

Essays in Causal Machine Learning

DISSERTATION
of the University of St.Gallen,
School of Management,
Economics, Law, Social Sciences,
International Affairs and Computer Science,
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Michael Zimmert

from

Germany

Approved on the application of

Prof. Dr. Michael Lechner

and

Prof. Dr. Martin Spindler

Dissertation no. 5162

Difo-Druck GmbH, Untersiemaun 2021

Essays in Causal Machine Learning

DISSERTATION
of the University of St.Gallen,
School of Management,
Economics, Law, Social Sciences,
International Affairs and Computer Science,
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Michael Zimmert

from

Germany

Approved on the application of

Prof. Dr. Michael Lechner

and

Prof. Dr. Martin Spindler
Prof. Francesco Audrino, PhD

Dissertation no. 5162

Difo-Druck GmbH, Untersiemaun 2021

The University of St.Gallen, School of Management, Economics, Law, Social Sciences, International Affairs and Computer Science, hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St.Gallen, May 28, 2021

The President:

Prof. Dr. Bernhard Ehrenzeller

Acknowledgments

First and foremost I want to thank Michael Lechner for giving me the opportunity to write this thesis under his supervision. I thank him for excellent advice, permanent approachability and fair, transparent and helpful feedback. Besides his qualities as an academic supervisor, I also want to thank him for creating a well-organized and yet cooperative and friendly working atmosphere at the chair as well as for always communicating working objectives in a highly transparent manner. I am grateful for good cooperation as a co-author and during many other projects.

I thank Franziska Zimmert for constant support, patience with the author of this thesis, being a productive and competent co-author and for much more than what could be written in the acknowledgement section of a thesis.

Further, I want to thank Michael Knaus for motivating me to write a thesis in the field of causal inference, for discussing all sorts of promising and less promising ideas, constant enthusiasm and for giving helpful and supporting advice.

I am also indebted to my colleagues at the Swiss Institute for Empirical Economic Research. In particular, I thank my former office mate Carina Steckenleiter for valuable personal and academic advice throughout my PhD studies. I thank Conny David for pragmatically handling all kinds of organisational issues and for her personal commitment that is way beyond her job description. I also want to thank Daniel Goller, Gabriel Okasa, Anthony Strittmatter, Benedikt Lennartz, Jana Mareckova, Sandro Heiniger, Hugo Bodory and Petyo Bonev for inspiring discussions and for making the institute such a great working place.

Finally, I thank the ‘Studienstiftung des deutschen Volkes’ for enabling me to study at the University of St. Gallen. The financial and ideational support during my master’s degree was the basis of this thesis.

Nänikon, August 2021

Michael Zimmert

Contents

List of Tables	iv
List of Figures	v
Summary	vi
1 Introduction	1
2 Efficient difference-in-differences estimation with high-dimensional common trend confounding	3
2.1 Introduction	4
2.2 Setting and notation	6
2.3 Identification and efficiency bounds	7
2.3.1 Repeated cross-sections	7
2.3.2 Panel data	9
2.4 Estimation and inference	12
2.4.1 High-dimensional data and important building blocks	12
2.4.2 Asymptotic results	13
2.4.3 Plug-in estimators	15
2.4.4 Redundancy of nuisance parameters	16
2.5 Application	19
2.5.1 Estimators considered	20
2.5.2 Placebo tests	21
2.5.3 Main results	22
2.6 Conclusion	22
2.A Proofs for Section 2.3	25
2.A.1 Proof of Theorem 2.1	25
2.A.2 Proof of Corollary 2.1	32
2.A.3 Proof of Theorem 2.2	32
2.A.4 Proof of Corollary 2.3	36
2.B Proofs for Section 2.4	36
2.B.1 Lemma 2.3	36
2.B.2 Proof of Theorem 2.3	37
2.B.3 Proof of Corollary 2.4	40
2.B.4 Proof of Corollary 2.5	43

2.B.5	Proof of Corollary 2.7	43
2.B.6	Proof of Corollary 2.8	43
3	Nonparametric estimation of causal heterogeneity under high-dimensional confounding	44
3.1	Introduction	45
3.2	Methodology	47
3.2.1	Notation	47
3.2.2	Semiparametric efficiency theory	48
3.2.3	Parameter identification	48
3.3	Main results	49
3.3.1	GATE estimation	49
3.3.2	Joint estimation of GATE and ATE	53
3.4	Implementation choices	55
3.5	Empirical Monte Carlo experiments	56
3.5.1	Results for GATE and ATE	57
3.5.2	LOOCV bandwidth selection	58
3.6	Illustrative example	58
3.6.1	Empirical results	59
3.6.2	Comparison with other estimators	61
3.7	Conclusion	62
3.A	Proof of Theorems	65
3.A.1	Proof of Theorem 3.1	66
3.A.2	Proof of Theorem 3.2	69
3.B	Details on the bandwidth ranges	72
3.C	Further results of the Empirical Monte Carlo experiment	73
3.C.1	Further results for different estimators	73
3.C.2	Further results on the bandwidth choice	75
3.D	Details on the empirical example	76
3.D.1	Covariates in the dataset	76
3.D.2	Additional sensitivity analysis	77
4	Paid parental leave and maternal reemployment: Do part-time subsidies help or harm?	78
4.1	Introduction	79
4.2	Institutional background	80
4.2.1	Regulations of parental benefits prior to the reform in 2015 (<i>Elterngeld EG</i>)	82
4.2.2	The reform in July 2015 (<i>ElterngeldPlus EG+</i>)	82
4.2.3	The expansion of subsidised child care	83
4.3	Theoretical effects of part-time subsidies	84
4.4	Empirical strategy	85
4.4.1	Identification	85
4.4.2	Estimation of average effects	87
4.4.3	Estimation of heterogeneous effects	89

4.5	Data	90
4.6	Results	94
4.6.1	Estimation results for ATET and sensitivity analysis	94
4.6.2	Estimation results for conditional effects	98
4.7	Conclusion	99
4.A	Figures and tables	104
4.B	Technical details for the proposed estimators	109
	Bibliography	112
	Curriculum Vitae	119

List of Tables

2.1	Covariate specifications	19
2.2	Subsample sizes	20
2.3	Placebo tests 1988/89 under (CS-5)	21
2.4	Placebo tests 1988/89 under (CS-1)-(CS-4)	23
2.5	Results for (CS-1)-(CS-4)	23
3.1	Performance of different GATE estimators	58
3.2	Performance of different ATE estimators	59
3.3	Averaged ATE estimators	60
3.4	Performance of different GATE estimators for $\alpha = 2$	73
3.5	Performance of different ATE estimators for $\alpha = 2$	74
3.6	Description of covariates in the dataset	76
4.1	Parental leave regulations over time	81
4.2	Calculation of benefit amount <i>Elterngeld</i> (EG) and <i>ElterngeldPlus</i> (EG+)	84
4.3	Treatment definition	85
4.4	Descriptive statistics of variables by treatment and control group	92
4.5	ATETs for job continuity and accumulated earnings	96
4.6	Large and small bandwidth results for job continuity and accumulated earnings	104
4.7	Conditional effects for job continuity and accumulated earnings	108

List of Figures

3.1	GATE and ATE MSEs for different factors of the LOOCV bandwidth choice	60
3.2	AIPW GATE estimator with ensemble first stages	61
3.3	Sensitivity to bandwidth choice (age)	62
3.4	Other GATE estimators	63
3.5	GATE and ATE MSEs for different factors of the LOOCV bandwidth choice with 10 % treatment share	75
3.6	Sensitivity to bandwidth choice (care visits)	77
4.1	Monthly share of births in 2014 and 2015	87
4.2	Outcome means of mothers in the treatment group before reform	94
4.3	Baseline estimation results	95
4.4	Estimation results of placebo reform	97
4.5	Estimation results for subsequent birth within next 24 months	97
4.6	Estimation results for middle income group	99
4.7	Estimation results for high income group	100
4.8	Estimation results for prior full-time	101
4.9	Estimation results for West Germany	102
4.10	Imputed and official birth numbers in 2015	104
4.11	Propensity scores by treatment status	105
4.12	Estimation results with small bandwidth	106
4.13	Estimation results with large bandwidth	107

Summary

The thesis comprises three essays in the field of causal econometrics. Each of the single essays is dedicated to the question how supervised machine learning algorithms can be integrated in standard causal designs. A first, introductory chapter puts the three following main chapters into perspective. The second chapter studies the link between semiparametric efficiency theory and causal machine learning in semiparametric difference-in-differences estimation. The third chapter investigates nonparametric and machine learning methods for estimating heterogeneous treatment effects. The last chapter analyses the implications of a parental leave reform in Germany using semiparametric difference-in-differences.

A major finding from Chapter 2 is that the variance lower bound that is asymptotically attainable for semiparametric difference-in-differences estimators crucially depends on the assumptions imposed on the statistical model. It turns out that lower efficiency bounds can be achieved under more restrictive assumptions. This implies a robustness-efficiency trade-off. Further, the derived efficiency results are useful for integrating machine learning methods in semiparametric difference-in-differences estimation. In particular, it is shown that moment conditions implied by the efficient influence functions lead to plug-in estimators that maintain desirable statistical properties like asymptotic normality and efficiency even when machine learning algorithms are used to estimate first stage nuisance parameters.

Chapter 3 studies nonparametric heterogeneous treatment effect estimation when the researcher is interested in effect heterogeneity with respect to a limited number of covariates. However, it turns out that estimation of first stage nuisance parameters with machine learning methods remains feasible even if the dimension of the covariates needed to account for confounding is high. In particular, coupled convergence conditions between the machine learning first stage and the nonparametric second stage are derived. Additionally, it is shown that averaging over the heterogeneous treatment effects estimators leads to an estimator for the average treatment effect that attains the semiparametric efficiency bound. The finite sample performance of the different estimators is compared in various Empirical Monte Carlo experiments.

Chapter 4 finds that a parental leave reform in Germany encourages mothers to return to the labour market earlier. As intended by the reform, mothers are mostly encouraged to return in part-time. The semiparametric difference-in-differences approach used allows to estimate first stage nuisance parameters with an ensemble learner. The empirical results are robust against the inclusion of a large number of covariates that are available from administrative data. Additionally, an identification result for heterogeneous treatment effects in semiparametric difference-in-differences is derived. It turns out that mothers with higher opportunity costs of working part-time have lower incentives to return to the labour market earlier.

Zusammenfassung

Die vorliegende Thesis umfasst drei Essays aus dem Bereich der kausalen Ökonometrie. Jedes der einzelnen Kapitel ist der Frage gewidmet wie maschinelles Lernen für Standardprobleme der Kausalanalyse verwendet werden kann. Ein erstes, einführendes Kapitel veranschaulicht die Zusammenhänge zwischen den drei folgenden Hauptkapiteln. Das zweite Kapitel untersucht die Verknüpfung zwischen semiparametrischer Effizienztheorie und kausalem maschinellem Lernen in Bezug auf semiparametrische Differenzen-von-Differenzen Schätzung. Das dritte Kapitel beleuchtet Methoden aus dem Bereich der Nichtparametrik und des maschinellen Lernens zur Schätzung heterogener Treatment-Effekte. Das letzte Kapitel analysiert, unter Verwendung eines semiparametrischen Differenzen-von-Differenzen Ansatzes, die Implikationen einer Elterngeldreform in Deutschland.

Die Abhängigkeit der niedrigst möglichen asymptotischen Varianz, die in semiparametrischen Differenzen-von-Differenzen Ansätzen erreicht werden kann, von den Annahmen des statistischen Modells ist eine wesentliche Erkenntnis aus Kapitel 2. Es stellt sich heraus, dass niedrigere Effizienzgrenzen mit restriktiveren Annahmen einhergehen. Dies impliziert einen Zielkonflikt zwischen der Robustheit des Modells und der niedrigst möglichen, asymptotisch erreichbaren Varianz. Desweiteren können die hergeleiteten Effizienzergebnisse zur Integration von Methoden des maschinellen Lernens in semiparametrische Differenzen-von-Differenzen Ansätze benutzt werden. Insbesondere wird gezeigt, dass Momentenbedingungen, die durch effiziente Influence-Funktionen impliziert werden, zu mehrstufigen Schätzverfahren führen, die wünschenswerte statistische Eigenschaften wie asymptotische Normalität und Effizienz behalten – selbst wenn maschinelles Lernen zur Schätzung der ersten Stufe benutzt wird. Kapitel 3 betrachtet die nichtparametrische Schätzung heterogener Treatment-Effekte in Situationen wo Effektheterogenität nur für eine begrenzte Anzahl von Kovariaten von Interesse ist. Jedoch stellt sich heraus, dass die Schätzung von Parametern mit Methoden des maschinellen Lernens auf der ersten Stufe selbst dann möglich bleibt, wenn die Dimension derjenigen Kovariaten, die nötig sind um für Selbstselektion zu kontrollieren, hoch ist. Insbesondere werden kombinierte Konvergenzraten zwischen den Methoden des maschinellen Lernens auf der ersten Stufe und der nichtparametrischen zweiten Stufe hergeleitet. Zusätzlich wird gezeigt, dass die Mittelwertbildung über die Effektheterogenitätsschätzer zu Schätzern für mittlere Effekte führt, die asymptotisch die niedrigst mögliche Varianz erreichen. Die Eigenschaften der verschiedenen Schätzer in kleinen Stichproben werden mittels empirischer Monte Carlo-Simulation verglichen.

Kapitel 4 zeigt, dass eine Elterngeldreform in Deutschland Mütter dazu bewegt frühzeitig auf den Arbeitsmarkt zurückzukehren. Wie durch die Reform intendiert, werden Mütter hauptsächlich dazu bewegt in Teilzeit zurückzukehren. Der verwendete semiparametrische Differenzen-von-Differenzen Ansatz ermöglicht die Schätzung von Parametern auf der ersten Stufe mittels eines Ensembleverfahrens. Die empirischen Ergebnisse sind gegenüber einer Vielzahl von verwendeten Kovariaten, die aus administrativen Daten verfügbar sind, robust. Zusätzlich wird ein Identifikationsergebnis für heterogene Treatment-Effekte in semiparametrischen Differenzen-von-Differenzen Ansätzen hergeleitet. Es stellt sich heraus, dass Mütter mit höheren Opportunitätskosten in Teilzeit zu arbeiten weniger starke Anreize haben früher auf den Arbeitsmarkt zurückzukehren.

Chapter 1

Introduction

Methods from the field of ‘machine learning’ have shown immense potential to uncover causal effects in a systematic manner. The successful incorporation of these methods in standard tools for causal inference, however, also leads to new statistical and practical challenges. Paradoxically, state-of-the-art ‘causal machine learning’ algorithms revived some classical frameworks from semiparametric theory that were already studied in the late eighties of the previous century. In particular, it turns out that there is a close connection between semiparametric efficiency theory and the rate double robustness or orthogonality properties needed to obtain causal machine learning estimators that have desirable asymptotic properties. This thesis is dedicated to study the theoretical properties of causal machine learning methods for different semiparametric causal inference frameworks and to apply these algorithms to answer empirical questions in the field of labour economics.

The thesis is not a ‘syn’-thesis in the sense that the single chapters build on one another. Rather, it is a collection of three chapters that can and probably should be read independently. Especially Chapters 2 and 3 share some common features. They are both concerned with the theoretical development of estimators for different statistical problems. Chapter 4 is an empirical study applying causal machine learning algorithms to a real case in the field of labour economics. It is inspired by the methods developed in the two preceding chapters.

The first, single-authored essay derives efficiency bounds for semiparametric difference-in-differences estimators considering different classes of data-generating processes. It turns out that unambiguously lower efficiency bounds may in principle be attained under more restrictive assumptions imposed on the semiparametric model. This leads to the conclusion that for semiparametric difference-in-differences models there is a robustness-efficiency trade-off. Further, under comparable assumptions it is shown that lower efficiency bounds can be achieved when having access to panel rather than cross-sectional data. The efficient influence functions for the different settings can be used to obtain moment conditions for the identified parameter of interest. This leads to plug-in estimators that allow to use supervised machine learning algorithms for first stage nuisance parameter estimation. Therefore, the study represents an example how standard semiparametric efficiency theory can be used to derive plug-in estimators that are both semiparametrically efficient *and* allow for machine learning first stage predictions. The study, however, also shows that while the links between efficiency theory and causal machine learning are close, the approach of deriving rate double robust estimators from efficient influence functions might have some limitations. Firstly, not all efficient influence functions imply moment conditions that work under primitive double robust rate conditions. Secondly, there might be some estimators that have the desired rate double robustness properties but do not achieve the variance lower bound. Thirdly, since semiparametric efficiency theory only applies to regular

problems, this kind of analysis is limited to rather standard, average treatment effect type of settings.

The latter limitation is the starting point of the analysis in Chapter 3. In contrast to Chapter 2, the non-regular problem of estimating group average treatment effects in a typical selection-on-observables framework is studied. A large number of covariates might be needed in order for the identifying assumption to be satisfied. Again estimators with a rate double robustness feature allow to use flexible machine learning algorithms to estimate first stage nuisance parameters. Here, semiparametric efficiency theory is only indirectly helpful for deriving estimators with a rate double robustness feature. However, by noticing that aggregating over group average treatment effects leads to average effects, one may use the efficient influence function of the underlying average treatment effect problem to derive rate double robust estimators for heterogeneous effects. It turns out that coupling nonparametric regression asymptotics with first stage machine learning estimators is feasible under quiet weak convergence rate conditions. Further, it is shown that averaging over this type of nonparametric group average treatment effect estimators leads to an average treatment effects estimator that asymptotically attains the semiparametric efficiency bound. This last theoretical result is conceptually interesting as it links back non-regular rate double robustness problems to semiparametric efficiency theory.

The estimation strategies suggested in Chapters 2 and 3 share the common issue of being computationally demanding. For example, in order to obtain difference-in-differences estimators that obey the rate double robustness feature and attain the semiparametric efficiency bound, one may have to estimate up to seven first stage nuisance parameters with sophisticated machine learning algorithms. The fact that several sample splitting algorithms have to be implemented in order to guarantee certain asymptotic properties, magnifies the problem. Hence, the methods developed in Chapters 2 and 3 are particularly designed for datasets with potentially very many variables but only moderately large sample sizes.

The application in Chapter 4 evaluates a parental leave reform using semiparametric difference-in-differences methods with flexibly estimated machine learning first stages in a large sample. It follows from Chapter 2 that the estimator used in the application is not semiparametrically efficient. In contrast to the semiparametrically efficient alternative, however, it remains practically feasible because it only relies on two first stage nuisance parameters. Further, heterogeneous effects are estimated for discrete groups such that one does not rely on the computationally demanding nonparametric regression step as suggested in Chapter 3. Hence, Chapter 4 reminds us that nice theoretical properties may only be partially usable in practice due to computational constraints.

The preceding discussion indicates that this thesis may have to be defended from itself. However, one may also take ‘essays’ in the title of this thesis literally and consider it as a series of attempts to make the most out of data and to adapt the empirical methods used to changing conditions in specific causal designs.

Chapter 2

Efficient difference-in-differences estimation with high-dimensional common trend confounding

This study considers various semiparametric difference-in-differences models under different assumptions on the relation between the treatment group identifier, time and covariates for cross-sectional and panel data. The variance lower bound is shown to be sensitive to the model assumptions imposed implying a robustness-efficiency trade-off. The obtained efficient influence functions lead to estimators that are rate double robust and have desirable asymptotic properties under weak first stage convergence conditions. This enables to use sophisticated machine learning algorithms that can cope with settings where common trend confounding is high-dimensional. The usefulness of the proposed estimators is assessed in an empirical example. It is shown that the efficiency-robustness trade-offs and the choice of first stage predictors can lead to divergent empirical results in practice.

JEL classification: C14, C21

Keywords: Semiparametric difference-in-differences, machine learning, semiparametric efficiency bound, high-dimensional data, employment protection.

Acknowledgements: I want to thank Michael Lechner, Michael Knaus, Bryan Graham, Anthony Strittmatter, Beatrix Eugster, Pedro Sant’Anna and Petyo Bonev for competent advice and Franziska Zimmert for motivating the research question. I also thank Gabriel Okasa for some very useful pieces of R code and Daniel Goller, Carina Steckenleiter and Jana Mareckova for helpful remarks and comments. Further, I want to thank all participants of the European Causal Inference Meeting 2019 in Bremen for constructive feedback and fruitful discussions.

Financial support from the Swiss National Science Foundation (SNSF) is gratefully acknowledged. The study is part of the project ‘Causal Analysis with Big Data’ of the Swiss National Research Programme ‘Big Data’ (NRP 75).

2.1 Introduction

In difference-in-differences identification is ensured by the fact that the subpopulation that will be exposed to the treatment (treatment group) and the subpopulation that will not be exposed to the treatment (control group) would have developed equally in the absence of treatment.¹ Identification often relies on the assumption that the common trend holds conditional on covariates. Any underlying factor differently shifting the potential outcomes under non-treatment for the treatment group and the control group needs to be controlled for. However, even if the researcher can credibly identify the factors that may lead to common trend confounding, it is still unclear in what form covariates should ultimately enter the statistical model for several reasons.

Crucially, the statistical model depends on assumptions about the relation between the treatment group identifier, time and observed covariates. With cross-sectional data covariates might be needed to account for imbalances between treatment and control group *and* across time. With some notable exceptions (Lechner, 2011; Hong, 2013; Stuart, Huskamp, Duckworth, Simmons, Song, Chernew, and Barry, 2014; Lu, Nie, and Wager, 2019) most studies in semiparametric difference-in-differences exclude time-varying treatment group compositions and covariates (e.g., Heckman, Ichimura, and Todd, 1997; Abadie, 2005; Sant’Anna and Zhao, 2020; Chang, 2020). This study investigates semiparametric difference-in-differences models under various assumptions on how covariates, time and treatment group composition are related. Efficient influence functions are derived under more or less restrictive assumptions and for different sampling schemes. We present various identification and efficiency results for low-dimensional semiparametric difference-in-differences models. Our results are sensitive to assumptions about how covariates enter the model. In particular, our results hint at a trade-off between the strength of the assumptions the researcher is willing to impose on the model and the efficiency bound that can be achieved under such assumptions. Further, our results suggest that there are cases where we might want to include covariates even if not needed for identification, as they could increase the precision of some of the estimators. A comparison of the efficiency bounds for repeated cross-sections and panel data allows to draw interesting conclusions about the efficiency loss when panel data is not available. We therefore contribute to the literature on semiparametric efficiency in causal inference settings (e.g., Hahn, 1998; Firpo, 2007; Frölich, 2007; Chen, Hong, and Tarozi, 2008; Cattaneo, 2010; Graham, Pinto, and Egel, 2016; Lee, 2018). Such an analysis is typically based on the approach developed by Newey (1990, 1994b) and Bickel, Klaassen, Ritov, and Wellner (1993). Chamberlain (1987, 1992) contributes an alternative approach based on moment conditions. Graham (2011) establishes an equivalence result between the moment condition based approach and the approach of Bickel et al. (1993) for the general missing data problem. In parallel work, Sant’Anna and Zhao (2020) also consider efficiency theory for semiparametric difference-in-differences problems to derive efficient score functions. Their results crucially rely on a relatively strong stationarity assumption and are included in this study as a special case. We also notice that a previous version of the present study was the first that proposed efficiency bounds for the semiparametric difference-in-differences problem using Graham’s (2011) equivalence result. It turns out that for the panel case the moment conditions exhaust all the information in the identifying assumptions while for the cross-sectional case they do not.²

The efficient influence functions derived imply plug-in estimators that allow to combine semiparametric difference-in-differences models with very flexible first stage estimators. This is important because there might be many

¹For textbook treatments see for example Athey and Imbens (2017), Lechner (2011) or Imbens and Wooldridge (2009).

²We do not follow Graham’s (2011) approach in this version. However, we consider the insufficiency of the first and second stage moment conditions to exhaust all information necessary to derive the efficiency bound for the cross-sectional difference-in-differences case as an interesting topic for further research.

different covariates that are supposed to measure the same economic channel for common trend confounding and it might be unclear in what functional form the covariates should be included in the model. These issues might be especially prevalent in difference-in-differences models. Often covariates like geographic or industry classifications are available for different levels of aggregation – making covariate selection an even more tedious task. For standard parametric models usually used for difference-in-differences estimation (e.g., Card, 1990; Card and Krueger, 1994; Eissa and Liebman, 1996) or semiparametric models with parametric or nonparametric first stages (e.g., Abadie, 2005; Sant’Anna and Zhao, 2020) a high-dimensional covariate space will cause the estimator to break down. Advances in the supervised machine learning literature³ showed an immense potential to approach this problem by choosing a data-driven trade-off between the covariate dimension and the sample size at hand and were successfully integrated in causal inference settings (e.g., Belloni, Chen, Chernozhukov, and Hansen, 2012; Zhang and Zhang, 2014; van de Geer, Bühlmann, Ritov, and Dezeure, 2014; Belloni, Chernozhukov, and Hansen, 2014; Athey, Imbens, and Wager, 2018). This study builds on the generic ‘double machine learning’ framework developed in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018a). A major insight from this work is that ‘single’-robust estimators based on the treatment mechanism (e.g., Horvitz and Thompson, 1952; Hirano, Imbens, and Ridder, 2003; Hahn and Ridder, 2013) or outcome based models (e.g., Hahn, 1998) are inappropriate with machine learning generated first stages while ‘double’-robust estimators (Robins, Rotnitzky, and Zhao, 1994; Scharfstein, Rotnitzky, and Robins, 1999) maintain good statistical properties. Chernozhukov et al. (2018a) show that the (rate) double robustness properties can be used such that there is no effect of first stage nuisance parameter estimation under relatively weak convergence conditions of the first stage parameters. We modify and extend the double machine learning framework in this study. In contrast to Chernozhukov et al. (2018a), this study does not rely on high-level conditions based on Gateaux differentiation to verify the rate double robustness properties. Alternatively, we provide easy-to-check conditions that cover a broad range of scores typically used in the causal inference literature. Focusing on a specific class of widely used score functions allows to derive generalisable convergence condition requirements. Crucially, this substantially reduces the computational burden when deriving the asymptotic properties of an estimator and requires fewer regularity conditions. Further, some of the derived efficient influence functions imply a new class of plug-in estimators whose convergence conditions depend on the existence of higher-order moments of the outcome. These results do not trivially follow from existing theory (e.g., Chernozhukov et al., 2018a, Section 5). Our theoretical results on double machine learning is therefore also useful beyond the scope of difference-in-differences estimation. Additionally, they allow us to derive first stage convergence conditions for different semiparametric difference-in-differences estimators. This enables to incorporate sophisticated supervised machine learning algorithms that can cope with settings where the dimension of the covariate space is potentially higher than the sample size. Plug-in estimators that follow from the derived efficient influence functions are shown to achieve the low-dimensional variance lower bound. Our results also indicate that for some cases there is a trade-off between estimation robustness and efficiency. Throughout the evolution of this study other related but independent work on semiparametric difference-in-differences estimation with machine learning appeared. Chang (2020) also considers difference-in-differences estimation under the strong stationarity assumption by directly applying the double machine learning results of Chernozhukov et al. (2018a). His estimator does generally not attain the semiparametric efficiency bound. Lu et al. (2019) propose estimators that are robust against the distortion of the stationarity assumption but focus on another parameter, consider alternative estimation methods and do not derive efficiency results.

Finally, we adopt the methods to a well-known application and investigate whether the efficiency-robustness trade-

³For an overview see e.g. Hastie, Tibshirani, and Friedman (2009).

offs matter in practice and the value added when using machine learning methods instead of standard parametric estimation methods.

The following section introduces the setting and required notation. The development of semiparametric theory for the difference-in-differences problem will be the starting point of our analysis in Section 2.3. Section 2.4 presents our results on estimation and inference. To assess the usability of the proposed methods, they are applied to real world data in Section 2.5. The last section concludes. Most technical proofs are relegated to the Appendix.

2.2 Setting and notation

Random variables like A are denoted by capital letters. They have realizations $A = a$ that are in the support of the random variable $\mathcal{A}$. $A = a$ has density $f_A(a)$. If A is discrete, we write $Pr(A = a) = f_A(a)$ as a shorthand. The cumulative distribution function is given by $F_A(a)$. Let B be another random variable. Then independence between two random variables A and B is denoted by $A \perp B$. The expectation operator is defined by $\mathbb{E}$ and Var is used as a shorthand for the variance. For a generic function $g = g(B) = g_A(B)$, we use A in the subscript to remind us of the mapping $g : b \mapsto a$. The L_p norm is denoted by $\|g(B)\|_p$. As a special case, we use $\sup_{b \in \mathcal{B}} |g(b)|$ and $\|g(B)\|_\infty$ interchangeably to denote the supremum of the function. The infimum is $\inf_{b \in \mathcal{B}} |g(b)|$. Let β be some parameter, then we denote $\dot{g}(B, \beta) = \frac{\partial g(B, \beta)}{\partial \beta}$. Generically, $C > 0$ denotes a constant.

Let D , T and G_τ be binary indicator variables such that $d, t, g_\tau \in \{0, 1\}$ where $\tau \in \mathcal{T}$ and either $\mathcal{T} = \{(d, t)\}$ or $\mathcal{T} = \{d\}$. In particular, $D = 1$ for observations that belong to the treatment group, $T = 1$ for observations that are observed in period 1 and $G_{d,t} = 1$ if $D = d$ and $T = t$ and 0 otherwise and $G_d = 1$ if $D = d$ and 0 otherwise.⁴ Denote the outcome variable by Y and some further observed variables by X . We follow the established literature (e.g., Roy, 1951; Rubin, 1974) and let $Y^d(t)$ be the potential outcome variable that contains the potentially unobserved realizations of Y for the state $D = d$ and $T = t$. The exposition additionally relies on the definition of some conditional expectations. In particular, we have $m_Y(d, t, x) = \mathbb{E}[Y|D = d, T = t, X = x]$ and $m_Y(x) = \sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} m_Y(d, t, x)$. Similarly, for $\Delta Y = Y(1) - Y(0)$ we have $m_{\Delta Y}(d, x) = \mathbb{E}[Y(1) - Y(0)|D = d, X = x]$ and $m_{\Delta Y}(x) = m_{\Delta Y}(1, x) - m_{\Delta Y}(0, x)$. Additionally, suppose that A is a binary variable. Then we generically define the probabilities $p_{A=a}(b) = Pr(A = a|B = b)$, $p_A(b) = Pr(A = 1|B = b)$, $p_{A=a} = Pr(A = a)$ and $p_A = Pr(A = 1)$. For example, we have $p_{D=d, T=t}(x) = Pr(D = d, T = t|X = x)$ and $p_{DT} = Pr(D = 1, T = 1)$. Note that the definition of G_τ allows us to flexibly write for example $m_Y(d, t, x) = m_Y(G_{d,t} = 1, x)$ and similarly for the other parameters. In difference-in-differences settings the researcher is generally interested in identifying the parameter

$$\theta = \mathbb{E}[Y^1(1) - Y^0(1)|D = 1, T = 1].$$

It can be described as an average treatment effect on the treated (ATET) because the parameter is defined for those who actually receive the treatment ($D = 1, T = 1$). An average population effect cannot be identified because this would require a subpopulation for which the treatment vanishes between period $T = 0$ and $T = 1$ (for a discussion on this see Lechner, 2011). We also note that when panel data is available, Abadie (2005) shows that $\mathbb{E}[Y^1(1) - Y^0(1)|D = 1]$ is identified. Notice that, without further assumptions, this parameter is more general than the ATET. It is an average treatment effect for the treatment group. Intuitively, panel data ensures that the composition of the treatment group does not depend on T and so $\mathbb{E}[Y^1(1) - Y^0(1)|D = 1, T = 0] =$

⁴Obviously for the latter case $G_d = D$. We introduce this notation to formulate results as general as possible throughout our exposition.

$$\mathbb{E}[Y^1(1) - Y^0(1)|D = 1, T = 1].$$

2.3 Identification and efficiency bounds

2.3.1 Repeated cross-sections

Assumption 2.1 (Data-generating process CS). *Let $W(t) = (Y(t), D(t), T = t, X(t))$. (i) The i.i.d. samples of $W(0)$ and $W(1)$ with sample sizes $N(0)$ and $N(1)$ are observed; (ii) The merged sample of size $N(0) + N(1) = N$ with observations $i = 1, \dots, N$ is drawn from $W = \{W(0), W(1)\} = (Y, D, T, X)$ and the distribution $F_{\{W(0), W(1)\}}(\{w(0), w(1)\}) = F_W(w)$ exists.*

Assumption 2.1 describes the data-generating process (DGP) for the repeated cross-sections. It guarantees that we can use the pseudo-sample W_i with observations $i = 1, \dots, N$ and emphasises that we have to cope with a merged sample problem where the sample sizes of $W(0)$ and $W(1)$, $N(0)$ and $N(1)$, obey $\frac{N(0)}{N(1)} \rightarrow C$. For a related discussion on this see Abadie and Imbens (2006) and Graham et al. (2016). Also notice that the existence of the joint distribution implies that the dimension of X is fixed.

In what follows we discriminate five settings (CS-1) to (CS-5) that describe different assumption sets on the relation between D , T and X . In (CS-1) we do not make any further assumptions. It is the fully robust setting. Settings (CS-2)-(CS-5) are comprised in Assumption 2.2.

Assumption 2.2 (Relation of D , T and X). *The variables D , T and X are assumed to be related in the following ways. (CS-2) Conditional independence of D and T , $D \perp T | X = x$ for every $x \in \mathcal{X}$; (CS-3) Independence of X and T , $X \perp T$; (CS-4) Joint independence of D and X from T , $(D, X) \perp T$; (CS-5) Mutual independence of D , T and X , $(D \perp T \perp X)$.*

(CS-2) allows for time varying D and X , but requires that all time variation in D is fully captured by X . (CS-3) does not allow for time-varying X , but D may still follow a time trend. (CS-4) implies the strong stationarity assumption used by Abadie (2005), Sant’Anna and Zhao (2020) and Chang (2020). It excludes time variation in D and X . (CS-5) is the ‘experimental’ setting. Even though the outcome might depend on X , there are no imbalances neither between treatment and control group nor across time.

Since Assumption 2.2 only contains conditions on the relation of observed random variables, it is in principal testable. Especially, a significant correlation between D and T rules out settings (CS-4) and (CS-5). Also notice that (CS-2) and (CS-3) are mutually exclusive and that the restrictiveness of the assumptions can be ordered as (CS-5), (CS-4), (CS-3)/(CS-2), (CS-1).

Identification

Assumption 2.3 (Identification CS). *For any $d, t \in \{0, 1\}$ and $x \in \mathcal{X}$,*

(i) (Observational Rule) For each observation i , the outcome

$$Y_i = D_i (T_i Y_i^1(1) + (1 - T_i) Y_i^1(0)) + (1 - D_i) (T_i Y_i^0(1) + (1 - T_i) Y_i^0(0))$$

is observed;

(ii) (Common Support) The propensity score $p_{D=d, T=t}(x)$ is bounded away from zero;

(iii) (No Anticipation) $\mathbb{E}[Y^1(0) - Y^0(0)|D = 1, T = 0, X = x] = 0$;

(iv) (Conditional Common Trends)

$$\begin{aligned} & \mathbb{E}[Y^0(1)|D = 0, T = 1, X = x] - \mathbb{E}[Y^0(0)|D = 0, T = 0, X = x] \\ &= \mathbb{E}[Y^0(1)|D = 1, T = 1, X = x] - \mathbb{E}[Y^0(0)|D = 1, T = 0, X = x]. \end{aligned}$$

Assumption 2.3 yields an identification result for θ with cross-sectional data. The Observational Rule underscores that we only observe Y and not $Y(0)$ and $Y(1)$ for every observation. It rules out that observations in the treatment group can be part of the control group or that observations in period $T = 0$ are again observed in $T = 1$. Notice that since we consider a pseudo-sample, this does *not* rule out individuals from an actual population from being re-sampled in the second cross-section. Common Support is necessary to guarantee the existence of conditional outcome expectations. The No Anticipation condition rules out an effect of the treatment in period $T = 0$ for the treatment group. The Conditional Common Trends condition requires that conditional on the covariates the treatment and the control group would have developed equally in the absence of the treatment. To allow for a more compact representation of the results, some further notation is introduced. Let $q_{CS;D=d, T=t}(X)$ denote the conditional probability function $p_{D=d, T=t}(X)$ under some of the specific assumptions on the relation of D , T and X in (CS-1) to (CS-5). For example, we have $q_{CS-2;D=1, T=1}(X) = p_D(X)p_T(X)$. Equivalently, denote by $q_{CS;DT}$ the unconditional probability p_{DT} under some specific assumption (CS-1) to (CS-5).

Lemma 2.1. *Under Assumptions 2.1 and 2.3, the parameter $\theta = \mathbb{E}[Y^1(1) - Y^0(1)|D = 1, T = 1]$ is identified as $\mathbb{E}\left[m_Y(X) \frac{q_{CS;D=1, T=1}(X)}{q_{CS;DT}}\right]$.*

Proof: Notice that

$$\begin{aligned} \mathbb{E}[Y^1(1) - Y^0(1)|D = 1, T = 1, X = x] &= m_Y(1, 1, x) - \mathbb{E}[Y^0(0)|D = 1, T = 0, X = x] \\ &\quad - \mathbb{E}[Y^0(1)|D = 0, T = 1, X = x] + \mathbb{E}[Y^0(0)|D = 0, T = 0, X = x] \\ &= m_Y(1, 1, x) - \mathbb{E}[Y^1(0)|D = 1, T = 0, X = x] \\ &\quad - m_Y(0, 1, x) + m_Y(0, 0, x) \\ &= m_Y(x) \end{aligned}$$

using Assumption 2.3. Further, for the conditional density function $f_{X|D, T}(x|d, t)$

$$\theta = \int m_Y(x) f_{X|D, T}(x|1, 1) dx = \int m_Y(x) \frac{p_{D=1, T=1}(x)}{p_{DT}} f_X(x) dx = \int m_Y(x) \frac{q_{CS;D=1, T=1}(x)}{q_{CS;DT}} f_X(x) dx.$$

Semiparametric efficiency bounds

Theorem 2.1 (Semiparametric efficiency bounds CS). *Suppose that Assumptions 2.1 and 2.3 hold. Then under each of the settings in (CS-1)-(CS-5) the efficient influence function is given by*

$$\psi_{CS}^*(W; \theta) = \frac{q_{CS;D=1, T=1}(X)}{q_{CS;DT}} \psi_{CS}^{*a}(W) + \psi_{CS}^{*b}(W) (m_Y(X) - \theta)$$

where

$$\psi_{CS}^{*a}(W) = \sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \frac{G_{d,t}}{q_{CS;D=d,T=t}(X)} (Y - m_Y(d,t,X))$$

and $\psi_{CS-1}^{*b}(W) = \frac{DT}{p_{DT}}$, $\psi_{CS-2}^{*b}(W) = \frac{Dp_T(X) + p_D(X)T - p_D(X)p_T(X)}{p_{DT}}$, $\psi_{CS-3}^{*b}(W) = \frac{T(D - p_D(1,X)) + p_D(1,X)p_T}{p_D(1)p_T}$, $\psi_{CS-4}^{*b}(W) = \frac{D}{p_D}$ and $\psi_{CS-5}^{*b}(W) = 1$. The semiparametric efficiency bound for settings (CS-1) to (CS-5) is $\mathbb{E} [\psi_{CS}^*(W; \theta)^2]$.

Proof: see Appendix 2.A.1.

Theorem 2.1 is our first main result. Under each of the settings (CS-1)-(CS-5) an efficient influence function with a different adjustment term $\psi_{CS}^{*b}(W)$ is derived implying different efficiency bounds for the different settings. An important implication is summarised in Corollary 2.1.

Corollary 2.1. *In terms of the asymptotic variance bound, the value of knowing that one of the assumed relations of D , T and X in (CS-2) to (CS-5) holds relative to (CS-1) is given by*

$$\begin{aligned} \Delta_{CS-1,CS-2} &= \mathbb{E} \left[\frac{p_D(X)^2 p_T(X)^2}{p_{DT}^2} (m_Y(X) - \theta)^2 \left(\frac{1}{p_D(X)p_T(X)} - \frac{1}{p_D(X)} - \frac{1}{p_T(X)} + 1 \right) \right] \\ \Delta_{CS-1,CS-3} &= \mathbb{E} \left[\frac{p_D(1,X)^2}{p_D(1)^2} (m_Y(X) - \theta)^2 \left(\frac{1}{p_T} - 1 \right) \right] \\ \Delta_{CS-1,CS-4} &= \mathbb{E} \left[\frac{p_D(X)^2}{p_D^2} (m_Y(X) - \theta)^2 \frac{1}{p_D(X)} \left(\frac{1}{p_T} - 1 \right) \right] \\ \Delta_{CS-1,CS-5} &= \mathbb{E} \left[(m_Y(X) - \theta)^2 \left(\frac{1}{p_D p_T} - 1 \right) \right]. \end{aligned}$$

Proof: see Appendix 2.A.2.

Unambiguously, $\Delta_{CS-1,CS-2}, \Delta_{CS-1,CS-3}, \Delta_{CS-1,CS-4}, \Delta_{CS-1,CS-5} > 0$. Hence, knowing that one of the assumptions (CS-2)-(CS-5) is true, results in lower efficiency bounds compared to not making any assumptions about the relation between D , T and X . Since (CS-2) and (CS-3) contain (CS-4) and (CS-4) contains (CS-5) similar results can be shown for these cases. To conclude, more restrictive assumptions about the relation between D , T and X imply lower efficiency bounds. This hints at a trade-off regarding the robustness with respect to the model assumptions imposed and the variance lower bound that is asymptotically achievable.

2.3.2 Panel data

Assumption 2.4 (Data-generating process PA). (i) *The i.i.d. sample of a two-period panel with $W = (Y(0), Y(1), D, X)$ and observations $i = 1, \dots, N$ is observed;* (ii) *The distribution $F_W(w)$ exists.*

Assumption 2.4 describes the DGP when panel data is available. In contrast to Assumption 2.1, the sample is not merged and contains $i = 1, \dots, N$ unique individuals. Also the outcomes $Y(0)$ and $Y(1)$ are directly observable in the sample for one observation and, thus, do not need to be inferred.

For panel data, we discriminate two settings (PA-1) and (PA-2) that describe different assumption sets on the relation between D and X . Since we do not need a time indicator to describe the sample, the analysis is limited to the relation between D and X . In (PA-1), we do not make any further assumptions. It is the fully robust setting.

Assumption 2.5 (Relation of D and X). *The variables D and X are assumed to be independent, $D \perp X$ (PA-2).*

(PA-2) is more restrictive in the sense that imbalances between the treatment and the control group are ruled out.

Identification

Assumption 2.6 (Identification PA). *For any $d, t \in \{0, 1\}$ and $x \in \mathcal{X}$,*

(i) *(Observational Rule) For each observation i , the outcomes*

$$Y_i(t) = D_i Y_i^1(t) + (1 - D_i) Y_i^0(t) \text{ are observed;}$$

(ii) *(Common Support) The propensity score $p_{D=d}(x)$ is bounded away from zero;*

(iii) *(No Anticipation) $\mathbb{E}[Y^1(0) - Y^0(0)|D = 1, X = x] = 0$;*

(iv) *(Conditional Common Trends)*

$$\mathbb{E}[Y^0(1) - Y^0(0)|D = 0, X = x] = \mathbb{E}[Y^0(1) - Y^0(0)|D = 1, X = x].$$

Similarly to Assumption 2.3, Assumption 2.6 ensures the identification of θ when panel data is available. The Observational Rule guarantees that for both outcomes $Y(0)$ and $Y(1)$ an observation cannot be part of the treatment and the control group. The other assumptions are adapted in a straightforward manner from the cross-sectional setting.

Again we let $q_{PA;D=d}(X)$ denote the conditional probability function $p_{D=d}(X)$ under either (PA-1) or (PA-2).

Lemma 2.2. *Under Assumptions 2.4 and 2.6, the parameter $\theta = \mathbb{E}[Y^1(1) - Y^0(1)|D = 1]$ is identified as $\mathbb{E}\left[m_{\Delta Y}(X) \frac{q_{PA;D}(X)}{p_D}\right]$.*

Proof: The proof follows similar to Lemma 2.1.

Efficiency bounds

Theorem 2.2 (Semiparametric efficiency bounds PA). *Suppose that Assumptions 2.4 and 2.6 hold. Then under each of the settings (PA-1) and (PA-2) the efficient influence function is given by*

$$\psi_{PA}^*(W; \theta) = \frac{q_{PA;D=1}(X)}{p_D} \psi_{PA}^{*a}(W) + \psi_{PA}^{*b}(W) (m_{\Delta Y}(X) - \theta)$$

where

$$\psi_{PA}^{*a}(W) = \frac{D}{q_{PA;D=1}(X)} (Y(1) - Y(0) - m_{\Delta Y}(1, X)) - \frac{1 - D}{q_{PA;D=0}(X)} (Y(1) - Y(0) - m_{\Delta Y}(0, X))$$

and $\psi_{PA-1}^{*b}(W) = \frac{D}{p_D}$ and $\psi_{PA-2}^{*b}(W) = 1$. The semiparametric efficiency bound for settings (PA-1) and (PA-2) is $\mathbb{E}[\psi_{PA}^*(W; \theta)^2]$.

Proof: see Appendix 2.A.3.

Theorem 2.2 is our second main result. Under (PA-1) and (PA-2) an influence function with a different adjustment term $\psi_{PA}^{*b}(W)$ is derived implying different efficiency bounds for (PA-1) and (PA-2). An important implication is summarised in Corollary 2.2.

Corollary 2.2. *In terms of the asymptotic variance bound, the value of knowing that $D \perp X$ is given by*

$$\Delta_{PA-1,PA-2} = \mathbb{E} \left[(m_{\Delta Y}(X) - \theta)^2 \left(\frac{1}{p_D} - 1 \right) \right]$$

Proof: The proof follows similarly to Corollary 2.1.

Again, since $\Delta_{PA-1,PA-2} > 0$, the more restrictive assumption (PA-2) is associated with a lower efficiency bound. Hence, the robustness-efficiency trade-off also materialises for panel data.

Having derived results for both cross-sectional and panel data, it might be of interest to compare the variance lower bounds under both sampling schemes. Corollary 2.3 summarises the implications of Theorems 2.1 and 2.2 for the relative efficiency between panel and cross-sectional data.

Corollary 2.3. *In terms of the asymptotic variance bound, the value of knowing the panel structure under no further assumptions is given by*

$$\Delta_{CS-1,PA-1} = \mathbb{E} \left[\frac{p_D(X)^2}{p_D^2} \left(\sum_{d=0}^1 \frac{\text{Var}(Y(1) + Y(0)|D=d, X)}{p_{D=d}(X)} + \frac{(m_{\Delta Y}(X) - \theta)^2}{p_D(X)} \right) \right].$$

The minimum value of knowing the panel structure is given by

$$\Delta_{CS-5,PA-1} = \mathbb{E} \left[\sum_{d=0}^1 \frac{\text{Var}(Y(1) + Y(0)|D=d, X)}{p_{D=d}} + (m_{\Delta Y}(X) - \theta)^2 \left(1 - \frac{1}{p_D} \right) \right].$$

Proof: see Appendix 2.A.4.

Since $\Delta_{CS-1,PA-1} > 0$, under no further condition on the relation between D , T and X the first result of the corollary shows that knowing the panel structure generally reduces the variance lower bound that is asymptotically achievable under comparable assumptions. This is an intuitive result because the information in a panel is unambiguously richer. $\Delta_{CS-1,PA-1}$ can therefore also be seen as the gain from the potentially more costly panel sampling scheme. Since the difference ΔY is observed with panel data, the variance lower bound only contains propensity score reweighted conditional variances. For cross-sectional data the conditional variances are additionally reweighted by $p_{T=i}$ resulting in an efficiency loss. Further, notice that the gain from observing the panel $\Delta_{CS-1,PA-1}$ is higher (lower) when the correlation between $Y(0)$ and $Y(1)$ is positive (negative). This can be explained by the fact that the variance of ΔY is lowest when $Y(0)$ and $Y(1)$ are positively correlated. Hence, the panel becomes more valuable when the observed difference ΔY is less volatile.

The second result of the corollary assesses the value of making assumptions relative to having access to panel versus cross-sectional data. It is hypothesised that the researcher knows that (CS-5) is correct, panel data is available and he uses the stronger than necessary setting (PA-1). From Theorem 2.1 and Corollary 2.1 we know that among all settings (CS-1)-(CS-5) the minimum variance lower bound with cross-sectional data is achieved when making assumption (CS-5). From Theorem 2.2 and Corollary 2.2 we know that among (PA-1) and (PA-2) the maximum variance lower bound with panel data is achieved when making assumption (PA-1). The difference among these settings when panel data becomes available $\Delta_{CS-5,PA-1}$ thus measures the value of making assumptions relative to having access to better data. Notice that $\Delta_{CS-5,PA-1} \leq 0$. Hence, having access to panel data does not necessarily lead to lower efficiency bounds. Rather, the corollary shows the importance of making adequate assumptions on the relation between D , (T) and X .

2.4 Estimation and inference

2.4.1 High-dimensional data and important building blocks

Denote by η a set of nuisance parameters that are generally unknown and have to be estimated in a first stage. From Section 2.3, η consists of functions of $X \in \mathbb{R}^{\lambda_X}$. By Assumptions 2.1 and 2.4 the results of Section 2.3 only hold when the dimension of the covariates λ_X is fixed. For estimation and inference we make this condition more specific and assume that potentially the dimension of λ_X can be much higher than the sample size observed. In particular, we allow for $\lambda_X \gg N$. Conditions in the next subsection will implicitly describe the concrete relation of λ_X and N .

Let ψ be a function of the observed variables W and the nuisance parameters η . W contains the generic outcome variable $\tilde{Y}$ and η contains some projection on $\tilde{Y}$ denoted by $m_{\tilde{Y}}(\cdot)$. In contrast to Section 2.3, we now explicitly indicate that the function ψ depends on η and consider ψ to be of the general form

$$\psi(W, \eta; \theta) = \psi^a(W, \eta) - \psi^b(W, \eta)\theta.$$

In particular, the function $\psi^a(W, \eta)$ is a sum over the index τ of terms in the form

$$\frac{q_1(X)}{q_1} \frac{G_\tau}{q_{G_\tau}(X)} (\tilde{Y} - m_{\tilde{Y}}(G_\tau = 1, X)) + \psi^b(W, \eta) m_{\tilde{Y}}(G_\tau = 1, X)$$

where $q_{G_\tau}(x) = \Pr(G_\tau = 1 | X = x)$ under some specific assumption on the relation between G_τ and X and $q_1(X)$ and q_1 are shorthand symbols for $q_{G_\tau}(X)$ and q_{G_τ} with $\tau = (1, 1)$ in the cross-sectional and $\tau = 1$ in the panel case. Suppose that $\mathbb{E}[\psi(W, \eta; \theta)] = 0$ such that θ is identified as

$$\theta = \frac{\mathbb{E} \left[\sum_\tau \left(\frac{q_1(X)}{q_1} \frac{G_\tau}{q_{G_\tau}(X)} (\tilde{Y} - m_{\tilde{Y}}(G_\tau = 1, X)) + \psi^b(W, \eta) m_{\tilde{Y}}(G_\tau = 1, X) \right) \right]}{\mathbb{E}[\psi^b(W, \eta)]}.$$

A plug-in estimator uses the ratio of sample averages with estimated nuisance parameters. Following the suggestions of Chernozhukov et al. (2018a), we apply a cross-fitting algorithm for the nuisance parameter estimation step in order to guarantee that the resulting estimators of $\psi^a(W, \eta)$ and $\psi^b(W, \eta)$ consist of independent observations. The details of the estimation strategy are outlined in the algorithm below. Our conditions on the first

Cross-fitting algorithm:

Suppose that the random variables W can be indexed by i such that the sample is described by W_i for $i = 1, \dots, N$. Randomly split the sample in K equal subsamples of size $n = \frac{N}{K}$. For each of the subsamples with index $k = 1, \dots, K$, define the set of sample indices in subsample k by $\mathcal{J}^k$ and the set of sample indices not in k by $\mathcal{J}^{-k}$. Then a cross-fitted estimator $\hat{\theta}$ is obtained by the following procedure.

for $k = 1$ **to** K :

1. Estimate all nuisance parameters η using $W_{i \in \mathcal{J}^{-k}}$ and define these estimators as $\hat{\eta}_{-k}$.
2. Use $W_{i \in \mathcal{J}^k}$ to obtain $\frac{1}{n} \sum_{i \in \mathcal{J}^k} \psi^a(W_i, \hat{\eta}_{-k})$ and $\frac{1}{n} \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \hat{\eta}_{-k})$.

endfor.

Finally, construct the estimator $\hat{\theta} = \frac{\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^a(W_i, \hat{\eta}_{-k})}{\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \hat{\eta}_{-k})}$.

stage nuisance parameter convergence rates should cover a wide range of estimators. We therefore rely on L_2 convergence rates. To reduce the notational burden, generically write L_2 -rates for cross-fitted nuisance parameters as $\varepsilon_\eta = \|\hat{\eta}_{-k} - \eta\|_2$.

2.4.2 Asymptotic results

In order to derive asymptotic results for the cross-fitting estimator described in the previous subsection, we have to make several assumptions.

Assumption 2.7 (Existence of higher-order moments). *(i) For $r \in \mathbb{N}$, $\mathbb{E}[\|\tilde{Y}\|^r]$ with $r \geq 2$ exists; (ii) For all $x \in \mathcal{X}$, $\tau \in \mathcal{T}$, the conditional variance $\text{Var}(\tilde{Y}|G_\tau = 1, X = x)$ exists.*

Notice that Assumption 2.7 at least requires the existence of the second moment of the outcome. We will later show that for some difference-in-differences estimators the existence of higher-order moments interacts with the requirements of first stage convergence conditions.

Assumption 2.8 (Boundedness of propensity scores). *For all $\tau \in \mathcal{T}$, conditional probabilities and their estimators obey $0 < \inf_{x \in \mathcal{X}} |q_{G_\tau}(x)| < \sup_{x \in \mathcal{X}} |q_{G_\tau}(x)| < 1$ and $0 < \inf_{x \in \mathcal{X}} |\hat{q}_{G_\tau}(x)_{-k}| < \sup_{x \in \mathcal{X}} |\hat{q}_{G_\tau}(x)_{-k}| < 1$.*

Assumption 2.8 is stronger than the usual Common Support conditions made in Assumptions 2.3 and 2.6 to obtain the identification results in Section 2.3. All conditional probabilities and their estimators are strictly bounded away from zero and one.

Assumption 2.9 (Behaviour of adjustment term). *The term $\psi^b(W, \eta)$ obeys,*

- (i) $\mathbb{E}[\psi^b(W, \eta)|X] = \frac{q_1(X)}{q_1}$;
- (ii) $0 < \inf \|\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \eta)\| \leq C$, $0 < \inf \|\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \hat{\eta}_{-k})\| \leq C$;
- (iii) for all $\tau \in \mathcal{T}$, $\left\| \frac{G_\tau}{q_1} \frac{q_1(X)}{q_{G_\tau}(X)} - \psi^b(W, \eta) \right\|_\infty \leq C$.

Assumption 2.9 contains some further regularity conditions on the behaviour of the term $\psi^b(W, \eta)$ that are easily satisfied for the difference-in-differences estimators to be considered. Condition (i) is in principle redundant because if it would not hold then $\mathbb{E}[\psi(W, \eta; \theta)] = 0$ would not be satisfied. We keep it, however, to remind us that the term cannot be of arbitrary form. Since the cross-fitted estimator $\hat{\theta}$ involves dividing by the sample plug-in average of $\psi^b(W, \eta)$, condition (ii) is needed to guarantee that $\hat{\theta}$ is prevented from being extremely scaled by its denominator. Condition (iii) is satisfied if term $\psi^b(W, \eta)$ is a bounded function.

Assumption 2.10 (First stage convergence conditions). *The convergence conditions*

- (i) $\|\psi^b(W, \hat{\eta}_{-k}) - \psi^b(W, \eta)\|_2 = o(1)$ and $\|\psi^b(W, \hat{\eta}_{-k}) - \psi^b(W, \eta)\|_\infty = O(1)$;
- (ii) $\mathbb{E}[\psi^b(W, \hat{\eta}_{-k}) - \psi^b(W, \eta)] = o(N^{-\frac{1}{2}})$;
- (iii) $\mathbb{E}[(\psi^b(W, \hat{\eta}_{-k}) - \psi^b(W, \eta)) \hat{m}_{\tilde{Y}}(G_\tau = 1, X)] = o(N^{-\frac{1}{2}})$ and $\|(\psi^b(W, \hat{\eta}_{-k}) - \psi^b(W, \eta)) m_{\tilde{Y}}(G_\tau = 1, X)\|_2 = o(1)$;
- (iv) $\|\hat{m}_{\tilde{Y}}(G_\tau = 1, X)_{-k} - m_{\tilde{Y}}(G_\tau = 1, X)\|_2 = o(1)$ and $\left\| \frac{\hat{q}_1(X)_{-k}}{\hat{q}_{G_\tau}(X)_{-k}} - \frac{q_1(X)}{q_{G_\tau}(X)} \right\|_2 = o(1)$;
- (v) $\left\| \frac{\hat{q}_1(X)_{-k}}{\hat{q}_{G_\tau}(X)_{-k}} - \frac{q_1(X)}{q_{G_\tau}(X)} \right\|_2 \times \|\hat{m}_{\tilde{Y}}(G_\tau = 1, X)_{-k} - m_{\tilde{Y}}(G_\tau = 1, X)\|_2 = o(N^{-\frac{1}{2}})$

are satisfied for all $\tau \in \mathcal{T}$.

Assumption 2.10 comprises the required coupled convergence conditions that are at the centre of our theoretical argument. The assumption ensures that the first stage nuisance parameter estimation has no effect on the asymptotic behaviour of the cross-fitted estimator. Conditions (i)-(iii) contain the convergence requirements for term $\psi^b(W, \eta)$. Notice that they are trivially satisfied whenever $\psi^b = \psi^b(W)$. Hence, when the term does not depend on nuisance parameters but just on observed variables, conditions (i)-(iii) always hold. However, some of the efficient influence functions considered for the cross-sectional case in Section 2.3 suggest that one may also use functions where $\psi^b = \psi^b(W, \eta)$. This represents the first extension to the work of Chernozhukov et al. (2018a) who just consider problems of the latter type when deriving first stage convergence rate conditions. It enables us to consider cross-fitted difference-in-differences estimators implied by efficient influence functions from settings (CS-2) or (CS-3). Condition (iv) requires that for all $\tau \in \mathcal{T}$, the nuisance parameters converge in L_2 . Condition (v) requires that for all $\tau \in \mathcal{T}$, the conditional probability and the outcome nuisances jointly achieve $\sqrt{N}$ -convergence. Conditions (iv) and (v) are easy-to-check conditions because they can be applied to all sorts of reweighting schemes. In particular, the specific convergence conditions for all $\tau \in \mathcal{T}$ can be directly retrieved from the conditions provided. This represents the second extension to the work of Chernozhukov et al. (2018a) who rely on Gateaux differentiation and implied regularity conditions to derive required convergence rate for specific settings. Our theory covers all functions $\psi(W, \eta; \theta)$ where $\psi^a(W, \eta)$ can be written as a sum over some index τ . This includes the usual parameters of interest and settings in causal econometrics such as selection-on-observables with binary or multiple treatments and difference-in-differences. For example for multiple treatments effects estimation one can directly infer the various rate conditions for the different possible parameters of interest using the efficient influence functions derived in Cattaneo (2010) and Lee (2018). Further, the framework extends to parameters where Gateaux differentiation is not trivial. This may in particular be useful for estimators with finite sample adjustments like trimming conditions that rely on index functions.

The L_2 rate conditions in Assumption 2.10 (iv) and (v) can be shown to be satisfied for many supervised machine learning algorithms. For example, Belloni and Chernozhukov (2013) and Bühlmann and van de Geer (2011) provide detailed L_2 convergence rate conditions for Lasso estimators. Similar rate conditions can also be shown for non-linear models like Random Forests (Wager and Walther, 2016), Honest Random Forests (Wager and Athey, 2018) or forms of Deep Neural Nets (Farrell, Liang, and Misra, 2019). Belloni and Chernozhukov (2013) and Wager and Walther (2016) even show that sufficient L_2 convergence rates are feasible when the dimension λ_X slowly grows with the sample size. We do not explicitly allow for this because Assumptions 2.1 (ii) and 2.4 (ii) require that the dimension of X is fixed. Still, the dimension can be much higher than the sample size.

Theorem 2.3 (Estimation and inference). *Suppose that $\psi^a(W, \eta)$ is a sum over the index τ of terms in the form*

$$\frac{q_1(X)}{q_1} \frac{G_\tau}{q_{G_\tau}(X)} (\tilde{Y} - m_{\tilde{Y}}(G_\tau = 1, X)) + \psi^b(W, \eta) m_{\tilde{Y}}(G_\tau = 1, X),$$

$\mathbb{E}[\psi(W, \eta; \theta)] = 0$ and Assumptions 2.7-2.10 hold. Then the cross-fitted estimator $\hat{\theta}$ obeys

$$\sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(W_i, \eta; \theta) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

where $\sigma^2 = \mathbb{E}[\psi(W, \eta; \theta)^2]$.

Proof: see Appendix 2.B.2.

Theorem 2.3 is our third main result. The cross-fitted estimator is asymptotically normal with influence function $\psi(W, \eta; \theta)$. Notice that the result implies that whenever we use the sample plug-in estimator of the efficient influence function the cross-fitted estimator asymptotically attains the low-dimensional variance lower bound if Assumptions 2.7-2.10 are satisfied. The following subsections contain several corollaries of Theorem 2.3 that summarise the convergence conditions and asymptotic behaviour of different cross-fitted semiparametric difference-in-differences estimators.

2.4.3 Plug-in estimators

Corollary 2.4 summarises the implications of Theorem 2.3 for cross-sectional difference-in-differences estimators in settings (CS-1)-(CS-5) that use the efficient influence functions derived in Theorem 2.1.

Corollary 2.4. *Suppose that Assumptions 2.1, 2.3, 2.7 and 2.8 hold then*

- (a) *under setting (CS-1), $r = 2$ and assuming that for all $d, t \in \{0, 1\}$ $\epsilon_{p_{D=d, T=t}(X)} = o(1)$ and for $(d, t) \in \{(0, 1), (1, 0), (0, 0)\}$ $\epsilon_{m_Y(d, t, X)} = o(1)$ and*

$$\left(\epsilon_{p_{D=1, T=1}(X)} + \epsilon_{p_{D=d, T=t}(X)} \right) \times \epsilon_{m_Y(d, t, X)} = o\left(N^{-\frac{1}{2}}\right)$$

- (b) *under the condition (CS-2) in Assumption 2.2, $r > 2$ and assuming that $\epsilon_{p_D(X)} = o(1)$, $\epsilon_{p_T(X)} = o(1)$ and $\epsilon_{m_Y(d, t, X)} = o(1)$, $\|\hat{m}_Y(d, t, X) - m_Y(d, t, X)\|_r = O(1)$ for all $d, t \in \{0, 1\}$ and*

$$\begin{aligned} \epsilon_{p_D(X)} \times \epsilon_{p_T(X)} &= o\left(N^{-\frac{1}{2} \frac{r}{r-1}}\right), \quad \epsilon_{p_D(X)} \times \epsilon_{m_Y(0, 1, X)} = o\left(N^{-\frac{1}{2}}\right), \\ \epsilon_{p_T(X)} \times \epsilon_{m_Y(1, 0, X)} &= o\left(N^{-\frac{1}{2}}\right), \quad (\epsilon_{p_D(X)} + \epsilon_{p_T(X)}) \times \epsilon_{m_Y(0, 0, X)} = o\left(N^{-\frac{1}{2}}\right) \end{aligned}$$

- (c) *under the condition (CS-3) in Assumption 2.2, $r > 2$ and assuming that $\epsilon_{p_D(1, X)} = o(1)$, $\epsilon_{p_D(0, X)} = o(1)$ and $\epsilon_{m_Y(d, t, X)} = o(1)$ for all $d, t \in \{0, 1\}$ and*

$$\epsilon_{p_D(1, X)} \times \epsilon_{m_Y(0, 1, X)} = o\left(N^{-\frac{1}{2}}\right), \quad (\epsilon_{p_D(1, X)} + \epsilon_{p_D(0, X)}) \times (\epsilon_{m_Y(1, 0, X)} + \epsilon_{m_Y(0, 0, X)}) = o\left(N^{-\frac{1}{2}}\right)$$

- (d) *under the condition (CS-4) in Assumption 2.2, $r = 2$ and assuming that $\epsilon_{p_D(X)} = o(1)$ and $\epsilon_{m_Y(d, t, X)} = o(1)$ for all $d, t \in \{0, 1\}$ and*

$$\epsilon_{p_D(X)} \times (\epsilon_{m_Y(0, 1, X)} + \epsilon_{m_Y(0, 0, X)}) = o\left(N^{-\frac{1}{2}}\right)$$

- (e) *under the condition (CS-5) in Assumption 2.2, $r = 2$ and assuming that $\epsilon_{m_Y(d, t, X)} = o(1)$ for all $d, t \in \{0, 1\}$*

the sample plug-in estimators implied by $\mathbb{E}[\psi_{CS}^(W; \theta)] = 0$ are efficient estimators in the sense that they obey*

$$\sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi_{CS}^*(W_i; \theta) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\psi_{CS}^*(W; \theta)^2]).$$

Proof: see Appendix 2.B.3.

The joint convergence conditions are more sophisticated for settings that are less restrictive with respect to the model assumptions on the relation between D , T and X imposed. For (CS-1) we need six, for (CS-2) and (CS-3) we need five, for (CS-4) we need two and for (CS-5) we need zero joint convergence rates to be satisfied. This hints at a trade-off between the robustness of the estimators with respect to the assumption on D , T and X and the robustness of the estimators with respect to first stage nuisance parameter convergence requirements. Further, we notice that for settings (CS-2) and (CS-3) higher-order moments $r > 2$ have to exist. For setting (CS-3) this is a mere regularity condition. For setting (CS-2) it has some implications on the joint convergence condition of the two propensity scores. The more moments of the outcome exist, the closer the joint convergence condition to the usual $\sqrt{N}$ condition. This implies that for a bounded outcome the condition becomes $\epsilon_{p_D(X)} \times \epsilon_{p_T(X)} = o\left(N^{-\frac{1}{2}}\right)$.⁵ If the researcher is only willing to assume that a second moment of Y exists, this implies that for both propensity scores parametric convergence rates are needed. We additionally notice that for the experimental setting (CS-5) the efficient estimator is not the simple difference in means estimator but a residualised version of it. We will discuss the implications of this finding in the next subsection and in Section 2.5.

Corollary 2.5 summarises the implications of Theorem 2.3 for panel difference-in-differences estimators in settings (PA-1) and (PA-2) that use the efficient influence functions derived in Theorem 2.2.

Corollary 2.5. *Suppose that Assumptions 2.4, 2.6, 2.7 and 2.8 hold then*

(a) *under the condition (PA-1), $r = 2$ and assuming that $\epsilon_{p_D(X)} = o(1)$, $\epsilon_{m_{\Delta Y}(0,X)} = o(1)$ and*

$$\epsilon_{p_D(X)} \times \epsilon_{m_{\Delta Y}(0,X)} = o\left(N^{-\frac{1}{2}}\right)$$

(b) *under the condition (PA-2) in Assumption 2.5, $r = 2$ and assuming that $\epsilon_{m_{\Delta Y}(d,X)} = o(1)$ for $d \in \{0, 1\}$*

the sample plug-in estimators implied by $\mathbb{E}[\psi_{PA}^(W; \theta)] = 0$ are efficient estimators in the sense that they obey*

$$\sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi_{PA}^*(W_i; \theta) + o_p(1) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}[\psi_{PA}^*(W; \theta)^2]\right).$$

Proof: see Appendix 2.B.4.

Compared to the cross-sectional case, the joint convergence condition are generally weaker for (PA-1) for two reasons. Firstly, the fact that we observe ΔY only requires one projection on the difference of the outcomes instead of the difference between two projections on each single outcome. Secondly, the efficient influence function in Theorem 2.2 implies that $m_{\Delta Y}(1, X)$ is redundant. This allows us to obtain efficient difference-in-differences estimators for the panel under relatively weak conditions.

2.4.4 Redundancy of nuisance parameters

For the cross-sectional difference-in-differences estimators considered so far some nuisance parameters are redundant. For example, in setting (CS-1) for all $x \in \mathcal{X}$ we have $\sum_{d=0}^1 \sum_{t=0}^1 p_{D=d, T=t}(x) = 1$. In principle, we could therefore infer one of the four propensity scores from the other three. A well-known problem is that if one estimates the three propensity scores without requiring that the implied four propensity scores sum to one, Assumption 2.8

⁵For most outcomes in labour market applications the outcome is bounded. The condition might, however, become relevant for distributions with ‘fat tails’ typically present in financial econometrics.

might be violated because the fourth, implied propensity score is not guaranteed to be strictly greater than zero. A solution could be to explicitly require that the four propensity scores estimators sum to one for all $x \in \mathcal{X}$. Multinomial versions of the standard Logit regression or the Logit Lasso are available. To the knowledge of the author, solutions do not exist for more sophisticated machine learning algorithms like Random Forests or Neural Nets. However, notice that $p_{D=d, T=t}(X) = p_{D=d}(t, X)p_{T=t}(X) = p_{T=t}(d, X)p_{D=d}(X)$ and that the propensity score estimators obey $\sum_{d=0}^1 \sum_{t=0}^1 \hat{p}_{D=d}(t, x) - k \hat{p}_{T=t}(x) - k = \sum_{d=0}^1 \sum_{t=0}^1 \hat{p}_{T=t}(d, x) - k \hat{p}_{D=d}(x) - k = 1$ for all $x \in \mathcal{X}$ by construction. Corollary 2.6 summarises the properties of the two implied estimators.

Corollary 2.6. *Suppose that Assumptions 2.1, 2.3, 2.7 and 2.8 hold then under the condition (CS-1) and $r = 2$, the estimators implied by the moment condition of the score functions*

(a)

$$\begin{aligned} \psi_{CS-1}^{**}(W, \eta; \theta) &= \frac{p_D(1, X)p_T(X)}{p_{DT}} \sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \frac{G_{d,t}}{p_{D=d}(t, X)p_{T=t}(X)} (Y - m_Y(d, t, X)) \\ &\quad + \frac{DT}{p_{DT}} (m_Y(X) - \theta) \end{aligned}$$

under the further conditions that $\varepsilon_{p_D(1, X)} = o(1)$, $\varepsilon_{p_D(0, X)} = o(1)$, $\varepsilon_{p_T(X)} = o(1)$ and for all $(d, t) \in \{(0, 1), (1, 0), (0, 0)\}$ $\varepsilon_{m_Y(d, t, X)} = o(1)$ and

$$\begin{aligned} \varepsilon_{p_D(1, X)} \times \varepsilon_{m_Y(0, 1, X)} &= o\left(N^{-\frac{1}{2}}\right) \quad \text{and} \\ (\varepsilon_{p_D(1, X)} + \varepsilon_{p_D(0, X)} + \varepsilon_{p_T(X)}) \times (\varepsilon_{m_Y(1, 0, X)} + \varepsilon_{m_Y(0, 0, X)}) &= o\left(N^{-\frac{1}{2}}\right) \end{aligned}$$

(b)

$$\begin{aligned} \psi_{CS-1}^{***}(W, \eta; \theta) &= \frac{p_T(1, X)p_D(X)}{p_{DT}} \sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \frac{G_{d,t}}{p_{T=t}(d, X)p_{D=d}(X)} (Y - m_Y(d, t, X)) \\ &\quad + \frac{DT}{p_{DT}} (m_Y(X) - \theta) \end{aligned}$$

under the further conditions that $\varepsilon_{p_T(1, X)} = o(1)$, $\varepsilon_{p_T(0, X)} = o(1)$, $\varepsilon_{p_D(X)} = o(1)$ and for all $(d, t) \in \{(0, 1), (1, 0), (0, 0)\}$ $\varepsilon_{m_Y(d, t, X)} = o(1)$ and

$$\begin{aligned} \varepsilon_{p_T(1, X)} \times \varepsilon_{m_Y(1, 0, X)} &= o\left(N^{-\frac{1}{2}}\right) \quad \text{and} \\ (\varepsilon_{p_T(1, X)} + \varepsilon_{p_T(0, X)} + \varepsilon_{p_D(X)}) \times (\varepsilon_{m_Y(0, 1, X)} + \varepsilon_{m_Y(0, 0, X)}) &= o\left(N^{-\frac{1}{2}}\right) \end{aligned}$$

obey

$$\sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi_{CS-1}^*(W_i; \theta) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\psi_{CS-1}^*(W; \theta)^2]).$$

Proof: The proof follows similarly to Corollary 2.4 (a).

Compared to Corollary 2.4 (a), the joint convergence conditions are increased from six to seven and the efficiency result is maintained. The scores rely on either $p_{D=d}(t, X)$ or $p_{T=t}(d, X)$. We therefore recommend to use $\psi_{CS-1}^{**}(W, \eta; \theta)$ when $|p_T - 0.5| < |p_D - 0.5|$ and $\psi_{CS-1}^{***}(W, \eta; \theta)$ otherwise.

The result of Corollary 2.4 (b)-(d) indicate that some of the outcome nuisances in settings (CS-2), (CS-3) and (CS-4) are redundant in the sense that they do not contribute to the joint convergence condition in Assumption 2.10 (v). However, the outcome nuisances still need to converge in L_2 in order to satisfy Assumption 2.10 (iv). The same applies for (CS-5) since the simple difference-in-means estimator represents an alternative estimator that does not rely on any first stage convergence conditions. We do not provide a separate result for case (CS-5). However, it can be checked that the implied estimator without any outcome nuisances is just the difference-in-means estimator. Notice that as long as Y is not independent of X the difference-in-means estimator is *not* an efficient estimator. This has some practical relevance as in difference-in-differences the credibility of the design is often assessed by using placebo tests in some pre-periods without relying on covariates. Not rejecting the null might, however, just due to a higher than necessary standard error when using the difference-in-means estimator. The same applies for setting (PA-2). Since extensions for (CS-3) follow in a very similar fashion and do not yield large simplifications, we focus on settings (CS-2) and (CS-4). Corollaries 2.7 and 2.8 summarise the implications when scores are used that do not contain some of the outcome nuisances. Generally, the implied cross-fitted sample plug-in estimators do not attain the variance lower bound but have otherwise desirable asymptotic properties under weaker conditions. This hints at another trade-off between the robustness of the estimator towards first stage convergence requirements and semiparametric efficiency.

Corollary 2.7. *Suppose that Assumptions 2.1, 2.3, 2.7 and 2.8 hold then under the condition (CS-2) in Assumption 2.2, $r = 2$ and assuming that $\epsilon_{pD(X)} = o(1)$, $\epsilon_{pT(X)} = o(1)$ and $\epsilon_{m_Y(d,t,X)} = o(1)$ for all $(d,t) \in \{(0,1), (1,0), (0,0)\}$ and*

$$\begin{aligned}\epsilon_{pD(X)} \times \epsilon_{m_Y(0,1,X)} &= o\left(N^{-\frac{1}{2}}\right), \quad \epsilon_{pT(X)} \times \epsilon_{m_Y(1,0,X)} = o\left(N^{-\frac{1}{2}}\right), \\ (\epsilon_{pD(X)} + \epsilon_{pT(X)}) \times \epsilon_{m_Y(0,0,X)} &= o\left(N^{-\frac{1}{2}}\right)\end{aligned}$$

*the sample plug-in estimator using the score $\psi'_{CS-2}(W; \theta) = \frac{p_D(X)p_T(X)}{p_{DT}} \psi_{CS-2}^{*a} + \frac{DT}{p_{DT}} (m_Y(X) - \theta)$ obeys*

$$\sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi'_{CS-2}(W_i; \theta) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\psi'_{CS-2}(W; \theta)^2])$$

with efficiency loss $\mathbb{E}[\psi'_{CS-2}(W; \theta)^2] - \mathbb{E}[\psi_{CS-2}^(W; \theta)^2] = \Delta_{CS-1, CS-2}$.*

Proof: see Appendix 2.B.5.

Corollary 2.7 shows that for (CS-2) one may in principle get rid off the restrictive existence of higher-order moments requirement and the L_2 convergence condition for $m_Y(1,1,x)$ in Corollary 2.4 (b). However, this results in an efficiency loss. Notice that from Corollary 2.1 $\Delta_{CS-1, CS-2}$ represents the value of knowing that (CS-2) in Assumption 2.2 is true relative to not making any assumptions. Hence, when using score $\psi'_{CS-2}(W; \theta)$ all efficiency gains from the stronger setting (CS-2) relative to (CS-1) are exchanged for relatively weak first stage convergence conditions.

Corollary 2.8. *Suppose that Assumptions 2.1, 2.3, 2.7 and 2.8 hold then under the condition (CS-4) in Assumption 2.2, $r = 2$ and assuming that $\epsilon_{pD(X)} = o(1)$ and $\epsilon_{m_Y(d,t,X)} = o(1)$ for all $(d,t) \in \{(0,1), (0,0)\}$ and*

$$\epsilon_{pD(X)} \times (\epsilon_{m_Y(0,1,X)} + \epsilon_{m_Y(0,0,X)}) = o\left(N^{-\frac{1}{2}}\right)$$

the sample plug-in estimator using the score

$$\psi'_{CS-4}(W; \theta) = \psi^*_{CS-4}(W; \theta) + \frac{D}{p_D} \left(\frac{T}{p_T} - 1 \right) m_Y(1, 1, X) - \frac{D}{p_D} \left(\frac{1-T}{1-p_T} - 1 \right) m_Y(1, 0, X)$$

obeys

$$\sqrt{N} (\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi'_{CS-4}(W_i; \theta) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \mathbb{E} [\psi'_{CS-4}(W; \theta)^2])$$

$$\text{with efficiency loss } \mathbb{E} [\psi'_{CS-4}(W; \theta)^2] - \mathbb{E} [\psi^*_{CS-4}(W; \theta)^2] = \mathbb{E} \left[\frac{p_D(X)^2}{p_D^2} \frac{\left(\sqrt{\frac{1-p_T}{p_T}} m_Y(1, 1, X) + \sqrt{\frac{p_T}{1-p_T}} m_Y(1, 0, X) \right)^2}{p_D(X)} \right].$$

Proof: see Appendix 2.B.6.

Corollary 2.8 provides the convergence conditions when we use a score where $m_Y(1, 1, x)$ and $m_Y(1, 0, x)$ are redundant. This results in weaker convergence conditions similar to those of the score $\psi^*_{PA-1}(W; \theta)$. However, also the efficiency loss can be relatively large for p_T close to zero or one and p_D close to zero.

2.5 Application

To illustrate the practical relevance of the proposed methods, we revisit Angrist and Acemoglu (2001). The paper is concerned with the theoretically ambiguous effect of increased employment protection for disabled workers on weeks worked (for more details see the paper). An empirical evaluation of the Americans with Disabilities Act reform introduced in 1991 is used to test the theory using data from the Current Population Survey (CPS), a repeated cross-section.

Table 2.1: Covariate specifications

specification	covariates used	# of covariates
original	sex, age, race group, education group, region	14
baseline	sex, age, race group, education group, marital status, class of worker, major industry, major occupation, state, central city MSA status	108
extended	sex, age, race group, education group, marital status, class of worker, major industry, major occupation, state, central city MSA status, longest job class of worker, longest job major occupation, longest job major industry, number of employers, unemployment compensation benefit value, supplemental security income amount received, public assistance or welfare value received, social security payments received, veteran status, veterans payment income, survivor's income received, value of other income, value of workers' compensation for job related illness or injury, retirement income, health insurance group, medicare coverage, medicaid coverage, coverage by military health care	161

As in the paper, we define D as being disabled. Y are the weeks worked in the respective year. The outcome is therefore bounded between 0 and 52. In the original paper several post-reform years are considered. Since the credibility of the common trend assumption might be questionable for years well after the reform, we focus on 1992 as the post-reform period ($T = 1$). As in the original paper we use the years 1988-1990 for $T = 0$. All CPS data is retrieved from Joshua Angrist's data archive⁶. We consider three different sets of covariates. An overview

⁶<https://economics.mit.edu/faculty/angrist/data1/data/aceang01>

of the different specifications is provided in Table 2.1. The variable set ‘original’ includes the covariates that are also used in the specifications of Angrist and Acemoglu (2001). The variable sets ‘baseline’ and ‘extended’ use some further covariates available from the CPS. The ‘baseline’ specification disaggregates the region variable used in the original dataset to control for geographically different common trends. Instead, we include state dummies and dummies indicating whether the individual lives in a metropolitan area. Additional controls on marital status, class of worker, industry and occupation should help to control for non-parallel trends between disabled and non-disabled. For example, if people who report a disability self-select more often into a certain sector or industry, any underlying structural change in this sector needs to be controlled for in order for the common trend assumption to be satisfied. This might be a point that is of more general interest in empirical economics. In labour market applications the common trend is often valid only after conditioning on variables that are available at different aggregation levels, e.g. geographic or sector dummies. The ‘extended’ specification includes further covariates on employment history, social welfare payments and health insurance status. Table 2.2 shows that the merged sample consists of 276,127 observations (for all specifications) and is highly imbalanced. We observe that the share of D varies between $T = 0$ and $T = 1$ indicating that assumptions (CS-4) and (CS-5) are likely to be violated.

Table 2.2: Subsample sizes

$N(0)$	$N(1)$	p_D	p_T	$p_D(1)$	$p_D(0)$	$p_{D=1,T=1}$	$p_{D=0,T=1}$	$p_{D=1,T=0}$	$p_{D=0,T=0}$
206,058	70,069	6.19 %	25.38 %	6.49 %	6.08 %	1.65 %	23.73 %	4.54 %	70.08 %

2.5.1 Estimators considered

We consider different estimators using the different score functions outlined in Section 2.4 with different first stage estimators. Our preferred estimator is an Ensemble Learner that weights Lasso and Random Forest predictions by using out-of-sample mean squared error (MSE) optimal weights. For the Lasso, we allow for polynomials up to order four and all two way interactions. The Random Forest is an ensemble of regression trees and therefore implicitly contains higher-order terms. For both estimators we use the default settings in the *glmnet* and *ranger* R-packages. Using an ensemble of machine learning estimators has several important advantages. Firstly, Lasso and Random Forests are designed for different DGPs. Whereas the Lasso allows for some form of smoothing, we expect a tree-based estimator to work well with strong non-linearities. Secondly, the Ensemble gives more weight to the single predictor that works best and therefore should be less dependent on the particular tuning parameter choices of the Lasso and the Random Forest. Thirdly, the strong imbalances in our sample indicate that using single propensity score estimators might result in extreme weights. By combining two predictors, the probability of generating extreme weights is reduced.

We compare the Ensemble Learner to the single predictors Lasso and Random Forest. Due to computational constraints, we restrict this analysis to the ‘baseline’ set of covariates. Further, we compare the Ensemble to parametric models for the propensity scores and the outcome nuisances. For the propensity scores we use Logit and for the outcome nuisances standard linear regression. We include covariates in levels. As a benchmark, we also consider Abadie’s (2005) Inverse Probability Weighting (IPW) difference-in-differences estimator with Logit regression for the propensity score.

2.5.2 Placebo tests

As usual in the difference-in-differences literature, we run some placebo experiments for period 1988/89. Table 2.3 shows the results under (CS-5). Notice that the simple differences-in-means estimator does not hint at a violation of the specification that does not account for any imbalances. This is not because the effect is economically small but due to the comparatively high standard error. If the efficient score $\psi_{CS-5}^*(W, \eta; \theta)$ is used, the point estimators are in the same range but – in line with theory in Section 2.3 – the standard errors are substantially decreased leading to significant results. This shows that using the efficient score function rather than the simple mean differences estimator may lead to opposite conclusions regarding the credibility of the design. The result also highlights the importance of discriminating the two roles covariates can have in semiparametric difference-in-differences estimation. First of all, they may be included to improve the reliability of the Conditional Common Trends (see Assumptions 2.3 and 2.6). Second of all, under some assumptions, they should be included to improve on the efficiency of the derived estimator. Further, we notice that the standard error for the estimator with the Ensemble Learner decreases when more covariates are added to the model, whereas the standard error when using the linear model increases for the ‘extended’ specification. The standard errors are also generally larger. This indicates that the Ensemble Learner is more effective in predicting the conditional outcomes leading to a stronger variance reduction of the residualised estimator. From these results we conclude that incorporating covariates seems to be necessary in our example. Notice that due to the strong imbalances in our application, especially propensity scores for $D = 1, T = 1$ are hard to predict. Results of placebo tests for our estimators are reported in Table 2.4. For estimators that rely on the Ensemble Learner none of the different specifications used hint at a violation of the conditional common trend assumption. Moreover, the standard errors are reduced when a higher number of covariates is included, indicating that the Ensemble is effective in extracting the additional information. When covariates in ‘baseline’ and ‘extended’ are used, estimators under (CS-2) and (CS-4) show substantially decreased standard errors compared to estimators that rely on (CS-1). This is in line with the theoretical argument in Corollary 2.1. When using the same scores with the same set of variables most estimators that rely on parametric models hint at a potential violation of the conditional common trend assumption. With the exception of IPW, estimators implied by scores that only rely on the propensity scores $p_D(X)$ and $p_T(X)$ (CS-2, CS-4) generally do not indicate a violation of the conditional common trend assumption. The same can be observed for estimators that rely solely on the Lasso or the Random Forest. We conclude that scores that rely on sophisticated convergence conditions are

Table 2.3: Placebo tests 1988/89 under (CS-5)

	η estimator	original	baseline	extended
$\psi_{CS-5}^*(W, \eta; \theta)$	Ensemble	-1.410 (0.464)	-0.956 (0.344)	-0.600 (0.218)
$\psi_{CS-5}^*(W, \eta; \theta)$	Linear	-1.141 (0.551)	-1.077 (0.506)	-1.257 (0.690)
Mean differences		-0.801 (0.652)	-0.801 (0.652)	-0.801 (0.652)

Notes: Results for estimators with the Ensemble were obtained using the cross-fitting procedure in Section 2.4 with $K = 2$. The Ensemble Learner comprises Lasso and Random Forest. For Lasso the penalty term was chosen such that the 10-fold cross-validation criterion was minimised. The Ensemble weights were chosen by minimising out-of-sample MSE. Results for estimators with Linear regression were obtained by using the sample plug-in estimator without cross-fitting. The sample for the placebo tests contains $N = 135,174$ observations. Standard errors are in parenthesis.

less robust to potential violations of these conditions in practice. However, it seems that the efficiency-estimation trade-off is less of a concern when high-quality predictors like the Ensemble are used. Moreover, even when using a parametric model and a large number of covariates, the double robust scores give results whereas IPW explodes for the ‘extended’ specification.

2.5.3 Main results

The placebo tests for the different estimators suggest some trustworthiness of estimators that use the Ensemble Learner for first stage prediction. Table 2.5 summarises the main results for the difference-in-differences estimators considered. For some of the specifications for the single predictors and the linear estimators we obtain extreme results. In general, the same conclusions as for the placebo tests apply. We notice that the particular first stage estimator and the estimation-robustness of the score used can drastically shift the results. Whereas estimators for (CS-1)-(CS-3) that use the Ensemble Learner unambiguously give negative or insignificant effects of reasonable size, again some of the other estimators explode. In addition, the Ensemble Learner based estimators mostly become insignificant when more covariates are included in the model. In contrast to estimators with parametric first stages, this is not due to an increase in standard errors when more covariates are included in the model. Rather, the decreased standard errors indicate that the Ensemble Learner allows to effectively exhaust the information of the increased covariate sets. Lastly, we notice that for specifications ‘baseline’ and ‘extended’ the efficiency-robustness trade-off from Corollary 2.1 becomes relevant. Scores under setting (CS-1) generally have higher standard errors compared to efficient scores in settings (CS-2), (CS-3) and (CS-4). While for (CS-4) using $\psi'_{CS-4}(W, \eta; \theta)$ leads to a substantially higher standard error compared to the efficient score, for (CS-2) there is barely a difference. Also the point estimators are similar. This is expected since the score $\psi'_{CS-2}(W, \eta; \theta)$ converges under similar conditions when the outcome is bounded.

2.6 Conclusion

Semiparametric difference-in-differences estimation is a non-trivial endeavour. In this study, we highlight the importance of different assumptions in semiparametric difference-in-differences models. Our results show that efficiency bounds may strongly depend on the model assumptions imposed and the data that is available. In particular, we show that there is a trade-off between the strength of the assumptions imposed and the variance lower bound that can be achieved. For estimation, we provide easy-to-check conditions to derive the required convergence rates for a broad class of estimation problems. Our theoretical results allow to integrate scores with sophisticated adjustment terms in the double machine learning framework. Further, we show that the different semiparametric models imply estimators with different properties. Estimators that are more robust against the model assumptions imposed also rely on more sophisticated conditions for first stage prediction. Some of these conditions can be relaxed when we give up on asymptotically attaining the efficiency bound. An empirical example shows that our proposed estimators are useful in practice. However, estimation results might be highly sensitive regarding the choice of the first stage nuisance parameter predictor. Placebo tests indicate that, in contrast to other choices, our proposed Ensemble Learner performs well.

Some interesting problems are beyond the scope of this study and have to be left for further research. The performance of the estimators proposed might depend on the parameter K . Some theoretical results how to optimally

Table 2.4: Placebo tests 1988/89 under (CS-1)-(CS-4)

	Ensemble			Linear/Logit regression			Forest	Lasso
	original	baseline	extended	original	baseline	extended	baseline	baseline
$\psi_{CS-1}^*(W, \eta; \theta)$	-0.368 (0.487)	0.223 (0.496)	-0.167 (0.457)	-1.338 (0.546)	-3.875 (0.435)	-4.832 (0.598)	7.564 (0.857)	-10.850 (0.410)
$\psi_{CS-1}^{**}(W, \eta; \theta)$	-0.359 (0.487)	0.255 (0.470)	-0.527 (0.367)	-1.425 (0.544)	-3.598 (0.443)	-8.640 (0.516)	24.452 (1.363)	-15.053 (0.365)
$\psi_{CS-1}^{***}(W, \eta; \theta)$	-0.390 (0.481)	0.084 (0.385)	-0.352 (0.287)	-1.296 (0.548)	-4.033 (0.431)	-6.802 (0.554)	4.205 (0.765)	-0.753 (0.618)
$\psi_{CS-2}^*(W, \eta; \theta)$	-0.470 (0.477)	-0.022 (0.379)	-0.391 (0.294)	-0.977 (0.559)	-0.994 (0.510)	0.849 (0.733)	9.365 (0.875)	-0.772 (0.605)
$\psi_{CS-2}'(W, \eta; \theta)$	-0.429 (0.477)	0.008 (0.381)	-0.406 (0.296)	-0.992 (0.556)	-1.070 (0.524)	0.870 (0.730)	11.121 (0.961)	-0.819 (0.616)
$\psi_{CS-3}^*(W, \eta; \theta)$	-0.317 (0.487)	0.271 (0.474)	-0.484 (0.371)	-1.847 (0.528)	-3.817 (0.433)	-10.331 (0.481)	7.275 (0.796)	-14.973 (0.364)
$\psi_{CS-4}^*(W, \eta; \theta)$	-0.685 (0.477)	-0.378 (0.385)	-0.312 (0.277)	-1.138 (0.542)	-1.271 (0.494)	-0.749 (0.680)	0.363 (0.594)	-0.762 (0.586)
$\psi_{CS-4}'(W, \eta; \theta)$	-0.645 (0.620)	-0.416 (0.631)	-0.450 (0.627)	-0.832 (0.604)	-1.027 (0.602)	-0.328 (0.602)	0.719 (0.603)	-0.413 (0.603)
IPW				-1.808 (0.686)	-0.558 (0.658)	-		

Notes: Results for estimators with the Ensemble, the Random Forest and the Lasso were obtained using the cross-fitting procedure in Section 2.4 with $K = 2$. The Ensemble Learner comprises Lasso and Random Forest. For Lasso the penalty term was chosen such that the 10-fold cross-validation criterion was minimised. The Ensemble weights were chosen by minimising out-of-sample MSE. Results for estimators with Linear/Logit regression were obtained by using the sample plug-in estimator without cross-fitting. The Logit was used for the propensity scores and linear regression for the outcome nuisances. The sample for the placebo tests contains $N = 135, 174$ observations. Standard errors are in parenthesis.

Table 2.5: Results for (CS-1)-(CS-4)

	Ensemble			Linear/Logit regression			Forest	Lasso
	original	baseline	extended	original	baseline	extended	baseline	baseline
$\psi_{CS-1}^*(W, \eta; \theta)$	-0.904 (0.376)	-0.855 (0.351)	0.031 (0.273)	1.049 (0.458)	2.118 (0.465)	13.959 (0.782)	13.123 (0.691)	17.558 (0.795)
$\psi_{CS-1}^{**}(W, \eta; \theta)$	-0.828 (0.377)	-0.689 (0.351)	0.017 (0.209)	0.785 (0.453)	0.928 (0.442)	10.466 (0.720)	14.793 (0.720)	11.546 (0.680)
$\psi_{CS-1}^{***}(W, \eta; \theta)$	-0.837 (0.375)	-0.356 (0.318)	-0.096 (0.184)	1.144 (0.459)	0.978 (0.443)	6.457 (0.650)	23.419 (0.887)	-0.120 (0.478)
$\psi_{CS-2}^*(W, \eta; \theta)$	-0.856 (0.365)	-0.422 (0.307)	-0.169 (0.183)	-0.456 (0.434)	-0.845 (0.429)	-1.746 (0.542)	-11.963 (0.432)	-0.987 (0.491)
$\psi_{CS-2}'(W, \eta; \theta)$	-0.792 (0.364)	-0.407 (0.309)	-0.123 (0.185)	-0.453 (0.432)	-0.819 (0.411)	-1.695 (0.517)	-7.549 (0.371)	-0.895 (0.466)
$\psi_{CS-3}^*(W, \eta; \theta)$	-0.934 (0.378)	-0.748 (0.345)	-0.061 (0.198)	-0.121 (0.434)	0.135 (0.426)	8.760 (0.688)	34.956 (1.127)	10.029 (0.648)
$\psi_{CS-4}^*(W, \eta; \theta)$	-0.869 (0.384)	-0.294 (0.322)	-0.118 (0.185)	-1.013 (0.440)	-1.334 (0.436)	-2.409 (0.553)	-1.385 (0.494)	-1.465 (0.502)
$\psi_{CS-4}'(W, \eta; \theta)$	0.903 (0.498)	0.800 (0.514)	0.284 (0.490)	0.834 (0.486)	0.501 (0.485)	0.142 (0.485)	0.842 (0.485)	0.793 (0.487)
IPW				1.470 (0.546)	0.148 (0.529)	-		

Notes: Results for estimators with the Ensemble, the Random Forest and the Lasso were obtained using the cross-fitting procedure in Section 2.4 with $K = 2$. The Ensemble Learner comprises Lasso and Random Forest. For Lasso the penalty term was chosen such that the 10-fold cross-validation criterion was minimised. The Ensemble weights were chosen by minimising out-of-sample MSE. Results for estimators with Linear/Logit regression were obtained by using the sample plug-in estimator without cross-fitting. The Logit was used for the propensity scores and linear regression for the outcome nuisances. The sample contains $N = 276, 127$ observations. Standard errors are in parenthesis.

choose this parameter would be helpful. Our theoretical results for the different estimators suggest that the finite sample performance of the point estimators and the coverage probabilities of the variance estimators might be relatively diverse and depend on the particular DGP considered. Monte Carlo simulations could shed some more light on this subject. Also this study is limited to two time periods and two groups. Some recent advances in the semiparametric difference-in-differences literature (e.g., Callaway and Sant'Anna, 2020; Goodman-Bacon, 2018) comprise extensions to more complicated adoption patterns. Further, in practice panel and cross-sectional data are often combined (e.g. in rotating panels). The results in Section 2.3 suggest that for these kinds of data combination problems efficiency gains are possible if neither the panel structure is neglected nor the cross-sectional data is thrown away. A generalization of the efficiency theory provided in this study to these settings represents yet another avenue for further research.

2.A Proofs for Section 2.3

2.A.1 Proof of Theorem 2.1

We observe the data $W = (Y, D, T, X)$. For the joint distribution of the data consider a regular parametric submodel indexed by β . The density under the submodel can then be written as

$$f_W(w; \beta) = f_{Y|D,T,X}(y|d, t, x; \beta) f_{D,T|X}(d, t|x; \beta) f_X(x; \beta)$$

which equals $f_W(w)$ at $\beta = \beta_0$.

The score function is defined as $S(y, d, t, x; \beta_0) = \frac{\partial \log f_W(w; \beta_0)}{\partial \beta}$ and we obtain

$$S(y, d, t, x; \beta_0) = \sum_{d=0}^1 \sum_{t=0}^1 g_{d,t} S_y(y, d, t, x; \beta_0) + S_{d,t}(d, t, x; \beta_0) + S_x(x; \beta_0)$$

with

$$\begin{aligned} S_y(y, d, t, x; \beta_0) &= \frac{\partial \log f_{Y|D,T,X}(y|d, t, x; \beta_0)}{\partial \beta} \\ S_{d,t}(d, t, x; \beta_0) &= \frac{\partial \log f_{D,T|X}(d, t|x; \beta_0)}{\partial \beta} \\ S_x(x; \beta_0) &= \frac{\partial \log f_X(x; \beta_0)}{\partial \beta} \end{aligned}$$

where $S_{d,t}(d, t, x; \beta_0)$ depends on the settings (CS-1) to (CS-5) on the relation between D , T and X .

For all regular parametric submodels the variance lower bound for a model is the second moment of the projection of a function $\psi_{CS}^*(W; \theta)$ (with $\mathbb{E}[\psi_{CS}^*(W; \theta)] = 0$ and an existing second moment) on the tangent space $\mathfrak{T}$ that satisfies

$$\frac{\partial \theta(\beta_0)}{\partial \beta} = \mathbb{E}[\psi_{CS}^*(W; \theta) S(Y, D, T, X; \beta_0)].$$

When $\psi_{CS}^*(W; \theta) \in \mathfrak{T}$, the projection on $\mathfrak{T}$ is the function itself and therefore the variance lower bound for the model is given by $\mathbb{E}[\psi_{CS}^*(W; \theta)^2]$.

Proof for CS-1

Under the conditions in (CS-1)

$$S_{d,t}(d, t, x; \beta_0) = \sum_{d=0}^1 \sum_{t=0}^1 \frac{g_{d,t}}{p_{D=d, T=t}(x)} \dot{p}_{D=d, T=t}(x; \beta_0).$$

The tangent space of the model is characterised by the set of functions that are mean zero and satisfy the structure of the score function

$$\mathfrak{T} = \left\{ \sum_{d=0}^1 \sum_{t=0}^1 \left(g_{d,t} S_y(y, d, t, x) + \frac{g_{d,t}}{p_{D=d, T=t}(x)} \dot{p}_{D=d, T=t}(x) \right) + S_x(x) \right\}$$

for any functions $S_y(y, d, t, x)$, $\dot{p}_{D=d, T=t}(x)$ and $S_x(x)$ that satisfy

$$\begin{aligned} \mathbb{E}[S_y(Y, d, t, X) | D = d, T = t, X] &= 0 \\ \mathbb{E} \left[\sum_{d=0}^1 \sum_{t=0}^1 \frac{G_{d,t}}{p_{D=d, T=t}(X)} \dot{p}_{D=d, T=t}(X) \right] &= \mathbb{E} \left[\sum_{d=0}^1 \sum_{t=0}^1 \dot{p}_{D=d, T=t}(X) \right] = 0 \\ \mathbb{E}[S_x(X)] &= 0 \end{aligned}$$

where the first and last equality follow by the mean zero property of the score function and the second equality by the fact that $\sum_{d=0}^1 \sum_{t=0}^1 \dot{p}_{D=d, T=t}(x) = 0$ since $\sum_{d=0}^1 \sum_{t=0}^1 p_{D=d, T=t}(x) = 1$.

The parameter θ is pathwise differentiable. For the parametric submodel we have

$$\theta(\beta) = \frac{\sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \int \int y f_{Y|D, T, X}(y | d, t, x; \beta) p_{D=1, T=1}(x; \beta) f_X(x; \beta) dy dx}{\int p_{D=1, T=1}(x; \beta) f_X(x; \beta) dx}$$

and at $\beta = \beta_0$

$$\begin{aligned} \frac{\partial \theta(\beta_0)}{\partial \beta} &= \frac{1}{p_{DT}} \left(\sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \int \int y S_y(y, d, t, x; \beta_0) f_{Y|D, T, X}(y | d, t, x) p_{D=1, T=1}(x) f_X(x) dy dx \right) \\ &\quad + \frac{1}{p_{DT}} \int (m_Y(x) - \theta) (\dot{p}_{D=1, T=1}(x; \beta_0) + p_{D=1, T=1}(x) S_x(x; \beta_0)) f_X(x) dx. \end{aligned}$$

We consider the function

$$\psi_{CS-1}^*(W; \theta) = \frac{p_{D=1, T=1}(X)}{p_{DT}} \left(\sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \frac{G_{d,t}}{p_{D=d, T=t}(X)} (Y - m_Y(d, t, X)) \right) + \frac{DT}{p_{DT}} (m_Y(X) - \theta).$$

Notice that $\psi_{CS-1}^*(W; \theta) \in \mathfrak{T}$. Also for any $D = d, T = t$ we obtain

$$\begin{aligned} &\mathbb{E} \left[\frac{p_{D=1, T=1}(X)}{p_{DT}} \frac{G_{d,t}}{p_{D=d, T=t}(X)} (Y - m_Y(d, t, X)) \times S(Y, D, T, X; \beta_0) \right] \\ &= \mathbb{E} \left[\frac{p_{D=1, T=1}(X)}{p_{DT}} \mathbb{E}[Y S_y(y, d, t, X; \beta_0) | D = d, T = t, X] \right] \\ &= \frac{1}{p_{DT}} \int \int y S_y(y, d, t, x; \beta_0) f_{Y|D, T, X}(y | d, t, x) p_{D=1, T=1}(x) f_X(x) dy dx \end{aligned}$$

which follows from the fact that $\mathbb{E}[S_y(y, d, t, X; \beta_0) | D = d, T = t, X] = 0$. Further,

$$\begin{aligned} &\mathbb{E} \left[\frac{DT}{p_{DT}} (m_Y(X) - \theta) \times S(Y, D, T, X; \beta_0) \right] \\ &= \frac{1}{p_{DT}} \mathbb{E}[(m_Y(X) - \theta) (\dot{p}_{D=1, T=1}(X; \beta_0) + p_{D=1, T=1}(X) S_x(X; \beta_0))] \\ &= \frac{1}{p_{DT}} \int (m_Y(x) - \theta) (\dot{p}_{D=1, T=1}(x; \beta_0) + p_{D=1, T=1}(x) S_x(x; \beta_0)) f_X(x) dx. \end{aligned}$$

It follows that $\psi_{CS-1}^*(W; \theta)$ is the efficient influence function and the variance lower bound for (CS-1) is $\mathbb{E}[\psi_{CS-1}^*(W; \theta)^2]$.

Proof for CS-2

Under the conditions in (CS-2)

$$S_{d,t}(d, t, x; \beta_0) = \left(\frac{d}{p_D(x)} - \frac{1-d}{1-p_D(x)} \right) \dot{p}_D(x; \beta_0) + \left(\frac{t}{p_T(x)} - \frac{1-t}{1-p_T(x)} \right) \dot{p}_T(x; \beta_0).$$

The tangent space of the model is characterised by the set of functions that are mean zero and satisfy the structure of the score function

$$\mathfrak{T} = \left\{ \sum_{d=0}^1 \sum_{t=0}^1 (g_{d,t} S_y(y, d, t, x)) + (d - p_D(x))a(x) + (t - p_T(x))b(x) + S_x(x) \right\}$$

for any functions $S_y(y, d, t, x)$ and $S_x(x)$ that satisfy

$$\begin{aligned} \mathbb{E}[S_y(Y, d, t, X) | D = d, T = t, X] &= 0 \\ \mathbb{E}[S_x(X)] &= 0 \end{aligned}$$

and any square integrable functions $a(x)$ and $b(x)$.

The parameter θ is pathwise differentiable. For the parametric submodel we have

$$\theta(\beta) = \frac{\sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \int \int y f_{Y|D,T,X}(y|d, t, x; \beta) p_D(x; \beta) p_T(x; \beta) f_X(x; \beta) dy dx}{\int p_D(x; \beta) p_T(x; \beta) f_X(x; \beta) dx}$$

and at $\beta = \beta_0$

$$\begin{aligned} \frac{\partial \theta(\beta_0)}{\partial \beta} &= \frac{1}{p_{DT}} \left(\sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \int \int y S_y(y, d, t, x; \beta_0) f_{Y|D,T,X}(y|d, t, x) p_D(x) p_T(x) f_X(x) dy dx \right) \\ &\quad + \frac{1}{p_{DT}} \int (m_Y(x) - \theta) (\dot{p}_D(x; \beta_0) p_T(x) + p_D(x) \dot{p}_T(x; \beta_0) + p_D(x) p_T(x) S_x(x; \beta_0)) f_X(x) dx. \end{aligned}$$

We consider the function

$$\begin{aligned} \psi_{CS-2}^*(W; \theta) &= \frac{p_D(X) p_T(X)}{p_{DT}} \left(\sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \frac{G_{d,t}}{p_{D=d}(X) p_{T=t}(X)} (Y - m_Y(d, t, X)) \right) \\ &\quad + \left(\frac{p_T(X)(D - p_D(X))}{p_{DT}} + \frac{p_D(X)(T - p_T(X))}{p_{DT}} + \frac{p_D(X) p_T(X)}{p_{DT}} \right) (m_Y(X) - \theta). \end{aligned}$$

Notice that $\psi_{CS-2}^*(W; \theta) \in \mathfrak{T}$. Also for any $D = d, T = t$ we obtain

$$\begin{aligned} \mathbb{E} \left[\frac{p_D(X) p_T(X)}{p_{DT}} \frac{G_{d,t}}{p_{D=d}(X) p_{T=t}(X)} (Y - m_Y(d, t, X)) \times S(Y, D, T, X; \beta_0) \right] \\ = \frac{1}{p_{DT}} \int \int y S_y(y, d, t, x; \beta_0) f_{Y|D,T,X}(y|d, t, x) p_D(x) p_T(x) f_X(x) dy dx. \end{aligned}$$

Further,

$$\begin{aligned}\mathbb{E}\left[\frac{p_T(X)D}{p_{DT}}(m_Y(X) - \theta) \times S(Y, D, T, X; \beta_0)\right] &= \frac{1}{p_{DT}}\mathbb{E}[(m_Y(X) - \theta)(\dot{p}_D(X; \beta_0)p_T(X) + p_D(X)p_T(X)S_x(X; \beta_0))] \\ \mathbb{E}\left[\frac{p_D(X)T}{p_{DT}}(m_Y(X) - \theta) \times S(Y, D, T, X; \beta_0)\right] &= \frac{1}{p_{DT}}\mathbb{E}[(m_Y(X) - \theta)(p_D(X)\dot{p}_T(X; \beta_0) + p_D(X)p_T(X)S_x(X; \beta_0))] \\ \mathbb{E}\left[\frac{p_D(X)p_T(X)}{p_{DT}}(m_Y(X) - \theta) \times S(Y, D, T, X; \beta_0)\right] &= \frac{1}{p_{DT}}\mathbb{E}[(m_Y(X) - \theta)(p_D(X)p_T(X)S_x(X; \beta_0))]\end{aligned}$$

such that

$$\begin{aligned}\mathbb{E}\left[\left(\frac{p_T(X)(D - p_D(X))}{p_{DT}} + \frac{p_D(X)(T - p_T(X))}{p_{DT}} + \frac{p_D(X)p_T(X)}{p_{DT}}\right)(m_Y(X) - \theta) \times S(Y, D, T, X; \beta_0)\right] \\ = \frac{1}{p_{DT}} \int (m_Y(x) - \theta)(\dot{p}_D(x; \beta_0)p_T(x) + p_D(x)\dot{p}_T(x; \beta_0) + p_D(x)p_T(x)S_x(x; \beta_0))f_X(x)dx.\end{aligned}$$

It follows that $\psi_{CS-2}^*(W; \theta)$ is the efficient influence function and the variance lower bound for (CS-2) is $\mathbb{E}[\psi_{CS-2}^*(W; \theta)^2]$.

Proof for CS-3

Under the conditions in (CS-3)

$$\begin{aligned}S_{d,t}(d, t, x; \beta_0) &= t \left(\frac{d}{p_D(1, x)} - \frac{1-d}{1-p_D(1, x)} \right) \dot{p}_D(1, x; \beta_0) + (1-t) \left(\frac{d}{p_D(0, x)} - \frac{1-d}{1-p_D(0, x)} \right) \dot{p}_D(0, x; \beta_0) \\ &\quad + \left(\frac{t}{p_T} - \frac{1-t}{1-p_T} \right) \dot{p}_T(\beta_0).\end{aligned}$$

The tangent space of the model is characterised by the set of functions that are mean zero and satisfy the structure of the score function

$$\mathfrak{T} = \left\{ \sum_{d=0}^1 \sum_{t=0}^1 (g_{d,t} S_y(y, d, t, x)) + t(d - p_D(1, x))a(x) + (1-t)(d - p_D(0, x))b(x) + (t - p_T)c + S_x(x) \right\}$$

for any functions $S_y(y, d, t, x)$ and $S_x(x)$ that satisfy

$$\mathbb{E}[S_y(Y, d, t, X) | D = d, T = t, X] = 0$$

$$\mathbb{E}[S_x(X)] = 0$$

and any square integrable functions $a(x)$ and $b(x)$ and some constant c .

The parameter θ is pathwise differentiable. For the parametric submodel we have

$$\theta(\beta) = \frac{\sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \int \int y f_{Y|D,T,X}(y|d, t, x; \beta) p_D(1, x; \beta) f_X(x; \beta) dy dx}{\int p_D(1, x; \beta) f_X(x; \beta) dx}$$

and at $\beta = \beta_0$

$$\begin{aligned} \frac{\partial \theta(\beta_0)}{\partial \beta} &= \frac{1}{p_D(1)} \left(\sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \int \int y S_y(y, d, t, x; \beta_0) f_{Y|D,T,X}(y|d, t, x) p_D(1, x) f_X(x) dy dx \right) \\ &\quad + \frac{1}{p_D(1)} \int (m_Y(x) - \theta) (\dot{p}_D(1, x; \beta_0) + p_D(1, x) S_x(x; \beta_0)) f_X(x) dx. \end{aligned}$$

We consider the function

$$\begin{aligned} \psi_{CS-3}^*(W; \theta) &= \frac{p_D(1, X)}{p_D(1)} \left(\sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \frac{G_{d,t}}{p_{D=d}(t, X) p_{T=t}} (Y - m_Y(d, t, X)) \right) \\ &\quad + \left(\frac{T(D - p_D(1, X))}{p_D(1) p_T} + \frac{p_D(1, X)}{p_D(1)} \right) (m_Y(X) - \theta). \end{aligned}$$

Notice that $\psi_{CS-3}^*(W; \theta) \in \mathfrak{T}$. Also for any $D = d, T = t$ we obtain

$$\begin{aligned} &\mathbb{E} \left[\frac{p_D(1, X)}{p_D(1)} \frac{G_{d,t}}{p_{D=d}(t, X) p_{T=t}} (Y - m_Y(d, t, X)) \times S(Y, D, T, X; \beta_0) \right] \\ &= \frac{1}{p_D(1)} \int \int y S_y(y, d, t, x; \beta_0) f_{Y|D,T,X}(y|d, t, x) p_D(1, x) f_X(x) dy dx. \end{aligned}$$

Further,

$$\begin{aligned} &\mathbb{E} \left[\frac{DT}{p_D(1) p_T} (m_Y(X) - \theta) \times S(Y, D, T, X; \beta_0) \right] \\ &= \frac{1}{p_D(1)} \mathbb{E} \left[(m_Y(X) - \theta) \left(\dot{p}_D(1, X; \beta_0) + p_D(1, X) \frac{\dot{p}_T(\beta_0)}{p_T} + p_D(1, X) S_x(X; \beta_0) \right) \right] \\ &\mathbb{E} \left[\frac{p_D(1, X) T}{p_D(1)} (m_Y(X) - \theta) \times S(Y, D, T, X; \beta_0) \right] = \frac{1}{p_D(1)} \mathbb{E} \left[(m_Y(X) - \theta) \left(p_D(1, X) \frac{\dot{p}_T(\beta_0)}{p_T} + p_D(1, X) S_x(X; \beta_0) \right) \right] \\ &\mathbb{E} \left[\frac{p_D(1, X)}{p_D(1)} (m_Y(X) - \theta) \times S(Y, D, T, X; \beta_0) \right] = \frac{1}{p_D(1)} \mathbb{E} [(m_Y(X) - \theta) (p_D(1, X) S_x(X; \beta_0))] \end{aligned}$$

such that

$$\begin{aligned} &\mathbb{E} \left[\left(\frac{T(D - p_D(1, X))}{p_D(1) p_T} + \frac{p_D(1, X)}{p_D(1)} \right) (m_Y(X) - \theta) \times S(Y, D, T, X; \beta_0) \right] \\ &= \frac{1}{p_D(1)} \int (m_Y(x) - \theta) (\dot{p}_D(1, x; \beta_0) + p_D(1, x) S_x(x; \beta_0)) f_X(x) dx. \end{aligned}$$

It follows that $\psi_{CS-3}^*(W; \theta)$ is the efficient influence function and the variance lower bound for (CS-3) is $\mathbb{E}[\psi_{CS-3}^*(W; \theta)^2]$.

Proof for CS-4

A similar proof can be found in Sant'Anna and Zhao (2020). We give it here for completeness.

Under the conditions in (CS-4)

$$S_{d,t}(d, t, x; \beta_0) = \left(\frac{d}{p_D(x)} - \frac{1-d}{1-p_D(x)} \right) \dot{p}_D(x; \beta_0) + \left(\frac{t}{p_T} - \frac{1-t}{1-p_T} \right) \dot{p}_T(\beta_0).$$

The tangent space of the model is characterised by the set of functions that are mean zero and satisfy the structure of the score function

$$\mathfrak{T} = \left\{ \sum_{d=0}^1 \sum_{t=0}^1 (g_{d,t} S_y(y, d, t, x)) + (d - p_D(x))a(x) + (t - p_T)b + S_x(x) \right\}$$

for any functions $S_y(y, d, t, x)$ and $S_x(x)$ that satisfy

$$\mathbb{E}[S_y(Y, d, t, X) | D = d, T = t, X] = 0$$

$$\mathbb{E}[S_x(X)] = 0$$

and any square integrable functions $a(x)$ and some constant b .

The parameter θ is pathwise differentiable. For the parametric submodel we have

$$\theta(\beta) = \frac{\sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \int \int y f_{Y|D,T,X}(y|d, t, x; \beta) p_D(x; \beta) f_X(x; \beta) dy dx}{\int p_D(x; \beta) f_X(x; \beta) dx}$$

and at $\beta = \beta_0$

$$\begin{aligned} \frac{\partial \theta(\beta_0)}{\partial \beta} &= \frac{1}{p_D} \left(\sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \int \int y S_y(y, d, t, x; \beta_0) f_{Y|D,T,X}(y|d, t, x) p_D(x) f_X(x) dy dx \right) \\ &\quad + \frac{1}{p_D} \int (m_Y(x) - \theta) (\dot{p}_D(x; \beta_0) + p_D(x) S_x(x; \beta_0)) f_X(x) dx. \end{aligned}$$

We consider the function

$$\psi_{CS-4}^*(W; \theta) = \frac{p_D(X)}{p_D} \left(\sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \frac{G_{d,t}}{p_{D=d}(X) p_{T=t}} (Y - m_Y(d, t, X)) \right) + \frac{D}{p_D} (m_Y(X) - \theta).$$

Notice that $\psi_{CS-4}^*(W; \theta) \in \mathfrak{T}$. Also for any $D = d, T = t$ we obtain

$$\begin{aligned} \mathbb{E} \left[\frac{p_D(X)}{p_D} \frac{G_{d,t}}{p_{D=d}(X) p_{T=t}} (Y - m_Y(d, t, X)) \times S(Y, D, T, X; \beta_0) \right] \\ = \frac{1}{p_D} \int \int y S_y(y, d, t, x; \beta_0) f_{Y|D,T,X}(y|d, t, x) p_D(x) f_X(x) dy dx. \end{aligned}$$

Further,

$$\mathbb{E} \left[\frac{D}{p_D} (m_Y(X) - \theta) \times S(Y, D, T, X; \beta_0) \right] = \frac{1}{p_D} \mathbb{E} [(m_Y(X) - \theta) (\dot{p}_D(X; \beta_0) + p_D(X) S_x(X; \beta_0))].$$

It follows that $\psi_{CS-4}^*(W; \theta)$ is the efficient influence function and the variance lower bound for (CS-4) is $\mathbb{E}[\psi_{CS-4}^*(W; \theta)^2]$.

Proof for CS-5

Under the conditions in (CS-5)

$$S_{d,t}(d, t, x; \beta_0) = \left(\frac{d}{p_D} - \frac{1-d}{1-p_D} \right) \dot{p}_D(\beta_0) + \left(\frac{t}{p_T} - \frac{1-t}{1-p_T} \right) \dot{p}_T(\beta_0).$$

The tangent space of the model is characterised by the set of functions that are mean zero and satisfy the structure of the score function

$$\mathfrak{T} = \left\{ \sum_{d=0}^1 \sum_{t=0}^1 (g_{d,t} S_y(y, d, t, x)) + (d - p_D)a + (t - p_T)b + S_x(x) \right\}$$

for any functions $S_y(y, d, t, x)$ and $S_x(x)$ that satisfy

$$\mathbb{E}[S_y(Y, d, t, X) | D = d, T = t, X] = 0$$

$$\mathbb{E}[S_x(X)] = 0$$

and the constants a and b .

The parameter θ is pathwise differentiable. For the parametric submodel we have

$$\theta(\beta) = \sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \int \int y f_{Y|D,T,X}(y|d, t, x; \beta) f_X(x; \beta) dy dx$$

and at $\beta = \beta_0$

$$\begin{aligned} \frac{\partial \theta(\beta_0)}{\partial \beta} &= \sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \int \int y S_y(y, d, t, x; \beta_0) f_{Y|D,T,X}(y|d, t, x) f_X(x) dy dx \\ &\quad + \sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \int \int y f_{Y|D,T,X}(y|d, t, x) S_x(X; \beta_0) f_X(x) dy dx. \end{aligned}$$

We consider the function

$$\psi_{CS-5}^*(W; \theta) = \sum_{d=0}^1 \sum_{t=0}^1 (-1)^{(d+t)} \frac{G_{d,t}}{p_{D=d} p_{T=t}} (Y - m_Y(d, t, X)) + m_Y(X) - \theta.$$

Notice that $\psi_{CS-5}^*(W; \theta) \in \mathfrak{T}$. Also for any $D = d, T = t$ we obtain

$$\begin{aligned} &\mathbb{E} \left[\frac{G_{d,t}}{p_{D=d} p_{T=t}} (Y - m_Y(d, t, X)) \times S(Y, D, T, X; \beta_0) \right] \\ &= \int \int y S_y(y, d, t, x; \beta_0) f_{Y|D,T,X}(y|d, t, x) f_X(x) dy dx. \end{aligned}$$

Further,

$$\mathbb{E}[(m_Y(X) - \theta) \times S(Y, D, T, X; \beta_0)] = \mathbb{E}[m_Y(X) S_x(X; \beta_0)].$$

It follows that $\psi_{CS-5}^*(W; \theta)$ is the efficient influence function and the variance lower bound for (CS-5) is $\mathbb{E}[\psi_{CS-5}^*(W; \theta)^2]$.

2.A.2 Proof of Corollary 2.1

From Theorem 2.1 we directly obtain

$$\begin{aligned} \mathbb{E} [\psi_{CS}^*(W; \theta)^2] &= \mathbb{E} \left[\frac{q_{CS;D=1,T=1}(X)^2}{q_{CS;DT}^2} \left(\mathbb{E} [\psi_{CS}^{*a}(W)^2|X] + \frac{q_{CS;DT}^2}{q_{CS;D=1,T=1}(X)^2} \mathbb{E} [\psi_{CS}^{*b}(W)^2|X] (m_Y(X) - \theta)^2 \right. \right. \\ &\quad \left. \left. + 2 \frac{q_{CS;DT}}{q_{CS;D=1,T=1}(X)} \mathbb{E} [\psi_{CS}^{*a}(W) \psi_{CS}^{*b}(W)|X] (m_Y(X) - \theta) \right) \right] \end{aligned}$$

where

$$\begin{aligned} \mathbb{E} [\psi_{CS}^{*a}(W)^2|X] &= \mathbb{E} \left[\sum_{d=0}^1 \sum_{t=0}^1 \frac{\text{Var}(Y|D=d, T=t, X)}{q_{CS;D=d, T=t}(X)} \right] \\ \mathbb{E} [\psi_{CS}^{*a}(W) \psi_{CS}^{*b}(W)|X] &= 0 \end{aligned}$$

for all settings (CS-1) to (CS-5). Further, we obtain

$$\begin{aligned} \mathbb{E} [\psi_{CS-1}^{*b}(W)^2|X] &= \frac{p_{D=1,T=1}(X)^2}{p_{DT}^2} \frac{1}{p_{D=1,T=1}(X)} \\ \mathbb{E} [\psi_{CS-2}^{*b}(W)^2|X] &= \frac{p_D(X)^2 p_T(X)^2}{p_{DT}^2} \left(\frac{1}{p_D(X)} + \frac{1}{p_T(X)} - 1 \right) \\ \mathbb{E} [\psi_{CS-3}^{*b}(W)^2|X] &= \frac{p_D(1, X)^2}{p_D(1)^2} \left(\frac{1}{p_D(1, X) p_T} - \frac{1}{p_T} + 1 \right) \\ \mathbb{E} [\psi_{CS-4}^{*b}(W)^2|X] &= \frac{p_D(X)^2}{p_D^2} \frac{1}{p_D(X)} \\ \mathbb{E} [\psi_{CS-5}^{*b}(W)^2|X] &= 1. \end{aligned}$$

Suppose that any of the assumptions in settings (CS-2)-(CS-5) is true then the efficiency bound of (CS-5) is higher by

$$\Delta_{CS-1, CS} = \mathbb{E} \left[\frac{q_{CS;D=1,T=1}(X)^2}{q_{CS;DT}^2} (m_Y(X) - \theta)^2 \left(\frac{1}{q_{CS;D=1,T=1}(X)} - \frac{q_{CS;DT}^2}{q_{CS;D=1,T=1}(X)^2} \mathbb{E} [\psi_{CS}^{*b}(W)^2|X] \right) \right].$$

Similar arguments can be made for the comparison of the other bounds.

2.A.3 Proof of Theorem 2.2

We observe the data $W = (Y(0), Y(1), D, X)$. For the distribution of the data consider a regular parametric submodel indexed by β . The density under the submodel can then be written as

$$f_W(w; \beta) = f_{Y(0), Y(1)|D, X}(y(0), y(1)|d, x; \beta) f_{D|X}(d|x; \beta) f_X(x; \beta)$$

which equals $f_W(w)$ at $\beta = \beta_0$.

The score function is defined as $S(y(0), y(1), d, x; \beta_0) = \frac{\partial \log f_W(w; \beta_0)}{\partial \beta}$ and we obtain

$$\begin{aligned} S(y(0), y(1), d, x; \beta_0) &= dS_{y(0), y(1)}(y(0), y(1), 1, x; \beta_0) + (1-d)S_{y(0), y(1)}(y(0), y(1), 0, x; \beta_0) \\ &\quad + S_d(d, x; \beta_0) + S_x(x; \beta_0) \end{aligned}$$

with

$$\begin{aligned} S_{y(0), y(1)}(y(0), y(1), d, x; \beta_0) &= \frac{\partial \log f_{Y(0), Y(1)|D, X}(y(0), y(1)|d, x; \beta_0)}{\partial \beta} \\ S_d(d, x; \beta_0) &= \frac{\partial \log f_{D|X}(d|x; \beta_0)}{\partial \beta} \\ S_x(x; \beta_0) &= \frac{\partial \log f_X(x; \beta_0)}{\partial \beta} \end{aligned}$$

where $S_d(d, x; \beta_0)$ again depends on the assumptions (PA-1) and (PA-2).

For all regular parametric submodels the variance lower bound for a model is the second moment of the projection of a function $\psi_{PA}^*(W; \theta)$ (with $\mathbb{E}[\psi_{PA}^*(W; \theta)] = 0$ and an existing second moment) on the tangent space $\mathfrak{T}$ that satisfies

$$\frac{\partial \theta(\beta_0)}{\partial \beta} = \mathbb{E}[\psi_{PA}^*(W; \theta) S(Y(0), Y(1), D, X; \beta_0)].$$

When $\psi_{PA}^*(W; \theta) \in \mathfrak{T}$, the projection on $\mathfrak{T}$ is the function itself and therefore the variance lower bound for the model is given by $\mathbb{E}[\psi_{PA}^*(W; \theta)^2]$.

Proof for PA-1

Under the conditions in (PA-1)

$$S_d(d, x; \beta_0) = \left(\frac{d}{p_D(x)} - \frac{1-d}{1-p_D(x)} \right) \dot{p}_D(x; \beta_0).$$

The tangent space of the model is characterised by the set of functions that are mean zero and satisfy the structure of the score function

$$\mathfrak{T} = \left\{ dS_{y(0), y(1)}(y(0), y(1), 1, x) + (1-d)S_{y(0), y(1)}(y(0), y(1), 0, x) + (d - p_D(x))a(x) + S_x(x) \right\}$$

for any functions $S_{y(0), y(1)}(y(0), y(1), d, x)$ and $S_x(x)$ that satisfy

$$\begin{aligned} \mathbb{E}[S_{y(0), y(1)}(Y(0), Y(1), d, X)|D = d, X] &= 0 \\ \mathbb{E}[S_x(X)] &= 0 \end{aligned}$$

and any square integrable function $a(x)$.

The parameter θ is pathwise differentiable. For the parametric submodel we have

$$\begin{aligned}\theta(\beta) &= \frac{\int \int \int (y(1) - y(0)) f_{Y(0), Y(1)|D, X}(y(0), y(1)|1, x; \beta) p_D(x; \beta) f_X(x; \beta) dy(0) dy(1) dx}{\int p_D(x; \beta) f_X(x; \beta) dx} \\ &\quad - \frac{\int \int \int (y(1) - y(0)) f_{Y(0), Y(1)|D, X}(y(0), y(1)|0, x; \beta) p_D(x; \beta) f_X(x; \beta) dy(0) dy(1) dx}{\int p_D(x; \beta) f_X(x; \beta) dx}\end{aligned}$$

and at $\beta = \beta_0$

$$\begin{aligned}\frac{\partial \theta(\beta_0)}{\partial \beta} &= \frac{1}{p_D} \int \int \int (y(1) - y(0)) S_{y(0), y(1)}(y(0), y(1), 1, x; \beta_0) f_{Y(0), Y(1)|D, X}(y(0), y(1)|1, x) p_D(x) \\ &\quad f_X(x) dy(0) dy(1) dx - \frac{1}{p_D} \int \int \int (y(1) - y(0)) S_{y(0), y(1)}(y(0), y(1), 0, x; \beta_0) \\ &\quad f_{Y(0), Y(1)|D, X}(y(0), y(1)|0, x) p_D(x) f_X(x) dy(0) dy(1) dx + \frac{1}{p_D} \int (m_{\Delta Y}(x) - \theta) \\ &\quad (\dot{p}_D(x; \beta_0) + p_D(x) S_x(x; \beta_0)) f_X(x) dx.\end{aligned}$$

We consider the function

$$\begin{aligned}\psi_{PA-1}^*(W; \theta) &= \frac{p_D(X)}{p_D} \left(\frac{D}{p_D(X)} (Y(1) - Y(0) - m_{\Delta Y}(1, X)) - \frac{1-D}{1-p_D(X)} (Y(1) - Y(0) - m_{\Delta Y}(0, X)) \right) \\ &\quad + \frac{D}{p_D} (m_{\Delta Y}(X) - \theta).\end{aligned}$$

Notice that $\psi_{PA-1}^*(W; \theta) \in \mathfrak{T}$. Also we obtain

$$\begin{aligned}\mathbb{E} &\left[\frac{p_D(X)}{p_D} \frac{D}{p_D(X)} (Y(1) - Y(0) - m_{\Delta Y}(1, X)) \times S(Y, D, X; \beta_0) \right] \\ &= \frac{1}{p_D} \int \int \int (y(1) - y(0)) S_{y(0), y(1)}(y(0), y(1), 1, x; \beta_0) f_{Y(0), Y(1)|D, X}(y(0), y(1)|1, x) p_D(x) f_X(x) \\ &\quad dy(0) dy(1) dx \\ \mathbb{E} &\left[\frac{p_D(X)}{p_D} \frac{1-D}{1-p_D(X)} (Y(1) - Y(0) - m_{\Delta Y}(0, X)) \times S(Y, D, X; \beta_0) \right] \\ &= \frac{1}{p_D} \int \int \int (y(1) - y(0)) S_{y(0), y(1)}(y(0), y(1), 0, x; \beta_0) f_{Y(0), Y(1)|D, X}(y(0), y(1)|0, x) p_D(x) f_X(x) \\ &\quad dy(0) dy(1) dx.\end{aligned}$$

Further,

$$\mathbb{E} \left[\frac{D}{p_D} (m_{\Delta Y}(X) - \theta) \times S(Y, D, X; \beta_0) \right] = \frac{1}{p_D} \int (m_{\Delta Y}(x) - \theta) (\dot{p}_D(x; \beta_0) + p_D(x) S_x(x; \beta_0)) f_X(x) dx.$$

It follows that $\psi_{PA-1}^*(W; \theta)$ is the efficient influence function and the variance lower bound for (PA-1) is $\mathbb{E}[\psi_{PA-1}^*(W; \theta)^2]$.

Proof for PA-2

Under the conditions in (PA-2)

$$S_d(d, x; \beta_0) = \left(\frac{d}{p_D} - \frac{1-d}{1-p_D} \right) \dot{p}_D(\beta_0).$$

The tangent space of the model is characterised by the set of functions that are mean zero and satisfy the structure of the score function

$$\mathfrak{T} = \left\{ dS_{y(0), y(1)}(y(0), y(1), 1, x) + (1-d)S_{y(0), y(1)}(y(0), y(1), 0, x) + (d-p_D)a + S_x(x) \right\}$$

for any functions $S_{y(0), y(1)}(y(0), y(1), d, x)$ and $S_x(x)$ that satisfy

$$\begin{aligned} \mathbb{E}[S_{y(0), y(1)}(Y(0), Y(1), d, X) | D = d, X] &= 0 \\ \mathbb{E}[S_x(X)] &= 0 \end{aligned}$$

and a constant a .

The parameter θ is pathwise differentiable. For the parametric submodel we have

$$\begin{aligned} \theta(\beta) &= \int \int (y(1) - y(0)) f_{Y(0), Y(1) | D, X}(y(0), y(1) | 1, x; \beta) f_X(x; \beta) dy(0) dy(1) dx \\ &\quad - \int \int (y(1) - y(0)) f_{Y(0), Y(1) | D, X}(y(0), y(1) | 0, x; \beta) f_X(x; \beta) dy(0) dy(1) dx \end{aligned}$$

and at $\beta = \beta_0$

$$\begin{aligned} \frac{\partial \theta(\beta_0)}{\partial \beta} &= \int \int (y(1) - y(0)) S_{y(0), y(1)}(y(0), y(1), 1, x; \beta_0) f_{Y(0), Y(1) | D, X}(y(0), y(1) | 1, x) f_X(x) dy(0) dy(1) dx \\ &\quad - \int \int (y(1) - y(0)) S_{y(0), y(1)}(y(0), y(1), 0, x; \beta_0) f_{Y(0), Y(1) | D, X}(y(0), y(1) | 0, x) f_X(x) dy(0) dy(1) dx \\ &\quad + \int m_{\Delta Y}(x) S_x(x; \beta_0) f_X(x) dx. \end{aligned}$$

We consider the function

$$\psi_{PA-2}^*(W; \theta) = \frac{D}{p_D} (Y(1) - Y(0) - m_{\Delta Y}(1, X)) - \frac{1-D}{1-p_D} (Y(1) - Y(0) - m_{\Delta Y}(0, X)) + m_{\Delta Y}(X) - \theta.$$

Notice that $\psi_{PA-2}^*(W; \theta) \in \mathfrak{T}$. Similarly to the proof for (PA-1) it can be shown that

$$\frac{\partial \theta(\beta_0)}{\partial \beta} = \mathbb{E}[\psi_{PA-2}^*(W; \theta) \times S(Y, D, X; \beta_0)].$$

It follows that $\psi_{PA-2}^*(W; \theta)$ is the efficient influence function and the variance lower bound for (PA-2) is $\mathbb{E}[\psi_{PA-2}^*(W; \theta)^2]$.

2.A.4 Proof of Corollary 2.3

Notice that for cross-sectional data we can write $Y = TY(1) + (1 - T)Y(0)$ and we have $m_Y(d, t, X) = \mathbb{E}[Y(t)|D = d, X]$. Further, under the conditions of a panel for case (CS-1) we obtain $p_{D=d, T=t}(X) = \frac{1}{2}p_D(X)$. Hence, when we observe the panel structure the efficiency bound for (CS-1) is

$$\mathbb{E} \left[2 \frac{p_D(X)^2}{p_D^2} \left(\sum_{d=0}^1 \sum_{t=0}^1 \frac{\text{Var}(Y(t)|D = d, X)}{p_{D=d}(X)} + \frac{(m_{\Delta Y}(X) - \theta)^2}{p_D(X)} \right) \right].$$

For

$$\Delta_{CS-1, PA-1} = \mathbb{E} \left[2 \frac{p_D(X)^2}{p_D^2} \left(\sum_{d=0}^1 \sum_{t=0}^1 \frac{\text{Var}(Y(t)|D = d, X)}{p_{D=d}(X)} + \frac{(m_{\Delta Y}(X) - \theta)^2}{p_D(X)} \right) \right] - \mathbb{E} [\psi_{PA-1}^*(W; \theta)^2]$$

the first claim immediately follows.

The second claim follows similarly by observing that when we observe the panel structure the efficiency bound for (CS-5) is

$$\mathbb{E} \left[2 \sum_{d=0}^1 \sum_{t=0}^1 \frac{\text{Var}(Y(t)|D = d, X)}{p_{D=d}} + (m_{\Delta Y}(X) - \theta)^2 \right].$$

2.B Proofs for Section 2.4

2.B.1 Lemma 2.3

The general result of the lemma can be similarly found for example in Chernozhukov et al. (2018a).

Lemma 2.3. *For a generic function $\psi(W, \cdot)$ let $\psi(W, \eta)$ be the version with true nuisance parameters and $\psi(W, \hat{\eta}_{-k})$ the version with cross-fitted nuisance parameters. Define $\bar{\psi} = \psi(W, \hat{\eta}_{-k}) - \psi(W, \eta)$. Then under the conditions that (i) $\mathbb{E}[\bar{\psi}] = o(N^{-\frac{1}{2}})$ and (ii) $\|\bar{\psi}\|_2 = o(1)$ the term $\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \bar{\psi}_i$ is bounded such that*

$$\left| \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \bar{\psi}_i \right| = o_p(N^{-\frac{1}{2}}).$$

Proof:

Write

$$\left| \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \bar{\psi}_i \right| \leq |\mathbb{E}[\bar{\psi}]| + \left| \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} (\bar{\psi}_i - \mathbb{E}[\bar{\psi}]) \right|$$

and notice that for every $k \in [1, \dots, K]$

$$\begin{aligned} \mathbb{E} \left[\left| \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{J}^k} (\bar{\psi}_i - \mathbb{E}[\bar{\psi}]) \right|^2 \right] &= \mathbb{E} \left[\left| \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{J}^k} (\bar{\psi}_i - \mathbb{E}[\bar{\psi}]) \right|^2 \middle| W_{i \in \mathcal{J}^k} \right] \\ &\leq \mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{J}^k} (\bar{\psi}_i - \mathbb{E}[\bar{\psi}]) \right)^2 \middle| W_{i \in \mathcal{J}^k} \right] \\ &\leq \mathbb{E} \left[(\psi(W, \hat{\eta}_{-k}) - \psi(W, \eta))^2 \right] \end{aligned}$$

such that by Markov's inequality

$$\left| \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} (\bar{\psi}_i - \mathbb{E}[\bar{\psi}]) \right| \leq \frac{1}{\sqrt{N}} \|\bar{\psi}\|_2.$$

Thus, $\sqrt{N} \mathbb{E}[\bar{\psi}] = o(1)$ and $\sqrt{N} \left| \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} (\bar{\psi}_i - \mathbb{E}[\bar{\psi}]) \right| = o_p(1)$ under the conditions (i) and (ii).

2.B.2 Proof of Theorem 2.3

We have to show that

$$\sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(W_i, \eta; \theta) + o_p(1).$$

This implies that $\hat{\theta}$ is asymptotically linear with influence function $\psi(W, \eta; \theta)$. Thus, given the above holds, combining Slutsky's Theorem and the Lindeberg-Lévy Central Limit Theorem implies the second claim of the Theorem.

In order to derive the result, write

$$\begin{aligned} \hat{\theta} - \theta &= \left(\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \hat{\eta}_{-k}) \right)^{-1} \times \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi(W_i, \hat{\eta}_{-k}; \theta) \\ &= \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi(W_i, \hat{\eta}_{-k}; \theta) + \left(\frac{1}{\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \eta)} - 1 \right) \times \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi(W_i, \hat{\eta}_{-k}; \theta) \\ &\quad + \left(\frac{1}{\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \hat{\eta}_{-k})} - \frac{1}{\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \eta)} \right) \times \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi(W_i, \hat{\eta}_{-k}; \theta) \end{aligned}$$

and define

$$\begin{aligned} i &= \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} (\psi(W_i, \hat{\eta}_{-k}; \theta) - \psi(W_i, \eta; \theta)) \\ ii &= \left(\frac{1}{\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \eta)} - 1 \right) \\ iii &= \left(\frac{1}{\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \hat{\eta}_{-k})} - \frac{1}{\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \eta)} \right). \end{aligned}$$

Suppose that $|i| = o_p(N^{-\frac{1}{2}})$, $|ii| = o_p(1)$, $|iii| = o_p(N^{-\frac{1}{2}})$ and notice that $\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(W_i, \eta; \theta) \right\|_2 = O(1)$ and

$$\sqrt{N}(\hat{\theta} - \theta) - \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(W_i, \eta; \theta) = \sqrt{N}i + \left(\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(W_i, \eta; \theta) + \sqrt{N}i \right) \times (ii + iii).$$

We obtain

$$\begin{aligned} \left| \sqrt{N}(\hat{\theta} - \theta) - \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(W_i, \eta; \theta) \right| &\leq o_p(1) + O_p(1)o_p(1) + O_p(1)o_p(N^{-\frac{1}{2}}) + o_p(1)o_p(1) \\ &\quad + o_p(1)o_p(N^{-\frac{1}{2}}) \\ &= o_p(1) \end{aligned}$$

using the fact that Markov's inequality yields $\left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(W_i, \eta; \theta) \right| = O_p(1)$. In the following we will derive the bounds for the different terms.

Bounding i

Notice that

$$\begin{aligned} \psi(W, \hat{\eta}_{-k}; \theta) - \psi(W, \eta; \theta) &= \underbrace{\psi^a(W, \hat{\eta}_{-k}) - \psi^a(W, \eta)}_{ia} + \underbrace{(\psi^b(W, \eta) - \psi^b(W, \hat{\eta}_{-k}))}_{ib} \theta. \end{aligned}$$

Using Assumption 2.10 (i) and (ii) and since

$$\begin{aligned} \mathbb{E}[\psi(W, \hat{\eta}_{-k}; \theta) - \psi(W, \eta; \theta)] &= \mathbb{E}[ia] + \mathbb{E}[ib]\theta \quad \text{and} \\ \|\psi(W, \hat{\eta}_{-k}; \theta) - \psi(W, \eta; \theta)\|_2 &\leq \|ia\|_2 + \|ib\|_2 \theta, \end{aligned}$$

Lemma 2.3 implies that $|i| = o_p(N^{-\frac{1}{2}})$ when $\mathbb{E}[ia] = o(N^{-\frac{1}{2}})$ and $\|ia\|_2 = o(1)$.

For a specific G_τ one can write

$$\begin{aligned} ia_{G_\tau} &= \underbrace{\frac{G_\tau}{q_1} \left(\frac{\hat{q}_1(X)_{-k}}{\hat{q}_{G_\tau}(X)_{-k}} - \frac{q_1(X)}{q_{G_\tau}(X)} \right) (m_{\tilde{Y}}(G_\tau = 1, X) - \hat{m}_{\tilde{Y}}(G_\tau = 1, X)_{-k})}_{ia_{G_\tau}1} \\ &\quad + \underbrace{\left(\frac{G_\tau}{q_1} \frac{q_1(X)}{q_{G_\tau}(X)} - \psi^b(W, \eta) \right) (m_{\tilde{Y}}(G_\tau = 1, X) - \hat{m}_{\tilde{Y}}(G_\tau = 1, X)_{-k})}_{ia_{G_\tau}2} \\ &\quad + \underbrace{\frac{G_\tau}{q_1} \left(\frac{\hat{q}_1(X)_{-k}}{\hat{q}_{G_\tau}(X)_{-k}} - \frac{q_1(X)}{q_{G_\tau}(X)} \right) (\tilde{Y} - m_{\tilde{Y}}(G_\tau = 1, X))}_{ia_{G_\tau}3} \\ &\quad + \underbrace{(\psi^b(W, \hat{\eta}_{-k}) - \psi^b(W, \eta)) \hat{m}_{\tilde{Y}}(G_\tau = 1, X)_{-k}}_{ia_{G_\tau}4}. \end{aligned}$$

Notice that

$$\begin{aligned}\mathbb{E}[ia_{G_\tau}1] &\leq C \left\| \frac{\hat{q}_1(X)_{-k}}{\hat{q}_{G_\tau}(X)_{-k}} - \frac{q_1(X)}{q_{G_\tau}(X)} \right\|_2 \times \|\hat{m}_{\tilde{Y}}(G_\tau = 1, X)_{-k} - m_{\tilde{Y}}(G_\tau = 1, X)\|_2 \\ &= o\left(N^{-\frac{1}{2}}\right)\end{aligned}$$

by Assumption 2.10 (v), $\mathbb{E}[ia_{G_\tau}2] = 0$ by Assumption 2.9 (i), $\mathbb{E}[ia_{G_\tau}3] = 0$ by the Law of Iterated Expectations and $\mathbb{E}[ia_{G_\tau}4] = o\left(N^{-\frac{1}{2}}\right)$ by Assumption 2.10 (iii).

Further, $\|ia_{G_\tau}\|_2 \leq \|ia_{G_\tau}1\|_2 + \|ia_{G_\tau}2\|_2 + \|ia_{G_\tau}3\|_2 + \|ia_{G_\tau}4\|_2$ and

$$\begin{aligned}\|ia_{G_\tau}1\|_2 &\leq \left\| \frac{G_\tau}{q_1} \left(\frac{\hat{q}_1(X)_{-k}}{\hat{q}_{G_\tau}(X)_{-k}} - \frac{q_1(X)}{q_{G_\tau}(X)} \right) \right\|_\infty \times \|\hat{m}_{\tilde{Y}}(G_\tau = 1, X)_{-k} - m_{\tilde{Y}}(G_\tau = 1, X)\|_2 \\ &\leq C \left\| \frac{\hat{q}_1(X)_{-k}}{\hat{q}_{G_\tau}(X)_{-k} q_{G_\tau}(X)} \right\|_\infty \times \|\hat{q}_{G_\tau}(X)_{-k} - q_{G_\tau}(X)\|_\infty \times \|\hat{m}_{\tilde{Y}}(G_\tau = 1, X)_{-k} - m_{\tilde{Y}}(G_\tau = 1, X)\|_2 \\ &\quad + C \left\| \frac{1}{q_{G_\tau}(X)} \right\|_\infty \times \|\hat{q}_1(X)_{-k} - q_1(X)\|_\infty \times \|\hat{m}_{\tilde{Y}}(G_\tau = 1, X)_{-k} - m_{\tilde{Y}}(G_\tau = 1, X)\|_2 \\ &= o(1)\end{aligned}$$

by Assumptions 2.8 and 2.10 (iv),

$$\begin{aligned}\|ia_{G_\tau}2\|_2 &\leq \left\| \frac{G_\tau}{q_1} \frac{q_1(X)}{q_{G_\tau}(X)} - \psi^b(W, \eta) \right\|_\infty \times \|\hat{m}_{\tilde{Y}}(G_\tau = 1, X)_{-k} - m_{\tilde{Y}}(G_\tau = 1, X)\|_2 \\ &= o(1)\end{aligned}$$

by Assumptions 2.9 (iii) and 2.10 (iv),

$$\begin{aligned}\|ia_{G_\tau}3\|_2 &\leq C \left\| \left(\frac{\hat{q}_1(X)_{-k}}{\hat{q}_{G_\tau}(X)_{-k}} - \frac{q_1(X)}{q_{G_\tau}(X)} \right) G_\tau (\tilde{Y} - m_{\tilde{Y}}(G_\tau = 1, X)) \right\|_2 \\ &= C \left(\mathbb{E} \left[\left(\frac{\hat{q}_1(X)_{-k}}{\hat{q}_{G_\tau}(X)_{-k}} - \frac{q_1(X)}{q_{G_\tau}(X)} \right)^2 G_\tau (\tilde{Y} - m_{\tilde{Y}}(G_\tau = 1, X))^2 \right] \right)^{\frac{1}{2}} \\ &\leq C \|q_{G_\tau}(X)\|_\infty \left(\mathbb{E} \left[\left(\frac{\hat{q}_1(X)_{-k}}{\hat{q}_{G_\tau}(X)_{-k}} - \frac{q_1(X)}{q_{G_\tau}(X)} \right)^2 \text{Var}(\tilde{Y} | G_\tau = 1, X) \right] \right)^{\frac{1}{2}} \\ &\leq C \left\| \frac{\hat{q}_1(X)_{-k}}{\hat{q}_{G_\tau}(X)_{-k}} - \frac{q_1(X)}{q_{G_\tau}(X)} \right\|_2 \\ &= o(1)\end{aligned}$$

by Assumptions 2.8, 2.7 (ii) and 2.10 (iv), and lastly

$$\begin{aligned}\|ia_{G_\tau}4\|_2 &\leq \left\| \psi^b(W, \hat{\eta}_{-k}) - \psi^b(W, \eta) \right\|_\infty \times \|\hat{m}_{\tilde{Y}}(G_\tau = 1, X)_{-k} - m_{\tilde{Y}}(G_\tau = 1, X)\|_2 \\ &\quad + \left\| \left(\psi^b(W, \hat{\eta}_{-k}) - \psi^b(W, \eta) \right) m_{\tilde{Y}}(G_\tau = 1, X) \right\|_2 \\ &= o(1)\end{aligned}$$

by Assumptions 2.10 (i), (iii) and (iv).

Bounding ii

$$\begin{aligned}
|ii| &\leq \left| 1 - \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \eta) \right| \times \left| \left(1 + \left(\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \eta) - 1 \right) \right)^{-1} \right| \\
&= o_p(1) (1 + o_p(1))^{-1} \\
&= o_p(1) O_p(1) \\
&= o_p(1)
\end{aligned}$$

since by a weak law of large numbers sample means should converge against their expectations and $\mathbb{E} [\psi^b(W, \eta)] = 1$ by Assumption 2.9 (i).

Bounding iii

$$\begin{aligned}
|iii| &\leq \left| \underbrace{\frac{1}{\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \hat{\eta}_{-k})}}_{iiia} \right| \times \left| \underbrace{\frac{1}{\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \eta)}}_{iiib} \right| \\
&\times \left| \underbrace{\frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \eta) - \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{J}^k} \psi^b(W_i, \hat{\eta}_{-k})}_{iiic} \right|
\end{aligned}$$

where $|iiia| \leq \|iiia\|_\infty \leq C$, $|iiib| \leq \|iiib\|_\infty \leq C$ by Assumption 2.9 (ii). Further, conditions in Assumption 2.10 (i) and (ii) in combination with Lemmma 2.3 implies that $|iiic| = o_p(N^{-\frac{1}{2}})$ with high probability. Therefore $|iii| = O(1)O(1)o_p(N^{-\frac{1}{2}}) = o_p(N^{-\frac{1}{2}})$.

2.B.3 Proof of Corollary 2.4

$\mathbb{E} [\psi_{CS}^*(W, \eta; \theta)] = 0$ using Assumptions 2.3 and the respective condition in settings (CS-1)-(CS-5). For all $\psi_{CS}^{*b}(W, \eta)$ it is easy to see that the conditions in Assumption 2.9 are satisfied. Therefore it remains to show that the conditions in Assumptions 2.10 hold under the convergence conditions stated in the corollary.

- (a) For $\psi_{CS-1}^{*b}(W, \eta) = \frac{DT}{pDT}$ the conditions in Assumption 2.10 (i)-(iii) are trivially satisfied. Notice that $m_Y(1, 1, X)$ is redundant and that for all $(d, t) \in \{(0, 1), (1, 0), (0, 0)\}$

$$\begin{aligned}
&\left\| \frac{\hat{p}_{D=1, T=1}(X)_{-k}}{\hat{p}_{D=d, T=t}(X)_{-k}} - \frac{p_{D=1, T=1}(X)}{p_{D=d, T=t}(X)} \right\|_2 \\
&\leq \left\| \frac{1}{\hat{p}_{D=d, T=t}(X)_{-k}} \right\|_\infty \times \|\hat{p}_{D=1, T=1}(X)_{-k} - p_{D=1, T=1}(X)\|_2 \\
&+ \left\| \frac{p_{D=1, T=1}(X)}{\hat{p}_{D=d, T=t}(X)_{-k} p_{D=d, T=t}(X)} \right\|_\infty \times \|\hat{p}_{D=d, T=t}(X)_{-k} - p_{D=d, T=t}(X)\|_2 \\
&\leq C \times (\epsilon_{p_{D=1, T=1}(X)} + \epsilon_{p_{D=d, T=t}(X)})
\end{aligned}$$

by Assumption 2.8. This immediately implies that Assumptions 2.10 (iv) and (v) are satisfied under the

conditions stated.

(b) For $\psi_{CS-2}^{*b}(W, \eta) = \frac{Dp_T(X) + Tp_D(X) - p_D(X)p_T(X)}{p_{DT}}$ write

$$\begin{aligned} & \psi_{CS-2}^{*b}(W, \hat{\eta}_{-k}) - \psi_{CS-2}^{*b}(W, \eta) \\ &= \frac{1}{p_{DT}} (D - p_D(X)) (\hat{p}_T(X)_{-k} - p_T(X)) + \frac{1}{p_{DT}} (T - p_T(X)) (\hat{p}_D(X)_{-k} - p_D(X)) \\ &+ \frac{1}{p_{DT}} (\hat{p}_T(X)_{-k} - p_T(X)) (p_D(X) - \hat{p}_D(X)_{-k}). \end{aligned}$$

Then $\|\psi_{CS-2}^{*b}(W, \hat{\eta}_{-k}) - \psi_{CS-2}^{*b}(W, \eta)\|_2 = o(1)$ by Assumption 2.8 and the convergence conditions stated. Similarly,

$$\left\| \psi_{CS-2}^{*b}(W, \hat{\eta}_{-k}) - \psi_{CS-2}^{*b}(W, \eta) \right\|_{\infty} \leq C + \|\hat{p}_T(X)_{-k} - p_T(X)\|_{\infty} \times \|\hat{p}_D(X)_{-k} - p_D(X)\|_{\infty} = O(1)$$

by Assumption 2.8. This verifies Assumption 2.10 (i).

Also $\mathbb{E} [\psi_{CS-2}^{*b}(W, \hat{\eta}_{-k}) - \psi_{CS-2}^{*b}(W, \eta)] \leq \varepsilon_{p_T(X)} \times \varepsilon_{p_D(X)} = o\left(N^{-\frac{1}{2} \frac{r}{r-1}}\right)$ which verifies Assumption 2.10 (ii).

To verify Assumption 2.10 (iii), notice that for $r > 2$

$$\begin{aligned} & \|(\hat{p}_T(X)_{-k} - p_T(X)) (p_D(X) - \hat{p}_D(X)_{-k})\|_{\frac{r}{r-1}} \\ &= \left(\mathbb{E} \left[|(\hat{p}_T(X)_{-k} - p_T(X))|^{\frac{r}{r-1}} |(p_D(X) - \hat{p}_D(X)_{-k})|^{\frac{r}{r-1}} \right] \right)^{\frac{r-1}{r}} \\ &\leq \|\hat{p}_D(X)_{-k} - p_D(X)\|_{\infty}^{\frac{1}{r}} \times \|\hat{p}_T(X)_{-k} - p_T(X)\|_{\infty}^{\frac{1}{r}} \times (\varepsilon_{p_T(X)} \times \varepsilon_{p_D(X)})^{\frac{r-1}{r}} \\ &= o\left(N^{-\frac{1}{2}}\right). \end{aligned}$$

Jensen's inequality implies that $\|m_Y(d, t, X)\|_r \leq C \|Y\|_r = O(1)$ by Assumption 2.7 (i). Then for any $d, t \in \{0, 1\}$

$$\begin{aligned} & \mathbb{E} \left[\left(\psi_{CS-2}^{*b}(W, \hat{\eta}_{-k}) - \psi_{CS-2}^{*b}(W, \eta) \right) (\hat{m}_Y(d, t, X)_{-k} - m_Y(d, t, X)) \right] \\ &\leq C \|(\hat{p}_T(X)_{-k} - p_T(X)) (p_D(X) - \hat{p}_D(X)_{-k})\|_{\frac{r}{r-1}} \times \|\hat{m}_Y(d, t, X)_{-k} - m_Y(d, t, X)\|_r \\ &= o\left(N^{-\frac{1}{2}}\right), \\ & \mathbb{E} \left[\left(\psi_{CS-2}^{*b}(W, \hat{\eta}_{-k}) - \psi_{CS-2}^{*b}(W, \eta) \right) m_Y(d, t, X) \right] \\ &\leq C \|(\hat{p}_T(X)_{-k} - p_T(X)) (p_D(X) - \hat{p}_D(X)_{-k})\|_{\frac{r}{r-1}} \times \|m_Y(d, t, X)\|_r \\ &= o\left(N^{-\frac{1}{2}}\right) \end{aligned}$$

under the conditions stated in the corollary. Further,

$$\begin{aligned} & \|(D - p_D(X)) (\hat{p}_T(X)_{-k} - p_T(X))\|_{\frac{2r}{r-2}} \leq C \|\hat{p}_T(X)_{-k} - p_T(X)\|_{\frac{2r}{r-2}} \\ &\leq C \|\hat{p}_T(X)_{-k} - p_T(X)\|_{\infty}^{\frac{2}{r}} \times \|\hat{p}_T(X)_{-k} - p_T(X)\|_2^{\frac{r-2}{r}} \\ &= o(1), \end{aligned}$$

similarly $\|(T - p_T(X))(\hat{p}_D(X)_{-k} - p_D(X))\|_{\frac{2r}{r-2}} = o(1)$ and

$$\begin{aligned} & \|(\hat{p}_T(X)_{-k} - p_T(X))(p_D(X) - \hat{p}_D(X)_{-k})\|_{\frac{2r}{r-2}} \\ & \leq \|\hat{p}_T(X)_{-k} - p_T(X)\|_{\infty}^{\frac{r+2}{2r}} \times \|\hat{p}_D(X)_{-k} - p_D(X)\|_{\infty}^{\frac{r+2}{2r}} \times (\varepsilon_{p_T(X)} \times \varepsilon_{p_D(X)})^{\frac{r-2}{2r}} \\ & = o(1) \end{aligned}$$

by Assumption 2.8, the fact that $r > 2$ and the convergence conditions stated. Therefore, for all $d, t \in \{0, 1\}$

$$\begin{aligned} \left\| \left(\psi_{CS-2}^{*b}(W, \hat{\eta}_{-k}) - \psi_{CS-2}^{*b}(W, \eta) \right) m_Y(d, t, X) \right\|_2 & \leq \left\| \psi_{CS-2}^{*b}(W, \hat{\eta}_{-k}) - \psi_{CS-2}^{*b}(W, \eta) \right\|_{\frac{2r}{r-2}} \times \|m_Y(d, t, X)\|_r \\ & = o(1). \end{aligned}$$

Assumption 2.10 (iv) and (v) directly follow from the conditions stated.

(c) For $\psi_{CS-3}^{*b}(W, \eta) = \frac{T(D - p_D(1, X))}{p_D(1)p_T} + \frac{p_D(1, X)}{p_D(1)}$ write

$$\psi_{CS-3}^{*b}(W, \hat{\eta}_{-k}) - \psi_{CS-3}^{*b}(W, \eta) = \frac{1}{p_D(1)p_T} ((p_T - T)(\hat{p}_D(1, X)_{-k} - p_D(1, X))).$$

Since

$$\begin{aligned} \left\| \psi_{CS-3}^{*b}(W, \hat{\eta}_{-k}) - \psi_{CS-3}^{*b}(W, \eta) \right\|_2 & \leq C \varepsilon_{p_D(1, X)} = o(1) \quad \text{and} \\ \left\| \psi_{CS-3}^{*b}(W, \hat{\eta}_{-k}) - \psi_{CS-3}^{*b}(W, \eta) \right\|_{\infty} & \leq C \|\hat{p}_D(1, X)_{-k} - p_D(1, X)\|_{\infty} = O(1), \end{aligned}$$

Assumption 2.10 (i) is verified.

Assumptions 2.10 (ii) holds since trivially $\mathbb{E} [\psi_{CS-3}^{*b}(W, \hat{\eta}_{-k}) - \psi_{CS-3}^{*b}(W, \eta)] = 0$ by the Law of Iterated Expectations. Similarly, for all $d, t \in \{0, 1\}$ the first part of Assumption 2.10 (iii) is verified by

$$\mathbb{E} \left[\left(\psi_{CS-3}^{*b}(W, \eta) - \psi_{CS-3}^{*b}(W, \hat{\eta}_{-k}) \right) \hat{m}_Y(d, t, X)_{-k} \right] = 0.$$

For the second part of Assumption 2.10 (iii) write

$$\begin{aligned} \left\| \left(\psi_{CS-3}^{*b}(W, \eta) - \psi_{CS-3}^{*b}(W, \hat{\eta}_{-k}) \right) m_Y(d, t, X) \right\|_2 & \leq C \|\hat{p}_D(1, X)_{-k} - p_D(1, X)\|_{\frac{2r}{r-2}} \times \|m_Y(d, t, X)\|_r \\ & \leq C \|\hat{p}_D(1, X)_{-k} - p_D(1, X)\|_{\infty}^{\frac{2}{r}} \times (\varepsilon_{p_D(1, X)})^{\frac{r-2}{r}} \\ & = o(1) \end{aligned}$$

for all $d, t \in \{0, 1\}$ using Assumption 2.8, the fact that $r > 2$ and the convergence conditions stated.

Assumptions 2.10 (iv) and (v) directly follow from the conditions stated.

(d) For $\psi_{CS-4}^{*b}(W, \eta) = \frac{D}{p_D}$ the conditions in Assumption 2.10 (i)-(iii) are trivially satisfied. Assumptions 2.10 (iv) and (v) directly follow from the conditions stated.

(e) For $\psi_{CS-5}^{*b}(W, \eta) = 1$ the conditions in Assumption 2.10 (i)-(iii) are trivially satisfied. Assumptions 2.10 (iv) and (v) directly follow from the conditions stated.

2.B.4 Proof of Corollary 2.5

$\mathbb{E}[\psi_{PA}^*(W, \eta; \theta)] = 0$ using Assumptions 2.6 and the respective conditions (PA-1) or (PA-2). For all $\psi_{PA}^{*b}(W, \eta)$ it is easy to see that the conditions in Assumption 2.9 and Assumption 2.10 (i)-(iii) are satisfied. Assumptions 2.10 (iv) and (v) directly follow from the conditions stated in the corollary.

2.B.5 Proof of Corollary 2.7

Assumption 2.9 and Assumption 2.10 (i)-(iii) are trivially satisfied when using $\psi'_{CS-2}(W; \theta)$. Notice that $m_Y(1, 1, X)$ is redundant. Assumptions 2.10 (iv) and (v) then directly follow from the conditions stated in the corollary.

2.B.6 Proof of Corollary 2.8

Assumption 2.9 and Assumption 2.10 (i)-(iii) follow similarly to the proof of Corollary 2.4 for $(d, t) \in \{(0, 1), (0, 0)\}$. Notice that $m_Y(1, 1, X)$ and $m_Y(1, 0, X)$ are redundant. Assumption 2.9 and Assumption 2.10 (i)-(iii) are then trivially satisfied for $(d, t) \in \{(1, 1), (1, 0)\}$. Since $m_Y(1, 1, X)$ and $m_Y(1, 0, X)$ are redundant, Assumptions 2.10 (iv) and (v) then directly follow from the conditions stated in the corollary.

Chapter 3

Nonparametric estimation of causal heterogeneity under high-dimensional confounding

jointly with Michael Lechner

This study considers the practically important case of nonparametrically estimating heterogeneous average treatment effects that vary with a limited number of discrete and continuous covariates in a selection-on-observables framework where the number of possible confounders is very large. We propose a two-step estimator for which the first step is estimated by machine learning. We show that this estimator has desirable statistical properties like consistency, asymptotic normality and rate double robustness. In particular, we derive the coupled convergence conditions between the nonparametric and the machine learning steps. We also show that estimating population average treatment effects by averaging the estimated heterogeneous effects is semiparametrically efficient. The resulting estimators are compared to other suggestions in the literature in Monte Carlo experiments that are inspired by real data. They are found to perform relatively better in most settings. The new estimators are applied to the empirical example of the effects of mothers' smoking during pregnancy on the resulting birth weight.

JEL classification: C14, C21

Keywords: causal machine learning, effect heterogeneity, group average treatment effects, semiparametric efficiency, ensemble learning, Empirical Monte Carlo

3.1 Introduction

Recently, new machine learning based estimators showed immense potential to systematically uncover causal effect heterogeneity leading to a rapidly growing literature on this topic (e.g., see the overviews in Athey and Imbens, 2017, 2019 and Knaus, Lechner, and Strittmatter, 2020b). In the context of heterogeneity, the respective aggregation levels for which the heterogeneity is estimated is playing an important role. Most papers focus on a selection-on-observable framework and investigate estimators for the heterogeneity at the lowest aggregation level to uncover possible heterogeneities to the largest extent possible. While this finest level of causal granularity is obviously of interest, Chernozhukov, Fernández-Val, and Luo (2018b) and Lechner (2018) argue to analyse heterogeneity also at higher levels, so called ‘Group Average Treatment Effects’ (GATEs). Such aggregates can be estimated more precisely, may be more easily interpretable by researchers in substantive terms, and may be more useful for decision makers. In particular, some fine-grained heterogeneities come with the limitation that the systematic drivers of the heterogeneity are difficult to understand. Furthermore, such rules may depend on variables that may be undesirable as decision variables for fairness reasons (e.g. sex, race, etc.), but are needed to control for confounding. Therefore, decision-makers are often interested in effect heterogeneities based on a rather small subset of a-priori determined covariates. This study suggests an approach that is based on statistical-learning assisted estimation of the GATEs for the various discrete and continuous variables of interest, and subsequent nonparametric aggregation of the GATEs to obtain ‘Average Treatment Effects’ (ATEs).¹

In effect heterogeneity analysis covariates do not (only) serve the purpose of making identifying assumptions credible. They become part of the outcome analysis by discriminating different subgroups of units for which an average effect is of interest. Further, whenever new observations enter the sample these covariates could be used to predict ‘their’ causal effect for a pre-specified aggregation level. The set of covariates to be included in the statistical model to explore effect heterogeneity is a statistical as well as a substantive decision.

The estimation of subgroup specific effects is a tedious task when there is confounding. In such settings, causal effects are typically only identified if the researcher includes the confounding covariates in the statistical model as well. Hence, the identifying assumptions dictate the inclusion of the set of covariates required. In empirical research based on selection-on-observables assumptions, the credibility of causal effects estimation often depends on a very large set of possible covariates with very many possible functional forms.

There are currently several suggestions to flexibly estimate heterogeneous effects when there is confounding. The general concept of averaging causal effects conditional on covariates² already dates back to Hahn (1998). He suggests to estimate a nonparametric outcome regression on the set of covariates that needs to be controlled for. Averaging over the conditional means leads to estimators of ATE that attain the semiparametric efficiency bound. In practice, however, nonparametric regression with many covariates is hardly feasible because the convergence rate of nonparametric methods exponentially decreases with the number of covariates included. Recently, Wager and Athey (2018) follow the same ideas as in Hahn (1998) but use Causal Forests instead of standard nonparametric regression. Athey, Tibshirani, and Wager (2019) and Lechner (2018) modify the Random Forest algorithm to better adjust for confounding and improve precision. Outcome-based models adjust for confounding and infer heterogeneous effects in a single estimation step. Therefore, in all of these approaches inference for effect heterogeneity relies on a dimension of the covariate space that is small relative to the sample size. Given the previous

¹Lechner (2018) also proposed this aggregation idea. However, that paper considered only a version of a Causal Forest while here we are in principle agnostic with respect to the machine learning method used. Furthermore, it considered only GATEs based on discrete variables and thus GATEs were obtained as unweighted within-cell means.

²We avoid the imprecise term ‘Conditional Average Treatment Effect’ because it is unclear which conditioning set is actually meant.

discussion, this might be a very strong assumption as, depending on the specific approach used, it might limit severely the way how different measurements and functional forms of the covariates enter the particular estimators.

In this study we follow an alternative approach in the literature. The two distinct roles of the covariates – adjusting for confounding and estimating heterogeneous effects – are explicitly reflected in a two-step estimation procedure for the GATEs. This idea is conceptionally not new in the literature.³ In the context of difference-in-differences estimation, Abadie (2005) shows that propensity scores weighted outcomes can be used as a dependent variable in a second stage regression on the covariates that are of interest for heterogeneous effects. Abrevaya, Hsu, and Lieli (2015) use a similar idea in the standard selection-on-observables setting. They provide inferential results for nonparametric and parametric propensity score first stages with nonparametric second stages. In line with other results in the literature on average effects (Hirano et al., 2003; Robins et al., 1994; Lunceford and Davidian, 2004), they show that the variance for Inverse Probability Weighting (IPW) estimators can be substantially decreased when the propensity score is estimated nonparametrically. Since their second stage also relies on nonparametric regression, the validity of their asymptotic results requires jointly choosing two kernel bandwidths which have to be in a rather small feasible interval. Lee, Okui, and Whang (2016) augment the model of Abrevaya et al. (2015) by including outcome projections (Augmented IPW, AIPW) and show that when both the propensity score and the outcome projections are estimated parametrically, one can treat the nuisance parameters as if they were known. Their asymptotic results with parametric first stages are then equivalent⁴ to those of Abrevaya et al. (2015) with nonparametric propensity score estimation. Still, the parametric nuisance models used could lead to substantial misspecification bias if the functional form does not coincide with the unknown data generating process (DGP). Moreover, if there are more potential regressors than observations, even parametric estimators collapse.

In their contributions to ATE estimation, Belloni et al. (2014) and Chernozhukov et al. (2018a) show how AIPW type estimators can be adapted to settings where the dimension of the relevant confounders grows with the sample size at a certain rate. They use various machine learning methods to estimate the propensity score and the outcome projections. Recently, Semenova and Chernozhukov (2020) use their framework to estimate effect heterogeneity based on linear models. They provide conditions under which their second-stage linear model can become increasingly flexible. In an alternative approach Nie and Wager (2020) propose to account for confounding using Robinson’s (1988) partially linear model.

Postulating nonparametric second stages, we do not assume any specific functional form of the GATE. We contribute to the literature on GATE estimation by allowing for flexible functional forms as well as a high-dimensional⁵ confounder space. This enables our proposed estimator to be robust against functional form misspecification and to remain consistent even if the number of covariates relative to the sample size is large. In particular, we provide a generic statistical framework such that the convergence rate requirements of the first stage nuisance estimation are coupled with the kernel bandwidth second stage nonparametric convergence.

³A few days before a first version of this work appeared on arXiv, Fan, Hsu, Lieli, and Zhang (2020) published their independent work on arXiv (up to that moment unknown to us) that uses similar ideas about aggregation and machine learning. We highlight the following differences between their and our contribution. While Fan et al. (2020) also motivate their approach with semiparametric efficiency theory, we explicitly derive the asymptotic properties of the implied ATE estimator. Further, we note that while their cross-fitting based estimator requires distinct nonparametric regressions in every fold, our GATE estimator is also based on cross-fitted first stage predictions but only requires a single second stage nonparametric regression. Additionally, we assess the performance of our proposed methods using Empirical Monte Carlo experiments where DGPs are to a large extent determined by real data. Arguably, this allows us to draw conclusions about the finite sample performance that do not depend on arbitrary choices of synthetic DGPs.

⁴While Abrevaya et al. (2015) use local constant nonparametric regression, Lee et al. (2016) show their results with local linear nonparametric regression.

⁵In general the term ‘high-dimensional’ refers to the fact that the dimension of the covariates in the model can be much larger than the sample size.

Additionally, we link our identification and estimation result to semiparametric efficiency theory by providing a new estimator for the ATE that can be estimated as a by-product of the GATEs. The estimator aggregates over all point estimates of the GATEs. We show that under certain convergence conditions for the kernel bandwidth, asymptotically it hits the variance lower bound of the semiparametric estimation problem. We therefore also contribute to the small literature on three-step semiparametric ATE estimation. Specifically, Hahn and Ridder (2013) (for an alternative theoretical development see also Mammen, Rothe, and Schienle, 2012) investigate a related set-up showing that nonparametric regression on an estimated propensity score can lead to efficient estimation of the ATE. To the best of our knowledge, this study is, however, the first that analyses the asymptotic properties of averaging a transformed outcome projection instead of the outcome projection on the propensity score. In a Monte Carlo experiment that is inspired by real data we show that our three-step estimator can have better finite-sample properties than two-step estimators that share the same first-order asymptotic properties (Robins et al., 1994; Hirano et al., 2003; Chernozhukov et al., 2018a) because the propensity score weights are subject to an additional smoothing step. Unlike propensity score matching, IPW or Hahn’s (1998) estimator, the proposed ATE estimator remains feasible when the dimension of the confounders entering the model is high.

After providing some more information on the theoretical background in Section 3.2, we present the details of our main asymptotic results in Section 3.3. We discuss some implementation choices for our proposed estimators in Section 3.4. The sensitivity towards these choices as well as the relative performance against other estimators is assessed in a Monte Carlo study in Section 3.5. An empirical example in Section 3.6 compares different alternative estimators for GATE and ATE and illustrates the applicability and usefulness of the new methods. The last section concludes. Formal proofs, some further results of the Monte Carlo study, and additional details on the empirical implementation are relegated to the Appendix.

3.2 Methodology

3.2.1 Notation

Denote with uppercase letters a variable and with lowercase letters its realizations. Then Y is the outcome variable and D is the binary treatment of interest. To describe causal effects, we use Rubin’s (1974) potential outcome notation such that Y^d is the outcome that would have been observed under treatment $D = d$. Further, X is a vector of observed covariates with support $\mathcal{X}$ and $Z \subseteq X$ as a set of predefined variables where the researcher is interested in effect heterogeneity with support $\mathcal{Z}$. Also let $X \in \mathbb{R}^{\dim X}$ and $Z \in \mathbb{R}^{\dim Z}$ and denote $\lambda_X = \dim X$ and $\lambda_Z = \dim Z$. Potentially we have that $\lambda_X \gg N$ whereas we require that $\lambda_Z \ll N$. Hence, we explicitly allow for models where the dimension of X is high-dimensional but the dimension of the subset of covariates that is of interest for the heterogeneity analysis Z is much smaller than the sample size. Similarly to the outcome, denote by X^d the potential value of X that would have been observed under treatment status $D = d$. Suppose that we observe an independent and identically distributed random sample $\{W_i\}_{i=1}^N$ with sample size N where $W_i = (Y_i, D_i, X_i, Z_i)$. We remain agnostic about the underlying cumulative distribution from which the sample of W is drawn $F = F(W)$ and just assume that it exists with density $f = f(W)$.

3.2.2 Semiparametric efficiency theory

The main parameter of interest in this study is the GATE defined as

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

Let $\psi(W, \cdot)$ be an outcome transforming function that adjust for non-random selection into the treatment (see the references mentioned in Section 3.1). Since we want to avoid usually unrealistic parametric assumptions on the underlying DGP, we allow for a flexible $\psi(W, \cdot)$ such that GATE is identified as

$$\tau(z) = \mathbb{E}[\psi(W, \cdot) | Z = z]$$

using the variables observed or functions of them. It is a priori unclear which one is a ‘good’ transformation $\psi(W, \cdot)$ in the sense that it achieves the variance lower bound for the problem. Hence, ideally our exposition would start by deriving the semiparametric efficiency bound for the problem at hand and then use the moment condition implied as an estimand for GATE. However, since parameters using ‘last stage’ nonparametric projections are not pathwise differentiable, standard semiparametric efficiency bounds cannot be derived following established theory (e.g. Bickel et al., 1993; Newey, 1994b; Hahn, 1998; Tsiatis, 2006). This was also noted for different problems in Rubin and Laan (2007) and Kennedy, Ma, McHugh, and Small (2017) and recently for similar problems in Semenova and Chernozhukov (2020), Foster and Syrgkanis (2020) and Kennedy (2020). Instead of directly relying on an efficiency result for GATE, we use that

$$\text{ATE} = \theta = \mathbb{E}[\mathbb{E}[\psi(W, \cdot) | Z]]$$

and plug in Hahn’s (1998) efficient score function for ATE

$$\psi(W, p, m_0, m_1) = \frac{D(Y - m_1(X))}{p(X)} - \frac{(1 - D)(Y - m_0(X))}{1 - p(X)} + m_1(X) - m_0(X)$$

where $p(X) = \mathbb{E}[D|X]$ denotes the propensity score and $m_d(X) = \mathbb{E}[Y|X, D = d]$ for $d \in \{0, 1\}$ denotes the conditional expectations of the outcome in the treatment-specific subpopulations.

3.2.3 Parameter identification

For identification of GATE and ATE we make the following assumptions that characterise the selection-on-observables design (e.g. Rosenbaum and Rubin, 1983).

Assumption 3.1 (Identification).

- (i) *(Conditional Independence)* For any $x \in \mathcal{X}$, $Y^0, Y^1 \perp D | X = x$;
- (ii) *(Observational Rule)* For each observation i , the outcome $Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$ is observed;
- (iii) *(Exogeneity of confounders)* For each observation i , $X_i = X_i^1 = X_i^0$;
- (iv) *(Common support)* For any $x \in \mathcal{X}$, the propensity score $p(x)$ is bounded away from zero and one.

Under Assumption 3.1 it is standard to show identification results for the parameters of interest. In particular, for GATE we have

$$\begin{aligned}
\tau(z) &= \mathbb{E} [\mathbb{E} (Y^1 - Y^0 | X) | Z = z] \\
&= \mathbb{E} \left[\frac{D(Y - m_1(X))}{p(X)} - \frac{(1-D)(Y - m_0(X))}{1-p(X)} + m_1(X) - m_0(X) \middle| Z = z \right] \\
&= \mathbb{E} \left[\frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)} \middle| Z = z \right] \\
&= \mathbb{E} [m_1(X) - m_0(X) | Z = z].
\end{aligned}$$

The exposition shows that IPW and outcome based estimands are embedded in the estimand based on $\psi(W, p, m_0, m_1)$. Finally, by noticing that

$$\theta = \mathbb{E} [\tau(Z)]$$

identification of ATE trivially follows from these considerations.

3.3 Main results

3.3.1 GATE estimation

Proposed estimator

The identification results from the preceding section suggest a two-step estimation strategy. The details of our proposed estimator are described in Procedure 1. In a first step a sample plug-in versions of $\psi(W, p, m_0, m_1)$ is obtained by estimating the nuisance parameters. In a second step the ψ -vector is projected on Z . Our goal

Procedure 1. GATE estimation

Introduce the subsample index $l = 1, \dots, L$ and denote the corresponding information set by $\mathcal{J}_l$ as well as its complement by $\mathcal{J}_l^C$.

1. Randomly split the sample in equally sized subsamples $1, \dots, L$.
2. **for** $l = 1$ **to** L **do**

Estimate the propensity score $p(X)$ and the outcome projections $m_0(X)$ and $m_1(X)$ in the sample with $\mathcal{J}_l^C$ using suitable machine learning methods.

Predict $\hat{p}(X)$, $\hat{m}_0(X)$ and $\hat{m}_1(X)$ in the sample with $\mathcal{J}_l$.

end

3. Denote $\hat{p} = \hat{p}_{l=1, \dots, L}$, $\hat{m}_0 = \hat{m}_{0, l=1, \dots, L}$ and $\hat{m}_1 = \hat{m}_{1, l=1, \dots, L}$. Construct the vector with elements $\hat{\psi} = \psi(W_i, \hat{p}, \hat{m}_0, \hat{m}_1)$ for $i = 1, \dots, N$. Estimate GATE as

$$\hat{\tau}(z) = \sum_{i=1}^N \frac{K\left(\frac{Z_i - z}{h}\right) \psi(W_i, \hat{p}, \hat{m}_0, \hat{m}_1)}{\sum_{i=1}^N K\left(\frac{Z_i - z}{h}\right)}$$

where $K = K(u)$ is a (product) kernel function that depends on a bandwidth h .

is to estimate both stages as flexible as possible and to avoid parametric assumptions. Further, our estimator can cope with settings where λ_X is very large which precludes classical nonparametric and parametric methods to estimate the first stage nuisance functions $p(X)$, $m_0(X)$ and $m_1(X)$. However, we can use a large class of supervised machine learning algorithms that have been shown to be very effective predictors for such types of tasks. Following the suggestions of for example Bickel and Ritov (1988), Robins, Li, Tchetgen, and van der Vaart (2008) and Chernozhukov et al. (2018a) we apply a cross-fitting algorithm for the nuisance parameter estimation step in order to guarantee that the resulting estimator of $\psi(W, p, m_0, m_1)$ consists of independent observations. The requirements for the second stage estimation step are more sophisticated as this estimator should allow for valid inference. To estimate GATE flexibly, we apply nonparametric local constant regression in the second step.

Asymptotic results

We now investigate the theoretical properties of our proposed estimation procedure. To ease the notational burden, we start with some definitions.

Definition 3.1 (Norms). Denote by $\|g(X)\|_p = (\mathbb{E}[|g(X)|^p])^{\frac{1}{p}}$ the L_p norm of the generic function $g(\cdot)$. Further, denote the supremum by $\sup_{x \in \mathcal{X}} |g(x)| = \|g(X)\|_\infty$ and the infimum by $\inf_{x \in \mathcal{X}} |g(x)|$.

Definition 3.2 (Rates). The nuisance parameter first stage estimates $\hat{p}$, $\hat{m}_0$ and $\hat{m}_1$ obtained by the sample splitting procedure described above obey the rates

$$\begin{aligned}\epsilon_{m_d} &= \|\hat{m}_d(X) - m_d(X)\|_2, \\ \epsilon_p &= \|\hat{p}(X) - p(X)\|_2, \\ \epsilon_{\max} &= \max\{\epsilon_{m_0}, \epsilon_{m_1}, \epsilon_p\}.\end{aligned}$$

Definition 3.3 (Scaling factor). For any function g define a scaling parameter δ_g that determines $g = O(N^{-\delta_g})$.

We then make the following standard assumptions on the kernel regression step (see for example Pagan and Ullah, 1999).

Assumption 3.2 (Kernel regression).

- (i) The support of Z , $\mathcal{Z}$ is a Cartesian product of compact intervals. $Z = z$ is an interior point in $\mathcal{Z}$.
- (ii) The density function is uniformly bounded away from zero such that $\inf_{z \in \mathcal{Z}} f(z) \geq C$ where $C > 0$ is a generic constant.
- (iii) The Kernel function $K(u)$ is r times continuously differentiable, symmetric and of order r in the sense $\int u^{r-1} K(u) du = 0$ and $\int u^r K(u) du = O(1)$ for $r \in \mathbb{N}$.
- (iv) $f(z)$ and $\mathbb{E}(\psi(W, p, m_0, m_1)|Z = z)$ are r times continuously differentiable. Additionally, $\mathbb{E} \left[|\psi(W, p, m_0, m_1) - \mathbb{E}(\psi(W, p, m_0, m_1)|Z = z)|^{2+C} \right] = O(1)$ for any $C > 0$.

(v) Further the Kernel function satisfies $\int K(u)du = 1$; $\int |K(u)|^{2+C}du = O(1)$ for any $C > 0$; $|u|K(u) \rightarrow 0$ as $|u| \rightarrow \infty$; $\|K(u)\|_\infty = O(1)$; and $\int K^2(u)du = O(1)$.

Assumption 3.2 comprises the standard nonparametric local constant regression assumptions allowing for multiple covariates and higher-order kernels. For illustrative purposes multivariate regression results are derived assuming the same bandwidth for every regressor. Further, we have to make boundedness assumptions on the second moment of the sample error of the outcome model and on the nuisance prediction errors.

Assumption 3.3 (Boundedness of conditional variances). *The conditional variances of the outcome are bounded such that for every $x \in \mathcal{X}$ and $d \in \{0, 1\}$ they obey*

$$\text{Var}(Y|D = d, X = x) = O(1).$$

Assumption 3.4 (Boundedness of propensity score). *The propensity score and its estimator obey*

$$0 < \inf_{x \in \mathcal{X}} |p(x)| < \sup_{x \in \mathcal{X}} |p(x)| < 1 \quad \text{and} \quad 0 < \inf_{x \in \mathcal{X}} |\hat{p}(x)| < \sup_{x \in \mathcal{X}} |\hat{p}(x)| < 1.$$

Additionally, the convergence rates of our first stage nuisance parameter prediction and the second stage nonparametric regression are assumed to be as follows:

Assumption 3.5 (Coupled convergence GATE). *The bandwidth h and the sample size N jointly converge such that*

$$(i) \quad h = o(1), N^{-1}h^{-\lambda_Z} = o(1) \text{ as } N \rightarrow \infty \text{ and}$$

$$(ii) \quad N^{\frac{1}{2}}h^{\frac{1}{2}\lambda_Z}h^r = o(1).$$

Further, N and h satisfy the joint convergence conditions with the nuisance parameter convergence rates

$$(iii) \quad h^{-\frac{1}{2}\lambda_Z}\epsilon_{\max} = o(1) \text{ and}$$

$$(iv) \quad N^{\frac{1}{2}}h^{-\frac{1}{2}\lambda_Z}\epsilon_{m_0}\epsilon_p + N^{\frac{1}{2}}h^{-\frac{1}{2}\lambda_Z}\epsilon_{m_1}\epsilon_p = o(1).$$

Assumption 3.5 comprises the coupled convergence rate assumptions that are at the centre of our theoretical results. Conditions (i) and (ii) quantify how the bandwidth has to converge to zero in relation to the sample size N and the number of regressors λ_Z . As usual, the bandwidth has to go to zero but slower than the sample size grows to infinity. Also the bandwidth has to be chosen such that the asymptotic bias term vanishes faster than the variance. This allows to apply the Central Limit Theorem and makes the estimator asymptotically unbiased. In particular, condition (ii) requires undersmoothing in the sense that the bandwidth has to be below the mean squared error (MSE) optimal rate.

Conditions (iii) and (iv) state that h has to be chosen such that first stage convergence rates vanish fast enough. In particular, by condition (iv) the joint convergence rates from propensity score and outcome projection estimation

have to vanish faster than $\sqrt{N}$ scaled with the kernel bandwidth. Since $h \rightarrow 0$, this is a more restrictive assumption compared to the rate conditions usually required for average effects estimation (see for example Chernozhukov et al., 2018a). In contrast to average effects, one is only interested in estimating the effect at a prespecified point $Z = z$. Thus, since sample observations enter the estimator in a weighted form, the prediction precision needed is for the lower effective sample size around $Z = z$. Hence, the first stage prediction rate guarantees need to adapt to this smaller sample conditions and therefore achieve a faster joint rate of convergence in terms of the sample size N . Condition (iii) additionally prevents the worst rate from becoming arbitrarily slow especially when λ_Z is larger than one. Still, our estimator has a ‘rate’ double robustness feature in the sense that joint rates can vanish relatively slowly but all single first stage rates are restricted from converging very slowly. L_2 convergence rates of many supervised machine learning methods satisfy these properties under sparsity conditions. For example Belloni and Chernozhukov (2013) show that the predictive error of the Lasso is of order $O\left(\sqrt{\frac{s \log \max(\lambda_X, N)}{N}}\right)$ where s the unknown number of true coefficients in the oracle model. Suppose that s and λ_X are equal in the outcome and the propensity score models then we require $\frac{s^2 \log^2 \max(\lambda_X, N)}{Nh^{\lambda_Z}} \rightarrow 0$. Hence, the dimension of the confounding variables λ_X can be much larger than the effective sample size Nh^{λ_Z} . Similar rates can be shown for L_2 boosting (Luo and Spindler, 2016) and nonlinear models like Random Forests (Wager and Walther, 2016) or forms of Deep Neural Nets (Farrell et al., 2019).⁶

A natural question is then if a bandwidth exists that satisfies the rate conditions in Assumption 3.5. Indeed, one can show (for more details see Appendix 3.B) that the theoretical range of possible bandwidth choices can be described by

$$\frac{1}{\lambda_Z + 2r} < \delta_h < \frac{2(\delta_{\epsilon_p} + \delta_{\epsilon_{m_d}}) - 1}{\lambda_Z}$$

and we achieve a condition for the order of the kernel

$$r > \lambda_Z \frac{1 - (\delta_{\epsilon_p} + \delta_{\epsilon_{m_d}})}{2(\delta_{\epsilon_p} + \delta_{\epsilon_{m_d}}) - 1}.$$

If we restrict ourselves on second order kernel functions, then for $\lambda_Z = 1$ we require $\delta_{\epsilon_p} + \delta_{\epsilon_{m_d}} = \frac{3}{5}$. Similarly, for $\lambda_Z = 2$ and $\lambda_Z = 3$, $\delta_{\epsilon_p} + \delta_{\epsilon_{m_d}} = \frac{2}{3}$ and $\delta_{\epsilon_p} + \delta_{\epsilon_{m_d}} = \frac{5}{7}$ are required respectively. Thus, for a growing dimension λ_Z the joint rate condition for the first stage nuisance parameters approaches the parametric rate.

The discussion indicates that the researcher can choose the bandwidth somehow below but not too much below the MSE optimal rate. One could therefore simply use a certain fraction (e.g. 0.9) of the cross-validation bandwidth choice. Hence, from a practical perspective our bandwidth choice problem is equivalent to nonparametric regression with undersmoothing.

Given these assumptions we can then derive the first main theoretical result.

Theorem 3.1. *Under Assumptions 3.1-3.5 our proposed estimation procedure for GATE obeys*

$$\sqrt{Nh^{\lambda_Z}} (\hat{\tau} - \tau) = \frac{1}{\sqrt{Nh^{\lambda_Z}}} \frac{1}{f(z)} \sum_{i=1}^N K\left(\frac{Z_i - z}{h}\right) (\psi(W_i, p, m_0, m_1) - \tau) + o_p(1)$$

⁶For the concrete dependence of sparsity conditions on the parameters of the predictors see the references mentioned.

and

$$\sqrt{Nh^{\lambda_Z}} (\hat{\tau} - \tau) \rightarrow_d N(0, \sigma_{GATE}^2)$$

$$\text{with } \sigma_{GATE}^2 = \frac{\int K(u)^2 du \times \mathbb{E}[(\psi(W_i, p, m_0, m_1) - \tau)^2 | Z=z]}{f(z)}.$$

Theorem 3.1 shows that under the assumptions discussed above the speed of convergence is determined only by the nonparametric regression step. In particular, it does not depend on the first stage estimation steps. An equivalent result can also be achieved when using IPW with nonparametric first stages (see Abrevaya et al., 2015). However, this requires an additional bandwidth choice for the first stage propensity score regression and is limited to the case when also λ_X is very small. The dimension of X can be increased under functional form assumptions for the first stage. However, as shown by the authors the price to pay is an increase in the asymptotic variance. This is not the case for the estimator proposed in this study. Also Theorem 3.1 is valid under generally weaker conditions compared to the results in Lee et al. (2016) for parametric first stages. Heuristically, if the first stage estimators converge at $\sqrt{N}$ then our conditions on the bandwidth are satisfied and our asymptotic results continue to apply.⁷ To this extent, our results comprise the result of Lee et al. (2016) as a special case.

3.3.2 Joint estimation of GATE and ATE

Proposed estimator

Given the considerations so far, it appears ‘natural’ to estimate ATE in three steps as an average of GATEs in the sample. The details of the proposed estimator are described in Procedure 2. As suggested in Chernozhukov et al. (2018a) one could also directly estimate ATE as the average of the vector $\psi(W, p, m_0, m_1)$ using the first stage nuisance parameter predictions. However, as a by-product of GATE estimation, using an additional kernel smoothing step may lead to an ATE estimator with better finite sample properties. In particular, the propensity score weights do not enter the last step of our estimator directly and our hope is that small misspecification errors of propensity scores close to zero or one are therefore smoothed out. The sensitivity of estimators incorporating inverse propensity score weights directly is the subject of many Monte Carlo experiments (e.g., Huber, Lechner, and Wunsch, 2013 and Frölich, 2004b). We notice that similar reasoning is also behind three-step propensity score estimators that apply nonparametric regression on an estimated propensity score often used in practice.

Procedure 2. ATE estimation

1. Follow steps 1-3 of Procedure 1.
2. Predict GATE at every observation in the sample as $\hat{\tau}(z_i)$.
3. Estimate ATE as

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^N \hat{\tau}(z_i).$$

⁷In contrast to Lee et al. (2016), we use local constant instead of local linear regression and introduce cross-fitting for nuisance parameter estimation. This should, however, not be relevant for the intuitive argument made.

Asymptotic results

To obtain our theoretical results, we have to modify Assumption 3.5 slightly.

Assumption 3.5' (Coupled convergence ATE). *The bandwidth h and the sample size N jointly converge such that*

- (i) $h = o(1)$, $\frac{\ln(N)}{Nh^{\lambda_Z}} = o(1)$ as $N \rightarrow \infty$,
- (ii) $Nh^{2r} = o(1)$ and $Nh^{2\lambda_Z} \rightarrow \infty$.

Further, N and h satisfy the joint convergence conditions with the nuisance parameter convergence rates

- (iii) $h^{-\lambda_Z} \varepsilon_{\max} = o(1)$ and
- (iv) $N^{\frac{1}{2}} h^{-\lambda_Z} \varepsilon_{m_0} \varepsilon_p + N^{\frac{1}{2}} h^{-\lambda_Z} \varepsilon_{m_1} \varepsilon_p = o(1)$.

We notice that averaging over the estimated projection $\mathbb{E}[\psi(W, p, m_0, m_1)|Z]$ is a full mean problem in the sense of Newey (1994a). Conditions (i) and (ii) guarantee the applicability of Newey's (1994a) framework.

Since we want ATE to converge with a rate of $\sqrt{N}$, the requirements on the first stage convergence rates in condition (iii) and (iv) are more restrictive than those in the respective condition of Assumption 3.5. The range of theoretically feasible bandwidth choices reduces to

$$\frac{1}{2r} < \delta_h < \frac{(\delta_{\varepsilon_p} + \delta_{\varepsilon_{m_d}}) - \frac{1}{2}}{\lambda_Z}.$$

In general this result indicates that one relies on a higher-order kernel function whenever $\lambda_Z > 1$ when GATE and ATE are estimated jointly. Under the stronger Assumption 3.5' we can then derive the following efficiency result.

Theorem 3.2. *Under Assumptions 3.1-3.4 and 3.5' and some further regularity conditions stated in Newey (1994a, Theorem 4.2) our proposed estimation procedure for ATE has the influence function*

$$\frac{D(Y - m_1(X))}{p(X)} - \frac{(1 - D)(Y - m_0(X))}{1 - p(X)} + m_1(X) - m_0(X) - \theta$$

and therefore obeys

$$\sqrt{N}(\hat{\theta} - \theta) \rightarrow_d N(0, \sigma_{ATE}^2)$$

where σ_{ATE}^2 is the semiparametric efficiency bound of Hahn (1998).

Conceptually, Theorem 3.2 underpins our intuition from semiparametric theory outlined in Section 3.2.2. The result shows that indeed every estimator that involves a nonparametric projection of the AIPW modified outcome on any low-dimensional subset of X is consistent, asymptotically normal and achieves the semiparametric efficiency bound. This asymptotic result has also been shown for other estimators already discussed in Section 3.1. In contrast to Hirano, Imbens, and Ridder's (2003) estimator, Hahn's (1998) estimator and matching on the propensity score (Hahn and Ridder, 2013), we do not rely on nonparametrically estimated first stages. Due to the fact that $\lambda_X \gg N$,

these estimators are of no practical use in our setting. Further, unlike AIPW with machine learning nuisance parameter estimation (Chernozhukov et al., 2018a), our estimator involves an additional step. Therefore the inverse propensity score does not directly enter our estimator but is smoothed through the additional nonparametric step. Asymptotically, this does not make any difference as the result in Theorem 3.2 shows. However, in finite samples this could be a major advantage over the usual AIPW estimator.

3.4 Implementation choices

The theoretical arguments in the previous sections are formulated very generally and in principal allow for a large variety of implementations of the local constant nonparametric regression step as well as the prediction of the nuisance parameters. We follow the common practice and use a second order Gaussian kernel. It is widely acknowledged that for nonparametric regression the bandwidth choice is most crucial. Several plug-in and data-driven methods exist (for a discussion see Li and Racine, 2007). We generally rely on leave-one-out cross-validation (LOOCV) to determine our bandwidth. A formal treatment of the properties of LOOCV for bandwidth selection in nonparametric regression models is given in Hall, Racine, and Li (2004) and Hall, Li, and Racine (2007). The authors show that under certain regularity conditions the LOOCV bandwidth converges at order $\frac{1}{\lambda_Z + 2r}$. To achieve undersmoothing, we multiply the bandwidth obtained by LOOCV with 0.9. Due to computational constraints, we do only cross-validate the nonparametric regression step and not the whole estimation procedure. This might be critical as the theoretical arguments made in Hall et al. (2004) and Hall et al. (2007) rely on the fact that the left-out observation is independent of the rest of the sample. The assumptions trivially holds for i.i.d. data, and it would also hold if we would observe $\psi(W, p, m_0, m_1)$. Since we rely on first stage estimators of the propensity score and the outcome projections, though, the left-out observation is constructed using the same first stage estimators that also appear in the rest of the sample. We conjecture, however, that $\psi(W, \hat{p}, \hat{m}_0, \hat{m}_1)$ is ‘close’ enough to $\psi(W, p, m_0, m_1)$ such that this is only a theoretical concern. A formal assessment of this question is beyond the scope of this study but we provide comprehensive Monte Carlo evidence showing a strong robustness of the LOOCV bandwidth selection in the next section.

Another implementation choice concerns the estimation of the propensity score and the outcome projections. We use an ensemble learner that weights Lasso and Random Forest predictions by using out-of-sample MSE optimal weights. The covariates enter our model very flexibly. For the Lasso we allow for polynomials up to order four and all two way interactions. The Random Forest has the particular advantage of being an ensemble of trees itself and is therefore very flexible by construction. While the Lasso allows for some form of extrapolation, the Random Forest implicitly weights observation around a point in a data-driven way to make predictions. We therefore conjecture that the Lasso performs well in areas that are particularly smooth whereas the Random Forest has an advantage in areas with strong non-linearities like jumps. Combining two very different predictors in one ensemble in a data-driven way, has the advantage of giving more weight to the estimator that is particularly suitable for the underlying DGP. Further, machine learning estimators typically rely on a variety of tuning parameters. The ensemble gives more weight to the single predictor that performs best and therefore should be more robust against the choice of tuning parameters than a single implementation of a predictor.

As suggested by Chernozhukov et al. (2018a), all estimators that involve the ensemble learner are implemented using 4-fold cross-fitting. Whether or not the presumably better predictive performance of the ensemble learner translates into a better performance of the GATE and ATE estimator compared to standard maximum likelihood

based methods is a finite sample concern that is investigated in the next section.

3.5 Empirical Monte Carlo experiments

In order to assess the finite sample performance of the estimators proposed in Section 3.3 and to assess some of our implementation choices discussed in the previous Section, we run several Empirical Monte Carlo (EMC) experiments (e.g. Huber et al., 2013; Knaus et al., 2020b). In contrast to standard simulation methods, the EMC approach intends to reduce arbitrary choices on the DGP. Rather, DGPs in the EMC approach are constructed by incorporating as many components as possible from real data. Arguably, this results in a more realistic assessment of the performance of the estimators in empirical studies. The EMC experiments in this study are all built on the Swiss administrative social security data previously used by Behncke, Frölich, and Lechner (2010a,b), Huber, Lechner, and Mellace (2017), Knaus, Lechner, and Strittmatter (2020a), Lechner (2018) and Knaus (2020) for the evaluation of the effects of participating in a job search programme (D) on months in employment after the programme start (Y). We assume that the researcher is interested in the variation of the effect with respect to previous earnings (Z). We only focus on observations with previous earnings larger than zero. Further, we focus on two different sets of covariates. The first set was already used in Huber et al. (2017) (HLM). The second set of covariates additionally includes some dummies on the previous sector the unemployed was employed (for more details see the papers referenced). The details of the EMC experiments are outlined in Procedure 3. In a first step the ‘true’ propensity score is estimated. We use both specifications – standard logistic regression and the ensemble learner outlined in Section 3.4 – in order not to give an advantage regarding a specific class of nuisance parameter estimators. Dropping the treated observations leaves us with a remaining sample with 85,789 observations from which we randomly draw the Monte Carlo samples. We choose the true GATE in our EMC experiment as

Procedure 3. Empirical Monte Carlo experiment

1. Take the full sample and estimate the propensity score using the Logit specification as in HLM and using an ensemble learner that comprises a Lasso and a Random Forest estimator on the extended dataset. Define the average of both estimators as $p^{sample}(x)$.
2. Remove the N^1 treated observations. Observe that for the remaining subsample $Y^0 = Y$. Specify the true GATE $\tau(z)$ and generate $Y^1 = Y^0 + \tau(z) + (1 - u) + v$ where $u \sim \text{Poisson}(1)$, $v = \mathbb{1}(v^* > \bar{v}) * (-\bar{v}) + \mathbb{1}(v^* \leq \bar{v}) * \bar{v}$ with $v^* \sim \text{Uniform}(0, 1)$ and $\bar{v} = \tau(z) + u - \text{floor}(\tau(z) + u)$.
3. Set aside a random validation sample.
4. Obtain $p^{mc}(x)$ as a modified version of $p^{sample}(x)$ that controls the desired ratio of treated and untreated observations.
5. Draw a random subsample of size N^{mc} .
 - Generate $D \sim \text{Bernoulli}(p^{mc}(x))$.
 - Generate $Y = DY^1 + (1 - D)Y^0$.
 - Estimate τ and θ with all estimators of interest.
 - Predict $\hat{\tau}(z)$ in the validation sample.
6. Repeat the previous steps many times.
7. Calculate performance measures for τ on every observation of the validation subsample and average them. Calculate performance measures for θ .

$\tau(z) = \alpha \times \tilde{\tau}(z)$ where $\tilde{\tau}(z)$ is a standardised version of $0.4 + 0.25 * \sin(8 * z - 5) + 0.4 * \exp(-16 * (4 * z - 2.5)^2)$. This functional form was initially proposed by Frölich (2004a) and is also used by Busso, DiNardo, and McCrary (2014) to model conditional expectations in a related theoretical setting. The parameter α determines the extent to which variation in the GATE explains Y^1 . Using the transformations in the error term v , we generate the potential outcome Y^1 such that it obeys the discrete nature of the underlying outcome. For the GATE, performance measures are calculated on the 1,000 points of the randomly drawn validation sample and then averaged.

We compare our estimators described in Procedures 1 and 2 with several other suggestions in the literature. The nuisance parameters are either estimated using parametric maximum likelihood methods (Logit for the propensity score and OLS for the outcome models) or the ensemble learner. Estimators that rely on local constant nonparametric regression are implemented using a second order Gaussian kernel function. For GATE estimation this leaves us with estimators that combine parametric nuisance parameter estimation for IPW or AIPW scores and nonparametric regression (for details see Abrevaya et al., 2015 and Lee et al., 2016). Alternatively, the nuisance parameters in the AIPW score can be estimated using the ensemble learner. In this class of estimators we compare the estimator described in Procedure 1 with the linear predictor approach covered by the theoretical framework in Semenova and Chernozhukov (2020). For the ATE we compare standard IPW and AIPW estimation with parametric first stages to the Double Machine Learning estimator (e.g. Chernozhukov et al., 2018a) and our suggestion described in Procedure 2.

We consider four different DGPs that vary the share of the treated (50 % and 10 %) and the number of observations (1,000 and 4,000). We choose $\alpha = 7$ such that variance of the GATE amounts to approximately 31 % of the variance of Y^0 . Results for some further DGPs with $\alpha = 2$ are relegated to Appendix 3.C.

3.5.1 Results for GATE and ATE

Table 3.1 shows the results for the GATE. For the benchmark design with 50 % treatment share all estimators are nicely behaved. Estimators based on nonparametric regression generally outperform those based on the linear predictor. Additionally, estimators that use the ensemble for nuisance parameter estimation unambiguously perform better than their counterparts that use parametric models. For the more complicated designs with the 10 % treatment share extreme inverse propensity score weights become more likely. The parametric models for this design result in estimators with exploding MSEs such that they mostly are not reported. Estimators that use the ensemble learner for nuisance parameter prediction also show a decreased performance but remain stable. Interestingly, the estimators based on the linear predictor are relatively close to their performance in the benchmark design for larger samples while our estimator converges much slower. This hints at the data-hungriness of the nonparametric regression compared to linear models. Still, the estimator in Procedure 1 outperforms all other alternatives.

The results for the ATE are summarised in Table 3.2. Again estimators that use the ensemble learner for nuisance parameter prediction unambiguously outperform standard IPW or AIPW estimation. For the benchmark DGP with 50 % treatment share the Double Machine Learning estimator dominates, even though the smoothed estimator suggested in Procedure 2 is pretty close especially for the larger sample. For the more complicated design with 10 % treatment share our estimator outperforms Double Machine Learning. Again for larger sample sizes the differences become negligible. This is in line with the theoretical results of Theorem 3.2 discussed in Section 3.3.2. While being asymptotically equivalent, potentially more extreme inverse propensity score weights (that are more likely to arise in the 10 % treatment share design) give a finite sample advantage of nonparametric smoothing relative to

Table 3.1: Performance of different GATE estimators

	$\overline{\text{MSE}}$	$ \overline{\text{Bias}} $	$\overline{\text{SD}}$	$\overline{\text{MSE}}$	$ \overline{\text{Bias}} $	$\overline{\text{SD}}$
50 % treatment share						
	$N = 1,000$			$N = 4,000$		
IPW maxlik lc	16.54	1.06	3.32	4.67	0.53	1.74
AIPW maxlik lp	44.17	5.68	1.20	42.94	5.69	0.55
AIPW maxlik lp squared	26.43	3.76	1.37	24.65	3.69	0.65
AIPW maxlik lc	7.26	0.96	2.20	2.07	0.44	1.20
AIPW ens lp	43.93	5.70	1.10	42.93	5.70	0.53
AIPW ens lp squared	26.00	3.64	1.27	24.59	3.68	0.62
AIPW ens lc	5.94	0.55	2.13	1.87	0.34	1.19
10 % treatment share						
	$N = 1,000$			$N = 4,000$		
IPW maxlik lc	-	-	-	17.12	1.19	3.31
AIPW maxlik lp	-	-	-	-	-	-
AIPW maxlik lp squared	-	-	-	-	-	-
AIPW maxlik lc	-	-	-	-	-	-
AIPW ens lp	61.94	5.71	3.94	43.60	5.71	0.93
AIPW ens lp squared	61.78	3.69	4.86	25.54	3.74	1.11
AIPW ens lc	20.91	1.18	4.08	4.93	0.63	1.93

Notes: Results were obtained as described in Procedure 3 for 1,000 Monte Carlo replications for $\alpha = 7$. The three performance measures are averages over the 1,000 out-of-sample evaluation points. *maxlik* and *ens* indicate that a maximum likelihood based procedure or an ensemble learner based on the Random Forest and the Lasso were used to estimate the nuisance parameters. *lc* and *lp* indicate that the conditional expectation was estimated using local constant nonparametric regression and the linear predictor respectively. *squared* indicates that a second order polynomial was included in the linear model.

relying on inverse propensity score weights directly.

3.5.2 LOOCV bandwidth selection

We assess the sensitivity with regard to the bandwidth by running Procedures 1 and 2 for 20 different multiples ranging from 0.1 to 2 of the LOOCV bandwidth choice. Figure 3.1 depicts the MSEs for the different bandwidth choices in the EMC design with 50 % treatment share for sample sizes 1,000 and 4,000. The results indicate that the estimators are indeed MSE optimal for bandwidth choices close to the LOOCV implied bandwidth. Interestingly, this seems to hold for both the GATE and the ATE. Additionally, the range of bandwidth choices for which the loss in MSE compared to the minimum is relatively small seems to be large for factors smaller than one. Similar results are obtained for the other EMC designs. They are relegated to Appendix 3.C.

3.6 Illustrative example

We investigate the applicability of our methods using a dataset originally used by Almond, Chay, and Lee (2005) to study the effect of cigarette smoking on birthweight. It was made available by Cattaneo (2010) and can be

Table 3.2: Performance of different ATE estimators

	MSE	Bias	SD	MSE	Bias	SD
50 % treatment share						
	<i>N</i> = 1,000			<i>N</i> = 4,000		
IPW maxlik	0.96	0.01	0.98	0.17	0.00	0.41
AIPW maxlik	0.79	0.05	0.89	0.17	0.01	0.41
AIPW ens	0.67	0.02	0.82	0.15	0.05	0.39
AIPW ens smoothed	0.71	0.16	0.83	0.16	0.04	0.39
10 % treatment share						
	<i>N</i> = 1,000			<i>N</i> = 4,000		
IPW maxlik	-	-	-	1.30	0.38	1.08
AIPW maxlik	-	-	-	-	-	-
AIPW ens	6.48	0.38	2.52	0.52	0.16	0.70
AIPW ens smoothed	4.51	0.01	2.13	0.50	0.00	0.71

Notes: Results were obtained as described in Procedure 3 for 1,000 Monte Carlo replications for $\alpha = 7$. *maxlik* and *ens* indicate that a maximum likelihood based procedure or an ensemble learner based on the Random Forest and the Lasso were used to estimate the nuisance parameters.

retrieved from the Stata website.⁸ The dataset contains the outcome variable birthweight in grams (Y), whether the mother smoked during pregnancy ($D = 1$) and several covariates on the mother's health and socio-economic background (X). A detailed description of all covariates in the dataset can be found in Appendix 3.D. Applied studies with different estimation approaches unambiguously find negative average effects (see Almond et al., 2005; Abrevaya, 2006; da Veiga and Wilder, 2008; Walker, Tekin, and Wallace, 2009). Conditional average treatment effects were investigated by Abrevaya et al. (2015) and Lee et al. (2016) who find that mother's age is associated with increasingly negative effects of smoking. We replicate their results and compare their estimators with ours. Clearly, this type of analysis is limited in its scope since the true DGP remains unknown. However, the dataset has the particular advantage that some strong hypothesis about the estimation results are plausible. (i) The effect of smoking on birthweight should be either negative or zero. (ii) The effect should be increasingly negative with mother's age.

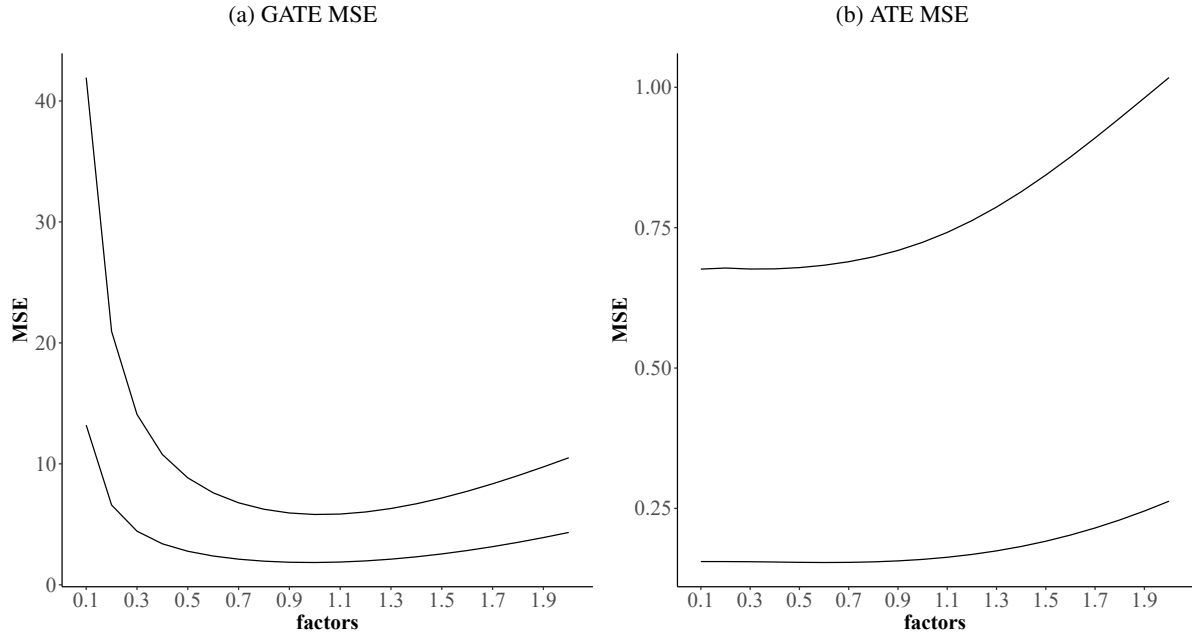
As a second example we consider how the effect changes with the number of prenatal care visits. On the one hand a very low number of care visits could indicate the mother's insufficient access to medical infrastructure and therefore could be associated with particularly negative effects. On the other hand a very high number of care visits could indicate a poor health situation. Hence, it is a priori unclear how the treatment effect and health care visits are related.

3.6.1 Empirical results

Figure 3.2 depicts the main results of our empirical analysis. We estimate the GATEs as described in Procedure 1 using the ensemble learner. X in our specification is an extended variable set ('alldata') and is exactly documented in Appendix 3.D. For example in contrast to Lee et al. (2016) we also include the available characteristics for the father of the child, since they could be a good predictor for the smoking behaviour of the mother. The results are

⁸The original dataset can be retrieved from <http://www.stata-press.com/data/r13/cattaneo2.dta>.

Figure 3.1: GATE and ATE MSEs for different factors of the LOOCV bandwidth choice



Notes: Results were obtained as described in Procedure 3 for the design with $\alpha = 7$ and 50 % treatment share. The upper line is for 1,000 and the lower line for 4,000 observations. Procedures 1 and 2 were implemented with a second-order Gaussian kernel function and the ensemble learner. The factors on the horizontal axis denote the different values with which the LOOCV bandwidth choice was multiplied in order to obtain the final bandwidth.

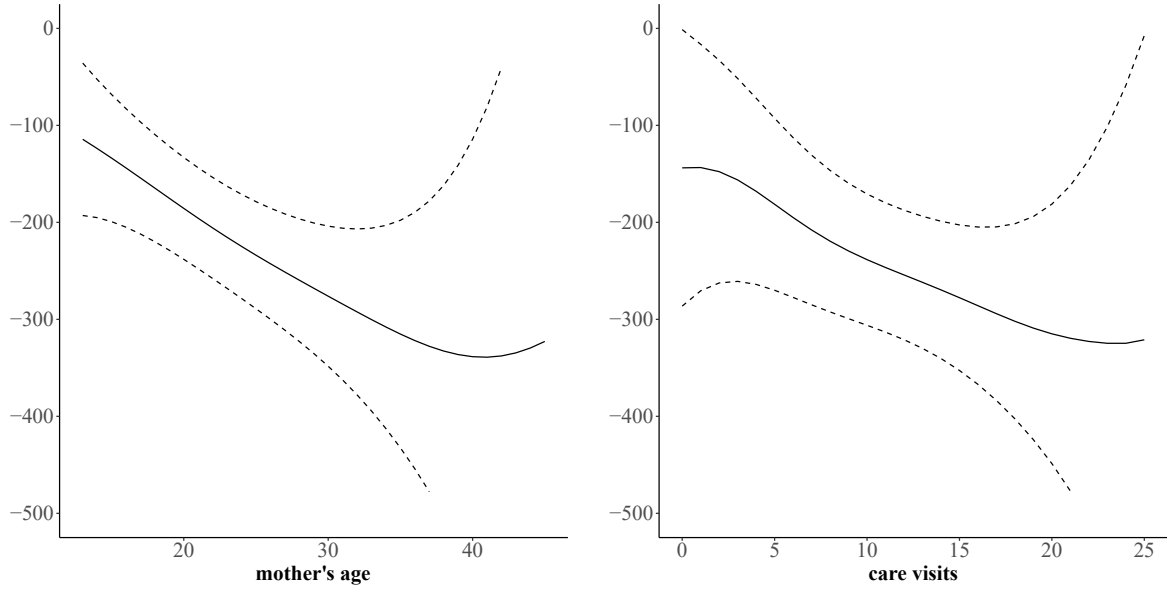
generally in line with the hypothesis made. In particular, the effect of smoking is unambiguously negative over the whole support of mother's age and prenatal care visits. As expected, the effect increases with age. Interestingly, a higher number of prenatal care visits seems to be associated with higher negative effects. We again assess the robustness of the results towards the $0.9 \times \text{LOOCV}$ bandwidth choice in Figure 3.3. A major change in the shape of the function only appears for massive oversmoothing. An equivalent analysis for prenatal care visits yielding the same conclusion is relegated to Appendix 3.D. Finally, the first two columns in Table 3.3 show the results for ATE estimation as described in Procedure 2. In line with the previous literature mentioned above, the average effect of smoking is estimated to be negative. Crucially, the estimated effect turns out to be very robust regarding the choice of the smoothing variable.

Table 3.3: Averaged ATE estimators

AIPW ens smooth (age)	AIPW ens smooth (care visits)	AIPW ens	AIPW maxlik	IPW maxlik
-239.051	-241.628	-238.271	-242.990	-295.388
(29.356)	(29.356)	(29.356)	(25.885)	(46.521)

Notes: Results for ATE estimation using Inverse Probability Weighting (IPW) and Augmented IPW (AIPW). For the ensemble learner nuisance parameters were estimated using an ensemble learner comprising Lasso and Random Forest. For Lasso the penalty term was chosen such that the 10-fold cross-validation criterion was minimised. The ensemble weights were chosen by minimising out-of-sample MSE. For the parametric specifications nuisance parameters were estimated using Logit for the propensity score and OLS for the outcome projections. Asymptotic standard errors are in parenthesis.

Figure 3.2: AIPW GATE estimator with ensemble first stages



Notes: Results were obtained as described in Procedure 1 with a second-order Gaussian kernel function and a $0.9 \times \text{LOOCV}$ bandwidth choice. Nuisance parameters were estimated using an ensemble learner comprising Lasso and Random Forest. For Lasso the penalty term was chosen such that the 10-fold cross-validation criterion was minimised. The ensemble weights were chosen by minimising out-of-sample MSE. Asymptotic confidence bands are at the 95% level.

3.6.2 Comparison with other estimators

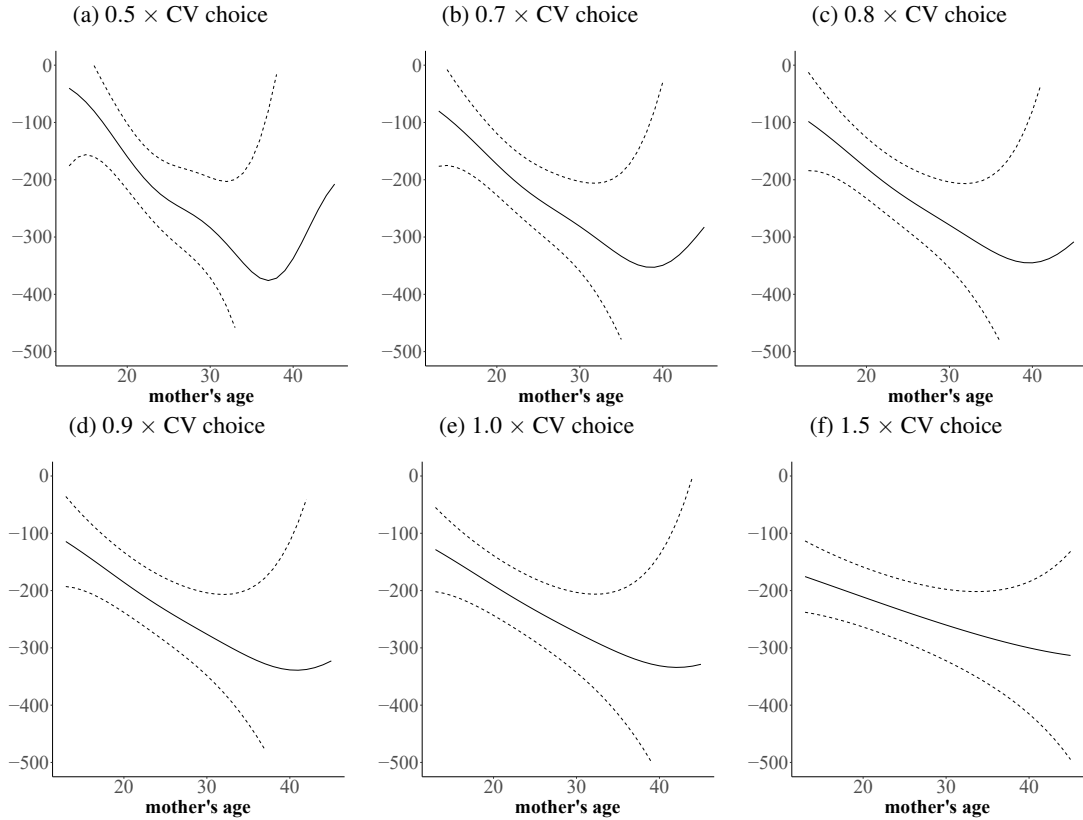
A ‘fair’ comparison with other estimators is hardly feasible because our approach does not require specific functional form assumptions.⁹ In other words, the related estimators of Abrevaya et al. (2015) and Lee et al. (2016) require knowledge on the true propensity score or outcome projection specifications. Since we cannot compare our estimator against every possible parametric specification, we use the specification selected by Lee et al. (2016) as a benchmark. Figure 3.4 depicts GATE estimation results using the benchmark models. Strikingly, the IPW based estimator gives implausible results. For mother’s age positive effects of smoking can almost nowhere be excluded. For care visits we do not obtain a GATE estimation result for our bandwidth choice. In fact, ATE is estimated indicating that the bandwidth is too large in order to obtain GATE estimates. In line with the theoretical result, standard errors are inflated compared to our estimation procedure. The AIPW based estimator with parametric models for the propensity score and the outcome projections gives plausible results for GATE with respect to mother’s age. Slight differences arise when comparing the GATE curves regarding the number of prenatal care visits.

Table 3.3 shows the results for ATE estimation. As expected, the results of Procedure 2 are roughly in line with the standard AIPW based ATE estimator with ensemble first stages.¹⁰ The relative bad performance of IPW based estimation for the GATEs is also reflected in the estimation of ATE. In particular, the standard error nearly doubles compared to AIPW based estimators and point estimates are reduced.

⁹We do not consider nonparametric propensity score estimation as suggested in Abrevaya et al. (2015) because it most likely does not allow to include all potential confounders in order to make Assumption 3.1 credible.

¹⁰This might also be seen as a implicit test for the credibility of the stronger conditions required for the smoothed estimator compared to the averaged efficient score as in Chernozhukov et al. (2018a).

Figure 3.3: Sensitivity to bandwidth choice (age)



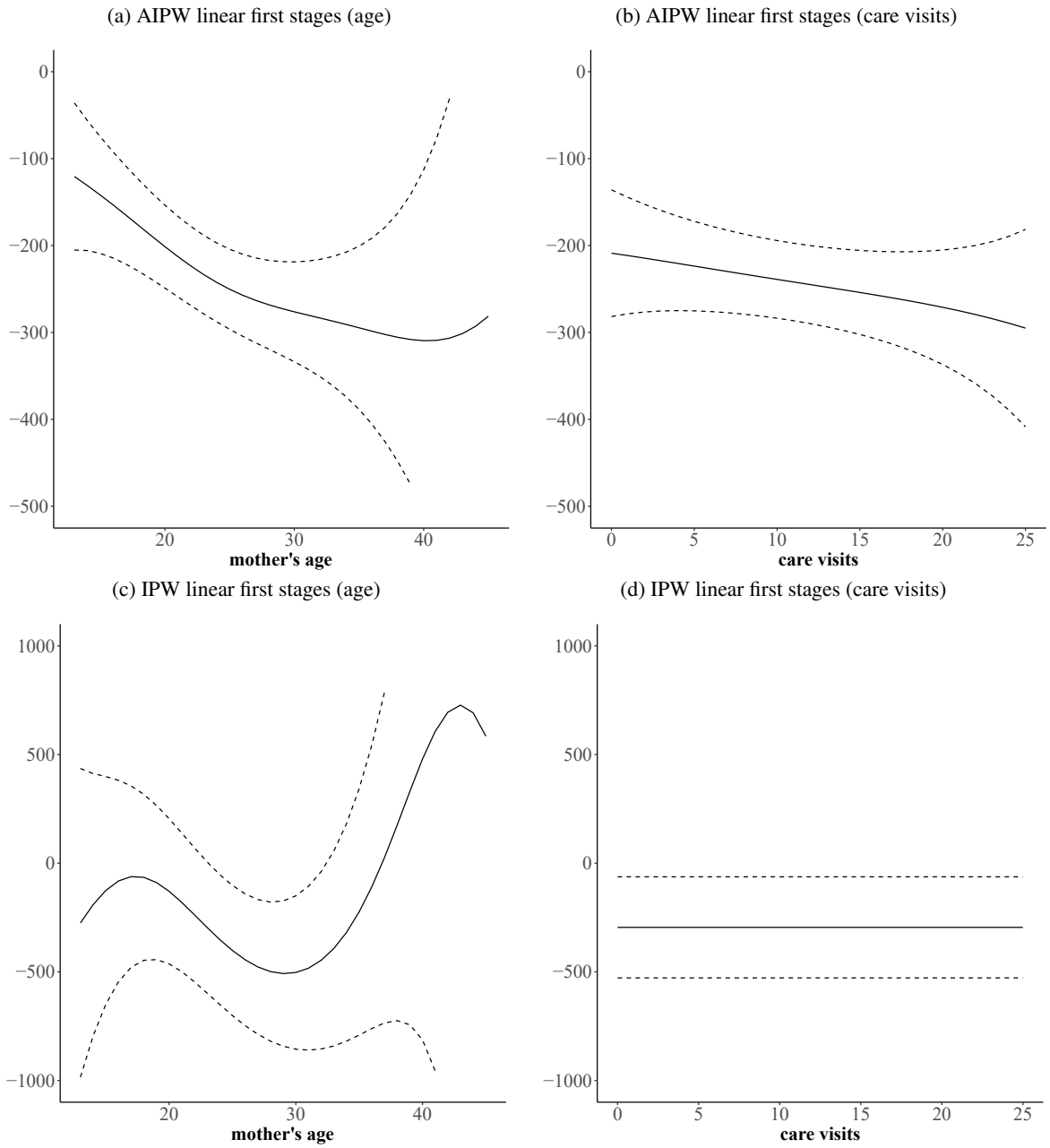
Notes: Results were obtained as described in Procedure 1 with a second-order Gaussian kernel function and different multiples of the LOOCV bandwidth choice. Nuisance parameters were estimated using an ensemble learner comprising Lasso and Random Forest. For Lasso the penalty term was chosen such that the 10-fold cross-validation criterion was minimised. The ensemble weights were chosen by minimising out-of-sample MSE. Asymptotic confidence bands are at the 95% level.

3.7 Conclusion

In this study we propose new estimators for specific conditional and average causal effects when the dimension of the covariate space is high. In particular, by discriminating the different roles of covariates (adjusting for confounding vs. measuring causal heterogeneity of interest) in our approach, they can be included very flexibly – not relying on functional form assumptions. Rather, we show coupled convergence conditions for the different steps involved. The procedures suggested are based on semiparametric efficiency theory. In this sense, our proposed three-step estimator for ATE estimation is shown to reach the semiparametric efficiency bound. EMC experiments show the relatively good performance of our proposed estimators. Relative to estimators that rely on specific functional form assumptions, combining nonparametric regression with flexible machine learning first stages can lead to significant performance gains for GATE estimation. Additionally, while the smoothed three-step estimator for the ATE is first order asymptotically equivalent to directly averaging the efficient score vector, it possesses better finite sample properties in settings that are prone to extreme propensity score weights. An empirical example shows that our estimators are useful in practice. Compared to other estimators, their desirable theoretical properties and increased flexibility could lead to divergent empirical results.

Some interesting problems and refinements are beyond the scope of this study and have to be left for further research. As already discussed, the formal properties of LOOCV for bandwidth selection with generated outcomes in nonparametric regression is an interesting problem. Further, as an alternative to classical nonparametric regres-

Figure 3.4: Other GATE estimators



Notes: Results were obtained following the procedures in Abrevaya et al. (2015) and Lee et al. (2016) with a second-order Gaussian kernel function and a $0.9 \times \text{LOOCV}$ bandwidth choice. Nuisance parameters were estimated using Logit for the propensity score and OLS for the outcome projections. Asymptotic confidence bands are at the 95% level.

sion, one could also investigate using methods from the toolbox of supervised machine learning. This might help to get reliable estimators even for cases when the dimension of Z is moderately higher than considered in this study, while sacrificing only little flexibility.

3.A Proof of Theorems

Before presenting the proofs of Theorem 3.1 and 3.2, Lemma 3.1 comprises some useful preliminary results.

Lemma 3.1. Write $\bar{\psi}_i$ as a shorthand for $\psi(W_i, \hat{p}, \hat{m}_0, \hat{m}_1) - \psi(W_i, p, m_0, m_1)$. Then under Assumptions 3.2, 3.3 and 3.4 and using the convergence rate definitions in Section 3.3 $K\left(\frac{Z-z}{h}\right) \bar{\psi}$ obeys (i) $|\mathbb{E}[K\left(\frac{Z-z}{h}\right) \bar{\psi}]| = O(\varepsilon_p \times (\varepsilon_{m_0} + \varepsilon_{m_1}))$; (ii) $\|K\left(\frac{Z-z}{h}\right) \bar{\psi}\|_2 = O(\varepsilon_{\max})$.

Proof:

By simple algebra we first of all notice that

$$\begin{aligned} \bar{\psi} = & \underbrace{\left(\frac{1}{\hat{p}(X)} - \frac{1}{p(X)}\right) D(Y - m_1(X))}_{ia} + \underbrace{\left(\frac{1}{1-p(X)} - \frac{1}{1-\hat{p}(X)}\right) (1-D)(Y - m_0(X))}_{ib} \\ & + \underbrace{(\hat{m}_1(X) - m(X)) \left(1 - \frac{D}{p(X)}\right)}_{iia} + \underbrace{(\hat{m}_0(X) - m_0(X)) \left(\frac{1-D}{1-p(X)} - 1\right)}_{iib} \\ & + \underbrace{\frac{D}{p(X)\hat{p}(X)} (\hat{p}(X) - p(X)) (\hat{m}_1(X) - m_1(X))}_{iiia} + \underbrace{\frac{1-D}{(1-p(X))(1-\hat{p}(X))} (\hat{p}(X) - p(X)) (\hat{m}_0(X) - m_0(X))}_{iiib}. \end{aligned}$$

The first claim follows because $\mathbb{E}[ia + ib + iia + iib|X] = 0$ and therefore

$$\mathbb{E}[\bar{\psi}|X] = \frac{1}{\hat{p}(X)} (\hat{p}(X) - p(X)) (\hat{m}_1(X) - m_1(X)) + \frac{1}{1-\hat{p}(X)} (\hat{p}(X) - p(X)) (\hat{m}_0(X) - m_0(X)).$$

Then by the Law of Iterated Expectations, using that $Z \subseteq X$ and Hölder's inequality

$$\begin{aligned} & \left| \mathbb{E} \left[K \left(\frac{Z-z}{h} \right) \mathbb{E}[(\psi(W, \hat{p}, \hat{m}_0, \hat{m}_1) - \psi(W, p, m_0, m_1)) | X] \right] \right| \\ & \leq \|K(u)\|_\infty \times \left\| \frac{1}{\hat{p}(X)} (\hat{p}(X) - p(X)) (\hat{m}_1(X) - m_1(X)) + \frac{1}{(1-\hat{p}(X))} (\hat{p}(X) - p(X)) (\hat{m}_0(X) - m_0(X)) \right\|_1 \\ & \leq \|K(u)\|_\infty \times \left(\left\| \frac{1}{\hat{p}(X)} \right\|_\infty \times \varepsilon_p \times \varepsilon_{m_0} + \left\| \frac{1}{1-\hat{p}(X)} \right\|_\infty \times \varepsilon_p \times \varepsilon_{m_1} \right) \\ & = O(\varepsilon_p \times \varepsilon_{m_0} + \varepsilon_p \times \varepsilon_{m_1}) \end{aligned}$$

where the last bound follows by $\|K(u)\|_\infty = O(1)$ in Assumption 3.2 and the boundedness of the propensity score estimator in Assumption 3.4.

For the second claim notice that

$$\|\bar{\psi}\|_2 \leq \|ia\|_2 + \|ib\|_2 + \|iia\|_2 + \|iib\|_2 + \|iiia\|_2 + \|iiib\|_2.$$

For the single terms we obtain

$$\begin{aligned} \|ia\|_2 &\leq \left\| \frac{1}{p(X)} \right\|_\infty \times \left\| \frac{1}{\hat{p}(X)} \right\|_\infty \times \|(\hat{p}(X) - p(X))D(Y - m_1(X))\|_2 \\ &= O(\varepsilon_p) \end{aligned}$$

by Assumptions 3.3 and 3.4,

$$\begin{aligned} \|iia\|_2 &\leq \left\| \frac{1}{p(X)} \right\|_\infty \times \|D - p(X)\|_\infty \times \|\hat{m}_1(X) - m_1(X)\|_2 \\ &= O(\varepsilon_{m_1}) \end{aligned}$$

by Assumption 3.4 and

$$\begin{aligned} \|iia\|_2 &\leq \left\| \frac{1}{p(X)} \right\|_\infty \times \left\| \frac{1}{\hat{p}(X)} \right\|_\infty \times \|\hat{p}(X) - p(X)\|_\infty \times \|\hat{m}_1(X) - m_1(X)\|_2 \\ &= O(\varepsilon_{m_1}) \end{aligned}$$

by Assumption 3.4. The bounds for terms ib , iib and $iiib$ follow similarly such that $\|\bar{\psi}\|_2 = O(\varepsilon_{\max})$. Immediately,

$$\left\| K\left(\frac{Z-z}{h}\right) \bar{\psi} \right\|_2 \leq \|K(u)\|_\infty \times \|\bar{\psi}\|_2 = O(\varepsilon_{\max})$$

by Assumption 3.2.

3.A.1 Proof of Theorem 3.1

Denote $\hat{f}(z) = \frac{1}{Nh^{\lambda_Z}} \sum_{i=1}^N K\left(\frac{Z_i - z}{h}\right)$ and

$$\begin{aligned} i &= f(z)^{-1} \times \frac{1}{Nh^{\lambda_Z}} \sum_{i=1}^N K\left(\frac{Z_i - z}{h}\right) (\psi(W_i, p, m_0, m_1) - \tau), \\ ii &= f(z)^{-1} \times \frac{1}{Nh^{\lambda_Z}} \sum_{i=1}^N K\left(\frac{Z_i - z}{h}\right) (\psi(W_i, \hat{p}, \hat{m}_0, \hat{m}_1) - \psi(W_i, p, m_0, m_1)) \\ &= \underbrace{f(z)^{-1} \times \frac{1}{Nh^{\lambda_Z}} \sum_{i=1}^N \mathbb{E} \left[K\left(\frac{Z - z}{h}\right) \bar{\psi} \right]}_{iia} + \underbrace{f(z)^{-1} \times \frac{1}{Nh^{\lambda_Z}} \sum_{i=1}^N \left(K\left(\frac{Z_i - z}{h}\right) \bar{\psi}_i - \mathbb{E} \left[K\left(\frac{Z - z}{h}\right) \bar{\psi} \right] \right)}_{iib}, \\ iii &= \left(\frac{f(z) - \hat{f}(z)}{f(z)} \right) \times \left(\frac{\hat{f}(z) - f(z)}{f(z)} + 1 \right)^{-1}. \end{aligned}$$

Then we can write

$$\hat{\tau} - \tau = (i + ii) + (i + ii) \times iii.$$

Influence function

Bounding i

Under the conditions stated in Assumption 3.2 and the convergence condition (i) in Assumption 3.5

$$\begin{aligned}
 |i| &= \left| \frac{\frac{1}{Nh^{\lambda_Z}} \sum_{i=1}^N K\left(\frac{Z_i - z}{h}\right) \psi(W_i, p, m_0, m_1)}{\frac{1}{Nh^{\lambda_Z}} \sum_{i=1}^N K\left(\frac{Z_i - z}{h}\right)} - \tau \right| \times \left| \frac{\hat{f}(z) - f(z)}{f(z)} + 1 \right| \\
 &= o_p(1)(1 + o_p(1)) \\
 &= o_p(1).
 \end{aligned}$$

Bounding ii_a

From condition (ii) in Assumption 3.2 and Lemma 3.1 it immediately follows that

$$|ii_a| = \frac{1}{h^{\lambda_Z}} \times |f(z)^{-1}| \times \left| \mathbb{E} \left[K\left(\frac{Z - z}{h}\right) \tilde{\psi} \right] \right| = O(h^{-\lambda_Z} \times \varepsilon_p \times (\varepsilon_{m_0} + \varepsilon_{m_1})).$$

Bounding ii_b

Let $\mathcal{J}_l$ be the set of observation in subsample l and $\mathcal{J}_l^c$ be the set of observations not in subsample l . We can then write

$$|ii_b| \leq \frac{\sqrt{L}}{h^{\lambda_Z} \sqrt{N}} |f(z)^{-1}| \times \frac{1}{L} \sum_{l=1}^L \left| \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{J}_l} \left(K\left(\frac{Z_i - z}{h}\right) \tilde{\psi}_i - \mathbb{E} \left[K\left(\frac{Z - z}{h}\right) \tilde{\psi} \right] \right) \right|.$$

Further, notice that under the sample splitting procedure we have

$$\begin{aligned}
 &\mathbb{E} \left[\left| \frac{1}{L} \sum_{l=1}^L \left| \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{J}_l} \left(K\left(\frac{Z_i - z}{h}\right) \tilde{\psi}_i - \mathbb{E} \left[K\left(\frac{Z - z}{h}\right) \tilde{\psi} \right] \right) \right| \right| \right] \\
 &= \mathbb{E} \left[\left| \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{J}_l} \left(K\left(\frac{Z_i - z}{h}\right) \tilde{\psi}_i - \mathbb{E} \left[K\left(\frac{Z - z}{h}\right) \tilde{\psi} \right] \right) \right| \right] \\
 &\leq \mathbb{E} \left[\left| \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{J}_l} \left(K\left(\frac{Z_i - z}{h}\right) \tilde{\psi}_i - \mathbb{E} \left[K\left(\frac{Z - z}{h}\right) \tilde{\psi} \right] \right) \right|^2 \middle| W_{i \in \mathcal{J}_l^c} \right]^{\frac{1}{2}} \\
 &\leq \mathbb{E} \left[\left| K\left(\frac{Z - z}{h}\right) (\psi(W, \hat{p}, \hat{m}_0, \hat{m}_1) - \psi(W, p, m_0, m_1)) \right|^2 \right]^{\frac{1}{2}} \\
 &= \left\| K\left(\frac{Z - z}{h}\right) \tilde{\psi} \right\|_2 \\
 &= O(\varepsilon_{\max})
 \end{aligned}$$

where the last bound follows from Lemma 3.1. Combining this result with Markov's inequality and condition (ii) in Assumption 3.2 and the fact that L is a constant independent of N implies that $|ii_b| = O_p(h^{-\lambda_Z} \times N^{-\frac{1}{2}} \times \varepsilon_{\max})$.

Bounding iii

We obtain

$$\begin{aligned}
|iii| &= \left| \frac{1}{f(z)} \right| \times |\hat{f}(z) - f(z)| \times \left| \left(\frac{\hat{f}(z) - f(z)}{f(z)} + 1 \right)^{-1} \right| \\
&= O(1) O_p \left(h^r + \frac{1}{\sqrt{Nh^{\lambda_Z}}} \right) O_p(1) \\
&= O_p \left(h^r + \frac{1}{\sqrt{Nh^{\lambda_Z}}} \right)
\end{aligned}$$

by the usual convergence in probability result of the kernel density estimator and the fact that $f(z)$ is bounded away from zero under Assumptions 3.2 and 3.5 (i).

Under the convergence conditions in Assumption 3.5 we collect terms and write

$$\sqrt{Nh^{\lambda_Z}} (\hat{\tau} - \tau) = \frac{1}{\sqrt{Nh^{\lambda_Z}}} \frac{1}{f(z)} \sum_{i=1}^N K \left(\frac{Z_i - z}{h} \right) (\psi(W_i, p, m_0, m_1) - \tau) + o_p(1).$$

Hence, the first claim of the theorem is verified.

Asymptotic normality

We can rewrite the previous result as

$$\begin{aligned}
\sqrt{Nh^{\lambda_Z}} (\hat{\tau} - \tau) &= \frac{1}{\sqrt{Nh^{\lambda_Z}}} \frac{1}{f(z)} \sum_{i=1}^N K \left(\frac{Z_i - z}{h} \right) (\psi(W_i, p, m_0, m_1) - \tau) + o_p(1) \\
&= \underbrace{\frac{1}{\sqrt{Nh^{\lambda_Z}}} \frac{1}{f(z)} \sum_{i=1}^N K \left(\frac{Z_i - z}{h} \right) (\psi(W_i, p, m_0, m_1) - \mathbb{E}[\psi(W_i, p, m_0, m_1) | Z = z_i])}_{ia} \\
&\quad + \underbrace{\frac{1}{\sqrt{Nh^{\lambda_Z}}} \frac{1}{f(z)} \sum_{i=1}^N K \left(\frac{Z_i - z}{h} \right) (\mathbb{E}[\psi(W_i, p, m_0, m_1) | Z = z_i] - \tau)}_{ib} \\
&\quad + o_p(1).
\end{aligned}$$

By following the steps in the proof of Theorem 3.6 in Pagan and Ullah (1999), it can be shown that $|ib| = O_p(N^{\frac{1}{2}} h^{\frac{1}{2} \lambda_Z} h^r)$. Thus, Assumption 3.5 (ii) implies $|ib| = o_p(1)$. Under the usual assumptions on the existence of higher order moments in Assumption 3.2, we can apply the Lyapunov Central Limit Theorem on ia as in Pagan and Ullah (1999, Chapter 3.4). Then

$$\sqrt{Nh^{\lambda_Z}} (\hat{\tau} - \tau) \rightarrow_d N \left(0, \frac{\int K(u)^2 du \times \mathbb{E} \left[(\psi(W_i, p, m_0, m_1) - \tau)^2 | Z = z \right]}{f(z)} \right).$$

3.A.2 Proof of Theorem 3.2

Similar to the proof in Theorem 3.1, we can write

$$\begin{aligned}\hat{\theta} - \theta &= \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \frac{\frac{1}{Nh^{\lambda_Z}} K\left(\frac{Z_j - z_i}{h}\right) \psi(W_j, \hat{p}, \hat{m}_0, \hat{m}_1)}{\frac{1}{Nh^{\lambda_Z}} \sum_{j=1}^N K\left(\frac{Z_j - z_i}{h}\right)} - \theta \\ &= \underbrace{\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \frac{\frac{1}{Nh^{\lambda_Z}} K\left(\frac{Z_j - z_i}{h}\right) \psi(W_j, p, m_0, m_1)}{\frac{1}{Nh^{\lambda_Z}} \sum_{j=1}^N K\left(\frac{Z_j - z_i}{h}\right)} - \theta}_i \\ &\quad + \underbrace{\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \frac{\frac{1}{Nh^{\lambda_Z}} K\left(\frac{Z_j - z_i}{h}\right) (\psi(W_j, \hat{p}, \hat{m}_0, \hat{m}_1) - \psi(W_j, p, m_0, m_1))}{\frac{1}{Nh^{\lambda_Z}} \sum_{j=1}^N K\left(\frac{Z_j - z_i}{h}\right)}}_{ii}.\end{aligned}$$

Bounding ii

Using similar arguments and the notation from the proof of Theorem 3.1, by the triangle inequality we write

$$|ii| \leq (iia + iib) \times iic$$

where

$$\begin{aligned}iia &= \frac{1}{N} \sum_{i=1}^N \left| \frac{1}{Nh^{\lambda_Z}} \sum_{j=1}^N \mathbb{E} \left[K\left(\frac{Z_j - z}{h}\right) \tilde{\psi} \right] \right|, \\ iib &= \frac{1}{N} \sum_{i=1}^N \left| \frac{1}{Nh^{\lambda_Z}} \sum_{j=1}^N \left(K\left(\frac{Z_j - z_i}{h}\right) \tilde{\psi}_j - \mathbb{E} \left[K\left(\frac{Z - z}{h}\right) \tilde{\psi} \right] \right) \right|, \\ iic &= \sup_i |\hat{f}(z_i)^{-1} - f(z_i)^{-1}| + \sup_i |f(z_i)^{-1}|.\end{aligned}$$

Then

$$\begin{aligned}\mathbb{E}[|iia|] &= \frac{1}{h^{\lambda_Z}} \left| \mathbb{E} \left[K\left(\frac{Z - z}{h}\right) \tilde{\psi} \right] \right| \\ &= O\left(h^{-\lambda_Z} \times \varepsilon_p \times (\varepsilon_{m_0} + \varepsilon_{m_1})\right)\end{aligned}$$

by Lemma 3.1.

Equivalently to the proof for Theorem 3.1 denote by $\mathcal{J}_l$ the set of observation in subsample l and by $\mathcal{J}_l^c$ the set of observations not in subsample l . Then

$$|iib| \leq \frac{\sqrt{L}}{h^{\lambda_Z} \sqrt{N}} \frac{1}{N} \sum_{i=1}^N \frac{1}{L} \sum_{l=1}^L \left| \frac{1}{\sqrt{n}} \sum_{j \in \mathcal{J}_l} \left(K\left(\frac{Z_i - z_j}{h}\right) \tilde{\psi}_j - \mathbb{E} \left[K\left(\frac{Z - z}{h}\right) \tilde{\psi} \right] \right) \right|$$

by the triangle inequality and

$$\begin{aligned}
& \mathbb{E} \left[\frac{1}{N} \sum_{i=1}^N \frac{1}{L} \sum_{l=1}^L \left| \frac{1}{\sqrt{n}} \sum_{j \in \mathcal{J}_l}^n \left(K \left(\frac{Z_i - z_j}{h} \right) \tilde{\psi}_j - \mathbb{E} \left[K \left(\frac{Z - z}{h} \right) \tilde{\psi} \right] \right) \right| \right] \\
&= \mathbb{E} \left[\left| \frac{1}{\sqrt{n}} \sum_{j \in \mathcal{J}_1}^n \left(K \left(\frac{Z_i - z_j}{h} \right) \tilde{\psi}_j - \mathbb{E} \left[K \left(\frac{Z - z}{h} \right) \tilde{\psi} \right] \right) \right| \right] \\
&\leq \left\| K \left(\frac{Z - z}{h} \right) \tilde{\psi} \right\|_2
\end{aligned}$$

by the same steps as in the proof of Theorem 3.1. Lemma 3.1 in combination with Markov's inequality and the fact that L is a constant independent of N implies that $|iib| = O_p \left(\frac{\varepsilon_{\max}}{h^{\lambda_Z} \sqrt{N}} \right)$.

Finally, for *iic* we notice that $\sup_i |f(z_i)^{-1}| < C$ by Assumption 3.2 (ii). Additionally, Assumption 3.2 (iv) allows us to write

$$\begin{aligned}
\sup_i |\hat{f}(z_i)^{-1} - f(z_i)^{-1}| &\leq \sup_i |\bar{f}(z_i)^{-2}| \times \sup_i |\hat{f}(z_i) - f(z_i)| \\
&= O_p \left(h^r + \sqrt{\frac{\ln(N)}{N h^{\lambda_Z}}} \right) \\
&= o_p(1)
\end{aligned}$$

by a mean value theorem where $\bar{f}(z_i)$ is a value between $\hat{f}(z_i)$ and $f(z_i)$. Since $\bar{f}(z_i)$ has to be larger than $f(z_i)$, $\sup_i |\bar{f}(z_i)^{-2}| \leq C^2$ by Assumption 3.2 (ii). The uniform convergence rate of the kernel density estimator is for example derived in Masry (1996) under the conditions in Assumption 3.2. Using the rate conditions in Assumption 3.5' (i) yields the convergence result.

Hence, for the overall term we have

$$|\sqrt{N}ii| = O_p \left(\frac{\sqrt{N}}{h^{\lambda_Z}} \varepsilon_p \times (\varepsilon_{m_0} + \varepsilon_{m_1}) + \frac{\varepsilon_{\max}}{h^{\lambda_Z}} \right) = o_p(1)$$

under the coupled convergence conditions of Assumption 3.5'.

Bounding i

For i notice that

$$\sqrt{N}i = \sqrt{N}(\tilde{\theta} - \theta),$$

where

$$\tilde{\theta} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \frac{\frac{1}{N h^{\lambda_Z}} K \left(\frac{Z_j - z_i}{h} \right) \psi(W_j, p, m_0, m_1)}{\frac{1}{N h^{\lambda_Z}} \sum_{j=1}^N K \left(\frac{z_j - z_i}{h} \right)}.$$

Thus, term i gives the contribution of estimating the nonparametric projection of the vector $\psi = \psi(W, p, m_0, m_1)$ with population nuisance parameters on $Z \in \mathcal{Z}$. We notice that the estimator $\tilde{\theta}$ is a special case of Newey's (1994a) full mean estimator. In particular, Theorem 4.2 in Newey (1994a) holds under the condition that the first stage nonparametric estimator is bounded by any norm (and some further regularity conditions). For example,

using Assumption 3.5' (ii) we have

$$N^{\frac{1}{4}} \left\| \sum_{j=1}^N \frac{\frac{1}{Nh^{\lambda_Z}} K\left(\frac{Z_j - z}{h}\right) \psi(W_j, p, m_0, m_1)}{\frac{1}{Nh^{\lambda_Z}} \sum_{j=1}^N K\left(\frac{Z_j - z}{h}\right)} - \tau(z) \right\|_2 = o(1)$$

for the L_2 norm. Therefore, by Theorem 4.2 in Newey (1994a) the influence function has the form

$$\frac{D(Y - m_1(X))}{p(X)} - \frac{(1 - D)(Y - m_0(X))}{1 - p(X)} + m_1(X) - m_0(X) - \theta.$$

Hence, combining the results for terms i and ii gives

$$\sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{D_i(Y_i - m_1(X_i))}{p(X_i)} - \frac{(1 - D_i)(Y_i - m_0(X_i))}{1 - p(X_i)} + m_1(X_i) - m_0(X_i) - \theta \right) + o_p(1)$$

such that by the Central Limit Theorem we obtain

$$\sqrt{N}(\hat{\theta} - \theta) \rightarrow_d N \left(0, \mathbb{E} \left[\frac{\text{Var}(Y|D=1, X)}{p(X)} + \frac{\text{Var}(Y|D=0, X)}{1 - p(X)} + (m_1(X) - m_0(X) - \theta)^2 \right] \right).$$

3.B Details on the bandwidth ranges

Using the notation implied by Definition 3.2, notice that the convergence conditions in Assumption 3.5 imply the following system of inequalities.

$$\begin{aligned}
\frac{1}{2}\lambda_Z\delta_h - \delta_{\varepsilon_{\max}} &< 0 \\
\frac{1}{2} + \frac{1}{2}\lambda_Z\delta_h - (\delta_p + \delta_{m_d}) &< 0 \\
\frac{1}{2} - \frac{1}{2}\lambda_Z\delta_h - r\delta_h &< 0 \\
-\delta_h &< 0 \\
1 - \delta_h\lambda_Z &> 0
\end{aligned}$$

The third and fourth inequality imply $\delta_h > \frac{1}{\lambda_Z + 2r} > 0$. Further, the other inequalities imply $\delta_h < \frac{2(\delta_p + \delta_{m_d}) - 1}{\lambda_Z} < \frac{2\delta_{\varepsilon_{\max}}}{\lambda_Z} < \frac{1}{\lambda_Z}$. It therefore follows that the possible range of the bandwidth can be described by $\frac{1}{\lambda_Z + 2r} < \delta_h < \frac{2(\delta_p + \delta_{m_d}) - 1}{\lambda_Z}$.

3.C Further results of the Empirical Monte Carlo experiment

3.C.1 Further results for different estimators

Table 3.4: Performance of different GATE estimators for $\alpha = 2$

	MSE	Bias	SD	MSE	Bias	SD
50 % treatment share						
	$N = 1,000$			$N = 4,000$		
IPW maxlik lc	7.04	0.92	2.23	2.63	0.50	1.34
AIPW maxlik lp	4.84	1.62	1.11	3.78	1.62	0.52
AIPW maxlik lp squared	3.93	1.07	1.30	2.40	1.05	0.61
AIPW maxlik lc	3.91	0.68	1.69	1.20	0.30	0.94
AIPW ens lp	4.71	1.63	1.06	3.78	1.63	0.52
AIPW ens lp squared	3.75	1.04	1.24	2.39	1.05	0.60
AIPW ens lc	3.79	0.56	1.71	1.17	0.28	0.95
10 % treatment share						
	$N = 1,000$			$N = 4,000$		
IPW maxlik lc	-	-	-	7.81	1.03	2.41
AIPW maxlik lp	-	-	-	-	-	-
AIPW maxlik lp squared	-	-	-	-	-	-
AIPW maxlik lc	-	-	-	-	-	-
AIPW ens lp	23.36	1.61	4.22	376.16	1.80	19.28
AIPW ens lp squared	23.94	1.04	4.48	616.39	1.48	23.13
AIPW ens lc	19.85	0.77	4.31	375.04	0.82	19.35

Notes: Results were obtained as described in Procedure 3 for 1,000 Monte Carlo replications for $\alpha = 2$. The three performance measures are averages over the 1,000 out-of-sample evaluation points. *maxlik* and *ens* indicate that a maximum likelihood based procedure or an ensemble learner based on the Random Forest and the Lasso were used to estimate the nuisance parameters. *lc* and *lp* indicate that the conditional expectation was estimated using local constant nonparametric regression and the linear predictor respectively. *squared* indicates that a second order polynomial was included in the linear model.

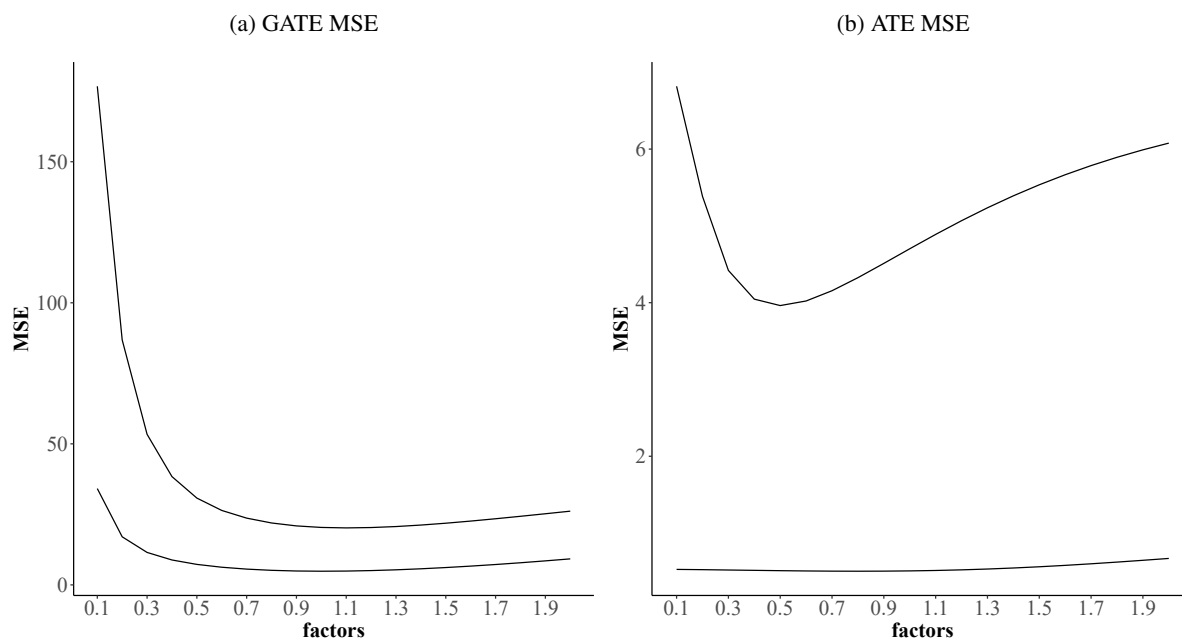
Table 3.5: Performance of different ATE estimators for $\alpha = 2$

	MSE	Bias	SD	MSE	Bias	SD
50 % treatment share						
	$N = 1,000$			$N = 4,000$		
IPW maxlik	0.78	0.02	0.89	0.15	0.03	0.39
AIPW maxlik	0.67	0.02	0.82	0.15	0.03	0.39
AIPW ens	0.63	0.04	0.79	0.15	0.04	0.38
AIPW ens smoothed	0.65	0.10	0.80	0.15	0.05	0.38
10 % treatment share						
	$N = 1,000$			$N = 4,000$		
IPW maxlik	-	-	-	1.11	0.38	0.98
AIPW maxlik	-	-	-	-	-	-
AIPW ens	12.63	0.05	3.55	372.53	0.77	19.30
AIPW ens smoothed	12.67	0.11	3.56	372.52	0.63	19.30

Notes: Results were obtained as described in Procedure 3 for 1,000 Monte Carlo replications for $\alpha = 2$. *maxlik* and *ens* indicate that a maximum likelihood based procedure or an ensemble learner based on the Random Forest and the Lasso were used to estimate the nuisance parameters.

3.C.2 Further results on the bandwidth choice

Figure 3.5: GATE and ATE MSEs for different factors of the LOOCV bandwidth choice with 10 % treatment share



Notes: Results were obtained as described in Procedure 3 for the design with $\alpha = 7$ and 10 % treatment share. The upper line is for 1,000 and the lower line for 4,000 observations. Procedures 1 and 2 were implemented with a second-order Gaussian kernel function and the ensemble learner. The factors on the horizontal axis denote the different values with which the LOOCV bandwidth choice was multiplied in order to obtain the final bandwidth.

3.D Details on the empirical example

3.D.1 Covariates in the dataset

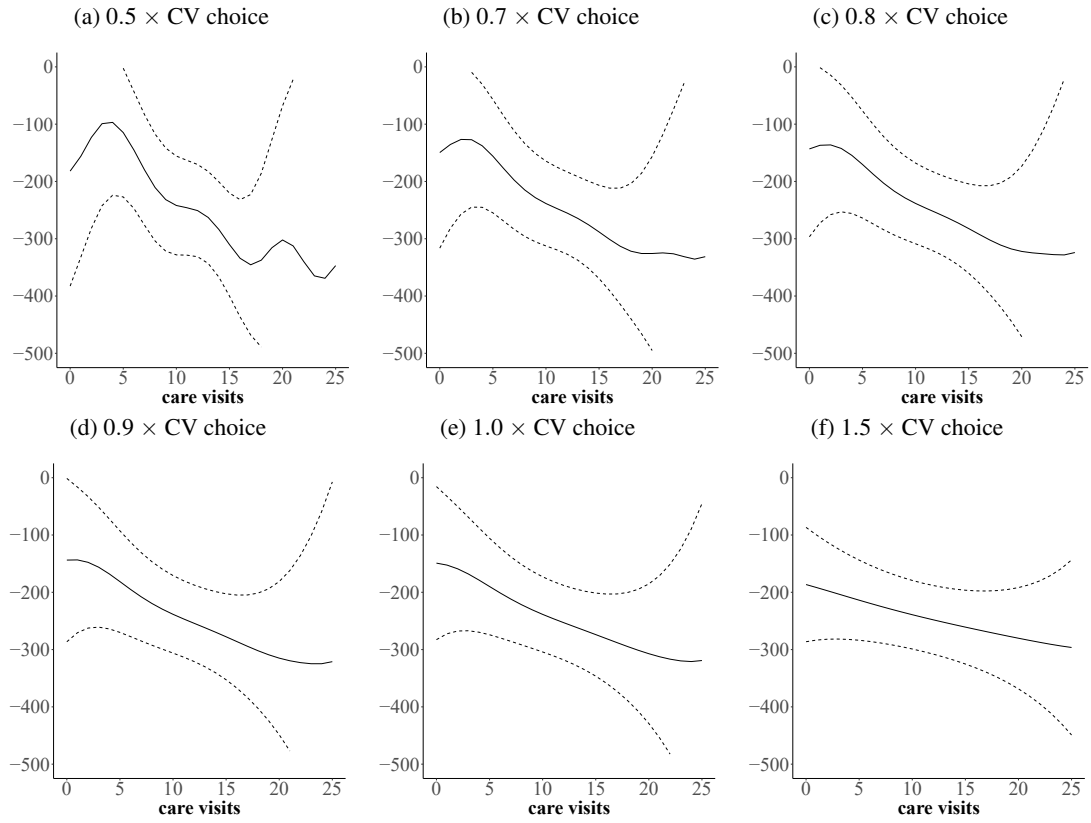
Table 3.6: Description of covariates in the dataset

variable	description	newly created	smalldata	alldata	mean
Y	infant birth weight in grams	no	yes	yes	3,361.68
D	=1 if mother smoked during pregnancy	no	yes	yes	0.19
mmarried	=1 if mother is married	no	no	yes	0.70
mhispanic	=1 if mother is hispanic	no	yes	yes	0.03
fhispanic	=1 if father is hispanic	no	no	yes	0.04
foreign	=1 if mother born abroad	no	no	yes	0.05
alcohol	=1 if alcohol consumed during pregnancy	no	yes	yes	0.03
deadkids	=1 if previous birth were newborn died	no	yes	yes	0.26
mage	mother's age	no	yes	yes	26.50
medu	mother's educational attainment	no	yes	yes	12.69
fage	father's age	no	no	yes	27.27
fedu	father's educational attainment	no	no	yes	12.31
nprenatal	number of prenatal care visits	no	yes	yes	10.76
mrace	=1 if mother is white	no	yes	yes	0.84
frace	=1 if father is white	no	no	yes	0.81
prenatal1	=1 if first prenatal visit in first trimester	no	yes	yes	0.80
prenatal2	=1 if first prenatal visit in second trimester	yes	no	yes	0.15
prenatal3	=1 if first prenatal visit in third trimester	yes	no	yes	0.05
order1	=1 if first infant	yes	yes	yes	0.44
order2	=1 if second infant	yes	no	yes	0.34
order3	=1 if j^{th} infant with $j \geq 3$	yes	no	yes	0.22
birthmonth1	=1 if birth in January	yes	no	yes	0.07
birthmonth2	=1 if birth in February	yes	no	yes	0.08
birthmonth3	=1 if birth in March	yes	no	yes	0.08
birthmonth4	=1 if birth in April	yes	no	yes	0.08
birthmonth5	=1 if birth in May	yes	no	yes	0.08
birthmonth6	=1 if birth in June	yes	no	yes	0.10
birthmonth7	=1 if birth in July	yes	no	yes	0.08
birthmonth8	=1 if birth in August	yes	no	yes	0.10
birthmonth9	=1 if birth in September	yes	no	yes	0.09
birthmonth10	=1 if birth in October	yes	no	yes	0.08
birthmonth11	=1 if birth in November	yes	no	yes	0.08
birthmonth12	=1 if birth in December	yes	no	yes	0.08

Notes: Sample with $N = 4,642$ observations with 864 treated and 3,778 non-treated. 'smalldata' indicates that the variable was also used in Lee et al. (2016). 'newly created' indicates that the variable was additionally created from the original dataset by the authors. 'alldata' contains the specification used for the estimation results in Section 3.6.

3.D.2 Additional sensitivity analysis

Figure 3.6: Sensitivity to bandwidth choice (care visits)



Notes: Results were obtained as described in Procedure 1 with a second-order Gaussian kernel function and different multiples of the LOOCV bandwidth choice. Nuisance parameters were estimated using an ensemble learner comprising Lasso and Random Forest. For Lasso the penalty term was chosen such that the 10-fold cross-validation criterion was minimised. The ensemble weights were chosen by minimising out-of-sample MSE. Asymptotic confidence bands are at the 95% level.

Chapter 4

Paid parental leave and maternal reemployment: Do part-time subsidies help or harm?

jointly with Franziska Zimmert

Employment subsidies can incentivise mothers to shorten employment interruptions after childbirth. We examine a German parental leave reform promoting an early return to work in part-time. Exploiting the exogenous variation in the benefit entitlement length defined by the child's birthday, we apply machine learning augmented semiparametric difference-in-difference estimation using administrative data. The reform yields positive average employment effects that are mainly driven by part-time employment. In particular, the increased attractiveness of part-time work does not cannibalise full-time employment. Conditional effects show that the policy creates heterogeneous incentives depending on the opportunity costs of working part-time.

JEL classification: J21, J22, C14

Keywords: causal machine learning, effect heterogeneity, maternal labour supply, parental leave, Germany

Acknowledgements: The authors are grateful to Michael Knaus, Michael Lechner, Gesine Stephan and Enzo Weber.

Franziska Zimmert gratefully acknowledges financial support by the Graduate Program of the IAB and FAU.

Michael Zimmert gratefully acknowledges financial support from the Swiss National Science Foundation (SNSF). The study is part of the project 'Causal Analysis with Big Data' of the Swiss National Research Program 'Big Data' (NRP 75).

4.1 Introduction

Motherhood and related employment interruptions are still one of the main causes for the gender wage gap and different labour market prospects for women (Lundborg, Plug, and Rasmussen, 2017). While gender roles converge and women catch up in terms of education and job choice (Goldin, 2014), women face statistical discrimination even *ex ante* in expectation of potential motherhood and its related costs to the firm (Jessen, Jessen, and Kluve, 2019). As long career breaks imply hiring and training costs for a new candidate, employers might anticipate motherhood in their recruiting process. Even if young women find a suitable job, starting a family will fundamentally change their choice set and potentially lead to a reduction of working time with the consequence of worsened employment prospects and lower wage expectations (Goldin, 2014). Hence, labour market interventions promoting an earlier return to employment may be regarded as a suitable tool to cushion these adverse effects. In particular, this study analyses if subsidised part-time work after child birth shortens employment breaks and affects the working time pattern of mothers sustainably.

Several studies investigate the effect of parental leave on maternal employment outcomes. For European countries and Canada many authors find that mothers delay their return to work for extended parental leave regulations (Baker and Milligan, 2008; Bergemann and Riphahn, 2015; Dahl, Løken, Mogstad, and Salvanes, 2016; Kluve and Schmitz, 2018; Kluve and Tamm, 2013; Lalive and Zweimüller, 2009; Schönberg and Ludsteck, 2014). In the United States, however, there is no nationwide paid leave period despite some states notably California agreed on a *paid* protection period (Rossin-Slater, Ruhm, and Waldfogel, 2013). Moreover, previous studies find that short *unpaid* protection periods like the Federal Maternal Legislation Act (FMLA) in the United States have small effects on maternal employment and wages (Baum, 2003; Waldfogel, 1999). Two articles analyse working hours after mothers returned to work. Schönberg and Ludsteck (2014) show for Germany that several extensions in paid leave coverage between 1979 and 1992 lead to short-term reductions in full-time employment, but do not have any long-run effects. In contrast, Kluve and Schmitz (2018) find even long-lasting positive effects on full-time employment for mothers from the upper tercile of the income distribution which is the group the most affected by a German parental leave reform in 2007.

A common feature of these articles is their focus on policies incentivising mothers to completely abstain from working during parental leave. Since prolonged employment interruptions might conflict with the longer-term labour market participation of mothers, these policies might be more harmful than helpful for mothers. Additionally, the policies investigated in these articles represent relatively strong intervention in the labour market situation of mothers when they were implemented. In contrast, we investigate the effect of part-time subsidies, promoting an earlier return to the labour market, on reemployment outcomes in an already generous parental leave regime. Understanding these effects are crucial as promoting the longer-term participation of mothers in the labour market and the economy is a fundamental goal of parental leave regulation. On a higher level, Goldin (2014) even considers length and timing of working hours to be the ‘last chapter’ for reducing the gender wage gap. From a theoretical perspective, however, the question arises if part-time subsidies are a suitable labour market instrument to achieve these goals. As part-time work becomes more attractive relative to extending parental leave *and* to working full-time, the overall employment effects are unclear. Even if the effect on the extensive employment margin is positive, the policy might foster the so-called part-time trap. In particular, employees might be unable to increase their agreed working hours to a full-time job at a later point in time.

This motivates to empirically assess the effects of an extended part-time subsidy. In detail, we examine the effect of

subsidising part-time work on the maternal working time pattern after the (2 month) period mothers are protected for health reasons after the birth of a child. Mothers affected by a new law coming into force for births from July 2015 onwards are encouraged to combine income from part-time work and public subsidies. The implementation of the reform enables to exploit exogenous variation in the entitlement length and benefit amount of parental leave affecting mothers with children born later than June 2015. We compare those treated mothers with women having children born shortly before the cut-off date. To account for seasonal effects resulting from different patterns for the start of the school year, we use difference-in-differences estimation. As different factors such as local differences in the economy or personal characteristics of women may differently shift the employment trends, we include a large list of covariates. In particular, we use administrative data covering the population of German employees with detailed longitudinal information. We apply recently proposed doubly-robust semiparametric difference-in-differences (DiD) estimation (Chang, 2020; Sant’Anna and Zhao, 2020; Zimmert, 2020). It allows to include covariates in a data-driven way using state of the art machine learning algorithms. We argue that the inclusion of a large set of covariates available from the administrative data makes our identifying assumptions more credible. In particular, our conditional model with control variables is very close to the unconditional model and we conclude that our results are robust against a large set of potentially confounding variables. Additionally, we avoid common problems in parametric DiD estimation like arbitrary functional form assumptions and misspecification errors of the included covariates. Moreover, as first shown in Abadie (2005), semiparametric DiD estimation allows to infer heterogeneous effects that uncover for which subgroups the reform was effective. We give an identification result that motivates a new estimator for heterogeneous treatment effects estimation in the DiD setting.

Our results show that women exposed to the reform have on average an about two percentage points higher probability to return to work within the first year which amounts to about 14 percent of the pre-reform level. Like the reform intended, this increase is mainly driven by part-time employment. However, these positive average effects do not continue after the child’s first birthday. Although limited to a two-year perspective, these findings cast doubt upon sustainably strengthening female employment prospects. Besides, on average we cannot confirm the existence of a part-time trap for this short time horizon. To some extent, the heterogeneous effects show a more refined pattern. Especially middle income mothers are willing to take up the new part-time subsidy. In turn, mothers with higher income expectations might fear future income losses in case they accept a lower-paid part-time job.

The study proceeds as follows. The next two sections describe the institutional setting and the theoretical implications of the reform. Section 4.4 and 4.5 explain the estimation strategy and the exploited data before presenting the estimation results and sensitivity checks in Section 4.6. The study ends with a discussion of the results and a conclusion.

4.2 Institutional background

The German system of birth-related legal work interruptions distinguishes two different forms: maternity protection and parental leave. The first concept describes a period of six weeks before and eight weeks after child birth in which mothers are not allowed to work due to health risks. The latter wants to facilitate the employment continuity of parents and especially of mothers by defining a period up to which parents have the right to return to their previous employer. Table 4.1 gives an overview of the two most important parental leave (*Elterngeld* abbreviated *EG*) regulations over the last years. This study will focus on the regulations for births from July 2015 onwards (last column).

Table 4.1: Parental leave regulations over time

births	$\geq 01/2007$: <i>Elterngeld (EG)</i>	$\geq 07/2015$: <i>Elterngeld (EG)</i> or <i>ElterngeldPlus (EG+)</i>
maximum unpaid entitlement length	36 months	36 months
maximum paid entitlement length	12 months extension option: 24 months (half amount)	<i>EG</i> : 12 months or <i>EG+</i> : 24 months
benefit amount Euro/month	<i>basic parental benefit amount</i> : 65 % of average net monthly income measured during 12 months before birth 67 - 100 % if average net monthly income < 1,200 Euro minimum 300 maximum 1,800	<i>EG</i> : full <i>basic parental benefit amount</i> <i>EG+</i> : up to 0.5 * <i>basic parental benefit amount</i>
means testing	no	no
employment during benefit receipt	≤ 30 hours/week	≤ 30 hours/week
benefit deduction in case of part-time work	yes	yes, but <i>EG+</i> may imply an equal total benefit amount (compare Table 4.2)
'daddy' months	2 additional months	<i>EG</i> : 2 additional months <i>EG+</i> : 4 additional months for each parent if both decide for <i>EG+</i>

Source: Own representation according to *Bundeselterngeld- und Elternzeitgesetz (BEEG)*.

4.2.1 Regulations of parental benefits prior to the reform in 2015 (*Elterngeld EG*)

Former regulations (see second column in Table 4.1) for births from January 2007 onwards aimed at facilitating motherhood for working women and engaging fathers in child care. It standardised the maximum benefit receipt duration to 12 months with two additional months if both parents are on leave (so-called daddy months, see Tamm, 2019 for his evaluation). Besides, a replacement rate λ of 65 percent (up to 100 percent for parents with low income) was introduced determining the *basic parental benefit amount* based on the average net monthly income measured during the twelve months before child birth denoted by $\bar{y}$. Previously none-working or low-income mothers receive a minimum of 300 Euro per month while the maximum was set to 1,800 Euro per month. Part-time work up to 30 hours per week is also possible, but reduces the benefit amount. For part-time working mothers the difference between former and current net income ($\bar{y} - y$) serves as reference value for the replacement rate τ that is also between 65 and 100 percent. Optionally, mothers could choose to double the benefit receipt duration for half of the benefit amount (*extension option*). About 18 percent made use of this option in 2014 (Federal Statistical Office, 2015c).¹

$$l_{EG} = \begin{cases} = 300 \leq \lambda \bar{y} \leq 1,800 \text{ if not working, paid for 12 months} \\ = 300 \leq \tau(\bar{y} - y) \leq 1,800 \text{ if part-time working,} \\ \text{paid for 12 months} \\ = 150 \leq \frac{\lambda}{2} \bar{y} \leq 900 \text{ if not working, paid for 24 months} \\ = 150 \leq \frac{\tau}{2}(\bar{y} - y) \leq 900 \text{ if part-time working,} \\ \text{paid for 24 months} \\ = 0 \text{ else} \end{cases}$$

Kluge and Schmitz (2018) show that mothers with income from the upper tercile of the distribution benefit the most from these parental leave regulations having positive effects on full-time employment up to the child's fifth birthday. Moreover, Kluge and Schmitz (2014, 2018) highlight the formation of a social norm to return to work at the end of the maximum entitlement length of 12 months which is challenged by the new regulations coming into force in July 2015.

4.2.2 The reform in July 2015 (*ElterngeldPlus EG+*)

The new regulations coming into effect for births from July 2015 onwards enable mothers to choose between the *EG* and the newly introduced *EG+* option. In *EG+* they receive up to half of the basic benefit amount for a period of 24 months. In particular, a new payment scheme for part-time work is added to the previous regulation:

$$l_{EG+} = \begin{cases} = 150 \leq \min(\tau(\bar{y} - y), \frac{\lambda \bar{y}}{2}) \leq 900 \text{ if part-time working,} \\ \text{paid for 24 months} \\ = 0 \text{ else} \end{cases}$$

Table 4.2 gives several examples for the calculation of the subsidy under the new and the old regime. Under the

¹Notice that even though the total benefit amount is the same for *EG* and the *extension option*, mothers may still prefer the *extension option* for tax reasons (progressive tax rates in Germany).

new regime mothers get higher subsidies for working part-time (examples II and III). Compared to the old regime, a return in full-time to the previous job (example IV) or not working at all during parental leave (example I) does not lead to different payment schemes. The regulations for births until July 2015 discourage mothers to return to work before parental benefits expire as current labour income is taken into account for the calculation of the subsidy. Official statistics show that the majority of female benefit recipients with children born in the third quarter 2015 chooses the *full basic amount* (81 percent) with an average benefit amount of monthly 757 Euro and in total 8,797 Euro (Federal Statistical Office, 2019). The remaining 19 percent decided for the second option (*EG+*) and received on average 492 Euro per month with a slightly higher total sum of 9,130 Euro compared to *EG*. Thus, the share of mothers with a receipt duration of up to 24 months remains constant compared to the extension option of the old regime. Parents can also share the parental leave period. The two additional daddy months in full-time result in four extra months under the new regime in part-time. Parents choosing this option are eligible for another four months of benefit receipt resulting in 32 months all together. About ten percent of all male benefit recipients decided to be on leave for at least four months in the relevant birth cohort (Federal Statistical Office, 2019) which amounts to about three percent of all births in the third quarter of 2015. Hence, we expect only small effects on paternal labour supply as a channel for maternal employment adjustments. Moreover, Tamm (2019) finds that the daddy months established by the previous reform in 2007 do not significantly affect paternal involvement in child care and housework on weekdays for those currently on leave.² Hence, paternal leave can rather be considered as shared family time than a promotion of maternal employment.

Unaffected by both reforms, the *unpaid* maximum parental leave duration amounts to 36 months from child birth onwards, i.e., mothers have the right to return to their previous employer until the child's third birthday.

4.2.3 The expansion of subsidised child care

The lack of suitable child care slots may prevent mothers from returning to work. Recent parental leave changes are part of different family policies notably the child care expansion for under three-year-olds starting in 2005 and culminating in a legal claim for a child care slot from August 2013 onwards (see Zimmert, 2019). Before the first parental leave reform in 2007, only 13.6 percent of children younger than three years old attended subsidised child care (Federal Statistical Office, 2006) increasing to 32.9 percent in 2015 (Federal Statistical Office, 2015b). Although also under one-year-olds have a claim for a child care slot if both parents are working, jobseeking or in education,³ only 2.6 percent of this age group attend child care in 2015 (Federal Statistical Office, 2015b). These numbers tend to be higher in urban areas and in East Germany (4.1 percent) as a legacy of the former GDR. As official statistics lack information on the number of authorised child care slots, we exploit data from the survey FiD (DIW Berlin/SOEP, 2014, wave 2013) to explore the reasons for the low early child care coverage. Similar to official statistics, the survey provides a coverage rate for under one-year-olds of 3.1 percent. It shows that 88 percent of parents not making use of a child care slot consider their child to be too young while five percent indicate the lack of suitable child care slots. Hence, we conclude that institutional restrictions do not determine the low coverage rate but attitudes towards external child care.

²In turn, Tamm (2019) and also Patnaik (2019) find that paternal leave can strengthen paternal involvement in child care and housework in the longer term, i.e., beyond the leave duration. Unfortunately, the data set does not allow to examine the policy from a comprehensive household context as information on paternal employment and subsidy receipt are unknown.

³This regulation is defined by §24 SGB VIII.

Table 4.2: Calculation of benefit amount *Elterngeld* (EG) and *ElterngeldPlus* (EG+)

Example: net monthly income before birth = 2,000	
Parental benefit amount	
I: net monthly income after birth = 0	
EG	$12 \times 2,000 \times 0.65 = 15,600$
extension option	$24 \times 0.5 \times 2,000 \times 0.65 = 15,600$
EG+	$24 \times 0.5 \times 2,000 \times 0.65 = 15,600$
II: net monthly income after birth = 1,200	
EG	$12 \times (2,000 - 1,200) \times 0.77 = 7,392$
extension option	$24 \times 0.5 \times (2,000 - 1,200) \times 0.77 = 7,392$
EG+	$24 \times (2,000 - 1,200) \times 0.77 = 14,784$
III: net monthly income after birth = 500	
EG	$12 \times (2,000 - 500) \times 0.65 = 11,700$
extension option	$24 \times 0.5 \times (2,000 - 500) \times 0.65 = 11,700$
EG+	$24 \times 0.5 \times 2,000 \times 0.65 = 24 \times 650 = 15,600$ (cap= $0.5 \times 2,000 \times 0.65 = 650$)
IV: net monthly income after birth = 2,000	
EG	$12 \times 300 = 3,600$
extension option	$24 \times 0.5 \times 300 = 3,600$ (minimum amount)
EG+	$24 \times 0.5 \times 300 = 3,600$ (minimum amount)

Notes: Income in Euro. The table gives several examples for the calculation of the benefit amount under the new regime. Mothers can optionally decide for the full basic benefit amount for a period of 12 months (EG) which amounts to 1,300 Euro per month or 15,600 Euro in total in the example. The total sum is not affected if the mother decides for half the amount (650 Euro per month) for the longer period of 24 months (15,600 Euro). However, if she has net earnings of 1,200 Euro per month (Example II), she will only receive additional 616 Euro for up to 24 months which is in total less than Example I. Under Example III with a monthly income of 500 Euro she can decide for 975 Euro for one year (in total 11,700 Euro) or 650 Euro for two years (in total 15,600 Euro). In case she has an equal income than before (Example IV) she will receive the minimum amount of 300 or 150 Euro respectively.

Source: Own representation based on Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ) (2018).

4.3 Theoretical effects of part-time subsidies

In parental leave mothers face an optimization problem with three possible options. They can stay in parental leave and receive income determined by the replacement rate λ . Alternatively, they can decide to return to the labour market either in part-time or in full-time. If the mother decides to return in part-time and is eligible to the part-time subsidy she receives income from part-time work and the part-time subsidy determined by the replacement rate τ . The introduction of EG+ increased the amount of the part-time subsidy. To see this, notice that if the mother receives part-time subsidies for 24 months the replacement rate doubles from $\frac{\tau}{2}$ to τ under the new regime. This has two implications. Firstly, on the intensive margin an increase in the part-time subsidy makes the part-time option more attractive relative to working full-time. Hence, the (full-time equivalent) income that makes the mother indifferent between working part-time and full-time will increase. It follows that given the mother decides to return to the labour market, it is more likely that she returns in part-time. Secondly, following the above argumentation, the reservation wage at the extensive margin may increase, as full-time working gets less attractive. However, an increased part-time subsidy may also lower the reservation wage because working part-time becomes more attractive relative to staying in parental leave.

To conclude, theory predicts a positive effect of the reform on part-time employment. Since the effects on the extensive margin are ambiguous, though, the question arises if the hypothesised increase in part-time work cannibalise full-time jobs or has positive overall employment effects. We will test these implications empirically in the next sections.

4.4 Empirical strategy

4.4.1 Identification

We exploit the exogenous variation induced by the parental leave reform to define an indicator variable $D \in \{0, 1\}$ such that $D = 0$ whenever a mother gave birth shortly before July 2015 and $D = 1$ whenever the mother gave birth shortly after that date. A possible identification strategy could be to compare mothers in $D = 0$ with those in $D = 1$ in case many data points are available exactly at the cut-off. Any difference in the employment outcomes of interest Y could now be attributed to the change of the part-time subsidy scheme if nothing else drives a potential difference in the outcomes.⁴ However, any estimation strategy, that exploits information away from the cut-off date to increase the sample size, might potentially be exposed to seasonal patterns and time trends. In our setting the starting date of the school year could invalidate such an analysis. Depending on the federal state, the school year usually starts in August or September and child care attendance follows this time plan. Children who are already one year old have better chances to get a child care slot. This implies that children born before the cut-off date who are slightly older than those born after the cut-off date might have a higher probability to attend child care. If employment decisions of mothers systematically differ shortly before and after the cut-off date due to the availability of public child care, then a measured difference in the outcomes cannot be purely attributed to the reform. We account for this by comparing the difference in outcomes in 2015 ($T = 1$) to the difference in the previous year ($T = 0$). Similar to prior articles (Cygan-Rehm, 2016; Cygan-Rehm, Kühnle, and Riphahn, 2018; Schönberg and Ludsteck, 2014), this DiD identification strategy yields an average treatment effect for the treatment group (ATET) under certain assumptions. To clarify things, consider the potential outcomes frame-

Table 4.3: Treatment definition

Cut-off 01/07/2015	Control group $D = 0$	Treatment group $D = 1$
$T = 0$	mid May - mid June 2014	mid July - mid August 2014
$T = 1$	mid May - mid June 2015	mid July - mid August 2015

Source: Own representation.

work proposed by Rubin (1973). In general, denote random variables with capital letters and its realizations with lowercase letters. Define the potential outcome for the two time periods as $Y^d(0)$ and $Y^d(1)$ such that for every observation in the sample only the potential outcome with $D = d$ of the realised outcome is observed. Further, we observe some covariates which we denote by X . Then, Heckman et al. (1997) show that our parameter of interest $ATET = \theta = \mathbb{E}[Y^1(1) - Y^0(1)|D = 1, T = 1]$ is identified under the following set of assumptions.

Assumption 1 (Common trends):

$$\mathbb{E}[Y^0(1)|X, D = 0, T = 1] - \mathbb{E}[Y^0(0)|X, D = 0, T = 0] = \mathbb{E}[Y^0(1)|X, D = 1, T = 1] - \mathbb{E}[Y^0(0)|X, D = 1, T = 0].$$

In words, conditional on X the potential outcomes for the control group would have followed parallel trends in the absence of the reform. In our setting this means that the difference in outcomes between mothers giving birth shortly before and after the cut-off date stays constant between 2014 and 2015 in the absence of the reform. In general, the assumption is empirically not testable and there might be evidence that can raise doubts concerning the validity of the assumption. Following the standard in the literature, in our sensitivity analysis we estimate effects

⁴This identification strategy was applied by some studies in the context of parental leave implementation using a cut-off date (e.g., Dahl et al., 2016).

for periods where we would not expect an effect (placebo reform). Crucially, the assumption might only hold conditional on covariates X . For example Buckles and Hungerman (2013) show that socio-economic characteristics (marital status, race, education and age) of mothers are strongly correlated with the quarter of birth. Our specification with covariates therefore includes covariates on the socio-economic background of mothers like education, age, migration background and the employment history. Common trend confounding due to seasonal effects is also mitigated by only using observations from a relatively narrow time window around the cut-off date. Also local economic differences or differences between occupations might affect the trends differently. We therefore also include geographic and occupational dummies as covariates.

The inclusion of a large set of control variables might be advisable to test the robustness of an unconditional model against a conditional model. If the unconditional mean differences drastically differ from an estimator with many included controls, this may at least be interpreted as a non-robustness against the chosen specification. In other words, if specifications with many control variables shift the results, it is very likely that some form of *observed* common trend confounding, that may or may not be fully adjusted for, takes place. Clearly, this argument does not rule out some form of common trend confounding, that is unrelated to the rich set of control variables included.

Assumption 2 (Observational rule): The outcome process follows the observational rule

$$Y = D(TY^1(1) + (1 - T)Y^1(0)) + (1 - D)(TY^0(1) + (1 - T)Y^0(0)).$$

Hence, we require that reform exposure of one mother does not affect the outcome of another mother. The assumption can be violated if being exposed to the reform has an impact on a colleague's or a friend's reemployment decision. While we cannot completely rule out this kind of peer effect, we argue that the narrow birth interval of four weeks for reform exposure makes the occurrence of peer effects very unlikely.⁵

Assumption 3 (No anticipation): $\mathbb{E}[Y^1(0) - Y^0(0)|X, D = 1, T = 0] = 0$.

This assumption requires that being exposed to the reform has no effect prior to the reform and thus rules out anticipation effects. It might be violated if mothers plan to give birth to their child in order to benefit from the new regime. This kind of anticipation can occur in two different forms. Firstly, women considering to become mothers could have tried to plan conception and secondly, they could have tried to postpone the birth shortly before the calculated date of birth. The first type of anticipation is relevant for those mothers knowing about the reform before they are pregnant. However, the German parliament approved the law only in November 2014. Hence, knowledge about the reform becoming definitely effective for births from July 2015 onwards was less than nine months before the cut-off date available when concerned mothers have already been pregnant. To rule out the possibility that mothers might have heard from the draft and waited for another few months, we analysed the monthly birth numbers relative to the overall births in 2015 and compared these shares to those in 2014. Both statistics from the German Federal Statistical Office and the imputed birth numbers from the *Employment History (BeH)* used in the subsequent analysis show a similar movement in 2014 and 2015 and one cannot detect any sudden increase in the second half of 2015 (see Figure 4.1). We handle the second type of anticipation, trying to postpone the birth-date that might especially relevant for planned Caesarian sections, by dropping individuals with births two weeks around the cut-off date. Obviously, the two weeks rule is arbitrary and we check the sensitivity against it in Section 4.6. Furthermore, the share of mothers wanting a Caesarian section is rather low in Germany (at maximum two to three percent, for details see Kolip, 2012). The concrete definitions of the indicators D and T are summarised in Table 4.3.

⁵Welteke and Wrohlich (2019) identify peer effects for mothers with births between a much longer period (July 2007 and December 2009).

Figure 4.1: Monthly share of births in 2014 and 2015



Source: Own calculations from *BeH* and the German Federal Statistical Office (2014, 2015a).

Assumption 4 (Common support): $p(X) < 1$ where $p(X) = \mathbb{E}[D|X]$.

Thus, we exclude perfect predictability for belonging to group $D = 1$. For our estimation procedure we avoid extreme weights by dropping observations with no overlap.

4.4.2 Estimation of average effects

Given these assumptions, different estimands can be shown to identify the ATET. The usually used linear DiD interaction model relies on several restrictive assumptions. As shown by Lechner (2011), ATET is only identified in the linear interaction model if additional linearity assumptions are imposed on the potential outcome models. This may as such be arbitrary, because under linearity, the effects of D and T are implicitly assumed to be constant, covariates can only affect Y linearly and effects are homogeneous across groups $D = 0$ and $D = 1$. Further, the linear interaction model is per se misspecified for binary outcome variables because predicted outcomes can be larger than one or smaller than zero. Therefore, we instead rely on results of the semiparametric DiD literature. For example, Heckman et al. (1997), Abadie (2005) and Lechner (2011) propose different variations of matching and inverse probability weighting type estimators. Recently, Chang (2020), Sant'Anna and Zhao (2020) and Zimmer (2020) propose DiD estimators that combine propensity score and outcome estimation (Augmented Inverse Probability Weighting, AIPW). It can be shown that

$$ATET = \theta = \mathbb{E} \left[\frac{1}{p_D} \frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{1 - p(X)} (Y - \gamma(X, T)) \right] \quad (4.4.1)$$

where $p_D = \mathbb{E}[D]$, $p_T = \mathbb{E}[T]$ and $\gamma(X, T) = \mathbb{E}[Y|X, T, D = 0]$ (see Appendix 4.B for details).

Our model depends on first stage estimators of the propensity score $p(X)$ and the outcome model $\gamma(X, T)$. A major advantage of this class of estimands is that they are doubly robust in the sense that when either the outcome model

or the propensity score is misspecified, the estimator is still consistent. Misspecification of the propensity score or the outcome model is a particular concern when using parametric models. Depending on the concrete setting, the researcher faces arbitrary decisions how to include covariates in the model. Firstly, given a set of potential controls, it is a priori unclear which ones to include in the model. E.g., as argued before, controlling for a large set of regional dummies might improve the credibility of the common trend assumption in our case. However, it remains for example unclear whether we should use dummies at the state or district level. Secondly, the functional form of the covariates (polynomials, interactions) that enter the model has to be manually chosen by the researcher. As employment trends might, e.g., differ by age, controlling for this variable could be necessary. Still, it is unclear if the covariate should enter the model in squared, some higher order polynomial form or interacted with another covariate. So-called supervised machine learning algorithms (for an overview see Hastie et al., 2009) partly avoid these problems. In our application a major advantage of using machine learning algorithms compared to standard parametric models is that we can exploit the rich information in the administrative data set more effectively.⁶ In particular, we do not rely on a certain specification but choose the covariates and their (implicit) functional form in a data-driven way. Combining machine learning first stages and the nonparametric second stage, we are able to reduce the sensitivity of our results towards functional form assumptions or arbitrary specification choices to a minimum. Building on the double machine learning results of Chernozhukov et al. (2018a) and the results in Chang (2020) and Zimmert (2020) it can be shown that the estimator based on the sample analogues of (4.4.1) converges with square-root-N to a normal distribution and has the asymptotic variance

$$\sigma^2 = \mathbb{E} \left[\left(\frac{1}{p_D} \frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{1 - p(X)} (Y - \gamma(X, T)) - \theta \right)^2 \right] \quad (4.4.2)$$

as long as the propensity score and the outcome model are consistent and the product of their convergence rates achieves $N^{-\frac{1}{2}}$ (see Appendix 4.B for a heuristic argument). These are much lower requirements than for example those needed for parametric models. Importantly, the rate conditions are satisfied for popular machine learning algorithms like Lasso (e.g., Belloni and Chernozhukov, 2013) or Random Forests (Wager and Walther, 2016) under particular forms of sparsity. Hence, the flexibility or dimensionality of the models used can grow with the sample size as long as it grows at a somewhat slower rate. Chernozhukov et al. (2018a) show that an additional requirement for the validity of the asymptotic results is that training and prediction sample need to be separated. This gives rise to the ATET estimation procedure.

The algorithm splits the sample in two different complementary sub-samples and the propensity score as well as the outcome projection are estimated in one of the samples. Then the values of the propensity score and the outcome projection are predicted in the other sample. Subsequently, this procedure is reverted such that one obtains a vector of propensity score and outcome projection predictions for the whole sample. These first step predictions are then plugged into the sample analogue of (4.4.1). Of course, one could extend this estimation principle and split the sample into much more sub-samples. This may increase the small sample efficiency of the estimator because more information could be used for the estimation of the first step parameters. However, it also drastically increases the computational burden of the procedure. In our application we argue that the sample is large enough such that estimation on the 50 percent sub-sample should not decrease efficiency too much.

For the prediction task we use a combinations of Lasso and Random Forests. While the Lasso as a form of penalised regression can be seen as a global nonparametric method, Random Forests are ensembles of regression trees and therefore a local nonparametric method. We merge the predictions from the two methods by choosing

⁶Parametric models can be regarded as submodels among the many options the algorithm can choose.

Procedure ATET estimation

Introduce the sub-sample index $l = 1, 2$ and denote the corresponding information set by $\mathcal{S}_l$ as well as its complement by $\mathcal{S}_l^C$.

1. Randomly split the sample in equally sized sub-samples 1 and 2.

2. **for** $l = 1$ **to** 2 **do**

Estimate the propensity score $p(X)$ and the outcome projection $\gamma(X, T)$ in the sample with $\mathcal{S}_l^C$ using any suitable machine learning method or an ensemble of them.

Predict $\hat{p}(X)$ and $\hat{\gamma}(X, T)$ in the sample with $\mathcal{S}_l$.

end

3. Denote $\hat{p}(X_i) = \hat{p}(X_i)_{l=1,2}$ and $\hat{\gamma}(X_i, T_i) = \hat{\gamma}(X_i, T_i)_{l=1,2}$. Then construct the vector with elements $\frac{1}{p_D} \frac{T_i - p_T}{p_T(1-p_T)} \frac{D_i - \hat{p}(X_i)}{1 - \hat{p}(X_i)} (Y_i - \hat{\gamma}(X_i, T_i))$ for $i = 1, \dots, N$ and estimate ATET as

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^N \frac{1}{p_D} \frac{T_i - p_T}{p_T(1-p_T)} \frac{D_i - \hat{p}(X_i)}{1 - \hat{p}(X_i)} (Y_i - \hat{\gamma}(X_i, T_i)).$$

out-of-sample mean squared error optimal weights. In this way, we obtain a purely data-driven procedure that assigns a high weight to the machine learner which shows a good predictive performance. This should make our procedure more robust against the tuning parameter choices of the two estimators in the ensemble.⁷

4.4.3 Estimation of heterogeneous effects

Yet another advantage of the identification result in (4.4.1) is its capability to infer subgroup specific average effects. In particular, denote a subset of the observed covariates by $Z \subseteq X$. In our case Z might for example include dummies for income groups or whether the mother worked part-time before parental leave. Then in order to assess how the effect of the reform varies among these subgroups we are interested in the parameter

$$\theta(z) = \mathbb{E} [Y^1(1) - Y^0(1) | Z = z, D = 1, T = 1].$$

The parameter represents a so-called conditional average treatment effect for the treatment group (CATET⁸). Define the propensity score conditional on Z as $\mathbb{E}[D|Z] = p(Z)$. Then we can show that under the assumptions in Section 4.4.1 and the further assumption that $p(Z) > 0$, the CATET is identified as

$$\theta(z) = \mathbb{E} \left[\frac{1}{p(Z)} \frac{T - p_T}{p_T(1-p_T)} \frac{D - p(X)}{1 - p(X)} (Y - \gamma(X, T)) \middle| Z = z \right]. \quad (4.4.3)$$

The details for this result are provided in Appendix 4.B. Equation (4.4.3) suggests to estimate CATET as a projection of the reweighted outcome on Z . A similar strategy was proposed in Abadie (2005) for DiD designs. However, the estimator of Abadie (2005) for the CATET relies on least squares regression weighted by the propensity score $p(X)$. Since we estimate $p(X)$ with our ensemble learner described in Section 4.4.2, inference for this estimator might be very complicated in our setting. We therefore reweight the transformed outcome also used in (4.4.1) for average effect estimation by $p(Z)$ instead of p_D to account for the fact that we are interested in a subpopulation that

⁷For the Lasso we choose the penalty term by 5-fold cross-validation and otherwise rely on the default values in the **R**-package `glmnet`. The Random Forest is estimated using the default values in the **R**-package `ranger`.

⁸For some recent contributions in other settings see Abrevaya et al. (2015), Fan et al. (2020), Lee et al. (2016), Semenova and Chernozhukov (2020), Wager and Athey (2018), and Zimmert and Lechner (2019).

is defined conditional on $D = 1$ and $Z = z$. In practice, Z is low-dimensional and hence $p(Z)$ can for example be estimated using Logit regression. Term (4.4.3) then suggests to simply use ordinary least squares (OLS) regression of the transformed outcome on the independent variables Z . Semenova and Chernozhukov (2020) show that this type of estimation strategy leads to valid inference for the OLS coefficients even when the first stages $p(X)$ and $\gamma(X, T)$ were predicted with sophisticated machine learning algorithms as described in the previous section. In particular, they demonstrate that the first stage estimations have no bearing on the asymptotic behaviour of the estimator. Given the results in Chang (2020), Zimmert (2020) and the argument made in Appendix 4.B, we postulate that this also holds for the DiD setting. A rigorous formal argument is, however, beyond the scope of this study. In contrast to the standard subgroup analysis, our procedure provides joint OLS inference on the coefficients. The method may therefore also be suitable to avoid the usual multiple testing problem when analysing heterogeneous effects with many subgroups.

4.5 Data

To empirically test our proposed theoretical considerations, we use data from the Federal Employment Agency covering the population of German employees (and provided by the research data centre (FDZ) of the Institute for Employment Research (IAB)). The exploitation of administrative in contrast to survey data like in previous studies (Bergemann and Riphahn, 2010, 2015; Cygan-Rehm, 2016; Cygan-Rehm et al., 2018; Kluve and Schmitz, 2018) has some major advantages: large sample size, mandatory notification by the employer and detailed longitudinal information on a daily basis (Müller and Strauch, 2017). Still, the fact that the data is not collected for scientific purposes implies that some information normally provided in surveys is not given. Concretely, we do not have exact information on child birth, but rely on a sophisticated imputation by Müller and Strauch (2017) which is explained in the following section. As a proof of quality, imputed birth numbers will show a movement over the year similar to official statistics. We use the population of mothers employed subject to social security contributions (SSC) before (potential) child birth given in the *Employment History (BeH, version 10.03.00)*. As we focus on the return to work, we neglect mothers previously not working, registered unemployed, in active labour market programs or receiving social assistance. Moreover, the data excludes self-employed and civil servants as they are not subject to SSC. We do not exclude mothers with more than one child but include the number of children as a control variable.

Child birth defines treatment

The *Employment History* covers all individual employment spells on a daily basis. While employers have to notify authorities at least once a year, notifications are furthermore only recorded if the employment spell ends. The imputation of the day of child birth is based on this information. Employers register when an expectant mother exits her job for the period of maternity protection and receives payment by the statutory health insurance. In general, maternity protection begins six weeks before the calculated birthday such that Müller and Strauch (2017) add six weeks of maternity protection to impute child birth. Unfortunately, the exit reason ‘receiving entitlements from statutory health insurance’ can also include long-term sickness ($\geq$ six weeks). However, misspecifications can be minimised by three restrictions. Firstly, the group of young women is more likely to have a child, but less likely to suffer from long-term sickness. Based on official birth statistics, Müller and Strauch (2017) restrict the childbearing age to 38 years for the first birth and to 40 years for subsequent births. Secondly, mothers are not

allowed to work during the 14 weeks lasting period of maternity protection. Any shorter job interruption period is more likely to specify a break due to illness. Thirdly, subsequent births are only possible after a period of about 40 weeks. As the pre-term rate was found to be 9.2 percent in 2010 (March of Dimes, PMNCH, Save the Children, and WHO, 2012), the authors limit the gap to 32 weeks. As we only observe births from mothers previously employed subject to SSC, total numbers are smaller compared to official statistics for whole Germany. However, Figure 4.10 in Appendix 4.A shows that the movement in the considered period is very similar for the official birth numbers and imputed births which highlights the quality of the imputation and the data in general. Additionally, we argue that the exclusion of a certain time window around the cut-off date should mitigate the problem. In section 4.6.1 we will show that our results are insensitive to the specific choice of the window width indicating that the imputation error is empirically a minor concern. We observe between 22,000 and 30,000 births per month while sample size in similar studies using survey data amounts to about 2,000 births for the same period.

Control and outcome variables

The *Employment History* includes a large set of other individual and job-related characteristics that we use to predict the propensity score and the outcome equation explained in Section 4.4.2. These are measured at the last employment spell, i.e., right before child birth. We include the individual age, the number of children, a binary indicator for having a migration background and the place of residence on the district level. Inserting regional fixed effects is especially important to control for the macroeconomic background or the availability of child care facilities showing large variation over German districts (compare Zimmert, 2019). With the inclusion of 402 districts the number of covariates gets large. In contrast to parametric models, our machine learning approach is able to flexibly include this large list of control variables.

Further information relate to education⁹ and occupational characteristics. We include six categories for the educational degree combined with information on the occupation (lower/middle secondary school with(out) vocational training, high school with(out) vocational training, university of applied sciences, university). Other covariates concern the requirement level (unskilled up to highly complex activities) and the occupational code both coming from the German classification of occupations KldB2010.

To control for the individual employment history the days spent in marginal, part- or full-time employment within the last five years and the working time pattern of the previous job (marginal, part- or full-time employment) are considered. The *Employment History* does not contain information on continuously measured working hours. Hence, we use the working time pattern and merge additional particularities, i.e., marginal employment as special form of a part-time contract. Marginal employment in Germany, so-called *Mini jobs*, do not exceed earnings of 450 Euro per month and are exempted from income taxation.

The gross monthly income is a generated variable that considers the duration of the employment spell. As employers only have to indicate income up to the SSC assessment ceiling,¹⁰ this variable is right censored. Special payments and misdeclarations can shift the upper ceiling such that we restrict the income range to up to 6,500 Euro per month. The type of working contract (fixed- or long-term) is also controlled for. Additionally, we also include information on the number of (female) employees coming from the *IAB Establishment History Panel (BHP, version 7516 v1)*.

⁹As the variable is characterised by a higher share of missing or inconsistent values compared to other information provided by the employer, we rely on an imputation procedure proposed by Fitzenberger, Osikominu, and Völter (2005).

¹⁰For the statutory pension and unemployment insurance, the assessment ceiling amounts to 6,050 Euro in 2015.

Table 4.4: Descriptive statistics of variables by treatment and control group

Variable	Control group		Treatment group		Standardised difference
	Mean	sd	Mean	sd	
T	0.51	0.50	0.50	0.50	2.30
Age	29.36	3.45	29.31	3.47	1.27
Number of children (including the current birth)	1.30	0.52	1.30	0.52	0.03
Migration background	0.07	0.25	0.07	0.25	0.24
Place of living (Federal state, <i>baseline</i> Schleswig-Holstein:)					
Hamburg	0.02	0.15	0.02	0.15	0.14
Lower Saxony	0.09	0.28	0.09	0.28	0.18
Bremen	0.01	0.08	0.01	0.08	0.97
North Rhine-Westphalia	0.19	0.39	0.19	0.39	0.70
Hessen	0.07	0.25	0.07	0.26	1.30
Rhineland-Palatinate	0.05	0.21	0.05	0.21	0.20
Baden-Wuerttemberg	0.13	0.34	0.13	0.34	0.30
Bavaria	0.17	0.37	0.17	0.38	1.32
Saarland	0.01	0.11	0.01	0.10	1.55
Berlin	0.05	0.21	0.04	0.21	0.82
Brandenburg	0.04	0.19	0.03	0.18	0.68
Mecklenburg-Western Pomerania	0.02	0.15	0.02	0.15	0.28
Saxony	0.07	0.25	0.07	0.25	0.01
Saxony-Anhalt	0.03	0.17	0.03	0.17	0.01
Thuringia	0.03	0.18	0.03	0.17	0.94
Education (<i>baseline</i> Lower/middle secondary school without vocational training):					
Lower/middle secondary school with vocational training	0.49	0.50	0.49	0.50	1.10
High school without vocational training	0.02	0.15	0.02	0.15	0.38
High school with vocational training	0.20	0.40	0.20	0.40	0.11
University of applied sciences	0.02	0.15	0.02	0.15	0.74
University	0.18	0.38	0.19	0.39	1.46
Days in					
marginal employment within last five years	122.56	259.24	122.60	260.33	0.01
part-time employment within last five years	327.67	497.69	331.20	500.38	0.71
full-time employment within last five years	979.62	652.75	982.60	655.75	0.45
Previous job:					
Employment pattern (<i>baseline</i> Marginal employment):					
Part-time	0.31	0.46	0.31	0.46	0.08
Full-time	0.65	0.48	0.64	0.48	0.21

Gross monthly income in Euro	2,195.92	1,241.25	2,219.98	1,245.32	1.93
Temporary contract	0.23	0.42	0.22	0.42	0.19
Requirement level (<i>baseline</i> Unskilled or semi-skilled activities):					
Specialist activities	0.66	0.47	0.66	0.47	0.39
Complex specialist activities	0.12	0.33	0.12	0.33	0.58
Highly complex activities	0.12	0.32	0.12	0.33	0.93
Occupational area (classification system Kldb2010 1-digit, <i>baseline</i> Agriculture, forestry, farming, and gardening):					
Production of raw materials and goods and manufacturing	0.06	0.24	0.06	0.24	0.71
Construction, architecture, surveying and technical building services	0.01	0.09	0.01	0.09	0.58
Natural sciences, geography and informatics	0.02	0.14	0.02	0.14	0.89
Traffic, logistics, safety and security	0.05	0.21	0.05	0.21	0.68
Commercial services, trading, sales, the hotel business and tourism	0.19	0.39	0.18	0.39	1.02
Business organization, accounting, law and administration	0.25	0.43	0.25	0.44	1.15
Health care, the social sector, teaching and education	0.38	0.49	0.38	0.49	0.26
Philology, literature, humanities, social sciences, economics, media, art, culture, and design	0.04	0.18	0.04	0.18	0.02
Establishment:					
Number of female employees	334	999	340	1,019	0.63
Total number of employees	695	3,049	712	3,173	0.54
<i>N</i>	46,263		48,212		

Notes: Instead of using federal states like presented in the table, districts are used for the prediction of the propensity score and the outcome equation. sd = standard deviation.

The standardised difference is calculated as $100 \times \left| \frac{\frac{1}{N^d} \sum_{i=1}^N D_i X_i - \frac{1}{N^0} \sum_{i=1}^N (1-D_i) X_i}{\sqrt{\frac{(sd^1)^2 + (sd^0)^2}{2}}} \right|$ where

N^d and sd^d are the sample size and the standard deviation in subsample $D = d$.

Source: Own calculations based on employee data from the *Employment History (BeH)*, establishment data from the *Establishment History Panel (BHP)*.

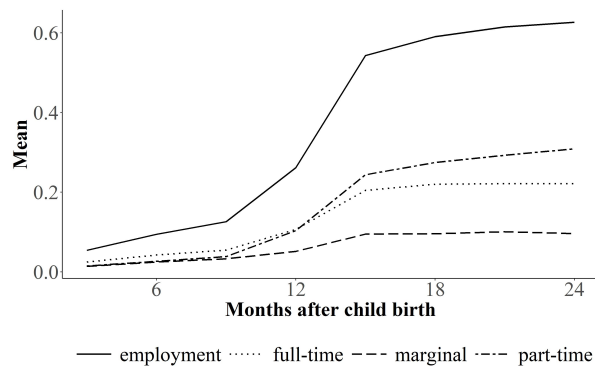
Table 4.4 shows mean values and their standard deviation of the previously described covariates by treatment and control group status. The last column gives the absolute standardised mean difference (Rosenbaum and Rubin, 1985) between these two groups and demonstrates that the sample is well balanced (values larger than 20 are typically considered as hinting at meaningful imbalances between subgroups).

Unfortunately, the data does not contain information on actual receipt of parental subsidies. So, our estimates are

intention to treat-effects. The outcome variables of interest refer to employment after child birth. Since employment spells are available until the end of 2017, we can analyse maternal labour market outcomes up to two years. We measure current employment (in full- or part-time as well as in marginal employment) as binary indicators every three months until the second birthday of the child, i.e., at eight different points in time. Figure 4.2 gives mean outcomes for the treatment group before the reform and shows that employment rates are increasing with the child's age. Before the first birthday the employment rate amounts to about 20 percent and is mainly characterised by full-time jobs. Interestingly, the employment rate sharply increases to about 60 percent until the second birthday with the highest share consisting of part-time contracts. It seems that the average pre-reform mother takes the maximum parental leave period of twelve months and returns in a part-time job.

We analyse other variables of job continuity depicted by a binary indicator for staying with the same employer and job quality in terms of earnings accumulated up to the first and second year. The second column of Table 4.5 shows that about 52 percent of treated mothers return to their previous employer before the reform while average earnings amount to 1,744 Euro in the first year and to 13,492 Euro up to the second birthday (including those with zero earnings who have not returned yet).

Figure 4.2: Outcome means of mothers in the treatment group before reform



Notes: $T = 0$ in 2014. $N = 23,993$.

Source: Own calculations based on employee data from the *Employment History (BeH)*.

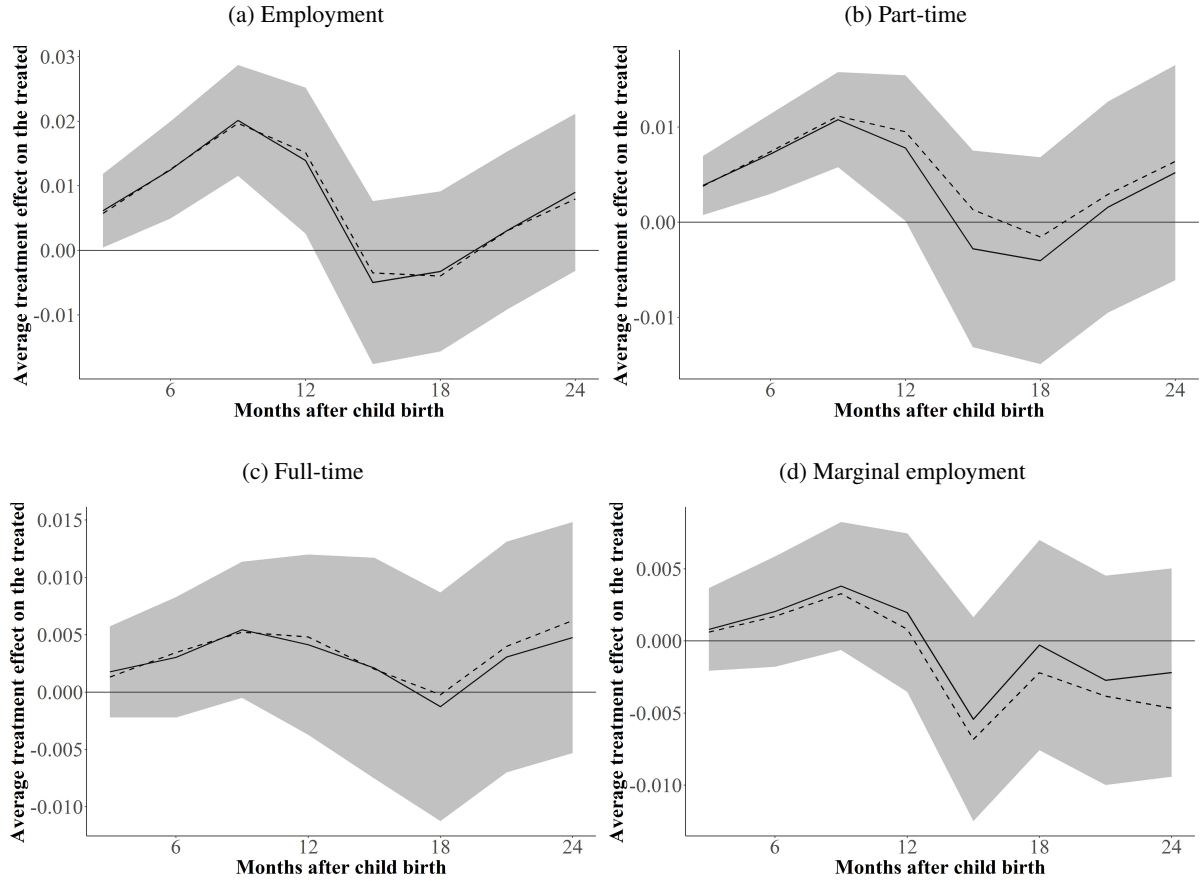
4.6 Results

4.6.1 Estimation results for ATET and sensitivity analysis

We present our main estimation results in graphs where the time in months after child birth is depicted on the horizontal axis and the ATET on the vertical axis. Apart from the machine learning augmented DiD estimator (solid line), we also show results of the unadjusted mean estimator without including any covariates (dashed line). We start by discussing the overall employment effect in Figure 4.3a. The reform gives increasing and positive employment effects up to nine months after birth amounting to statistically significant two percentage points at maximum. Although the effect size seems to be small, it accounts for about $\frac{0.020}{0.138} = 14$ percent of the pre-reform mean. Additionally, since take up only amounts to about 19 percent (Federal Statistical Office, 2019), the effect for those actually choosing the new regime should be much higher.¹¹ The effect vanishes when the child turns one year

¹¹We are cautious when interpreting DiD results in a Wald estimator kind of manner. For example Chaisemartin and D'Haultfoeuille (2018) show that such an argumentation may only be valid under very restrictive assumptions.

Figure 4.3: Baseline estimation results



Notes: Treatment status is measured six weeks around cut-off date excluding the two weeks on each side of the cut-off date. $T = 1$ in 2015, $T = 0$ in 2014. $N = 94,475$. Grey shaded areas depict pointwise 95 % confidence intervals. AIPW DiD is solid line. Unadjusted DiD is dashed line.

Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

old indicating that the first birthday remains a reference point for the majority of previously employed mothers. 18 months after child birth the ATET slightly increases again, but does not reach significance on conventional levels. How is this overall positive employment effect in the first year distributed over different employment patterns? Figures 4.3b to 4.3d show that part-time employment mainly drives this finding. At maximum the part-time effect equals about one percentage point. The reform's impact on full-time employment is insignificant and close to zero. Marginal employment as a form of part-time employment is also not significantly affected.

These empirical findings are in line with the mechanisms described in Section 4.3 predicting a decrease of the reservation income for a part-time job relative to a full-time job. Moreover, we find that the theoretically ambiguous effect on the extensive employment margin is empirically positive. Interestingly, the seemingly positive effect on the attractiveness of part-time employment is not associated by a drop in full-time employment. Instead, it is reflected by an increase in overall employment.

We do not identify any persistent employment patterns, i.e., the distribution into full-, part-time or marginal employment is not affected until the child's second birthday. This might suggest that those mothers incentivised to return before the child's first birthday under a part-time contract would have also returned in part-time under the pre-reform regime after the child's first birthday. In turn, mothers who would have returned in full-time employ-

Table 4.5: ATETs for job continuity and accumulated earnings

Outcome	$D = 1, T = 0$		Unadjusted DiD		AIPW	
	Mean	sd	ATET	se	ATET	se
<i>Panel A: Baseline</i>						
Same employer	0.522	0.500	0.008	0.006	0.006	0.006
Accumulated earnings						
1st year	1,743.86	5,370.63	282.38***	72.23	272.66***	71.40
2nd year	13,492.12	16,178.52	430.09**	211.53	314.32*	191.22
Notes: Treatment status is measured six weeks around the cut-off date excluding the two weeks on each side of the cut-off date. $T = 1$ in 2015, $T = 0$ in 2014. $N = 94,475$.						
<i>Panel B: Placebo</i>						
Same employer	0.561	0.496	-0.016**	0.007	-0.009	0.006
Accumulated earnings						
1st year	1,656.53	5,343.75	-76.78	73.17	-73.00	71.84
2nd year	12,556.42	15,652.03	-91.02	213.05	44.81	192.52
Notes: Treatment status is measured six weeks around the cut-off date excluding the two weeks on each side of the cut-off date. $T = 1$ in 2014, $T = 0$ in 2013. $N = 89,374$.						

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

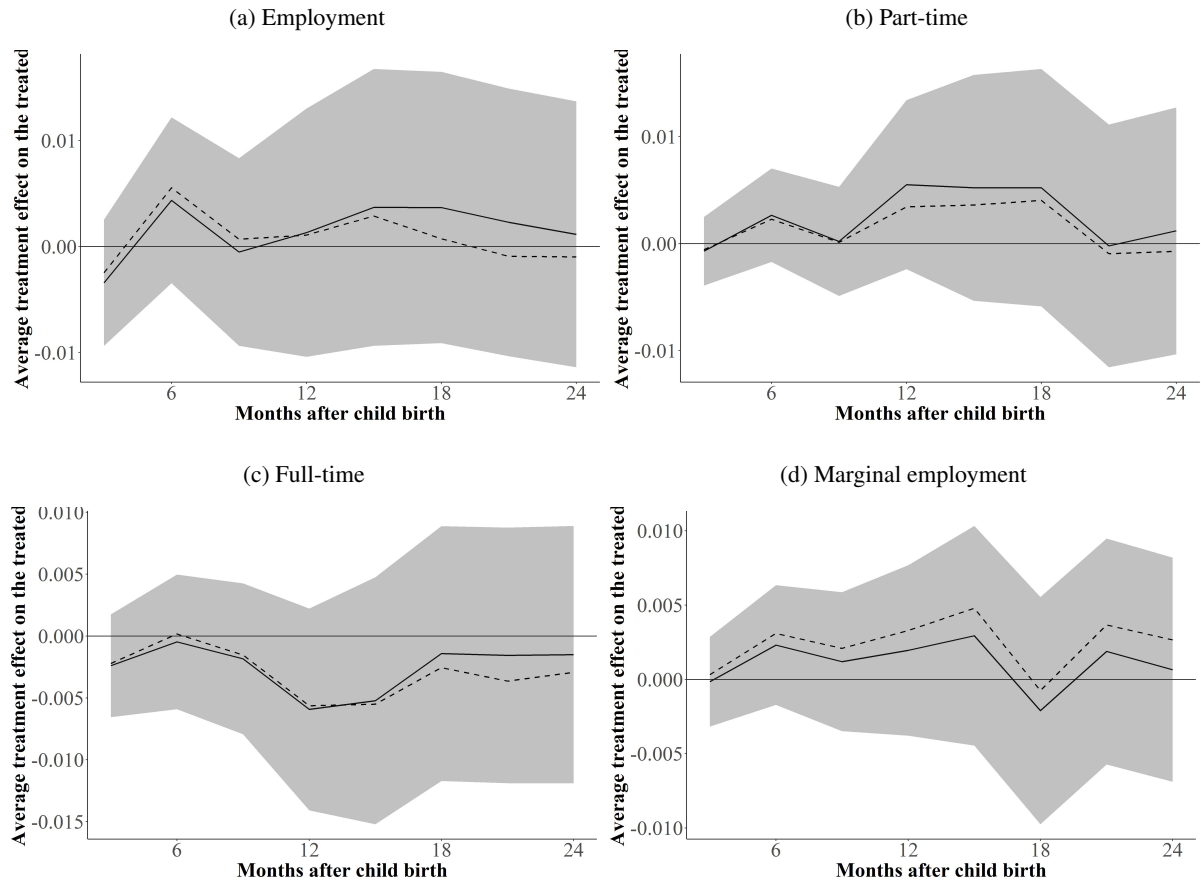
ment under the pre-reform regulations might not be willing to accept reduced working hours supplemented by parental benefits before the child's first birthday, as they might fear to get stuck in a part-time contract.

Panel A of Table 4.5 shows that the overall positive employment effects reflect in higher accumulated earnings within the first year (about 273 Euro) and less precisely estimated within the second year (about 314 Euro). Furthermore, the shorter employment break does not affect the probability to return to the previous employer. The validity of our findings is supported by the fact that the unconditional mean differences (dashed lines) are very close to our estimation results using a rich set of covariates. If our setting would be sensitive to confounding with respect to one of the observed covariates, we would expect different results for the simple differences in means estimator and our procedure. Moreover, we examine the plausibility of the common trend assumption between treatment and control group in the absence of the reform by postponing the reform year to 2014. While this kind of check cannot directly test the assumption, Figure 4.4 hints at similar employment trends before the reform reflected in ATETs that are precisely measured at around zero. Earnings and job continuity are as well not affected (compare Panel B of Table 4.5).

A second potential concern in our empirical strategy might be the arbitrary definition of the sampling periods around the cut-off date and the imputation error of the child's birthday (see Section 4.5). We check the sensitivity against these two threats by estimating the effects for different populations defined around the cut-off. The results of this analysis do not indicate crucial differences compared to our baseline results and are relegated to Appendix 4.A.

As a possible channel driving the employment outcomes, we look at the effect on subsequent births. E.g., Cygan-Rehm (2016) shows that the parental leave reform of 2007 incentivised mothers to postpone a subsequent pregnancy. Figure 4.5 does not indicate any effect on childbearing within the two year-horizon. However, this finding has to be interpreted with caution as we only observe women with subsequent births who have been employed in the meanwhile.

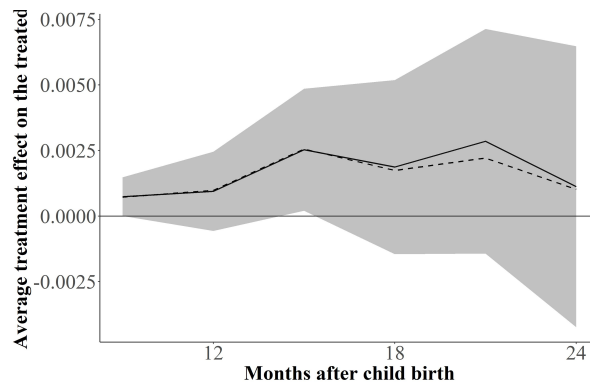
Figure 4.4: Estimation results of placebo reform



Notes: Treatment status is measured six weeks around cut-off date excluding the two weeks on each side of the cut-off date. $T = 1$ in 2014, $T = 0$ in 2013. $N = 89,374$. Grey shaded areas depict pointwise 95 % confidence intervals. AIPW DiD is solid line. Unadjusted DiD is dashed line.

Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

Figure 4.5: Estimation results for subsequent birth within next 24 months



Notes: Treatment status is measured six weeks around cut-off date excluding the two weeks on each side of the cut-off date. $T = 1$ in 2015, $T = 0$ in 2014. $N = 94,475$. Grey shaded areas depict pointwise 95 % confidence intervals. AIPW DiD is solid line. Unadjusted DiD is dashed line.

Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

4.6.2 Estimation results for conditional effects

To better understand the channels of the reform, we investigate how the treatment effects vary over different pre-specified subgroups. We investigate different income groups as well as heterogeneities concerning the prior working time pattern and the place of living (East and West Germany). The latter might yield interesting results as mothers growing up in the former GDR could have different attitudes towards maternal employment. Hence, in our setting Z contains dummies for the middle and high income groups, whether the mother worked full-time prior to child birth and a dummy for West Germany. In particular, we estimate the following specification using OLS regression

$$\tilde{y} = \beta_0 + \beta z + \varepsilon$$

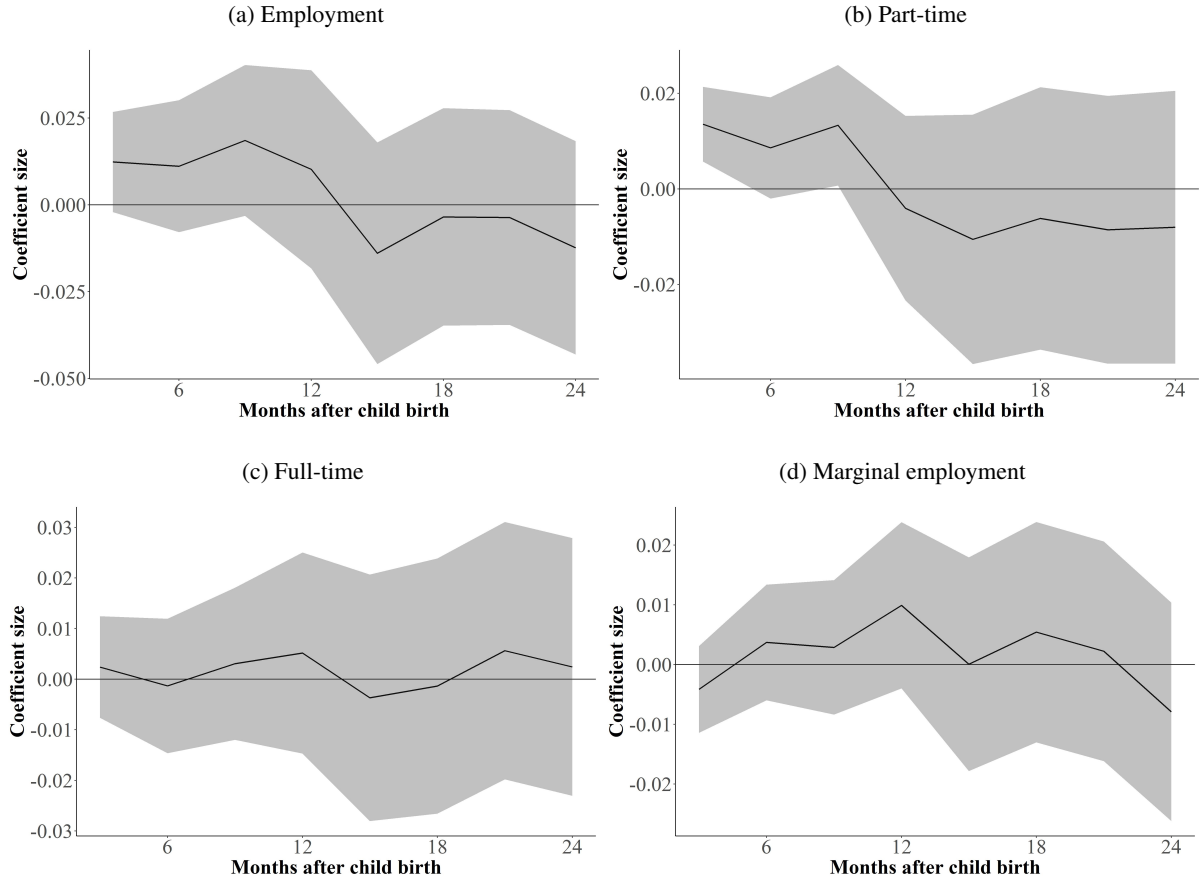
where $\tilde{y}$ represents the sample analogue of $\frac{1}{p(Z)} \frac{T-p_T}{p_T(1-p_T)} \frac{D-p(X)}{1-p(X)} (Y - \gamma(X, T))$ with first stages estimated as for the average effects and $p(Z)$ by Logit regression. The resulting OLS coefficients give the effect variation for the different subgroups. They are depicted on the vertical axis for the eight different periods in Figures 4.6 to 4.9. As discussed in Section 4.4.3 we report the usual OLS standard errors.

Figures 4.6a and 4.7a show that the effect size for employment does not differ with respect to income (low income is chosen as reference group). When we further differentiate the employment effects in part-time and full-time as well as marginal employment, Figure 4.6b reveals that the positive part-time effects are mainly driven by middle income earners. E.g., after nine months the part-time effect for middle income earners is about 1.3 percentage points higher compared with low income mothers.¹² For high-income mothers, the effects do not significantly differ from those with lower income (see Figure 4.7). Hence, we conclude that high-income and potentially more career-oriented mothers prefer not to return in part-time employment despite the simultaneous provision of parental subsidies. One potential explanation is that they fear the implications of reducing their working time. Mothers with middle income may be more willing to accept a part-time job because the future potential income loss after expiration of parental benefits is less severe. This argumentation is strongly supported when examining the subgroup of previously full-time employed women. Figure 4.8b demonstrates that they have a lower probability (-1.3 to -3.7 percentage points) to return in part-time employment in the first year after child birth. As their opportunity costs of taking up a part-time job are higher, they are characterised by a weaker response to the reform. We explain the similar effect size of low- and high-income mothers with the amount of the part-time subsidy. Low-income mothers are more likely to receive the minimum amount of 150 Euro such that the incentive to return before the child's first birthday is in general less pronounced. The effects for full-time and marginal employment do not significantly differ over subgroups.

Interestingly, the effect size does not significantly vary with the place of living, i.e., East and West Germany (see Figure 4.9). Bergemann and Riphahn (2015) and Kluge and Schmitz (2014) provide suggestive evidence that the parental leave reform of 2007 defines a social norm to return to work after the child's first birthday. In this regard, the new policy has the potential to further increase cultural acceptance for those mothers preferring a return even before the child turns one year old (Zoch and Hondralis, 2017). As a legacy of the German Democratic Republic, societal acceptance of maternal employment and the reliance on external child care are on general on a higher level in East Germany (e.g., Hanel and Riphahn, 2012). Consequently, a shift of social norms becomes more

¹²Table 4.7 in the appendix shows that these effects reflect in higher accumulated earnings within the first year. In the main analysis we concentrate on the CATETs for the different working time patterns. See Table 4.7 for a detailed presentation.

Figure 4.6: Estimation results for middle income group



Notes: Low (middle, high) income corresponds to 1st (2nd, 3rd) tercile of gross previous monthly income. $N_{low_income} = 31,170$, $N_{middle_income} = 32,139$ and $N_{high_income} = 31,166$. The coefficients give the effect variation for the respective subgroup. The reference group is low income. Grey shaded areas depict pointwise 95 % confidence intervals.

Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

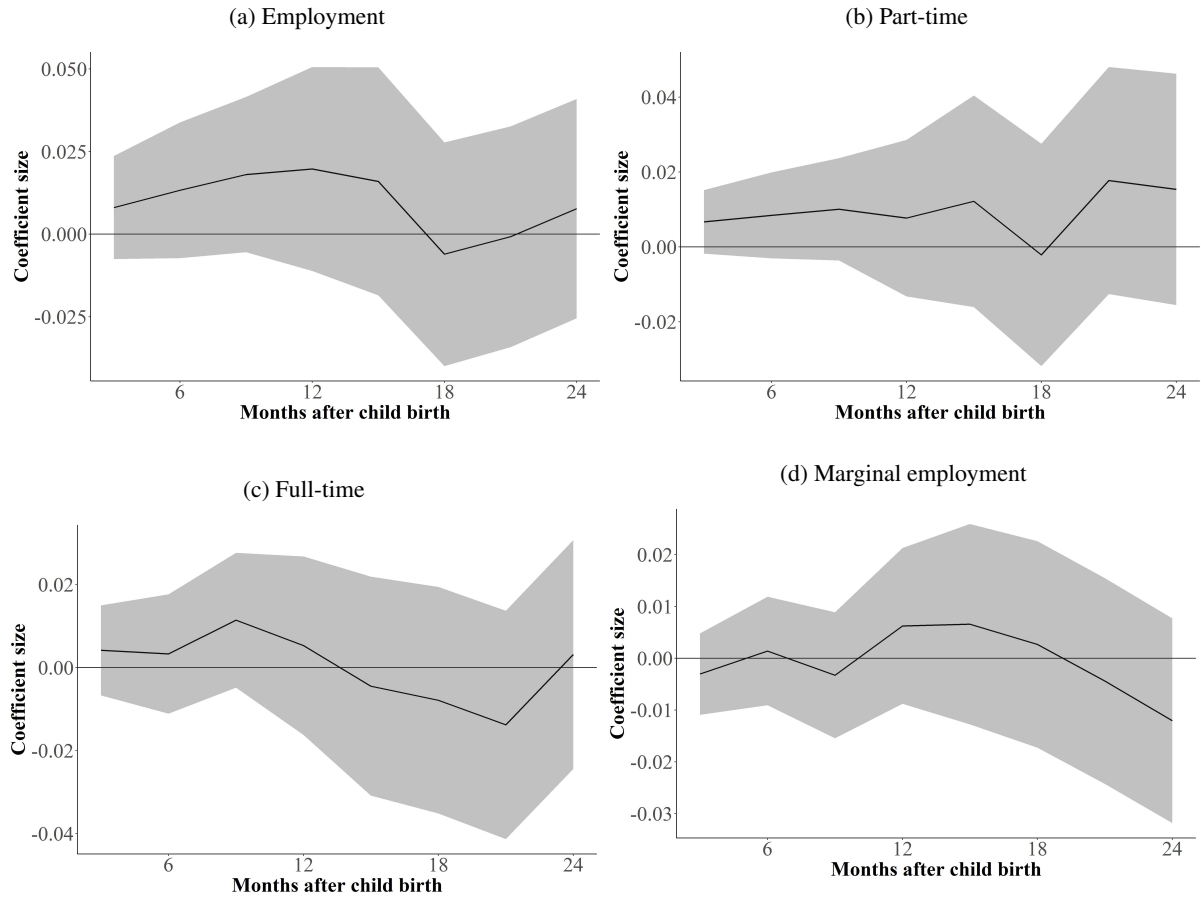
likely in the West German society where traditional approaches of the household's division of labour predominate. However, we do not observe statistically significant differences until the child's first birthday for mothers living in West Germany compared to East Germany (Figure 4.9a and 4.9b). Thus, we do not find suggestive evidence for a shift of social norms induced by the reform. As the administrative character of the data does not allow to follow up on this suggestion, we interpret it with caution.

Moreover, the financial incentive for part-time work, does not affect maternal employment outcomes after the child's first birthday for almost all subgroups. However, prior full-time working mothers have a lower probability for working part-time of up to 4.1 percentage points shortly before the child gets two years old. Hence, the reform may foster the path dependency of working part-time, at least for the short period of two years.

4.7 Conclusion

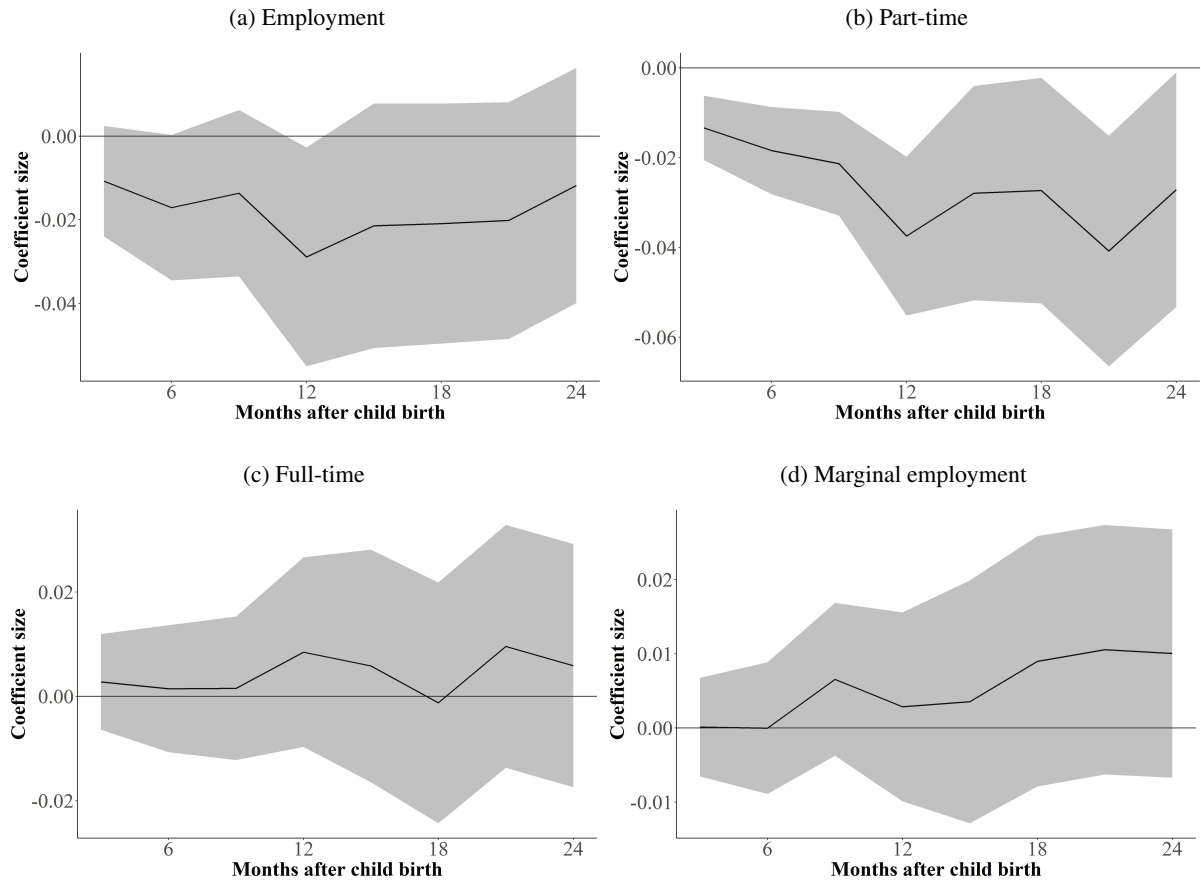
We analyse a German parental leave reform promoting a faster return to part-time work after child birth while receiving parental benefits. Our results indicate that part-time subsidies are indeed a successful tool incentivising mothers to return to the labour market earlier. Although shorter employment interruptions can improve career

Figure 4.7: Estimation results for high income group



Notes: Low (middle, high) income corresponds to 1st (2nd, 3rd) tercile of gross previous monthly income. $N_{low_income} = 31,170$, $N_{middle_income} = 32,139$ and $N_{high_income} = 31,166$. The coefficients give the effect variation for the respective subgroup. The reference group is low income. Grey shaded areas depict pointwise 95 % confidence intervals.
Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

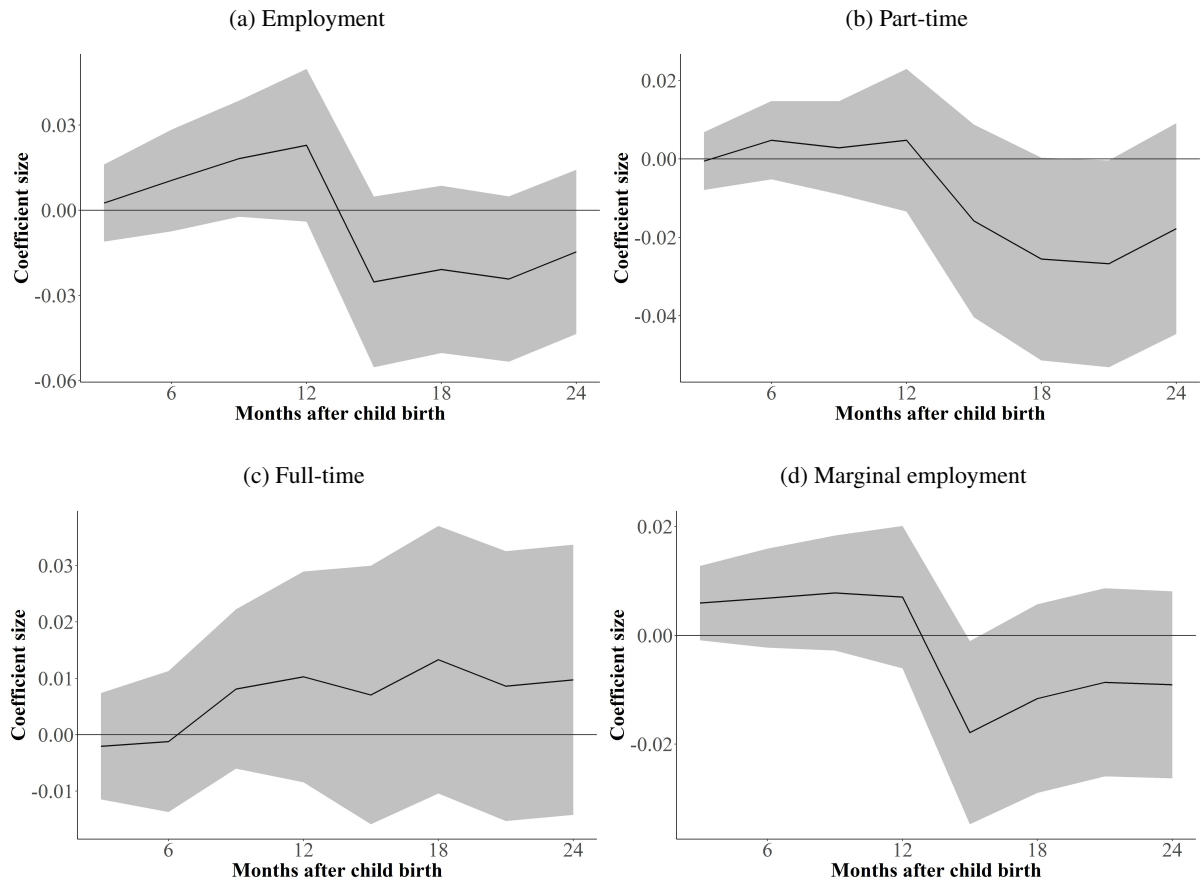
Figure 4.8: Estimation results for prior full-time



Notes: $N_{part_time} = 29,214$ and $N_{full_time} = 60,913$. Coefficient gives effect variation for respective subgroup. The reference group is prior part-time and marginal employment. Grey shaded areas depict pointwise 95 % confidence intervals.

Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

Figure 4.9: Estimation results for West Germany



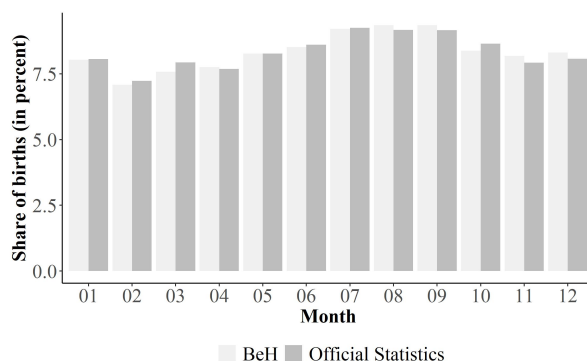
Notes: $N_{east} = 21,967$ and $N_{west} = 72,508$. The coefficients give the effect variation for the respective subgroup. The reference group is East Germany. Grey shaded areas depict pointwise 95 % confidence intervals.

Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

prospects, the policy could have pushed mothers to reduce working hours instead of returning to a full-time job when the child is older. Our results from semiparametric DiD estimation in combination with machine learning algorithms are robust towards a violation of the main identifying assumption and do not provide evidence for such a downside. The reform rather yields additional part-time effects before the child's birthday of up to one percentage point driven by mothers who would have also returned to a part-time job in absence of the reform. Heterogeneous effects support this argumentation. We find that mothers with lower opportunity costs of accepting a part-time job (i.e., those with middle previous income and prior part-time workers) show a stronger response to the reform. Previous regulations established the child's first birthday as a reference point for the parental leave duration reinforced by the legal claim for a child care slot introduced in 2013. Insignificant differences for West and East Germany found in this study do not hint at the potential to further change societal expectations when to return to work. Our findings have to be interpreted in the context of a labour market in which working mothers with children younger than one year old are a minority and the German tax and health insurance systems additionally promote an inequalitarian household division of paid working hours. This might also explain why the introduction of the new parental benefit system does not lead to better employment prospects in terms of working hours for those women deciding for an early return to work. To further support employees with a temporary preference for a working hour reduction, the German government recently enforced a legal claim to return to a full-time job which might especially be a good instrument for mothers after parental leave. Hence, it would be interesting to learn about long-term effects of the parental leave reform also in combination with the right to return in full-time.

4.A Figures and tables

Figure 4.10: Imputed and official birth numbers in 2015



Source: Own calculations from *BeH* and the German Federal Statistical Office (2015a).

Table 4.6: Large and small bandwidth results for job continuity and accumulated earnings

Outcome	$D = 1, T = 0$		Unadjusted DiD		AIPW	
	Mean	sd	ATET	se	ATET	se
<i>Panel C: Small bandwidth</i>						
Same employer	0.547	0.498	0.015*	0.009	0.016*	0.009
Accumulated earnings						
1st year	1,613.31	5,339.65	237.27**	99.82	242.32**	98.37
2nd year	12,909.01	15,939.58	171.54	294.15	189.89	264.31
Notes: Treatment status is measured four weeks around the cut-off date excluding the two weeks on each side of the cut-off date. $T = 1$ in 2015, $T = 0$ in 2014. $N = 48,544$.						
<i>Panel D: Large bandwidth</i>						
Same employer	0.553	0.497	0.005	0.006	-3.9E-04	0.006
Accumulated earnings						
1st year	1,731.44	5,483.04	262.78***	73.40	268.16***	73.73
2nd year	13,257.52	16,289.88	620.13***	214.52	337.54*	194.84
Notes: Treatment status is measured eight weeks around the cut-off date excluding the four weeks on each side of the cut-off date. $T = 1$ in 2015, $T = 0$ in 2014. $N = 94,493$.						

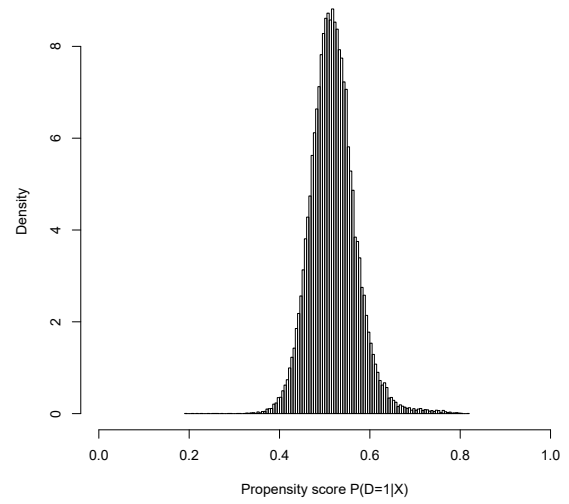
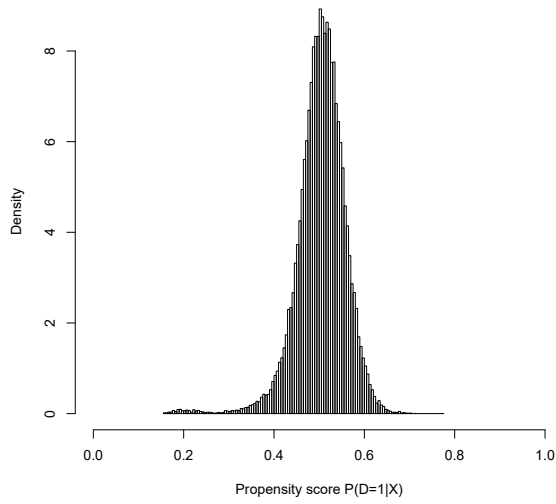
Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

Figure 4.11: Propensity scores by treatment status

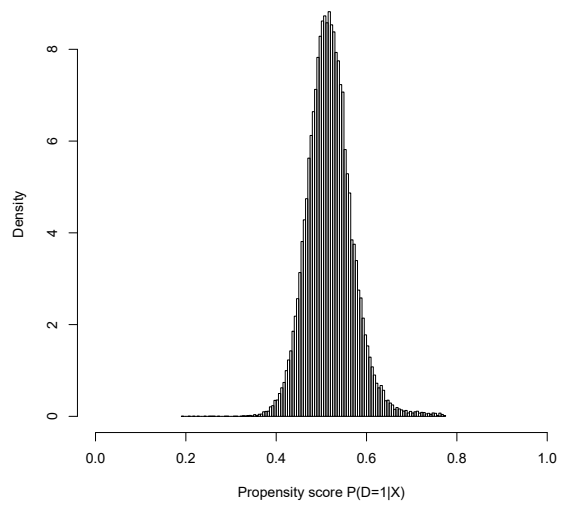
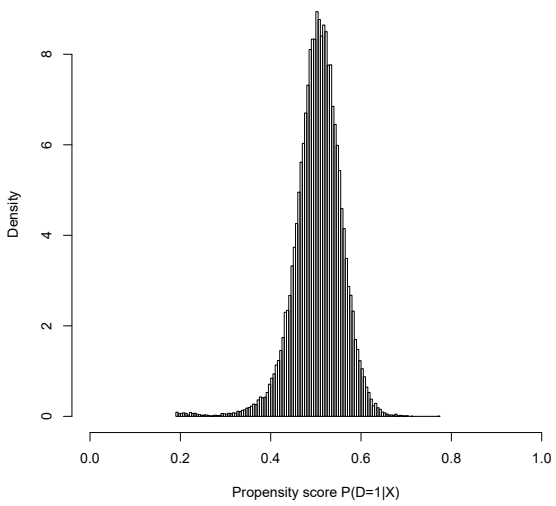
(a) $D = 0$ before trimming

(b) $D = 1$ before trimming



(c) $D = 0$ after trimming

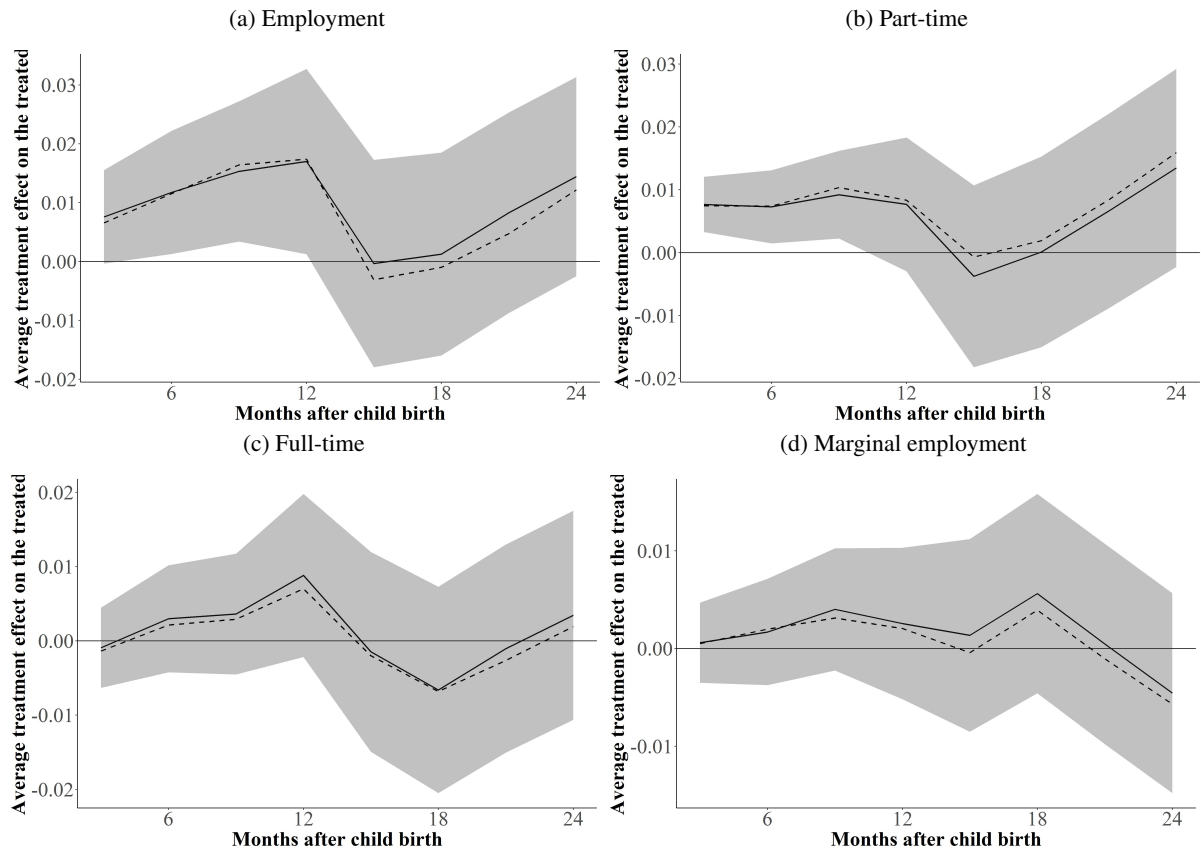
(d) $D = 1$ after trimming



Notes: $N_{D=0} = 46,263$ and $N_{D=1} = 48,212$ before trimming; $N_{D=0} = 46,191$ and $N_{D=1} = 48,184$ after trimming.

Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

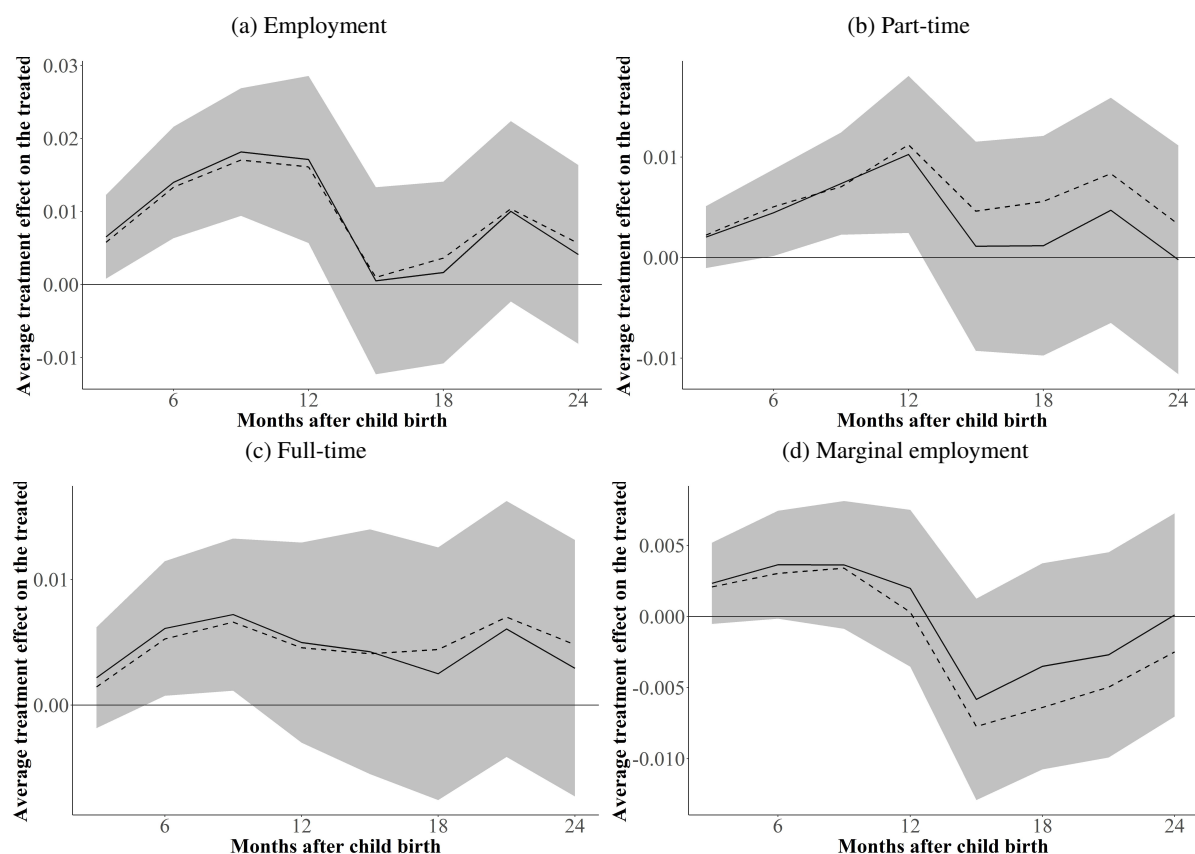
Figure 4.12: Estimation results with small bandwidth



Notes: Treatment status is measured four weeks around cut-off date excluding the two weeks on each side of the cut-off date. $T = 1$ in 2015, $T = 0$ in 2014. $N = 48,544$. Grey shaded areas depict pointwise 95 % confidence intervals. AIPW DiD is solid line. Unadjusted DiD is dashed line.

Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

Figure 4.13: Estimation results with large bandwidth



Notes: Treatment status is measured eight weeks around cut-off date excluding the four weeks on each side of the cut-off date. $T = 1$ in 2015, $T = 0$ in 2014. $N = 94,493$. Grey shaded areas depict pointwise 95 % confidence intervals. AIPW DiD is solid line. Unadjusted DiD is dashed line.

Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

Table 4.7: Conditional effects for job continuity and accumulated earnings

Outcome	$D = 1, T = 0$		AIPW	
	Mean	sd	coefficient	se
<i>Panel A: Middle income</i>				
Same employer	0.426	0.495	0.021	0.016
Accumulated earnings				
1st year	1,049.87	2,942.71	304.41*	179.73
2nd year	6,981.35	8,791.76	244.28	481.62
$N = 32,139$. Middle income corresponds to 2nd tercile of gross previous monthly income. Reference group is low income.				
<i>Panel B: High income</i>				
Same employer	0.563	0.496	0.022	0.017
Accumulated earnings				
1st year	1,382.94	4,359.20	474.98**	194.33
2nd year	11,984.46	12,674.35	337.69	520.74
$N = 31,166$. High income corresponds to 3rd tercile of gross previous monthly income. Reference group is low income.				
<i>Panel C: Prior full-time job</i>				
Same employer	0.563	0.496	-0.012	0.014
Accumulated earnings				
1st year	1,933.89	6,089.10	-270.84*	164.59
2nd year	14,687.05	17,480.42	-681.76	441.07
$N = 60,913$. Reference group is prior part-time or marginal employment.				
<i>Panel D: West Germany</i>				
Same employer	0.534	0.499	-0.031**	0.015
Accumulated earnings				
1st year	1,728.85	5,622.98	129.82	169.07
2nd year	11,834.38	16,158.99	-331.09	453.07
$N = 72,508$. Reference group is East Germany.				

Notes: Treatment status is measured six weeks around cut-off date excluding the two weeks on each side of the cut-off date. $T = 1$ in 2015, $T = 0$ in 2014. The coefficients give the effect variation for the respective subgroup. * $p < 0.1$, ** $p < 0.05$.

Source: Own calculations based on employee data from the *Employment History (BeH)* and establishment data from the *Establishment History Panel (BHP)*.

4.B Technical details for the proposed estimators

Identification of ATET

Under the condition that X and D are jointly independent of T , by the Law of Total Expectations we can write

$$\begin{aligned} & \mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{p(X)(1 - p(X))} Y \middle| X \right] \\ &= (\mathbb{E}[Y|X, D = 1, T = 1] - \mathbb{E}[Y|X, D = 1, T = 0]) \\ & \quad - (\mathbb{E}[Y|X, D = 0, T = 1] - \mathbb{E}[Y|X, D = 0, T = 0]) \end{aligned}$$

which follows from the fact that $P(T = t|X) = p_T$.

Additionally,

$$\begin{aligned} & \mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{p(X)(1 - p(X))} \gamma(X, T) \middle| X \right] \\ &= \mathbb{E} \left[\frac{D - p(X)}{p(X)(1 - p(X))} \gamma(X, 1) \middle| X, T = 1 \right] - \mathbb{E} \left[\frac{D - p(X)}{p(X)(1 - p(X))} \gamma(X, 0) \middle| X, T = 0 \right] \\ &= \mathbb{E}[D - p(X)|X] \frac{\gamma(X, 1) - \gamma(X, 0)}{p(X)(1 - p(X))} \\ &= 0. \end{aligned}$$

Analogous to the fundamental result of Heckman et al. (1997) we have

$$\begin{aligned} \mathbb{E}[Y^1(1) - Y^0(1)|X, D = 1, T = 1] &= \mathbb{E}[Y|X, D = 1, T = 1] - \mathbb{E}[Y^0(1)|X, D = 1, T = 1] \\ &= \mathbb{E}[Y|X, D = 1, T = 1] - \mathbb{E}[Y^0(1)|X, D = 0, T = 1] \\ & \quad + \mathbb{E}[Y^0(0)|X, D = 0, T = 0] - \mathbb{E}[Y^0(0)|X, D = 1, T = 0] \\ &= (\mathbb{E}[Y|X, D = 1, T = 1] - \mathbb{E}[Y|X, D = 1, T = 0]) \\ & \quad - (\mathbb{E}[Y|X, D = 0, T = 1] - \mathbb{E}[Y|X, D = 0, T = 0]) \end{aligned}$$

which follows by the Observational Rule, the No Anticipation and the Common Trend assumptions. Therefore,

$$\mathbb{E}[Y^1(1) - Y^0(1)|X, D = 1, T = 1] = \mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{p(X)(1 - p(X))} (Y - \gamma(X, T)) \middle| X \right].$$

Denote the conditional density function of a random variable $A = a$ given another random variable $B = b$ as $f_{A|B}(a|b)$. Further, denote the density function of $A = a$ as $f_A(a)$. Then using the previous finding and by the Law of Iterated Expectations similar to Abadie (2005), it follows that

$$\begin{aligned} \text{ATET} &= \mathbb{E}[Y^1(1) - Y^0(1)|D = 1, T = 1] \\ &= \int \mathbb{E}[Y^1(1) - Y^0(1)|X = x, D = 1, T = 1] f_{X|D, T}(x|1, 1) dx \\ &= \int \mathbb{E}[Y^1(1) - Y^0(1)|X = x, D = 1, T = 1] \frac{p(x)}{p_D} f_X(x) dx \\ &= \frac{1}{p_D} \mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{1 - p(X)} (Y - \gamma(X, T)) \right]. \end{aligned}$$

The existence of the expectation is guaranteed by the Common Support condition.

No effect of first stage estimators

Let $W = (Y, D, T, X)$ and the score

$$\psi(W, p, \gamma; \theta) = \frac{1}{p_D} \frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{1 - p(X)} (Y - \gamma(X, T)) - \theta.$$

To rule out the possibility that estimation of $p(X)$ and $\gamma(X, T)$ may have an first order effect on the asymptotic behaviour of the overall DiD estimator, the difference between the score with estimated nuisance parameters $\psi(W, \hat{p}, \hat{\gamma}; \theta)$ and hypothetically known nuisance parameters $\psi(W, p, \gamma; \theta)$ should be asymptotically negligible. The difference has therefore to converge at a rate (slightly) faster than $\sqrt{N}$. In order to guarantee that, some sophisticated conditions have to be satisfied (Chang, 2020; Chernozhukov et al., 2018a; Zimmert, 2020). To illustrate the basic mechanisms, we provide a heuristic argument here. Notice that

$$\begin{aligned} & \psi(W, \hat{p}, \hat{\gamma}; \theta) - \psi(W, p, \gamma; \theta) \\ &= \frac{T - p_T}{p_T(1 - p_T)} \underbrace{\left(\left(\frac{1 - D}{p_D} \frac{p(X)}{1 - p(X)} - \frac{D}{p_D} \right) (\hat{\gamma}(X, T) - \gamma(X, T)) \right)}_i \\ &+ \underbrace{\frac{1 - D}{p_D} \left(\frac{p(X)}{1 - p(X)} - \frac{\hat{p}(X)}{1 - \hat{p}(X)} \right) (Y - \gamma(X, T))}_{ii} \\ &+ \underbrace{\frac{1 - D}{p_D} \frac{1}{(1 - p(X))(1 - \hat{p}(X))} (\hat{p}(X) - p(X)) (\hat{\gamma}(X, T) - \gamma(X, T))}_{iii}. \end{aligned}$$

In expectations the difference vanishes relatively fast. For terms i and ii by the Law of Iterated Expectations we obtain $\mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} i \right] = 0$ and $\mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} ii \right] = 0$. Term iii mirrors the rate double robustness feature of the estimator. In particular, the error terms of the nuisance parameters enter the expression multiplicatively. The expectation of iii can be bounded if the conditional and unconditional probabilities p_D , p_T , $1 - p_T$, $1 - p(X)$ and $1 - \hat{p}(X)$ are bounded away from zero. Then using Hölder's inequality we obtain

$$\mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} iii \right] \leq C \sqrt{\mathbb{E} [(\hat{p}(X) - p(X))^2]} \times \sqrt{\mathbb{E} [(\hat{\gamma}(X, T) - \gamma(X, T))^2]}$$

where C is a generic constant. Hence, in expectation first stage estimation has no effect on the overall asymptotic behaviour if the joint convergence rate in L_2 of the two nuisance parameters is faster than $\sqrt{N}$.

Identification of CATET

From the identification of the ATET we know that

$$\mathbb{E} [Y^1(1) - Y^0(1) | X, D = 1, T = 1] = \mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{p(X)(1 - p(X))} (Y - \gamma(X, T)) \middle| X \right].$$

For the CATET we therefore obtain

$$\begin{aligned}
& \mathbb{E} [Y^1(1) - Y^0(1) | Z = z, D = 1, T = 1] \\
&= \mathbb{E} [\mathbb{E} [Y^1(1) - Y^0(1) | X, D = 1, T = 1] | Z = z, D = 1, T = 1] \\
&= \mathbb{E} \left[\mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{p(X)(1 - p(X))} (Y - \gamma(X, T)) \middle| X \right] \middle| Z = z, D = 1, T = 1 \right] \\
&= \int_X \mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{p(X)(1 - p(X))} (Y - \gamma(X, T)) \middle| X = x \right] f_{X|Z,D,T}(x|z, 1, 1) dx \\
&= \int_X \mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{p(X)(1 - p(X))} (Y - \gamma(X, T)) \middle| X = x \right] \frac{f_{X,Z|D,T}(x, z|1, 1)}{f_{Z|D,T}(z|1, 1)} dx \\
&= \int_X \mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{p(X)(1 - p(X))} (Y - \gamma(X, T)) \middle| X = x \right] \frac{f_{D,T|X,Z}(1, 1|x, z) f_{X,Z}(x, z)}{f_{D,T|Z}(1, 1|z) f_Z(z)} dx \\
&= \int_X \mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{p(X)(1 - p(X))} (Y - \gamma(X, T)) \middle| X = x \right] \frac{p(x)}{p(z)} f_{X|Z}(x|z) dx \\
&= \mathbb{E} \left[\mathbb{E} \left[\frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{1 - p(X)} (Y - \gamma(X, T)) \middle| X \right] \frac{1}{p(Z)} \middle| Z = z \right] \\
&= \mathbb{E} \left[\frac{1}{p(Z)} \frac{T - p_T}{p_T(1 - p_T)} \frac{D - p(X)}{1 - p(X)} (Y - \gamma(X, T)) \middle| Z = z \right]
\end{aligned}$$

which exists under the additional assumption that $p(Z) > 0$.

Bibliography

- Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies* 72 (1), 1–19.
- Abadie, A. & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* 74 (1), 235–267.
- Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics* 21 (4), 489–519.
- Abrevaya, J., Hsu, Y.-C., & Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business and Economic Statistics* 33 (4), 485–505.
- Almond, D., Chay, K., & Lee, D. (2005). The costs of low birth weight. *The Quarterly Journal of Economics* 120 (3), 1031–1083.
- Angrist, J. D. & Acemoglu, D. (2001). Consequences of employment protection? The case of the Americans with Disabilities Act. *Journal of Political Economy* 109 (5), 915–957.
- Athey, S. & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *The Journal of Economic Perspectives* 31 (2), 3–32.
- (2019). *Machine learning methods economists should know about*. Version 1. arXiv: [1903.10075v1](https://arxiv.org/abs/1903.10075v1).
- 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.
- Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics* 47 (2), 1148–1178.
- Baker, M. & Milligan, K. (2008). How does job-protected maternity leave affect mothers’ employment? *Journal of Labor Economics* 26 (4), 655–691.
- Baum, C. L. (2003). The effect of state maternity leave legislation and the 1993 Family and Medical Leave Act on employment and wages. *Labour Economics* 10 (5), 573–596.
- Behncke, S., Frölich, M., & Lechner, M. (2010a). A caseworker like me – does the similarity between the unemployed and their caseworkers increase job placements? *The Economic Journal* 120 (549), 1430–1459.
- (2010b). Unemployed and their caseworkers: Should they be friends or foes? *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173 (1), 67–92.
- Belloni, A. & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* 19 (2), 521–547.
- Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80 (6), 2369–2429.
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81 (2), 608–650.

- Bergemann, A. & Riphahn, R. T. (2010). Female labour supply and parental leave benefits – the causal effect of paying higher transfers for a shorter period of time. *Applied Economics Letters* 18(1), 17–20.
- (2015). Maternal employment effects of paid parental leave. *IZA Discussion Papers* 9073.
- Bühlmann, P. & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. 1st ed. Springer series in Statistics. Springer.
- Bickel, P. J. & Ritov, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, 381–393.
- Bickel, P. J., Klaassen, C. A., Ritov, Y., & Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. 1st ed. Springer.
- Buckles, K. S. & Hungerman, D. M. (2013). Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics* 95(3), 711–724.
- 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.
- Callaway, B. & Sant’Anna, P. H. (2020). *Difference-in-differences with multiple time periods*. Version 3. arXiv: [arXiv:1803.09015v3](https://arxiv.org/abs/1803.09015v3).
- Card, D. (1990). The impact of the mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review* 43(2), 245–247.
- Card, D. & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review* 84(4), 772–793.
- Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* 155(2), 138–154.
- Chaisemartin, C. de & D’Haultfoeuille, X. (2018). Fuzzy differences-in-differences. *Review of Economic Studies* 85(2), 999–1028.
- Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* 34(3), 305–334.
- (1992). Efficiency bounds for semiparametric regression. *Econometrica* 60(3), 567–596.
- Chang, N.-C. (2020). Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal* 23(2), 177–191.
- Chen, X., Hong, H., & Tarozi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics* 36(2), 808–843.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1), C1–C68.
- Chernozhukov, V., Fernández-Val, I., & Luo, Y. (2018b). The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica* 86(6), 1911–1938.
- Cygan-Rehm, K. (2016). Parental leave benefit and differential fertility responses: Evidence from a German reform. *Journal of Population Economics* 29(1), 73–103.
- Cygan-Rehm, K., Kühnle, D., & Riphahn, R. T. (2018). Paid parental leave and families’ living arrangements. *Labour Economics* 53, 182–197.
- da Veiga, P. V. & Wilder, R. P. (2008). Maternal smoking during pregnancy and birthweight: A propensity score matching approach. *Maternal and Child Health Journal* 12(2), 194–203.

- Dahl, G. B., Løken, K. V., Mogstad, M., & Salvanes, K. V. (2016). What is the case for paid maternity leave? *Review of Economics and Statistics* 98 (4), 655–670.
- DIW Berlin/SOEP (2014). *SOEP-RS FiD 2013 - Erhebungsinstrumente 2013 von 'Familien in Deutschland': Elternfragebogen 1: "Für alle Kinder der Jahrgänge 2012/2013"*. Deutsches Institut für Wirtschaftsforschung (DIW), Berlin.
- Eissa, N. & Liebman, J. B. (1996). Labor supply response to the earned income tax credit. *The Quarterly Journal of Economics* 111 (2), 605–637.
- Fan, Q., Hsu, Y.-C., Lieli, R. P., & Zhang, Y. (2020). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, 1–15.
- Farrell, M. H., Liang, T., & Misra, S. (2019). *Deep Neural Networks for estimation and inference*. Version 3. arXiv: [arXiv:1809.09953v3](https://arxiv.org/abs/1809.09953v3).
- Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ) (2018). *Elterngeld, Elterngeld-Plus und Elternzeit. Das Bundeselterngeld- und Elternzeitgesetz*. Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ).
- Federal Statistical Office (2006). *Kindertagesbetreuung regional 2006. Ein Vergleich aller 439 Kreise in Deutschland*. Federal Statistical Office, Wiesbaden.
- (2014). *Bevölkerung und Erwerbstätigkeit. Natürliche Bevölkerungsbewegung*. Federal Statistical Office, Wiesbaden.
- (2015a). *Bevölkerung und Erwerbstätigkeit. Natürliche Bevölkerungsbewegung*. Federal Statistical Office, Wiesbaden.
- (2015b). *Kindertagesbetreuung regional 2015. Ein Vergleich aller 402 Kreise in Deutschland*. Federal Statistical Office, Wiesbaden.
- (2015c). *Öffentliche Sozialleistungen. Statistik zum Elterngeld. Leistungsbezüge. 4. Vierteljahr 2014*. Federal Statistical Office, Wiesbaden.
- (2019). *Öffentliche Sozialleistungen. Statistik zum Elterngeld. Beendete Leistungsbezüge für im 3. Vierteljahr 2015 geborene Kinder*. Federal Statistical Office, Wiesbaden.
- Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica* 75 (1), 259–276.
- Fitzenberger, B., Osikominu, A., & Völter, R. (2005). Imputation rules to improve the education variable in the IAB employment subsample. *ZEW Discussion Paper 10/2005*.
- Foster, D. J. & Syrgkanis, V. (2020). *Optimal doubly robust estimation of heterogeneous causal effects*. Version 3. arXiv: [arXiv:1901.09036v3](https://arxiv.org/abs/1901.09036v3).
- Frölich, M. (2004a). A note on the role of the propensity score for estimating average treatment effects. *Econometric Reviews* 23 (2), 167–174.
- (2004b). Finite-sample properties of propensity-score matching and weighting estimators. *The Review of Economics and Statistics* 86 (1), 77–90.
- (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics* 139 (1), 35–75.
- Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review* 104 (4), 1091–1119.
- Goodman-Bacon, A. (2018). *Difference-in-differences with variation in treatment timing*. Tech. rep. National Bureau of Economic Research.
- Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica* 79 (2), 437–452.

- Graham, B. S., Pinto, C. C.d. X., & Egel, D. (2016). Efficient estimation of data combination models by the method of Auxiliary-to-Study Tilting (AST). *Journal of Business & Economic Statistics* 34 (2), 288–301.
- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66 (2), 315–331.
- Hahn, J. & Ridder, G. (2013). Asymptotic variance of semiparametric estimators with generated regressors. *Econometrica* 81 (1), 315–340.
- Hall, P., Racine, J., & Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* 99 (468), 1015–1026.
- Hall, P., Li, Q., & Racine, J. S. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *The Review of Economics and Statistics* 89 (4), 784–789.
- Hanel, B. & Riphahn, R. T. (2012). The employment of mothers – recent developments and their determinants in East and West Germany. *Jahrbücher für Nationalökonomie und Statistik* 232 (2), 146–176.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. Springer.
- Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies* 94 (4), 605–654.
- Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71 (4), 1161–1189.
- Hong, S.-H. (2013). Measuring the effect of Napster on recorded music sales: Difference-in-differences estimates under compositional changes. *Journal of Applied Econometrics* 28 (2), 297–324.
- 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., Lechner, M., & Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics* 175 (1), 1–21.
- Huber, M., Lechner, M., & Mellace, G. (2017). Why do tougher caseworkers increase employment? The role of program assignment as a causal mechanism. *Review of Economics and Statistics* 99 (1), 180–183.
- Imbens, G. W. & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47 (1), 5–86.
- Jessen, J., Jessen, R., & Kluve, J. (2019). Punishing potential mothers? Evidence for statistical employer discrimination from a natural experiment. *Labour Economics* 59, 164–172.
- Kennedy, E. H. (2020). *Optimal doubly robust estimation of heterogeneous causal effects*. Version 2. arXiv: [arXiv:2004.14497v2](https://arxiv.org/abs/2004.14497v2).
- Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 (4), 1229–1245.
- Kluve, J. & Schmitz, S. (2014). Social norms and mothers’ labor market attachment. *IZA Discussion Paper* 8115.
- (2018). Back to work: Parental benefits and mothers’ labor market outcomes in the medium run. *Industrial and Labor Relations Review* 71 (1), 143–173.
- Kluve, J. & Tamm, M. (2013). Parental leave regulations, mothers’ labor force attachment and fathers’ childcare involvement: Evidence from a natural experiment. *Journal of Population Economics* 26 (3), 983–1005.
- Knaus, M. C. (2020). *Double Machine Learning based program evaluation under unconfoundedness*. Version 1. arXiv: [arXiv:2003.03191v1](https://arxiv.org/abs/2003.03191v1).

- Knaus, M. C., Lechner, M., & Strittmatter, A. (2020a). Heterogeneous employment effects of job search programmes: A machine learning approach. *Journal of Human Resources*, 0718–9615R1.
- (2020b). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, forthcoming.
- Kolip, P. (2012). “Einflussfaktoren auf den Geburtsmodus: Kaiserschnitt vs. Spontangeburt”. *Gesundheitsmonitor* 2012.
- Lalive, R. & Zweimüller, J. (2009). How does parental leave affect fertility and return to work? Evidence from two natural experiments. *The Quarterly Journal of Economics* 124 (3), 1363–1402.
- Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics* 4 (3), 165–224.
- (2018). *Modified Causal Forests for estimating heterogeneous causal effects*. Version 2. arXiv: [1812.09487v2](https://arxiv.org/abs/1812.09487v2).
- Lee, S., Okui, R., & Whang, Y.-J. (2016). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics* 32 (7), 1207–1225.
- Lee, Y.-Y. (2018). Efficient propensity score regression estimators of multivalued treatment effects for the treated. *Journal of Econometrics* 204 (2), 207–222.
- Li, Q. & Racine, J. (2007). *Nonparametric Econometrics: Theory and practice*. Princeton University Press.
- Lu, C., Nie, X., & Wager, S. (2019). *Robust nonparametric difference-in-differences estimation*. Version 1. arXiv: [arXiv:1905.11622v1](https://arxiv.org/abs/1905.11622v1).
- Lunceford, J. & 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.
- Lundborg, P., Plug, E., & Rasmussen, A. W. (2017). Can women have children and a career? IV evidence from IVF treatments. *American Economic Review* 107 (6), 1611–37.
- Luo, Y. & Spindler, M. (2016). *High-dimensional L_2 boosting: Rate of convergence*. Version 2. arXiv: [1602.08927v2](https://arxiv.org/abs/1602.08927v2).
- Mammen, E., Rothe, C., & Schienle, M. (2012). Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics* 40 (2), 1132–1170.
- March of Dimes, PMNCH, Save the Children, & WHO (2012). *Born too soon: The global action report on preterm birth*. Ed. by World Health Organization. World Health Organization, Geneva.
- Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis* 17 (6), 571–599.
- Müller, D., Strauch, K., et al. (2017). Identifying mothers in administrative data. *FDZ-Methodenreport* 13/2017.
- Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5 (2), 99–135.
- (1994a). Kernel estimation of partial means and a general variance estimator. *Econometric Theory* 10 (2), 233–253.
- (1994b). The asymptotic variance of semiparametric estimators. *Econometrica* 62 (6), 1349–1382.
- Nie, X. & Wager, S. (2020). *Quasi-oracle estimation of heterogeneous treatment effects*. Version 4. arXiv: [arXiv:1712.04912v4](https://arxiv.org/abs/1712.04912v4).
- Pagan, A. & Ullah, A. (1999). *Nonparametric Econometrics*. 1st ed. Cambridge: Cambridge University Press.
- Patnaik, A. (2019). Reserving time for daddy: The consequences of fathers’ quotas. *Journal of Labor Economics* 37 (4), 1009–1059.

- Robins, J., Li, L., Tchetgen, E., van der Vaart, A., et al. (2008). “Higher order influence functions and minimax estimation of nonlinear functionals”. *Probability and statistics: essays in honor of David A. Freedman*. Institute of Mathematical Statistics, 335–421.
- Robins, J. M., Rotnitzky, A., & Zhao, P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89 (427), 846–866.
- Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica* 56 (4), 931–954.
- 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.
- (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39 (1), 33–38.
- Rossin-Slater, M., Ruhm, C. J., & Waldfogel, J. (2013). The effects of California’s paid family leave program on mothers’ leave-taking and subsequent labor market outcomes. *Journal of Policy Analysis and Management* 32 (2), 224–245.
- Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford economic papers* 3 (2), 135–146.
- Rubin, D. & Laan, M. J. van der (2007). A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics* 3 (1).
- Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* 29 (1), 185–203.
- (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66 (5), 688–701.
- Sant’Anna, P. H. & Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics* 219 (1), 101–122.
- Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 94 (448), 1096–1120.
- Schönberg, U. & Ludsteck, J. (2014). Expansions in maternity leave coverage and mothers’ labor market outcomes after childbirth. *Journal of Labor Economics* 32 (3), 469–505.
- Semenova, V. & Chernozhukov, V. (2020). *Debiased machine learning of conditional average treatment effects and other causal functions*. Version 6. arXiv: [1903.10075v6](https://arxiv.org/abs/1903.10075v6).
- Stuart, E. A., Huskamp, H. A., Duckworth, K., Simmons, J., Song, Z., Chernew, M. E., & Barry, C. L. (2014). Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Serv Outcomes Res Method* 14 (4), 166–182.
- Tamm, M. (2019). Fathers’ parental leave-taking, childcare involvement and labor market participation. *Labour Economics* 59, 184–197.
- Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. 1st ed. Springer Series in Statistics. Springer Science+Business Media.
- van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* 42 (3), 1166–1202.
- Wager, S. & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using Random Forests. *Journal of the American Statistical Association* 113 (523), 1228–1242.
- Wager, S. & Walther, G. (2016). *Adaptive concentration of Regression Trees, with application to Random Forests*. Version 3. arXiv: [arXiv:1503.06388v3](https://arxiv.org/abs/1503.06388v3).

- Waldfogel, J. (1999). The impact of the Family and Medical Leave Act. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* 18 (2), 281–302.
- Walker, M., Tekin, E., & Wallace, S. (2009). Teen smoking and birth outcomes. *Southern Economic Journal* 75 (3), 892–907.
- Welteke, C. & Wrohlich, K. (2019). Peer effects in parental leave decisions. *Labour Economics* 57, 146–163.
- Zhang, C.-H. & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (1), 217–242.
- Zimmert, F. (2019). Early child care and maternal employment: Empirical evidence from Germany. *IAB-Discussion Paper* 2/2019.
- Zimmert, M. (2020). *Efficient difference-in-differences estimation with high-dimensional common trend confounding*. Version 5. arXiv: [arXiv:1809.01643v5](https://arxiv.org/abs/1809.01643v5).
- Zimmert, M. & Lechner, M. (2019). *Nonparametric estimation of causal heterogeneity under high-dimensional confounding*. Version 1. arXiv: [arXiv:1908.08779v1](https://arxiv.org/abs/1908.08779v1).
- Zoch, G. & Hondralis, I. (2017). The expansion of low-cost, state-subsidized childcare availability and mothers' return-to-work behaviour in East and West Germany. *European Sociological Review* 33 (5), 693–707.

Curriculum Vitea

Education

- 2017 - 2020 PhD in Economics and Finance, University of St. Gallen
 Title: Essays in Causal Machine Learning
 Thesis supervisor: Prof. Dr. Michael Lechner
- 2015 - 2017 Master in Quantitative Economics and Finance, University of St. Gallen
- 2011 - 2015 B.Sc. International Economics, Eberhard Karls University Tübingen
- 2013 Singapore Management University, Singapore

Professional experience

- Since 2020 Business Analyst, Schweizerische Nationalbank (SNB)
- 2017 - 2020 Research Assistant to Professor Michael Lechner SEW Institute University of St. Gallen
- 2014 Intern, Portfolio Management (Multi Asset), Deutsche AWM
- 2014 Intern, Risk Controlling, KfW Group
- 2012 - 2013 Student Research Assistant, Department of Banking University Tübingen