Older Married Couples. *Journal of Labor Economics*, *16*(3), 595–629.

Bloemen, H. G., & Stancanelli, E. G. (2015). Toyboys or Supergirls? An Analysis of Partners' Employment Outcomes When She Outearns Him. *Review of Economics of the Household*, *13*(3), 501–530.

Börsch-Supan, A., & Schnabel, R. (1999). Social Security and Retirement in Germany. In J. Gruber & D. Wise (Eds.), *Social security and retirement around the world* (pp. 135–180).

Bundesamt für Sozialversicherungen. (2009). *AHV-Statistik* (Tech. Rep.).

Bütler, M. (2009). Switzerland: High replacement Rates and Generous Subsistence as a Barrier to Work in Old Age. *The Geneva Papers*, *34*, 561–577.

Cameron, C., & Trivedi, P. (2009). *Microeconometrics: Methods and Applications* (8th ed.). Cambridge: University Press.

Ciani, E. (2016). Retirement, Pension Eligibility and Home Production. *Labour Economics*, *38*, 106–120.

Coile, C. (2004). Retirement Incentives and Couples' Retirement Decisions. *Topics in Economic Analysis & Policy*, *4*(1).

Cragg, J. G. (1971). Some Statistical Models for Limited Dependent Variables With Application to the Demand for Durable Goods. *Econometrica*, *39*(5), 829–844.

Cribb, J., Emmerson, C., & Tetlow, G. (2013). Incentives, Shocks or Signals: Labour Supply Effects of Increasing the Female State Pension Age in the UK. *IFS Working Paper*, *W13/03*.

Duan, N. (1983). Smearing Estimate: A Nonparametric Retransformation Method. *Journal of the American Statistical Association*, *78*(383), 605–610.

Gustman, A. L., & Steinmeier, T. L. (2004). Social Security, Pensions and Retirement Behaviour Within the Family. *Journal of Applied Econometrics*, *19*, 723–737.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*(1), 153–161.

Hersche, M. (2018). Theoretical and Empirical Essays on Labor Supply of the Elderly. (Diss. ETH No. 25377).

Hersche, M., & Moor, E. (2018). Labor or Leisure? Labor Supply of Older Couples and the Role of Full Retirement Age. *Netspar Working Paper Series*(03/2018).

Hospido, L., & Zamarro, G. (2014). Retirement Patterns of Couples in Europe. *IZA Journal* *of European Labor Studies*, *3*(12).

Hurd, M. D. (1990). The Joint Retirement Decision of Husbands and Wives. In D. A. Wise (Ed.), *Issues in the economics of aging* (pp. 231–258). University of Chicago Press.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Kantarci, T., & Van Soest, A. (2008). Gradual Retirement: Preferences and Limitations. *De Economist*, *156*(2), 113–144.

Lalive, R., & Parrotta, P. (2017, 6). How Does Pension Eligibility Affect Labor Supply in Couples? *Labour Economics*, *46*, 177–188.

Laun, L. (2017). The Effect of Age-Targeted Tax Credits on Labor Force Participation of Older Workers. *Journal of Public Economics*, *152*, 102–118.

Mastrobuoni, G. (2009). Labor Supply Effects of the Recent Social Security Benefit Cuts: Empirical Estimates Using Cohort Discontinuities. *Journal of Public Economics*, *93*, 1224–1233.

Selin, H. (2017). What Happens to the Husband's Retirement Decision When the Wife's Retirement Incentives Change? *International Tax and Public Finance*, *24*(3), 432–458.

Stancanelli, E. (2017). Couples' Retirement Under Individual Pension Design: A Regression Discontinuity Study for France. *Labour Economics*, *49*, 14–26.

Stancanelli, E., & Van Soest, A. (2012). Retirement and Home Production: A Regression Discontinuity Approach. *American Economic Review*, *102*(3), 600–605.

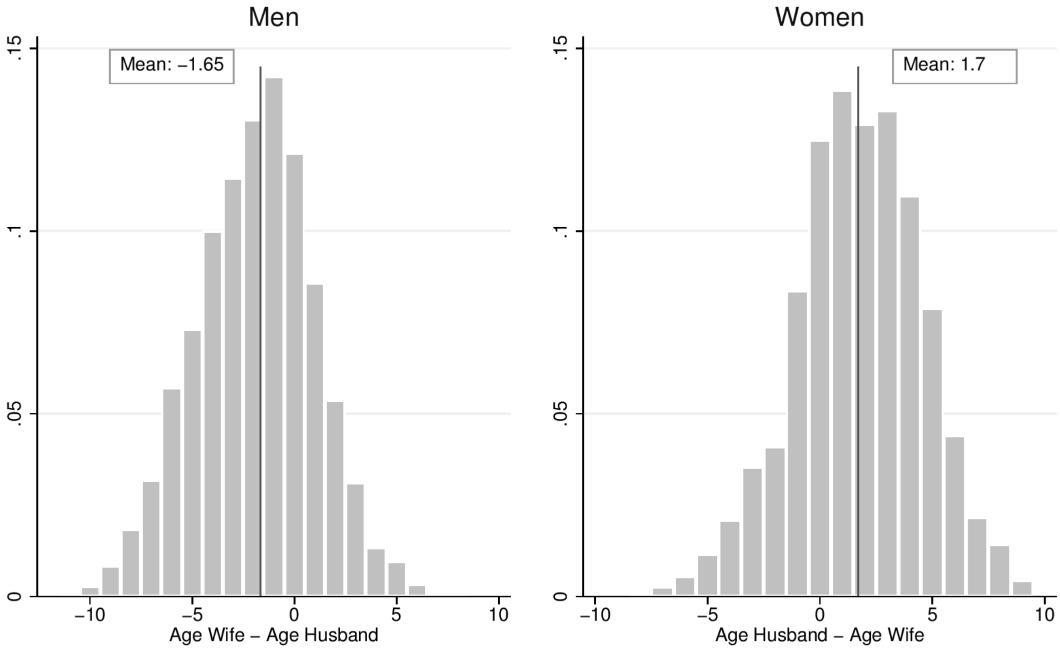
Staub, K. (2014). A Causal Interpretation of Extensive and Intensive Margin Effects in Generalized Tobit Models. *The Review of Economics and Statistics*, *96*(2), 371–375.

Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, *26*(1), 24–36.

Zweimüller, J., Winter-Ebmer, R., & Falkinger, J. (1996). Retirement of Spouses and Social Security Reform. *European Economic Review*, *40*, 449–472.

Appendix 3.A Identification

Figure 3.8: Identifying Variation in Age Difference Within Couples



Note: Data source: Own calculations based on SLFS data, FSO.

Table 3.5: Estimation Intensive Margin $t-1$

	Dependent variable: Δ(Working hours, t-1)			
	Men		Women	
	Coef.	SE(Coef.)	Coef.	SE(Coef.)
Spouse FRA reached	−0.537	(0.954)	0.343	(0.800)
FRA reached	No		Yes	
Age dummies	Yes		Yes	
Year dummies	Yes		Yes	
Education dummies	No		No	
Household size > 2	Yes		Yes	
Swiss citizenship	No		No	
Age spouse	Yes		Yes	
Age spouse squared	Yes		Yes	
Observations	2367		1751	

Note: Results difference-in-difference estimation on positive hours in period t-1. Interviewed individuals and spouses are aged between 58 and 70. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

Appendix 3.B Summary Statistics

Table 3.6: Summary Statistics Probit and Tobit Sample

	Men		Women	
	Treatment Mean / SD	Control Mean / SD	Treatment Mean / SD	Control Mean / SD
Age	65.98	63.06	64.57	61.57
	(2.90)	(2.96)	(2.84)	(2.65)
Year of Interview	2003.15	2003.80	2003.72	2002.42
	(4.10)	(4.34)	(3.97)	(4.43)
Age Spouse	65.73	60.15	67.40	61.73
	(2.18)	(1.55)	(1.68)	(1.85)
Education				
Lower Education	0.18	0.18	0.42	0.36
	(0.39)	(0.39)	(0.49)	(0.48)
Secondary Education	0.52	0.49	0.50	0.54
	(0.50)	(0.50)	(0.50)	(0.50)
Higher Education	0.29	0.32	0.08	0.10
	(0.45)	(0.47)	(0.27)	(0.30)
Household size > 2	0.08	0.17	0.09	0.11
	(0.26)	(0.37)	(0.28)	(0.32)
Swiss Citizenship	0.77	0.73	0.80	0.83
	(0.42)	(0.45)	(0.40)	(0.38)
Observations	7206	8477	8461	6182

Table 3.7: Summary Statistics Two-Part Sample (Intensive Margin)

	Men		Women	
	Treatment Mean / SD	Control Mean / SD	Treatment Mean / SD	Control Mean / SD
Age	64.26	61.83	62.62	60.39
	(3.09)	(2.50)	(2.77)	(2.04)
Year of Interview	2001.95	2003.01	2003.00	2001.85
	(4.72)	(4.71)	(4.46)	(4.52)
Age Spouse	65.15	59.91	67.07	61.46
	(2.18)	(1.50)	(1.67)	(1.86)
Education				
Lower Education	0.14	0.15	0.35	0.32
	(0.35)	(0.36)	(0.48)	(0.47)
Secondary Education	0.52	0.48	0.53	0.56
	(0.50)	(0.50)	(0.50)	(0.50)
Higher Education	0.35	0.37	0.12	0.12
	(0.48)	(0.48)	(0.33)	(0.32)
Household size > 2	0.10	0.19	0.11	0.14
	(0.30)	(0.39)	(0.32)	(0.35)
Swiss Citizenship	0.82	0.78	0.85	0.86
	(0.38)	(0.42)	(0.35)	(0.34)
Observations	2248	4476	1996	3195

Table 3.8: Summary Statistics Diff-in-Diff Sample

	Men		Women	
	Treatment Mean / SD	Control Mean / SD	Treatment Mean / SD	Control Mean / SD
Age	64.00	61.80	62.42	60.34
	(2.98)	(2.42)	(2.68)	(1.95)
Year of Interview	2001.94	2003.31	2003.16	2002.18
	(4.77)	(4.61)	(4.47)	(4.35)
Age Spouse	65.04	59.91	67.05	61.46
	(2.14)	(1.50)	(1.69)	(1.87)
Education				
Lower Education	0.13	0.14	0.33	0.30
	(0.33)	(0.34)	(0.47)	(0.46)
Secondary Education	0.52	0.47	0.54	0.57
	(0.50)	(0.50)	(0.50)	(0.49)
Higher Education	0.34	0.39	0.13	0.12
	(0.48)	(0.49)	(0.33)	(0.33)
Household size > 2	0.11	0.19	0.12	0.14
	(0.32)	(0.39)	(0.32)	(0.35)
Swiss Citizenship	0.83	0.80	0.86	0.87
	(0.38)	(0.40)	(0.35)	(0.34)
Observations	1267	2728	1138	1855

Appendix 3.C Discussion

Table 3.9: Estimation Intensive Margin Desired Hours Men

	Dependent variable: Desired Hours		
	Tobit	Two-part	Diff-in-diff
Spouse FRA reached	−0.189	−0.123	0.163
	(1.736)	(0.409)	(0.338)
Age dummies	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes
Age spouse	Yes	Yes	Yes
Age spouse squared	Yes	Yes	Yes
Education dummies	Yes	Yes	No
Household size > 2	Yes	Yes	Yes
Swiss citizenship	Yes	Yes	No
Observations	11520	2463	6648

Note: Results intensive margin effect for men with Tobit, two-part, and difference-in-difference estimator. Interviewed individuals and spouses are aged between 58 and 70. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

Table 3.10: Estimation Intensive Margin Desired Hours Women

	Dependent variable: Desired Hours		
	Tobit	Two-part	Diff-in-diff
Spouse FRA reached	−1.261	−0.217	0.030
	(0.865)	(0.307)	(0.274)
FRA reached	−9.708***	−2.978***	−2.156***
	(1.197)	(0.427)	(0.408)
Age dummies	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes
Age spouse	Yes	Yes	Yes
Age spouse squared	Yes	Yes	Yes
Education dummies	Yes	Yes	No
Household size > 2	Yes	Yes	Yes
Swiss citizenship	Yes	Yes	No
Observations	13546	3946	7964

Note: Results intensive margin effect for women with Tobit, two-part, and difference-in-difference estimator. Interviewed individuals and spouses are aged between 58 and 70. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

Table 3.11: Anticipation Effects Extensive Margin

	Dep. variable: Indicator $\mathbb{1}$(weekly working hours > 0)			
	Men		Women	
	APE	SE(APE)	APE	SE(APE)
Spouse FRA reached	0.023	(0.018)	−0.036**	(0.017)
Spouse 1 year younger than FRA	0.019	(0.014)	−0.004	(0.013)
Age dummies	Yes		Yes	
Year dummies	Yes		Yes	
Age spouse	Yes		Yes	
Age spouse squared	Yes		Yes	
Education dummies	Yes		Yes	
Household size > 2	Yes		Yes	
Swiss citizenship	Yes		Yes	
Observations	15683		14643	

Note: Results Probit estimation. Average partial effects (APE) reported. The dummy ”Spouse 1 year younger than FRA” captures the anticipation effect of having a spouse being one year before reaching FRA. Interviewed individuals and spouses are aged between 58 and 70. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

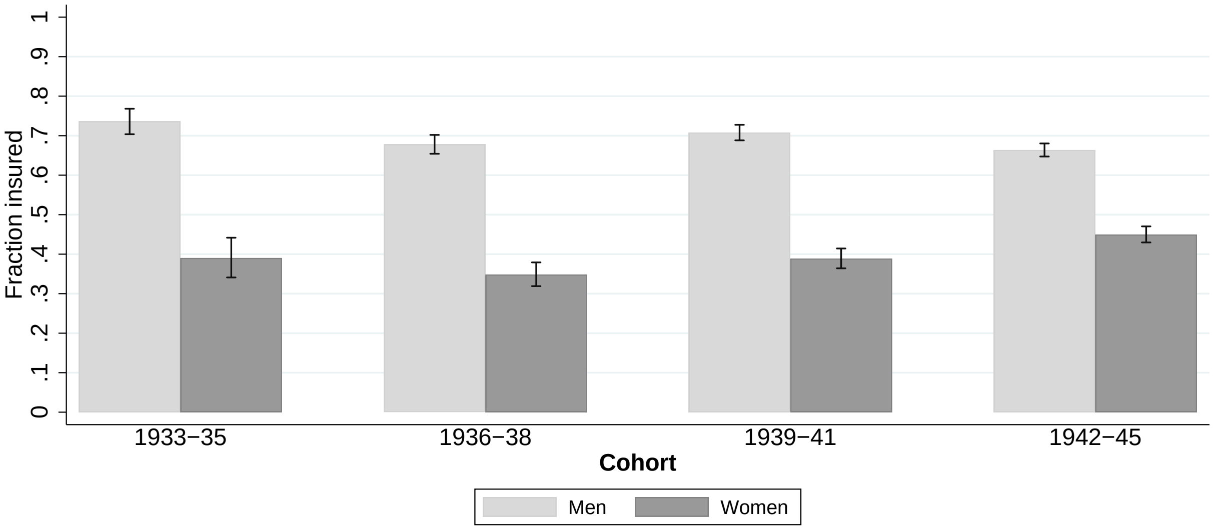
Table 3.12: Delayed Effects Extensive Margin

	Dependent variable: Indicator $\mathbb{1}$(weekly working hours > 0)			
	Men		Women	
	APE	SE(APE)	APE	SE(APE)
Spouse age = FRA	0.012	(0.014)	−0.027**	(0.013)
Spouse age $\geq$ FRA + 1	0.013	(0.018)	−0.047***	(0.017)
Age dummies	Yes		Yes	
Year dummies	Yes		Yes	
Age spouse	Yes		Yes	
Age spouse squared	Yes		Yes	
Education dummies	Yes		Yes	
Household size > 2	Yes		Yes	
Swiss citizenship	Yes		Yes	
Observations	15683		14643	

Note: Results Probit estimation. Average partial effects (APE) reported. Interviewed individuals and spouses are aged between 58 and 70. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

Appendix 3.D Institutional Background

Figure 3.9: Fraction of Individuals Insured in the 2nd Pillar



Note: Fraction of individuals insured in 2nd pillar by gender and cohort. Error bars indicate 95% confidence interval for the mean estimate. Data source: Own calculations based on special module on social security in Swiss Labor Force Survey (SLFS).

Appendix 3.E Sensitivity Analysis

3.E.1 Alternative Model: Linear Probability Model

Table 3.13: Estimation Extensive Margin Linear Probability Model

	Dep. variable: Indicator $\mathbb{1}$(weekly working hours > 0)			
	Men		Women	
	Coef.	SE(Coef.)	Coef.	SE(Coef.)
Spouse FRA reached	0.013	(0.015)	−0.039***	(0.014)
FRA reached			−0.168***	(0.019)
Age dummies	Yes		Yes	
Year dummies	Yes		Yes	
Age spouse	Yes		Yes	
Age spouse squared	Yes		Yes	
Education dummies	Yes		Yes	
Household size > 2	Yes		Yes	
Swiss citizenship	Yes		Yes	
Observations	15683		14643	

Note: Results linear probability model estimation. Interviewed individuals and spouses are aged between 58 and 70. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

3.E.2 Alternative Dependent Variable: Hours Worked Last Week

Table 3.14: Estimation Extensive Margin Hours Worked Last Week

	Dep. variable: $\mathbb{1}$(hours worked last week > 0)			
	Men		Women	
	APE	SE(APE)	APE	SE(APE)
Spouse FRA reached	0.016	(0.014)	−0.027**	(0.013)
FRA reached			−0.116***	(0.016)
Age dummies	Yes		Yes	
Year dummies	Yes		Yes	
Age spouse	Yes		Yes	
Age spouse squared	Yes		Yes	
Education dummies	Yes		Yes	
Household size > 2	Yes		Yes	
Swiss citizenship	Yes		Yes	
Observations	15683		14643	

Note: Results probit estimation. Average partial effects (APE) reported. Interviewed individuals and spouses are aged between 58 and 70. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

Table 3.15: Estimation Intensive Margin Hours Worked Last Week Men

	Dependent variable: Hours worked last week		
	Tobit	Two-part	Diff-in-diff
Spouse FRA reached	0.888 (1.532)	−0.864 (0.706)	−0.789 (0.761)
Age dummies	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes
Age spouse	Yes	Yes	Yes
Age spouse squared	Yes	Yes	Yes
Education dummies	Yes	Yes	No
Household size > 2	Yes	Yes	Yes
Swiss citizenship	Yes	Yes	No
Observations	15093	5927	3189

Note: Results intensive margin effect for men with Tobit, two-part, and difference-in-difference estimator. Interviewed individuals and spouses are aged between 58 and 70. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

Table 3.16: Estimation Intensive Margin Hours Worked Last Week Women

	Dependent variable: Hours worked last week		
	Tobit	Two-part	Diff-in-diff
Spouse FRA reached	−2.131*	−0.151	−1.518
	(1.134)	(0.439)	(1.017)
FRA reached	−11.171***	−3.269***	−3.101**
	(1.487)	(0.627)	(1.399)
Age dummies	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes
Age spouse	Yes	Yes	Yes
Age spouse squared	Yes	Yes	Yes
Education dummies	Yes	Yes	No
Household size > 2	Yes	Yes	Yes
Swiss citizenship	Yes	Yes	No
Observations	14065	4357	2226

Note: Results intensive margin effect for women with Tobit, two-part, and difference-in-difference estimator. Interviewed individuals and spouses are aged between 58 and 70. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

3.E.3 Restricting the Age of the Spouse

Table 3.17: Estimation Extensive Margin

	Dep. variable: Indicator $\mathbb{1}$(weekly working hours > 0)			
	Men		Women	
	APE	SE(APE)	APE	SE(APE)
Spouse FRA reached	0.009	(0.013)	−0.044***	(0.011)
FRA reached			−0.147***	(0.021)
Age dummies	Yes		Yes	
Year dummies	Yes		Yes	
Age spouse	No		No	
Age spouse squared	No		No	
Education dummies	Yes		Yes	
Household size > 2	Yes		Yes	
Swiss citizenship	Yes		Yes	
Observations	7455		8053	

Note: Results probit estimation. Average partial effects (APE) reported. Interviewed individuals are aged between 58 and 70. Spouse of the interviewed person aged between 2 years prior and 2 years after reaching FRA. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

Table 3.18: Estimation Intensive Margin Men

	Dependent variable: Working hours		
	Tobit	Two-part	Diff-in-diff
Spouse FRA reached	0.500	−0.761	−0.738
	(1.217)	(0.571)	(0.811)
Age dummies	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes
Age spouse	Yes	Yes	Yes
Age spouse squared	Yes	Yes	Yes
Education dummies	Yes	Yes	No
Household size > 2	Yes	Yes	Yes
Swiss citizenship	Yes	Yes	No
Observations	7455	3291	1971

Note: Results intensive margin effect for men with Tobit, two-part, and difference-in-difference estimator. Interviewed individuals are aged between 58 and 70. Spouse of the interviewed person aged between 2 years prior and 2 years after reaching FRA. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

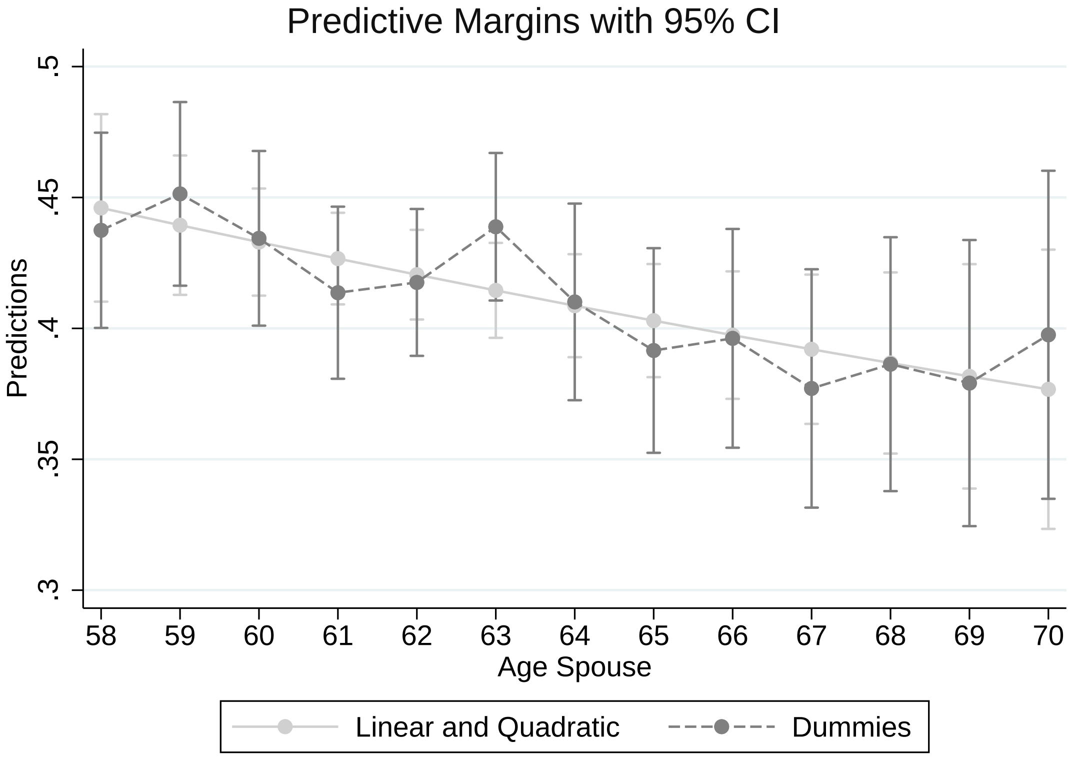
Table 3.19: Estimation Intensive Margin Women

	Dependent variable: Working hours		
	Tobit	Two-part	Diff-in-diff
Spouse FRA reached	−3.204***	−0.894***	−1.163
	(0.817)	(0.314)	(0.854)
FRA reached	−11.298***	−3.305***	−2.249
	(1.729)	(0.778)	(1.511)
Age dummies	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes
Age spouse	Yes	Yes	Yes
Age spouse squared	Yes	Yes	Yes
Education dummies	Yes	Yes	No
Household size > 2	Yes	Yes	Yes
Swiss citizenship	Yes	Yes	No
Observations	8053	2875	1666

Note: Results intensive margin effect for women with Tobit, two-part, and difference-in-difference estimator. Interviewed individuals are aged between 58 and 70. Spouse of the interviewed person aged between 2 years prior and 2 years after reaching FRA. Standard errors bootstrapped (1000 replications) and clustered at the individual level. $^{*}p < 0.1$. $^{**}p < 0.05$. $^{***}p < 0.01$.

3.E.4 Comparison Functional Form for the Age of the Spouse

Figure 3.10: Comparison Functional Form Age Spouse: Linear and Quadratic Term vs. Dummies



Note: Results extensive margin probit model. Predictive margins at different ages of the spouse, and at the mean values of age dummies, year dummies, education dummies, household size > 2 dummy, and Swiss dummy. The light-gray line indicates the predictive margins from a model with age of the spouse included with a linear and a quadratic term, the dark-gray dashed line indicates the predictive margins from a model with age of the spouse included with dummies.