Universität St.Gallen

Treatment evaluation with multiple outcome periods under endogeneity and attrition

Markus Frölich, Martin Huber

February 2014 Discussion Paper no. 2014-04

School of Economics and Political Science,
Department of Economics

University of St. Gallen

Editor: Martina Flockerzi
University of St.Gallen
School of Economics and Political Science
Department of Economics
Bodanstrasse 8
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35
Email seps@unisg.ch

Publisher: School of Economics and Political Science
Department of Economics
University of St.Gallen
Bodanstrasse 8
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35

Electronic Publication: http://www.seps.unisg.ch

Treatment evaluation with multiple outcome periods under endogeneity and attrition[1]

Markus Frölich, Martin Huber

Author's address:

Prof Dr. Markus Frölich
University of Mannheim
L7, 3-5
D-68131 Mannheim
Phone +49 621 181 3647
Fax +49 621 181 1931
Email froelich@uni-mannheim.de
Website http://froelich.vwl.uni-mannheim.de/

Martin Huber, Ph.D.
University of St. Gallen, SEW
Varnbüelstr 14
CH-9000 St. Gallen
Phone +41 71 224 2300
Fax +41 71 224 2302
Email martin.huber@unisg.ch
Website http://www.sew.unisg.ch/en/Assistenzprofessoren

[1] We have benefited from comments by Michael Lechner, Blaise Melly, Conny Wunsch, the editor and associate editor, and three anonymous referees. The first author acknowledges financial support from the Research Center (SFB) 884 "Political Economy of Reforms" Project B5, funded by the German Research Foundation (DFG). The second author acknowledges financial support from the Swiss National Science Foundation grant PBSGP1_138770.

Abstract

This paper develops a nonparametric methodology for treatment evaluation with multiple outcome periods under treatment endogeneity and missing outcomes. We use instrumental variables, pre-treatment characteristics, and short-term (or intermediate) outcomes to identify the average treatment effect on the outcomes of compliers (the subpopulation whose treatment reacts on the instrument) in multiple periods based on inverse probability weighting. Treatment selection and attrition may depend on both observed characteristics and the unobservable compliance type, which is possibly related to unobserved factors. We also provide a simulation study and apply our methods to the evaluation of a policy intervention targeting college achievement, where we find that controlling for attrition considerably affects the effect estimates.

Keywords

Treatment effect, attrition, endogeneity, panel data, weighting.

JEL Classification

C14, C21, C23, C24, C26.

1 Introduction

We develop a nonparametric methodology for evaluating the effect of an endogenous binary variable (referred to as treatment) in multiple outcomes periods where some outcomes are missing non-randomly due to non-response and attrition (e.g. survey non-response or truncation by death). Our identification strategy exploits an instrument (to control for treatment endogeneity), baseline covariates, and short-term (or intermediate) post-treatment variables to tackle the dynamic nature of the attrition problem. This in principle allows us to estimate the treatment effects also in later periods where the attrition problem is typically particularly severe.

The proposed methods appear important in the light of two fundamental trends that are currently observed in applied research in social sciences: First, the increasing use of randomized experiments and second, a growing interest in medium to long-term treatment effects of interventions, in order to see whether effects are sustainable. Even randomized experiments, which are frequently regarded as the gold standard for causal inference, are often plagued by imperfections such as noncompliance with treatment assignment and outcome attrition due to loss to follow-up. The noncompliance issue can be solved if it can be plausibly assumed that random treatment assignment provides a credible instrument for (endogenous) treatment take-up. While this is common practice for the identification of complier average causal effects (CACE) (also known as local average treatment effects, LATE) in experiments, see Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996), our approach also tackles the attrition problem. The latter appears particularly relevant when noting the increasing importance of long-term evaluations of policy interventions, as e.g. in the assessment of active labor market policies, e.g. Lechner, Miquel, and Wunsch (2011), or of educational interventions, e.g. Angrist, Bettinger, and Kremer (2006).

To see the contribution of this paper, it appears useful to review previously suggested approaches to correct for attrition. The very common missing at random (MAR) restriction assumes non-response or attrition to be conditionally ignorable (i.e., independent of the potential outcomes) given observed characteristics, see for instance Rubin (1976), Little and Rubin (1987), Robins, Rotnitzky, and Zhao (1994), Robins, Rotnitzky, and Zhao (1995), Carroll, Ruppert, and Stefanski (1995), Shah, Laird, and Schoenfeld (1997), Fitzgerald, Gottschalk, and Moffitt (1998), and Abowd, Crepon, and Kramarz (2001). Frangakis and Rubin (1999) suggest a relaxation of MAR in experiments which they call latent ignorability (LI). Non-response is assumed to be ignorable conditional on observed characteristics *and* the latent (compliance) type, characterizing how an individual's treatment state reacts on some instrument. See Barnard, Frangakis, Hill, and Rubin (2003), Frangakis, Brookmeyer, Varadhan, Safaeian, Vlahov, and Strathdee (2004), and Mealli, Imbens, Ferro, and Biggeri (2004) for related applications.

Approaches other than MAR and LI, permitting attrition to be related to unobservables in a general way, are referred to as non-ignorable non-response models. The earlier work, e.g. Heckman (1976), Hausman and Wise (1979), Bollinger and David (2001), and Chen, Wong, Dominik, and Steiner (2000), focussed on fully parameterized maximum likelihood estimation with identification often achieved only via functional form restrictions, see Little (1995) for an intuitive example. Instrumental variables for non-response and attrition offer an additional source of identification, see DiNardo, McCrary, and Sanbonmatsu (2006) for an application in an experimental context. In particular, such models allow for non-parametric identification and more flexible estimation, including the series regression approach of Das, Newey, and Vella (2003) and inverse probability weighting based on instruments for attrition as outlined in Huber (2012, 2013). While the standard framework consists of just one follow-up period, panel data sample selection models as suggested by Kyriazidou (1997, 2001) can be used to consider multiple periods as in this paper. In addition to dynamic attrition, Semykina and Wooldridge (2006) even allow for endogenous regressors, given that sufficiently many instruments to control for attrition and endogeneity are available.

An alternative to the assumptions discussed so far are methods that do not require a fully specified model for attrition, however, at the cost of sacrificing point identification. E.g., building on the partial identification literature (Robins, 1989, Manski, 1989, 1990), Zhang and Rubin (2003), Zhang, Rubin, and Mealli (2008), Imai (2008), and Lee (2009), among others, bound treatment effects in the presence of non-response under comparably mild restrictions. Another approach is multiple imputation of missing values, which goes back to Rubin (1977, 1978). Based on Bayesian techniques, multiple attrition models are used to impute multiple sets of plausible values for the missing data in order to obtain a probability interval for the parameter of interest. Finally, Rotnitzky, Robins, and Scharfstein (1998), Scharfstein, Rotnitzky, and Robins (1999), and Xie and Qian (2012) (who even allow for non-monotone non-response), among others, propose sensitivity checks for violations of MAR related to unobservables by varying the nuisance term causing non-ignorable attrition over a relevant range to examine the robustness of the results. By not considering arguably implausible attrition mechanisms, this approach likely yields more informative results than Manski-style bounds analysis and therefore provides a middle ground between the latter and point identification.

In this paper, we propose a new nonparametric approach for point identification of the average causal effect on the compliers (those who are responsive to the instrument). We rely on pre-treatment covariates and (endogenous) post-treatment variables to control for attrition in a panel data framework as well as a single instrument (e.g., random assignment in an experiment) to tackle treatment endogeneity. (We only require a *single* instrument, which is important because instrumental variables are often hard to find in applications.) Our method for the evaluation of binary treatments provides three improvements compared to standard MAR. Firstly, we do not control for pre-treatment covariates only. That would ignore information about the intermediate variables, which presumably are important predictors of non-response in many empirical contexts. Secondly, we allow for treatment endogeneity which has rarely been considered under MAR. Exceptions are Yau and Little (2001) and Ding and Lehrer (2010), who, however, rely on considerably stronger functional form assumptions and in the latter case, on a difference-in-difference strategy rather than an instrument. Thirdly, in our main identification theorem, we develop a panel data extension of LI by permitting that attrition does not only depend on observables but also on the latent types.

It is also interesting to compare our framework to the literature on dynamic treatment regimes, e.g. Robins, Greenland, and Hu (1999), Murphy, van der Laan, and Robins (2001), and Lok, Gill, van der Vaart, and Robins (2004). If one were to consider attrition as a dynamic treatment regime, those methods could be adjusted to our situation. However, they are all based on a type of dynamic ignorability condition, which would correspond to a MAR assumption in our context. In contrast, we also allow for selection on the latent types and make use of an instrumental variable to overcome the endogeneity problems.

Our framework is also more general than the original LI assumption of Frangakis and Rubin (1999). Firstly, we permit two-sided noncompliance (i.e. the existence of never takers, who are never treated irrespective of the instrument, and of always takers, who are always treated) and extend LI to conditional LI given observables. Secondly, we consider multiple periods under comparably weak assumptions, whereas the literature conventionally imposes more structure and assesses only one outcome period, see for instance Peng, Little, and Raghunathan (2004). Note, however, that the identification problem considered in this paper is distinct from non-ignorable non-response and panel data sample selection models. I.e., we assume that conditional on observed characteristics *and* the latent type, there are no further unobservables that are jointly related to attrition and the potential outcomes. Therefore, we do *not* require any additional instruments for non-response, which are typically hard to find in applications, see the discussion in Fitzgerald, Gottschalk, and Moffitt (1998). All in all, the methods proposed in this paper use less severe functional form and/or identifying assumptions than many non-response models invoked in recent empirical applications, see the examples in Preisser, Galecki, Lohman, and Wagenknecht (2000), Mattei and Mealli (2007), Shepherd, Redman, and Ankerst (2008), Zhang, Rubin, and Mealli (2009), Frumento, Mealli, Pacini, and Rubin (2012), and Wang, Rotnitzky, Lin, Millikan, and Thall (2012).

The remainder of this paper is organized as follows. Section 2 introduces a treatment effect model with endogeneity and multiple outcome periods and shows nonparametric identification under two distinct forms of attrition. For the ease of exposition, only two outcome periods are considered in the main text. A simulation study is provided in Section 3. Section 4 presents an application to a policy intervention aiming to increase college achievement previously analyzed by Angrist, Lang, and Oreopoulos (2009). Section 5 concludes. The (separate) online appendix presents identification in the more general case with several outcome periods along with the identification proofs, discusses the implications of our identifying assumptions in a parametric benchmark model, provides nonparametric and $\sqrt{n}$-consistent estimators based on kernel regression along with the proofs of their asymptotic properties, and includes an extended range of simulation studies.

2 Model and identification

Suppose we are interested in estimating the treatment effect of a binary variable $D \in \{0, 1\}$ on an outcome Y_t, where the subscript t denotes the period ($t = 1, 2, 3, ...$) after the start of the treatment. All variables observed *prior to* the treatment are indexed by period zero and are denoted as X_0. The *potential* outcomes Y_t^1 and Y_t^0 are the outcomes that would have been realized if D had been set to 1 or 0, respectively, by external intervention. (To avoid confusion between subscripts and superscripts we sometimes write $Y_{t=1}^0$ instead of Y_1^0 when referring to a specific time period.) In our nonparametric identification framework, two major issues have to be dealt with: endogenous treatment selection and missing outcome data due to attrition or non-response. The indicator variable R_t will denote whether in time period t outcome data is observed ($R_t = 1$) or missing

($R_t = 0$). We assume that information on the treatment D and baseline covariates X_0 is available for all individuals, but that individuals may not respond or drop out at follow-up data collection. In most applications, non-response increases at later follow-up periods.

2.1 Treatment endogeneity without attrition

Consider first the case without missing data. Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) have shown that in the presence of an instrumental variable (denoted by Z) satisfying particular assumptions, treatment effects are nonparametrically identified for a subset of the population, the so-called compliers. Adhering to their terminology, let D_i^z denote the *potential* treatment status of some individual i if Z_i were hypothetically set to z. For ease of exposition we will focus on a binary Z, which often occurs in experiments, even though the framework could be extended easily to non-binary discrete instruments, see e.g. Frölich (2007). The two binary potential treatment states D_i^0 and D_i^1 partition the population into four different types of individuals according to treatment behavior: the always takers (a) who are treated irrespective of the instrument ($D_i^1 = 1, D_i^0 = 1$), the never takers (n) who are never treated ($D_i^1 = 0, D_i^0 = 0$), the compliers (c) who only attend treatment if the instrument takes the value one ($D_i^1 = 1, D_i^0 = 0$), and the defiers (d) who only attend treatment if the instrument takes the value zero ($D_i^1 = 0, D_i^0 = 1$). As shortcut notation we will henceforth use $\mathcal{T}_i$ for 'type' with $\mathcal{T}_i \in \{a, n, c, d\}$. Note that the type of any individual is only partially observed, i.e. *latent*, because the observed D and Z do not uniquely determine $\mathcal{T}$, as discussed in the appendix.

Abadie (2003) shows the nonparametric identification of the CACE (or LATE)

$$E[Y_t^1 - Y_t^0 | \mathcal{T} = c],$$

i.e. the effect for the compliers, under conditions implying conditional validity of the instrument given observed baseline characteristics, which we denote by X_0:

$$\begin{aligned} &\{Y_t^d, \mathcal{T}\} \perp\!\!\!\perp Z | X_0 && \text{for } d \in \{0, 1\}, \\ &\Pr(\mathcal{T} = d | X_0) = 0 && \Pr(\mathcal{T} = c | X_0) > 0. \end{aligned}$$

The first line assumes independence between the instrument and the type and potential outcomes, conditional on X_0. (Note that '$\perp\!\!\!\perp$' denotes statistical independence). It thus assumes random assignment of Z and an exclusion restriction with respect to the potential outcomes for given values of the baseline covariates X_0. The second line states that the treatment is (weakly) monotonic in the instrument conditional on X_0 so that defiers are ruled out and compliers do exist.

In the subsequent sections, we extend the CACE framework to allow for missing values in the outcome variables Y_t. We focus on the case of attrition (i.e. missingness as an absorbing state), which is the most frequent concern in empirical applications, particularly in impact evaluation. However, our approach does also permit intermittent missingness, implying that intermediate outcomes are missing while later ones are observed, but in this case does not exploit the information from later waves. With this respect, it is interesting to note that several contributions considering parametric missing data models distinguish explicitly between attrition and intermittent missingness, see for instance Xie and Qian (2012). In those approaches, however, one either has to additionally model the re-entry process after non-response or specify attrition and intermittent missingness as two separate processes. Under additional assumptions, also our nonparametric approach could use information from the re-entrants in order to permit more precise estimates (given that re-entry occurs sufficiently often), but the identification expressions and estimators would become less tractable. Since we aim at imposing as few restrictions as possible and do not make use of additional instruments (other than for treatment), which are often not available in applications, we therefore only model the non-response process and ignore any information after the first non-response. Hence, we do permit that individuals have missing data in only one or several waves and then re-enter the panel after periods of non-observability, but we do not exploit this information.

In the following, we denote by X_t the observed characteristics for any $t > 0$, i.e. *after* treatment. Note that X_t usually also contains the outcome Y_t. In contrast to X_0, these variables X_t may possibly already be causally affected by the treatment, and we refer to them as (endogenous) post-treatment characteristics. (Note that whereas X_0 is permitted to be endogenous in the sense of Frölich (2008), i.e. that X_0 may be correlated with baseline unobservables, X_0 is not permitted to be causally influenced by treatment D, e.g. due to anticipation.) Furthermore, define $\underline{X}_t = \{X_1, ..., X_t\}$ to be the history of the characteristics up to time t, where we do *not* include X_0 here in order to make the distinction between pre-treatment and endogenous post-treatment variables explicit. Accordingly, X_t^d and $\underline{X}_t^d$ denote the *potential* values of the characteristics and of their history, respectively, at time t, if the treatment had been set to d by external intervention. Furthermore, let R_t be the response indicator in period t. I.e., X_t and Y_t are only observed if $R_t = 1$. Our setup permits that R_1 is zero for some individuals, such that outcome data is completely missing for those subjects. The history of response indicators over the post-treatment periods up to t is denoted by $\underline{R}_t = \{R_1, ..., R_t\}$. The *potential* values of response and the response history are denoted by R_t^d and $\underline{R}_t^d$, respectively.

The occurrence of attrition and non-response may have many reasons. In the simplest and least realistic case, it is only triggered by random events happening after treatment, such that outcomes are missing completely at random (MCAR), see e.g. Rubin (1976) and Heitjan and Basu (1996). However, it is more likely that attrition depends also on observed and/or unobserved characteristics of the individuals. In particular, attrition may depend on Y_{t-1}, which is an endogenous variable that has been causally affected by the treatment. In addition, attrition could also be directly causally affected by the treatment itself, e.g. due to side effects or adverse events of a drug treatment in a medical intervention. Finally, attrition could also be caused directly by the instrumental variable Z.

Our identification strategy requires us to restrict the missing data process in two ways: First, we assume that non-response in time t is not *simultaneously* related to the outcome variable in time t. This implies that while any variables measured in the past may trigger non-response today, current and future values of the outcome variable are *not* permitted to do so. Non-response is thus considered to be predetermined. Second, we need to impose some restrictions on the relationship between the instrument and non-response. In the following two subsections, we will discuss two different identification assumptions. The first approach permits non-response to depend on unobservables, but requires it to be ignorable given the observed characteristics *and* the latent types (*conditional LI*). The second approach assumes that non-response is missing at random (MAR) given the observed pre- and post-treatment characteristics. While the first setup appears to be more general in most applications than the second one, they are not strictly nested. I.e. while the first approach is less restrictive with respect to the non-response process, the second one imposes weaker (albeit only mildly weaker) assumptions on the instrument.

Our analysis covers four cases. First, it includes randomized experiments with full compliance. Then, the exclusion restriction is valid with X_0 being the empty set (i.e. not controlling for any covariates) and using D as its own instrument, i.e. defining $Z_i \equiv D_i$. Second, under random assignment but imperfect compliance, we may use the randomization Z as an instrumental variable for the actual treatment receipt D. If the randomization probability is the same for everyone, X_0 may again be the empty set. Third, the framework also includes observational studies, where the instrumental variable assumption is often plausible only after conditioning on some variables X_0, see Abadie (2003), Tan (2006), and Frölich (2007). Finally, when controlling for some X_0 and using D as its own instrument, i.e. defining $Z_i \equiv D_i$, we impose what is referred to in the literature as the selection on observables, unconfoundedness, ignorable assignment, or conditional independence assumption, see for instance Rosenbaum and Rubin (1983), Lechner (1999), and Imbens (2004). Hence, although we focus on instrument-based identification, our identification results are also directly applicable to the selection on observables framework with missing outcome data.

2.2 Non-response under conditional latent ignorability

This section presents the identifying assumptions for the case of *conditional latent ignorability*. We permit that the response process at time R_t is related to all observed variables in the past *and* that it is a function of the latent type $\mathcal{T}$. Hence, the response process is supposed to be predetermined, which means that past values of the outcomes and further observed characteristics may affect the response behavior today. However, conditional on these past values, the instrument and conditional on the latent type, current and future outcomes must be independent of non-response in period t. This is, for instance, different to Xie and Qian (2012), who permit response and contemporaneous outcomes to be related and propose various sensitivity checks. Assumption 1 formalizes predetermined non-response under conditional LI. As already mentioned, $\underline{X}_{t-1}$ may contain both intermediate outcomes $\underline{Y}_{t-1}$ as well as other observed characteristics.

Assumption 1: Predetermined non-response

$$Y_{t+s} \perp\!\!\!\perp R_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z, \mathcal{T}, \qquad \text{for } s \geq 0. \tag{1}$$

The plausibility of predetermined non-response (not related to contemporaneous outcomes) needs to be judged in the light of the application at hand. Some statistical support that this may be an empirically relevant case comes from Hirano, Imbens, Ridder, and Rubin (2001), who provide conditions implied by non-response related to (i) past information and (ii) contemporaneous outcomes that can be tested if a refreshment sample is available. Applying their test to a Dutch household survey, they reject attrition related to contemporaneous outcomes, but do not reject predetermined non-response at any conventional level. Our assumption may for instance appear plausible in the context of educational outcomes, where Y_t denotes a measure of cognitive skills (e.g. test scores or grades) at the end of some academic year t and R_t is an indicator for (not) having dropped out of school. Predetermined non-response is (closely) satisfied if individuals decide to remain in or leave education (mainly) based on their academic performance in the previous academic year, Y_{t-1}, so that the drop-out decision R_t is taken shortly after that, e.g. during or at the end of summer vacation.

In addition to Assumption 1, we invoke exclusion, monotonicity, and common support restrictions, as stated in Assumptions 2 and 3. The latter are similar to Abadie (2003), apart from that we have to strengthen the instrumental exclusion restriction for the always and never takers.

Assumption 2: Exclusion restriction: For $d \in \{0, 1\}$

$$(Y_t, \underline{X}_{t-1}, \underline{R}_{t-1}) \perp\!\!\!\perp Z | X_0, \mathcal{T} \in \{a, n\}$$
$$Y_t^d \perp\!\!\!\perp Z | X_0, \mathcal{T} = c$$
$$\mathcal{T} \perp\!\!\!\perp Z | X_0.$$

Assumption 2 requires that conditional on the observed baseline characteristics, the instrumental variable Z affects neither the histories of characteristics (possibly including intermediate potential outcomes) nor of responses of the always and never takers up to one period prior to the outcome period considered. In the two outcome periods case for instance, only X_1, R_1 are restricted in this way, but not R_2. Furthermore, note that for the always takers, the exclusion restriction only refers to the potential outcome under treatment, because $(Y_t, \underline{X}_{t-1}, \underline{R}_{t-1}) = (Y_t^1, \underline{X}_{t-1}^1, \underline{R}_{t-1}^1)$ for $\mathcal{T} = a$, while $(Y_t^0, \underline{X}_{t-1}^0, \underline{R}_{t-1}^0)$ is not restricted. An analogous statement holds for the never takers.

Assuming that Z does not affect the response behavior of always and never takers may appear reasonable in double-blind randomized medical trials, where individuals are not even aware of their treatment assignment. In non-blinded trials, this assumption seems generally less innocuous. Consider e.g. a non-blinded randomized drug-trial, where a never taker does not take the new drug irrespective of being assigned to treatment or control. Under assignment to treatment she actively decides to not comply with the protocol, whereas she would comply when being assigned to the control group. It is conceivable that the decision to not comply might affect response behavior. In other cases it may however be less of a problem. E.g. assume the randomization of a school voucher (for tuition fees) where the outcome of interest is some test score in the final grade and non-response is characterized by dropping out from school. Here, it appears more reasonable that mere voucher assignment does not affect the drop out decision of never takers, who would not use the school voucher anyway.

The stronger exclusion restriction is only required for the always and never takers, *not* for the compliers. Concerning the latter, only the standard exclusion restriction $Y_t^d \perp\!\!\!\perp Z|X_0, \mathcal{T} = c$ is imposed (see second line of Assumption 2) such that non-response may be arbitrarily related to and thus, affected by the instrument. This may happen either directly, e.g. when Z is treatment assignment and the notification of having been assigned to the treatment or control group itself changes the response behavior, or indirectly via treatment choice, e.g. due to the side effects or adverse events of a drug treatment which influences attrition.

Assumption 3: Monotonicity and support restrictions

$$\begin{array}{ll} \text{Existence of compliers:} & \Pr(\mathcal{T} = c) > 0 \\ \text{Monotonicity:} & \Pr(\mathcal{T} = d) = 0 \\ \text{Common support:} & 0 < \Pr(Z = 1|X_0) < 1. \end{array}$$

Assumption 3 invokes weak monotonicity, i.e. the existence of compliers and the non-existence of defiers (or vice versa). For nonparametric identification, common support in the baseline characteristics X_0 across the populations receiving and not receiving the instrument must also hold. This is e.g. satisfied in randomized experiments, where $\Pr(Z = 1|X_0)$ is often a constant.

Theorem 1 shows the identification of the mean potential outcomes of the compliers. For ease of exposition, only two outcome periods are considered here, i.e. $t \in \{1, 2\}$, while the general result for more than two periods is provided in the online appendix. For a concise exposition of the results, we define the following conditional probabilities:

$$\begin{aligned} \pi &= \Pr(Z = 1|X_0) \\ P_t &= \Pr(Z = 1|X_0, \underline{X}_t, \underline{R}_t = 1, D = 1) \\ P_t' &= \Pr\left(Z = 1|X_0, \underline{X}_t, \underline{R}_{t+1} = 1, D = 1\right) \\ \Xi_t &= \Pr(R_{t+1} = 1|X_0, \underline{X}_t, \underline{R}_t = 1, D = 1) \\ \Xi_{t,Z=z} &= \Pr(R_{t+1} = 1|X_0, \underline{X}_t, \underline{R}_t = 1, D = 1, Z = z). \end{aligned}$$

Identification is based on a weighting representation in which four conditional probabilities enter multiplicatively: The probability that Z takes the value one, conditional on three different sets of regressors, and a time-varying conditional response probability. For identification, Ξ_t has to be larger than zero, i.e. for each value of the covariates (X_0,$\underline{X}_t$), the probability of attrition must not be one. Then, the treatment effect on the compliers is identified as $E\left[Y_t^1|\mathcal{T} = c\right] - E\left[Y_t^0|\mathcal{T} = c\right]$.

The intuition underlying Theorem 1 is as follows. By the independence of Z and $\mathcal{T}$ given X_0 stated in Assumption 2, the proportions of compliers, always takers, and never takers in groups defined by D and Z are identified. By Assumption 1, the first period potential outcomes are independent of first period response conditional on X_0, Z, and $\mathcal{T}$, and in the second period, independence of Y_2 and R_2 holds by additionally conditioning on X_1 and R_1. Together with the exclusion restrictions on the compliers' potential outcomes as well as the potential outcomes and pre-period responses (only relevant for the second period) of always and never takers postulated in Assumption 2, this ultimately allows isolating the mean potential outcomes of compliers in the mixed groups with $(Z=1, D=1)$ and $(Z=0, D=0)$, so that the CACE is identified. Finally, it is worth noting that if there was no attrition, the CACE based on the expressions in Theorem 1 would simplify to equation (11) in Frölich (2007), which provides a representation of the CACE based on inverse probability weighting in the absence of the missing outcomes problem.

Theorem 1 *Under Assumptions 1, 2 and 3, the potential outcomes in periods* $t \in \{1,2\}$ *are identified as*

$$\begin{aligned} E\left[Y^1_{t=1}|\mathcal{T}=c\right] &= E\left[Y_{t=1}R_{t=1}\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\frac{1}{\Xi_0}\frac{P_0-\pi}{P'_0-\pi}\right] \times \frac{1}{E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right]} \\ E\left[Y^1_{t=2}|\mathcal{T}=c\right] &= E\left[Y_{t=2}R_{t=1}R_{t=2}\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\frac{1}{\Xi_0\Xi_1}\frac{P_0-\pi}{P'_0-\pi}\frac{P_1-\pi}{P'_1-\pi}\right] \times \frac{1}{E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right]}. \qquad (2) \end{aligned}$$

An equivalent expression for $E\left[Y^0_t|\mathcal{T}=c\right]$ *is obtained by replacing* D *with* $1-D$ *and* $D=1$ *with* $D=0$ *everywhere.*

2.3 Non-response under the missing at random assumption

In this section we consider an alternative identification approach, where the response process is assumed to be ignorable conditional on observed characteristics, which corresponds to a type of MAR assumption. I.e., we do no longer permit that the unobserved type $\mathcal{T}$ is related to response behavior. This implies that only unobservables that are not related to the potential outcomes are allowed to affect attrition. Again, past values of Y may trigger non-response in the current period, but neither present nor future values of Y. As stated in Assumption 1', response behavior might depend on all past values of X, which itself could be endogenous, i.e. causally affected by the treatment.

Assumption 1': Predetermined non-response

$$Y_{t+s} \perp\!\!\!\perp R_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z, D \qquad \text{for } s \geq 0. \tag{3}$$

The key difference between Assumption 1' and Assumption 1 is that the latter permits the response behavior to depend on the latent type $\mathcal{T}$, while the former does not. Still, Assumption 1' allows response to be a function of the received treatment, which is a relevant scenario e.g. if the treatment leads to dissatisfaction and reduces the willingness to provide outcome data. On the other hand, one can think of many frameworks where it is not the treatment receipt alone that determines response behavior but rather the unobserved type $\mathcal{T}$ of an individual, as permitted in Assumption 1. Consider e.g. an educational intervention as analyzed in Angrist, Lang, and Oreopoulos (2009) where college students are randomly provided with services and/or financial incentives to obtain better grades. In this context, never takers who do not comply when offered a treatment might have a higher probability to drop out due to a lower commitment to this particular college or to higher education in general. Assumption 1' therefore appears to be more restrictive than Assumption 1 in many empirical applications.

On the other hand, since we need no longer condition on the latent type, the restrictions on the instrument can be relaxed somewhat. The following Assumption 2' is thus a little weaker than Assumption 2 because exclusion restrictions of the instrument on the response behavior do not have to be imposed for any type. This may be of practical relevance in randomized trials e.g. if those always takers who were not randomized into the treatment ($Z = 0$) are less inclined to respond than those with $Z = 1$ due to their discontent about having to organize the treatment receipt through alternative means. In this case, Assumption 2 is violated while Assumption 2' may still hold.

Assumption 2': Exclusion restriction: For $d \in \{0, 1\}$

$$(Y_t^d, \mathcal{T}) \perp\!\!\!\perp Z | X_0.$$

Theorem 2 gives the identification results for the compliers under MAR for the case of two outcome periods, while the general result for more outcome periods is provided in the appendix.

Theorem 2 *Under Assumptions 1', 2' and 3, the potential outcomes in periods $t \in \{1, 2\}$ are* *identified as*

$$E\left[Y^1_{t=1}|\mathcal{T}=c\right] = E\left[\frac{Y_{t=1}R_{t=1}DZ}{\pi\Xi_{0,Z=1}} - \frac{Y_{t=1}R_{t=1}D(1-Z)}{(1-\pi)\Xi_{0,Z=0}}\right] \times \frac{1}{E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right]},$$

$$E\left[Y^1_{t=2}|\mathcal{T}=c\right] = E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}} - \frac{Y_{t=2}R_{t=1}R_{t=2}D(1-Z)}{(1-\pi)\Xi_{0,Z=0}\Xi_{1,Z=0}}\right] \times \frac{1}{E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right]}. \quad (4)$$

The expression for $E\left[Y^0_t|\mathcal{T}=c\right]$ *is obtained by replacing* D *with* $1-D$ *and* $D=1$ *with* $D=0$ *everywhere.*

Note that the assumptions underlying Theorems 1 and 2 are partly testable. Consider first the case that attrition is zero in some outcome period (e.g. zero attrition in the first follow-up period). Our setup then collapses to the standard LATE assumptions, for which tests have been proposed by Huber and Mellace (2013) and Kitagawa (2013). Similar tests could be derived for the case with attrition. By straightforward modifications of Theorems 1 and 2 the distribution functions of the potential outcomes among compliers are identified and therefore, also the density functions. As in Kitagawa (2013), a testable implication is that the estimated potential outcome densities of compliers must not be significantly negative at any point in the outcome support, because this would indicate the failure of our identifying assumptions. As a further possibility to validate the MAR assumptions underlying Theorem 2, one may consider the approach of Hirano, Imbens, Ridder, and Rubin (2001) for testing MAR models in the presence of a refreshment sample. We leave the detailed derivations and analyses of such tests for future research.

3 Finite sample properties

To illustrate the behavior of the proposed estimators in finite samples we examine a small simulation study in this section. We consider the following data generating process (DGP) with, for the sake of simplicity, parsimonious specifications of the instrument, treatment, covariate, response, and outcome equations that nevertheless give an idea about which forms of attrition can be controlled for based on our identification results:

$$
\begin{aligned}
X_0 &\sim \text{uniform}(0,1), \qquad Z = I\{0.25X_0 + W > 0\}, \qquad D = I\{\alpha Z - 0.25X_0 + U_0 > 0.5\},\\
Y_1 &= 0.5X_0 + 0.5D + \kappa D U_0 + U_1,\\
X_1 &= 0.5Y_1 + 0.5Q,\\
R_1 &= I\{0.25X_0 + 0.25D + \beta Z + \gamma I\{0.5 - \alpha < U_0 - 0.25X_0 \leq 0.5\} + \delta Y_1 + V > 0\},\\
Y_2 &= 0.5X_0 + X_1 + D + \kappa D U_0 + U_2,\\
R_2 &= R_1 I\{0.25X_0 + 0.25X_1 + 0.25D + \beta Z + \gamma I\{0.5 - \alpha < U_0 - 0.25X_0 \leq 0.5\} - \delta Y_2 + \epsilon > 0\},\\
& \quad \text{each of } Q, V, W, \epsilon \sim N(0,1), \text{ independent of each other,}\\
& \quad \text{and} \begin{pmatrix} U_0 \\ U_1 \\ U_2 \end{pmatrix} \sim N(\mu, \sigma), \text{ where } \mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \text{ and } \sigma = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}.
\end{aligned}
$$

α, β, γ, δ, κ are parameters of the DGP that will be varied later. $I\{\cdot\}$ denotes the indicator function which is one if its argument is true and zero otherwise. Q, V, W, ϵ are random nuisance variables that are standard normal with zero correlation. U_0, U_1, U_2 are unobserved terms in the treatment and outcome equations in various periods. Correlation among these variables causes the endogeneity problem we have to deal with: Endogeneity is caused by the fact that U_0 affects the treatment D and is also associated with the outcomes Y_1, Y_2 through its correlation with U_1 and U_2. The response indicators R_1, R_2 are equal to one if the outcome is observed in the respective period. Attrition is modeled as an absorbing state, i.e., R_2 is necessarily zero if $R_1 = 0$. X_0, X_1 are observed covariates. The uniformly distributed X_0 confounds the instrument because of its impact on Z and Y_1 and Y_2. Therefore, the instrument is only conditionally valid given X_0. The latter also affects response in both periods, thus causing attrition bias if not controlled for. Similarly, X_1 jointly influences R_2 and Y_2, creating further bias in the second period. Note that X_1 is a function of Y_1, which incorporates the idea that previous outcomes or functions thereof might be used to model attrition in the current period.

Our set up contains several tuning parameters: α, β, γ, δ, κ. In the treatment equation, α determines the strength of the instrument and thus the share of compliers. The compliers are those individuals with values of U_0 and X_0 such that $0.5-\alpha < U_0 - 0.25X_0 \leq 0.5$. The larger α, the more individuals react to a change in the instrument by switching their treatment status. We consider two values of α: α=0.68 and α=1.35, resulting in complier shares of roughly 25% and 50% under effect homogeneity, respectively. β in the response equations gauges the effect of the instrument on R_1 and R_2. If $\beta \neq 0$, the exclusion restriction of Z on response as postulated in Assumption

2 is violated and estimators based on Theorem 1 are inconsistent. γ defines the extent to which the compliers' response behavior differs from the remainder of the population (i.e., the never and always takers). To see this, remember that $I\{0.5 - \alpha < U_0 - 0.25X_0 \leq 0.5\}$ is an indicator for being a complier. For $\gamma \neq 0$, Assumption 1' is violated because response then depends on the latent types. In this case, estimators based on Theorem 2 are inconsistent. δ determines whether response is related to the outcomes of the current period, as for instance considered in Xie and Qian (2012). I.e., if $\delta \neq 0$, then R_t depends on Y_t such that neither Assumption 1 nor Assumption 1' are satisfied. Hence, estimators based on Theorems 1 or 2 are all inconsistent. Finally, κ determines whether the treatment effects are homogeneous or heterogeneous as a function of the unobservables U_0. For $\kappa = 0$, the treatment effects are homogeneous, i.e. identical for everyone. In this case, the treatment effect is 0.5 for everyone in the first period and 1.25 in the second period. (The effect of 1.25 consists of the direct effect of D on Y_2, which is 1.00, and the indirect effect of $0.5 \cdot 0.5 = 0.25$ running through X_1.) For $\kappa \neq 0$, the treatment effects differ depending on the values of U_0. Therefore, the CACE differs from the average effect in the total population because of different distributions of U_0.

We simulate the DGP 1000 times with a sample size of 5000 observations, which is representative for many recently conducted field experiments in social sciences, see for instance Angrist, Bettinger, and Kremer (2006) and Bertrand and Mullainathan (2004). (The separate online appendix also examines other sample sizes.) We investigate the performance of the following estimators: (i) naive estimation based on mean differences in observed treated and non-treated outcomes that ignores both treatment endogeneity and attrition, (ii) CACE estimation based on equation (11) in Tan (2006) or equation (12) in Frölich (2007) that controls for endogeneity, but ignores attrition (denoted by $\hat{\omega}$), (iii) CACE estimation using expression (2) of Theorem 1 (denoted by $\hat{\theta}$), and (iv) CACE estimation using expression (4) of Theorem 2 (denoted by $\hat{\phi}$). The propensity scores in $\hat{\omega}$, $\hat{\theta}$, and $\hat{\phi}$ are estimated by local constant kernel regression (with Gaussian kernel). The bandwidths were chosen according to the nearest-neighbor-based default smoothing parameter in the R-package 'locfit', which was 0.7. (The results were similar when using a different kernel function such as the Epanechnikov kernel and/or when using other bandwidth values such as 0.6 and 0.8. However, values smaller than 0.6 considerably increased the variance of $\hat{\theta}$, whereas the estimates and standard errors were fairly robust for larger bandwidth values, e.g. 1.0 or larger.)

We also consider trimmed versions of $\hat{\theta}$ and $\hat{\phi}$ in order to prevent denominators from being

Table 1: Simulation 1 - treatment endogeneity and conditional LI

	Homogeneous effects											
	α=0.68, β=0, γ=0.5, δ=0, κ=0						α=1.35, β=0, γ=0.5, δ=0, κ=0					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.70	0.03	0.70	0.92	0.06	0.92	0.56	0.03	0.56	0.69	0.06	0.69
$\hat{\omega}$	0.05	0.11	0.13	0.12	0.16	0.20	0.03	0.06	0.07	0.07	0.08	0.11
$\hat{\theta}$	0.01	0.12	0.12	-0.03	0.17	0.17	0.01	0.06	0.06	-0.01	0.09	0.09
$\hat{\theta}_{\text{trim}}(0.15)$	0.01	0.12	0.12	-0.03	0.17	0.17	0.01	0.06	0.06	-0.01	0.09	0.09
$\hat{\theta}_{\text{trim}}(0.01)$	0.01	0.12	0.12	-0.03	0.17	0.17	0.01	0.06	0.06	-0.01	0.09	0.09
$\hat{\phi}$	-0.14	0.14	0.20	-0.41	0.23	0.47	-0.09	0.07	0.12	-0.25	0.11	0.27
$\hat{\phi}_{\text{trim}}(0.15)$	-0.14	0.14	0.20	-0.41	0.23	0.47	-0.09	0.07	0.12	-0.25	0.11	0.27
$\hat{\phi}_{\text{trim}}(0.01)$	-0.14	0.14	0.20	-0.41	0.23	0.47	-0.09	0.07	0.12	-0.25	0.11	0.27
MAR G-comp.	0.71	0.03	0.71	0.94	0.06	0.94	0.56	0.03	0.56	0.70	0.06	0.70
Heckman	0.36	1.59	1.63	0.03	1.61	1.61	0.33	3.30	3.32	-0.22	2.63	2.64
true CACE		0.50			1.25			0.50			1.25	
mean response		0.63			0.41			0.69			0.49	
	Heterogeneous effects											
	α=0.68, β=0, γ=0.5, δ=0, κ=0.5						α=1.35, β=0, γ=0.5, δ=0, κ=0.5					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.99	0.04	0.99	1.34	0.06	1.34	0.83	0.04	0.83	1.10	0.06	1.10
$\hat{\omega}$	0.05	0.13	0.14	0.12	0.18	0.22	0.03	0.07	0.08	0.08	0.10	0.13
$\hat{\theta}$	0.01	0.13	0.13	-0.03	0.20	0.20	0.01	0.07	0.07	-0.01	0.10	0.10
$\hat{\theta}_{\text{trim}}(0.15)$	0.01	0.13	0.13	-0.03	0.20	0.20	0.01	0.07	0.07	-0.01	0.10	0.10
$\hat{\theta}_{\text{trim}}(0.01)$	0.01	0.13	0.13	-0.03	0.20	0.20	0.01	0.07	0.07	-0.01	0.10	0.10
$\hat{\phi}$	-0.20	0.16	0.25	-0.52	0.26	0.58	-0.14	0.08	0.16	-0.35	0.13	0.37
$\hat{\phi}_{\text{trim}}(0.15)$	-0.20	0.16	0.25	-0.52	0.26	0.58	-0.14	0.08	0.16	-0.35	0.13	0.37
$\hat{\phi}_{\text{trim}}(0.01)$	-0.20	0.16	0.25	-0.52	0.26	0.58	-0.14	0.08	0.16	-0.35	0.13	0.37
MAR G-comp.	1.00	0.04	1.00	1.36	0.06	1.37	0.84	0.04	0.84	1.10	0.06	1.10
Heckman	0.63	2.27	2.36	0.25	2.07	2.09	0.41	5.04	5.06	-0.18	2.85	2.86
true CACE		0.64			1.45			0.48			1.22	
mean response		0.63			0.43			0.69			0.50	

Note: Results are based on 1000 simulations and 5000 observations.

close to zero, which may imply arbitrarily large weights for some observations. Propensity score trimming is discussed e.g. in Frölich (2004), Heckman, Ichimura, and Todd (1997), Dehejia and Wahba (1999), Busso, DiNardo, and McCrary (2009), and Crump, Hotz, Imbens, and Mitnik (2009). Yet, a trimming rule that is optimal in the sense that it minimizes the mean square error of the estimator does not appear to be available in the literature. Here, we follow Huber, Lechner, and Wunsch (2013) and discard observations whose relative weights within subgroups defined by Z and D exceed a particular threshold. As trimming thresholds we consider relative weights of 15 and 1%, resulting in the trimmed estimators $\hat{\theta}_{\text{trim}}(0.15)$, $\hat{\phi}_{\text{trim}}(0.15)$, $\hat{\theta}_{\text{trim}}(0.01)$, $\hat{\phi}_{\text{trim}}(0.01)$. The appendix provides additional results for further trimming levels (10, 5, and 2%).

We also consider an estimator that controls for attrition under the assumption of MAR but ignores treatment endogeneity due to U_0, U_1, U_2, while controlling for confounding related to X_0. To be specific, we use the MLE-based G-computation procedure of Robins (1986), in which the outcomes and response processes are modeled parametrically by linear and logit specifications, respectively. The appendix also provides the results for estimation based on targeted MLE, see van der Laan and Rubin (2006), inverse probability weighting (see e.g. Horvitz and Thompson (1952) and Hirano, Imbens, and Ridder (2003)), and augmented IPW (AIPW) (as in Robins, Rotnitzky, and Zhao (1995) and Scharfstein, Rotnitzky, and Robins (1999)) which yield very similar results. Finally, parametric Heckman (1976) MLE estimation of sample selection models assuming jointly normally distributed unobserved terms in the response and the outcome equations is also considered. The latter estimator controls for X_0, D, and Z in the estimation of response and can therefore account for attrition related to unobservables if R_1 and R_2 are functions of Z and if Z does not have a direct effect on the outcomes conditional on X_0 and D. However, it does not allow for treatment endogeneity related to U_0, U_1, U_2 and additionally presumes treatment effects to be homogeneous.

Table 1 provides the bias, standard deviation and root mean squared error (rmse) of the various estimators in periods 1 and 2 under treatment endogeneity and conditional LI with $\gamma = 0.5$, and β, δ equal to zero. $\hat{\theta}$, which is consistent in this scenario, performs very well in terms of bias and rmse irrespective of the period, share of compliers and effect homogeneity or heterogeneity. In contrast, the naive approach, the MAR-based G-computation procedure not controlling for treatment endogeneity, and the Heckman estimator are severely biased in any specification. Also $\hat{\phi}$ (and its trimmed versions) and $\hat{\omega}$ are prone to non-negligible bias, even though the latter performs

Table 2: Simulation 2 - treatment endogeneity and MAR (with the instrument affecting response)

	Homogeneous effects											
	α=0.68, β=0.5, γ=0, δ=0, κ=0						α=1.35, β=0.5, γ=0, δ=0, κ=0					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.75	0.03	0.75	1.08	0.06	1.09	0.64	0.03	0.64	0.93	0.06	0.93
$\hat{\omega}$	0.21	0.09	0.23	0.47	0.12	0.48	0.12	0.05	0.13	0.28	0.07	0.29
$\hat{\theta}$	-0.37	41.34	41.34	307.31	9507.59	9512.56	-0.05	0.08	0.10	-0.40	4.80	4.82
$\hat{\theta}_{\text{trim}}(0.15)$	-0.93	5.08	5.17	1.76	32.97	33.02	-0.05	0.08	0.10	-0.49	0.64	0.81
$\hat{\theta}_{\text{trim}}(0.01)$	-0.86	4.30	4.38	2.10	41.21	41.26	-0.05	0.08	0.10	-0.43	0.84	0.95
$\hat{\phi}$	0.02	0.14	0.14	-0.09	0.23	0.25	0.01	0.07	0.07	-0.05	0.12	0.13
$\hat{\phi}_{\text{trim}}(0.15)$	0.02	0.14	0.14	-0.09	0.23	0.25	0.01	0.07	0.07	-0.05	0.12	0.13
$\hat{\phi}_{\text{trim}}(0.01)$	0.02	0.14	0.14	-0.09	0.23	0.25	0.01	0.07	0.07	-0.05	0.12	0.13
MAR G-comp.	0.76	0.03	0.77	1.11	0.05	1.11	0.65	0.03	0.65	0.95	0.06	0.95
Heckman	0.86	0.04	0.86	1.30	0.06	1.30	0.83	0.04	0.83	1.26	0.09	1.26
true CACE		0.50			1.25			0.50			1.25	
mean response		0.68			0.49			0.69			0.51	

	Heterogeneous effects											
	α=0.68, β=0.5, γ=0, δ=0, κ=0.5						α=1.35, β=0.5, γ=0, δ=0, κ=0.5					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	1.05	0.04	1.05	1.54	0.06	1.54	0.94	0.04	0.94	1.39	0.06	1.39
$\hat{\omega}$	0.29	0.10	0.31	0.65	0.13	0.67	0.18	0.06	0.19	0.43	0.09	0.44
$\hat{\theta}$	-0.28	41.34	41.34	307.47	9507.59	9512.56	0.01	0.09	0.09	-0.27	4.80	4.81
$\hat{\theta}_{\text{trim}}(0.15)$	-0.85	5.08	5.15	1.92	32.97	33.03	0.01	0.09	0.09	-0.36	0.65	0.74
$\hat{\theta}_{\text{trim}}(0.01)$	-0.77	4.23	4.37	2.26	41.21	41.27	0.01	0.09	0.09	-0.30	0.84	0.89
$\hat{\phi}$	0.02	0.15	0.15	-0.08	0.25	0.26	0.01	0.08	0.08	-0.04	0.13	0.14
$\hat{\phi}_{\text{trim}}(0.15)$	0.02	0.15	0.15	-0.08	0.25	0.26	0.01	0.08	0.08	-0.04	0.13	0.14
$\hat{\phi}_{\text{trim}}(0.01)$	0.02	0.15	0.15	-0.08	0.25	0.26	0.01	0.08	0.08	-0.04	0.13	0.14
MAR G-comp.	1.07	0.04	1.07	1.57	0.06	1.58	0.95	0.04	0.95	1.41	0.06	1.41
Heckman	1.21	0.04	1.21	1.88	0.07	1.88	1.19	0.04	1.19	2.02	0.44	2.06
true CACE		0.64			1.45			0.48			1.22	
mean response		0.68			0.49			0.69			0.51	

Note: Results are based on 1000 simulations and 5000 observations.

comparably well in the first time period. Note that trimming does neither affect $\hat{\theta}$, nor $\hat{\phi}$, implying that large relative weights do not occur.

In the second simulation (Table 2), $\gamma = 0$ such that the assumptions underlying $\hat{\phi}$ hold. At the same time $\beta = 0.5$, implying a direct effect of the instrument on the response process and a violation of Assumption 2 required for the consistency of $\hat{\theta}$. Hence, estimators based on Theorem 2 are consistent, whereas the assumptions for Theorem 1 are not met. As expected, $\hat{\phi}$ now dominates any other estimator with respect to bias and low rmse and is unchanged by trimming. The naive approach, $\hat{\theta}$, G-computation, the Heckman estimator, and (to a lesser extent) $\hat{\omega}$ are substantially biased in most cases. $\hat{\theta}$ performs particularly poorly under the smaller complier share ($\alpha = 0.68$) due to a large increase of the variance. Yet, already moderate trimming using the 15% threshold ($\hat{\theta}_{\text{trim}}(0.15)$) reduces the variance (and the rmse) considerably, even though it remains at comparably high levels. More trimming further decreases the rmse in the first period, but increases it in the second one. In the latter case, the rmse is relatively stable for 15% and 10%, but grows more strongly for 2% and 1%.

In the third simulation (Table 3) we consider a scenario where all estimators are inconsistent: γ is set to zero, while $\beta = 0.5$ and $\delta = 0.25$, implying that the instrument directly affects non-response, which in addition is also related to the outcomes of the current period. $\hat{\theta}$ and $\hat{\phi}$ are biased because they ignore attrition related to contemporaneous outcomes, while G-computation ignores both treatment endogeneity and attrition related to contemporaneous outcomes, and the Heckman estimator does not account for treatment endogeneity. Trimming again reduces the variance of $\hat{\theta}$ in several cases, but smaller threshold values tend to increase the rmse relative to larger thresholds when α=0.68. All in all, no method performs convincingly in this last set-up considered.

4 Application to a policy intervention in college

In this section, we apply our methods to data from the Student Achievement and Retention Project assessed in Angrist, Lang, and Oreopoulos (2009), a randomized program providing academic services and financial incentives to first year students at a Canadian campus which aimed at improving the academic performance. To this end, all students who entered in September 2005 and had a high school grade point average (GPA) lower than the upper quartile were randomly assigned either to one of three different treatments provided in the first year, namely academic support services, financial incentives, or both, or otherwise to a control group. The services

Table 3: Simulation 3 - treatment endogeneity and selection on current outcomes

	Homogeneous effects											
	α=0.68, β=0.5, γ=0, δ=0.25, κ=0						α=1.35, β=0.5, γ=0, δ=0.25, κ=0					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.68	0.03	0.68	0.85	0.05	0.85	0.57	0.03	0.57	0.67	0.06	0.67
$\hat{\omega}$	0.10	0.10	0.13	0.36	0.11	0.37	0.05	0.05	0.07	0.19	0.07	0.20
$\hat{\theta}$	-1.55	5.09	5.32	1.48	572.82	572.82	-0.19	0.08	0.21	-0.68	0.47	0.83
$\hat{\theta}_{\text{trim}}(0.15)$	-1.18	4.65	4.80	-1.21	20.30	20.34	-0.19	0.08	0.21	-0.67	0.24	0.71
$\hat{\theta}_{\text{trim}}(0.01)$	-1.21	4.90	5.05	-0.45	24.37	24.37	-0.19	0.08	0.21	-0.65	0.20	0.69
$\hat{\phi}$	-0.23	0.14	0.27	-0.59	0.22	0.63	-0.13	0.07	0.15	-0.36	0.11	0.37
$\hat{\phi}_{\text{trim}}(0.15)$	-0.23	0.14	0.27	-0.59	0.22	0.63	-0.13	0.07	0.15	-0.36	0.11	0.37
$\hat{\phi}_{\text{trim}}(0.01)$	-0.23	0.14	0.27	-0.59	0.22	0.63	-0.13	0.07	0.15	-0.36	0.11	0.37
MAR G-comp.	0.70	0.03	0.70	0.89	0.05	0.89	0.58	0.03	0.58	0.70	0.05	0.70
Heckman	0.87	0.04	0.88	1.23	0.09	1.23	0.84	0.07	0.84	2.11	1.02	2.34
true CACE		0.50			1.25			0.50			1.25	
mean response		0.71			0.55			0.72			0.57	
	Heterogeneous effects											
	α=0.68, β=0.5, γ=0, δ=0.25, κ=0.5						α=1.35, β=0.5, γ=0, δ=0.25, κ=0.5					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	1.00	0.03	1.00	1.33	0.06	1.33	0.89	0.04	0.89	1.18	0.06	1.18
$\hat{\omega}$	0.13	0.11	0.17	0.43	0.14	0.45	0.09	0.06	0.11	0.28	0.09	0.29
$\hat{\theta}$	-1.51	5.09	5.31	1.55	572.82	572.82	-0.15	0.09	0.18	-0.60	0.47	0.76
$\hat{\theta}_{\text{trim}}(0.15)$	-1.14	4.65	4.79	-1.13	20.30	20.33	-0.15	0.09	0.18	-0.59	0.25	0.64
$\hat{\theta}_{\text{trim}}(0.01)$	-1.16	4.90	5.04	-0.38	24.37	24.37	-0.15	0.09	0.18	-0.57	0.21	0.61
$\hat{\phi}$	-0.22	0.15	0.27	-0.56	0.24	0.61	-0.11	0.08	0.13	-0.29	0.12	0.31
$\hat{\phi}_{\text{trim}}(0.15)$	-0.22	0.15	0.27	-0.56	0.24	0.61	-0.11	0.08	0.13	-0.29	0.12	0.31
$\hat{\phi}_{\text{trim}}(0.01)$	-0.22	0.15	0.27	-0.56	0.24	0.61	-0.11	0.08	0.13	-0.29	0.12	0.31
MAR G-comp.	1.02	0.03	1.02	1.37	0.06	1.37	0.90	0.04	0.90	1.21	0.06	1.21
Heckman	1.30	0.04	1.30	1.95	0.27	1.97	1.67	0.57	1.77	2.65	1.27	2.94
true CACE		0.64			1.45			0.48			1.22	
mean response		0.72			0.56			0.73			0.59	

Note: Results are based on 1000 simulations and 5000 observations.

contained both access to peer advisors, i.e., trained upper-class students supposed to provide academic support, and class-specific sessions targeted at improving study habits without focusing on specific course content. The financial incentives consisted of cash payments between 1,000 and 5,000 dollars that were conditional on attaining particular GPA targets in college, where the targets were a function of the high school GPA.

While the intervention appeared to be generally ineffective for males, Angrist, Lang, and Oreopoulos (2009) found positive effects of the combined treatment (academic support and financial incentives) on the college performance of females in the first and second year. For this reason, we will only focus on the subsample of 948 female students in the subsequent discussion. As the number of observations assigned to a particular treatment arm is rather low, we aggregate the academic services and financial incentives to a binary treatment that takes the value one if any form of intervention took place and zero otherwise in order to avoid small sample problems. For the same reason, we use (parametric) probit regressions (rather than nonparametric methods) to estimate the conditional probabilities involved in the identification results, which entails semiparametric estimators of the CACE. Inference is based on the bootstrap.

Albeit treatment assignment was random, identification may be flawed by both endogeneity and attrition. The endogeneity issue stems from the fact that only 274 (or 73%) of the 374 students who were offered any treatment actually signed up for it, which gives rise to potential selection bias into treatment. Furthermore, GPA scores, one of the outcomes measuring college success, are not observed for all students. Whereas they are missing for only 56 students (or 6%) in the first year, non-response amounts to a non-negligible 169 (or 18%) in the second year. If attrition is selective so that e.g. the probability to drop out decreased in both the treatment state and unobserved ability, the treatment effect is biased due to positive selection into observed GPA scores. Angrist, Lang, and Oreopoulos (2009) use instrumental variable estimation to control for endogeneity, where the random assignment indicator serves as instrument. They, however, do not correct for attrition in the GPA outcomes, but merely base their analysis on all those observations without missing GPAs, see the note underneath Table 6 in their paper. Here, we apply the methods outlined in Sections 2.2 and 2.3 to control for both endogeneity *and* attrition.

We are interested in the effect of having signed up for any of the three treatments ($D = 1$) vs. no treatment ($D = 0$) on the GPA scores at the end of the first and second year. We estimate the CACE based on Theorem 1 to allow attrition to be related to the latent types, as compliers with the treatment assignment may be more motivated to stay in college than the never takers, whose reluctance to take the treatment even when offered may be associated with a higher inclination to drop out of college. This motivates our higher confidence in Assumption 1 rather than the stronger Assumption 1' (which does not permit LI conditional on observables). At the same time, it seems likely that mere assignment does not affect the drop out decision of never takers, who would not take advantage of the treatment anyway. We therefore suspect Assumption 2 to be satisfied, albeit somewhat stronger than Assumption 2'. Nevertheless, we also consider estimation based on Theorem 2 imposing MAR given the observed variables and the treatment, which allows checking the sensitivity of the results to the presumed form of attrition. If one obtains similar results under both methods, this may imply that (the respective stronger assumption of) both sets of assumptions are satisfied, i.e. Assumption 1' and Assumption 2. We use both untrimmed and trimmed versions of the respective estimators. As in the simulations, trimming discards observations whose relative weights in subgroups defined by Z and D exceed a certain threshold, which is set to 10% in the application.

The data set contains a range of pre-treatment variables measuring performance and ambition as well as socioeconomic characteristics that allow us to model the response process in the first year. E.g., we observe the GPA score in high school, the fall grade of the first year, and the attempted maths and science courses, which are most likely correlated with both GPA scores in later periods and the probability to drop out. Indeed, the empirical relevance of academic performance in high school and in the first semester of college as a predictor for attrition is well documented in the literature on higher education, see e.g. Leppel (2002), Herzog (2005), and Tinto (1997). Furthermore, the data includes self-assessed measures of effort and ambition, e.g., whether the student wants to finish in four years, or strives for a higher degree than a BA. Learning habits are reflected by the information on how often a student leaves studying until the last minute. The data also comprises important characteristics reflecting the socioeconomic background, such as age, parents' education, and indicators for living at home and English mother tongue. Finally, it contains dummies for whether the student is at the first choice college and whether she completed the base line survey which may be correlated with the likelihood to be observed in later periods.

Table 4 gives the results of a probit regression of first year response on the baseline covariates X_0 and the treatment indicator D in order to estimate $\Xi_0 = \Pr(R_1 = 1|X_0, D = 1)$. The main specification (1) contains all regressors, and is used for the estimation of the CACE. Specification

Table 4: Probit coefficients and marginal effects of the model for 1st year response

	Coefficients			Marginal effects		
	(1)	(2)	(3)	(1)	(2)	(3)
Constant	5.495	6.196	1.445			
	(2.896)	(2.721)	(0.072)			
Treatment D	0.335	0.529	0.571	0.003	0.004	0.052
	(0.265)	(0.262)	(0.184)	(0.002)	(0.003)	(0.013)
High school GPA $\leq$ 75.2	-0.574	-0.503		-0.008	-0.007	
	(0.312)	(0.300)		(0.008)	(0.007)	
High school GPA	-0.078	-0.076		-0.001	-0.001	
	(0.036)	(0.034)		(0.001)	(0.001)	
Fall grade	0.028	0.029		0.000	0.000	
	(0.003)	(0.003)		(0.000)	(0.000)	
Attempted math/science credits	0.878	0.970		0.008	0.009	
	(0.239)	(0.247)		(0.004)	(0.005)	
Wants more than B.A.	0.235			0.002		
	(0.223)			(0.003)		
Last minute learning (usual/often)	-0.054			-0.000		
	(0.244)			(0.002)		
Age < 20	0.311			0.004		
	(0.396)			(0.008)		
Father has college degree	-0.129			-0.001		
	(0.230)			(0.002)		
At first choice college	0.208			0.002		
	(0.258)			(0.002)		
Completed baseline survey	0.778			0.018		
	(0.269)			(0.016)		
Pseudo R^2	0.494	0.452	0.027	0.494	0.452	0.027

Note: (1)-(3) give the probit coefficients and marginal effects, respectively, when estimating $\Xi_0 = \Pr(R_1 = 1|X_0, D = 1)$ under different specifications: (1) is the main specification with all regressors, (2) contains a subset of regressors, (3) contains only D and a constant. The marginal effects are evaluated at the means of all other regressors. Standard errors are given in brackets. The sample size is 948.

(2) presents a more parsimonious model consisting of D and pre-treatment outcomes (high school GPA, fall grade, attempted math/science credits). Finally, specification (3) only contains D and a constant as regressors. Comparing the results for the different specifications, we find that the pre-treatment outcomes and the dummy for survey completed clearly have the highest predictive power, whereas socioeconomic variables are less important.

For modeling response in the second period, we use in addition to the covariates X_0 of specification (1) three intermediate outcomes (X_1) at the end of the first year: the GPA as well as the number of credits earned in the first year and an indicator for good standing, all of which are highly correlated with response in the second year.

Table 5 provides descriptive statistics (means and standard deviations) of the variables used in our analysis for all females, as well as for subsamples with $D = 1$ and $D = 0$. The variables measured after the first or second year are only observed if $R_1 = 1$ and $R_2 = 1$, respectively. Note that treated females have on average higher pre-treatment outcomes (high school GPA and fall grade) and higher aspirations (wanting more than a B.A.) than the non-treated. This points to selectivity and motivates the use of random treatment assignment Z as an instrument for actual treatment take-up D.

Table 5: Descriptive statistics

	total sample (948 obs.)		$D = 1$ (274 obs.)		$D = 0$ (674 obs.)	
Regressor	mean	std. dev	mean	std. dev	mean	std. dev
High school GPA (multi-valued)	78.88	4.29	79.10	4.30	78.80	4.28
Fall grade (multi-valued)	53.69	25.71	58.55	23.04	51.71	26.48
Attempted math/science credits (multi-valued)	1.00	1.16	1.05	1.19	0.97	1.15
Wants more than B.A. (binary)	0.52	0.50	0.58	0.49	0.49	0.50
Last minute learning (binary)	0.28	0.45	0.30	0.46	0.28	0.45
At first choice college (binary)	0.24	0.43	0.26	0.44	0.23	0.42
Age $<$ 20 (binary)	0.97	0.17	0.99	0.12	0.97	0.18
Father has college degree (binary)	0.37	0.48	0.40	0.49	0.36	0.48
Completed baseline survey (binary)	0.90	0.30	0.95	0.23	0.89	0.32
First year response R_1 (binary)	0.94	0.24	0.98	0.15	0.93	0.26
First year GPA Y_1 (multi-valued)	1.76	0.90	1.81	0.88	1.74	0.91
First year good standing for $R_1 = 1$ (binary)	0.48	0.50	0.54	0.50	0.46	0.50
First year credits earned for $R_1 = 1$ (multi-valued)	2.36	0.93	2.47	0.94	2.32	0.92
Second year response R_2 (binary)	0.82	0.38	0.83	0.37	0.82	0.39
Second year GPA Y_2 for $R_2 = 1$ (multi-valued)	2.07	0.87	2.19	0.86	2.01	0.87

Note: Descriptive statistics for baseline covariates X_0, response indicators R_1 and R_2 and outcomes Y_1 and Y_2, if observed.

Table 6 presents the estimated treatment effects of the intervention on the GPA one and two years later. The top panel provides the estimates for the full sample. The subsequent panels show the results for various subsamples defined by age and parental background. In each panel, the first line gives the CACE estimates, the second line the bootstrap standard errors, and the third line the bootstrap p-values based on the quantiles of the resampled distribution of the CACE estimates, see equation (6) in MacKinnon (2006). We provide the quantile-based p-values (rather than those based on the t-statistic) to account for the problem that in finite samples the moments of instrumental variable estimators may not exist such that t-statistics may be misleading, which might even be aggravated by attrition. The first and sixth columns labelled "Wald" show the Wald estimates, i.e. the instrumental variable estimator without any covariates. The estimates based on Theorem 1 are denoted by $\hat{\theta}$ and $\hat{\theta}_{\text{trim}}$, where the latter represents the trimmed version. The estimates based on Theorem 2 are denoted by $\hat{\phi}$ and $\hat{\phi}_{\text{trim}}$. We find that large weights rarely occur such that the trimmed and untrimmed point estimates are always very similar, if not the same. Note, however, that trimming reduces the standard errors of the estimates based on Theorem 1 by disciplining outliers in the bootstrap samples.

Both $\hat{\theta}$ and $\hat{\theta}_{\text{trim}}$ are nevertheless less precise than $\hat{\phi}$ and $\hat{\phi}_{\text{trim}}$. We would generally (and specifically in moderate samples) expect this to be the case at least if both theorems are (closely) satisfied, because Theorem 1 contains more conditional probabilities to be estimated, e.g. $P'_0 - \pi$ and $P'_1 - \pi$ in the denominator, which may potentially decrease precision in small samples. In particular, if the latter differences are small (which likely occurs if Z only weakly shifts D so that few compliers exist) the variance might be large. Furthermore, in the current application, $\hat{\phi}$ and $\hat{\phi}_{\text{trim}}$ appear to rest on stronger assumptions than $\hat{\theta}$ and $\hat{\theta}_{\text{trim}}$, which again suggests lower standard errors of the former: Whereas we argued in Section 2 that Assumption 2' is generally weaker than Assumption 2, they are, however, very similar in the application at hand. This is because Assumption 2 only restricts the response process in time period 1, where we have in fact very little non-response. (Non-response is larger in time period 2, but this does not enter Assumption 2.) On the other hand, Assumption 1' is clearly much stronger than Assumption 1. The former imposes independence within each stratum defined by Z and D (and other predetermined observables), whereas the latter additionally requires conditioning on the (unobserved) type. Therefore, estimators based on Assumption 1' exploit more restrictions and can (figuratively speaking) use coarser strata with more information than methods relying on Assumption 1, which have to operate within finer strata additionally defined upon the type. Therefore, $\hat{\phi}$ and $\hat{\phi}_{\text{trim}}$ can exploit more information.

Examining first the estimates for the whole population, we do not find any significant effects in the first year. In contrast, the simple Wald estimates for the second year are significant (at the 5% level) and suggest that the GPA of compliers increases by 0.164 points. However, when using the attrition corrected estimators, the effect shrinks considerably to 0.077 or 0.071, respectively, and becomes insignificant. Therefore, our results suggest that attrition, if ignored, may lead to an overestimation of the effects in education experiments.

In the remainder of Table 6, we investigate effect heterogeneity for subsamples stratified by age, prior academic achievement and parental background. E.g., we separately consider students in the lower and the upper half of the high school GPA distribution (median: 78.5 points) to see whether high or low achievers particularly benefit from the intervention. Indeed, the Wald estimate for the second year GPA of low achievers amounts to 0.225 points, indicating that the less capable students benefit most when taking advantage of the services and incentives. However, after controlling for attrition, the effect becomes much smaller and insignificant, irrespective of trimming. When we split the sample by age groups (17 & 18 years versus older than 18), we also cannot draw reliable conclusions as the estimates are generally rather noisy.

Finally, we examine whether the effects differ by parents' education, which might be regarded as a proxy for family background. Interestingly, the second year Wald estimate in the subsample with mothers that have a college degree is negative and large. When controlling for attrition, the estimate shrinks in magnitude (in the case of $\hat{\theta}$, $\hat{\theta}_{\text{trim}}$ quite considerably) and becomes even less significant. In contrast, for those students whose mother has no degree, the Wald estimate is significantly positive (at the 5% level) in both periods. Furthermore, correcting for attrition does *not* substantially reduce the estimate in the second year, even though the precision decreases. The estimates $\hat{\phi}$ and $\hat{\phi}_{\text{trim}}$ remain significant at the 5% level. A similar pattern appears when stratifying on the father's degree status. While the Wald estimate in the second year is insignificant in the subpopulation with fathers having a degree, it is large and significant in the subsample without college degree. Furthermore, the effect is almost the same when using $\hat{\theta}$, albeit less precisely estimated, and $\hat{\theta}_{\text{trim}}$, $\hat{\phi}$, $\hat{\phi}_{\text{trim}}$ are significant at the 10% and 5% levels, respectively.

In summary, our findings suggest that the empirical evidence about the effectiveness of the intervention considered by Angrist, Lang, and Oreopoulos (2009) is much weaker once attrition is acknowledged. Nevertheless, female students with a less favorable family background seem to gain

Table 6: Effectiveness of the school intervention on GPA: all females and subsamples

	1st year effect of intervention on GPA					2nd year effect of intervention on GPA				
	Wald estimate (no covariates)	$\hat{\theta}$ (Theorem 1)	$\hat{\theta}_{\text{trim}}$	$\hat{\phi}$ (Theorem 2)	$\hat{\phi}_{\text{trim}}$	Wald estimate (no covariates)	$\hat{\theta}$ (Theorem 1)	$\hat{\theta}_{\text{trim}}$	$\hat{\phi}$ (Theorem 2)	$\hat{\phi}_{\text{trim}}$
	Full sample: all females (948 obs.)									
effect	0.074	0.022	0.022	-0.047	-0.047	0.164	0.077	0.077	0.071	0.071
s.e.	0.079	0.291	0.129	0.075	0.076	0.083	8.694	0.214	0.090	0.093
p-val	0.399	0.709	0.679	0.517	0.577	0.040	0.817	0.770	0.419	0.340
	Subsample: high school GPA $\leq$ 78.5 (467 obs.)									
effect	0.156	0.107	0.107	0.057	0.057	0.225	0.143	0.143	0.043	0.043
s.e.	0.110	2.340	0.274	0.112	0.112	0.128	10.937	0.378	0.130	0.131
p-val	0.149	0.502	0.456	0.595	0.565	0.069	0.451	0.382	0.698	0.669
	Subsample: high school GPA $>$ 78.5 (481 obs.)									
effect	0.023	-0.093	-0.093	-0.170	-0.170	0.099	0.036	-0.066	-0.001	-0.122
s.e.	0.106	36.995	0.382	0.135	0.128	0.098	24.073	0.968	0.180	0.156
p-val	0.812	0.668	0.712	0.220	0.382	0.300	0.941	0.670	0.816	0.456
	Subsample: 17 and 18 years old (741 obs.)									
effect	0.042	-0.024	-0.024	-0.050	-0.050	0.132	0.093	0.093	0.043	0.043
s.e.	0.090	2.472	0.310	0.089	0.088	0.092	28.780	0.322	0.103	0.103
p-val	0.632	0.816	0.848	0.547	0.556	0.145	0.573	0.469	0.680	0.668
	Subsample: 19-23 years old (207 obs.)									
effect	0.131	0.064	0.002	-0.051	-0.059	0.226	0.275	0.171	0.125	-0.021
s.e.	0.192	1530.530	0.627	0.194	0.195	0.218	733.224	0.889	0.379	0.306
p-val	0.484	0.822	0.776	0.772	0.939	0.259	0.615	0.448	0.656	0.703
	Subsample: mother has college degree (304 obs.)									
effect	-0.178	-0.171	-0.249	-0.201	-0.201	-0.253	-0.085	-0.197	-0.211	-0.211
s.e.	0.143	2.745	0.329	0.167	0.135	0.155	64.916	0.481	0.193	0.157
p-val	0.212	0.388	0.148	0.157	0.224	0.105	0.936	0.675	0.208	0.281
	Subsample: mother has no college degree (644 obs.)									
effect	0.197	0.135	0.135	0.080	0.080	0.383	0.345	0.345	0.259	0.259
s.e.	0.097	5.960	0.098	0.111	0.098	0.099	5.095	0.271	0.155	0.131
p-val	0.034	0.130	0.117	0.366	0.290	0.000	0.147	0.096	0.050	0.018
	Subsample: father has college degree (355 obs.)									
effect	0.035	-0.029	-0.029	-0.038	-0.038	-0.078	0.065	0.065	-0.127	-0.127
s.e.	0.133	58.422	0.216	0.187	0.123	0.141	37.131	0.406	0.206	0.134
p-val	0.827	0.839	0.886	0.717	0.833	0.566	0.457	0.212	0.342	0.443
	Subsample: father has no college degree (593 obs.)									
effect	0.092	0.066	0.066	-0.008	-0.008	0.317	0.307	0.307	0.250	0.250
s.e.	0.101	8.636	0.133	0.099	0.101	0.108	6.418	0.225	0.140	0.143
p-val	0.371	0.491	0.479	0.979	0.904	0.002	0.150	0.085	0.063	0.029

Note: Treatment effects of the intervention (support and/or financial services) on GPA outcomes one and two years later, respectively. The top panel displays the results for the full sample (on the left, the effect after one year; on the right, the effect after two years). The subsequent panels show estimates for subpopulations stratified by age, parental background, and prior academic achievement. P-values are given in brackets and are based on 1999 bootstrap replications. Trimming in $\hat{\theta}_{\text{trim}}$ and $\hat{\phi}_{\text{trim}}$ is based on dropping observations that have a relative weight larger than 10%.

from the services and financial incentives.

5 Conclusions

In this paper, we proposed a novel approach for the identification and estimation of local average treatment effects in multiple outcome periods which controls for both treatment endogeneity and outcome attrition. We showed how pre-treatment information can be combined with intermediate outcomes in order to correct more plausibly for non-response bias in later periods, while an instrument was used to tackle endogenous treatment selection. Two sets of identifying assumptions were presented. The first one, which we call conditional latent ignorability, permits attrition to depend on observables and the latent treatment compliance type, which may be related to unobservables. The second one imposes randomness given observed variables only, which amounts to a dynamic missing at random assumption. The proposed methods were applied to a policy intervention aimed at increasing academic performance in college, where ignoring attrition was found to lead to upwardly biased estimates.

References

Abadie, A. (2003): "Semiparametric instrumental Variable estimation of treatment response models," *Journal of Econometrics*, 113, 231–263.

Abowd, J., B. Crepon, and F. Kramarz (2001): "Moment Estimation With Attrition: An Application to Economic Models," *Journal of the American Statistical Association*, 96, 1223–1230.

Angrist, J., E. Bettinger, and M. Kremer (2006): "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia," *American Economic Review*, 96(3), 847–862.

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of Causal Effects using Instrumental Variables," *Journal of American Statistical Association*, 91, 444–472 (with discussion).

Angrist, J., D. Lang, and P. Oreopoulos (2009): "Incentives and Services for College Achievement: Evidence from a Randomized Trial," *American Economic Journal: Applied Economics*, 1, 136–163.

Barnard, J., C. Frangakis, J. Hill, and D. Rubin (2003): "A Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City," *Journal of the American Statistical Association*, 98, 299–323.

Bertrand, M., and S. Mullainathan (2004): "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *The American Economic Review*, 94, 991–1013.

Bollinger, C., and M. David (2001): "Estimation With Response Error and Nonresponse: Food-Stamp Participation in the SIPP," *Journal of Business and Economic Statistics*, 19, 129–141.

Busso, M., J. DiNardo, and J. McCrary (2009): "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," *mimeo.*

Carroll, R., D. Ruppert, and L. Stefanski (1995): *Measurement Error in Nonlinear Models.* Chapman and Hall, London.

Chen, P., E. Wong, R. Dominik, and M. Steiner (2000): "A transitional model of barrier methods compliance with unbalanced loss to follow-up," *Statistics in Medicine*, 19, 71–82.

Crump, R., J. Hotz, G. Imbens, and O. Mitnik (2009): "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96, 187–199.

Das, M., W. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

Dehejia, R., and S. Wahba (1999): "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programmes," *Journal of American Statistical Association*, 94, 1053–1062.

DiNardo, J., J. McCrary, and L. Sanbonmatsu (2006): "Constructive Proposals for Dealing with Attrition: An Empirical Example," *Working paper, University of Michigan.*

Ding, W., and S. Lehrer (2010): "Estimating Treatment Effects from Contaminated Multi-period Education Experiments: The Dynamic Impacts of Class Size Reductions," *Review of Economics and Statistics*, 92, 31–42.

Fitzgerald, J., P. Gottschalk, and R. Moffitt (1998): "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics," *Journal of Human Resources*, 33, 251–299.

Frangakis, C., R. Brookmeyer, R. Varadhan, M. Safaeian, D. Vlahov, and S. Strathdee (2004): "Methodology for Evaluating a Partially Controlled Longitudinal Treatment Using Principal Stratification, With Application to a Needle Exchange Program," *Journal of the American Statistical Association*, 99, 239–249.

Frangakis, C., and D. Rubin (1999): "Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes," *Biometrika*, 86, 365–379.

Frumento, P., F. Mealli, B. Pacini, and D. B. Rubin (2012): "Evaluating the Effect of Training on Wages in the Presence of Noncompliance, Nonemployment, and Missing Outcome Data," *Journal of the American Statistical Association*, 107, 450–466.

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

——— (2007): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75.

——— (2008): "Parametric and Nonparametric Regression in the Presence of Endogenous Control Variables," *International Statistical Review*, 76, 214–227.

Hausman, J., and D. Wise (1979): "Attrition Bias In Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, 47(2), 455–473.

Heckman, J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

Heckman, J., H. Ichimura, and P. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.

Heitjan, D., and S. Basu (1996): "Distinguishing Missing at Random and Missing Completely at Random," *The American Statistician*, 50, 207–213.

Herzog, S. (2005): "Measuring Determinants of Student Return vs. Dropout/Stopout vs. Transfer: A First-to-Second Year Analysis of New Freshmen," *Research in Higher Education*, 46, 883–928.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Hirano, K., G. W. Imbens, G. Ridder, and D. B. Rubin (2001): "Combining Panel Data Sets with Attrition and Refreshment Samples," *Econometrica*, 69, 1645–1659.

Horvitz, D., and D. Thompson (1952): "A Generalization of Sampling without Replacement from a Finite Population," *Journal of American Statistical Association*, 47, 663–685.

Huber, M. (2012): "Identification of average treatment effects in social experiments under alternative forms of attrition," *Journal of Educational and Behavioral Statistics*, 37, 443–474.

——— (2013): "Treatment evaluation in the presence of sample selection," *forthcoming in Econometric Reviews.*

Huber, M., M. Lechner, and C. Wunsch (2013): "The performance of estimators based on the propensity score," *Journal of Econometrics*, 175, 1–21.

Huber, M., and G. Mellace (2013): "Testing instrument validity for LATE identification based on inequality moment constraints," *forthcoming in the Review of Economics and Statistics.*

Imai, K. (2008): "Sharp bounds on the causal effects in randomized experiments with 'truncation-by-death'," *Statistics & Probability Letters*, 78, 144–149.

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

Imbens, G., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Kitagawa, T. (2013): "A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model," *UCL working paper.*

Kyriazidou, E. (1997): "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 1335–1364.

——— (2001): "Estimation of Dynamic Panel Data Sample Selection Models," *The Review of Economic Studies*, 68, 543–572.

Lechner, M. (1999): "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification," *Journal of Business and Economic Statistics*, 17, 74–90.

Lechner, M., R. Miquel, and C. Wunsch (2011): "Long-Run Effects of Public Sector Sponsored Training in West Germany," *Journal of the European Economic Association*, 9, 742–784.

Lee, D. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Leppel, K. (2002): "Similarities and differences in the college persistence of men and women," *Review of Higher Education*, 25, 433–450.

Little, R., and D. Rubin (1987): *Statistical Analysis with Missing Data*. Wiley, New York.

Little, R. J. A. (1995): "Modeling the Drop-Out Mechanism in Repeated-Measures Studies," *Journal of the American Statistical Association*, 90, 1112–1121.

Lok, J., R. Gill, A. van der Vaart, and J. Robins (2004): "Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models," *Statistica Neerlandica*, 58, 271–295.

MacKinnon, J. G. (2006): "Bootstrap Methods in Econometrics," *The Economic Record*, 82, S2–S18.

Manski, C. (1989): "Anatomy of the Selection Problem," *Journal of Human Resources*, 24, 343–360.

——— (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review, Papers and Proceedings*, 80, 319–323.

Mattei, A., and F. Mealli (2007): "Application of the Principal Stratification Approach to the Faenza Randomized Experiment on Breast Self-Examination," *Biometrics*, 63, 437–446.

Mealli, F., G. Imbens, S. Ferro, and A. Biggeri (2004): "Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes," *Biostatistics*, 5, 207–222.

Murphy, S., M. van der Laan, and J. Robins (2001): "Marginal mean models for dynamic regimes," *Journal of the American Statistical Association*, 96, 1410–1423.

NEWEY, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

PENG, Y., R. J. A. LITTLE, AND T. E. RAGHUNATHAN (2004): "An Extended General Location Model for Causal Inferences from Data Subject to Noncompliance and Missing Values," *Biometrics*, 60, 598–607.

POWELL, J., J. STOCK, AND T. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.

PREISSER, J. S., A. T. GALECKI, K. K. LOHMAN, AND L. E. WAGENKNECHT (2000): "Analysis of Smoking Trends with Incomplete Longitudinal Binary Responses," *Journal of the American Statistical Association*, 95, 1021–1031.

RACINE, J., AND Q. LI (2004): "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data," *Journal of Econometrics*, 119, 99–130.

ROBINS, J. (1989): "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS*, ed. by L. Sechrest, H. Freeman, and A. Mulley, pp. 113–159. U.S. Public Health Service, Washington, DC.

ROBINS, J., S. GREENLAND, AND F.-C. HU (1999): "Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome," *Journal of the American Statistical Association*, 94, 687–700.

ROBINS, J., A. ROTNITZKY, AND L. ZHAO (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of American Statistical Association*, 90, 106–121.

ROBINS, J. M. (1986): "A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect," *Mathematical Modelling*, 7, 1393–1512.

ROBINS, J. M., A. ROTNITZKY, AND L. ZHAO (1994): "Estimation of Regression Coefficients When Some Regressors Are not Always Observed," *Journal of the American Statistical Association*, 90, 846–866.

ROSENBAUM, P., AND D. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

ROTNITZKY, A., J. ROBINS, AND D. SCHARFSTEIN (1998): "Semiparametric Regression for Repeated Outcomes With Nonignorable Nonresponse," *Journal of the American Statistical Association*, 93, 1321–1339.

RUBIN, D. (1976): "Inference and Missing Data," *Biometrika*, 63, 581–592.

——— (1977): "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–543.

——— (1978): "Multiple Imputations in Sample Surveys-A Phenomenological Bayesian Approach to Nonresponse," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 20–34.

Scharfstein, D., A. Rotnitzky, and J. Robins (1999): "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94, 1096–1120.

Semykina, A., and J. Wooldridge (2006): "Estimating Panel Data Models in the Presence of Endogeneity and Selection: Theory and Application," *unpublished manuscript*.

Serfling, R. (1980): *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Shah, A., N. Laird, and D. Schoenfeld (1997): "A Random-Effects Model for Multiple Characteristics With Possibly Missing Data," *Journal of the American Statistical Association*, 92, 775–779.

Shepherd, B. E., M. W. Redman, and D. P. Ankerst (2008): "Does Finasteride Affect the Severity of Prostate Cancer? A Causal Sensitivity Analysis," *Journal of the American Statistical Association*, 103, 1392–1404.

Tan, Z. (2006): "Regression and Weighting Methods for Causal Inference Using Instrumental Variables," *Journal of the American Statistical Association*, 101, 1607–1618.

Tinto, V. (1997): "Classroom as communities: Exploring the educational character of student persistence," *Journal of Higher Education*, 68, 599–623.

van der Laan, M., and D. Rubin (2006): "Targeted Maximum Likelihood Learning," *The International Journal of Biostatistics*, 2.

Wang, L., A. Rotnitzky, X. Lin, R. E. Millikan, and P. F. Thall (2012): "Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer," *Journal of the American Statistical Association*, 107, 493–508.

Xie, H., and Y. Qian (2012): "Measuring the impact of nonignorability in panel data with non-monotone nonresponse," *Journal of Applied Econometrics*, 27, 129–159.

Yau, L., and R. Little (2001): "Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed," *Journal of American Statistical Association*, 96, 1232–1244.

Zhang, J., and D. Rubin (2003): "Estimation of causal effects via principal stratification when some outcome are truncated by death," *Journal of Educational and Behavioral Statistics*, 28, 353–368.

Zhang, J., D. Rubin, and F. Mealli (2008): "Evaluating The Effects of Job Training Programs on Wages through Principal Stratification," in *Advances in Econometrics: Modelling and Evaluating Treatment Effects in Econometrics*, ed. by D. Millimet, J. Smith, and E. Vytlacil, vol. 21, pp. 117–145. Elsevier Science Ltd.

——— (2009): "Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification," *Journal of the American Statistical Association*, 104, 166–176.

Supplementary appendix

This appendix is not intended to be published.
It will be made available on the authors' homepages.

It is included here for the convenience of the reviewers in case they want to consult this Supplementary Appendix.

Supplementary appendix

A Relation between observed D, Z and latent types

This section discusses the relation between observed values of D, Z and the latent types. D_i^z denotes the potential treatment status of individual i if Z_i were hypothetically set to z. As discussed in Angrist, Imbens, and Rubin (1996) and summarized in Table 7, the population can be categorized into four types (denoted by $T \in \{a, c, d, n\}$), depending on how the treatment status changes with the instrument. The compliers react on the instrument in the intended way by taking the treatment when $Z = 1$ and abstaining from it when $Z = 0$. For the remaining three types, $D_i^z \neq z$ for either $Z_i = 1$, or $Z_i = 0$, or both: The always takers are always treated irrespective of the instrument state, the never takers are never treated, and the defiers only take the treatment when $Z_i = 0$. Note that any *observed* combination of the treatment and the instrument does not uniquely determine the type, because it provides us only with either D_i^1 or D_i^0, whereas the other potential treatment status remains unobserved. Hence, while the type of an individual is uniquely defined, it cannot be inferred for each individual from the observed values of D and Z alone, i.e. while we observe $D_i^{Z_i}$, we cannot observe $D_i^{1-Z_i}$. From observing $D_i^{Z_i}$ we know for any observation that it belongs to one of two types, which is demonstrated in Table 8. The type is hence only partially known or *latent*.

Table 7: Types

Type	D_i^1	D_i^0	Notion
a	1	1	Always takers
c	1	0	Compliers
d	0	1	Defiers
n	0	0	Never takers

Table 8: Observed instrument/treatment and latent types

Observed values of instrument and treatment	latent types
$\{Z_i = 1, D_i = 1\}$	i belongs to either a or c
$\{Z_i = 1, D_i = 0\}$	i belongs to either d or n
$\{Z_i = 0, D_i = 1\}$	i belongs to either a or d
$\{Z_i = 0, D_i = 0\}$	i belongs to either c or n

However, if defiers ($\mathcal{T} = d$) can be plausibly ruled out, which corresponds to the (weak) monotonicity assumption of Imbens and Angrist (1994), the mean potential outcomes of the compliers ($\mathcal{T} = c$) may be identified despite the latent nature of $\mathcal{T}$. This follows from the fact that the second and third lines in Table 8 then only consist of n and a, respectively. Under particular assumptions, this in turn allows isolating the proportions and mean outcomes of c in the mixed groups given in the first line (under treatment) and the fourth line (under non-treatment). This intuition underlies the identification of the complier average causal effect (CACE), also known as local average treatment effect (LATE) on the compliers:

$$E[Y_t^1 - Y_t^0 | \mathcal{T} = c].$$

B Implications of the identifying assumptions in a parametric latent index framework

To gain more intuition about the identifying assumptions of Sections 2.2 and 2.3, we subsequently discuss their implications for a parametric benchmark model with latent index equations for the instrument, treatment, and response. For ease of exposition, we again only consider two outcome periods. Furthermore, for reasons of tractability, we assume that Y_t is the only time-varying regressor in the model so that $\underline{X}_1 = (Y_1, X_0)$ (even though we could have permitted other variables X_1 in $\underline{X}_1$ as well). We consider the following equations for the instrument and the treatment:

$$\begin{aligned} Z &= I\{\alpha X_0 > W\} \\ D &= I\{\gamma_0 X_0 + \gamma_1 Z > U_0\}, \end{aligned}$$

where $I\{\cdot\}$ denotes the indicator function (which is one if its argument is true and zero otherwise), X_0 are the baseline characteristics, W and U_0 are unobservables, and $\alpha, \gamma_0, \gamma_1$ are coefficients. Note that monotonicity (see Assumption 3) holds by construction in the latent index treatment model, unless γ_1 is zero. Assuming γ_1 to be positive, an individual is an always taker if $\gamma_0 X_0 > U_0$, a never taker if $\gamma_0 X_0 + \gamma_1 \leq U_0$, and a complier if $\gamma_0 X_0 \leq U_0 < \gamma_0 X_0 + \gamma_1$. To simplify notation, we define the following indicators for these types: $T_a = I\{\gamma_0 X_0 > U_0\}$, $T_n = I\{\gamma_0 X_0 + \gamma_1 \leq U_0\}$, and $T_c = I\{\gamma_0 X_0 \leq U_0 < \gamma_0 X_0 + \gamma_1\}$. I.e. conditional on X_0 there are three strata defined by the type. (For Assumption 3, we further require that $0 < \Pr(Z = 1 | X_0) < 1$, which is satisfied for instance if W has unbounded support.)

We examine the content of our identifying assumptions in the following linear outcome equations and latent index response models:

$$\begin{aligned} Y_1 &= \beta_1 D + \zeta_1 Z + \delta_1 X_0 + \kappa_{a1} T_a + \kappa_{n1} T_n + \kappa_{c1} T_c + U_1, \\ Y_2 &= \beta_2 D + \zeta_2 Z + \delta_2 X_0 + \kappa_{a2} T_a + \kappa_{n2} T_n + \kappa_{c2} T_c + U_2 + \lambda Y_1, \end{aligned}$$

and

$$\begin{aligned} R_1 &= I\{\theta_1 D + \rho_1 Z + \xi_1 X_0 + \tau_{a1} T_a + \tau_{n1} T_n + \tau_{c1} T_c > V_1\}, \\ R_2 &= I\{\theta_2 D + \rho_2 Z + \xi_2 X_0 + \tau_{a2} T_a + \tau_{n2} T_n + \tau_{c2} T_c + \phi Y_1 > V_2\}. \end{aligned}$$

Note that in this simple benchmark model, the treatment effect in the first period is β_1, whereas in the second period it is $\beta_2+\lambda\beta_1$, consisting of the direct effect in the second period plus the effect channeled through Y_1. The various κ and τ coefficients are type-specific effects on outcomes and responses, i.e. represent characteristics of the (unobserved) types. U_1, U_2, V_1, V_2 represent further unobservables that have not yet been captured by the types. (Note that we could have augmented the outcome equations by interaction terms between the types and D in order to incorporate effect heterogeneity of the intervention across types, as it is permitted by our assumptions and in most of the nonparametric literature on the CACE. For reasons of tractability, we examine the more parsimonious model here.)

Consider first the case without non-response, i.e. where $R_1 = R_2 = 1$ for everyone. The usual endogeneity problem arises through correlation between U_0 and U_1. Following Abadie (2003) we can estimate the CACE using the instrumental variable if $\{Y_t^d, \mathcal{T}\} \perp\!\!\!\perp Z | X_0$, which in our parametric model requires, first, that $\zeta_1 = \zeta_2 = 0$ (i.e. no direct effect of the instrument on the outcomes), and second, that W is independent of U_0, U_1 and U_2, given X_0.

We now examine the scenario with missing data and illustrate the identifying assumptions in the context of a schooling example. Consider an intensive remedial education intervention for poorly performing children during the summer vacations before the last year of *compulsory* schooling. Among the target population of weak students, some children receive a voucher for participation in this summer course (e.g. waiver of tuition fees and/or free lodging plus lodging allowance plus pocket money or other incentives to participate). The voucher is randomly assigned (where the assignment probability may depend on various baseline characteristics X_0) and thus serves as instrument Z. Participation in the summer course is denoted by D. We are interested in the effects of remedial summer education on cognitive skills at the end of *compulsory* schooling (Y_1) and at some later period Y_2, e.g. one year later or at the end of high school. Estimating the treatment effects at the end of *compulsory* schooling is less an issue because administrative student data e.g. on exam results are usually available for nearly all students, so that $R_1 = 1$ for a high share of individuals. However, after compulsory schooling many (weak) students may drop out of school, where drop-out is likely to depend on Y_1, e.g. the examination results at the end of compulsory schooling. This will hamper assessing medium-term treatment effects on Y_2. To those students who remain in school one can easily administer cognitive skills tests, whereas drop-outs will be more difficult to track for further follow-up data collection, such that $R_2 = 0$ for many (if not all) of them.

In this example, the coefficients κ and τ reflect type specific influences on cognitive skills and attrition related to unobservables such as motivation and willingness to pay. In particular, the compliers are those who only participate in summer vacation remedial education if the costs are heavily subsidized. On the other hand, the always and never takers do not care about the price. As the former participate in any case, we can think of this group as being most motivated and committed to complete school successfully. The never takers, on the other hand, are not willing to spend their summer vacations in school even if they would get pocket money and lodging allowance.

In the following discussion, we focus on time period 2, where the dynamic nature of the attrition problem is most clearly visible. For estimating the treatment effect on Y_2, Assumption 1 requires that U_2 is independent of V_2 conditional on the baseline characteristics, the instrument, the types, Y_1, and previous response. This is equivalent to assuming that

$$U_2 \perp\!\!\!\perp V_2 | U_1, V_1, \mathcal{T}, X_0, Z. \tag{5}$$

(This relationship is obtained by noting that $\mathcal{T}$ and Z jointly uniquely determine D.) Hence, Assumption 1 permits U_2 and V_2 to be arbitrarily related to other unobservables in the model, implying that Y_2 and R_2 may depend on past (endogenous) outcomes Y_1. However, conditional on what happened in prior periods, U_2 and V_2 must be independent, which rules out unobserved events in the second period that affect outcomes and responses simultaneously.

In the context of our remedial education intervention, predetermined non-response appears plausible if individuals decide to stay in or leave school at the end of compulsory schooling mostly based on the exam results Y_1, so that drop-out occurs shortly after the results are released, e.g. during or at the end of summer vacation after the last compulsory year of schooling. Then, drop-out *during* the next school year should be very low or non-existent (so that R_2 is expected to be one for (almost) all students who did *not* already drop out during summer), indicating that cognitive developments U_2 do not play an important additional role for R_2. Note that Assumption 1 would not be valid if we had not conditioned on Y_1 and the type, because both variables affect Y_2 as well as R_2.

A sufficient condition for (5) is

$$U_2 \perp\!\!\!\perp V_2 | U_1, V_1, U_0, W, X_0,$$

because Z and $\mathcal{T}$ are deterministic functions of W, U_0 and X_0.

Assumption 1' is stronger than Assumption 1 by requiring independence without conditioning on the unobserved type. In general, Assumption 1' is not satisfied in the above model because the types appear in both Y_2 and R_2. It would only be valid if either $\kappa_{a2} = \kappa_{n2} = \kappa_{c2}$ or $\tau_{a2} = \tau_{n2} = \tau_{c2}$. Hence, the various types need to be homogeneous with respect to either their outcome distributions (which also rules out treatment-type interaction terms) or their response behavior. Homogeneity in responses appears unlikely if the reason why never takers turn down the subsidy for remedial summer education is for instance a systematically lower level of motivation. In this case, one would expect a higher drop-out probability, i.e. a low τ_{n2}, as well as a lower growth rate in cognitive skills, i.e. a low κ_{n2}, for this group. Hence, in the remedial education example considered Assumption 1 appears clearly more credible than Assumption 1'.

Assumption 2' is identical to the exclusion restriction in the absence of attrition: $(Y_t^d, \mathcal{T}) \perp\!\!\!\perp Z | X_0$, which requires that $\zeta_1 = \zeta_2 = 0$ (i.e. no direct effect of Z) and W to be independent of U_0, U_1, and U_2, given X_0.

Assumption 2 is stronger by additionally implying that W is independent of V_1, the unobservable in R_1. Furthermore, ρ_1 has to be zero (no direct effect of the instrument on response). Note that when estimating the treatment effect on Y_2, Assumption 2 only restricts first period response R_1, but does *not* impose constraints on R_2, and thus neither on ρ_2 nor on V_2. This is particularly relevant in the remedial education example when $R_1 = 1$ with probability one, i.e. if exam results are observed for everyone during compulsory schooling and attrition only commences afterwards. In this case, Assumption 2 does not impose any restrictions on the response process.

Only when non-response already occurs in time period 1, it has to be required that W is independent of V_1 and $\rho_1 = 0$, which in our model refers to all types. However, if one allowed for differential effects of Z on R_1 across types (e.g. by including interaction terms between Z and the types), then only the effects of Z on response among never and always takers had to be zero, whereas no restriction is needed for the compliers. In the context of our example, it appears likely that mere voucher assignment does not affect the never takers' decisions to drop out right after compulsory schooling, because they would not make use of the voucher anyway. For always takers, the absence of a direct effect may be harder to justify if we, for instance, believed that already the mere assignment of the voucher can have an affirmative effect (e.g. through a feeling of appreciation or financial relief) that decreases the probability to drop out after compulsory schooling despite the fact that always takers would have participated in the summer course anyway. However, in many randomized trials individuals are perfectly excluded from the treatment if randomized out so that always takers (and the issues just discussed) do not exist. This is also the case for the empirical application considered in Section 4.

C Identification with more than 2 time periods

For ease of exposition, in the main paper we had only considered the situation with two time periods of data available after treatment, which is rather common e.g. with randomized trials. In this appendix we extend Theorems 1 and 2 to the general case with more than 2 time periods, i.e. when more than two time periods *after* treatment are available. To make this appendix self-contained, we first repeat the assumptions of the main paper for $t = 1, 2, 3, ...$ after the start of the treatment. As before, $t = 0$ refers to the time before the treatment.

Assumption 1: Predetermined non-response

$$Y_{t+s} \perp\!\!\!\perp R_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z, \mathcal{T}, \qquad \text{for } s \geq 0$$

Assumption 1': Predetermined non-response

$$Y_{t+s} \perp\!\!\!\perp R_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z, D \qquad \text{for } s \geq 0$$

Assumption 2: Exclusion restriction: For $d \in \{0, 1\}$

$$(Y_t, \underline{X}_{t-1}, \underline{R}_{t-1}) \perp\!\!\!\perp Z | X_0, \mathcal{T} \in \{a, n\}$$
$$Y_t^d \perp\!\!\!\perp Z | X_0, \mathcal{T} = c$$
$$\mathcal{T} \perp\!\!\!\perp Z | X_0$$

Assumption 2': Exclusion restriction: For almost every X_0 and For $d \in \{0,1\}$

$$(Y_t^d, \mathcal{T}) \perp\!\!\!\perp Z | X_0$$

Assumption 3: Monotonicity and support restrictions

Existence of compliers:	$\Pr(\mathcal{T} = c) > 0$
Monotonicity:	$\Pr(\mathcal{T} = d) = 0$
Common support:	$0 < \Pr(Z = 1 \| X_0) < 1$

Theorems 1 and 2 for general t, i.e. t possibly larger than 2, now become:

Theorem 1 (Non-response under conditional latent ignorability): Under Assumptions 1, 2 and 3, the potential outcomes are identified as

$$E\left[Y_t^1 | \mathcal{T} = c\right] = E\left[Y_t R_1 ... R_t \frac{D}{\pi} \frac{Z-\pi}{1-\pi} \prod_{l=1}^{t} \frac{1}{\Xi_{l-1}} \frac{P_{l-1}-\pi}{P'_{l-1}-\pi}\right] \times \frac{1}{E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right]}.$$

Theorem 2 (Non-response under the missing at random assumption): Under Assumptions 1', 2' and 3 we identify the potential outcomes as

$$\begin{aligned} E\left[Y_t^1 | \mathcal{T} = c\right] &= E\left[Y_t R_1 \cdots R_t D \prod_{l=1}^{t} \frac{1}{\Xi_{l-1}} \left(\frac{Z}{\pi} \prod_{l=1}^{t} \frac{P_{l-1}}{P'_{l-1}} - \frac{1-Z}{1-\pi} \prod_{l=1}^{t} \frac{1-P_{l-1}}{1-P'_{l-1}}\right)\right] \times \frac{1}{E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right]} \\ &= E\left[Y_t R_1 \cdots R_t D \left(\frac{Z}{\pi} \prod_{l=1}^{t} \frac{1}{\Xi_{l-1,Z=1}} - \frac{1-Z}{1-\pi} \prod_{l=1}^{t} \frac{1}{\Xi_{l-1,Z=0}}\right)\right] \times \frac{1}{E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right]}. \end{aligned}$$

The expressions for $E\left[Y_t^0 | \mathcal{T} = c\right]$ are obtained by replacing D with $1-D$ and $D=1$ with $D=0$ everywhere.

The proofs for these theorems are given in the following section. Theorem 1 and 2 in the main paper are special cases of these results.

We repeat the definition of the symbols for the conditional probabilities below:

$$
\begin{aligned}
\pi &= \Pr\left(Z=1|X_0\right)\\
P_t &= \Pr\left(Z=1|X_0,\underline{X}_t,\underline{R}_t=1,D=1\right)\\
P_t' &= \Pr\left(Z=1|X_0,\underline{X}_t,\underline{R}_{t+1}=1,D=1\right)\\
\Xi_t &= \Pr\left(R_{t+1}=1|X_0,\underline{X}_t,\underline{R}_t=1,D=1\right)\\
\Xi_{t,Z=z} &= \Pr\left(R_{t+1}=1|X_0,\underline{X}_t,\underline{R}_t=1,D=1,Z=z\right).
\end{aligned}
$$

D Proofs of Identification Theorems:

D.1 Proof of Theorem 1

Preliminaries, Part a

By Assumption 2 it follows:

$$
\begin{aligned}
\Pr\left(\mathcal{T}=a|X_0,Z,\underline{X}_{t-1},\underline{R}_{t-1}\right) &= \frac{dF\left(\underline{X}_{t-1},\underline{R}_{t-1}|X_0,Z,\mathcal{T}=a\right)\Pr\left(\mathcal{T}=a|X_0,Z\right)}{dF\left(\underline{X}_{t-1},\underline{R}_{t-1}|X_0,Z\right)}\\
&= \frac{dF\left(\underline{X}^1_{t-1},\underline{R}^1_{t-1}|X_0,Z,\mathcal{T}=a\right)\Pr\left(\mathcal{T}=a|X_0,Z\right)}{dF\left(\underline{X}_{t-1},\underline{R}_{t-1}|X_0,Z\right)}\\
&= \frac{dF\left(\underline{X}^1_{t-1},\underline{R}^1_{t-1}|X_0,\mathcal{T}=a\right)\Pr\left(\mathcal{T}=a|X_0\right)}{dF\left(\underline{X}_{t-1},\underline{R}_{t-1}|X_0,Z\right)}
\end{aligned}
$$

Hence,

$$
\frac{\Pr\left(\mathcal{T}=a|X_0,Z=1,\underline{X}_{t-1},\underline{R}_{t-1}\right)}{\Pr\left(\mathcal{T}=a|X_0,Z=0,\underline{X}_{t-1},\underline{R}_{t-1}\right)}=\frac{dF\left(\underline{X}_{t-1},\underline{R}_{t-1}|X_0,Z=0\right)}{dF\left(\underline{X}_{t-1},\underline{R}_{t-1}|X_0,Z=1\right)}=\frac{\Pr\left(Z=0|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right)}{\Pr\left(Z=1|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right)}\frac{\Pr\left(Z=1|X_0\right)}{\Pr\left(Z=0|X_0\right)} \tag{6}
$$

where the last part follows from using $\Pr\left(Z=0|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right)=\frac{dF\left(\underline{X}_{t-1},\underline{R}_{t-1}|X_0,Z=0\right)\Pr(Z=0|X_0)}{dF\left(\underline{X}_{t-1},\underline{R}_{t-1}|X_0\right)}$ because of Bayes' theorem.

We can also derive an analogous expression for never takers:

$$
\frac{\Pr\left(\mathcal{T}=n|X_0,Z=1,\underline{X}_{t-1},\underline{R}_{t-1}\right)}{\Pr\left(\mathcal{T}=n|X_0,Z=0,\underline{X}_{t-1},\underline{R}_{t-1}\right)}=\frac{dF\left(\underline{X}_{t-1},\underline{R}_{t-1}|X_0,Z=0\right)}{dF\left(\underline{X}_{t-1},\underline{R}_{t-1}|X_0,Z=1\right)}=\frac{\Pr\left(Z=0|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right)}{\Pr\left(Z=1|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right)}\frac{\Pr\left(Z=1|X_0\right)}{\Pr\left(Z=0|X_0\right)}.
$$

Note that no such simple expression exists for the compliers because a change in Z also entails a change in D for them.

Similarly, we obtain

$$
\frac{\Pr\left(\mathcal{T}=a|X_0,Z=1,\underline{X}_{t-2},\underline{R}_{t-1}\right)}{\Pr\left(\mathcal{T}=a|X_0,Z=0,\underline{X}_{t-2},\underline{R}_{t-1}\right)}=\frac{dF\left(\underline{X}_{t-2},\underline{R}_{t-1}|X_0,Z=0\right)}{dF\left(\underline{X}_{t-2},\underline{R}_{t-1}|X_0,Z=1\right)}=\frac{\Pr\left(Z=0|X_0,\underline{X}_{t-2},\underline{R}_{t-1}\right)}{\Pr\left(Z=1|X_0,\underline{X}_{t-2},\underline{R}_{t-1}\right)}\frac{\Pr\left(Z=1|X_0\right)}{\Pr\left(Z=0|X_0\right)}. \tag{7}
$$

Preliminaries, Part b

Consider the following term

$$E\left[Y_t R_t D | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0\right]$$

$$\begin{aligned}
&= E\left[Y_t R_t D | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0, \mathcal{T}=c\right] \Pr\left(\mathcal{T}=c | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0\right) \\
&\quad + E\left[Y_t R_t D | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0\right) \\
&\quad + E\left[Y_t R_t D | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0, \mathcal{T}=n\right] \Pr\left(\mathcal{T}=n | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0\right)
\end{aligned}$$

$$= E\left[Y_t R_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0\right)$$

$$\begin{aligned}
&= E\left[Y_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0, \mathcal{T}=a\right] E\left[R_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0\right) \\
&= E\left[Y_t^1 | X_0, \underline{X}_{t-1}^1, \underline{R}_{t-1}^1, Z=0, \mathcal{T}=a\right] E\left[R_t^1 | X_0, \underline{X}_{t-1}^1, \underline{R}_{t-1}^1, Z=0, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0\right) \\
&= E\left[Y_t^1 | X_0, \underline{X}_{t-1}^1, \underline{R}_{t-1}^1, \mathcal{T}=a\right] E\left[R_t^1 | X_0, \underline{X}_{t-1}^1, \underline{R}_{t-1}^1, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=0\right)
\end{aligned}$$

Analogously, we obtain

$$E\left[Y_t R_t D | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1\right]$$

$$\begin{aligned}
&= E\left[Y_t R_t D | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1, \mathcal{T}=c\right] \Pr\left(\mathcal{T}=c | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1\right) \\
&\quad + E\left[Y_t R_t D | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1\right) \\
&\quad + E\left[Y_t R_t D | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1, \mathcal{T}=n\right] \Pr\left(\mathcal{T}=n | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1\right)
\end{aligned}$$

$$\begin{aligned}
&= E\left[Y_t R_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1, \mathcal{T}=c\right] \Pr\left(\mathcal{T}=c | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1\right) \\
&\quad + E\left[Y_t R_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1\right)
\end{aligned}$$

$$\begin{aligned}
&= E\left[Y_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1, \mathcal{T}=c\right] E\left[R_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1, \mathcal{T}=c\right] \Pr\left(\mathcal{T}=c | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1\right) \\
&\quad + E\left[Y_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1, \mathcal{T}=a\right] E\left[R_t | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1\right)
\end{aligned}$$

$$\begin{aligned}
&= E\left[Y_t^1 | X_0, \underline{X}_{t-1}^1, \underline{R}_{t-1}^1, Z=1, \mathcal{T}=c\right] E\left[R_t^1 | X_0, \underline{X}_{t-1}^1, \underline{R}_{t-1}^1, Z=1, \mathcal{T}=c\right] \Pr\left(\mathcal{T}=c | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1\right) \\
&\quad + E\left[Y_t^1 | X_0, \underline{X}_{t-1}^1, \underline{R}_{t-1}^1, \mathcal{T}=a\right] E\left[R_t^1 | X_0, \underline{X}_{t-1}^1, \underline{R}_{t-1}^1, \mathcal{T}=a\right] \Pr\left(\mathcal{T}=a | X_0, \underline{X}_{t-1}, \underline{R}_{t-1}, Z=1\right).
\end{aligned}$$

Putting these results together we obtain

$$E\left[Y_tR_tD|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=1\right]-E\left[Y_tR_tD|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=0\right]\frac{\Pr\left(\mathcal{T}=a|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=1\right)}{\Pr\left(\mathcal{T}=a|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=0\right)}$$
$$=E\left[Y_t^1|X_0,\underline{X}_{t-1}^1,\underline{R}_{t-1}^1,Z=1,\mathcal{T}=c\right]E\left[R_t^1|X_0,\underline{X}_{t-1}^1,\underline{R}_{t-1}^1,Z=1,\mathcal{T}=c\right]\Pr\left(\mathcal{T}=c|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=1\right).$$

With analogous derivations we also obtain that

$$E\left[R_tD|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=1\right]-E\left[R_tD|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=0\right]\frac{\Pr\left(\mathcal{T}=a|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=1\right)}{\Pr\left(\mathcal{T}=a|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=0\right)}$$
$$=E\left[R_t^1|X_0,\underline{X}_{t-1}^1,\underline{R}_{t-1}^1,Z=1,\mathcal{T}=c\right]\Pr\left(\mathcal{T}=c|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=1\right)$$

and therefore that

$$\frac{E\left[Y_tR_tD|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=1\right]-E\left[Y_tR_tD|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=0\right]\frac{\Pr\left(\mathcal{T}=a|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=1\right)}{\Pr\left(\mathcal{T}=a|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=0\right)}}{E\left[R_tD|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=1\right]-E\left[R_tD|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=0\right]\frac{\Pr\left(\mathcal{T}=a|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=1\right)}{\Pr\left(\mathcal{T}=a|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=0\right)}}$$
$$=E\left[Y_t^1|X_0,\underline{X}_{t-1}^1,\underline{R}_{t-1}^1,Z=1,\mathcal{T}=c\right].$$

Combining these results with (6) we obtain after some algebra

$$E\left[Y_t^1|X_0,\underline{X}_{t-1}^1,\underline{R}_{t-1}^1,Z=1,\mathcal{T}=c\right]$$
$$=\frac{E\left[Y_tR_tD|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=1\right]-E\left[Y_tR_tD|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=0\right]\frac{\Pr\left(Z=0|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right)}{\Pr\left(Z=1|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right)}\frac{\Pr(Z=1|X_0)}{\Pr(Z=0|X_0)}}{E\left[R_tD|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=1\right]-E\left[R_tD|X_0,\underline{X}_{t-1},\underline{R}_{t-1},Z=0\right]\frac{\Pr\left(Z=0|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right)}{\Pr\left(Z=1|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right)}\frac{\Pr(Z=1|X_0)}{\Pr(Z=0|X_0)}}$$
$$=\frac{\frac{E\left[Y_tR_tDZ|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right]}{\Pr(Z=1|X_0)}-\frac{E\left[Y_tR_tD(1-Z)|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right]}{\Pr(Z=0|X_0)}}{\frac{E\left[R_tDZ|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right]}{\Pr(Z=1|X_0)}-\frac{E\left[R_tD(1-Z)|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right]}{\Pr(Z=0|X_0)}}$$
$$=\frac{E\left[\frac{Y_tR_tD}{\pi(X_0)}\frac{Z-\pi(X_0)}{1-\pi(X_0)}|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right]}{E\left[\frac{R_tD}{\pi(X_0)}\frac{Z-\pi(X_0)}{1-\pi(X_0)}|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right]}=\frac{E\left[Y_tR_tD\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right]}{E\left[R_tD\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-1},\underline{R}_{t-1}\right]}$$

Preliminaries, Part c

We will need an expression for

$$dF_{X_{t-1}^1|X_0,\underline{X}_{t-2}^1,\underline{R}_{t-1}^1=1,Z=1,\mathcal{T}=c}$$

Let us examine the following expression and make use of the Assumptions 1 and 2 and also make use of (7), we obtain after tedious calculations

$$E\left[1(X_{t-1}\le u)D|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1,Z=1\right]$$
$$-E\left[1(X_{t-1}\le u)D|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1,Z=0\right]\frac{\Pr\left(Z=0|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1\right)}{\Pr\left(Z=1|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1\right)}\frac{\Pr\left(Z=1|X_0\right)}{\Pr\left(Z=0|X_0\right)}$$

$$= E\left[1(X^1_{t-1} \leq u)|X_0, \underline{X}^1_{t-2}, \underline{R}^1_{t-1} = 1, Z = 1, \mathcal{T} = c\right] \Pr\left(\mathcal{T} = c|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1, Z = 1\right).$$

After various calculations and making use of (7) we obtain

$$\frac{E\left[1(X_{t-1} \leq u)D|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1, Z = 1\right] - E\left[1(X_{t-1} \leq u)D|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1, Z = 0\right] \frac{\frac{\Pr\left(Z=0|X_0, \underline{X}_{t-2}, \underline{R}_{t-1}=1\right)}{\Pr(Z=0|X_0)}}{\frac{\Pr\left(Z=1|X_0, \underline{X}_{t-2}, \underline{R}_{t-1}=1\right)}{\Pr(Z=1|X_0)}}}{E\left[D|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1, Z = 1\right] - E\left[D|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1, Z = 0\right] \frac{\Pr\left(Z=0|X_0, \underline{X}_{t-2}, \underline{R}_{t-1}=1\right)}{\Pr\left(Z=1|X_0, \underline{X}_{t-2}, \underline{R}_{t-1}=1\right)} \frac{\Pr(Z=1|X_0)}{\Pr(Z=0|X_0)}}$$

$$= E\left[1(X^1_{t-1} \leq u)|X_0, \underline{X}^1_{t-2}, \underline{R}^1_{t-1} = 1, Z = 1, \mathcal{T} = c\right].$$

After some algebra we then obtain

$$\begin{aligned} F_{X^1_{t-1}|X_0, \underline{X}^1_{t-2}, \underline{R}^1_{t-1}=1, Z=1, \mathcal{T}=c}(u) &= \frac{E\left[\frac{1(X_{t-1} \leq u)D}{\pi(X_0)} \frac{Z - \pi(X_0)}{1 - \pi(X_0)}|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right]}{E\left[\frac{D}{\pi(X_0)} \frac{Z - \pi(X_0)}{1 - \pi(X_0)}|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right]} \\ &= \frac{E\left[1(X_{t-1} \leq u)D\left(Z - \pi(X_0)\right)\ |X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right]}{E\left[D\left(Z - \pi(X_0)\right)\ |X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right]} \qquad (8) \end{aligned}$$

i.e. we have expressed the distribution function of the potential values of X_{t-1} conditional on potential values of $\underline{X}_{t-2}$ as a function of observed variables.

By noting that

$$E\left[1(X_{t-1} \leq u)D\left(Z - \pi(X_0)\right)|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right]$$

$$\begin{aligned} = \ & E\left[1(X_{t-1} \leq u)D\left(Z - \pi(X_0)\right)|D = 1, Z = 1, X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right] \Pr\left(D = 1, Z = 1|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right) \\ & + E\left[1(X_{t-1} \leq u)D\left(Z - \pi(X_0)\right)|D = 1, Z = 0, X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right] \Pr\left(D = 1, Z = 0|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right) \\ & + E\left[1(X_{t-1} \leq u)D\left(Z - \pi(X_0)\right)|D = 0, X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right] \Pr\left(D = 0|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right) \end{aligned}$$

$$\begin{aligned} = \ & E\left[1(X_{t-1} \leq u)|D = 1, Z = 1, X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right] \Pr\left(D = 1, Z = 1|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right)\left(1 - \pi(X_0)\right) \\ & + E\left[1(X_{t-1} \leq u)|D = 1, Z = 0, X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right] \Pr\left(D = 1, Z = 0|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right)\left(-\pi(X_0)\right) \end{aligned}$$

$$\begin{aligned} = \ & \left(1 - \pi(X_0)\right) F_{X_{t-1}|D=1, Z=1, X_0, \underline{X}_{t-2}, \underline{R}_{t-1}=1}(u) \Pr\left(D = 1, Z = 1|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right) \\ & - \pi(X_0) F_{X_{t-1}|D=1, Z=0, X_0, \underline{X}_{t-2}, \underline{R}_{t-1}=1}(u) \Pr\left(D = 1, Z = 0|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right) \end{aligned}$$

we obtain the density function

$$\begin{aligned} & dF_{X^1_{t-1}|X_0, \underline{X}^1_{t-2}, \underline{R}^1_{t-1}=1, Z=1, \mathcal{T}=c}(u) \\ & \quad = \frac{\left(1 - \pi(X_0)\right) dF_{X_{t-1}|D=1, Z=1, X_0, \underline{X}_{t-2}, \underline{R}_{t-1}=1}(u) \Pr\left(D = 1, Z = 1|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right)}{E\left[D\left(Z - \pi(X_0)\right)|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right]} \\ & \qquad - \frac{\pi(X_0) dF_{X_{t-1}|D=1, Z=0, X_0, \underline{X}_{t-2}, \underline{R}_{t-1}=1}(u) \Pr\left(D = 1, Z = 0|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right)}{E\left[D\left(Z - \pi(X_0)\right)|X_0, \underline{X}_{t-2}, \underline{R}_{t-1} = 1\right]}. \end{aligned}$$

Now we make use of Bayes' theorem applied to

$$\begin{aligned}&\Pr\left(D=1,Z=1|X_0,\underline{X}_{t-1},\underline{R}_{t-1}=1\right)\\&\qquad=\frac{dF\left(X_{t-1}|D=1,Z=1,X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1\right)\Pr\left(D=1,Z=1|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1\right)}{dF\left(X_{t-1}|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1\right)}.\end{aligned}$$

When we insert this expression in the previous expression for $dF_{X^1_{t-1}|X_0,\underline{X}^1_{t-2},\underline{R}^1_{t-1}=1,Z=1,\mathcal{T}=c}(u)$ we obtain after some algebra:

$$\begin{aligned}&dF_{X^1_{t-1}|X_0,\underline{X}^1_{t-2},\underline{R}^1_{t-1}=1,Z=1,\mathcal{T}=c}\\&=\left\{(1-\pi(X_0))\Pr\left(D=1,Z=1|X_0,\underline{X}_{t-1},\underline{R}_{t-1}=1\right)-\pi(X_0)\Pr\left(D=1,Z=0|X_0,\underline{X}_{t-1},\underline{R}_{t-1}=1\right)\right\}\\&\qquad\times\frac{dF_{X_{t-1}|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1}}{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1\right]}\end{aligned}$$

and finally after some algebra

$$dF_{X^1_{t-1}|X_0,\underline{X}^1_{t-2},\underline{R}^1_{t-1}=1,Z=1,\mathcal{T}=c}=\frac{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-1},\underline{R}_{t-1}=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1\right]}dF_{X_{t-1}|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1}. \tag{9}$$

Preliminaries, Part d

Finally, we need an expression for

$$dF_{X_0|\mathcal{T}=c}=\frac{\Pr\left(\mathcal{T}=c|X_0\right)dF_{X_0}}{\Pr\left(\mathcal{T}=c\right)}=\frac{\Pr\left(\mathcal{T}=c|X_0\right)dF_{X_0}}{\Pr\left(\mathcal{T}=c\right)}=\frac{E\left[\frac{D}{\pi(X_0)}\frac{Z-\pi(X_0)}{1-\pi(X_0)}|X_0\right]dF_{X_0}}{E\left[\frac{D}{\pi(X_0)}\frac{Z-\pi(X_0)}{1-\pi(X_0)}\right]} \tag{10}$$

where we used that $\Pr(\mathcal{T}=c|X_0)=E\left[D|X_0,Z=1\right]-E\left[D|X_0,Z=0\right]=E\left[\frac{D}{\pi(X_0)}\frac{Z-\pi(X_0)}{1-\pi(X_0)}|X_0\right]$ and $\Pr(\mathcal{T}=c)=E\left[\frac{D}{\pi(X_0)}\frac{Z-\pi(X_0)}{1-\pi(X_0)}\right]$.

Preliminaries, Part e

By making use of iterated expectation and the assumptions

$$E\left[Y^1_t|X_0,Z=1,\mathcal{T}=c\right]=E\left[Y^1_t|X_0,R^1_1=1,Z=1,\mathcal{T}=c\right]$$

$$
\begin{aligned}
&= E\left[E\left[Y_t^1|X_0, X_1^1, R_1^1=1, Z=1, \mathcal{T}=c\right]|X_0, R_1^1=1, Z=1, \mathcal{T}=c\right]\\
&= \int E\left[Y_t^1|X_0, X_1^1, R_1^1=1, Z=1, \mathcal{T}=c\right]\cdot dF_{X_1^1|X_0,R_1^1=1,Z=1,\mathcal{T}=c}\\
&= \int E\left[Y_t^1|X_0, X_1^1, \underline{R}_2^1=1, Z=1, \mathcal{T}=c\right]\cdot dF_{X_1^1|X_0,R_1^1=1,Z=1,\mathcal{T}=c}\\
&= \int\cdots\int E\left[Y_t^1|X_0, \underline{X}_2^1, \underline{R}_2^1=1, Z=1, \mathcal{T}=c\right]\cdot dF_{X_2^1|X_0,X_1^1,\underline{R}_2^1=1,Z=1,\mathcal{T}=c}\cdot dF_{X_1^1|X_0,R_1^1=1,Z=1,\mathcal{T}=c}\\
&= \int\cdots\int E\left[Y_t^1|X_0, \underline{X}_2^1, \underline{R}_3^1=1, Z=1, \mathcal{T}=c\right]\cdot dF_{X_2^1|X_0,X_1^1,\underline{R}_2^1=1,Z=1,\mathcal{T}=c}\cdot dF_{X_1^1|X_0,R_1^1=1,Z=1,\mathcal{T}=c}
\end{aligned}
$$

$$
\begin{aligned}
= \int\cdots\int E\left[Y_t^1|X_0, \underline{X}_3^1, \underline{R}_3^1=1, Z=1, \mathcal{T}=c\right] dF_{X_3^1|X_0,\underline{X}_2^1,\underline{R}_3^1=1,Z=1,\mathcal{T}=c}\\
\cdot\, dF_{X_2^1|X_0,X_1^1,\underline{R}_2^1=1,Z=1,\mathcal{T}=c} dF_{X_1^1|X_0,R_1^1=1,Z=1,\mathcal{T}=c}
\end{aligned}
$$

$$
\begin{aligned}
&= \int\cdots\int E\left[Y_t^1|X_0, \underline{X}_{t-1}^1, \underline{R}_{t-1}^1=1, Z=1, \mathcal{T}=c\right]\cdot dF_{X_{t-1}^1|X_0,\underline{X}_{t-2}^1,\underline{R}_{t-1}^1=1,Z=1,\mathcal{T}=c}\cdot\ldots\cdot dF_{X_4^1|X_0,\underline{X}_3^1,\underline{R}_4^1=1,Z=1,\mathcal{T}=c}\\
&\quad dF_{X_3^1|X_0,\underline{X}_2^1,\underline{R}_3^1=1,Z=1,\mathcal{T}=c}\cdot dF_{X_2^1|X_0,X_1^1,\underline{R}_2^1=1,Z=1,\mathcal{T}=c}\cdot dF_{X_1^1|X_0,R_1^1=1,Z=1,\mathcal{T}=c} \qquad (11)
\end{aligned}
$$

Final proof:

Making use of all the previous calculations we obtain

$$
\begin{aligned}
E\left[Y_t^1|\mathcal{T}=c\right] &= \int E\left[Y_t^1|X_0, \mathcal{T}=c\right] dF_{X_0|\mathcal{T}=c}\\
&= \int E\left[Y_t^1|X_0, Z=1, \mathcal{T}=c\right] dF_{X_0|\mathcal{T}=c}
\end{aligned}
$$

by Assumption 1.

$$
\begin{aligned}
&= \int\cdots\int E\left[Y_t^1|X_0, \underline{X}_{t-1}^1, \underline{R}_{t-1}^1=1, Z=1, \mathcal{T}=c\right]\cdot dF_{X_{t-1}^1|X_0,\underline{X}_{t-2}^1,\underline{R}_{t-1}^1=1,Z=1,\mathcal{T}=c}\cdot\ldots\cdot dF_{X_4^1|X_0,\underline{X}_3^1,\underline{R}_4^1=1,Z=1,\mathcal{T}=c}\\
&\quad dF_{X_3^1|X_0,\underline{X}_2^1,\underline{R}_3^1=1,Z=1,\mathcal{T}=c}\cdot dF_{X_2^1|X_0,X_1^1,\underline{R}_2^1=1,Z=1,\mathcal{T}=c}\cdot dF_{X_1^1|X_0,R_1^1=1,Z=1,\mathcal{T}=c}\cdot dF_{X_0|\mathcal{T}=c}
\end{aligned}
$$

$$
\begin{aligned}
= \quad & \int \cdots \int \frac{E\left[Y_t R_t D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_{t-1}=1\right]}{E\left[R_t D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_{t-1}=1\right]} \\
& \cdot \frac{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_{t-1}=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-2}, \underset{\sim}{R}_{t-1}=1\right]} dF_{X_{t-1}|X_0, \underset{\sim}{X}_{t-2}, \underset{\sim}{R}_{t-1}=1} \\
& \cdot \frac{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-2}, \underset{\sim}{R}_{t-2}=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-3}, \underset{\sim}{R}_{t-2}=1\right]} dF_{X_{t-2}|X_0, \underset{\sim}{X}_{t-3}, \underset{\sim}{R}_{t-2}=1} \\
& \cdot \ldots \cdot \frac{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_3, \underset{\sim}{R}_3=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_2, \underset{\sim}{R}_3=1\right]} dF_{X_3|X_0, \underset{\sim}{X}_2, \underset{\sim}{R}_3=1} \\
& \cdot \frac{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_2, \underset{\sim}{R}_2=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0, X_1, \underset{\sim}{R}_2=1\right]} dF_{X_2|X_0, X_1, \underset{\sim}{R}_2=1} \\
& \cdot \frac{E\left[D\left(Z-\pi(X_0)\right)|X_0, X_1, R_1=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0, R_1=1\right]} dF_{X_1|X_0, R_1=1} \cdot \frac{E\left[\frac{D}{\pi(X_0)}\frac{Z-\pi(X_0)}{1-\pi(X_0)}|X_0\right] dF_{X_0}}{E\left[\frac{D}{\pi(X_0)}\frac{Z-\pi(X_0)}{1-\pi(X_0)}\right]}.
\end{aligned}
$$

By re-arranging terms we can write this as

$$
\begin{aligned}
= \quad & \int \cdots \int \frac{E\left[Y_t R_1 \ldots R_t D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-1}\right]}{\Pr\left(\underset{\sim}{R}_{t-1}=1|X_0, \underset{\sim}{X}_{t-1}\right)} \\
& \frac{1}{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_t=1\right] \Pr\left(R_t=1|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_{t-1}=1\right)} \\
& \cdot \frac{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_{t-1}=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-2}, \underset{\sim}{R}_{t-1}=1\right]} \frac{\Pr\left(\underset{\sim}{R}_{t-1}=1|X_0, \underset{\sim}{X}_{t-1}\right)}{\Pr\left(\underset{\sim}{R}_{t-1}=1|X_0, \underset{\sim}{X}_{t-2}\right)} dF_{X_{t-1}|X_0, \underset{\sim}{X}_{t-2}} \\
& \cdot \frac{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-2}, \underset{\sim}{R}_{t-2}=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-3}, \underset{\sim}{R}_{t-2}=1\right]} \frac{\Pr\left(\underset{\sim}{R}_{t-2}=1|X_0, \underset{\sim}{X}_{t-2}\right)}{\Pr\left(\underset{\sim}{R}_{t-2}=1|X_0, \underset{\sim}{X}_{t-3}\right)} dF_{X_{t-2}|X_0, \underset{\sim}{X}_{t-3}} \\
& \cdot \ldots \cdot \frac{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_3, \underset{\sim}{R}_3=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_2, \underset{\sim}{R}_3=1\right]} \frac{\Pr\left(\underset{\sim}{R}_3=1|X_0, \underset{\sim}{X}_3\right)}{\Pr\left(\underset{\sim}{R}_3=1|X_0, \underset{\sim}{X}_2\right)} dF_{X_3|X_0, \underset{\sim}{X}_2} \\
& \cdot \frac{E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_2, \underset{\sim}{R}_2=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0, X_1, \underset{\sim}{R}_2=1\right]} \frac{\Pr\left(\underset{\sim}{R}_2=1|X_0, \underset{\sim}{X}_2\right)}{\Pr\left(\underset{\sim}{R}_2=1|X_0, X_1\right)} dF_{X_2|X_0, X_1} \\
& \frac{E\left[D\left(Z-\pi(X_0)\right)|X_0, X_1, R_1=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0, R_1=1\right]} \frac{\Pr\left(R_1=1|X_0, X_1\right)}{\Pr\left(R_1=1|X_0\right)} dF_{X_1|X_0} \cdot \frac{E\left[\frac{D}{\pi(X_0)}\frac{Z-\pi(X_0)}{1-\pi(X_0)}|X_0\right] dF_{X_0}}{E\left[\frac{D}{\pi(X_0)}\frac{Z-\pi(X_0)}{1-\pi(X_0)}\right]}.
\end{aligned}
$$

where we also had used that

$$
dF_{X_{t-1}|X_0, \underset{\sim}{X}_{t-2}, \underset{\sim}{R}_{t-1}=1} = \frac{\Pr\left(\underset{\sim}{R}_{t-1}=1|X_0, \underset{\sim}{X}_{t-1}\right) dF_{X_{t-1}|X_0, \underset{\sim}{X}_{t-2}}}{\Pr\left(\underset{\sim}{R}_{t-1}=1|X_0, \underset{\sim}{X}_{t-2}\right)}.
$$

For the next step we note that

$$
\begin{aligned}
& E\left[D\left(Z-\pi(X_0)\right)|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_{t-1}=1\right] \\
& \qquad = \left\{\Pr\left(Z=1|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_{t-1}=1, D=1\right) - \pi(X_0)\right\} \Pr\left(D=1|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_{t-1}=1\right)
\end{aligned}
$$

and also note that

$$
\Pr\left(D=1|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_{t-1}=1\right) = \frac{\Pr\left(R_{t-1}=1|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_{t-2}=1, D=1\right) \Pr\left(D=1|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_{t-2}=1\right)}{\Pr\left(R_{t-1}=1|X_0, \underset{\sim}{X}_{t-1}, \underset{\sim}{R}_{t-2}=1\right)}
$$

which together implies that

$$\frac{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-1},\underline{R}_{t-1}=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-1},\underline{R}_{t}=1\right]}$$

$$=\frac{\left\{\Pr\left(Z=1|X_0,\underline{X}_{t-1},\underline{R}_{t-1}=1,D=1\right)-\pi(X_0)\right\}\Pr\left(R_t=1|X_0,\underline{X}_{t-1},\underline{R}_{t-1}=1\right)}{\left\{\Pr\left(Z=1|X_0,\underline{X}_{t-1},\underline{R}_{t}=1,D=1\right)-\pi(X_0)\right\}\Pr\left(R_t=1|X_0,\underline{X}_{t-1},\underline{R}_{t-1}=1,D=1\right)}.$$

Inserting these calculations into the expression for $E\left[Y_t^1|\mathcal{T}=c\right]$ we obtain

$$\begin{aligned}
= \quad & \int\cdots\int\frac{E\left[Y_tR_1...R_tD\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-1}\right]}{\Pr\left(\underline{R}_{t-1}=1|X_0,\underline{X}_{t-1}\right)}\frac{1}{\Pr\left(R_t=1|X_0,\underline{X}_{t-1},\underline{R}_{t-1}=1\right)}\\
&\cdot\frac{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-1},\underline{R}_{t-1}=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-1},\underline{R}_{t}=1\right]}\frac{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-2},\underline{R}_{t-2}=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1\right]}\\
&\cdot\frac{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-3},\underline{R}_{t-3}=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-3},\underline{R}_{t-2}=1\right]}\cdot...\cdot\frac{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{2},\underline{R}_{2}=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{2},\underline{R}_{3}=1\right]}\\
&\cdot\frac{E\left[D\left(Z-\pi(X_0)\right)|X_0,X_1,R_1=1\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0,X_1,\underline{R}_2=1\right]}\frac{E\left[D\left(Z-\pi(X_0)\right)|X_0\right]}{E\left[D\left(Z-\pi(X_0)\right)|X_0,R_1=1\right]}\\
&\frac{\Pr\left(\underline{R}_{t-1}=1|X_0,\underline{X}_{t-1}\right)}{\Pr\left(\underline{R}_{t-1}=1|X_0,\underline{X}_{t-2}\right)}\frac{\Pr\left(\underline{R}_{t-2}=1|X_0,\underline{X}_{t-2}\right)}{\Pr\left(\underline{R}_{t-2}=1|X_0,\underline{X}_{t-3}\right)}\cdot...\cdot\frac{\Pr\left(\underline{R}_{3}=1|X_0,\underline{X}_{3}\right)}{\Pr\left(\underline{R}_{3}=1|X_0,\underline{X}_{2}\right)}\frac{\Pr\left(\underline{R}_{2}=1|X_0,\underline{X}_{2}\right)}{\Pr\left(\underline{R}_{2}=1|X_0,X_1\right)}\\
&\frac{\Pr\left(R_1=1|X_0,X_1\right)}{\Pr\left(R_1=1|X_0\right)}\cdot\frac{\frac{1}{\pi(X_0)(1-\pi(X_0))}}{E\left[\frac{D}{\pi(X_0)}\frac{Z-\pi(X_0)}{1-\pi(X_0)}\right]}dF_{X_0,\underline{X}_{t-1}}
\end{aligned}$$

$$\begin{aligned}
&=\int\cdots\int E\left[Y_tR_1...R_tD\left(Z-\pi(X_0)\right)|X_0,\underline{X}_{t-1}\right]\\
&\prod_{l=1}^{t}\frac{\Pr\left(Z=1|X_0,\underline{X}_{l-1},\underline{R}_{l-1}=1,D=1\right)-\pi(X_0)}{\Pr\left(Z=1|X_0,\underline{X}_{l-1},\underline{R}_{l}=1,D=1\right)-\pi(X_0)}\frac{1}{\Pr\left(R_l=1|X_0,\underline{X}_{l-1},\underline{R}_{l-1}=1,D=1\right)}\\
&\qquad\times\frac{\frac{1}{\pi(X_0)(1-\pi(X_0))}}{E\left[\frac{D}{\pi(X_0)}\frac{Z-\pi(X_0)}{1-\pi(X_0)}\right]}dF_{X_0,\underline{X}_{t-1}}
\end{aligned}$$

where we also had used that

$$\frac{\Pr\left(\underline{R}_{t-2}=1|X_0,\underline{X}_{t-2}\right)}{\Pr\left(\underline{R}_{t-1}=1|X_0,\underline{X}_{t-2}\right)}=\frac{1}{\Pr\left(R_{t-1}=1|X_0,\underline{X}_{t-2},\underline{R}_{t-2}=1\right)}.$$

We can write this expression more concisely as

$$E\left[Y_t^1|\mathcal{T}=c\right]=E\left[Y_tR_1...R_t\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\prod_{l=1}^{t}\frac{1}{\Xi_{l-1}}\frac{P_{l-1}-\pi}{P'_{l-1}-\pi}\right]\times\frac{1}{E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right]}$$

where

$$
\begin{aligned}
\pi &= \pi(X_0) = \Pr(Z=1|X_0) \\
\Xi_t &= \Pr(R_{t+1}=1|X_0, \underline{X}_t, \underline{R}_t=1, D=1) \\
P_t &= \Pr(Z=1|X_0, \underline{X}_t, \underline{R}_t=1, D=1) \\
P_t' &= \Pr\left(Z=1|X_0, \underline{X}_t, \underline{R}_{t+1}=1, D=1\right).
\end{aligned}
$$

D.2 Proof of Theorem 2

The proof consists of three steps and is somewhat different from the one of Theorem 1 since we need not consider unobservable variables in the attrition process. First, we make repeated use of Assumption 2' to obtain an expression for $E\left[Y_t^1|X_0, \mathcal{T}=c\right]$. Second, we make repeated use of Assumption 1' to characterize the missing data process. Third, we simplify the expressions. This last part does not rest on any identifying assumptions and contains only algebraic calculations.

Part 1: Identification based on Assumption 2'

$$E[Y_tD|X_0, Z=1]$$

$$
\begin{aligned}
&= E[Y_tD|X_0, Z=1, \mathcal{T}=c]\Pr(\mathcal{T}=c|X_0, Z=1) \\
&\quad + E[Y_tD|X_0, Z=1, \mathcal{T}=a]\Pr(\mathcal{T}=a|X_0, Z=1) \\
&\quad + E[Y_tD|X_0, Z=1, \mathcal{T}=n]\Pr(\mathcal{T}=n|X_0, Z=1)
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[Y_t^1|X_0, Z=1, \mathcal{T}=c\right]\Pr(\mathcal{T}=c|X_0, Z=1) + E\left[Y_t^1|X_0, Z=1, \mathcal{T}=a\right]\Pr(\mathcal{T}=a|X_0, Z=1) \\
&= E\left[Y_t^1|X_0, \mathcal{T}=c\right]\Pr(\mathcal{T}=c|X_0) + E\left[Y_t^1|X_0, \mathcal{T}=a\right]\Pr(\mathcal{T}=a|X_0).
\end{aligned}
$$

With analogous calculations for $E[Y_tD|X_0, Z=0]$ and $E[D|X_0, Z=1]$ and $E[D|X_0, Z=0]$ we obtain

$$E[D|X_0, Z=1] - E[D|X_0, Z=0] = \Pr(\mathcal{T}=c|X_0) \tag{12}$$

and

$$\frac{E[Y_tD|X_0, Z=1] - E[Y_tD|X_0, Z=0]}{E[D|X_0, Z=1] - E[D|X_0, Z=0]} = E\left[Y_t^1|X_0, \mathcal{T}=c\right] \tag{13}$$

and

$$
\begin{aligned}
E\left[Y_t^1|\mathcal{T}=c\right] &= \int E\left[Y_t^1|X_0, \mathcal{T}=c\right] dF_{X_0|\mathcal{T}=c} = \int E\left[Y_t^1|X_0, \mathcal{T}=c\right] \frac{\Pr(\mathcal{T}=c|X_0)}{\Pr(\mathcal{T}=c)} dF_{X_0} \\
&= \frac{\int (E[Y_tD|X_0, Z=1] - E[Y_tD|X_0, Z=0])\, dF_{X_0}}{\int (E[D|X_0, Z=1] - E[D|X_0, Z=0])\, dF_{X_0}}. \qquad (14)
\end{aligned}
$$

Part 2: Missing data process

We consider the missing data process and derive an expression for $E\left[Y_t|X_0,Z=1,D=1\right]$ by making repeated use of Assumption 1'

$$
\begin{aligned}
& E\left[Y_t|X_0,Z=1,D=1\right] \\
= \; & E\left[Y_t|X_0,R_1=1,Z=1,D=1\right] \\
= \; & \int E\left[Y_t|X_0,X_1,R_1=1,Z=1,D=1\right] dF_{X_1|X_0,R_1=1,Z=1,D=1} \\
= \; & \int E\left[Y_t|X_0,X_1,\underline{R}_2=1,Z=1,D=1\right] dF_{X_1|X_0,R_1=1,Z=1,D=1} \\
= \; & \int \cdots \int E\left[Y_t|X_0,\underline{X}_2,\underline{R}_2=1,Z=1,D=1\right] dF_{X_2|X_0,X_1,\underline{R}_2=1,Z=1,D=1} dF_{X_1|X_0,R_1=1,Z=1,D=1} \\
= \; & \int \cdots \int E\left[Y_t|X_0,\underline{X}_2,\underline{R}_3=1,Z=1,D=1\right] dF_{X_2|X_0,X_1,\underline{R}_2=1,Z=1,D=1} dF_{X_1|X_0,R_1=1,Z=1,D=1}
\end{aligned}
$$

$$
\begin{aligned}
= & \int \cdots \int E\left[Y_t|X_0,\underline{X}_{t-1},\underline{R}_{t-1}=1,Z=1,D=1\right] dF_{X_{t-1}|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1,Z=1,D=1} \cdots dF_{X_3|X_0,\underline{X}_2,\underline{R}_3=1,Z=1,D=1} \\
& \times dF_{X_2|X_0,X_1,\underline{R}_2=1,Z=1,D=1} dF_{X_1|X_0,R_1=1,Z=1,D=1} \\
= & \int \cdots \int E\left[Y_t|X_0,\underline{X}_{t-1},\underline{R}_t=1,Z=1,D=1\right] dF_{X_{t-1}|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1,Z=1,D=1} \cdots dF_{X_3|X_0,\underline{X}_2,\underline{R}_3=1,Z=1,D=1} \\
& \times dF_{X_2|X_0,X_1,\underline{R}_2=1,Z=1,D=1} dF_{X_1|X_0,R_1=1,Z=1,D=1}. \qquad (15)
\end{aligned}
$$

Part 3: Simplification and identification We have obtained all parts for identification. The following calculations are made to obtain simpler expressions. By noting that

$$
dF_{X_{t-1}|X_0,\underline{X}_{t-2},\underline{R}_{t-1}=1,Z=1,D=1} = \frac{\Pr\left(\underline{R}_{t-1}=1|X_0,\underline{X}_{t-1},Z=1,D=1\right)}{\Pr\left(\underline{R}_{t-1}=1|X_0,\underline{X}_{t-2},Z=1,D=1\right)} dF_{X_{t-1}|X_0,\underline{X}_{t-2},Z=1,D=1}
$$

we obtain

$$
E\left[Y_t|X_0,Z=1,D=1\right] =
$$

$$
\begin{aligned}
= & \int \cdots \int E\left[Y_t|X_0,\underline{X}_{t-1},\underline{R}_t=1,Z=1,D=1\right] \\
& \frac{\Pr\left(\underline{R}_{t-1}=1|X_0,\underline{X}_{t-1},Z=1,D=1\right)}{\Pr\left(\underline{R}_{t-1}=1|X_0,\underline{X}_{t-2},Z=1,D=1\right)} \frac{\Pr\left(\underline{R}_{t-2}=1|X_0,\underline{X}_{t-2},Z=1,D=1\right)}{\Pr\left(\underline{R}_{t-2}=1|X_0,\underline{X}_{t-3},Z=1,D=1\right)} \\
& \cdots \\
& \frac{\Pr\left(\underline{R}_3=1|X_0,\underline{X}_3,Z=1,D=1\right)}{\Pr\left(\underline{R}_3=1|X_0,\underline{X}_2,Z=1,D=1\right)} \frac{\Pr\left(\underline{R}_2=1|X_0,\underline{X}_2,Z=1,D=1\right)}{\Pr\left(\underline{R}_2=1|X_0,X_1,Z=1,D=1\right)} \\
& \times \frac{\Pr\left(R_1=1|X_0,X_1,Z=1,D=1\right)}{\Pr\left(R_1=1|X_0,Z=1,D=1\right)} dF_{\underline{X}_{t-1}|X_0,Z=1,D=1}
\end{aligned}
$$

after re-arranging terms we obtain

$$= \int \cdots \int E\left[Y_t R_1 \cdots R_t | X_0, \underline{X}_{t-1}, Z=1, D=1\right] \frac{1}{\Pr\left(R_t = 1 | X_0, \underline{X}_{t-1}, \underline{R}_{t-1} = 1, Z=1, D=1\right)}$$

$$\frac{1}{\Pr\left(R_{t-1} = 1 | X_0, \underline{X}_{t-2}, \underline{R}_{t-2} = 1, Z=1, D=1\right)} \frac{1}{\Pr\left(R_{t-2} = 1 | X_0, \underline{X}_{t-3}, \underline{R}_{t-3} = 1, Z=1, D=1\right)}$$

$$\cdots \frac{1}{\Pr\left(R_3 = 1 | X_0, \underline{X}_2, \underline{R}_2 = 1, Z=1, D=1\right)}$$

$$\frac{1}{\Pr\left(R_2 = 1 | X_0, \underline{X}_1, \underline{R}_1 = 1, Z=1, D=1\right)} \frac{dF_{\underline{X}_{t-1}|X_0, Z=1, D=1}}{\Pr\left(R_1 = 1 | X_0, Z=1, D=1\right)}$$

where we used that

$$\frac{\Pr\left(\underline{R}_{t-2} = 1 | X_0, \underline{X}_{t-2}, Z=1, D=1\right)}{\Pr\left(\underline{R}_{t-1} = 1 | X_0, \underline{X}_{t-2}, Z=1, D=1\right)} = \frac{1}{\Pr\left(R_{t-1} = 1 | X_0, \underline{X}_{t-2}, \underline{R}_{t-2} = 1, Z=1, D=1\right)}.$$

Hence,

$$E\left[Y_t | X_0, Z=1, D=1\right] = E\left[Y_t R_1 \cdots R_t \prod_{l=1}^{t} \frac{1}{\Pr\left(R_l = 1 | X_0, \underline{X}_{l-1}, \underline{R}_{l-1} = 1, Z=1, D=1\right)} | X_0, Z=1, D=1\right]$$

$$E\left[Y_t | X_0, Z=1, D=1\right] = E\left[Y_t R_1 \cdots R_t \prod_{l=1}^{t} \frac{1}{\Xi_{l-1}} \frac{P_{l-1}}{P'_{l-1}} | X_0, Z=1, D=1\right]$$

where we used that $\Pr\left(R_l = 1 | X_0, \underline{X}_{l-1}, \underline{R}_{l-1} = 1, Z=1, D=1\right) =$

$$\frac{\Pr\left(Z=1 | X_0, \underline{X}_{l-1}, \underline{R}_l = 1, D=1\right) \Pr\left(R_l = 1 | X_0, \underline{X}_{l-1}, \underline{R}_{l-1} = 1, D=1\right)}{\Pr\left(Z=1 | X_0, \underline{X}_{l-1}, \underline{R}_{l-1} = 1, D=1\right)} = \frac{P'_{l-1}}{P_{l-1}} \Xi_{l-1}$$

and where we defined

$$\begin{aligned}
\pi &= \pi(X_0) = \Pr\left(Z=1 | X_0\right) \\
\Xi_t &= \Pr\left(R_{t+1} = 1 | X_0, \underline{X}_t, \underline{R}_t = 1, D=1\right) \\
P_t &= \Pr\left(Z=1 | X_0, \underline{X}_t, \underline{R}_t = 1, D=1\right) \\
P'_t &= \Pr\left(Z=1 | X_0, \underline{X}_t, \underline{R}_{t+1} = 1, D=1\right).
\end{aligned}$$

Hence,

$$E\left[Y_t D Z | X_0\right] = E\left[Y_t R_1 \cdots R_t D Z \prod_{l=1}^{t} \frac{1}{\Xi_{l-1}} \frac{P_{l-1}}{P'_{l-1}} | X_0\right]. \tag{16}$$

By analogous calculations we obtain for

$$E\left[Y_t | X_0, Z=0, D=1\right] = E\left[Y_t R_1 \cdots R_t \prod_{l=1}^{t} \frac{1}{\Xi_{l-1}} \frac{1 - P_{l-1}}{1 - P'_{l-1}} | X_0, Z=0, D=1\right]$$

where we used that $\Pr(Z=0|X_0,\underline{X}_t,\underline{R}_t=1,D=1)=1-P_t$ and $\Pr(Z=0|X_0,\underline{X}_t,\underline{R}_{t+1}=1,D=1)=1-P_t'$ because Z is binary. Also

$$E[Y_tD(1-Z)|X_0]=E\left[Y_tR_1\cdots R_tD(1-Z)\prod_{l=1}^{t}\frac{1}{\Xi_{l-1}}\frac{1-P_{l-1}}{1-P_{l-1}'}|X_0\right]. \tag{17}$$

Now inserting all intermediate expressions we obtain

$$\begin{aligned}
E\left[Y_t^1|\mathcal{T}=c\right] &= \frac{\int(E[Y_tD|X_0,Z=1]-E[Y_tD|X_0,Z=0])\,dF_{X_0}}{\int(E[D|X_0,Z=1]-E[D|X_0,Z=0])\,dF_{X_0}}\\
&= \frac{\int\left(E\left[Y_tD\frac{Z}{\pi}|X_0\right]-E\left[Y_tD\frac{1-Z}{1-\pi}|X_0\right]\right)dF_{X_0}}{E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right]}\\
&= E\left[Y_tR_1\cdots R_tD\prod_{l=1}^{t}\frac{1}{\Xi_{l-1}}\left(\frac{Z}{\pi}\prod_{l=1}^{t}\frac{P_{l-1}}{P_{l-1}'}-\frac{1-Z}{1-\pi}\prod_{l=1}^{t}\frac{1-P_{l-1}}{1-P_{l-1}'}\right)\right]/E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right].
\end{aligned}$$

An alternative expression can be obtained by noting that via Bayes theorem

$$\frac{P_t}{P_t'}=\frac{\Xi_t}{\Pr(R_{t+1}=1|X_0,\underline{X}_t,\underline{R}_t=1,Z=1,D=1)}$$

and

$$\frac{1-P_t}{1-P_t'}=\frac{\Xi_t}{\Pr(R_{t+1}=1|X_0,\underline{X}_t,\underline{R}_t=1,Z=0,D=1)}$$

such that we can also write

$$\begin{aligned}
&E\left[Y_t^1|\mathcal{T}=c\right]\\
=\;& E\left[Y_tR_1\cdots R_tD\left(\frac{Z}{\pi}\prod_{l=1}^{t}\frac{1}{\Xi_{l-1,Z=1}}-\frac{1-Z}{1-\pi}\prod_{l=1}^{t}\frac{1}{\Xi_{l-1,Z=0}}\right)\right]/E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right],
\end{aligned}$$

where

$$\Xi_{t,Z=z}=\Pr(R_{t+1}=1|X_0,\underline{X}_t,\underline{R}_t=1,D=1,Z=z).$$

E Estimation and asymptotic properties

The potential outcomes (2) and (4) can be nonparametrically estimated by replacing the outer expectations by sample averages and the conditional probabilities by nonparametric (e.g. kernel regression) estimators thereof. For the sake of brevity, we focus on $E\left[Y_{t=2}^1|\mathcal{T}=c\right]$, while estimation of $E\left[Y_{t=2}^0|\mathcal{T}=c\right]$ is obtained by replacing D with $1-D$ and $D=1$ with $D=0$ everywhere. Estimators of the potential outcomes in period $t=1$ can be derived analogously and the corresponding formulae follow naturally from the discussion below.

E.1 Estimation under conditional latent ignorability

To simplify the exposition, we write expression (2) in Theorem 1 as

$$E\left[Y^1_{t=2}|\mathcal{T}=c\right]=\theta=\frac{\Delta}{\Gamma}=\frac{\Delta_1-\Delta_2}{\Gamma} \tag{18}$$

where

$$\Delta_1=E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}DZ}{\pi\Xi_0\Xi_1\frac{P_0'-\pi}{P_0-\pi}\frac{P_1'-\pi}{P_1-\pi}}\right]\qquad \Delta_2=E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}D(1-Z)}{(1-\pi)\,\Xi_0\Xi_1\frac{P_0'-\pi}{P_0-\pi}\frac{P_1'-\pi}{P_1-\pi}}\right]\qquad \Gamma=E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right].$$

Hence, Δ_1 and Δ_2 are weighted averages of $Y_{t=2}$ conditional on $D=1$, $Z=1$ and $D=1$, $Z=0$, respectively. We first examine Δ_1 and note that the other terms can be analyzed analogously. In particular, the expressions for Δ_2 are obtained by redefining Z as $1-Z$ in all expressions involved in Δ_1. We estimate Δ_1 by the sample average

$$\hat{\Delta}_1=\frac{1}{n}\sum_i\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\hat{\pi}(X_{0i})\hat{\Xi}_0(X_{0i})\hat{\Xi}_1(X_{0i},X_{1i})\frac{\hat{P}_0'(X_{0i})-\hat{\pi}(X_{0i})}{\hat{P}_0(X_{0i})-\hat{\pi}(X_{0i})}\frac{\hat{P}_1'(X_{0i},X_{1i})-\hat{\pi}(X_{0i})}{\hat{P}_1(X_{0i},X_{1i})-\hat{\pi}(X_{0i})}},$$

where $\hat{\pi}$, $\hat{\Xi}_0$, $\hat{\Xi}_1$, $\hat{P}_0'$, $\hat{P}_0$, $\hat{P}_1'$, and $\hat{P}_1'$ are nonparametric estimators.

For attaining $\sqrt{n}$-consistency, all nonparametric components have to be estimated with a convergence rate faster than $n^{-\frac{1}{4}}$ such that their bias is sufficiently small. To achieve this rate we use kernel regression (local constant or local linear estimation) with *higher order* product kernels, which in practice is easier to implement than higher-order local polynomial regression, in particular when there are many covariates in X. Note, however, that the asymptotic distribution does not depend on the specific nonparametric estimator used as long as particular regularity conditions are met. We define the kernel regression estimators

$$\hat{\Xi}_0(x_0)=\frac{\sum R_{1i}K_{0i}(x_0)D_i}{\sum K_{0i}(x_0)D_i}\qquad \hat{\Xi}_1(x_0,x_1)=\frac{\sum R_{2i}K_{1i}(x_0,x_1)R_{1i}D_i}{\sum K_{1i}(x_0,x_1)R_{1i}D_i}$$
$$\hat{P}_0(x_0)=\frac{\sum Z_iK_{0i}(x_0)D_i}{\sum K_{0i}(x_0)D_i}\qquad \hat{P}_1(x_0,x_1)=\frac{\sum Z_iK_{1i}(x_0,x_1)R_{1i}D_i}{\sum K_{1i}(x_0,x_1)R_{1i}D_i}$$
$$\hat{P}_0'(x_0)=\frac{\sum Z_iK_{0i}(x_0)R_{1i}D_i}{\sum K_{0i}(x_0)R_{1i}D_i}\qquad \hat{P}_1'(x_0,x_1)=\frac{\sum Z_iK_{1i}(x_0,x_1)R_{1i}R_{2i}D_i}{\sum K_{1i}(x_0,x_1)R_{1i}R_{2i}D_i}$$
$$\hat{\pi}(x_0)=\frac{\sum Z_iK_{0i}(x_0)}{\sum K_{0i}(x_0)}$$

where

$$K_{0i}\left(x_0\right)=\prod_{l=1}^{\dim(x_0)}\kappa_0\left(\frac{X_{0i,l}-x_{0,l}}{h_0}\right) \tag{19}$$

with $x_{0,l}$ being the l-th component of the vector x_0 and $X_{0i,l}$ being the l-th component of the vector X_{0i}, and

$$K_{1i}\left(x_0,x_1\right)=\left\{\prod_{l=1}^{\dim(x_0)}\kappa_1\left(\frac{X_{0i,l}-x_{0,l}}{h_1}\right)\right\}\left\{\prod_{l=1}^{\dim(x_1)}\kappa_1\left(\frac{X_{1i,l}-x_{1,l}}{h_1}\right)\right\} \tag{20}$$

where κ_0 and κ_1 are univariate kernel functions. (Alternatively we could also use local linear regression estimators.)

Note that different conditions are required for the estimation of $\pi(x_0)$, $P_0'(x_0)$, $P_0(x_0)$, $\Xi_0(x_0)$ and $P_1'(x_0, x_1)$, $P_1'(x_0, x_1)$, $\Xi_1(x_0, x_1)$, because estimation of P_1', P_1', and Ξ_1 involves a larger number of covariates as we regress on X_0 *and* X_1. For estimating P_1', P_1', and Ξ_1 we thus require stricter bandwidth conditions and possibly a kernel of higher order. Let L_0 be the number of continuous covariates in X_0 and L_1 the number of continuous covariates in (X_0, X_1). The asymptotic theory only depends on *continuous* covariates. If X contains both continuous and discrete regressors, we follow the generalized kernel approach of Racine and Li (2004) and extend (19) and (20) to multiply with smoothing weights for the discrete regressors. To ease the exposition, we assume that all covariates are continuously distributed in the proofs.

Define the order of a kernel function κ as λ if $\int u^t \kappa(u)du = 0$ for $0 < t < \lambda$ and $\int u^t \kappa(u)du > 0$ for $t = \lambda$. Conventional kernels are of order 2. We will require that $nh^{L_0}/\ln n \to \infty$ and $nh^{L_1}/\ln n \to \infty$ and also that $nh^{2\lambda_0} \to 0$ and $nh^{2\lambda_1} \to 0$. These conditions jointly require that $\lambda_0 > L_0/2$ and $\lambda_1 > L_1/2$. Hence, if (X_0, X_1) contain four or more continuous regressors, a higher order kernel κ_1 is required. The same applies to κ_0 if X_0 includes four or more continuous regressors. Obviously, the bandwidth and kernel conditions are stricter for the later period than for the earlier one. In practice, though, we may use the kernel of the later period for both periods. I.e. for convenience, we may choose $\kappa_0 = \kappa_1$ and $h_0 = h_1$. Assumption 4 summarizes the regularity conditions for kernel regression and along with those required for estimation under Theorem 2.

Assumption 4: Regularity conditions

- (i) The data are iid with $E[|Y_t|] < \infty$ and X_0 and X_1 being supported on a compact set.

- (ii) Support conditions: All conditional probabilities π, Ξ_0, Ξ_1, $\Xi_{0,Z=1}$, $\Xi_{1,Z=1}$, $\Xi_{0,Z=0}$, $\Xi_{1,Z=0}$ and $P_0' - \pi$ and $P_1' - \pi$ are *strictly* bounded away from zero.

- (iii) Smoothness:

- $\pi(x_0)$, $P_0'(x_0)$, $P_0(x_0)$, $\Xi_0(x_0)$, $\Xi_{0,Z=1}(x_0)$ are $\lambda_0 - 1$ times continuously differentiable with the $(\lambda_0 - 1)$-th derivative Hölder continuous.

- $P_1'(x_0, x_1)$, $P_1'(x_0, x_1)$, $\Xi_1(x_0, x_1)$, $\Xi_{1,Z=1}(x_0, x_1)$ are $\lambda_1 - 1$ times continuously differentiable with the $(\lambda_1 - 1)$-th derivative Hölder continuous.

(iv) Kernel functions: κ_0 and κ_1 are compactly supported, bounded, Lipschitz, integrating to one, univariate kernel functions, and of order λ_0 and λ_1, respectively.

(v) Bandwidth conditions: $nh^{L_0}/\ln n \to \infty$ and $nh^{L_1}/\ln n \to \infty$ and $nh^{2\lambda_0} \to 0$ and $nh^{2\lambda_1} \to 0$.

To derive the asymptotic properties of $\hat{\Delta}_1$, we follow Newey (1994) and first derive the influence function and then use his Lemma 5.1 to verify that the remainder terms are sufficiently small. From Newey (1994) it then also follows that our proposed estimators are efficient. The derivations and proofs are given in the appendix, which imply that under our regularity conditions, the estimator $\hat{\Delta}_1$ is asymptotically linear:

$$\sqrt{n}(\hat{\Delta}_1 - \Delta_1) = \frac{1}{\sqrt{n}} \sum \psi_{\Delta_1,i} + o_p(1)$$

with the mean-zero influence function ψ_{Δ_1} being

$$\begin{aligned}
\psi_{\Delta_1,i} = & \left(\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i \Xi_{0i}\Xi_{1i}\frac{P'_{0i}-\pi_i}{P_{0i}-\pi_i}\frac{P'_{1i}-\pi_i}{P_{1i}-\pi_i}} - \Delta_1 \right) \\
& + E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}} \left(-\frac{1}{\pi} + \frac{P_0-P'_0}{(P'_0-\pi)(P_0-\pi)} + \frac{P_1-P'_1}{(P'_1-\pi)(P_1-\pi)} \right) |X_{0i} \right] (Z_i - \pi_i) \\
& \quad + E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}} |X_{0i} \right] \frac{D_i\,(Z_i-\pi_i)}{E\,[R_1D|X_{0i}]} \left(\frac{\Xi_{0i}}{P_{0i}-\pi_i} - \frac{R_{1i}}{P'_{0i}-\pi_i} \right) \\
& \quad + E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}} |X_{0i}, X_{1i} \right] \frac{R_{1i}D_i\,(Z_i-\pi_i)}{E\,[R_2R_1D|X_{0i}, X_{1i}]} \left(\frac{\Xi_{1i}}{P_{1i}-\pi_i} - \frac{R_{2i}}{P'_{1i}-\pi_i} \right)
\end{aligned}$$

where $\pi_i = \pi(X_{0i})$, $\Xi_{0i} = \Xi_0(X_{0i})$, $\Xi_{1i} = \Xi_1(X_{0i}, X_{1i})$, $P'_{0i} = P'_0(X_{0i})$, $P_{0i} = P_0(X_{0i})$, $P'_{1i} = P'_1(X_{0i}, X_{1i})$, $P_{1i} = P_1(X_{0i}, X_{1i})$. Using a central limit theorem for iid observations it thus follows that

$$\sqrt{n}(\hat{\Delta}_1 - \Delta_1) \longrightarrow N(0, E[\psi^2_{\Delta_1,i}]).$$

The previous variance expression appears complicated, which is due to the large number of terms that have to be nonparametrically estimated for $\hat{\Delta}_1$. The influence function ψ_{Δ_1} consists of four terms. If the conditional probabilities π, Ξ_0, Ξ_1, P_0, P'_0, P_1, P'_1 were all known, the last three terms would be zero and the variance would simplify to $E\left[\left(\frac{Y_{t=2}R_{t=1}R_{t=2}DZ}{\pi\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}} - \Delta_1 \right)^2 \right]$, which is the variance stemming from replacing the expectation in $\Delta_1 = E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}DZ}{\pi\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}} \right]$ by a sample mean. The other terms in ψ_{Δ_1} come from the nonparametric estimation of the conditional probabilities. The last term in ψ_{Δ_1} stems from the estimation of Ξ_1, the third term from Ξ_0, and the second term from the estimation of $\pi\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}$. This second term contains three sub-terms in the inner brackets: The first arises through the nonparametric estimation of π, whereas the second and third represent the influences through estimation of P_0 and P_0' as well P_1 and P_1' , respectively.

We can obtain analogous results for Δ_2 and Γ in (18) and combine these expressions to compute the influence function for $E\left[Y^1_{t=2}|\mathcal{T}=c\right]$. As a consequence, $\hat{E}\left[Y^1_{t=2}|\mathcal{T}=c\right]$ can be shown to be $\sqrt{n}$-consistent and asymptotically normal under Assumptions 1 to 4, as stated in Theorem 3.

Theorem 3 *Under Assumptions 1 to 4, the estimator* $\hat{E}\left[Y^1_{t=2}|\mathcal{T}=c\right]$ *defined as* $\frac{\hat{\Delta}_1-\hat{\Delta}_2}{\hat{\Gamma}}$ *in (18) of* $E\left[Y^1_{t=2}|\mathcal{T}=c\right]=\frac{\Delta}{\Gamma}$ *is asymptotically normal:*

$$\sqrt{n}\left(\hat{E}\left[Y^1_{t=2}|\mathcal{T}=c\right]-E\left[Y^1_{t=2}|\mathcal{T}=c\right]\right)\longrightarrow N\left(0,E\left[\left(\frac{1}{\Gamma}\psi_{\Delta_i}-\frac{\Delta}{\Gamma^2}\psi_{\Gamma_i}\right)^2\right]\right),$$

where

$$\begin{aligned}
\psi_{\Delta_i} &= \frac{Y_{2i}R_{2i}R_{1i}D_i}{\Xi_{0i}\Xi_{1i}\frac{P'_{0i}-\pi_i}{P_{0i}-\pi_i}\frac{P'_{1i}-\pi_i}{P_{1i}-\pi_i}}\frac{Z_i-\pi_i}{\pi_i(1-\pi_i)}-\Delta \\
&+E\left[\frac{Y_2R_2R_1D}{\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}}\frac{Z-\pi}{\pi(1-\pi)}\left(\frac{P_0-P'_0}{(P'_0-\pi)(P_0-\pi)}+\frac{P_1-P'_1}{(P'_1-\pi)(P_1-\pi)}-\frac{Z-\pi}{\pi(1-\pi)}\right)|X_{0i}\right](Z_i-\pi_i) \\
&\quad+E\left[\frac{Y_2R_2R_1D}{\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}}\frac{Z-\pi}{\pi(1-\pi)}|X_{0i}\right]\frac{D_i(Z_i-\pi_i)}{E[R_1D|X_{0i}]}\left(\frac{\Xi_{0i}}{P_{0i}-\pi_i}-\frac{R_{1i}}{P'_{0i}-\pi_i}\right) \\
&\quad+E\left[\frac{Y_2R_2R_1D}{\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}}\frac{Z-\pi}{\pi(1-\pi)}|X_{0i},X_{1i}\right]\frac{R_{1i}D_i(Z_i-\pi_i)}{E[R_2R_1D|X_{0i},X_{1i}]}\left(\frac{\Xi_{1i}}{P_{1i}-\pi_i}-\frac{R_{2i}}{P'_{1i}-\pi_i}\right)
\end{aligned}$$

and

$$\psi_{\Gamma_i}=D_i\frac{Z_i-\pi_i}{\pi_i(1-\pi_i)}-\Gamma\quad-\quad E\left[D\frac{(Z-\pi)^2}{\pi^2(1-\pi)^2}|X_{0i}\right](Z_i-\pi_i). \tag{21}$$

E.2 Estimation under missing at random

We can write expression (4) in Theorem 2 as

$$E\left[Y^1_{t=2}|\mathcal{T}=c\right]=\phi=\frac{\Lambda}{\Gamma}=\frac{\Lambda_1-\Lambda_2}{\Gamma}, \tag{22}$$

where

$$\Lambda_1=E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}\right]\qquad \Lambda_2=E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}D(1-Z)}{(1-\pi)\Xi_{0,Z=0}\Xi_{1,Z=0}}\right]\qquad \Gamma=E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right].$$

We briefly sketch the estimation of Λ_1 and note that the other terms can be analyzed analogously: the expressions for Λ_2 are obtained by redefining Z as $1-Z$ in all terms involved in Λ_1.

The term Γ has already been analyzed in the previous subsection. We estimate Λ_1 by the sample average

$$\hat{\Lambda}_1 = \frac{1}{n}\sum \frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\hat{\pi}(X_{0i})\hat{\Xi}_{0,Z=1}(X_{0i})\hat{\Xi}_{1,Z=1}(X_{0i},X_{1i})},$$

where $\hat{\pi}$, $\hat{\Xi}_{0,Z=1}$, $\hat{\Xi}_{1,Z=1}$ are nonparametric kernel estimators. $\hat{\pi}$ is defined as in the previous subsection. The estimators $\hat{\Xi}_{0,Z=1}$ and $\hat{\Xi}_{1,Z=1}$ are similar to the previously used $\hat{\Xi}_0$ and $\hat{\Xi}_1$, but also condition on Z:

$$\hat{\Xi}_{0,Z=1}(x_0) = \frac{\sum R_{1i}K_{0i}(x_0)D_iZ_i}{\sum K_{0i}(x_0)D_iZ_i} \qquad \hat{\Xi}_{1,Z=1}(x_0,x_1) = \frac{\sum R_{2i}K_{1i}(x_0,x_1)R_{1i}D_iZ_i}{\sum K_{1i}(x_0,x_1)R_{1i}D_iZ_i}.$$

The derivations of the asymptotic properties of $\hat{E}\left[Y^1_{t=2}|\mathcal{T}=c\right]$ stated in Theorem 4 are similar to those of the previous subsection and again based on Newey (1994). Details are given in the appendix.

Theorem 4 *Under Assumptions 1', 2', 3, and 4, the estimator* $\hat{E}\left[Y^1_{t=2}|\mathcal{T}=c\right]$ *defined as* $\frac{\hat{\Lambda}_1-\hat{\Lambda}_2}{\hat{\Gamma}}$ *in (22) of* $E\left[Y^1_{t=2}|\mathcal{T}=c\right]=\frac{\Lambda}{\Gamma}$ *is asymptotically normal:*

$$\sqrt{n}\left(\hat{E}\left[Y^1_{t=2}|\mathcal{T}=c\right]-E\left[Y^1_{t=2}|\mathcal{T}=c\right]\right)\longrightarrow N\left(0,E\left[\left(\frac{1}{\Gamma}\psi_{\Lambda_i}-\frac{\Lambda}{\Gamma^2}\psi_{\Gamma_i}\right)^2\right]\right),$$

where

$$\begin{aligned}\psi_{\Lambda_i} = {} & \frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i\Xi_{0i,Z=1}\Xi_{1i,Z=1}} - \frac{Y_{2i}R_{2i}R_{1i}D_i(1-Z_i)}{(1-\pi_i)\Xi_{0i,Z=0}\Xi_{1i,Z=0}} - \Lambda \\ & + E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]\left\{\frac{\pi_i-Z_i}{\pi_i}+D_iZ_i\frac{\Xi_{0i,Z=1}-R_{1i}}{E\left[R_1DZ|X_{0i}\right]}\right\} \\ & - E\left[\frac{Y_2R_2R_1D(1-Z)}{(1-\pi)\Xi_{0,Z=0}\Xi_{1,Z=0}}|X_{0i}\right]\left\{\frac{Z_i-\pi_i}{1-\pi_i}+D_i(1-Z_i)\frac{\Xi_{0i,Z=0}-R_{1i}}{E\left[R_1D(1-Z)|X_{0i}\right]}\right\} \\ & + E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i},X_{1i}\right]R_{1i}D_iZ_i\frac{\Xi_{1i,Z=1}-R_{2i}}{E\left[R_2R_1DZ|X_{0i},X_{1i}\right]} \\ & - E\left[\frac{Y_2R_2R_1D(1-Z)}{(1-\pi)\Xi_{0,Z=0}\Xi_{1,Z=0}}|X_{0i},X_{1i}\right]R_{1i}D_i(1-Z_i)\frac{\Xi_{1i,Z=0}-R_{2i}}{E\left[R_2R_1D(1-Z)|X_{0i},X_{1i}\right]},\end{aligned}$$

with $\Xi_{0i,Z=z}=\Xi_{0,Z=z}(X_{0i})$ *and* $\Xi_{1i,Z=z}=\Xi_{1,Z=z}(X_{0i},X_{1i})$ *and* ψ_{Γ_i} *as defined in (21).*

Similarly to Section 3.1, the structure of the variance term reflects the large number of conditional probabilities that are estimated nonparametrically. Again, the first term in ψ_{Λ_i} arises because the expectation in $\Lambda = E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}-\frac{Y_{t=2}R_{t=1}R_{t=2}D(1-Z)}{(1-\pi)\Xi_{0,Z=0}\Xi_{1,Z=0}}\right]$ is replaced by a sample mean. The other terms in ψ_{Λ_i} reflect the influences due to the nonparametric estimation of π, $\Xi_{0,Z=1}$, $\Xi_{1,Z=1}$, $\Xi_{0,Z=0}$ and $\Xi_{1,Z=0}$.

F Proofs of Asymptotic Distributions

To derive the asymptotic properties of the proposed estimators, we follow the procedure of Newey (1994): We first derive the influence function, and thereafter specify regularity conditions which ensure that the remainder terms are of sufficiently small order. From Newey (1994) it thus also follows that our proposed estimators are efficient. The expressions of Theorems 1 and 2 can be written as sample averages of Y_t weighted by several plug-in nonparametric estimates of conditional probabilities. The precise expressions differ for Theorem 1 and 2.

F.1 Derivation of the influence functions

F.1.1 Estimation under conditional latent ignorability

To simplify notation, we write expression (2) of Theorem 1 as

$$E\left[Y_{t=2}^{1}|\mathcal{T}=c\right]=\frac{\Delta}{\Gamma}=\frac{\Delta_1-\Delta_2}{\Gamma} \tag{23}$$

where

$$\Delta_1=E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}DZ}{\pi\Xi_0\Xi_1\frac{P_0'-\pi}{P_0-\pi}\frac{P_1'-\pi}{P_1-\pi}}\right]\qquad \Delta_2=E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}D(1-Z)}{(1-\pi)\,\Xi_0\Xi_1\frac{P_0'-\pi}{P_0-\pi}\frac{P_1'-\pi}{P_1-\pi}}\right]\qquad \Gamma=E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right].$$

We first examine Δ_1 and note that the other terms can analogously be analyzed. Note that we can write Δ_1 as

$$\Delta_1=E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_0\Xi_1\frac{P_0'-\pi}{P_0-\pi}\frac{P_1'-\pi}{P_1-\pi}}\right]=E\left[\frac{Y_2R_2R_1DZ}{\pi\frac{E[ZR_1D|X_0]-\pi E[R_1D|X_0]}{E[ZD|X_0]-\pi E[D|X_0]}\frac{E[ZR_2R_1D|X_0,X_1]-\pi E[R_2R_1D|X_0,X_1]}{E[ZR_1D|X_0,X_1]-\pi E[R_1D|X_0,X_1]}}\right]$$

where we used

$$\begin{aligned}
\Xi_0&=\tfrac{E[R_1D|X_0]}{E[D|X_0]} & \Xi_1&=\tfrac{E[R_2R_1D|X_0,X_1]}{E[R_1D|X_0,X_1]}\\
P_0'&=\tfrac{E[ZR_1D|X_0]}{E[R_1D|X_0]} & P_1'&=\tfrac{E[ZR_2R_1D|X_0,X_1]}{E[R_2R_1D|X_0,X_1]}\\
P_0&=\tfrac{E[ZD|X_0]}{E[D|X_0]} & P_1&=\tfrac{E[ZR_1D|X_0,X_1]}{E[R_1D|X_0,X_1]}
\end{aligned}\;.$$

Hence, we have expressed Δ_1 as a sample average weighted by various conditional probabilities, either conditioning on X_0 or on X_0, X_1. Hence, we can directly follow Newey (1994) to calculate the influence function $\psi_{\Delta_1,i}$ with $E[\psi_{\Delta_1,i}]=0$ as:

$$\psi_{\Delta_1,i} = \frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i\Xi_{0i}\Xi_{1i}\frac{P'_{0i}-\pi_i}{P_{0i}-\pi_i}\frac{P'_{1i}-\pi_i}{P_{1i}-\pi_i}} - \Delta_1$$

$$+ E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}}\left(-\frac{1}{\pi}+\frac{P_0-P'_0}{(P'_0-\pi)(P_0-\pi)}+\frac{P_1-P'_1}{(P'_1-\pi)(P_1-\pi)}\right)|X_{0i}\right](Z_i-\pi_i)$$

$$+ E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}}|X_{0i}\right]\frac{D_i(Z_i-\pi_i)}{E[R_1D|X_{0i}]}\left(\frac{\Xi_{0i}}{P_{0i}-\pi_i}-\frac{R_{1i}}{P'_{0i}-\pi_i}\right)$$

$$+ E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}}|X_{0i},X_{1i}\right]\frac{R_{1i}D_i(Z_i-\pi_i)}{E[R_2R_1D|X_{0i},X_{1i}]}\left(\frac{\Xi_{1i}}{P_{1i}-\pi_i}-\frac{R_{2i}}{P'_{1i}-\pi_i}\right)$$

where $\pi_i=\pi(X_{0i})$, $\Xi_{0i}=\Xi_0(X_{0i})$, $\Xi_{1i}=\Xi_1(X_{0i},X_{1i})$, $P'_{0i}=P'_0(X_{0i})$, $P_{0i}=P_0(X_{0i})$, $P'_{1i}=P'_1(X_{0i},X_{1i})$, $P_{1i}=P_1(X_{0i},X_{1i})$.

Having obtained the influence function for Δ_1 we obtain by analogous calculations the influence function for Δ_2 by replacing Z with $1-Z$ everywhere. Note that the weights $\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}$ are unchanged if we replace Z with $1-Z$. After some calculations we obtain

$$\psi_{\Delta_2,i} = \frac{Y_{2i}R_{2i}R_{1i}D_i(1-Z_i)}{(1-\pi_i)\Xi_{0i}\Xi_{1i}\frac{P'_{0i}-\pi_i}{P_{0i}-\pi_i}\frac{P'_{1i}-\pi_i}{P_{1i}-\pi_i}} - \Delta_2$$

$$+ E\left[\frac{Y_2R_2R_1D(1-Z)}{(1-\pi)\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}}\left(\frac{1}{1-\pi}+\frac{P_0-P'_0}{(P'_0-\pi)(P_0-\pi)}+\frac{P_1-P'_1}{(P'_1-\pi)(P_1-\pi)}\right)|X_{0i}\right](Z_i-\pi_i)$$

$$+ E\left[\frac{Y_2R_2R_1D(1-Z)}{(1-\pi)\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}}|X_{0i}\right]\frac{D_i(Z_i-\pi_i)}{E[R_1D|X_{0i}]}\left(\frac{\Xi_{0i}}{P_{0i}-\pi_i}-\frac{R_{1i}}{P'_{0i}-\pi_i}\right)$$

$$+ E\left[\frac{Y_2R_2R_1D(1-Z)}{(1-\pi)\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}}|X_{0i},X_{1i}\right]\frac{R_{1i}D_i(Z_i-\pi_i)}{E[R_2R_1D|X_{0i},X_{1i}]}\left(\frac{\Xi_{1i}}{P_{1i}-\pi_i}-\frac{R_{2i}}{P'_{1i}-\pi_i}\right)$$

The influence function for the numerator in (23) is accordingly

$$\psi_{\Delta,i} = \psi_{\Delta_1,i} - \psi_{\Delta_2,i} \tag{24}$$

$$= \frac{Y_{2i}R_{2i}R_{1i}D_i}{\Xi_{0i}\Xi_{1i}\frac{P'_{0i}-\pi_i}{P_{0i}-\pi_i}\frac{P'_{1i}-\pi_i}{P_{1i}-\pi_i}}\frac{Z_i-\pi_i}{\pi_i\left(1-\pi_i\right)}-\Delta$$

$$+E\left[\frac{Y_2R_2R_1D}{\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}}\frac{Z-\pi}{\pi\left(1-\pi\right)}\left(\frac{P_0-P'_0}{(P'_0-\pi)(P_0-\pi)}+\frac{P_1-P'_1}{(P'_1-\pi)(P_1-\pi)}-\frac{Z-\pi}{\pi\left(1-\pi\right)}\right)|X_{0i}\right]\left(Z_i-\pi_i\right)$$

$$+E\left[\frac{Y_2R_2R_1D}{\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}}\frac{Z-\pi}{\pi\left(1-\pi\right)}|X_{0i}\right]\frac{D_i\left(Z_i-\pi_i\right)}{E\left[R_1D|X_{0i}\right]}\left(\frac{\Xi_{0i}}{P_{0i}-\pi_i}-\frac{R_{1i}}{P'_{0i}-\pi_i}\right)$$

$$+E\left[\frac{Y_2R_2R_1D}{\Xi_0\Xi_1\frac{P'_0-\pi}{P_0-\pi}\frac{P'_1-\pi}{P_1-\pi}}\frac{Z-\pi}{\pi\left(1-\pi\right)}|X_{0i},X_{1i}\right]\frac{R_{1i}D_i\left(Z_i-\pi_i\right)}{E\left[R_2R_1D|X_{0i},X_{1i}\right]}\left(\frac{\Xi_{1i}}{P_{1i}-\pi_i}-\frac{R_{2i}}{P'_{1i}-\pi_i}\right).$$

Influence function for denominator in (23): By analogous calculations we can derive the influence function for the denominator in (23)

$$\Gamma=E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right]=E\left[\frac{DZ}{\pi}\right]-E\left[D\frac{1-Z}{1-\pi}\right]$$

which is

$$\psi_{\Gamma_i} = \frac{D_iZ_i}{\pi(X_{0i})}-\frac{D_i(1-Z_i)}{1-\pi(X_{0i})}-\Gamma - E\left[\frac{DZ}{\pi(X_0)^2}+\frac{D(1-Z)}{(1-\pi(X_0))^2}|X_{0i}\right]\left(Z_i-\pi(X_{0i})\right) \quad (25)$$

$$= D_i\frac{Z_i-\pi_i}{\pi_i\left(1-\pi_i\right)}-\Gamma - E\left[D\frac{\left(Z-\pi\right)^2}{\pi^2\left(1-\pi\right)^2}|X_{0i}\right]\left(Z_i-\pi(X_{0i})\right) \quad (26)$$

with $E[\psi_{\Gamma_i}]=0$.

Influence function for (23): Having obtained the influence function for numerator and denominator in (23) we obtain the asymptotic distribution for $E\left[Y^1_{t=2}|\mathcal{T}=c\right]=\frac{\Delta}{\Gamma}$ via the delta formula. The leading term in the approximation to $\sqrt{n}\left(\frac{\hat\Delta}{\hat\Gamma}-\frac{\Delta}{\Gamma}\right)$ is thus

$$\sqrt{n}\left(\frac{\hat\Delta}{\hat\Gamma}-\frac{\Delta}{\Gamma}\right)=\sqrt{n}\left(\frac{\hat\Delta-\Delta}{\Gamma}-\frac{\Delta}{\Gamma}\cdot\frac{\hat\Gamma-\Gamma}{\Gamma}\right)$$

such that the influence function of $E\left[Y^1_{t=2}|\mathcal{T}=c\right]$ is

$$\frac{1}{\Gamma}\psi_{\Delta,i}-\frac{\Delta}{\Gamma^2}\psi_{\Gamma_i}$$

and thus the asymptotic variance of $\hat{E}\left[Y^1_{t=2}|\mathcal{T}=c\right]$ is

$$E\left[\left(\frac{1}{\Gamma}\psi_{\Delta,i}-\frac{\Delta}{\Gamma^2}\psi_{\Gamma_i}\right)^2\right].$$

F.1.2 Estimation under missing at random

Now we examine estimation when using Theorem 2. To simplify notation, we write the expression of Theorem 2 as

$$E\left[Y^1_{t=2}|\mathcal{T}=c\right]=\frac{\Lambda}{\Gamma}=\frac{\Lambda_1-\Lambda_2}{\Gamma}$$

where

$$\Lambda_1=E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}\right]\qquad\Lambda_2=E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}D(1-Z)}{(1-\pi)\Xi_{0,Z=0}\Xi_{1,Z=0}}\right]\qquad\Gamma=E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right].$$

We first examine Λ_1 and note that the other terms can analogously be analyzed. For deriving the influence function, it is useful to note that we can write

$$\Xi_{0,Z=1}=\frac{E\left[R_1DZ|X_0\right]}{E\left[DZ|X_0\right]}\qquad\Xi_{1,Z=1}=\frac{E\left[R_2R_1DZ|X_0,X_1\right]}{E\left[R_1DZ|X_0,X_1\right]}$$

such that we can express Λ_1 as

$$\Lambda_1=E\left[\frac{Y_2R_2R_1DZ}{\pi\frac{E[R_1DZ|X_0]}{E[DZ|X_0]}\frac{E[R_2R_1DZ|X_0,X_1]}{E[R_1DZ|X_0,X_1]}}\right].$$

Hence, Λ_1 can be expressed as a weighted average weighted by various conditional probabilities, either conditioning on X_0 or on X_0, X_1. Hence, we can directly follow Newey (1994) to calculate the influence function:

$$\begin{aligned}
\psi_{\Lambda_1,i}=&\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi(X_{0i})\Xi_{0,Z=1}(X_{0i})\Xi_{1,Z=1}(X_{0i},X_{1i})}-\Lambda_1\\
&+E\left[-\frac{Y_2R_2R_1DZ}{\pi(X_0)^2\cdot\Xi_{0,Z=1}(X_0)\cdot\Xi_{1,Z=1}(X_0,X_1)}|X_{0i}\right](Z_i-\pi(X_{0i}))\\
&+E\left[-\frac{Y_2R_2R_1DZ}{\pi(X_0)\cdot\Xi_{0,Z=1}(X_0)E\left[R_1DZ|X_0\right]\cdot\Xi_{1,Z=1}(X_0,X_1)}|X_{0i}\right](R_{1i}D_iZ_i-E\left[R_1DZ|X_{0i}\right])\\
&+E\left[\frac{Y_2R_2R_1DZ}{\pi(X_0)E\left[R_1DZ|X_0\right]\cdot\Xi_{1,Z=1}(X_0,X_1)}|X_{0i}\right](D_iZ_i-E\left[DZ|X_{0i}\right])\\
&+E\left[-\frac{Y_2R_2R_1DZ}{\pi(X_0)\Xi_{0,Z=1}(X_0)\cdot\Xi_{1,Z=1}(X_0,X_1)E\left[R_2R_1DZ|X_0,X_1\right]}|X_{0i},X_{1i}\right](R_{2i}R_{1i}D_iZ_i-E\left[R_2R_1DZ|X_{0i},X_{1i}\right])\\
&+E\left[\frac{Y_2R_2R_1DZ}{\pi(X_0)\Xi_{0,Z=1}(X_0)\cdot E\left[R_2R_1DZ|X_0,X_1\right]}|X_{0i},X_{1i}\right](R_{1i}D_iZ_i-E\left[R_1DZ|X_{0i},X_{1i}\right])
\end{aligned}$$

with $E[\psi_{\Lambda_1,i}]=0$.

After simplification we obtain

$$\psi_{\Lambda_1,i} = \frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i\Xi_{0i,Z=1}\Xi_{1i,Z=1}} - \Lambda_1 + E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]\left\{\frac{\pi_i - Z_i}{\pi_i} + D_iZ_i\frac{\Xi_{0i,Z=1} - R_{1i}}{E\left[R_1DZ|X_{0i}\right]}\right\} + E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i},X_{1i}\right]R_{1i}D_iZ_i\frac{\Xi_{1i,Z=1} - R_{2i}}{E\left[R_2R_1DZ|X_{0i},X_{1i}\right]}.$$

By analogous calculations we obtain the influence function for Λ_2 as

$$\psi_{\Lambda_2,i} = \frac{Y_{2i}R_{2i}R_{1i}D_i\left(1-Z_i\right)}{\left(1-\pi_i\right)\Xi_{0i,Z=0}\Xi_{1i,Z=0}} - \Lambda_2 + E\left[\frac{Y_2R_2R_1D\left(1-Z\right)}{\left(1-\pi\right)\Xi_{0,Z=0}\Xi_{1,Z=0}}|X_{0i}\right]\left\{\frac{Z_i-\pi_i}{1-\pi_i} + D_i\left(1-Z_i\right)\frac{\Xi_{0i,Z=0} - R_{1i}}{E\left[R_1D(1-Z)|X_{0i}\right]}\right\} + E\left[\frac{Y_2R_2R_1D\left(1-Z\right)}{\left(1-\pi\right)\Xi_{0,Z=0}\Xi_{1,Z=0}}|X_{0i},X_{1i}\right]R_{1i}D_i\left(1-Z_i\right)\frac{\Xi_{1i,Z=0} - R_{2i}}{E\left[R_2R_1D(1-Z)|X_{0i},X_{1i}\right]}.$$

The influence function for Λ is thus

$$\psi_{\Lambda,i} = \psi_{\Lambda_1,i} - \psi_{\Lambda_2,i}$$

$$= \frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i\Xi_{0i,Z=1}\Xi_{1i,Z=1}} - \frac{Y_{2i}R_{2i}R_{1i}D_i\left(1-Z_i\right)}{\left(1-\pi_i\right)\Xi_{0i,Z=0}\Xi_{1i,Z=0}} - \Lambda + E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]\left\{\frac{\pi_i - Z_i}{\pi_i} + D_iZ_i\frac{\Xi_{0i,Z=1} - R_{1i}}{E\left[R_1DZ|X_{0i}\right]}\right\} - E\left[\frac{Y_2R_2R_1D\left(1-Z\right)}{\left(1-\pi\right)\Xi_{0,Z=0}\Xi_{1,Z=0}}|X_{0i}\right]\left\{\frac{Z_i-\pi_i}{1-\pi_i} + D_i\left(1-Z_i\right)\frac{\Xi_{0i,Z=0} - R_{1i}}{E\left[R_1D(1-Z)|X_{0i}\right]}\right\} + E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i},X_{1i}\right]R_{1i}D_iZ_i\frac{\Xi_{1i,Z=1} - R_{2i}}{E\left[R_2R_1DZ|X_{0i},X_{1i}\right]} - E\left[\frac{Y_2R_2R_1D\left(1-Z\right)}{\left(1-\pi\right)\Xi_{0,Z=0}\Xi_{1,Z=0}}|X_{0i},X_{1i}\right]R_{1i}D_i\left(1-Z_i\right)\frac{\Xi_{1i,Z=0} - R_{2i}}{E\left[R_2R_1D(1-Z)|X_{0i},X_{1i}\right]}.$$

F.2 Analysis of the remainder terms

In the previous section we had derived the influence functions which represent the first order approximations to the estimators. In this section we need to ensure that second order terms are sufficiently small. A direct calculation of the expansions of the estimator would be very involved as the estimators of $E\left[Y_{t=2}^1|\mathcal{T}=c\right]$ contain at least three nonparametric kernel regression components such that sample averages would involve four-fold summations with some additional summation terms appearing in the denominator. This would require higher-order U-statistic projection theorems leading to a rather involved notation. Instead, we follow Lemma 5.1 of Newey (1994), which makes use of a linearization that permits a sequential analysis of the nonparametric components. Therefore we have to verify Assumptions 5.1 to 5.3 of Newey (1994). Theorems 3 and 4 then follow immediately from Lemma 5.1 of Newey (1994). For verifying Assumptions 5.1 to 5.3 we now follow more closely the notation of Newey (1994). Let $\hat{\eta}$ denote the vector of nonparametric regression estimators and η_0 the true regression lines. (Note that Newey (1994) uses the symbol h instead of η, but since h is frequently used as a symbol for the bandwidth value in a kernel regression context, as we also do, we deviate from Newey's notation here.)

For ease of exposition we first analyze the expressions of Theorem 4. The derivations for Theorem 3 are analogous, but more cumbersome.

F.2.1 Estimation under missing at random

Under the missing at random assumption we had identified the potential outcomes in Theorem 2 as

$$E\left[Y^1_{t=2}|\mathcal{T}=c\right] = \frac{\Lambda}{\Gamma} = \frac{\Lambda_1 - \Lambda_2}{\Gamma}$$

where

$$\Lambda_1 = E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}\right] \qquad \Lambda_2 = E\left[\frac{Y_{t=2}R_{t=1}R_{t=2}D(1-Z)}{(1-\pi)\Xi_{0,Z=0}\Xi_{1,Z=0}}\right] \qquad \Gamma = E\left[\frac{D}{\pi}\frac{Z-\pi}{1-\pi}\right].$$

For ease of exposition we only show the derivations for Λ_1 and mention that the other terms Λ_2 and Γ have the same structure such that analogous derivations can be applied.

Using the notation of Newey (1994), the estimator of Λ_1 can be written as the value $\hat{\Lambda}_1$ that solves

$$\frac{1}{n}\sum m(\mathcal{Z}_i, \hat{\Lambda}_1, \hat{\eta}) = 0$$

where $\mathcal{Z}_i = (Y_{2i}, Y_{1i}, R_{2i}, R_{1i}, D_i, Z_i, X_{0i}, X_{1i})$ contains all observed covariates of individual i and the function m is given as

$$m_i(\Lambda_1, \eta) = m(\mathcal{Z}_i, \Lambda_1, \eta) = \frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\eta_{1i}\eta_{2i}\eta_{3i}} - \Lambda_1$$

where η_{1i}, η_{2i} and η_{3i} refer to the respective elements of the vector of nonparametric components:

$$\eta_i = (\pi(X_{0i})\ ,\ \Xi_{0,Z=1}(X_{0i})\ ,\ \Xi_{1,Z=1}(X_{0i}, X_{1i}))\,.$$

Note that the function m_i only depends on η at the values X_{0i} and X_{1i}, which helps to simplify the derivations. Note further that at the true regression lines η_0 we have

$$E\left[m(\mathcal{Z}_i, \Lambda_1, \eta_0)\right] = 0.$$

For verifying Assumptions 5.1 to 5.3 of Newey (1994) we need to linearize. Let $\eta_0 = \left(\eta_{1,0}\ ,\ \eta_{2,0}\ ,\ \eta_{3,0}\right)$ be the vector of the nonparametric components. The linearization term $\mathcal{D}(\mathcal{Z}, \eta - \eta_0)$ is obtained via differentiation with respect to each nonparametric component as:

$$\mathcal{D}(\mathcal{Z}_i, \eta - \eta_0) = -\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\eta_{1i,0}^2\eta_{2i,0}\eta_{3i,0}}\left(\eta_{1i} - \eta_{1i,0}\right) - \frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\eta_{1i,0}\eta_{2i,0}^2\eta_{3i,0}}\left(\eta_{2i} - \eta_{2i,0}\right) - \frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\eta_{1i,0}\eta_{2i,0}\eta_{3i,0}^2}\left(\eta_{3i} - \eta_{3i,0}\right). \tag{27}$$

Assumption 5.1: For verifying Assumption 5.1 of Newey (1994) we have to show that

$$\text{(i)} \qquad \|m(\mathcal{Z}, \Lambda_1, \eta) - m(\mathcal{Z}, \Lambda_1, \eta_0) - \mathcal{D}(\mathcal{Z}, \eta - \eta_0)\| \leq b(\mathcal{Z})\, \|\eta - \eta_0\|^2$$

$$\text{(ii)} \qquad E\left[b(\mathcal{Z})\right]\sqrt{n}\, \|\hat{\eta} - \eta_0\|^2 \xrightarrow{p} 0$$

for all η with $\|\eta - \eta_0\|$ small enough.

We show this for the sup-norm. First note that

$$\|m(\mathcal{Z}, \Lambda_1, \eta) - m(\mathcal{Z}, \Lambda_1, \eta_0) - \mathcal{D}(\mathcal{Z}, \eta - \eta_0)\|$$

$$= \left\| Y_{2i}R_{2i}R_{1i}D_iZ_i\left(\frac{1}{\eta_{1i}\eta_{2i}\eta_{3i}} - \frac{1}{\eta_{1i,0}\eta_{2i,0}\eta_{3i,0}} + \frac{\eta_{1i} - \eta_{1i,0}}{\eta_{1i,0}^2\eta_{2i,0}\eta_{3i,0}} + \frac{\eta_{2i} - \eta_{2i,0}}{\eta_{1i,0}\eta_{2i,0}^2\eta_{3i,0}} + \frac{\eta_{3i} - \eta_{3i,0}}{\eta_{1i,0}\eta_{2i,0}\eta_{3i,0}^2}\right)\right\|$$

$$= ||Y_{2i}R_{2i}R_{1i}D_iZ_i\Big[\frac{1}{\eta_{1i}\eta_{2i}\eta_{3i}} - \frac{1}{\eta_{1i,0}\eta_{2i}\eta_{3i}} + \frac{1}{\eta_{1i,0}\eta_{2i}\eta_{3i}} - \frac{1}{\eta_{1i,0}\eta_{2i,0}\eta_{3i}} + \frac{1}{\eta_{1i,0}\eta_{2i,0}\eta_{3i}} - \frac{1}{\eta_{1i,0}\eta_{2i,0}\eta_{3i,0}}$$
$$+ \frac{\eta_{1i} - \eta_{1i,0}}{\eta_{1i,0}^2\eta_{2i,0}\eta_{3i,0}} + \frac{\eta_{2i} - \eta_{2i,0}}{\eta_{1i,0}\eta_{2i,0}^2\eta_{3i,0}} + \frac{\eta_{3i} - \eta_{3i,0}}{\eta_{1i,0}\eta_{2i,0}\eta_{3i,0}^2}\Big]||$$

and after some calculations we obtain

$$= \quad ||Y_{2i}R_{2i}R_{1i}D_iZ_i\Big[\frac{\left(\eta_{1i,0} - \eta_{1i}\right)^2}{\eta_{1i,0}^2\eta_{1i}\eta_{2i}\eta_{3i}} + \frac{\left(\eta_{1i,0} - \eta_{1i}\right)\left(\eta_{2i,0} - \eta_{2i}\right)}{\eta_{1i,0}^2\eta_{2i}\eta_{2i,0}\eta_{3i}} + \frac{\left(\eta_{1i,0} - \eta_{1i}\right)\left(\eta_{3i,0} - \eta_{3i}\right)}{\eta_{1i,0}^2\eta_{2i,0}\eta_{3i,0}\eta_{3i}}$$
$$+ \frac{\left(\eta_{2i,0} - \eta_{2i}\right)^2}{\eta_{1i,0}\eta_{2i,0}^2\eta_{2i}\eta_{3i}} + \frac{\left(\eta_{2i,0} - \eta_{2i}\right)\left(\eta_{3i,0} - \eta_{3i}\right)}{\eta_{1i,0}\eta_{2i,0}^2\eta_{3i}\eta_{3i,0}} + \frac{\left(\eta_{3i,0} - \eta_{3i}\right)^2}{\eta_{1i,0}\eta_{2i,0}\eta_{3i,0}^2\eta_{3i}}\Big]||$$

$$\leq \quad |Y_{2i}R_{2i}R_{1i}D_iZ_i| \cdot ||\Big[\frac{\left(\eta_{1i,0} - \eta_{1i}\right)^2}{\eta_{1i,0}^2\eta_{1i}\eta_{2i}\eta_{3i}} + \frac{\left(\eta_{1i,0} - \eta_{1i}\right)\left(\eta_{2i,0} - \eta_{2i}\right)}{\eta_{1i,0}^2\eta_{2i}\eta_{2i,0}\eta_{3i}} + \frac{\left(\eta_{1i,0} - \eta_{1i}\right)\left(\eta_{3i,0} - \eta_{3i}\right)}{\eta_{1i,0}^2\eta_{2i,0}\eta_{3i,0}\eta_{3i}}$$
$$+ \frac{\left(\eta_{2i,0} - \eta_{2i}\right)^2}{\eta_{1i,0}\eta_{2i,0}^2\eta_{2i}\eta_{3i}} + \frac{\left(\eta_{2i,0} - \eta_{2i}\right)\left(\eta_{3i,0} - \eta_{3i}\right)}{\eta_{1i,0}\eta_{2i,0}^2\eta_{3i}\eta_{3i,0}} + \frac{\left(\eta_{3i,0} - \eta_{3i}\right)^2}{\eta_{1i,0}\eta_{2i,0}\eta_{3i,0}^2\eta_{3i}}\Big]||$$

$$\leq \quad |Y_{2i}R_{2i}R_{1i}D_iZ_i| \cdot \Big[|| \frac{\left(\eta_{1i,0}-\eta_{1i}\right)^2}{\eta_{1i,0}^2\eta_{1i}\eta_{2i}\eta_{3i}} || + || \frac{\left(\eta_{1i,0}-\eta_{1i}\right)\left(\eta_{2i,0}-\eta_{2i}\right)}{\eta_{1i,0}^2\eta_{2i}\eta_{2i,0}\eta_{3i}} || + || \frac{\left(\eta_{1i,0}-\eta_{1i}\right)\left(\eta_{3i,0}-\eta_{3i}\right)}{\eta_{1i,0}^2\eta_{2i,0}\eta_{3i,0}\eta_{3i}} ||$$
$$+ || \frac{\left(\eta_{2i,0}-\eta_{2i}\right)^2}{\eta_{1i,0}\eta_{2i,0}^2\eta_{2i}\eta_{3i}} || + || \frac{\left(\eta_{2i,0}-\eta_{2i}\right)\left(\eta_{3i,0}-\eta_{3i}\right)}{\eta_{1i,0}\eta_{2i,0}^2\eta_{3i}\eta_{3i,0}} || + || \frac{\left(\eta_{3i,0}-\eta_{3i}\right)^2}{\eta_{1i,0}\eta_{2i,0}\eta_{3i,0}^2\eta_{3i}} || \Big]$$

$$\leq |Y_{2i}R_{2i}R_{1i}D_iZ_i| \cdot C \cdot \|\eta-\eta_0\|^2 \tag{28}$$

with $C < \infty$ because $\eta_{1,0}$, $\eta_{2,0}$, $\eta_{3,0}$ are all strictly bounded away from zero because by Assumption 4(ii) there exists a positive c such that $\inf_{x_0} \pi(x_0) > c$ and $\inf_{x_0} \Xi_{0,Z=1}(x_0) > c$ and $\inf_{x_0,x_1} \Xi_{1,Z=1}(x_0,x_1) > c$. Hence, the inverses will be bounded by a positive finite constant: $\sup_{x_0} \frac{1}{\pi(x_0)} < \frac{1}{c}$ and $\sup_{x_0} \frac{1}{\Xi_{0,Z=1}(x_0)} < \frac{1}{c}$ and $\sup_{x_0,x_1} \frac{1}{\Xi_{1,Z=1}(x_0,x_1)} < \frac{1}{c}$. Note further that for $\|\eta-\eta_0\|$ small enough, also $\sup ||\eta_{1,0}-\eta_1||$ and $\sup ||\eta_{2,0}-\eta_2||$ and $\sup ||\eta_{3,0}-\eta_3||$ will be bounded by some ε. For η very close to η_0 we will have that $\varepsilon < c$. This implies $\sup_{x_0} \frac{1}{\eta_1} < \frac{1}{c-\varepsilon}$ and analogously for $\frac{1}{\eta_2}$ and $\frac{1}{\eta_3}$. Hence choosing C as

$$C = \frac{1}{c^2} \frac{1}{(c-\varepsilon)^3}$$

will satisfy (28) when η is close enough to η_0.

Now we can verify Assumption 5.1 (ii):

$$E\left[b(\mathcal{Z})\right] \sqrt{n}\, \|\hat{\eta}-\eta_0\|^2 \xrightarrow{p} 0$$

By assumption 4(i) Y_t has finite moments. Also note that R_t and D and Z are all binary and thus bounded by definition. In addition, the constant C defined in (28) is finite. Further, by assumption 4 the nonparametric estimators are uniformly consistent and the bandwidth conditions in Assumption 4(v) ensure that the convergence rate is sufficiently rapid. Hence, Assumption 5.1 of Newey (1994) is met.

Assumption 5.2+5.3:

Next we have to verify **Assumption 5.2** and **Assumption 5.3**. We there have to verify that the linearization term captures the influence function previously derived. Given the structure of our estimation problem it will be easiest to verify Assumption 5.2 and 5.3 together. Therefore we have to show that:

$$plim \sqrt{n}\frac{1}{n}\sum \mathcal{D}(\mathcal{Z}_i, \hat{\eta}-\eta_0) - \sqrt{n}\frac{1}{n}\sum \alpha(\mathcal{Z}_i) = 0 \tag{29}$$

for some function $\alpha(\mathcal{Z}_i)$ with $E[\alpha(\mathcal{Z})] = 0$ and $E[||\alpha(\mathcal{Z})||^2] < \infty$.

When constructing the influence function we had followed the procedure of Newey (1994) and obtained the influence function $\psi_{\Lambda_1,i}$ as the sum of $m(\mathcal{Z}_i, \Lambda_1, \eta_0)$ and the correction terms $\alpha(\mathcal{Z}_i)$. Consequently we can obtain the candidate function for $\alpha(\mathcal{Z}_i)$ as $\psi_{\Lambda_1,i} - m(\mathcal{Z}_i, \Lambda_1, \eta_0)$, that is

$$\alpha(\mathcal{Z}_i) = E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]\left\{\frac{\pi_i - Z_i}{\pi_i} + D_iZ_i\frac{\Xi_{0i,Z=1} - R_{1i}}{E\left[R_1DZ|X_{0i}\right]}\right\} + E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}, X_{1i}\right] R_{1i}D_iZ_i\frac{\Xi_{1i,Z=1} - R_{2i}}{E\left[R_2R_1DZ|X_{0i}, X_{1i}\right]}.$$

Inserting $\mathcal{D}(\mathcal{Z}_i, \hat{\eta} - \eta_0)$ as well as the candidate function $\alpha(\mathcal{Z}_i)$ in (29) we obtain:

$$\sqrt{n}\frac{1}{n}\sum \mathcal{D}(\mathcal{Z}_i, \hat{\eta} - \eta_0) - \sqrt{n}\frac{1}{n}\sum \alpha(\mathcal{Z}_i)$$

$$\begin{aligned}
= & \frac{1}{\sqrt{n}}\sum \frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\eta_{1i,0}\eta_{2i,0}\eta_{3i,0}}\left(\frac{\eta_{1i,0} - \hat{\eta}_{1i}}{\eta_{1i,0}} + \frac{\eta_{2i,0} - \hat{\eta}_{2i}}{\eta_{2i,0}} + \frac{\eta_{3i,0} - \hat{\eta}_{3i}}{\eta_{3i,0}}\right) \\
& - \frac{1}{\sqrt{n}}\sum E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]\left\{\frac{\pi_i - Z_i}{\pi_i} + D_iZ_i\frac{\Xi_{0i,Z=1} - R_{1i}}{E\left[R_1DZ|X_{0i}\right]}\right\} \\
& - \frac{1}{\sqrt{n}}\sum E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}, X_{1i}\right] R_{1i}D_iZ_i\frac{\Xi_{1i,Z=1} - R_{2i}}{E\left[R_2R_1DZ|X_{0i}, X_{1i}\right]}
\end{aligned}$$

$$\begin{aligned}
= & \frac{1}{\sqrt{n}}\sum\left\{\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\eta_{1i,0}\eta_{2i,0}\eta_{3i,0}}\frac{\eta_{1i,0} - \hat{\eta}_{1i}}{\eta_{1i,0}} - E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]\left(\frac{\pi_i - Z_i}{\pi_i}\right)\right\} \\
& + \frac{1}{\sqrt{n}}\sum\left\{\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\eta_{1i,0}\eta_{2i,0}\eta_{3i,0}}\frac{\eta_{2i,0} - \hat{\eta}_{2i}}{\eta_{2i,0}} - E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]\left(D_iZ_i\frac{\Xi_{0i,Z=1} - R_{1i}}{E\left[R_1DZ|X_{0i}\right]}\right)\right\} \\
& + \frac{1}{\sqrt{n}}\sum\left\{\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\eta_{1i,0}\eta_{2i,0}\eta_{3i,0}}\frac{\eta_{3i,0} - \hat{\eta}_{3i}}{\eta_{3i,0}} - E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}, X_{1i}\right] R_{1i}D_iZ_i\frac{\Xi_{1i,Z=1} - R_{2i}}{E\left[R_2R_1DZ|X_{0i}, X_{1i}\right]}\right\}
\end{aligned}$$

$$= \frac{1}{\sqrt{n}}\sum\left\{\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\frac{\pi - \hat{\pi}_i}{\pi} - E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]\left(\frac{\pi_i - Z_i}{\pi_i}\right)\right\} \tag{30}$$

$$+ \frac{1}{\sqrt{n}}\sum\left\{\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\frac{\Xi_{0,Z=1} - \hat{\Xi}_{0i,Z=1}}{\Xi_{0,Z=1}} - E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]\left(D_iZ_i\frac{\Xi_{0i,Z=1} - R_{1i}}{E\left[R_1DZ|X_{0i}\right]}\right)\right\} \tag{31}$$

$$+ \frac{1}{\sqrt{n}}\sum\left\{\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\frac{\Xi_{1,Z=1} - \hat{\Xi}_{1i,Z=1}}{\Xi_{1,Z=1}} - E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}, X_{1i}\right] R_{1i}D_iZ_i\frac{\Xi_{1i,Z=1} - R_{2i}}{E\left[R_2R_1DZ|X_{0i}, X_{1i}\right]}\right\} \tag{32}$$

If each of these three terms (30)-(32) converges to zero, then (29) follows. We focus in the following on the expression (30) and note that the corresponding calculations for (31) and (32) are analogous.

Consider the first term in (30)

$$= \frac{1}{n}\sum \frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i^2\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\left(\hat{\pi}_i - \pi(X_{0i})\right) \tag{33}$$

where $\hat{\pi}$ is estimated either by kernel regression or by local linear regression. We display below the proof for $\hat{\pi}$ estimated by local linear regression. The analogous derivations for kernel regression are immediately obtained by deleting the linear term. The results for (33) are the same irrespective of whether one uses kernel regression or local linear regression. Making use of expression (35) and defining

$$\varsigma_{ij} = \left(\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i^2\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\right)\cdot e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{0j,i}$$
$$\times\left((Z_j-\pi_j)+\frac{1}{2}\left(X_{0j}-X_{0i}\right)'\frac{\partial^2\pi\left(X_{0i}\right)}{\partial x\partial x'}\left(X_{0j}-X_{0i}\right)+O\left(h^3\right)\right)$$

we obtain that (33) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i^2\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\right)(\hat{\pi}_i-\pi_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\varsigma_{ij} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\varsigma_{ij}+\varsigma_{ji}}{2}.$$

The latter term is a nondegenerate symmetric von Mises statistic. The von Mises statistic is asymptotically equivalent to the corresponding U-statistic, and its projection is

$$= \frac{2}{n}\sum_{i=1}^{n}\left(E\left[\frac{\varsigma_{ij}+\varsigma_{ji}}{2}|X_{0i},X_{1i},Z_i,D_i,Y_{2i},R_{2i},R_{1i}\right]-E\left[\frac{\varsigma_{ij}+\varsigma_{ji}}{2}\right]\right)+E\left[\frac{\varsigma_{ij}+\varsigma_{ji}}{2}\right]+o_p\left(\frac{1}{\sqrt{n}}\right) \tag{34}$$

under the condition that $E\left[\left(\frac{\varsigma_{ij}+\varsigma_{ji}}{2}\right)^2\right] \le o(n)$, see Serfling (1980, p.190) and Powell, Stock, and Stoker (1989). To verify this condition note that $E[\varsigma_{ij}^2] \le o(n)$ by (41) and that $E\left|\varsigma_{ij}\varsigma_{ji}\right| \le \sqrt{E[\varsigma_{ij}^2]\cdot E[\varsigma_{ji}^2]}$ by Hölder's inequality. From (38) and (40) we obtain that $E\left[\varsigma_{ij}+\varsigma_{ji}|X_{0i},X_{1i},Z_i,D_i,Y_{2i},R_{2i},R_{1i}\right] = \left(E\left[\frac{Y_2R_2R_1DZ}{\pi^2\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]+O(h)\right)\cdot(Z_i-\pi_i)+O(h^\lambda)$ and $E\left[\varsigma_{ij}+\varsigma_{ji}\right]=O(h^\lambda)$. This gives

$$= \frac{1}{n}\sum_{i=1}^{n}\left(E\left[\frac{Y_2R_2R_1DZ}{\pi^2\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]+O_p(h)\right)\cdot(Z_i-\pi_i)+O_p(h^\lambda)+o_p\left(\frac{1}{\sqrt{n}}\right).$$

Now we can use these results to show that the term (30) converges to zero. Inserting the previously obtained expressions into (30) we obtain

$$\sqrt{n}\frac{1}{n}\sum\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\frac{\pi-\hat{\pi}_i}{\pi}-\sqrt{n}\frac{1}{n}\sum E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]\left(\frac{\pi_i-Z_i}{\pi_i}\right)$$

$$= \sqrt{n}\frac{1}{n}\sum(-1)\left\{\left(E\left[\frac{Y_2R_2R_1DZ}{\pi^2\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]+O_p(h)\right)\cdot(Z_i-\pi_i)+O_p(h^\lambda)+o_p\left(\frac{1}{\sqrt{n}}\right)\right\}$$

$$-\sqrt{n}\frac{1}{n}\sum E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]\left(\frac{\pi_i-Z_i}{\pi_i}\right)$$

$$= \sqrt{n}\frac{1}{n}\sum\left(O_p(h)\cdot(Z_i-\pi_i)+O_p(h^\lambda)+o_p\left(\frac{1}{\sqrt{n}}\right)\right)\longrightarrow 0$$

under the bandwidth conditions of Assumption 4.

We thus have shown that the term (30) converges to zero. With analogous derivations one can also show that the terms (31) and (32) converge to zero. This implies that (29) converges to zero. Hence, Assumption 5.2+5.3 of Newey (1994) is met. Theorem 4 thus follows directly from Lemma 5.1 of Newey (1994).

The following subsections provide some preliminaries that have been used in the previous derivations.

F.2.2 Local linear regression

In this subsection we derive some properties of local linear regression that have been used in the previous U-statistics projection theorems. To simplify the notation, we often drop the subscript 0. The following kernel constants will be used later:

$$\mu_t = \int u^t\kappa(u)du \qquad \text{and} \qquad \bar{\mu}_t = \int u^t\kappa^2(u)du.$$

A kernel function being of order λ means that $\mu_t = 0$ for $0 < t < \lambda$ and $\mu_\lambda \neq 0$.

Consider estimation of $\pi(x)$ at a location x. Define the regressor matrices $\mathbb{X}_j = \left(1, \left(\frac{X_{0j}-x}{h}\right)'\right)'$ and $\mathbb{X} = (\mathbb{X}_1, \mathbb{X}_2, ..., \mathbb{X}_n)'$ and the matrix of kernel weights $\mathbb{K} = diag(K_1, K_2, ..., K_n)$. Since $\pi(x)$ is estimated by a weighted least squares regression, we can write the solution as

$$\hat{\pi}(x) = e_1'\left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1}\sum_{j=1}^n \mathbb{X}_jK_jZ_j = e_1'\left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1}\sum_{j=1}^n \mathbb{X}_jK_j\left(Z_j-\pi_j+\pi_j\right)$$

where e_1 is a column vector of zeros with first element being one and $\pi_j = \pi(X_{0j})$. A series expansion gives

$$= e_1'\left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1}\sum_{j=1}^n \mathbb{X}_jK_j\left(Z_j-\pi_j\right)$$

$$+ e_1'\left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1}\sum_{j=1}^n \mathbb{X}_jK_j\left(\pi(x)+(X_{0j}-x)'\frac{\partial\pi(x)}{\partial x}+(X_{0j}-x)'\frac{1}{2}\frac{\partial^2\pi(x)}{\partial x\partial x'}(X_{0j}-x)+R_j\right)$$

where $\frac{\partial \pi(x)}{\partial x}$ is the $L \times 1$ vector of first derivatives and $\frac{\partial^2 \pi(x)}{\partial x \partial x'}$ the $L \times L$ matrix of second derivatives and R_j is the remainder term of all third order derivatives multiplied with the respective third order interaction terms of $X_{0j} - x$. Since K_j has bounded support, the remainder term premultiplied with K_j is of order $O(K_j \cdot h^3)$. We thus obtain after some derivations that

$$= e_1' \left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1} \sum_{j=1}^{n} \mathbb{X}_j K_j \left(Z_j - \pi_j\right) + \pi\left(x\right) + e_1' \left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)^{-1} \sum_{j=1}^{n} \mathbb{X}_j K_j \left(\left(X_{0j} - x\right)' \frac{1}{2} \frac{\partial^2 \pi\left(x\right)}{\partial x \partial x'} \left(X_{0j} - x\right) + O\left(h^3\right) \right).$$

Now we replace x with X_{0i} to obtain the expression when estimating at a location X_{0i}

$$\hat{\pi}(X_{0i}) - \pi(X_{0i}) = e_1' \left(\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1} \sum_{j=1}^{n} \mathbb{X}_{j,i} K_{j,i} \left(\left(Z_j - \pi_j\right) + \left(X_{0j} - X_{0i}\right)' \frac{1}{2} \frac{\partial^2 \pi\left(X_{0i}\right)}{\partial x \partial x'} \left(X_{0j} - X_{0i}\right) + O\left(h^3\right) \right), \tag{35}$$

where $\mathbb{X}_{j,i} = \left(1, \left(\frac{X_{0j} - X_{0i}}{h}\right)'\right)'$ and $K_{j,i} = \prod_{l=1}^{L} \frac{1}{h} \kappa \left(\frac{X_{0,jl} - X_{0,il}}{h}\right)$ and $\mathbb{X}_i = (\mathbb{X}_{1,i}, \mathbb{X}_{2,i}, ..., \mathbb{X}_{n,i})'$ and $\mathbb{K}_i = diag(K_{1,i}, K_{2,i}, ..., K_{n,i})$.

F.2.3 Denominator of the local linear estimator

Under the assumption that $nh^L \to \infty$ and $h \to 0$ one can show that for a kernel of order λ:

$$\frac{1}{n} \left(\mathbb{X}'\mathbb{K}\mathbb{X}\right) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{X}_j \mathbb{X}_j' \prod_{l=1}^{L} \frac{1}{h} \kappa \left(\frac{X_{0,jl} - x_l}{h}\right)$$

$$= \begin{bmatrix} f(x) + O(h^\lambda) & h^{\lambda-1} \frac{\mu_\lambda}{(\lambda-1)!} \frac{\partial^{\lambda-1} f(x)}{\partial x_1^{\lambda-1}} + O(h^\lambda) & \cdots & \cdots \\ h^{\lambda-1} \frac{\mu_\lambda}{(\lambda-1)!} \frac{\partial^{\lambda-1} f(x)}{\partial x_1^{\lambda-1}} + O(h^\lambda) & h^{\lambda-2} \frac{\mu_\lambda}{(\lambda-2)!} \frac{\partial^{\lambda-2} f(x)}{\partial x_1^{\lambda-2}} + h^{\lambda-1} \frac{\mu_{\lambda+1}}{(\lambda-1)!} \frac{\partial^{\lambda-1} f(x)}{\partial x_1^{\lambda-1}} + O(h^\lambda) & O(h^{2\lambda-2}) & \cdots \\ \vdots & O(h^{2\lambda-2}) & \ddots & O(h^{2\lambda-2}) \\ \vdots & \vdots & O(h^{2\lambda-2}) & \ddots \end{bmatrix} \tag{36}$$

This can be shown element-wise via mean square convergence. Only the derivations for the $(2,2)$ element are shown here, with the derivations for the other elements being analogous. Consider the $(2,2)$ element of $\frac{1}{n} \left(\mathbb{X}'\mathbb{K}\mathbb{X}\right)$ and denote it by ξ

$$\xi = \frac{1}{nh^L} \sum_{j=1}^{n} \left(\frac{X_{0,j1} - x_1}{h}\right)^2 \prod_{l=1}^{L} \kappa \left(\frac{X_{0,jl} - x_l}{h}\right)$$

which has the expected value:

$$E[\xi] = \frac{1}{h^L} \int \cdots \int \left(\frac{X_{0,j1} - x_1}{h}\right)^2 \prod_{l=1}^{L} \kappa \left(\frac{X_{0,jl} - x_l}{h}\right) f(X_{0j}) dX_{0j}.$$

With a change in variables: $u_l = \frac{X_{0,jl} - x_l}{h}$ and $u = (u_1, ..., u_L)'$ and a Taylor series expansion and noting that κ is a kernel of order λ we obtain

$$
\begin{aligned}
&= \int \cdots \int u_1^2 \prod_{l=1}^{L} \kappa(u_l) f(x+uh) du \\
&= \int \cdots \int u_1^2 \prod_{l=1}^{L} \kappa(u_l) \left(\frac{u_1^{\lambda-2} h^{\lambda-2}}{(\lambda-2)!} \frac{\partial^{\lambda-2} f(x)}{\partial u_1^{\lambda-2}} + \frac{u_1^{\lambda-1} h^{\lambda-1}}{(\lambda-1)!} \frac{\partial^{\lambda-1} f(x)}{\partial u_1^{\lambda-1}} + O(h^\lambda) \right) du \\
&= h^{\lambda-2} \frac{\mu_\lambda}{(\lambda-2)!} \frac{\partial^{\lambda-2} f(x)}{\partial x_1^{\lambda-2}} + h^{\lambda-1} \frac{\mu_{\lambda+1}}{(\lambda-1)!} \frac{\partial^{\lambda-1} f(x)}{\partial x_1^{\lambda-1}} + O(h^\lambda)
\end{aligned}
$$

by bounded convergence.

To show convergence in mean square, it also needs to be shown that $Var(\xi)$ converges to zero

$$
\begin{aligned}
Var(\xi) &= \frac{1}{n^2 h^{2L}} \sum_{j=1}^{n} Var\left(\left(\frac{X_{0,j1} - x_1}{h} \right)^2 \prod_{l=1}^{L} \kappa\left(\frac{X_{0,jl} - x_l}{h} \right) \right) \\
&= \frac{1}{nh^{2L}} E\left[\left(\left(\frac{X_{0,j1} - x_1}{h} \right)^2 \prod_{l=1}^{L} \kappa\left(\frac{X_{0,jl} - x_l}{h} \right) \right)^2 \right] - \frac{1}{nh^{2L}} \left(E\left[\left(\frac{X_{0,j1} - x_1}{h} \right)^2 \prod_{l=1}^{L} \kappa\left(\frac{X_{0,jl} - x_l}{h} \right) \right] \right)^2 \\
&= \frac{1}{nh^{2L}} \int h^L u^4 \prod_{l=1}^{L} \kappa^2(u_l) f(x+uh) du - \frac{1}{nh^{2L}} \left(h^L \int \cdots \int u^2 \prod_{l=1}^{L} \kappa(u_l) f(x+uh) du \right)^2 \\
&= O\left(\frac{1}{nh^L} \right) - O\left(\frac{h^{2\lambda-4}}{n} \right),
\end{aligned}
$$

by bounded convergence and Taylor series expansion. As it has been assumed that $nh^L \to \infty$, the variance of ξ converges to zero. Hence, mean square convergence has been shown, which implies convergence in probability by Chebyshev's inequality.

From (36) one can derive after some tedious calculations that

$$
\begin{aligned}
e_1' \left(\frac{1}{n} \mathbb{X}' \mathbb{K} \mathbb{X} \right)^{-1} &= \frac{1}{f(x) + O(h)} \begin{pmatrix} 1 + h \frac{\mu_{\lambda+1}}{\mu_\lambda} \frac{(\lambda-2)!}{(\lambda-1)!} \sum_{l=1}^{L} \frac{\partial^{\lambda-1} f(x)}{\partial x_l^{\lambda-1}} / \frac{\partial^{\lambda-2} f(x)}{\partial x_l^{\lambda-2}} \\ -h \left(\frac{\partial^{\lambda-1} f(x)}{\partial x_1^{\lambda-1}} / \frac{\partial^{\lambda-2} f(x)}{\partial x_1^{\lambda-2}} \right) \frac{(\lambda-2)!}{(\lambda-1)!} \\ \vdots \\ -h \left(\frac{\partial^{\lambda-1} f(x)}{\partial x_L^{\lambda-1}} / \frac{\partial^{\lambda-2} f(x)}{\partial x_L^{\lambda-2}} \right) \frac{(\lambda-2)!}{(\lambda-1)!} \end{pmatrix}' + O(h^2) \\
&= \frac{1}{f(x)} \begin{pmatrix} 1 + O(h) \\ -h \left(\frac{\partial^{\lambda-1} f(x)}{\partial x_1^{\lambda-1}} / \frac{\partial^{\lambda-2} f(x)}{\partial x_1^{\lambda-2}} \right) \frac{(\lambda-2)!'}{(\lambda-1)!} + O(h^2) \\ \vdots \\ -h \left(\frac{\partial^{\lambda-1} f(x)}{\partial x_L^{\lambda-1}} / \frac{\partial^{\lambda-2} f(x)}{\partial x_L^{\lambda-2}} \right) \frac{(\lambda-2)!'}{(\lambda-1)!} + O(h^2) \end{pmatrix}' \qquad (37)
\end{aligned}
$$

F.2.4 Further properties of local linear regression

For deriving the asymptotic distribution the expressions appearing in equation (34) are needed, in particular $E[\varsigma_{ij}|X_{0i},X_{1i},Z_i,D_i,Y_{2i},R_{2i},R_{1i}]$ and $E[\varsigma_{ji}|X_{0i},X_{1i},Z_i,D_i,Y_{2i},R_{2i},R_{1i}]$. These are derived below. To simplify notation, we frequently write $\frac{\partial^{\lambda-1}f(X_i)/\partial x^{\lambda-1}}{\partial^{\lambda-2}f(X_i)/\partial x^{\lambda-2}}$ as a shorthand notation for the column vector $\left(\left(\frac{\partial^{\lambda-1}f(x)}{\partial x_1^{\lambda-1}}/\frac{\partial^{\lambda-2}f(x)}{\partial x_1^{\lambda-2}}\right),...,\left(\frac{\partial^{\lambda-1}f(x)}{\partial x_L^{\lambda-1}}/\frac{\partial^{\lambda-2}f(x)}{\partial x_L^{\lambda-2}}\right)\right)'$.

Derive first $E[\varsigma_{ij}|X_{0i},X_{1i},Z_i,D_i,Y_{2i},R_{2i},R_{1i}]$ which is:

$$E\Big[\left(\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i^2\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\right)\cdot e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{0j,i}$$
$$\times\left((Z_j-\pi_j)+\frac{1}{2}(X_{0j}-X_{0i})'\frac{\partial^2\pi(X_{0i})}{\partial x\partial x'}(X_{0j}-X_{0i})+O\left(h^3\right)\right)|X_{0i},X_{1i},Z_i,D_i,Y_{2i},R_{2i},R_{1i}\Big]$$

$$=\left(\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i^2\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\right)\cdot h^\lambda\frac{\mu_\lambda}{f(X_{0i})}\frac{1}{2}\sum_{l=1}^{L}\frac{\partial^2\pi(X_{0i})}{\partial x_l^2}$$
$$\times\left(\frac{\partial^{\lambda-2}f(X_{0i})/\partial x_l^{\lambda-2}}{(\lambda-2)!}-\frac{(\lambda-2)!}{(\lambda-1)!\,(\lambda-3)!}\frac{\partial^{\lambda-1}f(X_{0i})/\partial x_l^{\lambda-1}}{\partial^{\lambda-2}f(X_{0i})/\partial x_l^{\lambda-2}}\frac{\partial^{\lambda-3}f(X_{0i})}{\partial x_l^{\lambda-3}}\right)=O(h^\lambda)\quad(38)$$

by (39).

In the last expression, the following term has been used, where we make use of (37):

$$E\Big[e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{0j,i}$$
$$\times\left((Z_j-\pi_j)+(X_{0j}-X_{0i})'\frac{1}{2}\frac{\partial^2\pi(X_{0i})}{\partial x\partial x'}(X_{0j}-X_{0i})+O\left(h^3\right)\right)|X_{0i},X_{1i},Z_i,D_i,Y_{2i},R_{2i},R_{1i}\Big]$$
$$=E\Big[e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{0j,i}\left((X_{0j}-X_{0i})'\frac{1}{2}\frac{\partial^2\pi(X_{0i})}{\partial x\partial x'}(X_{0j}-X_{0i})+O\left(h^3\right)\right)|X_{0i}\Big]$$
$$=E\Big[\left(1+O(h)-h\frac{(\lambda-2)!}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(X_{0i})/\partial x^{\lambda-1}}{\partial^{\lambda-2}f(X_{0i})/\partial x^{\lambda-2}}'\frac{X_{0j}-X_{0i}}{h}(1+O(h))\right)\frac{K_{0j,i}}{f(X_{0i})}$$
$$\times h^2\left(\frac{X_{0j}-X_{0i}}{h}'\frac{1}{2}\frac{\partial^2\pi(X_{0i})}{\partial x\partial x'}\frac{X_{0j}-X_{0i}}{h}+O(h)\right)|X_{0i}\Big]$$
$$=\int\left(1+O(h)-h\frac{(\lambda-2)!}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(X_{0i})/\partial x^{\lambda-1}}{\partial^{\lambda-2}f(X_{0i})/\partial x^{\lambda-2}}'\frac{X_{0j}-X_{0i}}{h}(1+O(h))\right)\frac{K_{0j,i}}{f(X_{0i})}$$
$$\times h^2\left(\frac{X_{0j}-X_{0i}}{h}'\frac{1}{2}\frac{\partial^2\pi(X_{0i})}{\partial x\partial x'}\frac{X_{0j}-X_{0i}}{h}+O(h)\right)f(X_{0j})dX_{0j}$$
$$=\frac{h^2}{f(X_{0i})}\int\left(1+O(h)-h\frac{(\lambda-2)!}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(X_{0i})/\partial x^{\lambda-1}}{\partial^{\lambda-2}f(X_{0i})/\partial x^{\lambda-2}}'u(1+O(h))\right)$$
$$\times\prod_{l=1}^{L}\kappa(u_l)\left(u'\frac{1}{2}\frac{\partial^2\pi(X_{0i})}{\partial x\partial x'}u+O(h)\right)f(X_{0i}+uh)du$$

where $u = \frac{X_{0j} - X_{0i}}{h}$. For $\lambda > 2$ and using a Taylor series expansion we obtain:

$$= \frac{h^{\lambda}\mu_{\lambda}}{f(X_{0i})}\frac{1}{2}\sum_{l=1}^{L}\frac{\partial^2\pi(X_{0i})}{\partial x_l^2}\left(\frac{\partial^{\lambda-2}f(X_{0i})/\partial x_l^{\lambda-2}}{(\lambda-2)!} - \frac{(\lambda-2)!}{(\lambda-1)!\,(\lambda-3)!}\frac{\partial^{\lambda-1}f(X_{0i})}{\partial x_l^{\lambda-1}}\frac{\partial^{\lambda-3}f(X_{0i})}{\partial x_l^{\lambda-3}}/\frac{\partial^{\lambda-2}f(X_{0i})}{\partial x_l^{\lambda-2}}\right) \tag{39}$$

and for $\lambda = 2$ this term would be

$$= \frac{h^{\lambda}\mu_{\lambda}}{2}\sum_{l=1}^{L}\frac{\partial^2\pi(X_{0i})}{\partial x_l^2}.$$

In both cases this term is of order $O(h^{\lambda})$.

Now we derive $E[\varsigma_{ji}|X_{0i}, X_{1i}, Z_i, D_i, Y_{2i}, R_{2i}, R_{1i}]$ where we make use of iterated expectations and condition on the X_0 observations in the entire sample, i.e.$X_{01}, \ldots.., X_{0n}$

$$E[\varsigma_{ji}|X_{0i}, X_{1i}, Z_i, D_i, Y_{2i}, R_{2i}, R_{1i}]$$

$$= E\Big[\left(\frac{Y_{2j}R_{2j}R_{1j}D_jZ_j}{\pi_j^2\Xi_{0j,Z=1}\Xi_{1j,Z=1}}\right)\cdot e_1'\left(\frac{1}{n}\mathbb{X}_j'\mathbb{K}_j\mathbb{X}_j\right)^{-1}\mathbb{X}_{i,j}K_{0i,j}$$
$$\times\left((Z_i-\pi_i)+\frac{1}{2}(X_{0i}-X_{0j})'\frac{\partial^2\pi(X_{0j})}{\partial x\partial x'}(X_{0i}-X_{0j})+O\left(h^3\right)\right)\ |X_{0i}, X_{1i}, Z_i, D_i, Y_{2i}, R_{2i}, R_{1i}\Big]$$

$$= E\Big[E\Big[\left(\frac{Y_{2j}R_{2j}R_{1j}D_jZ_j}{\pi_j^2\Xi_{0j,Z=1}\Xi_{1j,Z=1}}\right)\cdot e_1'\left(\frac{1}{n}\mathbb{X}_j'\mathbb{K}_j\mathbb{X}_j\right)^{-1}\mathbb{X}_{i,j}K_{0i,j}$$
$$\times\left((Z_i-\pi_i)+\frac{1}{2}(X_{0i}-X_{0j})'\frac{\partial^2\pi(X_{0j})}{\partial x\partial x'}(X_{0i}-X_{0j})+O\left(h^3\right)\right)$$
$$|X_{01}, \ldots.., X_{0n}, X_{1i}, Z_i, D_i, Y_{2i}, R_{2i}, R_{1i}\Big]\ |X_{0i}, X_{1i}, Z_i, D_i, Y_{2i}, R_{2i}, R_{1i}\Big]$$

$$= E\Big[E\Big[\left(\frac{Y_{2j}R_{2j}R_{1j}D_jZ_j}{\pi_j^2\Xi_{0j,Z=1}\Xi_{1j,Z=1}}\right)\ |X_{01}, \ldots.., X_{0n}, X_{1i}, Z_i, D_i, Y_{2i}, R_{2i}, R_{1i}\Big]$$
$$\times e_1'\left(\frac{1}{n}\mathbb{X}_j'\mathbb{K}_j\mathbb{X}_j\right)^{-1}\mathbb{X}_{i,j}K_{0i,j}$$
$$\times\left((Z_i-\pi_i)+\frac{1}{2}(X_{0i}-X_{0j})'\frac{\partial^2\pi(X_{0j})}{\partial x\partial x'}(X_{0i}-X_{0j})+O\left(h^3\right)\right)\ |X_{0i}, X_{1i}, Z_i, D_i, Y_{2i}, R_{2i}, R_{1i}\Big]$$

$$= E\Big[\vartheta(X_{0j})\times e_1'\left(\frac{1}{n}\mathbb{X}_j'\mathbb{K}_j\mathbb{X}_j\right)^{-1}\mathbb{X}_{i,j}K_{0i,j}$$
$$\times\left((Z_i-\pi_i)+\frac{1}{2}(X_{0i}-X_{0j})'\frac{\partial^2\pi(X_{0j})}{\partial x\partial x'}(X_{0i}-X_{0j})+O\left(h^3\right)\right)\ |X_{0i}, X_{1i}, Z_i, D_i, Y_{2i}, R_{2i}, R_{1i}\Big]$$

where we define the shortcut notation

$$\vartheta_j = \vartheta\left(X_{0j}\right) = E\left[\frac{Y_2 R_2 R_1 D Z}{\pi^2 \Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0j}\right].$$

Now we enter (37) to obtain

$$\begin{aligned}
= \quad & E\Big[\vartheta\left(X_{0j}\right)\left(1+O(h)-h\frac{(\lambda-2)!}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(X_{0j})/\partial x^{\lambda-1}{}'}{\partial^{\lambda-2}f(X_{0j})/\partial x^{\lambda-2}}\frac{X_{0i}-X_{0j}}{h}\left(1+O(h)\right)\right)\frac{K_{0i,j}}{f(X_{0j})} \\
& \times\left(\left(Z_i-\pi_i\right)+\left(X_{0i}-X_{0j}\right)'\frac{1}{2}\frac{\partial^2\pi\left(X_{0j}\right)}{\partial x\partial x'}\left(X_{0i}-X_{0j}\right)+O\left(h^3\right)\right)\ |X_{0i},X_{1i},Z_i,D_i,Y_{2i},R_{2i},R_{1i}\Big] \\
= \quad & \int\vartheta\left(X_{0i}-vh\right)\left(1+O(h)-h\frac{(\lambda-2)!}{(\lambda-1)!}\frac{\partial^{\lambda-1}f(X_{0i}-vh)/\partial x^{\lambda-1}{}'}{\partial^{\lambda-2}f(X_{0i}-vh)/\partial x^{\lambda-2}}v\left(1+O(h)\right)\right) \\
& \times\left(\left(Z_i-\pi_i\right)+\frac{h^2}{2}v'\frac{\partial^2\pi\left(X_{0i}-vh\right)}{\partial x\partial x'}v+O\left(h^3\right)\right)\prod_{l=1}^{L}\kappa\left(v_l\right)dv
\end{aligned}$$

where $v=\frac{X_{0i}-X_{0j}}{h}$ and by bounded convergence. With $\partial^2\pi\left(x\right)/\partial x\partial x'$ Hölder continuous, the term $Z_i-\pi_i$ clearly dominates the last expression and we obtain by bounded convergence

$$= \left(\vartheta(X_{0i})+O(h)\right)\cdot\left(Z_i-\pi_i\right)$$

$$= \left(E\left[\frac{Y_2R_2R_1DZ}{\pi^2\Xi_{0,Z=1}\Xi_{1,Z=1}}|X_{0i}\right]+O(h)\right)\cdot\left(Z_i-\pi_i\right). \tag{40}$$

For an application of the projection theorem in (34) we need to show that $E[\varsigma_{ij}^2]\leq o(n)$. Therefore, consider the term

$$\begin{aligned}
& E\Big[\Big[\left(\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i^2\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\right)e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{0j,i} \\
& \qquad\times\left(\left(Z_j-\pi_j\right)+\frac{1}{2}\left(X_{0j}-X_{0i}\right)'\frac{\partial^2\pi\left(X_{0i}\right)}{\partial x\partial x'}\left(X_{0j}-X_{0i}\right)+O\left(h^3\right)\right)\Big]^2\Big] \\
= \quad & E\Big[\left(\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i^2\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\right)^2\Big[e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{0j,i} \\
& \qquad\times\left(\left(Z_j-\pi_j\right)+\frac{1}{2}\left(X_{0j}-X_{0i}\right)'\frac{\partial^2\pi\left(X_{0i}\right)}{\partial x\partial x'}\left(X_{0j}-X_{0i}\right)+O\left(h^3\right)\right)\Big]^2\Big]
\end{aligned}$$

and using iterated expectations

$$\begin{aligned}
= E\Big[E\Big[&\left(\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i^2\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\right)^2\Big[e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{0j,i} \\
& \times\left(\left(Z_j-\pi_j\right)+\frac{1}{2}\left(X_{0j}-X_{0i}\right)'\frac{\partial^2\pi\left(X_{0i}\right)}{\partial x\partial x'}\left(X_{0j}-X_{0i}\right)+O\left(h^3\right)\right)\Big]^2 \\
& \qquad\qquad |X_{0i},X_{1i},Z_i,D_i,Y_{2i},R_{2i},R_{1i}\Big]\Big]
\end{aligned}$$

$$= E\Big[E\Big[\Big[e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{0j,i}$$

$$\times\bigg((Z_j-\pi_j)+\frac{1}{2}(X_{0j}-X_{0i})'\frac{\partial^2\pi(X_{0i})}{\partial x\partial x'}(X_{0j}-X_{0i})+O\left(h^3\right)\bigg)\Big]^2|X_{0i},X_{1i},Z_i,D_i,Y_{2i},R_{2i},R_{1i}\Big]$$

$$\times\left(\frac{Y_{2i}R_{2i}R_{1i}D_iZ_i}{\pi_i^2\Xi_{0i,Z=1}\Xi_{1i,Z=1}}\right)^2\Big]$$

$$= o(n) \tag{41}$$

because of (42) and since $\pi(x_0)$ and $\Xi_{0,Z=1}(x_0)$ and $\Xi_{1i,Z=1}(x_0,x_1)$ are bounded away from zero as has been assumed.

Here we have used that

$$= E\Big[\Big[e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{0j,i}$$

$$\times\bigg((Z_j-\pi_j)+\frac{1}{2}(X_{0j}-X_{0i})'\frac{\partial^2\pi(X_{0i})}{\partial x\partial x'}(X_{0j}-X_{0i})+O\left(h^3\right)\bigg)\Big]^2|X_{0i},X_{1i},Z_i,D_i,Y_{2i},R_{2i},R_{1i}\Big]$$

$$\begin{aligned}
= \;& E\left[\left[e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{0j,i}\left((Z_j-\pi_j)+(X_{0j}-X_{0i})'\frac{1}{2}\frac{\partial^2\pi(X_{0i})}{\partial x\partial x'}(X_{0j}-X_{0i})+O\left(h^3\right)\right)\right]^2|X_{0i}\right]\\
= \;& E\left[\left(e_1'\left(\frac{1}{n}\mathbb{X}_i'\mathbb{K}_i\mathbb{X}_i\right)^{-1}\mathbb{X}_{j,i}K_{0j,i}\right)^2\left(\pi_j(1-\pi_j)+h^4\left(\frac{X_{0j}-X_{0i}}{h}'\frac{1}{2}\frac{\partial^2\pi(X_{0i})}{\partial x\partial x'}\frac{X_{0j}-X_{0i}}{h}+O(h)\right)^2\right)|X_{0i}\right]\\
= \;& E\left[\left(1+O(h)-h\frac{\partial^{\lambda-1}f(X_{0i})/\partial x^{\lambda-1}}{\partial^{\lambda-2}f(X_{0i})/\partial x^{\lambda-2}}'\frac{X_{0j}-X_{0i}}{h}\frac{(\lambda-2)!}{(\lambda-1)!}\right)^2\frac{K_{0j,i}^2}{f^2(X_{0i})}\right.\\
&\left.\times\left(\pi_j(1-\pi_j)+h^4\left(\left(\frac{X_{0j}-X_{0i}}{h}\right)'\frac{1}{2}\frac{\partial^2\pi(X_{0i})}{\partial x\partial x'}\left(\frac{X_{0j}-X_{0i}}{h}\right)+O(h)\right)^2\right)|X_{0i}\right]\\
= \;& \int\frac{1}{h^{2L}}\left(1+O(h)-h\frac{\partial^{\lambda-1}f(X_{0i})/\partial x^{\lambda-1}}{\partial^{\lambda-2}f(X_{0i})/\partial x^{\lambda-2}}'u\frac{(\lambda-2)!}{(\lambda-1)!}\right)^2\frac{1}{f^2(X_{0i})}\prod_{l=1}^{L}\kappa^2(u_l)\\
&\times\left(\pi(X_{0i}+uh)(1-\pi(X_{0i}+uh))+h^4\left(u'\frac{1}{2}\frac{\partial^2\pi(X_{0i})}{\partial x\partial x'}u+O(h)\right)^2\right)f(X_{0i}+uh)h^Ldu
\end{aligned}$$

where $u=\frac{X_{0j}-X_{0i}}{h}$ and by bounded convergence

$$= \frac{1}{h^L}\frac{\pi(X_{0i})(1-\pi(X_{0i}))}{f(X_{0i})}\bar{\mu}_0^L(1+O(h))=O\left(\frac{1}{h^L}\right)=o(n) \tag{42}$$

as it has been assumed that $nh^L\to\infty$.

F.2.5 Estimation under conditional latent ignorability

The derivations for Theorem 3 are analogous to those of the previous subsections. The candidate function for $\alpha(\mathcal{Z}_i)$ is

$$\alpha(\mathcal{Z}_i) = E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_0\Xi_1\frac{P_0'-\pi}{P_0-\pi}\frac{P_1'-\pi}{P_1-\pi}}\left(-\frac{1}{\pi}+\frac{P_0-P_0'}{(P_0'-\pi)(P_0-\pi)}+\frac{P_1-P_1'}{(P_1'-\pi)(P_1-\pi)}\right)|X_{0i}\right](Z_i-\pi_i)$$
$$+E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_0\Xi_1\frac{P_0'-\pi}{P_0-\pi}\frac{P_1'-\pi}{P_1-\pi}}|X_{0i}\right]\frac{D_i\,(Z_i-\pi_i)}{E\left[R_1D|X_{0i}\right]}\left(\frac{\Xi_{0i}}{P_{0i}-\pi_i}-\frac{R_{1i}}{P_{0i}'-\pi_i}\right)$$
$$+E\left[\frac{Y_2R_2R_1DZ}{\pi\Xi_0\Xi_1\frac{P_0'-\pi}{P_0-\pi}\frac{P_1'-\pi}{P_1-\pi}}|X_{0i},X_{1i}\right]\frac{R_{1i}D_i\,(Z_i-\pi_i)}{E\left[R_2R_1D|X_{0i},X_{1i}\right]}\left(\frac{\Xi_{1i}}{P_{1i}-\pi_i}-\frac{R_{2i}}{P_{1i}'-\pi_i}\right).$$

With analogous derivations as in the previous subsection the Assumptions 5.1 to 5.3 of Newey (1994) can be verified.

G Full set of simulation results for sample size 5000

In the main text we had described the first four estimators we examined in the simulations: (i) naive estimation based on mean differences in observed treated and non-treated outcomes (i.e. ignoring both treatment endogeneity and attrition), (ii) CACE estimation based on Tan (2006) and Frölich (2007) that controls for endogeneity, but ignores attrition (denoted by $\hat{\omega}$), (iii) CACE estimation based on the sample analog of (2) in Theorem 1 (denoted by $\hat{\theta}$), and (iv) CACE estimation based on the sample analog of (4) in Theorem 2 (denoted by $\hat{\phi}$). In Tables 1 to 3 we displayed the results for these estimators. The following three Tables 9 to 11 repeat those results and provide additionally the estimates for trimming levels 0.10, 0.05 and 0.02. In addition, they also show the results for several additional estimators described below. Finally, Table 12 examines an additional design with treatment exogeneity as described further below.

In addition to the aforementioned estimators, we also consider four other estimators that control for attrition under the assumption of MAR but ignore treatment endogeneity due to U_0, U_1, U_2, while controlling for confounding related to X_0. These estimators are: (v) the MLE-based G-computation procedure of Robins (1986) in which the outcomes and response processes are modeled parametrically (by linear and logit specifications, respectively), (vi) targeted MLE (TMLE) as introduced by van der Laan and Rubin (2006), an efficient doubly robust semi-parametric estimator where G-computation provides the initial estimator of the conditional mean outcomes given D and X_0 that in a second step are updated based on the efficient influence function of the ATE for bias reduction, (vii) semi-parametric inverse probability weighting (IPW, see for instance Horvitz and Thompson (1952) and Hirano, Imbens, and Ridder (2003)) of observations by estimates of the treatment propensity score $\Pr(D = 1|X_0)$ and the respective response propensity scores under MAR ($\Pr(R_1 = 1|D, X_0)$, $\Pr(R_2 = 1|D, X_0, X_1)$), where the propensity scores are estimated by probit specifications and the weights in either treatment group are normalized to add up to one, and (viii) augmented inverse probability weighting (AIPW) as discussed in Robins, Rotnitzky, and Zhao (1995) and Scharfstein, Rotnitzky, and Robins (1999), which augments IPW by an adjustment term related to the conditional mean outcomes (in our case estimated by a linear regression) and is (similarly to TMLE) doubly robust and efficient if both the outcomes and the propensity scores are correctly specified. Finally, (ix) Heckman (1976) fully parametric MLE of sample selection models assuming jointly normally distributed unobserved terms in the response and the outcome equations is also considered. The latter estimator controls for X_0, D, and Z in the estimation of response and can therefore account for attrition related to unobservables if R_1 and R_2 are functions of Z and if Z does not have a direct effect on the outcomes conditional on X_0 and D. However, it does not allow for treatment endogeneity related to U_0, U_1, U_2 and additionally presumes treatment effects to be homogeneous.

Note that the Heckman estimator is only consistent if the treatment is exogenous given X_0 and if treatment effects are homogeneous ($\kappa = 0$). This case is illustrated in Table 12. There we change the covariance matrix of the unobservables to

$$\sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix},$$

such that U_0 in the treatment equation is uncorrelated with U_1 and U_2 in the outcome equations. The remaining parameters are the same as under the homogeneous treatment effects scenario in Table 3. Now, in this particular set up, the Heckman estimator dominates the other methods, as it is the only one that accounts for attrition related to contemporaneous outcomes using Z. Our simulations demonstrated that the presumed form of attrition and the presence or absence of treatment endogeneity and/or effect heterogeneity is crucial for the choice of the estimation approach.

Table 9: Simulation 1 - treatment endogeneity and conditional LI, n=5000

	Homogeneous effects											
	α=0.68, β=0, γ=0.5, δ=0, κ=0						α=1.35, β=0, γ=0.5, δ=0, κ=0					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.696	0.034	0.697	0.920	0.059	0.922	0.555	0.033	0.556	0.691	0.055	0.693
$\hat{\omega}$	0.053	0.113	0.125	0.117	0.156	0.195	0.031	0.059	0.067	0.069	0.083	0.108
$\hat{\theta}$	0.011	0.120	0.121	-0.026	0.172	0.174	0.008	0.061	0.062	-0.014	0.088	0.089
$\hat{\theta}_{\text{trim}}(0.15)$	0.011	0.120	0.121	-0.026	0.172	0.174	0.008	0.061	0.062	-0.014	0.088	0.089
$\hat{\theta}_{\text{trim}}(0.10)$	0.011	0.120	0.121	-0.026	0.172	0.174	0.008	0.061	0.062	-0.014	0.088	0.089
$\hat{\theta}_{\text{trim}}(0.05)$	0.011	0.120	0.121	-0.026	0.172	0.174	0.008	0.061	0.062	-0.014	0.088	0.089
$\hat{\theta}_{\text{trim}}(0.02)$	0.011	0.120	0.121	-0.026	0.172	0.174	0.008	0.061	0.062	-0.014	0.088	0.089
$\hat{\theta}_{\text{trim}}(0.01)$	0.011	0.120	0.121	-0.026	0.172	0.174	0.008	0.061	0.062	-0.014	0.088	0.089
$\hat{\phi}$	-0.143	0.140	0.201	-0.409	0.232	0.470	-0.093	0.068	0.115	-0.251	0.108	0.274
$\hat{\phi}_{\text{trim}}(0.15)$	-0.143	0.140	0.201	-0.409	0.232	0.470	-0.093	0.068	0.115	-0.251	0.108	0.274
$\hat{\phi}_{\text{trim}}(0.10)$	-0.143	0.140	0.201	-0.409	0.232	0.470	-0.093	0.068	0.115	-0.251	0.108	0.274
$\hat{\phi}_{\text{trim}}(0.05)$	-0.143	0.140	0.201	-0.409	0.232	0.470	-0.093	0.068	0.115	-0.251	0.108	0.274
$\hat{\phi}_{\text{trim}}(0.02)$	-0.143	0.140	0.201	-0.409	0.232	0.470	-0.093	0.068	0.115	-0.251	0.108	0.274
$\hat{\phi}_{\text{trim}}(0.01)$	-0.143	0.140	0.201	-0.409	0.232	0.470	-0.093	0.068	0.115	-0.251	0.108	0.274
MAR G-comp.	0.709	0.034	0.709	0.939	0.058	0.941	0.558	0.032	0.559	0.695	0.055	0.698
MAR TMLE	0.708	0.034	0.709	0.975	0.058	0.977	0.558	0.032	0.559	0.720	0.055	0.722
MAR IPW	0.710	0.034	0.711	0.981	0.059	0.982	0.561	0.033	0.562	0.725	0.055	0.727
MAR AIPW	0.708	0.034	0.709	0.976	0.059	0.978	0.558	0.032	0.559	0.719	0.055	0.721
Heckman	0.355	1.586	1.626	0.027	1.609	1.610	0.332	3.301	3.318	-0.215	2.634	2.643
true CACE		0.500			1.250			0.500			1.250	
mean response		0.633			0.411			0.687			0.485	

	Heterogeneous effects											
	α=0.68, β=0, γ=0.5, δ=0, κ=0.5						α=1.35, β=0, γ=0.5, δ=0, κ=0.5					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.986	0.037	0.986	1.342	0.064	1.344	0.833	0.036	0.834	1.096	0.061	1.098
$\hat{\omega}$	0.054	0.128	0.139	0.123	0.183	0.221	0.032	0.070	0.077	0.078	0.102	0.128
$\hat{\theta}$	0.009	0.134	0.134	-0.030	0.195	0.198	0.008	0.072	0.072	-0.011	0.104	0.104
$\hat{\theta}_{\text{trim}}(0.15)$	0.009	0.134	0.134	-0.030	0.195	0.198	0.008	0.072	0.072	-0.011	0.104	0.104
$\hat{\theta}_{\text{trim}}(0.10)$	0.009	0.134	0.134	-0.030	0.195	0.198	0.008	0.072	0.072	-0.011	0.104	0.104
$\hat{\theta}_{\text{trim}}(0.05)$	0.009	0.134	0.134	-0.030	0.195	0.198	0.008	0.072	0.072	-0.011	0.104	0.104
$\hat{\theta}_{\text{trim}}(0.02)$	0.009	0.134	0.134	-0.030	0.195	0.198	0.008	0.072	0.072	-0.011	0.104	0.104
$\hat{\theta}_{\text{trim}}(0.01)$	0.009	0.134	0.134	-0.030	0.195	0.198	0.008	0.072	0.072	-0.011	0.104	0.104
$\hat{\phi}$	-0.201	0.156	0.254	-0.523	0.258	0.583	-0.142	0.079	0.163	-0.345	0.125	0.367
$\hat{\phi}_{\text{trim}}(0.15)$	-0.201	0.156	0.254	-0.523	0.258	0.583	-0.142	0.079	0.163	-0.345	0.125	0.367
$\hat{\phi}_{\text{trim}}(0.10)$	-0.201	0.156	0.254	-0.523	0.258	0.583	-0.142	0.079	0.163	-0.345	0.125	0.367
$\hat{\phi}_{\text{trim}}(0.05)$	-0.201	0.156	0.254	-0.523	0.258	0.583	-0.142	0.079	0.163	-0.345	0.125	0.367
$\hat{\phi}_{\text{trim}}(0.02)$	-0.201	0.156	0.254	-0.523	0.258	0.583	-0.142	0.079	0.163	-0.345	0.125	0.367
$\hat{\phi}_{\text{trim}}(0.01)$	-0.201	0.156	0.254	-0.523	0.258	0.583	-0.142	0.079	0.163	-0.345	0.125	0.367
MAR G-comp.	0.999	0.036	0.999	1.363	0.063	1.365	0.837	0.036	0.838	1.101	0.061	1.103
MAR TMLE	0.998	0.036	0.999	1.379	0.063	1.381	0.836	0.036	0.837	1.100	0.061	1.101
MAR IPW	1.000	0.037	1.001	1.384	0.063	1.386	0.839	0.036	0.840	1.104	0.061	1.105
MAR AIPW	0.998	0.036	0.999	1.379	0.063	1.381	0.836	0.036	0.837	1.098	0.061	1.099
Heckman	0.625	2.271	2.355	0.251	2.070	2.085	0.409	5.040	5.057	-0.182	2.849	2.855
true CACE		0.636			1.454			0.479			1.218	
mean response		0.632			0.426			0.689			0.501	

Note: Results are based on 1000 simulations and 5000 observations.

Table 10: Simulation 2 - treatment endogeneity and MAR (with instrument affecting response), n=5000

	Homogeneous effects											
	α=0.68, β=0.5, γ=0, δ=0, κ=0						α=1.35, β=0.5, γ=0, δ=0, κ=0					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.749	0.033	0.749	1.084	0.055	1.085	0.642	0.033	0.643	0.932	0.057	0.933
$\hat{\omega}$	0.207	0.093	0.227	0.468	0.116	0.482	0.119	0.053	0.130	0.282	0.070	0.290
$\hat{\theta}$	-0.365	41.338	41.340	307.305	9507.592	9512.557	-0.048	0.083	0.096	-0.401	4.803	4.819
$\hat{\theta}_{\text{trim}}\,(0.15)$	-0.931	5.084	5.169	1.756	32.972	33.019	-0.048	0.083	0.096	-0.489	0.644	0.809
$\hat{\theta}_{\text{trim}}\,(0.10)$	-0.965	4.869	4.964	1.972	33.730	33.787	-0.048	0.083	0.096	-0.484	0.681	0.836
$\hat{\theta}_{\text{trim}}\,(0.05)$	-0.917	4.772	4.860	1.974	35.823	35.877	-0.048	0.083	0.096	-0.480	0.727	0.871
$\hat{\theta}_{\text{trim}}\,(0.02)$	-0.903	4.107	4.205	2.050	38.673	38.727	-0.048	0.083	0.096	-0.462	0.790	0.915
$\hat{\theta}_{\text{trim}}\,(0.01)$	-0.855	4.299	4.383	2.097	41.206	41.259	-0.048	0.083	0.096	-0.432	0.841	0.946
$\hat{\phi}$	0.017	0.138	0.139	-0.085	0.230	0.245	0.009	0.070	0.071	-0.050	0.115	0.126
$\hat{\phi}_{\text{trim}}\,(0.15)$	0.017	0.138	0.139	-0.085	0.230	0.245	0.009	0.070	0.071	-0.050	0.115	0.126
$\hat{\phi}_{\text{trim}}\,(0.10)$	0.017	0.138	0.139	-0.085	0.230	0.245	0.009	0.070	0.071	-0.050	0.115	0.126
$\hat{\phi}_{\text{trim}}\,(0.05)$	0.017	0.138	0.139	-0.085	0.230	0.245	0.009	0.070	0.071	-0.050	0.115	0.126
$\hat{\phi}_{\text{trim}}\,(0.02)$	0.017	0.138	0.139	-0.085	0.230	0.245	0.009	0.070	0.071	-0.050	0.115	0.126
$\hat{\phi}_{\text{trim}}\,(0.01)$	0.017	0.138	0.139	-0.085	0.230	0.245	0.009	0.070	0.071	-0.050	0.115	0.126
MAR G-comp.	0.764	0.032	0.765	1.113	0.054	1.114	0.650	0.032	0.651	0.951	0.057	0.953
MAR TMLE	0.764	0.032	0.765	1.148	0.054	1.149	0.650	0.032	0.651	0.992	0.056	0.994
MAR IPW	0.764	0.033	0.764	1.147	0.055	1.149	0.650	0.033	0.651	0.991	0.057	0.992
MAR AIPW	0.764	0.032	0.765	1.148	0.054	1.149	0.650	0.032	0.651	0.991	0.057	0.993
Heckman	0.860	0.036	0.860	1.300	0.062	1.302	0.828	0.037	0.828	1.258	0.086	1.261
true CACE		0.500			1.250			0.500			1.250	
mean response		0.684			0.489			0.694			0.507	

	Heterogeneous effects											
	α=0.68, β=0.5, γ=0, δ=0, κ=0.5						α=1.35, β=0.5, γ=0, δ=0, κ=0.5					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	1.053	0.036	1.053	1.543	0.059	1.544	0.941	0.036	0.941	1.388	0.063	1.389
$\hat{\omega}$	0.291	0.103	0.309	0.653	0.132	0.666	0.181	0.062	0.191	0.432	0.085	0.441
$\hat{\theta}$	-0.282	41.337	41.338	307.469	9507.593	9512.563	0.014	0.090	0.091	-0.270	4.803	4.811
$\hat{\theta}_{\text{trim}}\,(0.15)$	-0.848	5.084	5.154	1.920	32.972	33.028	0.014	0.090	0.091	-0.358	0.646	0.738
$\hat{\theta}_{\text{trim}}\,(0.10)$	-0.881	4.869	4.948	2.136	33.730	33.797	0.014	0.090	0.091	-0.353	0.683	0.768
$\hat{\theta}_{\text{trim}}\,(0.05)$	-0.834	4.772	4.844	2.138	35.823	35.886	0.014	0.090	0.091	-0.348	0.728	0.807
$\hat{\theta}_{\text{trim}}\,(0.02)$	-0.819	4.107	4.188	2.214	38.672	38.736	0.014	0.090	0.091	-0.331	0.791	0.857
$\hat{\theta}_{\text{trim}}\,(0.01)$	-0.772	4.299	4.367	2.262	41.206	41.268	0.014	0.090	0.091	-0.300	0.842	0.894
$\hat{\phi}$	0.015	0.151	0.152	-0.082	0.251	0.264	0.008	0.080	0.080	-0.041	0.130	0.136
$\hat{\phi}_{\text{trim}}\,(0.15)$	0.015	0.151	0.152	-0.082	0.251	0.264	0.008	0.080	0.080	-0.041	0.130	0.136
$\hat{\phi}_{\text{trim}}\,(0.10)$	0.015	0.151	0.152	-0.082	0.251	0.264	0.008	0.080	0.080	-0.041	0.130	0.136
$\hat{\phi}_{\text{trim}}\,(0.05)$	0.015	0.151	0.152	-0.082	0.251	0.264	0.008	0.080	0.080	-0.041	0.130	0.136
$\hat{\phi}_{\text{trim}}\,(0.02)$	0.015	0.151	0.152	-0.082	0.251	0.264	0.008	0.080	0.080	-0.041	0.130	0.136
$\hat{\phi}_{\text{trim}}\,(0.01)$	0.015	0.151	0.152	-0.082	0.251	0.264	0.008	0.080	0.080	-0.041	0.130	0.136
MAR G-comp.	1.069	0.035	1.069	1.574	0.059	1.575	0.949	0.036	0.950	1.408	0.062	1.409
MAR TMLE	1.068	0.035	1.069	1.595	0.058	1.596	0.949	0.036	0.949	1.429	0.062	1.430
MAR IPW	1.068	0.035	1.068	1.595	0.059	1.596	0.949	0.036	0.949	1.427	0.062	1.428
MAR AIPW	1.068	0.035	1.069	1.595	0.059	1.596	0.949	0.036	0.949	1.427	0.063	1.428
Heckman	1.208	0.041	1.209	1.876	0.068	1.877	1.193	0.043	1.193	2.018	0.436	2.064
true CACE		0.636			1.454			0.479			1.218	
mean response		0.684			0.494			0.694			0.510	

Note: Results are based on 1000 simulations and 5000 observations.

Table 11: Simulation 3 - treatment endogeneity and selection on current outcomes, n=5000

	Homogeneous effects											
	α=0.68, β=0.5, γ=0, δ=0.25, κ=0						α=1.35, β=0.5, γ=0, δ=0.25, κ=0					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.682	0.032	0.683	0.852	0.052	0.853	0.565	0.033	0.566	0.667	0.055	0.670
$\hat{\omega}$	0.095	0.095	0.134	0.355	0.114	0.373	0.048	0.053	0.071	0.193	0.068	0.204
$\hat{\theta}$	-1.549	5.089	5.319	1.479	572.819	572.821	-0.194	0.080	0.210	-0.680	0.470	0.827
$\hat{\theta}_{\text{trim}}(0.15)$	-1.177	4.651	4.798	-1.209	20.300	20.336	-0.194	0.080	0.210	-0.667	0.239	0.708
$\hat{\theta}_{\text{trim}}(0.10)$	-1.179	4.657	4.804	-1.182	20.913	20.946	-0.194	0.080	0.210	-0.667	0.226	0.704
$\hat{\theta}_{\text{trim}}(0.05)$	-1.188	4.719	4.867	-0.985	21.978	22.000	-0.194	0.080	0.210	-0.664	0.207	0.695
$\hat{\theta}_{\text{trim}}(0.02)$	-1.197	4.776	4.924	-0.613	23.241	23.249	-0.194	0.080	0.210	-0.661	0.211	0.694
$\hat{\theta}_{\text{trim}}(0.01)$	-1.206	4.900	5.047	-0.452	24.367	24.371	-0.194	0.080	0.210	-0.651	0.203	0.682
$\hat{\phi}$	-0.234	0.137	0.271	-0.590	0.215	0.628	-0.132	0.069	0.149	-0.355	0.106	0.370
$\hat{\phi}_{\text{trim}}(0.15)$	-0.234	0.137	0.271	-0.590	0.215	0.628	-0.132	0.069	0.149	-0.355	0.106	0.370
$\hat{\phi}_{\text{trim}}(0.10)$	-0.234	0.137	0.271	-0.590	0.215	0.628	-0.132	0.069	0.149	-0.355	0.106	0.370
$\hat{\phi}_{\text{trim}}(0.05)$	-0.234	0.137	0.271	-0.590	0.215	0.628	-0.132	0.069	0.149	-0.355	0.106	0.370
$\hat{\phi}_{\text{trim}}(0.02)$	-0.234	0.137	0.271	-0.590	0.215	0.628	-0.132	0.069	0.149	-0.355	0.106	0.370
$\hat{\phi}_{\text{trim}}(0.01)$	-0.234	0.137	0.271	-0.590	0.215	0.628	-0.132	0.069	0.149	-0.355	0.106	0.370
MAR G-comp.	0.700	0.032	0.700	0.887	0.051	0.889	0.576	0.032	0.577	0.695	0.054	0.697
MAR TMLE	0.699	0.032	0.700	1.009	0.051	1.010	0.575	0.032	0.576	0.831	0.054	0.833
MAR IPW	0.699	0.032	0.700	1.012	0.052	1.013	0.575	0.032	0.576	0.837	0.056	0.839
MAR AIPW	0.699	0.032	0.700	1.012	0.053	1.013	0.575	0.032	0.576	0.838	0.058	0.840
Heckman	0.874	0.037	0.875	1.228	0.091	1.232	0.837	0.067	0.840	2.105	1.015	2.337
true CACE		0.500			1.250			0.500			1.250	
mean response		0.709			0.545			0.723			0.571	
	Heterogeneous effects											
	α=0.68, β=0.5, γ=0, δ=0.25, κ=0.5						α=1.35, β=0.5, γ=0, δ=0.25, κ=0.5					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.999	0.034	1.000	1.328	0.055	1.329	0.892	0.036	0.893	1.182	0.059	1.184
$\hat{\omega}$	0.134	0.110	0.173	0.431	0.140	0.453	0.085	0.064	0.107	0.280	0.089	0.293
$\hat{\theta}$	-1.507	5.091	5.309	1.554	572.819	572.821	-0.154	0.089	0.178	-0.599	0.473	0.764
$\hat{\theta}_{\text{trim}}(0.15)$	-1.135	4.652	4.789	-1.134	20.302	20.334	-0.154	0.089	0.178	-0.586	0.246	0.635
$\hat{\theta}_{\text{trim}}(0.10)$	-1.137	4.658	4.795	-1.107	20.914	20.944	-0.154	0.089	0.178	-0.586	0.235	0.631
$\hat{\theta}_{\text{trim}}(0.05)$	-1.146	4.721	4.858	-0.910	21.980	21.999	-0.154	0.089	0.178	-0.583	0.216	0.621
$\hat{\theta}_{\text{trim}}(0.02)$	-1.155	4.777	4.915	-0.538	23.243	23.249	-0.154	0.089	0.178	-0.580	0.219	0.620
$\hat{\theta}_{\text{trim}}(0.01)$	-1.164	4.902	5.038	-0.377	24.369	24.372	-0.154	0.089	0.178	-0.570	0.212	0.608
$\hat{\phi}$	-0.222	0.151	0.269	-0.561	0.235	0.608	-0.105	0.078	0.131	-0.289	0.119	0.313
$\hat{\phi}_{\text{trim}}(0.15)$	-0.222	0.151	0.269	-0.561	0.235	0.608	-0.105	0.078	0.131	-0.289	0.119	0.313
$\hat{\phi}_{\text{trim}}(0.10)$	-0.222	0.151	0.269	-0.561	0.235	0.608	-0.105	0.078	0.131	-0.289	0.119	0.313
$\hat{\phi}_{\text{trim}}(0.05)$	-0.222	0.151	0.269	-0.561	0.235	0.608	-0.105	0.078	0.131	-0.289	0.119	0.313
$\hat{\phi}_{\text{trim}}(0.02)$	-0.222	0.151	0.269	-0.561	0.235	0.608	-0.105	0.078	0.131	-0.289	0.119	0.313
$\hat{\phi}_{\text{trim}}(0.01)$	-0.222	0.151	0.269	-0.561	0.235	0.608	-0.105	0.078	0.131	-0.289	0.119	0.313
MAR G-comp.	1.018	0.034	1.018	1.367	0.055	1.368	0.903	0.035	0.904	1.210	0.059	1.212
MAR TMLE	1.017	0.034	1.017	1.488	0.055	1.489	0.903	0.035	0.903	1.334	0.059	1.335
MAR IPW	1.017	0.034	1.018	1.493	0.056	1.494	0.903	0.036	0.903	1.343	0.062	1.345
MAR AIPW	1.017	0.034	1.018	1.493	0.057	1.495	0.903	0.035	0.903	1.346	0.064	1.348
Heckman	1.297	0.041	1.297	1.953	0.274	1.972	1.672	0.570	1.766	2.651	1.273	2.940
true CACE		0.636			1.454			0.479			1.218	
mean response		0.719			0.562			0.732			0.586	

Note: Results are based on 1000 simulations and 5000 observations.

Table 12: Simulation 4 - treatment exogeneity and selection on current outcomes, n=5000

	selection on current outcomes											
	α=0.68, β=0.5, γ=0, δ=0.25, κ=0						α=1.35, β=0.5, γ=0, δ=0.25, κ=0					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	-0.068	0.034	0.076	-0.234	0.053	0.240	-0.077	0.034	0.084	-0.272	0.055	0.277
$\hat{\omega}$	0.102	0.096	0.140	0.338	0.118	0.358	0.058	0.053	0.079	0.202	0.070	0.214
$\hat{\theta}$	0.343	23.756	23.758	2.403	299.225	299.235	-0.158	0.077	0.175	-0.460	0.147	0.483
$\hat{\theta}_{\text{trim}}(0.15)$	-0.521	1.105	1.222	0.911	55.800	55.808	-0.158	0.077	0.175	-0.460	0.147	0.483
$\hat{\theta}_{\text{trim}}(0.10)$	-0.545	1.250	1.364	1.338	56.241	56.257	-0.158	0.077	0.175	-0.460	0.147	0.483
$\hat{\theta}_{\text{trim}}(0.05)$	-0.540	1.011	1.146	1.094	57.364	57.374	-0.158	0.077	0.175	-0.460	0.147	0.483
$\hat{\theta}_{\text{trim}}(0.02)$	-0.538	0.892	1.041	0.595	60.179	60.181	-0.158	0.077	0.175	-0.460	0.147	0.483
$\hat{\theta}_{\text{trim}}(0.01)$	-0.534	0.993	1.128	0.373	62.805	62.807	-0.158	0.077	0.175	-0.460	0.147	0.483
$\hat{\phi}$	-0.259	0.136	0.293	-0.657	0.213	0.690	-0.149	0.069	0.165	-0.397	0.106	0.411
$\hat{\phi}_{\text{trim}}(0.15)$	-0.259	0.136	0.293	-0.657	0.213	0.690	-0.149	0.069	0.165	-0.397	0.106	0.411
$\hat{\phi}_{\text{trim}}(0.10)$	-0.259	0.136	0.293	-0.657	0.213	0.690	-0.149	0.069	0.165	-0.397	0.106	0.411
$\hat{\phi}_{\text{trim}}(0.05)$	-0.259	0.136	0.293	-0.657	0.213	0.690	-0.149	0.069	0.165	-0.397	0.106	0.411
$\hat{\phi}_{\text{trim}}(0.02)$	-0.259	0.136	0.293	-0.657	0.213	0.690	-0.149	0.069	0.165	-0.397	0.106	0.411
$\hat{\phi}_{\text{trim}}(0.01)$	-0.259	0.136	0.293	-0.657	0.213	0.690	-0.149	0.069	0.165	-0.397	0.106	0.411
MAR G-comp.	-0.054	0.034	0.064	-0.210	0.052	0.216	-0.067	0.034	0.075	-0.252	0.055	0.258
MAR TMLE	-0.054	0.034	0.064	-0.105	0.053	0.117	-0.068	0.034	0.076	-0.129	0.056	0.140
MAR IPW	-0.054	0.034	0.064	-0.110	0.055	0.122	-0.068	0.034	0.076	-0.134	0.058	0.146
MAR AIPW	-0.054	0.034	0.064	-0.109	0.055	0.123	-0.068	0.034	0.076	-0.134	0.059	0.146
Heckman	-0.002	0.047	0.047	-0.024	0.067	0.071	-0.002	0.055	0.055	-0.038	0.075	0.084
true CACE		0.500			1.250			0.500			1.250	
mean response		0.713			0.549			0.726			0.575	

Note: Results are based on 1000 simulations and 5000 observations.

H Simulation results for sample size 1000

The results presented in this section are based on the same simulation design as in Section 4, now with a sample size of 1000 (instead of 5000).

Table 13: Simulation 1 - treatment endogeneity and conditional LI, n=1000

	Homogeneous effects											
	α=0.68, β=0, γ=0.5, δ=0, κ=0						α=1.35, β=0, γ=0.5, δ=0, κ=0					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.694	0.075	0.698	0.926	0.131	0.935	0.553	0.075	0.558	0.694	0.125	0.705
$\hat{\omega}$	0.036	0.261	0.263	0.116	0.365	0.383	0.025	0.135	0.137	0.072	0.196	0.209
$\hat{\theta}$	-0.019	0.282	0.283	-0.091	0.606	0.613	-0.000	0.140	0.140	-0.017	0.205	0.206
$\hat{\theta}_{\text{trim}}(0.15)$	-0.019	0.282	0.283	-0.072	0.462	0.467	-0.000	0.140	0.140	-0.017	0.205	0.206
$\hat{\theta}_{\text{trim}}(0.10)$	-0.019	0.282	0.283	-0.069	0.458	0.464	-0.000	0.140	0.140	-0.017	0.205	0.206
$\hat{\theta}_{\text{trim}}(0.05)$	-0.019	0.282	0.283	-0.067	0.459	0.464	-0.000	0.140	0.140	-0.016	0.206	0.206
$\hat{\theta}_{\text{trim}}(0.02)$	-0.018	0.284	0.284	0.565	0.692	0.894	0.001	0.141	0.141	0.371	0.395	0.542
$\hat{\theta}_{\text{trim}}(0.01)$	1.262	0.328	1.304	2.109	0.617	2.197	0.809	0.157	0.824	1.183	0.246	1.208
$\hat{\phi}$	-0.160	0.323	0.360	-0.391	0.533	0.661	-0.097	0.154	0.182	-0.230	0.249	0.339
$\hat{\phi}_{\text{trim}}(0.15)$	-0.160	0.323	0.360	-0.391	0.533	0.661	-0.097	0.154	0.182	-0.230	0.249	0.339
$\hat{\phi}_{\text{trim}}(0.10)$	-0.160	0.323	0.360	-0.391	0.533	0.661	-0.097	0.154	0.182	-0.230	0.249	0.339
$\hat{\phi}_{\text{trim}}(0.05)$	-0.160	0.323	0.360	-0.391	0.533	0.661	-0.097	0.154	0.182	-0.224	0.250	0.336
$\hat{\phi}_{\text{trim}}(0.02)$	-0.158	0.324	0.360	0.387	0.990	1.063	-0.092	0.156	0.181	0.397	0.494	0.634
$\hat{\phi}_{\text{trim}}(0.01)$	1.437	0.366	1.483	2.834	0.657	2.909	0.904	0.172	0.920	1.526	0.251	1.546
MAR G-comp.	0.707	0.074	0.711	0.945	0.130	0.954	0.557	0.074	0.561	0.699	0.124	0.709
MAR TMLE	0.707	0.074	0.711	0.980	0.131	0.989	0.556	0.075	0.561	0.721	0.125	0.732
MAR IPW	0.709	0.074	0.713	0.987	0.132	0.995	0.559	0.075	0.564	0.726	0.126	0.737
MAR AIPW	0.707	0.074	0.711	0.982	0.133	0.991	0.556	0.075	0.561	0.721	0.126	0.732
Heckman	0.558	2.550	2.610	0.579	2.550	2.959	0.519	4.722	4.751	0.217	4.722	5.101
true CACE		0.500			1.250			0.500			1.250	
mean response		0.633			0.422			0.688			0.496	

	Heterogeneous effects											
	α=0.68, β=0, γ=0.5, δ=0, κ=0.5						α=1.35, β=0, γ=0.5, δ=0, κ=0.5					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.984	0.081	0.988	1.350	0.141	1.358	0.832	0.082	0.836	1.102	0.136	1.110
$\hat{\omega}$	0.029	0.295	0.297	0.110	0.426	0.440	0.023	0.158	0.160	0.077	0.237	0.249
$\hat{\theta}$	-0.031	0.312	0.313	-0.088	1.323	1.326	-0.004	0.162	0.162	-0.021	0.241	0.242
$\hat{\theta}_{\text{trim}}(0.15)$	-0.031	0.312	0.313	-0.101	0.539	0.549	-0.004	0.162	0.162	-0.021	0.241	0.242
$\hat{\theta}_{\text{trim}}(0.10)$	-0.031	0.312	0.313	-0.105	0.550	0.560	-0.004	0.162	0.162	-0.021	0.241	0.242
$\hat{\theta}_{\text{trim}}(0.05)$	-0.031	0.312	0.313	-0.098	0.534	0.542	-0.004	0.162	0.162	-0.021	0.241	0.242
$\hat{\theta}_{\text{trim}}(0.02)$	-0.029	0.314	0.316	0.604	0.846	1.040	-0.002	0.163	0.163	0.362	0.478	0.600
$\hat{\theta}_{\text{trim}}(0.01)$	1.810	0.391	1.852	2.848	0.848	2.972	1.086	0.187	1.102	1.582	0.283	1.607
$\hat{\phi}$	-0.224	0.357	0.421	-0.511	0.588	0.779	-0.148	0.177	0.231	-0.323	0.283	0.429
$\hat{\phi}_{\text{trim}}(0.15)$	-0.224	0.357	0.421	-0.511	0.588	0.779	-0.148	0.177	0.231	-0.323	0.283	0.429
$\hat{\phi}_{\text{trim}}(0.10)$	-0.224	0.357	0.421	-0.511	0.588	0.779	-0.148	0.177	0.231	-0.323	0.283	0.429
$\hat{\phi}_{\text{trim}}(0.05)$	-0.224	0.357	0.421	-0.511	0.588	0.779	-0.148	0.177	0.231	-0.317	0.285	0.426
$\hat{\phi}_{\text{trim}}(0.02)$	-0.222	0.359	0.422	0.281	1.171	1.204	-0.143	0.180	0.230	0.308	0.590	0.665
$\hat{\phi}_{\text{trim}}(0.01)$	2.077	0.450	2.125	3.778	0.753	3.853	1.199	0.209	1.217	1.961	0.278	1.981
MAR G-comp.	0.998	0.080	1.001	1.371	0.140	1.378	0.836	0.082	0.840	1.107	0.136	1.115
MAR TMLE	0.997	0.080	1.000	1.386	0.141	1.393	0.835	0.082	0.839	1.104	0.136	1.112
MAR IPW	0.999	0.080	1.002	1.393	0.142	1.400	0.838	0.082	0.842	1.108	0.137	1.117
MAR AIPW	0.997	0.080	1.000	1.387	0.143	1.394	0.835	0.082	0.839	1.102	0.137	1.111
Heckman	0.761	3.698	3.775	0.957	3.698	4.244	0.807	6.907	6.954	0.682	6.907	5.387
true CACE		0.637			1.455			0.479			1.218	
mean response		0.633			0.427			0.688			0.500	

Note: Results are based on 1000 simulations and 1000 observations.

Table 14: Simulation 2 - treatment endogeneity and MAR (with the instrument affecting response), n=1000

	Homogeneous effects											
	α=0.68, β=0.5, γ=0, δ=0, κ=0						α=1.35, β=0.5, γ=0, δ=0, κ=0					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.748	0.072	0.751	1.087	0.122	1.094	0.641	0.075	0.645	0.932	0.129	0.941
$\hat{\omega}$	0.197	0.209	0.288	0.472	0.267	0.543	0.114	0.117	0.164	0.284	0.162	0.327
$\hat{\theta}$	-38.9	648.1	649.3	-260.8	4642.1	4649.4	-0.066	0.185	0.197	0.937	37.9	37.9
$\hat{\theta}_{\text{trim}}(0.15)$	-0.413	14.416	14.422	-0.553	128.208	128.210	-0.066	0.185	0.197	-0.329	2.851	2.870
$\hat{\theta}_{\text{trim}}(0.10)$	-0.442	14.390	14.397	-0.787	129.622	129.625	-0.066	0.185	0.197	-0.309	2.898	2.915
$\hat{\theta}_{\text{trim}}(0.05)$	-0.391	13.794	13.799	-0.208	141.792	141.793	-0.066	0.185	0.197	-0.309	2.438	2.457
$\hat{\theta}_{\text{trim}}(0.02)$	-0.132	13.628	13.629	0.397	150.516	150.517	-0.066	0.186	0.197	0.178	2.519	2.525
$\hat{\theta}_{\text{trim}}(0.01)$	1.056	13.705	13.746	1.793	150.420	150.431	0.761	0.208	0.788	0.899	2.556	2.709
$\hat{\phi}$	0.005	0.310	0.310	-0.060	0.517	0.520	0.006	0.154	0.154	-0.029	0.258	0.260
$\hat{\phi}_{\text{trim}}(0.15)$	0.005	0.310	0.310	-0.060	0.517	0.520	0.006	0.154	0.154	-0.029	0.258	0.260
$\hat{\phi}_{\text{trim}}(0.10)$	0.005	0.310	0.310	-0.060	0.517	0.520	0.006	0.154	0.154	-0.029	0.258	0.260
$\hat{\phi}_{\text{trim}}(0.05)$	0.005	0.310	0.310	-0.060	0.517	0.520	0.006	0.154	0.154	-0.029	0.258	0.260
$\hat{\phi}_{\text{trim}}(0.02)$	0.006	0.311	0.311	0.691	0.972	1.193	0.007	0.155	0.155	0.387	0.506	0.637
$\hat{\phi}_{\text{trim}}(0.01)$	1.605	0.366	1.646	2.483	0.737	2.590	0.963	0.184	0.980	1.512	0.353	1.553
MAR G-comp.	0.764	0.071	0.767	1.116	0.121	1.122	0.650	0.074	0.654	0.952	0.128	0.960
MAR TMLE	0.764	0.071	0.767	1.150	0.121	1.157	0.650	0.075	0.654	0.991	0.128	1.000
MAR IPW	0.763	0.071	0.767	1.151	0.122	1.157	0.649	0.075	0.654	0.991	0.130	0.999
MAR AIPW	0.764	0.071	0.767	1.150	0.122	1.157	0.650	0.075	0.654	0.991	0.130	1.000
Heckman	0.858	0.080	0.862	1.302	0.080	1.311	0.831	0.103	0.837	1.392	0.103	1.460
true CACE		0.500			1.250			0.500			1.250	
mean response		0.684			0.490			0.694			0.508	

	Heterogeneous effects											
	α=0.68, β=0.5, γ=0, δ=0, κ=0.5						α=1.35, β=0.5, γ=0, δ=0, κ=0.5					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	1.052	0.078	1.055	1.547	0.133	1.553	0.941	0.082	0.944	1.390	0.140	1.397
$\hat{\omega}$	0.277	0.232	0.361	0.651	0.300	0.717	0.175	0.138	0.223	0.432	0.194	0.474
$\hat{\theta}$	-38.860	648.1	649.3	-260.6	4642.1	4649.4	-0.006	0.201	0.201	1.065	37.9	37.9
$\hat{\theta}_{\text{trim}}(0.15)$	-0.335	14.418	14.422	-0.396	128.210	128.211	-0.006	0.201	0.201	-0.201	2.852	2.859
$\hat{\theta}_{\text{trim}}(0.10)$	-0.364	14.392	14.397	-0.630	129.624	129.625	-0.006	0.201	0.201	-0.181	2.899	2.905
$\hat{\theta}_{\text{trim}}(0.05)$	-0.312	13.796	13.800	-0.051	141.794	141.794	-0.006	0.201	0.201	-0.182	2.440	2.447
$\hat{\theta}_{\text{trim}}(0.02)$	-0.053	13.630	13.630	0.561	150.529	150.530	-0.006	0.201	0.201	0.296	2.529	2.546
$\hat{\theta}_{\text{trim}}(0.01)$	1.603	13.705	13.798	2.529	150.432	150.454	1.077	0.229	1.101	1.359	2.557	2.896
$\hat{\phi}$	-0.003	0.341	0.341	-0.062	0.564	0.567	0.003	0.176	0.176	-0.020	0.289	0.290
$\hat{\phi}_{\text{trim}}(0.15)$	-0.003	0.341	0.341	-0.062	0.564	0.567	0.003	0.176	0.176	-0.020	0.289	0.290
$\hat{\phi}_{\text{trim}}(0.10)$	-0.003	0.341	0.341	-0.062	0.564	0.567	0.003	0.176	0.176	-0.020	0.289	0.290
$\hat{\phi}_{\text{trim}}(0.05)$	-0.003	0.341	0.341	-0.062	0.564	0.567	0.003	0.176	0.176	-0.020	0.289	0.290
$\hat{\phi}_{\text{trim}}(0.02)$	-0.001	0.343	0.343	0.703	1.151	1.348	0.004	0.177	0.177	0.401	0.601	0.722
$\hat{\phi}_{\text{trim}}(0.01)$	2.306	0.455	2.351	3.552	0.829	3.647	1.311	0.220	1.329	2.056	0.377	2.090
MAR G-comp.	1.069	0.077	1.072	1.578	0.131	1.584	0.950	0.081	0.954	1.411	0.139	1.418
MAR TMLE	1.068	0.077	1.071	1.599	0.131	1.604	0.950	0.082	0.953	1.430	0.140	1.437
MAR IPW	1.068	0.077	1.071	1.599	0.133	1.605	0.949	0.082	0.953	1.429	0.141	1.436
MAR AIPW	1.068	0.077	1.071	1.599	0.132	1.604	0.950	0.082	0.953	1.429	0.141	1.436
Heckman	1.208	0.091	1.211	1.897	0.091	1.910	1.248	0.222	1.268	2.357	0.222	2.472
true CACE		0.636			1.454			0.479			1.218	
mean response		0.684			0.495			0.694			0.511	

Note: Results are based on 1000 simulations and 1000 observations.

Table 15: Simulation 3 - treatment endogeneity and selection on current outcomes, n=1000

	Homogeneous effects											
	α=0.68, β=0.5, γ=0, δ=0.25, κ=0						α=1.35, β=0.5, γ=0, δ=0.25, κ=0					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.680	0.069	0.684	0.854	0.110	0.861	0.563	0.073	0.568	0.671	0.119	0.681
$\hat{\omega}$	0.085	0.215	0.231	0.364	0.269	0.452	0.044	0.118	0.126	0.200	0.161	0.257
$\hat{\theta}$	12.4	357.6	357.9	-193.6	4602.4	4606.5	-0.210	0.179	0.276	-0.323	10.4	10.4
$\hat{\theta}_{\text{trim}}(0.15)$	-0.087	19.007	19.007	-3.874	94.212	94.291	-0.210	0.179	0.276	-0.681	1.062	1.261
$\hat{\theta}_{\text{trim}}(0.10)$	-0.377	17.696	17.700	-4.419	97.841	97.941	-0.210	0.179	0.276	-0.657	1.076	1.261
$\hat{\theta}_{\text{trim}}(0.05)$	-0.333	16.454	16.457	-4.170	97.883	97.972	-0.210	0.179	0.276	-0.603	0.990	1.159
$\hat{\theta}_{\text{trim}}(0.02)$	-0.340	13.998	14.002	-3.848	105.264	105.335	-0.210	0.179	0.276	-0.314	1.019	1.067
$\hat{\theta}_{\text{trim}}(0.01)$	0.833	12.172	12.200	-1.258	106.525	106.532	0.541	0.285	0.611	0.856	1.045	1.350
$\hat{\phi}$	-0.249	0.310	0.397	-0.567	0.477	0.741	-0.137	0.151	0.204	-0.335	0.235	0.410
$\hat{\phi}_{\text{trim}}(0.15)$	-0.249	0.310	0.397	-0.567	0.477	0.741	-0.137	0.151	0.204	-0.335	0.235	0.410
$\hat{\phi}_{\text{trim}}(0.10)$	-0.249	0.310	0.397	-0.567	0.477	0.741	-0.137	0.151	0.204	-0.335	0.235	0.410
$\hat{\phi}_{\text{trim}}(0.05)$	-0.249	0.310	0.397	-0.567	0.477	0.741	-0.137	0.151	0.204	-0.335	0.235	0.410
$\hat{\phi}_{\text{trim}}(0.02)$	-0.249	0.310	0.397	-0.563	0.483	0.741	-0.137	0.151	0.204	-0.250	0.247	0.351
$\hat{\phi}_{\text{trim}}(0.01)$	1.106	0.617	1.267	2.361	0.691	2.460	0.674	0.318	0.745	1.220	0.324	1.262
MAR G-comp.	0.698	0.068	0.701	0.890	0.110	0.897	0.574	0.072	0.579	0.698	0.119	0.708
MAR TMLE	0.698	0.068	0.701	1.009	0.110	1.015	0.574	0.073	0.578	0.829	0.119	0.838
MAR IPW	0.697	0.068	0.701	1.013	0.113	1.019	0.573	0.073	0.578	0.836	0.124	0.846
MAR AIPW	0.698	0.068	0.701	1.013	0.115	1.019	0.574	0.073	0.578	0.837	0.127	0.847
Heckman	0.872	0.086	0.876	1.370	0.086	1.450	0.999	0.405	1.078	2.037	0.405	2.346
true CACE		0.500			1.250			0.500			1.250	
mean response		0.710			0.546			0.724			0.572	

	Heterogeneous effects											
	α=0.68, β=0.5, γ=0, δ=0.25, κ=0.5						α=1.35, β=0.5, γ=0, δ=0.25, κ=0.5					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	0.997	0.075	1.000	1.330	0.118	1.336	0.891	0.080	0.895	1.187	0.127	1.194
$\hat{\omega}$	0.119	0.249	0.276	0.434	0.327	0.543	0.080	0.144	0.165	0.285	0.208	0.353
$\hat{\theta}$	12.4	357.6	357.9	-193.5	4602.4	4606.5	-0.172	0.199	0.263	-0.247	10.4	10.4
$\hat{\theta}_{\text{trim}}(0.15)$	-0.050	19.009	19.009	-3.807	94.204	94.281	-0.172	0.199	0.263	-0.603	1.073	1.231
$\hat{\theta}_{\text{trim}}(0.10)$	-0.340	17.698	17.701	-4.351	97.832	97.928	-0.172	0.199	0.263	-0.580	1.088	1.232
$\hat{\theta}_{\text{trim}}(0.05)$	-0.296	16.456	16.459	-4.103	97.873	97.959	-0.172	0.199	0.263	-0.525	1.003	1.132
$\hat{\theta}_{\text{trim}}(0.02)$	-0.303	14.000	14.003	-3.781	105.251	105.319	-0.172	0.199	0.263	-0.238	1.030	1.058
$\hat{\theta}_{\text{trim}}(0.01)$	1.069	12.165	12.212	-0.551	106.483	106.484	0.673	0.469	0.821	1.311	1.128	1.730
$\hat{\phi}$	-0.242	0.341	0.418	-0.546	0.523	0.756	-0.110	0.172	0.205	-0.272	0.266	0.380
$\hat{\phi}_{\text{trim}}(0.15)$	-0.242	0.341	0.418	-0.546	0.523	0.756	-0.110	0.172	0.205	-0.272	0.266	0.380
$\hat{\phi}_{\text{trim}}(0.10)$	-0.242	0.341	0.418	-0.546	0.523	0.756	-0.110	0.172	0.205	-0.272	0.266	0.380
$\hat{\phi}_{\text{trim}}(0.05)$	-0.242	0.341	0.418	-0.546	0.523	0.756	-0.110	0.172	0.205	-0.272	0.266	0.380
$\hat{\phi}_{\text{trim}}(0.02)$	-0.242	0.341	0.418	-0.547	0.524	0.757	-0.110	0.172	0.205	-0.189	0.275	0.334
$\hat{\phi}_{\text{trim}}(0.01)$	1.276	0.991	1.615	2.906	1.162	3.130	0.782	0.522	0.940	1.552	0.604	1.666
MAR G-comp.	1.017	0.074	1.019	1.370	0.117	1.375	0.903	0.079	0.906	1.215	0.128	1.222
MAR TMLE	1.016	0.074	1.018	1.489	0.117	1.494	0.902	0.080	0.905	1.334	0.128	1.341
MAR IPW	1.016	0.074	1.018	1.495	0.120	1.500	0.902	0.080	0.905	1.345	0.133	1.352
MAR AIPW	1.016	0.074	1.018	1.496	0.123	1.501	0.902	0.080	0.905	1.348	0.137	1.355
Heckman	1.306	0.144	1.314	2.215	0.144	2.311	1.872	0.710	2.002	2.800	0.710	3.442
true CACE		0.636			1.454			0.479			1.218	
mean response		0.720			0.564			0.733			0.587	

Note: Results are based on 1000 simulations and 1000 observations.

Table 16: Simulation 4 - treatment exogeneity and selection on current outcomes, n=1000

	selection on current outcomes											
	α=0.68, β=0.5, γ=0, δ=0.25, κ=0						α=1.35, β=0.5, γ=0, δ=0.25, κ=0					
	time period 1			time period 2			time period 1			time period 2		
	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse	bias	stddev	rmse
naive	-0.069	0.073	0.100	-0.232	0.116	0.259	-0.078	0.074	0.107	-0.269	0.121	0.295
$\hat{\omega}$	0.100	0.219	0.240	0.354	0.284	0.454	0.057	0.119	0.132	0.211	0.163	0.266
$\hat{\theta}$	-22.3	571.7	572.2	8.213	1116.4	1116.4	-0.163	0.170	0.235	0.448	33.4	33.4
$\hat{\theta}_{\text{trim}}\,(0.15)$	-0.483	13.256	13.265	3.488	220.794	220.821	-0.163	0.170	0.235	-0.446	0.664	0.799
$\hat{\theta}_{\text{trim}}\,(0.10)$	-0.606	13.179	13.193	3.099	220.029	220.051	-0.163	0.170	0.235	-0.440	0.633	0.771
$\hat{\theta}_{\text{trim}}\,(0.05)$	-0.547	12.627	12.639	3.893	212.300	212.336	-0.163	0.170	0.235	-0.438	0.634	0.770
$\hat{\theta}_{\text{trim}}\,(0.02)$	-0.573	12.802	12.815	3.996	213.017	213.054	-0.163	0.170	0.235	-0.374	0.642	0.743
$\hat{\theta}_{\text{trim}}\,(0.01)$	0.292	13.621	13.624	6.228	225.061	225.147	0.088	0.211	0.229	-0.083	0.646	0.652
$\hat{\phi}$	-0.262	0.310	0.406	-0.617	0.478	0.780	-0.151	0.153	0.215	-0.375	0.236	0.443
$\hat{\phi}_{\text{trim}}\,(0.15)$	-0.262	0.310	0.406	-0.617	0.478	0.780	-0.151	0.153	0.215	-0.375	0.236	0.443
$\hat{\phi}_{\text{trim}}\,(0.10)$	-0.262	0.310	0.406	-0.617	0.478	0.780	-0.151	0.153	0.215	-0.375	0.236	0.443
$\hat{\phi}_{\text{trim}}\,(0.05)$	-0.262	0.310	0.406	-0.617	0.478	0.780	-0.151	0.153	0.215	-0.375	0.236	0.443
$\hat{\phi}_{\text{trim}}\,(0.02)$	-0.262	0.310	0.406	-0.568	0.497	0.755	-0.151	0.153	0.215	-0.342	0.247	0.422
$\hat{\phi}_{\text{trim}}\,(0.01)$	0.706	0.373	0.799	1.478	0.580	1.588	0.258	0.210	0.332	0.338	0.274	0.435
MAR G-comp.	-0.054	0.073	0.091	-0.207	0.115	0.237	-0.068	0.074	0.100	-0.249	0.121	0.277
MAR TMLE	-0.055	0.073	0.091	-0.103	0.116	0.155	-0.069	0.074	0.101	-0.129	0.123	0.178
MAR IPW	-0.055	0.073	0.091	-0.106	0.119	0.159	-0.069	0.074	0.101	-0.133	0.128	0.184
MAR AIPW	-0.055	0.073	0.091	-0.106	0.120	0.161	-0.069	0.074	0.101	-0.133	0.130	0.186
Heckman	-0.006	0.105	0.105	-0.043	0.105	0.180	-0.011	0.123	0.123	-0.071	0.123	0.221
true CACE		0.500			1.250			0.500			1.250	
mean response		0.714			0.550			0.727			0.576	

Note: Results are based on 1000 simulations and 1000 observations.