Identifying Marginal Treatment Effects in the Presence of Sample Selection*

Otávio Bartalotti
Iowa State University and IZA

Désiré Kédagni
Iowa State University

Vitor Possebom
Yale University

Abstract

This article presents identification results for the marginal treatment effect (MTE) when there is sample selection. We show that the MTE is partially identified for individuals who are always observed regardless of treatment, and we derive sharp bounds on this parameter under four sets of assumptions. The first identification result combines the standard MTE assumptions without any restrictions to the sample selection mechanism. The second result imposes monotonicity of the sample selection variable with respect to the treatment, considerably shrinking the identified set. Third, we incorporate a stochastic dominance assumption which tightens the lower bound for the MTE. Finally, we provide a set of conditions that allows point identification for completeness. Our analysis extends to discrete instruments and distributional MTE. All the results rely on a mixture reformulation of the problem where the mixture weights are identified. We therefore extend the Lee (2009) trimming procedure to the MTE context. We propose some preliminary estimators for the bounds derived, provide a numerical example and simulations that corroborate the bounds feasibility and usefulness as an empirical tool. In future drafts, we plan to highlight the practical relevance of the results by analyzing the impacts of managed health care options on health outcomes and expenditures, following Deb, Munkin, and Trivedi (2006).

Keywords: Sample Selection, Instrumental Variable, Marginal Treatment Effect, Partial Identification, Principal Stratification, Program Evaluation, Mixture Models.
JEL Codes: C14, C31, C35.

*The present version is of September 15, 2019. First draft: August, 2018. We thank Joseph Altonji, Nathan Barker, Michael Bates, Ivan Canay, Xuan Chen, John Finlay, Carlos A. Flores, Thomas Fujiwara, John Eric Humphries, Yuichi Kitamura, Marianne Köhli, Helena Laneuville, Jaewon Lee, Ismael Mourifié, Yusuke Narita, Pedro Sant'anna, Masayuki Sawada, Azeem Shaikh, Edward Vytlacil, Stephanie Weber, Siuyuat Wong and seminar participants at Iowa State University, University of Iowa, Yale University, UC-Riverside, CEME Conference for Young Econometricians 2019, and the Bristol Econometrics Study Group for helpful discussions and Seung Jin Cho for excellent research assistance. Email addresses: bartalot@iastate.edu, dkedagni@iastate.edu and vitoraugusto.possebom@yale.edu. This paper replaces Bartalotti and Kedagni (2019) under the same title and Possebom (2019):"Sharp Bounds for the MTE with Sample Selection."

1 Introduction

Many interesting applications in the treatment effects literature involve two simultaneous identification challenges: endogenous selection into treatment and sample selection. For instance, in labor economics, when a researcher wishes to evaluate the college wage premium, she has to consider the individual's decision to attend college as well as their decision to participate in the labor market.

Similarly, when analyzing the effect of a job training program on wages, one needs to consider the individual's decision to attend college as well as their participation in the labor market. In the health sciences, the same identification challenges appear when analyzing the effect of a drug on well-being because the outcome of interest, health status, is observed only for those who survived since taking the drug. Moreover, in randomized control trials, non-compliance and differential attrition rates between treated and control groups lead to the same identification concerns. This double selection problem is also present when analyzing the effect of an educational intervention on short- and long-term outcomes and the effect of procedural laws on litigation outcomes.[1]

To address both identification challenges, we analyze a generalized sample selection model in which the realized outcome (e.g., wages) is observed only if the individual self-selects into the sample (e.g., employment status), and the treatment choice (e.g., education) is observed for all individuals in the data being analyzed. Furthermore, the choice of treatment is allowed to be endogenous, and can be related to the sample selection mechanism. In this paper, we derive novel sharp bounds on the marginal treatment effect (MTE) for individuals who would self-select into the sample regardless of the treatment status (MTE^{OO}). To do so, we propose four identification strategies under increasingly restrictive sets of assumptions that extend

[1] Training programs are studied by Heckman, LaLonde, and Smith (1999), Lee (2009) and Chen and Flores (2015). The college wage premium is analyzed by Altonji (1993), Card (1999) and Carneiro, Heckman, and Vytlacil (2011). Scarring effects are discussed by Heckman and Borjas (1980), Farber (1993) and Jacobson, LaLonde, and Sullivan (1993). Some education interventions are studied by Krueger and Whitmore (2001), Angrist, Bettinger, and Kremer (2006), Angrist, Lang, and Oreopoulos (2009), Chetty et al. (2011) and Dobbie and Jr. (2015). Medical treatments are analyzed by CASS (1984), Sexton and Hebel (1984) and U.S. Department of Health and Human Services (2004). Litigation outcomes are discussed by Helland and Yoon (2017). RCT with attrition are illustrated by DeMel, McKenzie, and Woodruff (2013) and Angelucci, Karlan, and Zinman (2015).

MTE identification to scenarios with endogenous sample selection.

Before detailing our identification strategies, it is important to understand the intuition and the importance of the MTE and of the MTE^{OO}. To do so, consider the college wage premium example (Carneiro, Heckman, and Vytlacil, 2011). College attendance influences both the likelihood of employment and wages, and wages are observed only for individuals who are employed. The MTE reflects the return across different levels of the (latent) cost of going to college. As a consequence, this parameter can be used to shed light on the heterogeneity of college wage premia, i.e., to understand who would benefit from going to college.[2] This knowledge can, then, be used to optimally design policies focusing on college affordability. Moreover, common parameters evaluated in the literature (e.g., ATE, ATT, ATU and $LATE$) could be positive even when most people are affected adversely by a policy, masking its effects. Furthermore, the MTE^{OO} reflects the returns to college at the intensive margin, i.e., to the group of individuals that would participate in the labor force even if they had not attended college.

Our first partial identification strategy leaves sample selection unrestricted, relying on standard constraints on the mechanism that governs selection into treatment: the instrument is exogenous and excluded from the outcome determination, the treatment choice is monotone in the instrument, and the propensity score is continuous.[3]

Our second partial identification result tightens the bounds around the MTE^{OO} by exploiting a "monotonicity in selection" assumption. This condition requires individuals to be at least as likely to be observed in the sample if they are treated. In the college wage premium example, this additional assumption imposes that the treatment can induce workers to join the labor force, but not the opposite.

Our third partial identification strategy further reduces the identified set by imposing a stochastic dominance assumption. This condition mandates that the subpopulation who self-selects into the sample regardless of the treatment status has better potential outcomes than the the subpopulation who self-select into the sample only when treated. Intuitively,

[2] as recently illustrated by Cornelissen et al. (2018), Bhuller et al. (2019) and Humphries et al. (2019)

[3] See Bjorklund and Moffitt (1987) and Heckman and Vytlacil (1999, 2001a, 2005) for a detailed discussion about the MTE when there is no sample selection.

workers that would be employed regardless of college attendance would earn higher wages after attending college than those that would choose to participate in the labor force only if they attended college.

Finally, point-identification can be achieved by imposing the strong assumption that the treatment has no impact on the sample selection behavior. In that case, the hypothetical workers' employment status would not be affected by college attendance.

Importantly, our identification argument relies on a reformulation of the conditional probabilities of the potential outcomes as a mixture between the latent groups of individuals who are "always observed" and "observed only when treated." This reformulation extends to the MTE case the trimming procedure proposed by Imai (2008), Lee (2009) and Chen and Flores (2015) in the context of identifying the average treatment effect (ATE) and the local average treatment effect (LATE). Crucially, since we are interested in the MTE, the trimming is based on the distribution of the potential outcome conditional on unobserved individual characteristics related to treatment receipt.

Our results can be used to construct bounds for any treatment effect parameter that can be written as a weighted average of the MTE^{OO}. For instance, one can immediately obtain sharp bounds on the ATE, the average treatment effect on the treated (ATT), any LATE (Imbens and Angrist, 1994) and any policy-relevant treatment effect (PRTE, Heckman and Vytlacil, 2001b) within the always-observed subpopulation.[4]

We extend our main results by deriving sharp bounds to a more general object of interest, the distributional marginal treatment effect (DMTE), which is the effect of the treatment on the distribution of the outcome for individuals at the margin for participation. From these, one can derive bounds on the quantile version of the marginal treatment effect, and possibly other parameters.

We also extend the results to cases where researchers only have access to multi-valued discrete instruments. We derive nonparametric sharp bounds on a weighted MTE for that case, which is conceptually similar to Chen and Flores (2015), and can be seen as an extension

[4]The weights that combine the MTE to generate other treatment effects parameters are discussed by Heckman and Vytlacil (2005), Carneiro and Lee (2009), and Carneiro, Heckman, and Vytlacil (2011).

of their results which focused on binary instruments only.

We contribute to the literature on identification of treatment effect parameters using an instrument in the presence of sample selection, which can be organized in three groups: methods focusing on self-selection into treatment, methods focusing on sample selection and methods addressing both problems.[5]

In the literature focusing on self-selection into treatment, Imbens and Angrist (1994) establish conditions under which we can identify the LATE. Heckman and Vytlacil (1999), Heckman and Vytlacil (2005) and Heckman, Urzua, and Vytlacil (2006) define the MTE and explain how to compute a wide array of treatment effect parameters as a weighted average of the MTE. However, if the support of the propensity score is not the unit interval, then it is not possible to non-parametrically point-identify some common treatment effects, such as the ATE, the ATT and the ATU. A parametric solution to this problem is given by Brinch, Mogstad, and Wiswall (2017), while a nonparametric solution is given by Mogstad, Santos, and Torgovitsky (2018).[6]

In the literature on identification of treatment effect parameters with sample selection, two well established solutions are (i) the control function approach (Heckman, 1979, Ahn and Powell, 1993, and Das, Newey, and Vella, 2003), and (ii) the use of auxiliary data (Chen, Hong, and Tarozzi, 2008). Alternatively, one can partially identify the parameter of interest while imposing weak monotonicity assumptions. For example, Lee (2009) imposes that sample selection is monotone on treatment assignment and exploits that restriction to sharply bound the ATE for the subpopulation of always-observed individuals (ATE^{OO}).[7]

In the intersection of both topics, a few authors address the problem of sample selection and endogenous treatment simultaneously. By using one instrument for selection into treatment and one instrument for selection into the sample, Fricke et al. (2015) and Lee and

[5] Those literatures are vast and we only briefly summarize them here.

[6] Other important contributions are made by Manski (1990), Manski (1997), Manski and Pepper (2000), Heckman and Vytlacil (2001a), Bhattacharya, Shaikh, and Vytlacil (2008), Chesher (2010), Chiburis (2010), Shaikh and Vytlacil (2011), Bhattacharya, Shaikh, and Vytlacil (2012), Chen, Flores, and Flores-Lagunes (2017), Huber, Laffers, and Mellace (2017), Kowalski (2018), Mourifie, Henry, and Meango (2018) and Zhou and Xie (2019).

[7] Other relevant contributions are made by Frangakis and Rubin (2002), Blundell et al. (2007), Imai (2008), Lechner and Mell (2010), Blanco, Flores, and Flores-Lagunes (2013), Mealli and Pacini (2013), Huber (2014), Behaghel et al. (2015) and Huber and Mellace (2015).

Salanie (2018) identify the LATE and the ATE. However, since finding a credible instrument for sample selection is challenging in some cases, developing alternative tools that only require an instrument for selection into treatment is important. Frolich and Huber (2014) point identify the LATE by assuming that there is no contemporaneous relationship between the censored outcomes and the sample selection problem conditioning on past non-censored outcome variables. Chen and Flores (2015) derive bounds for ATE within the always-observed compliers ($LATE^{OO}$) by combining one instrument with a double exclusion restriction and monotonicity assumptions on both the selection into treatment and the sample.[8]

We contribute to this literature by partially identify the MTE within the always-observed subpopulation, a treatment effect parameter that has not been previous analyzed in the presence of sample selection. Differently from Frolich and Huber (2014), we allow for a contemporaneous relationship between the potential censored outcomes and the sample selection problem. Distinctively from Fricke et al. (2015) and Lee and Salanie (2018), we require only one instrument for selection into the treatment. In comparison with Chen and Flores (2015), we additionally consider two different sets of assumptions: the case without monotone sample selection and the case when the treatment has no effect on selection.

The remainder of the paper is organized as follows. Section 2 presents the structural model and sample selection mechanism considered, followed by a discussion of our four sets of identifying assumptions. In Section 3, we provide the identification results for the MTE bounds in the case of a continuous instrument under our four sets of assumptions. Section 4 presents a numerical illustration, while Section 5 proposes a novel estimator for the bounds proposed. Two relevant extensions are provided in Section 6, with identification results for (i) distributional MTE, and (ii) discrete instruments. Section 7 concludes. The proofs, sharp testable implications of the model, and a Monte Carlo simulation for the estimator's performance are presented in the appendix.

[8]Other important contributions are made by Steinmayr (2014), Blanco et al. (2017), Kédagni (2018) and Semykina and Wooldrige (2018).

2 Analytical Framework

Following Heckman and Vytlacil (1999, 2001a, 2005), Lee (2009) and Chen and Flores (2015), we consider the generalized sample selection model, described in the potential outcomes framework:

$$\begin{cases} Y^* &= Y_1^* D + Y_0^*(1-D) \\ D &= \mathbb{1}\{V \leq P(Z)\} \\ S &= S_1 D + S_0(1-D) \\ Y &= Y^* S \end{cases} \tag{2.1}$$

where Z is a vector of observable instrumental variables (e.g., local unemployment rate at age 17, local earnings at age 17, local tuition at age 17) with support given by $\mathcal{Z} \subset \mathbb{R}^{d_z}$, D is the treatment status indicator (e.g., college education). The variable Y^* is the (censored) realized outcome variable (e.g., wages) with support $\mathcal{Y} \subset \mathbb{R}$, while Y_0^* and Y_1^* are the potential outcomes when the person is untreated and treated, respectively. Similarly, S is the realized sample selection indicator (employment status), and S_0 and S_1 are potential sample selection indicators when individuals are untreated and treated. Finally, Y is the observed outcome (e.g., labor earnings), and V represents unobserved individual characteristics (e.g., cognitive costs of attending college). The researcher observes only the vector (Y, D, S, Z), while Y_1^*, Y_0^*, S_1, S_0 and V are latent variables.[9]

The treatment status, D, is connected to the instrument Z and the unobserved characteristics V through the unknown function $P : \mathcal{Z} \to \mathbb{R}$. In this model, we assume that the individual takes the treatment when its cost V is less than or equal to a threshold $P(Z)$, i.e., we impose monotonicity of the treatment in the instrument Z (Imbens and Angrist, 1994) as shown by Vytlacil (2002). This setup is similar to the one proposed by Heckman and Vytlacil (2005) and leads to the definition of the marginal treatment effect (MTE),

$$MTE(p) = \mathbb{E}[Y_1^* - Y_0^* | V = p].$$

[9] For simplicity, we drop exogenous covariates from the model. All results derived in the paper hold conditionally on covariates.

In the setting analyzed here, the task of learning about the MTE is further complicated by the potential for nonrandom sample selection. As pointed out by Lee (2009), point identification of ATE is no longer possible, even if if treatment is randomly assigned. This problem leads him to develop bounds for the ATE. This paper combines the insights of these literatures to develop sharp bounds for the MTE under sample selection while allowing for treatment to be endogenously determined.

Similarly to the compliance groups defined by Imbens and Angrist (1994), we can also define four latent groups based on the potential sample selection indicators. The always-observed subpopulation is composed of individuals for whom $S_0 = 1$ and $S_1 = 1$, the observed-only-when-treated subpopulation is defined by $S_0 = 0$ and $S_1 = 1$, the observed-only-when-untreated subpopulation is defined by $S_0 = 1$ and $S_1 = 0$, and the never-observed subpopulation is defined by $S_0 = 0$ and $S_1 = 0$.[10] Those subgroups are summarized in Table 1.

Table 1: Employment status subgroups

subgroups	S_0	S_1	Designation
OO	1	1	Always-observed
ON	1	0	Observed-only-when-untreated
NO	0	1	Observed-only-when-treated
NN	0	0	Never-observed

Following Zhang, Rubin, and Mealli (2008) and Lee (2009), we focus on the subpopulation who is always-observed ($S_0 = 1, S_1 = 1$). This is the only group whose censored potential outcomes are observed in both treatment arms. For the other three subpopulations, their treatment effect are not point identified or bounded in a non-trivial way without further functional form assumptions, since at least one of their potential outcomes (Y_0^* or Y_1^*) is never observed.[11]

Hence, our target parameter is the MTE within the subpopulation who is always observed

[10] Since the conditioning subpopulation is determined by post-treatment outcomes, our work is also connected to the statistical literature known as principal stratification (Frangakis and Rubin, 2002). In particular, the four latent groups in this framework are called strata in the principal stratification literature.

[11] Note that, in some applications (e.g., analyzing the impact of a medical treatment on a health quality measure where selection is given by whether the patient is alive), the potential censored outcome Y_d^* is not even properly defined when $S_d = 0$ for $d \in \{0, 1\}$.

(MTE^{OO}):

$$MTE^{OO}(p) \coloneqq \mathbb{E}\left[Y_1^* - Y_0^* \,|V = p, S_0 = 1, S_1 = 1\right], \tag{2.2}$$

for any $p \in [0,1]$. In labor market applications where sample selection is due to observing wages only when agents are employed, this is the effect on wages for the subpopulation who is always employed. In medical applications where selection is due to the death of a patient, this is the effect on health quality for the subpopulation who survives regardless of treatment status. In the education literature where sample selection is due to students quitting school, it is the effect on test scores for the subpopulation who do not drop out of school regardless of treatment status. In all those cases, the target parameter captures the intensive margin of the treatment effect.[12]

Analogously to Lee (2009), identification of MTE^{OO} is complex because sample selection is nonrandom and is possibly impacted by the treatment. To address this issue, we consider four sets of increasingly restrictive assumptions that allow us to partially identify the target parameter. The identified sets shrinks as the assumptions strenghten, leading to point identification under the fourth set of assumptions.[13]

Following Imai (2008), Assumptions 1-5 are sufficient to partially identify MTE^{OO}.

Assumption 1 (Random Assignment)**.** *The vector of instruments Z is independent of all latent variables, i.e., $Z \perp\!\!\!\perp (Y_0^*, Y_1^*, S_0, S_1, V)$.*

Assumption 2 (Propensity Score is Continuous)**.** *$P(z)$ is a nontrivial function of z and the random variable $P(Z)$ is absolutely continuous with support given by $\mathcal{P} \subseteq [0,1]$.*

Assumption 3 (Positive Mass)**.** *Both treatment groups and the always-observed subpopulation exist i.e., $0 < \mathbb{P}[D=1] < 1$ and $\mathbb{P}[S_0 = 1, S_1 = 1 | V = p] > 0$ for any $p \in \mathcal{P}$.*

[12] If the researcher is interested in the extensive margin of the treatment effect, captured by the MTE on the observed outcome ($\mathbb{E}[Y_1 - Y_0 \,|V = p]$) and by the MTE on the selection indicator ($\mathbb{E}[S_1 - S_0 \,|V = p]$), she can apply the identification strategies described by Heckman, Urzua, and Vytlacil (2006), Brinch, Mogstad, and Wiswall (2017) and Mogstad, Santos, and Torgovitsky (2018).

[13] According to Tamer (2010, p. 167), this approach to identification "characterizes the informational content of various assumptions by providing a menu of estimates, each based on different sets of assumptions, some of which are plausible and some of which are not." Empirically, this approach is also illustrated by Kline and Tartari (2016).

Assumption 4 (Finite Moments)**.** *The first population moment of the potential outcomes for the always-observed subpopulation is finite, i.e.,* $\mathbb{E}\left[|Y_d^*| \,|\, S_0 = 1, S_1 = 1\right] < +\infty$ *for any* $d \in \{0, 1\}$.

Assumption 5 (Uniform Distribution of V)**.** *The conditional distribution of* V *is uniform over* $[0, 1]$*, i.e.,* $V \sim \mathcal{U}_{[0,1]}$.

Assumptions 2-4 are technical assumptions to ensure that our objects of interest are well-defined and are common in the literature about marginal treatment effects (Heckman, Urzua, and Vytlacil, 2006). Assumption 1 is a standard IV independence assumption. Intuitively, we rely on changes in Z shifting treatment status and, hence, sample participation to identify the marginal treatment effect bounds. Assumption 3 is crucial for the identification results and requires that there are always-observed individuals for all possible values of the unobserved heterogeneity V. This can be restrictive in practice, ruling out identification of the MTE for ranges of V in which receipt of treatment determines sample participation heavily. Assumption 5 can be seen as a normalization if one assumes that the latent variable V is absolutely continuous. Under the same normalization, the image of the function $P : \mathcal{Z} \to \mathbb{R}$ is contained in the unit interval.

Assumptions 1-5 form our first set of assumptions required for identification of the MTE for the always-observed individuals. Under those assumptions, the function $P(z)$ is identified and is equal to the propensity score $\mathbb{P}\left[D = 1|Z = z\right]$ as described in (Heckman and Vytlacil, 2005, p. 677). Indeed, $\mathbb{P}\left[D = 1|Z = z\right] = \mathbb{P}\left[V \leq P(z)|Z = z\right] = \mathbb{P}\left[V \leq P(z)\right] = P(z)$, where the second equality holds under Assumption 1 and the last holds under Assumption 5. Moreover, this first set of restrictions partially identifies the target parameter MTE^{OO}, as presented in Section 3.3.

We also stress that the identified set can be substantially tightened by imposing that the sample selection mechanism is monotone on the treatment.

Assumption 6 (Monotone Sample Selection)**.** *Treatment has a non-negative effect on the sample selection indicator for all individuals, i.e.,* $S_1 \geq S_0$.

This monotonicity assumption rules out the existence of the observed-only-when-untreated subpopulation and is commonly used in the literature about sample selection (Lee, 2009, Chen and Flores, 2015).[14] To obtain some intuition on the mechanisms behind this assumption, consider a the college wage premium example. An individual is employed when her job search skills $\vartheta(D)$, a function of college attendance, are above a threshold U_S so that

$$S = \mathbb{1}\{\vartheta(D) \geq U_S\}.$$

Additionally, suppose that college attendance does not decrease someone's job search skills, i.e., $\vartheta(1) \geq \vartheta(0)$, making it more likely that college graduates would be observed in the data. In such a case, Assumption 6 holds. However, if college attendance raises the agents' reservation wages, this assumption may not hold.

Assumptions 1-6 form our second set of identification assumptions and lead to the bounds for MTE^{OO} that are the main result of this paper, presented in Theorem 1. Importantly, this second set of assumptions has a testable implication: the treatment positively affects sample selection, i.e., $E[S_1 - S_0|V = p] \geq 0$, implying

$$\frac{\partial \mathbb{P}[S = 1|P(Z) = p]}{\partial p} \geq 0 \text{ for all } p \in \mathcal{P}. \tag{2.3}$$

In other words, the share of the population for which the outcome is observed rises with p. In Section 3, we discuss further testable implications, while in Appendix Section B we formally characterize sharp testable implications arising from those assumptions. In the college wage premium example, the likelihood of employment increases with the probability of taking attending college.

We can further shrink the identified set around the MTE^{OO}, by adding Assumption 7 and completing the third set of identifying assumptions.

Assumption 7 (Stochastic Dominance)**.** *The distribution of the potential outcome when*

[14] As in Lee (2009), this assumption can be stated as $S_1 \geq S_0$ with probability 1. For the sake of simplicity, we assume it to hold for all individuals. Manski (1997) and Manski and Pepper (2000) refer to this assumption as the "monotone treatment response" assumption. All results can be stated with some straightforward changes if the inequality in Assumption 6 holds in the opposite direction.

treated for the always-observed subpopulation first-order stochastically dominates the distribution of the same random variable for the observed-only-when-treated subpopulation, i.e.,

$$\mathbb{P}\left[Y_1^* \leq y | \, V = p, S_0 = 1, S_1 = 1\right] \leq \mathbb{P}\left[Y_1^* \leq y | \, V = p, S_0 = 0, S_1 = 1\right]$$

for any $y \in \mathcal{Y}$ *and any* $p \in \mathcal{P}$.

This dominance assumption imposes that the always-observed subpopulation has higher potential censored outcomes than the observed-only-when-treated group conditional on V. This type of assumption is common in the literature (Imai, 2008, Blanco, Flores, and Flores-Lagunes, 2013, Huber and Mellace, 2015, and Huber, Laffers, and Mellace, 2017) and is intuitively based on the argument that some population sub-groups have more favorable underlying characteristics than others.[15]

While this assumption is not directly testable, Chen and Flores (2015, Section 2.3) propose an indirect test for this assumption that compares average baseline characteristic between the always-observed and the observed-only-when-treated latent groups. If the always-observed group has worse characteristics at baseline than does the observed-only-when-treated group, than Assumption 7 is less likely to hold in the data. Intuitively, workers that would be employed regardless of college attendance would earn higher wages after attending college than those that would choose to participate in the labor force only if they attended college.

For completeness, note that adding Assumption 8 to Assumptions 1-5 allow us to *point-identify* the MTE^{OO}. These six restrictions form the fourth set of identifying assumptions considered in this paper.

Assumption 8 (Unit Mass)**.** *The always-observed and never-observed subpopulations are the only groups that exist, i.e.,* $\mathbb{P}\left[S_0 = S_1\right] = 1$.

This unit mass assumption imposes that the treatment has no impact on sample selection. Even though it is likely too strong for many applied contexts, it is stated here for theoretical

[15] All of our results can be stated if the inequality in Assumption 7 holds in the opposite direction, as it is the case if larger values of the outcome harms the agent. For example, the researcher might be interested on the effect of a drug on cholesterol levels and the selection is based on whether the patient is alive.

completeness. In the context of the college wage premium, the workers' employment status would not be affected by college attendance, indicating that workers tend to have very high or very low attachment to the labor force.

3 Identification Results

This section presents the identification results for $MTE^{OO}(p)$ under the different sets of assumptions described in Section 2 that are the main results of this paper. As stepping stones for our main identification results, Subsection 3.1 shows identification of the conditional joint distribution of $(Y_d^*, S_d)|\, V$ for any $d \in \{0,1\}$, while Subsection 3.2 shows that the distribution of the potential outcomes can be seen as a mixture of latent groups, an important feature of the model. In the following subsections, we sharply bound the MTE^{OO} under increasingly restrictive assumptions. First, we bound the MTE^{OO} without imposing any assumption on the selection mechanism (Subsection 3.3). We then tighten those bounds by additionally imposing the monotone sample selection assumption (Subsection 3.4) and the stochastic dominance assumption (Subsection 3.5). For completeness, we also show that the MTE^{OO} is point-identified under the unit mass assumption (Subsection 3.6). Finally, in Subsection 3.7, we discuss how to sharply bound treatment effect parameters that can be written as weighted averages of the $MTE^{OO}(p)$.

3.1 Identifying the Joint Distribution of Potential Outcome and Selection

Before we discuss the identification of MTE^{OO}, we need to point-identify the conditional joint distribution of each potential outcome and sample selection for different levels of individual heterogeneity, $(Y_d^*, S_d)|\, V$ for $d \in \{0,1\}$.

Under Assumptions 1-5, for any $p \in \text{int}\, \mathcal{P}$ and any Borel set $A \subseteq \mathcal{Y}$, we have that

$$\begin{aligned}
\mathbb{P}\left[Y \in A, S = 1, D = 1 | P(Z) = p\right] &= \mathbb{P}\left[Y_1^* \in A, S_1 = 1, V \leq p | P(Z) = p\right] \\
&= \mathbb{P}\left[Y_1^* \in A, S_1 = 1, V \leq p\right] \\
&= \mathbb{P}\left[Y_1^* \in A, S_1 = 1 | V \leq p\right] \mathbb{P}\left[V \leq p\right]
\end{aligned}$$

$$= \left(\int_0^p \mathbb{P}\left[Y_1^* \in A, S_1 = 1 | V = v\right] \frac{f_V(v)}{\mathbb{P}\left[V \leq p\right]} dv \right) \cdot \mathbb{P}\left[V \leq p\right]$$
$$= \int_0^p \mathbb{P}\left[Y_1^* \in A, S_1 = 1 | V = v\right] dv,$$

where the second equality follows from Assumption 1, the third and fourth equalities follow from the Law of Iterated Expectations, and the last equality follows from Assumption 5. By taking the derivative of each side of the above derived equality, we point-identify the conditional joint distribution of $(Y_1^*, S_1) | V = p$:

$$\mathbb{P}\left[Y_1^* \in A, S_1 = 1 | V = p\right] = \frac{\partial \mathbb{P}\left[Y \in A, S = 1, D = 1 | P(Z) = p\right]}{\partial p}. \tag{3.1}$$

Similarly, we can show that

$$\mathbb{P}\left[Y_0^* \in A, S_0 = 1 | V = p\right] = -\frac{\partial \mathbb{P}\left[Y \in A, S = 1, D = 0 | P(Z) = p\right]}{\partial p}. \tag{3.2}$$

Note that equations (3.1) and (3.2) generate two testable implications for Assumptions 1 and 5:

$$0 \leq \frac{\partial \mathbb{E}[\mathbb{1}\left\{Y \in A\right\} SD | P(Z) = p]}{\partial p} \leq 1, \tag{3.3}$$
$$0 \leq -\frac{\partial \mathbb{E}[\mathbb{1}\left\{Y \in A\right\} S(1-D) | P(Z) = p]}{\partial p} \leq 1, \tag{3.4}$$

for all Borel sets $A \subset \mathbb{R}$ and $p \in (0, 1)$. Intuitively, when looking at people for whom observable characteristics (Z) indicate a higher likelihood of taking treatment, the share of treated (untreated) individuals that self-select into the sample should increase (decrease) for any range of the outcome.[16]

Similarly to the Local Instrumental Variable approach proposed by Heckman and Vytlacil (2005), equations (3.1) and (3.2) can be used to point-identify the MTE on the probability of being observed ($\mathbb{E}[S_1 - S_0 | V = p]$), capturing the extensive margin of the treatment. Note

[16] A formal characterization of the sharp testable implications implied by our model is presented in Appendix Section B.

that, for $A = \mathcal{Y}$,

$$\mathbb{P}\left[S_1 = 1 | V = p\right] = \frac{\partial \mathbb{P}\left[S = 1, D = 1 | P(Z) = p\right]}{\partial p}, \tag{3.5}$$

$$\mathbb{P}\left[S_0 = 1 | V = p\right] = -\frac{\partial \mathbb{P}\left[S = 1, D = 0 | P(Z) = p\right]}{\partial p}, \tag{3.6}$$

implying that $\mathbb{E}[S_1 - S_0 | V = p] = \frac{\partial \mathbb{E}[S|P(Z)=p]}{\partial p}$. This effect could be of interest in itself: for example, the researcher may want to evaluate whether a training program increases employment levels. We can also identify the MTE on the observed outcome,

$$\mathbb{E}[Y_1^* S_1 - Y_0^* S_0 | V = p] = \frac{\partial \mathbb{E}[YS|P(Z) = p]}{\partial p}.$$

However, in general, potential outcomes are dependent on the sample selection status even conditional on V. For example, labor force participation and potential wages both depend on the individual's reservation wage regardless of the education cost.

We would like to disentangle the marginal treatment on the observed outcome into the extensive margin and the intensive margin. While the extensive margin is point-identified, we show that the intensive margin (MTE^{OO}) is partially identified by exploiting the fact that the distribution of potential outcomes is a mixture of latent groups.

3.2 Potential Outcomes as Mixtures of Latent Groups

Fundamental to our identification strategy is recognizing that the treated (untreated) group is composed only by OO and NO (ON) types, as described in Table 1. Hence, the conditional distribution $Y_1^* | S_1 = 1, V = p$ can be written as the mixture of these latent distributions. For notational simplicity, let $\alpha(p) \equiv \frac{\mathbb{P}[OO|V=p]}{\mathbb{P}[S_1=1|V=p]}$ be the share of always-observed individuals among those for which $S_1 = 1$ conditional on $V = p$. Naturally, the remainder, $\frac{\mathbb{P}[NO|V=p]}{\mathbb{P}[S_1=1|V=p]}$, can be described as $1 - \alpha(p)$. By the Law of Total Probability, we have that:

$$\begin{aligned}\mathbb{P}\left[Y_1^* \in A | S_1 = 1, V = p\right] = \alpha(p) \cdot \mathbb{P}\left[Y_1^* \in A | S_0 = 1, S_1 = 1, V = p\right] \\ + \left(1 - \alpha\left(p\right)\right) \cdot \mathbb{P}\left[Y_1^* \in A | S_0 = 0, S_1 = 1, V = p\right].\end{aligned} \tag{3.7}$$

As a consequence, the expectation $\mathbb{E}[Y_1^*|S_1 = 1, V = p]$ is also a mixture of the expectation of Y_1^* for the always-observed and for observed-only-when-treated given the unobserved characteristic $V = p$

$$\begin{aligned}\mathbb{E}\left[Y_1^* \middle| S_1 = 1, V = p\right] = \alpha\left(p\right) \cdot \mathbb{E}\left[Y_1^* \middle| S_0 = 1, S_1 = 1, V = p\right] \\ + \left(1 - \alpha\left(p\right)\right) \cdot \mathbb{E}\left[Y_1^* \middle| S_0 = 0, S_1 = 1, V = p\right].\end{aligned}$$

Similarly, the conditional distribution of $Y_0^* | S_0 = 1, V = p$ can be written as the mixture of $Y_d^* | V = p$ for two latent groups, the always-observed and the observed-only-when-untreated group.

$$\begin{aligned}\mathbb{E}\left[Y_0^* \middle| S_0 = 1, V = p\right] = \beta\left(p\right) \cdot \mathbb{E}\left[Y_0^* \middle| S_0 = 1, S_1 = 1, V = p\right] \\ + \left(1 - \beta\left(p\right)\right) \cdot \mathbb{E}\left[Y_0^* \middle| S_0 = 1, S_1 = 0, V = p\right],\end{aligned}$$

where $\beta(p) \equiv \frac{\mathbb{P}[OO|V=p]}{\mathbb{P}[S_0=1|V=p]}$.

We exploit these mixture representations to bound the marginal treatment response of the censored treated outcome within the always-observed subpopulation ($\mathbb{E}\left[Y_1^* \middle| S_0 = 1, S_1 = 1, V = p\right]$) by considering the tails of the observed outcomes' distribution for treated individuals. The smallest attainable value of $\mathbb{E}[Y_1^*|S_0 = 1, S_1 = 1, V = p]$ is obtained when we consider the scenario in which the always-observed individuals are contained entirely in the left tail of mass $\alpha(p)$ of the outcome distribution, i.e., the lowest values of Y_1^* among the subpopulation $\{S_1 = 1\}$ conditional on V being equal to p. Respectively, the largest attainable value of $\mathbb{E}[Y_1^*|S_0 = 1, S_1 = 1, V = p]$ is obtained in the case that the always-observed individuals would be the right tail of the same distribution, getting the highest values of Y_1^* on that subpopulation. This is the same intuition behind the trimming procedure suggested by Lee (2009) and Chen and Flores (2015).

Hence, $\mathbb{E}\left[Y_1^* \middle| S_0 = 1, S_1 = 1, V = p\right]$ lies within the interval $[LB_1(p), UB_1(p)]$, where

$$LB_1(p) = \mathbb{E}\left[Y_1^* | S_1 = 1, V = p, Y_1^* \leq F^{-1}_{Y_1^*|S_1=1,V=p}\left(\alpha(p)\right)\right], \tag{3.8}$$

$$UB_1(p) = \mathbb{E}\left[Y_1^* | S_1 = 1, V = p, Y_1^* > F^{-1}_{Y_1^*|S_1=1,V=p}\left(1 - \alpha(p)\right)\right] \tag{3.9}$$

and $F^{-1}_{Y_d^*|S_d=1,V=p}(\cdot)$ is the quantile function of the distribution of Y_d^* given $S_d = 1$ and $V = p$.

Similarly, the conditional distribution of $Y_0^* | S_0 = 1, V = p$ can be written as the mixture of $Y_d^* | V = p$ for two latent groups, the always-observed and the observed-only-when-untreated group. Analogously to the treated outcome, the marginal treatment response of the untreated outcome within the always-observed subpopulation ($\mathbb{E}[Y_0^* | S_0 = 1, S_1 = 1, V = p]$) lies within the interval $[LB_0(p), UB_0(p)]$, where

$$LB_0(p) = \mathbb{E}\left[Y_0^* | S_0 = 1, V = p, Y_0^* \leq F^{-1}_{Y_0^*|S_0=1,V=p}\left(\beta(p)\right)\right], \tag{3.10}$$

$$UB_0(p) = \mathbb{E}\left[Y_0^* | S_0 = 1, V = p, Y_0^* > F^{-1}_{Y_0^*|S_0=1,V=p}\left(1 - \beta(p)\right)\right]. \tag{3.11}$$

Combining the bounds around $\mathbb{E}[Y_1^* | S_0 = 1, S_1 = 1, V = p]$ and $\mathbb{E}[Y_0^* | S_0 = 1, S_1 = 1, V = p]$, we find that $MTE^{OO}(p)$ lies within the interval

$$[LB_1(p) - UB_0(p), UB_1(p) - LB_0(p)].$$

Remark 1. *The issue central to identification of the target parameter is what can be learned about the mixture weights $(\alpha(p), \beta(p))$, and $\mathbb{E}[Y_d^* | S_d = 1, V = p]$. Note that the bounds on the MTE of interest will be tighter for higher values of $\alpha(p)$ and $\beta(p)$ because we learn about $E[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1, V = p]$ by considering that the worst- and best-case outcomes of observed treated and untreated individuals are fully attributed to the always-observed. So, as $\alpha(p)$ increases, the share of the observed sample of treated individuals that are from our group of interest increases, providing more information about their conditional expectation of the outcomes. In the extreme case in which $\alpha(p) \to 1$, the $\mathbb{E}[Y_1^* | S_0 = 1, S_1 = 1, V = p]$ will be point identified. Similarly, if $\alpha(p) \to 0$, the observed sample is uninformative about the always-observed group. A similar intuition holds regarding $\beta(p)$.*

In the next subsections we investigate the bounds that are generated under four alternative sets of assumptions, described in Section 2. Intuitively, those assumptions impose different

restrictions on the possible values of the mixture weights $(\alpha(p), \beta(p))$, providing different sets of information about $E[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1, V = p]$.

3.3 Identification with No Assumption on the Sample Selection Mechanism

Initially, consider the case in which the researcher is only willing to consider Assumptions 1-5, leaving the sample selection mechanism unrestricted. To learn about MTE^{OO}, we need information about the share of always-observed individuals in the total population, $\mathbb{P}[S_0 = 1, S_1 = 1]$ and, hence, the conditional joint distribution of $(S_0, S_1) | V = p$. However, we only have information about the conditional marginal distributions $S_0 | V = p$ and $S_1 | V = p$ based on equations (3.5) and (3.6). According to Imai (2008) and Mullahy (2018), the following Boole-Fréchet bounds are sharp around the share of always-observed individuals:

$$\begin{aligned}\mathbb{P}[S_0 = 1, S_1 = 1 | V = p] \in [&\max\{\mathbb{P}[S_0 = 1 | V = p] + \mathbb{P}[S_1 = 1 | V = p] - 1, 0\}, \\ &\min\{\mathbb{P}[S_0 = 1 | V = p], \mathbb{P}[S_1 = 1 | V = p]\}]. \end{aligned} \tag{3.12}$$

Combining this information with Equations (3.5) and (3.6), leads to:

Lemma 1. *Under Assumptions 1-5, the share of always-observed individuals is partially identified:*

$$\begin{aligned}&\mathbb{P}[S_0 = 1, S_1 = 1 | V = p] \\ &\quad \in \left[\max\left\{-\frac{\partial \mathbb{P}[S = 1, D = 0 | P(Z) = p]}{\partial p} + \frac{\partial P[S = 1, D = 1 | P(Z) = p]}{\partial p} - 1, 0\right\}, \right. \\ &\qquad \left. \min\left\{-\frac{\partial \mathbb{P}[S = 1, D = 0 | P(Z) = p]}{\partial p}, \frac{\partial \mathbb{P}[S = 1, D = 1 | P(Z) = p]}{\partial p}\right\}\right] \\ &\quad =: \Upsilon(p).\end{aligned}$$

These bounds are sharp.

Note that, since $\Upsilon(p)$ provides the identified set of possible values for the share of always-observed individuals, we can obtain the equivalent range of possible values for $\alpha(p)$ and $\beta(p)$, the mixture weights described in Subsection 3.2. For brevity, let $\mathbb{P}[S_0 = 1, S_1 = 1 | V = p]$ take

any particular value, $\upsilon \in \Upsilon(p)$. Define,

$$\alpha(p,\upsilon) \coloneqq \mathbb{P}[S_0 = 1 | S_1 = 1, V = p] = \frac{\upsilon}{\mathbb{P}[S_1 = 1 | V = p]} = \frac{\upsilon}{\frac{\partial \mathbb{P}[S=1, D=1|P(Z)=p]}{\partial p}},$$
$$\beta(p,\upsilon) \coloneqq \mathbb{P}[S_1 = 1 | S_0 = 1, V = p] = \frac{\upsilon}{\mathbb{P}[S_0 = 1 | V = p]} = -\frac{\upsilon}{\frac{\partial \mathbb{P}[S=1, D=0|P(Z)=p]}{\partial p}}.$$

Let the bounds in Equations (3.8)-(3.11), for specific values of $\alpha(p,\upsilon)$ and $\beta(p,\upsilon)$ in the identified set be written as:

$$LB_1(p,\upsilon) = \mathbb{E}\left[Y_1^* | S_1 = 1, V = p, Y_1^* \leq F^{-1}_{Y_1^*|S_1=1,V=p}(\alpha(p,\upsilon))\right],$$
$$UB_1(p,\upsilon) = \mathbb{E}\left[Y_1^* | S_1 = 1, V = p, Y_1^* > F^{-1}_{Y_1^*|S_1=1,V=p}(1-\alpha(p,\upsilon))\right],$$
$$LB_0(p,\upsilon) = \mathbb{E}\left[Y_0^* | S_0 = 1, V = p, Y_0^* \leq F^{-1}_{Y_0^*|S_0=1,V=p}(\beta(p,\upsilon))\right],$$
$$UB_0(p,\upsilon) = \mathbb{E}\left[Y_0^* | S_0 = 1, V = p, Y_0^* > F^{-1}_{Y_0^*|S_0=1,V=p}(1-\beta(p,\upsilon))\right].$$

Combining the bounds around $\mathbb{E}[Y_1^* | S_0 = 1, S_1 = 1, V = p]$ and $\mathbb{E}[Y_0^* | S_0 = 1, S_1 = 1, V = p]$, we find that $MTE^{OO}(p)$ lies within the interval $[LB_1(p,\upsilon) - UB_0(p,\upsilon), UB_1(p,\upsilon) - LB_0(p,\upsilon)]$ for a particular $\mathbb{P}[S_0 = 1, S_1 = 1 | V = p] = \upsilon$.

To bound the target parameter, we find worst- and best-case scenarios for the target parameter by varying the value υ. Explicitly, $MTE^{OO}(p)$ is partially identified and lies within the interval

$$\left[\min_{\upsilon \in \Upsilon(p)} \{LB_1(p,\upsilon) - UB_0(p,\upsilon)\}, \max_{\upsilon \in \Upsilon(p)} \{UB_1(p,\upsilon) - LB_0(p,\upsilon)\}\right].$$

Note that υ has a monotone relationship to the mixture weights, which define the trimming points in the bounds. As previously discussed, higher values for $\alpha(p)$ ($\beta(p)$) indicate that a bigger share of the observed treated (untreated) population belongs to the always-observed latent group, thus providing more information and tighter bounds for the parameter of interest. Hence, we only need to focus on the scenario that generates the wider bounds, that is, the

smallest admissible $\alpha(p)$ and $\beta(p)$. Let υ^{ℓ} be the lower bound of $\Upsilon(p)$. We have:

$$\min_{\upsilon\in\Upsilon(p)} LB_1(p,\upsilon) - \max_{\upsilon\in\Upsilon(p)} UB_0(p,\upsilon) \leq \min_{\upsilon\in\Upsilon(p)} \{LB_1(p,\upsilon) - UB_0(p,\upsilon)\},$$

$$\min_{\upsilon\in\Upsilon(p)} LB_1(p,\upsilon) = LB_1(p,\upsilon^{\ell}), \text{ and } \max_{\upsilon\in\Upsilon(p)} UB_0(p,\upsilon) = UB_0(p,\upsilon^{\ell}).$$

Making the same argument to the upper bound, we can rewrite them as,

$$\min_{\upsilon\in\Upsilon(p)} \{LB_1(p,\upsilon) - UB_0(p,\upsilon)\} = LB_1(p,\upsilon^{\ell}) - UB_0(p,\upsilon^{\ell}),$$

$$\max_{\upsilon\in\Upsilon(p)} \{UB_1(p,\upsilon) - LB_0(p,\upsilon)\} = UB_1(p,\upsilon^{\ell}) - LB_0(p,\upsilon^{\ell}),$$

greatly simplifying our bounds because the bounds need only to be evaluated at the end points of $\Upsilon(p)$.

We can combine these facts with equations (3.1), (3.2), (3.5) and (3.6) to propose the first identification result for MTE^{OO}, which does not impose meaningful restrictions on the sample selection mechanism:

Proposition 1. *Under Assumptions 1-5, the MTE is partially identified for the always-observed:*

$$\begin{aligned}
\underline{\Delta}_1(p) \coloneqq\ & \mathbb{E}\left[\tilde{Y}_1 | S=1, D=1, P(Z)=p, \tilde{Y}_1 \leq F^{-1}_{\tilde{Y}_1|S=1,D=1,P(Z)=p}\left(\alpha(p,\upsilon^{\ell})\right)\right] \\
& - \mathbb{E}\left[\tilde{Y}_0 | S=1, D=0, P(Z)=p, \tilde{Y}_0 > F^{-1}_{\tilde{Y}_0|S=1,D=0,P(Z)=p}\left(1-\beta(p,\upsilon^{\ell})\right)\right] \\
\leq\ & MTE^{OO}(p) \\
\leq\ & \mathbb{E}\left[\tilde{Y}_1 | S=1, D=1, P(Z)=p, \tilde{Y}_1 > F^{-1}_{\tilde{Y}_1|S=1,D=1,P(Z)=p}\left(1-\alpha(p,\upsilon^{\ell})\right)\right] \\
& - \mathbb{E}\left[\tilde{Y}_0 | S=1, D=0, P(Z)=p, \tilde{Y}_0 \leq F^{-1}_{\tilde{Y}_0|S=1,D=0,P(Z)=p}\left(\beta(p,\upsilon^{\ell})\right)\right] =: \overline{\Delta}_1(p),
\end{aligned}$$

where the conditional distribution of $\tilde{Y}_d$ is given by

$$\tilde{Y}_d | S=1, D=d, P(Z)=p \ \sim F_{\tilde{Y}_d|S=1,D=d,P(Z)=p}(y) = \frac{\frac{\partial \mathbb{P}[Y\leq y, S=1, D=d|P(Z)=p]}{\partial p}}{\frac{\partial \mathbb{P}[S=1, D=d|P(Z)=p]}{\partial p}}$$

for any $d \in \{0,1\}$,

$$\alpha\left(p, \upsilon^{\ell}\right) = \frac{\max\left\{-\frac{\partial \mathbb{P}[S=1,D=0|P(Z)=p]}{\partial p} + \frac{P[S=1,D=1|P(Z)=p]}{\partial p} - 1, 0\right\}}{\frac{\partial \mathbb{P}[S=1,D=1|P(Z)=p]}{\partial p}}$$

and

$$\beta\left(p, \upsilon^{\ell}\right) = -\frac{\max\left\{-\frac{\partial \mathbb{P}[S=1,D=0|P(Z)=p]}{\partial p} + \frac{P[S=1,D=1|P(Z)=p]}{\partial p} - 1, 0\right\}}{\frac{\partial \mathbb{P}[S=1,D=0|P(Z)=p]}{\partial p}}.$$

Moreover, these bounds are sharp.

Remark 2. *The definition of sharpness used here follows the definition of sharpness given by Canay and Shaikh (2017, Remark 2.1.). Intuitively, for any value* $\delta \in \left[\underline{\Delta}_1, \overline{\Delta}_1\right]$*, it is possible to construct random variables* $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ *that satisfy the restrictions imposed on the data by Assumptions 1-5 (equations (3.3) and (3.4)), induce the joint distribution on the data* (Y, S, D, Z) *and achieve* $\delta = \mathbb{E}\left[\tilde{Y}_1^* - \tilde{Y}_0^* \middle| \tilde{S}_0 = 1, \tilde{S}_1 = 1, \tilde{V} = p\right]$.

Remark 3. *In the special case in which the treatment* D *is independent of the potential outcomes for* Y^* *and* S *and that* $\Upsilon(p) = \Upsilon$ *for any* $p \in \mathcal{P} = [0,1]$*, the MTE on the censored outcome will be constant and equal to the average treatment effect for the always-observed and the bounds derived in Proposition 1 simplify to the ones derived by Imai (2008, Proposition 1).*

3.4 Bounds under the Monotonicity Assumption

In this subsection, we introduce monotonicity of treatment on the sample selection (Assumption 6) which considerably shrinks the identified set for MTE^{OO}. As discussed in Section 2, under the monotonicity assumption, individuals who self-select into the sample when untreated ($S_0 = 1$) would also be observed if they had been treated, ruling out the subgroup ON. In other words, any untreated individuals observed on the sample are members of the always-observed latent subpopulation ($S_0 = 1, S_1 = 1$).

Formally, we have that the following two events are identical: $\{S_0 = 1\} = \{S_0 = 1, S_1 = 1\}$ and the mixture weight for the untreated group, $\beta(p)$, equals one.

Consequently, $\mathbb{P}[S_0 = 1, S_1 = 1|V = p]$ is point-identified by equation (3.6), and we do not

need to rely on the partial identification results in Lemma 1 anymore. Specifically, we have that

$$\begin{aligned}\mathbb{P}\left[S_0 = 1, S_1 = 1 | V = p\right] &= -\frac{\partial \mathbb{P}\left[S = 1, D = 0 | P(Z) = p\right]}{\partial p} \\ &= \frac{\partial \mathbb{P}\left[S = 1, D = 1 | P(Z) = p\right]}{\partial p} - \frac{\partial \mathbb{P}\left[S = 1 | P(Z) = p\right]}{\partial p}.\end{aligned} \tag{3.13}$$

This result connects the conditional share of always observed individuals to changes on the conditional mass of observed untreated individuals when the propensity score increases. Looking at the second equality, we find that the conditional probability of being always observed is the difference between the increase in the share of observed treated individuals and the increase in the share of observed individuals when the propensity score is equal to p.

Since $\{S_0 = 1\} = \{S_0 = 1, S_1 = 1\}$ ($\beta(p) = 1$), the distribution of $(Y_0^*, S_0 = 1, S_1 = 1) | V$ is equal to the distribution of $(Y_0^*, S_0 = 1) | V$, implying that

$$\mathbb{P}\left[Y_0^* \in A | S_0 = 1, S_1 = 1, V = p\right] = \frac{\mathbb{P}\left[Y_0^* \in A, S_0 = 1 | V = p\right]}{\mathbb{P}\left[S_0 = 1, S_1 = 1 | V = p\right]}. \tag{3.14}$$

Note that the right-hand side of equation (3.14) is point-identified according to equations (3.2) and (3.13). Consequently, the expectation $\mathbb{E}[Y_0^* | S_0 = 1, S_1 = 1, V = p]$ is also point-identified. Note that, monotonicity also leads to point identification of the mixture weight, $\alpha(p) = \dfrac{\mathbb{P}\left[S_0 = 1, S_1 = 1 | V = p\right]}{\mathbb{P}\left[S_1 = 1 | V = p\right]}$ by Equations (3.5) and (3.13).

Then, under monotonicity, the researcher has to obtain bounds only for the potential outcomes under treatment, which still can be written as a mixture of the always-observed and observed-only-when-treated latent subpopulations.

As discussed in Section 3.2, the expectation $\mathbb{E}[Y_1^* | S_0 = 1, S_1 = 1, V = p]$ lies within the interval $[LB_1(p), UB_1(p)]$, given in Equations (3.8)-(3.9). We can combine the bounds with the identification results in equations (3.1), (3.2), (3.5), (3.13) and (3.14) to derive the following theorem:

Theorem 1. *Under Assumptions 1-6, the MTE is partially identified for the always-observed:*

$$\begin{aligned}\underline{\Delta}_2(p) &:= \mathbb{E}\left[\tilde{Y}_1 | S=1, D=1, P(Z)=p, \tilde{Y}_1 \leq F^{-1}_{\tilde{Y}_1|S=1,D=1,P(Z)=p}(\alpha(p))\right] \\ &\quad - \mathbb{E}\left[\tilde{Y}_0 | S=1, D=0, P(Z)=p\right] \\ &\leq MTE^{OO}(p) \\ &\leq \mathbb{E}\left[\tilde{Y}_1 | S=1, D=1, P(Z)=p, \tilde{Y}_1 > F^{-1}_{\tilde{Y}_1|S=1,D=1,P(Z)=p}(1-\alpha(p))\right] \\ &\quad - \mathbb{E}\left[\tilde{Y}_0 | S=1, D=0, P(Z)=p\right] =: \overline{\Delta}_2(p),\end{aligned}$$

where the conditional distribution of $\tilde{Y}_d$ is given by

$$\tilde{Y}_d | S=1, D=d, P(Z)=p \ \sim F_{\tilde{Y}_d|S=1,D=d,P(Z)=p}(y) = \frac{\frac{\partial \mathbb{P}[Y \leq y, S=1, D=d | P(Z)=p]}{\partial p}}{\frac{\partial \mathbb{P}[S=1, D=d | P(Z)=p]}{\partial p}}$$

for any $d \in \{0,1\}$ and

$$\alpha(p) = -\frac{\frac{\partial \mathbb{P}[[S=1, D=0 | P(Z)=p]}{\partial p}}{\frac{\partial \mathbb{P}[S=1, D=1 | P(Z)=p]}{\partial p}}.$$

Moreover, these bounds are sharp.

Remark 4. *The definition of sharpness used here follows the definition of sharpness given by Canay and Shaikh (2017, Remark 2.1.). Intuitively, for any value $\delta \in \left[\underline{\Delta}_2, \overline{\Delta}_2\right]$, it is possible to construct random variables $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ that satisfy the restrictions imposed on the data by Assumptions 1-6 (equations (2.3), (3.3) and (3.4)), induce the joint distribution on the data (Y, S, D, Z) and achieve $\delta = \mathbb{E}\left[\tilde{Y}_1^* - \tilde{Y}_0^* \middle| \tilde{S}_0 = 1, \tilde{S}_1 = 1, \tilde{V} = p\right]$.*

Remark 5. *Notice that, by adding the Monotonicity Assumption, we increase the lower bound and decrease the upper bound stated in Proposition 1. The length of the identified set here is strictly shorter than the length of the identified set in Proposition 1 when the mixture weights are not point identified.*[17] *This improvement clearly shows the identifying power of Assumption 6.*

[17] That is, the set $\Upsilon(p)$ in Lemma 1 is not a singleton and the distributions of $Y_0^* | S_1 = 1, V$ and $Y_1^* | S_1 = 1, V$ are not degenerate.

Remark 6. *In the special case in which the treatment D is independent of the potential outcomes for Y^* and S, and $\alpha(p) = \alpha$ for any $p \in \mathcal{P} = [0,1]$, the MTE will be constant and equal to the average treatment effect for the always-observed and the bounds proposed in Theorem 1 simplify to the ones proposed in Lee (2009, Proposition 1a).*

3.5 Bounds under the Monotonicity and Dominance Assumptions

In this subsection, we add the stochastic mean dominance assumption (Assumption 7) to tighten the identified set for MTE^{OO} under Assumptions 1-7. The stochastic dominance assumption and equation (3.7) imply that

$$\mathbb{P}\left[Y_1^* \leq y | \, S_1 = 1, V = p\right] \geq \mathbb{P}\left[Y_1^* \leq y | \, S_0 = 1, S_1 = 1, V = p\right]$$

for any $y \in \mathcal{Y}$. As a consequence, the following inequality holds

$$\mathbb{E}\left[Y_1^* | \, S_1 = 1, V = p\right] \leq \mathbb{E}\left[Y_1^* | \, S_0 = 1, S_1 = 1, V = p\right]$$

This tightens the lower bound for $\mathbb{E}[Y_1^*|S_0 = 1, S_1 = 1, V = p]$ as we no longer need to focus on the lowest $\alpha(p)$ mass of outcomes as the lower bound, since the stochastic dominance assumption guarantees that the expectation of outcomes for the always observed subpopulation will be larger than the one of the observed treated individuals which mixes OO and NO types. Hence, $\mathbb{E}[Y_1^*|S_0 = 1, S_1 = 1, V = p]$ lies within the interval $[LB_3(p), UB_3(p)]$, where

$$\begin{aligned} LB_3(p) &= \mathbb{E}\left[Y_1^* | \, S_1 = 1, V = p\right], \\ UB_3(p) &= \mathbb{E}\left[Y_1^* | S_1 = 1, V = p, Y_1^* > F^{-1}_{Y_1^*|S_1=1,V=p}(1 - \alpha(p))\right]. \end{aligned}$$

Note that the upper bound remains unchanged. Naturally, that leads to tighter identified sets relative to the ones in Theorem 1, which are presented in the following theorem:

Theorem 2. *Under Assumptions 1-7, the MTE is partially identified for the always-observed:*

$$\begin{aligned}\underline{\Delta}_3(p) &:= \mathbb{E}\left[\tilde{Y}_1 | S=1, D=1, P(Z)=p\right] - \mathbb{E}\left[\tilde{Y}_0 | S=1, D=0, P(Z)=p\right] \\ &\leq MTE^{OO}(p) \\ &\leq \mathbb{E}\left[\tilde{Y}_1 | S=1, D=1, P(Z)=p, \tilde{Y}_1 > F^{-1}_{\tilde{Y}_1|S=1,D=1,P(Z)=p}(1-\alpha(p))\right] \\ &\quad - \mathbb{E}\left[\tilde{Y}_0 | S=1, D=0, P(Z)=p\right] =: \overline{\Delta}_3(p),\end{aligned}$$

where the conditional distribution of $\tilde{Y}_d$ is given by

$$\tilde{Y}_d | S=1, D=d, P(Z)=p \ \sim F_{\tilde{Y}_d|S=1,D=d,P(Z)=p}(y) = \frac{\frac{\partial \mathbb{P}[Y \leq y, S=1, D=d | P(Z)=p]}{\partial p}}{\frac{\partial \mathbb{P}[S=1, D=d | P(Z)=p]}{\partial p}}$$

for any $d \in \{0,1\}$. and

$$\alpha(p) = -\frac{\frac{\partial \mathbb{P}[S=1, D=0 | P(Z)=p]}{\partial p}}{\frac{\partial \mathbb{P}[S=1, D=1 | P(Z)=p]}{\partial p}}.$$

Moreover, these bounds are sharp.

Remark 7. *The definition of sharpness used here follows the definition of sharpness given by Canay and Shaikh (2017, Remark 2.1.). Intuitively, for any value $\delta \in \left[\underline{\Delta}_3, \overline{\Delta}_3\right]$, it is possible to construct random variables $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ that satisfy the restrictions imposed on the data by Assumptions 1-7 (equations (2.3), (3.3) and (3.4)), satisfy the stochastic dominance assumption, induce the joint distribution on the data (Y, S, D, Z) and achieve $\delta = \mathbb{E}\left[\tilde{Y}_1^* - \tilde{Y}_0^* \middle| \tilde{S}_0 = 1, \tilde{S}_1 = 1, \tilde{V} = p\right]$.*

Remark 8. *The lower bound proposed here is strictly greater than the one proposed in Theorem 1 when $\alpha(p) \in (0,1)$ and the distribution of $Y_1^* | S_1 = 1, V$ is not degenerate. This improvement shows the identifying power of Assumption 7.*

Remark 9. *In the special case in which the treatment D is independent of the potential outcomes for Y^* and S, and $\alpha(p) = \alpha$ for any $p \in \mathcal{P} = [0,1]$, the MTE will be constant and equal to the average treatment effect for the always-observed and the bounds proposed in Theorem 1 simplify to the ones proposed in Imai (2008, Equation (8)).*

3.6 Point-identification under the Unit Mass Assumption

In this subsection we provide point-identification results for MTE^{OO} under Assumptions 1-5 and 8. The key difference from the last three subsections is that, under these assumptions, the distributions of $Y_0^* | S_1 = 1, V$ and $Y_1^* | S_1 = 1, V$ are not mixtures of the distributions of two latent groups. Now, both distributions are exclusively composed of always-observed individuals, that is $\alpha(p) = \beta(p) = 1$. Consequently, we have that

$$\mathbb{P}\left[Y_0^* \in A | S_0 = 1, S_1 = 1, V = p\right] = \frac{\mathbb{P}\left[Y_0^* \in A, S_0 = 1 | V = p\right]}{\mathbb{P}\left[S_0 = 1 | V = p\right]}$$
$$\mathbb{P}\left[Y_1^* \in A | S_0 = 1, S_1 = 1, V = p\right] = \frac{\mathbb{P}\left[Y_1^* \in A, S_1 = 1 | V = p\right]}{\mathbb{P}\left[S_0 = 1 | V = p\right]},$$

where the right-hand sides of both equations are point-identified according to equations (3.1), (3.2), (3.5) and (3.6). Consequently, the expectations $\mathbb{E}[Y_0^* | S_0 = 1, S_1 = 1, V = p]$ and $\mathbb{E}[Y_1^* | S_0 = 1, S_1 = 1, V = p]$ are also point-identified. Using this result, we can derive the following proposition:

Proposition 2. *Under Assumptions 1-5 and 8, the MTE is point-identified for the always-observed:*

$$MTE^{OO}(p) = \frac{\frac{\partial \mathbb{E}[YS|P(Z)=p]}{\partial p}}{\frac{\partial \mathbb{E}[SD|P(Z)=p]}{\partial p}}.$$

Remark 10. *If we additionally impose that all individuals in the population are always-observed, then the unconditional MTE is equal to MTE^{OO} and is point-identified. Moreover, the unconditional MTE is equal to the Heckman and Vytlacil (2005) estimand:*

$$\mathbb{E}[Y_1^* - Y_0^* | V = p] = \frac{\partial \mathbb{E}[Y | P(Z) = p]}{\partial p}.$$

An alternative to achieve point-identification of the unconditional MTE is to assume that potential sample selection status (S_0, S_1) is independent of the potential outcomes (Y_0^*, Y_1^*) given the unobserved characteristics V, i.e., $(S_0, S_1) \perp\!\!\!\perp (Y_0^*, Y_1^*) | V$. In this case, the distributions $\mathbb{P}\left[Y_1^* \leq y | V = p\right]$ and $\mathbb{P}\left[Y_0^* \leq y | V = p)\right]$ are point-identified since $\mathbb{P}\left[Y_d^* \leq y | V = p\right] =$

$\mathbb{P}\left[Y_d^* \leq y | S_d = 1, V = p\right]$ *for* $d \in \{0,1\}$, *which is identified from Equations* (3.1) *and* (3.2). *Consequently, the unconditional MTE is point-identified as follows:*

$$\mathbb{E}[Y_1^* - Y_0^* | V = p] = \frac{\frac{\partial \mathbb{E}[YSD|P(Z)=p]}{\partial p}}{\frac{\partial \mathbb{E}[SD|P(Z)=p]}{\partial p}} - \frac{\frac{\partial \mathbb{E}[YS(1-D)|P(Z)=p]}{\partial p}}{\frac{\partial \mathbb{E}[S(1-D)|P(Z)=p]}{\partial p}}.$$

Even though both assumptions are likely too strong for many applied contexts, they are stated here for completeness.

3.7 Empirical Relevance of Bounds for the MTE^{OO}

The novel partial identification results for MTE^{OO} presented in this section are relevant for a vast array of empirical objectives. First, bounds for the MTE^{OO} can illuminate the heterogeneity of the treatment effect, allowing researchers to better understand who would benefit from a specific treatment. This is important because common parameters (e.g., ATE, ATT, ATU and $LATE$ within the always-observed subpopulation) can be positive even when most people are adversely affected by a policy. Moreover, knowing, even partially, the MTE^{OO} function can be useful to optimally design policies that provide incentives to agents to take some treatment. Second, the MTE^{OO} bounds can be used to partially identify alternative treatment effect parameters that are described as a weighted integral of $MTE^{OO}(p)$ because

$$\int_{\mathcal{P}} \left(\underline{\Delta}_t(p) \cdot \omega(p)\right) \mathrm{d}p \leq \int_{\mathcal{P}} \left(MTE^{OO}(p) \cdot \omega(p)\right) \mathrm{d}p \leq \int_{\mathcal{P}} \left(\overline{\Delta}_t(p) \cdot \omega(p)\right) \mathrm{d}p,$$

where $t \in \{1,2,3\}$, $\underline{\Delta}_t$ and $\overline{\Delta}_t$ are described in Proposition 1, Theorem 1 or Theorem 2, and $\omega(\cdot)$ is a known or identifiable weighting function. Moreover, those bounds are sharp as summarized in the following proposition:

Proposition 3. *Let* $\omega \colon \mathcal{P} \to \mathbb{R}$ *be a known or identifiable weighting function and define the treatment effect parameter* $TE := \int_{\mathcal{P}} MTE^{OO}(p) \cdot \omega(p)\ dp$.

Under Assumptions 1-5, the treatment effect parameter TE is partially identified:

$$\int_{\mathcal{P}} \left(\underline{\Delta}_1(p) \cdot \omega(p)\right) \, dp \leq TE \leq \int_{\mathcal{P}} \left(\overline{\Delta}_1(p) \cdot \omega(p)\right) \, dp.$$

Under Assumptions 1-6, the treatment effect parameter TE is partially identified:

$$\int_{\mathcal{P}} \left(\underline{\Delta}_2(p) \cdot \omega(p)\right) \, dp \leq TE \leq \int_{\mathcal{P}} \left(\overline{\Delta}_2(p) \cdot \omega(p)\right) \, dp.$$

Under Assumptions 1-7, the treatment effect parameter TE is partially identified:

$$\int_{\mathcal{P}} \left(\underline{\Delta}_3(p) \cdot \omega(p)\right) \, dp \leq TE \leq \int_{\mathcal{P}} \left(\overline{\Delta}_3(p) \cdot \omega(p)\right) \, dp.$$

Moreover, these bounds are sharp according to the definition given by Canay and Shaikh (2017, Remark 2.1).

Tables 2 and 3 show some of the treatment effect parameters that can be partially identified using Proposition 3. More examples are given by Heckman, Urzua, and Vytlacil (2006, Tables 1A and 1B) and Mogstad, Santos, and Torgovitsky (2018, Table 1).

Table 2: Treatment Effects as Weighted Integrals of the MTE

$ATE^{OO} = \mathbb{E}\left[Y_1^* - Y_0^* \,\middle|\, S_0 = 1, S_1 = 1\right] = \int_0^1 MTE^{OO}(p) \, \mathrm{d}p$
$ATT^{OO} = \mathbb{E}\left[Y_1^* - Y_0^* \,\middle|\, D = 1, S_0 = 1, S_1 = 1\right] = \int_0^1 MTE^{OO}(p) \cdot \omega_{ATT}(p) \, \mathrm{d}p$
$ATU^{OO} = \mathbb{E}\left[Y_1^* - Y_0^* \,\middle|\, D = 0, S_0 = 1, S_1 = 1\right] = \int_0^1 MTE^{OO}(p) \cdot \omega_{ATU}(p) \, \mathrm{d}p$
$LATE^{OO}(\underline{p}, \overline{p}) = \mathbb{E}\left[Y_1^* - Y_0^* \,\middle|\, V \in \left[\underline{p}, \overline{p}\right], S_0 = 1, S_1 = 1\right] = \int_0^1 MTE^{OO}(p) \cdot \omega_{LATE}(p) \, \mathrm{d}p$

Source: Heckman, Urzua, and Vytlacil (2006) and Mogstad, Santos, and Torgovitsky (2018).

Table 3: Weights

$\omega_{ATT}(p) = \dfrac{\int_p^1 f_{P(Z)}(u)\,\mathrm{d}u}{\mathbb{E}[P(Z)]}$
$\omega_{ATU}(p) == \dfrac{\int_0^p f_{P(Z)}(u)\,\mathrm{d}u}{1-\mathbb{E}[P(Z)]}$
$\omega_{LATE}(p) = \dfrac{\mathbf{1}\{p \in [\underline{p}, \overline{p}]\}}{\overline{p}-\underline{p}}$

Source: Heckman, Urzua, and Vytlacil (2006) and Mogstad, Santos, and Torgovitsky (2018).

4 Numerical Illustration

In this section, we highlight the feasibility and usefulness of the bounds proposed in Section 3 by considering a numerical illustration of a simple structural model with endogenous treatment and sample selection. We present the bounds for MTE^{OO} based on Proposition 1 and on Theorem 1. The illustration provides insights on the functioning of the bounds as well as the mechanisms driving how informative those bounds are. Consider the following data generating process (DGP):

$$\begin{cases} Y &= Y^*S \\ Y^* &= Y_1^*D + Y_0^*(1-D) \\ S &= \mathbb{1}\{U_S \leq \delta_0 + \delta_1 D\} \\ D &= \mathbb{1}\{V \leq \Phi(Z)\} \end{cases} \tag{4.1}$$

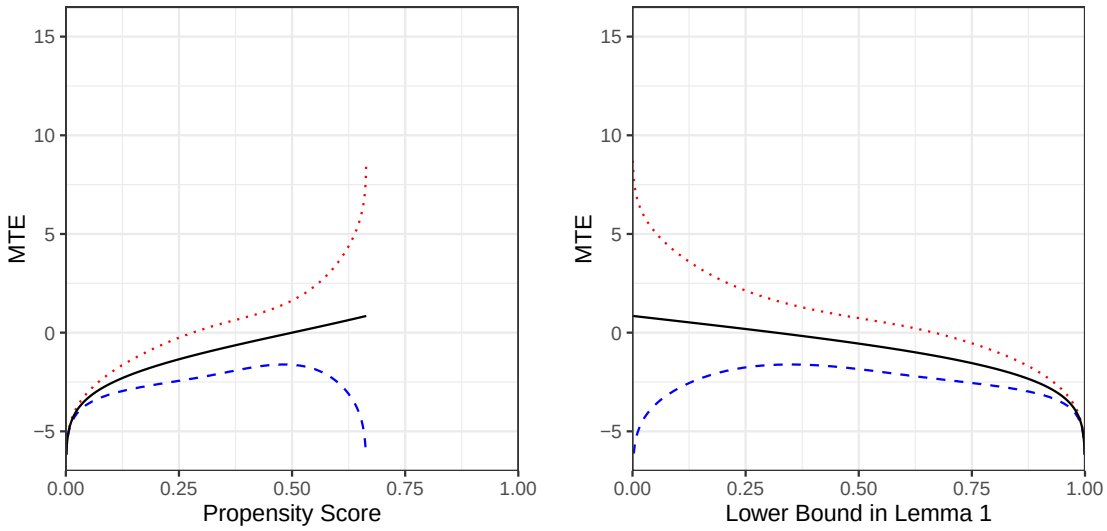
We set

$$\begin{cases} V &= \Phi(\theta) \\ U_S &= \frac{1}{\sqrt{2}}(\theta + \epsilon_S) \\ Y_0^* &= T \cdot \beta_{0,1}\theta + \gamma_0 + (1-T)\cdot(-\beta_{0,0}\theta - \gamma_0) \\ Y_1^* &= T \cdot \beta_{1,1}\theta + \gamma_1 + (1-T)\cdot(-\beta_{1,0}\theta - \gamma_1) \end{cases} \tag{4.2}$$

where $(\theta, \epsilon_S, \gamma_0, \gamma_1, Z, \xi)' \sim N(0, I)$, $T = \mathbb{1}\{\xi \geq 0\}$, I is the identity matrix, $\Phi(.)$ is the standard normal CDF, and $\Phi^{-1}(\cdot)$ its inverse. The potential outcomes equations has random

coefficients in this illustration, which can be intuitively understood as having two sets of individuals that might face different returns to treatment due to, for example, their gender or race. From a technical perspective, this choice guarantees reasonable overlap for treated and untreated groups in the observed population over the support of the the outcome conditional on V.

We present the bounds for the MTE^{OO} described in Proposition 1 $(\underline{\Delta}_1, \overline{\Delta}_1)$ in Figure 1, and Theorem 1 $(\underline{\Delta}_2, \overline{\Delta}_2)$ in Figure 2 for parameters $\delta_0 = 0.1$, $\delta_1 = 0.4$, $\beta_{0,0} = \beta_{0,1} = \beta_{1,0} = 1$ and $\beta_{1,1} = 5$.[18]



(a) Bounds as a function of the propensity score

(b) Bounds as a function of $\mathbb{P}[S_0 = 1|V = p] + \mathbb{P}[S_1 = 1|V = p] - 1$

Notes: The solid lines are the true values of the MTE^{OO}. The red dotted lines and the blue dashed lines are, respectively, the values of the upper and lower bounds around the MTE^{OO} computed by numerical integration using 100,000 simulated points for each value of the propensity score.

Figure 1: Numerical Bounds based on Proposition 1

As can be seen on Subfigure 1(a), the bounds based on Proposition 1 are not very informative for a large part of the support of V. In this DGP, when the propensity score is small (p is close to zero), v^ℓ — the lower bound on the proportion of the always-observed $(\mathbb{P}[S_0 = 1|V = p] + \mathbb{P}[S_1 = 1|V = p] - 1)$ — approaches 1 and the MTE^{OO} is almost point-identified as the bounds are close to each other. On the other hand, when p is larger than

[18] Note that these bounds can be computed using numeric integration. See Appendix C for more details.

0.664, the lower bound on the proportion of the always-observed becomes exactly zero, the MTE^{OO} is not identified and the bounds diverge. Nevertheless, the sign of MTE^{OO} is identified for propensity scores smaller than 0.28.

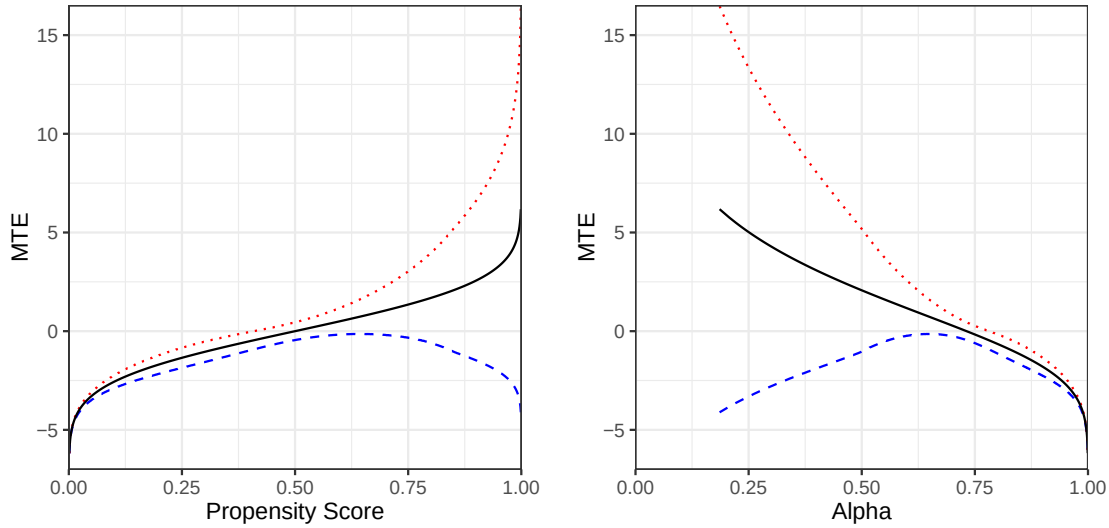
Subfigure 1(b) plots the identified interval for MTE^{OO} against υ^{ℓ} on the horizontal axis, emphasizing the important role of the lower bound on the proportion of the always-observed. As υ^{ℓ} increases, the observed expectation of the outcomes conditional on $V = p$ is fully composed by the always-observed group, leading to point identification of the MTE^{OO}. Moreover, they also illustrate that we can only non-trivially bound the MTE^{OO} when the lower bound on the proportion of the always-observed is strictly positive.

Figure 2 plots the MTE^{OO} and the its bounds based on Theorem 1, i.e., under the monotonicity assumption. The bounds presented on Subfigure 2(a) are in general informative. Similarly to the discussion above, when the propensity score is small (p is close to zero), the proportion of the always-observed ($\alpha(p)$) approaches 1 and the MTE^{OO} is almost point-identified. When p is close to 1, the proportion of the always-observed decreases and the identified set around the MTE^{OO} increases. Moreover, The sign of the MTE^{OO} is identified for $p < 0.409$, illustrating that Assumption 6 allow us to identify the sign of the MTE^{OO} in more cases than in Figure 1.

Subfigure 2(b) plots the same curves with $\alpha(p)$ on the horizontal axis, emphasizing the importance of the trimming proportion. As $\alpha(p) \to 1$, the observed expectation of the outcomes conditional on $V = p$ is fully composed by the always-observed group, leading to point identification of the MTE^{OO}. Moreover, they also illustrate that, under Assumption 6, $\alpha(p)$ never reaches zero, allowing us to non-trivially bound the MTE^{OO} for all values of the propensity score.

Figure 3 is a zoomed version of Subfigures 1(a) and 2(a). Note that the bounds based on Theorem 1 are much tighter than the bounds based on Proposition 1, especially for larger values of p. This is expected as the difference in trimming proportions in Proposition 1 and Theorem 1 increases with the propensity score for this DGP.[19]

[19] Assumption 7 holds with equality in this DGP, implying that the lower bound is equal to the true MTE^{OO} for the case considered in Theorem 2.

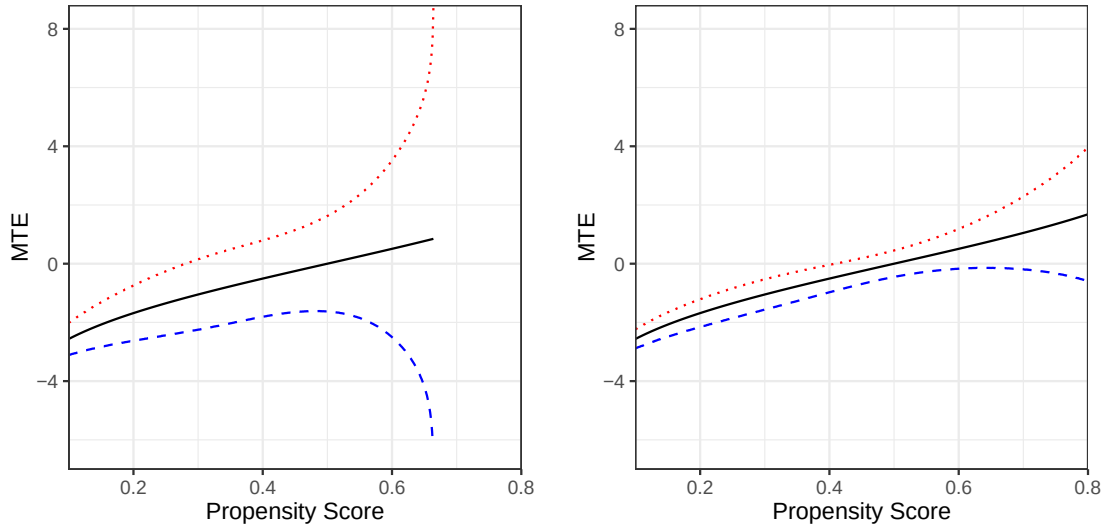


(a) Bounds as a function of the propensity score

(b) Bounds as a function of $\alpha\left(p\right)$

Notes: The solid lines are the true values of the MTE^{OO}. The red dotted lines and the blue dashed lines are, respectively, the values of the upper and lower bounds around the MTE^{OO} computed by numerical integration using 100,000 simulated points for each value of the propensity score.

Figure 2: Numerical Bounds based on Theorem 1



(a) Proposition 1

(b) Theorem 1

Notes: The solid lines are the true values of the MTE^{OO}. The red dotted lines and the blue dashed lines are, respectively, the values of the upper and lower bounds around the MTE^{OO} computed by numerical integration using 100,000 simulated points for each value of the propensity score.

Figure 3: Comparing Proposition 1 and Theorem 1

5 Estimation

This section presents an estimator for the bounds proposed in Theorem 1. For brevity, we focus on the bounds identified under the monotonicity of treatment on sample selection case (Assumptions 1-6), as it is the most relevant (and feasible) case empirically. Estimators for the bounds proposed in Proposition 1, Theorem 2 and Proposition 2 are natural extensions of the estimator discussed here. Subsections 5.1-5.4 present the estimation procedure. Appendix D presents a Monte Carlo Simulation that evaluates the small sample properties of our estimator.

In order to estimate the bounds in Theorem 1, we need to obtain the cumulative distribution functions:

$$\tilde{Y}_d|S=1,D=d,P(Z)=p \ \sim F_{\tilde{Y}_d|S=1,D=d,P(Z)=p}(y) = \frac{\frac{\partial \mathbb{P}[Y\leq y,S=1,D=d|P(Z)=p]}{\partial p}}{\frac{\partial \mathbb{P}[S=1,D=d|P(Z)=p]}{\partial p}}$$

for any $d \in \{0,1\}$ and

$$\alpha(p) = -\frac{\frac{\partial \mathbb{P}[S=1,D=0|P(Z)=p]}{\partial p}}{\frac{\partial \mathbb{P}[S=1,D=1|P(Z)=p]}{\partial p}}.$$

Consequently, we need to estimate:

$$\Gamma_1(p,y) \coloneqq \frac{\partial \mathbb{P}[Y\leq y,S=1,D=1|P(Z)=p]}{\partial p}, \quad \pi_1(p) \coloneqq \frac{\partial \mathbb{P}[S=1,D=1|P(Z)=p]}{\partial p},$$
$$\Gamma_0(p,y) \coloneqq -\frac{\partial \mathbb{P}[Y\leq y,S=1,D=0|P(Z)=p]}{\partial p}, \quad \pi_0(p) \coloneqq -\frac{\partial \mathbb{P}[S=1,D=0|P(Z)=p]}{\partial p}.$$

Furthermore, the estimation of the propensity score $P(Z)$ is necessary to obtain the moments of the conditional distribution of the observed outcome.

5.1 Estimating the Propensity Score $P(Z)$

The procedures proposed by Carneiro and Lee (2009) to estimate $P(z) = \mathbb{P}(D=1|Z=z)$ apply directly to our case, since treatment status D is observed for all individuals, and are summarized here. We model the probability as a partially linear additive regression model to improve precision of the estimates while avoiding the curse of dimensionality:

$\mathbb{P}[D=1|Z=z] = z^{pc}\vartheta + \sum_{j=1}^{d}\varphi_j(z_j^c)$, where z is composed by nonparametric (z^c) and parametric (z^{pc}) components, z^c is a continuous random vector of dimension d, ϑ is a vector of unknown parameters and $\varphi_j(\cdot)$ are unknown functions.

Let $\{p_\kappa : \kappa = 1, 2...\}$ be the basis for smooth functions that we will use to approximate $\varphi_j(\cdot)$ more closely as the number of approximating functions increases. For a given $\kappa > 0$, define $P_\kappa(z) = [z^{pc}, p_1(z_1^c), \ldots, p_\kappa(z_1^c), \ldots, p_1(z_d^c), \ldots, p_\kappa(z_d^c)]'$. Then, using Carneiro and Lee (2009) notation, we have that $\tilde{\mathbb{P}}(Z_i) = P_\kappa(Z_i)'\hat{\theta}_\kappa$, where $\hat{\theta}_\kappa = [\sum_{i=1}^n P_\kappa(Z_i)P_\kappa(Z_i)']^{-1}[\sum_{i=1}^n P_\kappa(Z_i)D_i]$. As discussed in Carneiro and Lee (2009), the estimated probabilities might fall outside of the $[0,1]$ interval in finite samples. As a consequence, it is preferable to use the trimmed version, $\hat{P}_i \equiv \hat{\mathbb{P}}(Z_i) = \tilde{\mathbb{P}}(Z_i) + (1-\lambda-\tilde{\mathbb{P}}(Z_i))\mathbb{1}(\tilde{\mathbb{P}}(Z_i) > 1) + (\lambda - \tilde{\mathbb{P}}(Z_i))\mathbb{1}(\tilde{\mathbb{P}}(Z_i) < 0)$, for a suitably small positive λ. Alternatively, a typical conditional probability estimator of $P(\cdot)$ based on a logit or probit model could be used, so that the fitted probability always lies between 0 and 1.

5.2 Estimating $\pi_1(p)$, $\pi_0(p)$ and $\alpha(p)$

In order to estimate $\pi_1(p)$ and $\pi_0(p)$, we consider the local polynomial estimators (Fan and Gijbels, 1996):

$$\hat{\pi}_1(p) \coloneqq e_2 \underset{c_0,c_1,c_2}{\operatorname{argmin}} \sum_{i=1}^{n} \left[S_i D_i - c_0 - c_1(\hat{P}_i - p) - c_2(\hat{P}_i - p)^2\right]^2 K\left(\frac{\hat{P}_i - p}{h}\right),$$
$$\hat{\pi}_0(p) \coloneqq -e_2 \underset{c_0,c_1,c_2}{\operatorname{argmin}} \sum_{i=1}^{n} \left[S_i(1-D_i) - c_0 - c_1(\hat{P}_i - p) - c_2(\hat{P}_i - p)^2\right]^2 K\left(\frac{\hat{P}_i - p}{h}\right)$$

where e_g is a conformable row vector of zeros with g-th element equal to one, $K(\cdot)$ is a kernel function and h is a bandwidth. Recent developments in Calonico, Cattaneo, and Farrell (2018) for nonparametric inference establish higher-order improvements that can be obtained by utilizing a robust bias-corrected (RBC) procedure for the estimation of those terms, which we adopt. Furthermore, Calonico, Cattaneo, and Farrell (2019a) develop optimal bandwidth and kernel selection for optimal coverage error and interval lenght for such RBC methods. Hence, we implement the RBC version of these estimators with optimal bandwidth choice as

conveniently implemented in the software *R*, using the package *nprobust*. [20]

To estimate $\alpha(p)$, we simply take its sample analog: $\hat{\alpha}(p) \coloneqq \dfrac{\hat{\pi}_0(p)}{\hat{\pi}_1(p)}$.

5.3 Estimating $\Gamma_1(p,y)$ and $\Gamma_0(p,y)$

In order to estimate $\Gamma_1(p,y)$ and $\Gamma_0(p,y)$, we choose a grid for the outcome variable ($\{y_1,\ldots,y_{K_n}\}$) and estimate the conditional density of the outcome for each bin in the grid, leading to the following local polynomial regression (Fan and Gijbels, 1996),

$$\hat{\gamma}_1(p,k) \coloneqq e_2 \underset{c_0,c_1,c_2}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ \left[\mathbb{1}\{y_{k-1} \leq Y_i \leq y_k\} S_i D_i - c_0 - c_1(\hat{P}_i - p) - c_2(\hat{P}_i - p)^2 \right]^2 \cdot K\left(\frac{\hat{P}_i - p}{h}\right) \right\},$$

$$\hat{\gamma}_0(p,k) \coloneqq -e_2 \underset{c_0,c_1,c_2}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ \left[\mathbb{1}\{y_{k-1} \leq Y_i \leq y_k\} S_i (1-D_i) - c_0 - c_1(\hat{P}_i - p) - c_2(\hat{P}_i - p)^2 \right]^2 \cdot K\left(\frac{\hat{P}_i - p}{h}\right) \right\}$$

for any $k \in \{2,\ldots,K_N\}$. Once more, we use the RBC procedure and the optimal bandwidth selection mechanism proposed by Calonico, Cattaneo, and Farrell (2018) and Calonico, Cattaneo, and Farrell (2019a).

Natural estimators for $\Gamma_1(p,y)$ and $\Gamma_0(p,y)$ are given by $\hat{\Gamma}_1(p,y_k) \coloneqq \sum_{j=2}^{k} \hat{\gamma}_1(p,j)$ and $\hat{\Gamma}_0(p,y_k) \coloneqq \sum_{j=2}^{k} \hat{\gamma}_0(p,j)$ for any $k \in \{2,\ldots,K_N\}$.

The estimation of $\hat{\gamma}_d(p,k)$ is a crucial step and can be adversely affected by several features of the population DGP and the available data. For example, these estimators will perform well in situations for which the available data about the observed outcome covers the whole range of possible values of Y for the values of p being considered and for both treated and untreated individuals. One can mitigate the challenges to feasibility of the estimator by choosing wider bins $[y_{k-1} \leq Y_i \leq y_k]$, at the cost of obtaining a coarse description of the distribution of $Y_d^*|V=p$. That can be particularly harmful in the region around the trimming points, and should be considered carefully.

[20] See Calonico, Cattaneo, and Farrell (2019b) for details.

5.4 Estimating $MTE^{OO}(p)$ Bounds

The estimators for the bounds $LB_2(p)$ and $UB_2(p)$ can be obtained as

$$\widehat{LB}_1(p) \coloneqq \sum_{k=2}^{K_N} \overline{y}_k \cdot \mathbb{1}\left\{\hat{\Gamma}_1(p, y_k) \leq \hat{\alpha}(p)\right\} \cdot \frac{\hat{\gamma}_1(p,k)}{\hat{\alpha}(p)} \tag{5.1}$$

$$\widehat{UB}_1(p) \coloneqq \sum_{k=2}^{K_N} \overline{y}_k \cdot \mathbb{1}\left\{1 - \hat{\Gamma}_1(p, y_k) < \hat{\alpha}(p)\right\} \cdot \frac{\hat{\gamma}_1(p,k)}{\hat{\alpha}(p)}, \tag{5.2}$$

where $\overline{y}_k$ is the center point of each bin $[y_{k-1}, y_k]$ for any $k \in \{2, \ldots, K_N\}$. Moreover, we can estimate $\mathbb{E}\left[\tilde{Y}_0 | S=1, D=0, P(Z)=p\right]$ in Theorem 1 using $\hat{\Xi}_{OO,0}(p) \coloneqq \sum_{k=2}^{K_N} \overline{y}_k \cdot \hat{\gamma}_1(p,k)$.

Note that the estimators for $\widehat{LB}_1(p), \widehat{UB}_1(p)$ and $\hat{\Xi}_{OO,0}(p)$ rely on the estimates for densities and trimming points obtained in previous steps, not relying on the original data for the observed outcomes.

Naturally, the estimated MTE^{OO} bounds can, then, be obtained by $\hat{\underline{\Delta}}_2(p) \coloneqq \widehat{LB}_1(p) - \hat{\Xi}_{OO,0}(p)$ and $\hat{\overline{\Delta}}_2(p) \coloneqq \widehat{UB}_1(p) - \hat{\Xi}_{OO,0}(p)$.

Analyzing the inference procedures and asymptotic properties of the proposed estimators is beyond the scope of this paper and an exciting area for future work.

We summarize the estimation procedure in the following steps:

STEP 1. Estimate $\mathbb{P}(D=1|Z=z)$ and obtain $\hat{P}_i = \hat{\mathbb{P}}(Z_i)$ for all observations.

STEP 2. Estimate $\hat{\gamma}_0(p, y_k)$, $\hat{\gamma}_1(p, y_k)$, $\hat{\pi}_0(p)$ and $\hat{\pi}_1(p)$ for the value p of interest.

STEP 3. Estimate $\hat{\alpha}(p)$ for p.

STEP 4. Implement $\widehat{LB}_1(p)$, $\widehat{UB}_1(p)$ and $\hat{\Xi}_{OO,0}(p)$.

STEP 5. Calculate the bounds for $MTE^{OO}(p)$ using $\hat{\underline{\Delta}}_2(p)$ and $\hat{\overline{\Delta}}_2(p)$.

6 Extensions

We provide two extensions to our main results. In Subsection 6.1, we extend Theorem 1 to bound the distributional marginal treatment effect for the always-observed sub-population.

Subsection 6.2 extends Theorem 1 to the case with multi-valued discrete instruments, implying that we identify many LATE parameters for the always-observed subpopulation. In both subsections, we focus on the case under the monotonicity restriction (Assumption 1-6) for brevity. Similar results hold under our other identifying sets of assumptions.

6.1 Bounds for the distributional marginal treatment effect (DMTE)

In this section, we derive sharp bounds on the distributional marginal treatment effect defined as

$$DMTE^{OO}(A;p) \coloneqq \mathbb{P}\left[Y_1^* \in A | S_0 = 1, S_1 = 1, V = p\right] - \mathbb{P}\left[Y_0^* \in A | S_0 = 1, S_1 = 1, V = p\right].$$

Carneiro and Lee (2009) show point identification results for $P\left[Y_1^* \in A | V = p\right] - \mathbb{P}\left[Y_0^* \in A | V = p\right]$ when there is no sample selection. However, in the presence of sample selection, the DMTE is only partially identified.

Combining Equation (3.7) and Corollary 1.2 by Horowitz and Manski (1995), we obtain functionally sharp bounds on the $DMTE^{OO}(A;p)$.

Proposition 4. *Under Assumptions 1-6, sharp bounds on the DMTE are given by:*

$$\begin{aligned}
&\max\left\{0, \frac{\mathbb{P}\left[Y_1^* \in A | S_1 = 1, V = p\right] - (1 - \alpha(p))}{\alpha(p)}\right\} - \mathbb{P}\left[Y_0^* \in A | S_0 = 1, S_1 = 1, V = p\right] \\
&\quad \leq DMTE^{OO}(A;p) \leq \\
&\min\left\{1, \frac{\mathbb{P}\left[Y_1^* \in A | S_1 = 1, V = p\right]}{\alpha(p)}\right\} - \mathbb{P}\left[Y_0^* \in A | S_0 = 1, S_1 = 1, V = p\right].
\end{aligned}$$

6.2 Identification with discrete instruments

In many applications, the only instruments available are discrete, e.g, treatment eligibility, number of children in the household, quarter of birth. In this section, we provide additional identification results when the instrument is multi-valued discrete, implying the the support of the propensity score is finite. The results here can be seen as an extension of the work developed by Chen and Flores (2015), who analyze the case for binary instrument.

Assumption 9. *The instrument Z is discrete with support $\{z_1, z_2, \ldots, z_K\}$ and the propensity score $p_\ell \equiv \mathbb{P}[D = 1|Z = z_\ell]$ satisfies $0 < p_1 < p_2 < \ldots < p_K < 1$.*

Assumption 9 requires that one can rank the probabilities of receiving treatment for the points of $P(Z)$ that are available, allowing the researcher to partition the $[0, 1]$ interval into regions $[p_\ell - p_{\ell-1}]$ for $\ell = 2, ..., K$. The researcher will only be able to identify an average of the MTE within each region, i.e., a LATE. Naturally, if the instrument has more points of positive mass (providing finer partitions of the probabilities), we are able to obtain averages of the MTE for more specific ranges of the unobservable characteristic V.[21]

Remark 11. *Multivalued qualitative instruments, e.g., profession, location, race, can be used by sorting the probabilities associated with each category in ascending order for analysis.*

The identification argument is similar to the one presented for the continuous instrument case in Subsection 3.4.

Under Assumption 1, we have $p_\ell = \mathbb{P}[V \leq P(z_\ell)]$ and $\mathbb{P}[P(z_{\ell-1}) < V \leq P(z_\ell)] = p_\ell - p_{\ell-1}$. If Assumptions 1 and 5 hold, then $P(z_\ell) = p_\ell$. To ease the exposition, we use the shorthand $P \coloneqq P(Z)$.

We have $\mathbb{P}[Y \in A, S = 1, D = 1|P = p_\ell] = \mathbb{P}[Y_1^* \in A, S_1 = 1, V \leq p_\ell]$. Therefore,

$$\begin{aligned}\mathbb{P}[Y_1^* \in A, S_1 = 1, p_{\ell-1} < V \leq p_\ell] \quad &= \quad \mathbb{P}[Y \in A, S = 1, D = 1|P = p_\ell] \\ &\quad -\mathbb{P}[Y \in A, S = 1, D = 1|P = p_{\ell-1}],\end{aligned}$$

which implies that

$$\mathbb{P}[Y_1^* \in A, S_1 = 1|p_{\ell-1} < V \leq p_\ell] = \frac{\mathbb{P}[Y \in A, S = 1, D = 1|P = p_\ell] - \mathbb{P}[Y \in A, S = 1, D = 1|P = p_{\ell-1}]}{p_\ell - p_{\ell-1}}.$$

[21] If for some values of Z, the probabilities are the same, $p_\ell = p_{\ell-1}$, we cannot refine the partition of the unit interval describing the probabilities and, hence, cannot improve on the detail level of the MTE identified.

Similarly, we have

$$\mathbb{P}\left[Y_0^* \in A, S_0 = 1 | p_{\ell-1} < V \leq p_\ell\right] = \\ -\frac{\mathbb{P}\left[Y \in A, S = 1, D = 0 | P = p_\ell\right] - \mathbb{P}\left[Y \in A, S = 1, D = 0 | P = p_{\ell-1}\right]}{p_\ell - p_{\ell-1}}.$$

Thus for $A = \mathcal{Y}$, we can write

$$\begin{aligned}
\mathbb{P}\left[S_1 = 1 | p_{\ell-1} < V \leq p_\ell\right] &= \frac{\mathbb{P}\left[S = 1, D = 1 | P = p_\ell\right] - \mathbb{P}\left[S = 1, D = 1 | P = p_{\ell-1}\right]}{p_\ell - p_{\ell-1}}, \\
\mathbb{P}\left[S_0 = 1 | p_{\ell-1} < V \leq p_\ell\right] &= -\frac{\mathbb{P}\left[S = 1, D = 0 | P = p_\ell\right] - \mathbb{P}\left[S = 1, D = 0 | P = p_{\ell-1}\right]}{p_\ell - p_{\ell-1}}.
\end{aligned}$$

We know that $\mathbb{P}\left[Y_d^* \in A | S_d = 1, p_{\ell-1} < V \leq p_\ell\right] = \frac{\mathbb{P}\left[Y_d^* \in A, S_d = 1 | p_{\ell-1} < V \leq p_\ell\right]}{\mathbb{P}\left[S_d = 1 | p_{\ell-1} < V \leq p_\ell\right]}$ for $d \in \{0, 1\}$. Under Assumption 6, we identify $\mathbb{P}\left[S_0 = 1, S_1 = 1 | p_{\ell-1} < V \leq p_\ell\right]$ as $\mathbb{P}\left[S_0 = 1 | p_{\ell-1} < V \leq p_\ell\right]$.

To implement the trimming in this setting, we define the discrete case analog of $\alpha(p)$, denoted by $\tilde{\alpha}(p_{\ell-1}, p_\ell)$,

$$\tilde{\alpha}(p_{\ell-1}, p_\ell) \coloneqq \frac{\mathbb{P}\left[S_0 = 1, S_1 = 1 | p_{\ell-1} < V \leq p_\ell\right]}{\mathbb{P}\left[S_1 = 1 | p_{\ell-1} < V \leq p_\ell\right]} = \frac{\mathbb{E}[S(1-D) | P = p_{\ell-1}] - \mathbb{E}[S(1-D) | P = p_\ell]}{\mathbb{E}[SD | P = p_\ell] - \mathbb{E}[SD | P = p_{\ell-1}]}.$$

Using the same steps as in Subsection 3.4, we derive bounds on $\mathbb{E}[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1, p_{\ell-1} < V \leq p_\ell]$.

Proposition 5. *Under Assumptions 1-6 and 9, the following holds: for $\ell = 2, \ldots, K$*

$$\begin{aligned}
&\mathbb{E}\left[Y_1^* | S_1 = 1, p_{\ell-1} < V \leq p_\ell, Y_1^* \leq F^{-1}_{Y_1^* | S_1 = 1, p_{\ell-1} < V \leq p_\ell}\left(\tilde{\alpha}(p_{\ell-1}, p_\ell)\right)\right] - \mathbb{E}[Y_0^* | S_0 = 1, S_1 = 1, p_{\ell-1} < V \leq p_\ell] \\
&\qquad \leq \mathbb{E}[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1, p_{\ell-1} < V \leq p_\ell] \leq \\
&\mathbb{E}\left[Y_1^* | S_1 = 1, p_{\ell-1} < V \leq p_\ell, Y_1^* > F^{-1}_{Y_1^* | S_1 = 1, p_{\ell-1} < V \leq p_\ell}\left(1 - \tilde{\alpha}(p_{\ell-1}, p_\ell)\right)\right] - \mathbb{E}[Y_0^* | S_0 = 1, S_1 = 1, p_{\ell-1} < V \leq p_\ell].
\end{aligned}$$

These bounds are sharp.

As mentioned above, the quantity for which we derive bounds in Proposition 5 is a average of the $MTE(p)$ evaluated at levels of p in the interval $(p_{\ell-1}, p_\ell]$, i.e., we partially identify a LATE. More can be said about the MTE if additional assumptions are made. For example, if we assume that the MTE is flat within each interval, then Theorem 5 provides sharp bounds

on the MTE for the always-observed. Note that Brinch, Mogstad, and Wiswall (2017) showed how a discrete instrument can be used to identify the marginal treatment effects under some functional structure in the absence of sample selection. An extension of their results to the current framework is an interesting question for future research.

7 Conclusion

This paper derives sharp bounds for the marginal treatment effect for the always-observed individuals when there is sample selection. We achieve partial identification results under four increasingly restrictive sets of assumptions. First, we impose standard MTE assumptions without any restrictions to the sample selection mechanism. The second case, which is the main result of this work, imposes monotonicity of the sample selection variable with respect to the treatment, considerably shrinking the identified set. Then, we consider a strong stochastic dominance assumption which tightens the lower bound for the MTE. Finally, we provide a set of conditions that allows point identification for completeness. All the results rely on the insight that the treated individuals observed in the sample will be of two possible groups, the ones that would always be observed regardless of treatment status and the ones that would self-select into the sample only when treated. Hence, we can rewrite the distribution of the observed population as a mixture of these groups, and the mixture weights can be identified. This leads to a trimming procedure that partially identifies the target parameter, extending Imai (2008), Lee (2009) and Chen and Flores (2015) to the context of MTE. Moreover, we derive testable implications of our identifying assumptions. We present a numerical example that the bounds can be informative in relevant settings. A feasible nonparametric estimator is proposed and simulation evidence of its performance in estimating the bounds for the parameter of interest is presented.

References

Ahn, Hyungtaik and James L. Powell. 1993. "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism." *Journal of Econometrics* 58:pp.

3–29.

Altonji, Joseph. 1993. "The Demand for and Return to Education When Education Outcomes are Uncertain." *Journal of Labor Economics* 11 (1):pp. 48–83.

Angelucci, Manuela, Dean Karlan, and Jonathan Zinman. 2015. "Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Comportamos Banco." *American Economic Journal: Applied Economics* 7 (1):pp. 151–182.

Angrist, Joshua, Eric Bettinger, and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *The American Economic Review* 96 (3):847–862. URL http://www.jstor.org/stable/30034075.

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2009. "Incentives and Services for College Achievement: Evidence from a Randomized Trial." *American Economic Journal: Applied Economics* 1 (1):pp. 1–28.

Balke, Alexander and Judea Pearl. 1997. "Bounds on Treatment Effects from Studies with Imperfect Compliance." *Journal of the American Statistical Association* 92 (439):1171–1176.

Behaghel, Luc, Bruno Crepon, Marc Gurgand, and Thomas Le Barbanchon. 2015. "Please Call Again: Correcting Nonresponse Bias in Treatment Effect Models." *The Review of Economics and Statistics* 97 (5):pp. 1070–1080.

Bhattacharya, Jay, Azeem M. Shaikh, and Edward Vytlacil. 2008. "Treatment Effect Bounds under Monotonicity Assumptions: An Application to Swan-Ganz Catheterization." *The American Economic Review: Papers and Proceedings* 98 (2):pp. 351–356.

———. 2012. "Treatment Effect Bounds: An Application to Swan-Ganz Catheterization." *Journal of Econometrics* 168 (2):pp. 223–243.

Bhuller, Manudeep, Gordon B. Dahl, Katrine V. Loken, and Magne Mogstad. 2019. "Incaceration, Recidivism, and Employment." Available at http://bit.ly/2XbGVOd.

Bjorklund, Anders and Robert Moffitt. 1987. "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models." *The Review of Economics and Statistics* 69 (1):42–49.

Blanco, German, Xuan Chen, Carlos A. Flores, and Alfonso Flores-Lagunes. 2017. "Bounding Average and Quantile Effects of Training on Employment and Unemployment Durations under Selection, Censoring, and Noncompliance." Available at: http://conference.iza.org/conference_files/EVAL_2017/blanco_g7367.pdf.

Blanco, German, Carlos A. Flores, and Alfonso Flores-Lagunes. 2013. "Bounds on Average and Quantile Treatment Effects of Job Corps Training on Wages." *Journal of Human Resources* 48 (3):pp. 659–701.

Blundell, Richard, Amanda Gosling, Hidehiko Ichimura, and Costas Meghir. 2007. "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds." *Econometrica* 75 (2):pp. 323–363.

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall. 2017. "Beyond LATE with a Discrete Instrument." *Journal of Political Economy* 125 (4):pp. 985–1039.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. 2018. "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference." *Journal of the American Statistical Association* 113 (522):767–779. URL https://doi.org/10.1080/01621459.2017.1285776.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. 2019a. "Coverage Error Optimal Confidence Intervals for Local Polynomial Regression." *arXiv e-prints* :arXiv:1808.01398.

———. 2019b. "nprobust: Nonparametric Kernel-Based Estimation and Robust Bias-Corrected Inference." *arXiv e-prints* :arXiv:1906.00198.

Canay, Ivan A. and Azeem M. Shaikh. 2017. "Practical and Theoretical Advances for Inference in Partially Identified Models." In *Advances in Economics and Econometrics*, *Econometric Society Monographs*, vol. 11th World Congress, edited by Bo Honore, Ariel Pakes, Monika Piazzesi, and Larry Samuelson. pp. 271–306.

Card, David. 1999. "The Causal Effect of Education on Earnings." In *Handbook of Labor Economics*, vol. 3A, edited by Orley Ashenfelter and David Card. Elsevier, pp. 1801–1863.

Carneiro, Pedro, James J. Heckman, and Edward J Vytlacil. 2011. "Estimating marginal returns to education." *American Economic Review* 101 (6):2754–81.

Carneiro, Pedro and Sokbae Lee. 2009. "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality." *Journal of Econometrics* 149 (2):191–208.

CASS. 1984. "Myocardial Infarction and Mortality in the Coronary Artery Surgery Study (CASS) Randomized Trial." *The New England Journal of Medicine* 310 (12):pp. 750–758.

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi. 2008. "Semiparametric Efficiency in GMM Models with Auxiliary Data." *The Annals of Statistics* 36 (2):808–843.

Chen, Xuan and Carlos A Flores. 2015. "Bounds on treatment effects in the presence of sample selection and noncompliance: the wage effects of job corps." *Journal of Business & Economic Statistics* 33 (4):523–540.

Chen, Xuan, Carlos A. Flores, and Alfonso Flores-Lagunes. 2017. "Going beyond LATE: Bounding Average Treatment Effects of Job Corps Training." *Journal of Human Resources* .

Chesher, Andrew. 2010. "Instrumental Variable Models for Discrete Outcomes." *Econometrica* 78 (2):pp. 575–601.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *The Quarterly Journal of Economics* 126 (4):1593–1660. URL http://dx.doi.org/10.1093/qje/qjr041.

Chiburis, Richard C. 2010. "Semiparametric Bounds on Treatment Effects." *Journal of Econometrics* 159 (2):267–275.

Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schonberg. 2018. "Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance." *Journal of Polical Economy* 126 (6):pp. 2356–2409.

Das, Mitali, Whitney K. Newey, and Francis Vella. 2003. "Nonparametric Estimation of Sample Selection Models." *The Review of Economic Studies* 70 (1):pp. 33–58.

Deb, Partha, Murat K. Munkin, and Pravin K. Trivedi. 2006. "Bayesian Analysis of the Two-part Model with Endogeneity: Application to Health Care Expenditure." *Journal of Applied Econometrics* 21:pp. 1081–1099.

DeMel, Suresh, David McKenzie, and Christopher Woodruff. 2013. "The Demand for, and Consequences of, Formalization among Informal First in Sri Lanka." *American Economic Journal: Applied Economics* 5 (2):122–150.

Dobbie, Will and Roland G. Fryer Jr. 2015. "The Medium-Term Impacts of High-Achieving Charter Schools." *Journal of Political Economy* 123 (5):pp. 985–1037.

Fan, Jianqing and Irene Gijbels. 1996. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, vol. 66. CRC Press.

Farber, Henry S. 1993. "The Incidence and Costs of Job Loss: 1982-91." *Brookings Papers: Microeconomics* :pp. 73–132.

Frangakis, Constantine E and Donald B Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58 (1):21–29.

Fricke, Hans, Markus Frolich, Martin Huber, and Michael Lechner. 2015. "Endogeneity and Non-Response Bias in Treatment Evaluation: Nonparametric Identification of Causal Effects by Instruments." Available at: https://www.iza.org/publications/dp/9428/endogeneity-and-non-response-bias-in-treatment-evaluation-nonparametric-identification-of

Frolich, Markus and Martin Huber. 2014. "Treatment Evaluation with Multiple Outcome Periods Under Endogeneity and Attrition." *Journal of the American Statistical Association* 109 (508):1697–1711.

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

Heckman, James J. and George J. Borjas. 1980. "Does Unemployment Cause Future Unemployment? Definitions, Questions and Answers from a Continuous Time Model of Heterogeneity and State Dependence." *Economica* 47 (187):247–283. URL http://www.jstor.org/stable/2553150.

Heckman, James J., Robert LaLonde, and Jeff Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics*, vol. 3A, edited by Orley Ashenfelter and David Card. Elsevier, pp. 1865–2097.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics* 88 (3):pp. 389–432.

Heckman, James J. and Edward Vytlacil. 1999. "Local Instrumental Variable and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences* 96:4730–4734.

———. 2001a. "Local Instrumental Variables." *in Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya* :1–46.

———. 2001b. "Policy-Relevant Treatment Effects." *American Economic Review: Papers and Proceedings* 91 (2):pp. 107–111.

———. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation 1." *Econometrica* 73 (3):669–738.

Helland, Eric and Jungmo Yoon. 2017. "Estimating the Effects of the English Rule on Litigation Outcomes." *The Review of Economics and Statistics* 99 (4):pp. 678–682.

Horowitz, Joel L. and Charles F. Manski. 1995. "Identification and Robustness with Contaminated and Corrupted Data." *Econometrica* 63 (2):pp. 281–302.

Huber, Martin. 2014. "Treatment Evaluation in the Presence of Sample Selection." *Econometric Reviews* 33 (8):pp. 869–905.

Huber, Martin, Lukas Laffers, and Giovanni Mellace. 2017. "Sharp IV Bounds on Average Treatment Effects on the Treated and Other Populations under Endogeneity and Noncompliance." *Journal of Applied Econometrics* 32:pp. 56–79.

Huber, Martin and Giovanni Mellace. 2015. "Sharp Bounds on Causal Effects under Sample Selection." *Oxford Bulletin of Economics and Statistics* 77 (1):pp. 129–151.

Humphries, John Eric, Nicholas Mader, Daniel Tannenbaum, and Winnie van Dijk. 2019. "Does Eviction Cause Poverty? Quasi-Experimental Evidence from Cook County, IL." Available at: https://johnerichumphries.com/Evictions_draft_2019.pdf.

Imai, Kosuke. 2008. "Sharp Bounds on the Causal Effects in Randomized Experiments with Truncation- by- Death." *Statistics and Probability Letters* 78 (2):pp. 144–149.

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2):467–475. URL http://www.jstor.org/stable/2951620.

Jacobson, Louis S., Robert J. LaLonde, and Daniel G. Sullivan. 1993. "Earnings Losses of Displaced Workers." *The American Economic Review* 83 (4):pp. 685–709.

Kédagni, D. 2018. "Identifying Treatment Effects in the Presence of Confounded Types." *Economics Working Papers, Iowa State University* 18014.

Kitagawa, Toru. 2015. "A Test for Instrument Validity." *Econometrica* 83 (5):pp. 2043–2063.

Kline, Patrick and Melissa Tartari. 2016. "Bounding the Labor Supply Responses to a Randomized Welfare Experiment: A Revealed Preference Approach." *The American Economic Review* 106 (4):pp. 972–1014.

Kowalski, Amanda E. 2018. "Extrapolation Using Selection and Moral Hazard Heterogeneity from within the Oregon Health Insurance Experiment." NBER Working Paper n. 24647.

Krueger, Alan B. and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *The Economic Journal* 111 (468):pp. 1–28.

Lechner, Michael and Blaise Mell. 2010. "Partial Idendification of Wage Effects of Training Programs." Available at: https://ideas.repec.org/p/bro/econwp/2010-8.html.

Lee, David S. 2009. "Training, wages, and sample selection: Estimating sharp bounds on treatment effects." *The Review of Economic Studies* 76 (3):1071–1102.

Lee, Sokbae and Bernard Salanie. 2018. "Identifying Effects of Multivalued Treatments." *Econometrica* 86 (6):pp. 1939–1963.

Machado, Cecilia, Azeem M. Shaikh, and Edward Vytlacil. 2018. "Instrumental Variables and the Sign of the Average Treatment Effect." Forthcoming on the Journal of Econometrics. Available at: http://bit.ly/2MDnRFF.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80 (2):pp. 319–323.

———. 1997. "Monotone Treatment Response." *Econometrica* 65 (6):pp. 1311–1334.

Manski, Charles F. and John V. Pepper. 2000. "Monotone Instrumental Variables: With an Application to the Returns to Schooling." *Econometrica* 68 (4):pp. 997–1010.

Mealli, Fabrizia and Barbara Pacini. 2013. "Using Secondary Outcomes to Sharpen Inference in Randomized Experiments with Noncompliance." *Journal of the American Statistical Association* 108 (503):pp. 1120–1131.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using Instrumental Variables for Inference about Policy Relevant Treatment Effects." *Econometrica* 86 (5):pp. 1589–1619. NBER Working Paper 23568.

Mourifié, I. and Y. Wan. 2017. "Testing Local Average Treatment Effect Assumptions." *The Review of Economics and Statistics* 99 (2):305–313.

Mourifie, Ismael, Marc Henry, and Romuald Meango. 2018. "Sharp Bounds and Testability of a Roy Model of STEM Major Choices." Available at: http://bit.ly/2RXq9Dn.

Mullahy, John. 2018. "Individual Results May Vary: Inequality-probability Bounds for Some Health-outcome Treatment Effects." *Journal of Health Economics* 61:pp. 151–162.

Semykina, Anastasia and Jeffrey M. Wooldrige. 2018. "Binary Response Panel Data Models with Sample Selection and Self-selection." *Journal of Applied Econometrics* 33:pp. 179–197.

Sexton, Mary and Richard Hebel. 1984. "A Clinical Trial of Change in Maternal Smoking and its Effects on Birth Weight." *Journal of the American Medical Association* 251 (7):pp. 911–915.

Shaikh, Azeem and Edward Vytlacil. 2011. "Partial Identification in Triangular Systems of Equations with Binary Dependent Variables." *Econometrica* 79 (3):pp. 949–955.

Steinmayr, Andreas. 2014. "When a Random Sample is not Random: Bounds on the Effect of Migration on Children Left Behind." Working Paper.

Tamer, Elie. 2010. "Partial Identification in Econometrics." *The Annual Review of Economics* 2:pp. 167–195.

U.S. Department of Health and Human Services. 2004. *The Health Consequences of Smoking: A Report of the Surgeon General*. U.S. Department of Health and Human Services, Public Health Service, Office on Smoking and Health. Available at: http://bit.ly/2Kb52Ih.

Vytlacil, E. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica* 70 (1):331–341.

Zhang, Junni L., Donald B. Rubin, and Fabrizia Mealli. 2008. "Evaluating the Effects of Job Training Programs on Wages through Principal Stratification." In *Modelling and Evaluating Treatment Effects in Econometrics*. Emerald Group Publishing Limited, 117–145.

Zhou, Xiang and Yu Xie. 2019. "Marginal Treatment Effects from a Propensity Score Perspective." *Journal of Political Economy* .

Supporting Information
(Online Appendix)

A Proofs

A.1 Proof of Proposition 1

The validity of the bounds is proven in the main text. It remains to show that the bounds are sharp. Given the restrictions that Assumptions 1-5 impose on the data (i.e., equations (3.3) and (3.4)), we need to find joint distributions on $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ that satisfy these restrictions, induce the joint distribution on the data (Y, S, D, Z), and achieve any value $\delta \in \left[\underline{\Delta}_1, \overline{\Delta}_1\right]$.

Assume that Y^* is absolutely continuous and has a strictly positive density. Fix $p \in [0,1]$ arbitrarily. Define $\overline{\upsilon} \coloneqq \text{argmin}_{\upsilon \in \Upsilon(p)} \{LB_1(p,\upsilon) - UB_0(p,\upsilon)\}$, where $LB_1(p,\upsilon)$ and $UB_0(p,\upsilon)$ are defined in Subsection 3.3. For brevity, denote the strata by OO = always observed, NO = observed only when treated, ON = observed only when untreated and NN = never observed, and the probability of the stratum k conditional on $(\tilde{V} = p, Z = z)$ by $\tilde{\pi}_{k|p,z}$. The probabilities $\tilde{\pi}_{k|p,z}$ are given by:

$$
\begin{aligned}
\tilde{\pi}_{OO|p,z} &= \overline{\upsilon} \\
\tilde{\pi}_{NO|p,z} &= \frac{\partial \mathbb{P}(S=1, D=1 | P(Z)=p)}{\partial p} - \overline{\upsilon} \\
\tilde{\pi}_{ON|p,z} &= -\frac{\partial \mathbb{P}(S=1, D=0 | P(Z)=p)}{\partial p} - \overline{\upsilon} \\
\tilde{\pi}_{NN|p,z} &= 1 - \tilde{\pi}_{OO|p,z} - \tilde{\pi}_{NO|p,z} - \tilde{\pi}_{ON|p,z}.
\end{aligned}
$$

According to Lemma 1, the above quantities are positive. Note also that they add up to 1 by construction.

Now, we will show that the the lower bound $\underline{\Delta}_1$ is attainable. Define

$$
F_{\tilde{Y}_1^* | \tilde{S}_1=1, \tilde{V}=p, Z=z}(y_1) = \frac{\frac{\partial \mathbb{P}(Y \leq y_1, S=1, D=1 | P(Z)=p)}{\partial p}}{\frac{\partial \mathbb{P}(S=1, D=1 | P(Z)=p)}{\partial p}}
$$

and note that this function has to be a C.D.F. under the identifying assumptions as it is a mixture of two distributions $F_{Y_1^*|OO,V=p}$ and $F_{Y_1^*|NO,V=p}$. Similarly, define

$$F_{\tilde{Y}_0^*|\tilde{S}_0=1,\tilde{V}=p,Z=z}(y_0) = \frac{\frac{\partial \mathbb{P}(Y\leq y_0,S=1,D=0|P(Z)=p)}{\partial p}}{\frac{\partial \mathbb{P}(S=1,D=0|P(Z)=p)}{\partial p}}$$

and note that this function has to be a C.D.F. under the identifying assumptions as it is a mixture of two distributions $F_{Y_0^*|OO,V=p}$ and $F_{Y_0^*|ON,V=p}$.

Suppose that $\tilde{Y}_0 \sim F_{\tilde{Y}_0^*|\tilde{S}_0=1,\tilde{V}=p,Z=z}$ and $\tilde{Y}_1 \sim F_{\tilde{Y}_1^*|\tilde{S}_1=1,\tilde{V}=p,Z=z}$. Define $\tilde{V} = F_{\tilde{Y}_1}(\tilde{Y}_1)$ and

$$\begin{aligned}
\mathbb{P}(\tilde{Y}_1^* \leq y_1|OO,\tilde{V}=p,Z=z) &= \mathbb{P}\left(\tilde{Y}_1 \leq y_1|\tilde{Y}_1 \leq F_{\tilde{Y}_1}^{-1}\left(\frac{\tilde{\pi}_{OO|p,z}}{\tilde{\pi}_{OO|p,z}+\tilde{\pi}_{NO|p,z}}\right)\right),\\
\mathbb{P}(\tilde{Y}_1^* \leq y_1|NO,\tilde{V}=p,Z=z) &= \mathbb{P}\left(\tilde{Y}_1 \leq y_1|\tilde{Y}_1 > F_{\tilde{Y}_1}^{-1}\left(\frac{\tilde{\pi}_{OO|p,z}}{\tilde{\pi}_{OO|p,z}+\tilde{\pi}_{NO|p,z}}\right)\right),\\
\mathbb{P}(\tilde{Y}_1^* \leq y_1|k,\tilde{V}=p,Z=z) &= \frac{\frac{\partial \mathbb{P}(Y\leq y_1,S=1,D=1|P(Z)=p)}{\partial p}}{\frac{\partial \mathbb{P}(S=1,D=1|P(Z)=p)}{\partial p}}, \quad k\in\{ON,NN\}\\
\mathbb{P}(\tilde{Y}_0^* \leq y_0|OO,\tilde{V}=p,Z=z) &= \mathbb{P}\left(\tilde{Y}_0 \leq y_0|\tilde{Y}_0 > F_{\tilde{Y}_0}^{-1}\left(\frac{\tilde{\pi}_{ON|p,z}}{\tilde{\pi}_{OO|p,z}+\tilde{\pi}_{ON|p,z}}\right)\right),\\
\mathbb{P}(\tilde{Y}_0^* \leq y_0|NO,\tilde{V}=p,Z=z) &= \mathbb{P}\left(\tilde{Y}_0 \leq y_0|\tilde{Y}_0 \leq F_{\tilde{Y}_0}^{-1}\left(\frac{\tilde{\pi}_{ON|p,z}}{\tilde{\pi}_{OO|p,z}+\tilde{\pi}_{ON|p,z}}\right)\right),\\
\mathbb{P}(\tilde{Y}_0^* \leq y_0|k,\tilde{V}=p,Z=z) &= \frac{\frac{\partial \mathbb{P}(Y\leq y_0,S=1,D=0|P(Z)=p)}{\partial p}}{\frac{\partial \mathbb{P}(S=1,D=0|P(Z)=p)}{\partial p}}, \quad k\in\{NO,NN\}.
\end{aligned}$$

Finally, define the joint density (mass) function on $(\tilde{Y}_0^*,\tilde{Y}_1^*,\tilde{S}_0,\tilde{S}_1,\tilde{V},Z)$:

$$\begin{aligned}
f_{\tilde{Y}_0^*,\tilde{Y}_1^*,(\tilde{S}_0,\tilde{S}_1),\tilde{V},Z}(y_0,y_1,k,p,z) = &\frac{\partial \mathbb{P}(\tilde{Y}_0^* \leq y_0|k,\tilde{V}=p,Z=z)}{\partial y_0}\cdot\\
&\frac{\partial \mathbb{P}(\tilde{Y}_1^* \leq y_1|k,\tilde{V}=p,Z=z)}{\partial y_1}\cdot\tilde{\pi}_{k|p,z}\cdot f_Z(z)
\end{aligned}$$

for $k\in\{OO,NO,ON,NN\}$, where $f_Z(z)$ is the density function of Z.

Notice that the lower bound in Proposition 1 is attained by the distributions of $\tilde{Y}_0^*\Big|OO,\tilde{V},Z$ and $\tilde{Y}_1^*\Big|OO,\tilde{V},Z$, i.e.,

$$\underline{\Delta}_1 = \mathbb{E}(\tilde{Y}_1^* - \tilde{Y}_0^*|OO,\tilde{V}=p,Z=z).$$

Similar reasoning holds for the upper bound, $\overline{\Delta}_1$. To attain any value $\delta \in \left(\underline{\Delta}_1, \overline{\Delta}_1\right)$, we can use convex combinations of the joint distributions that attain the lower and upper bounds. Moreover, note that the joint distribution of $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ satisfy the restrictions imposed by Assumptions 1-5 and induce the joint distribution on the data (Y, S, D, Z) by construction.

A.2 Proof of Theorem 1

The validity of the bounds is proven in the main text. It remains to show that the bounds are sharp. Given the restrictions that Assumptions 1-6 impose on the data (i.e., equations (2.3), (3.3) and (3.4)), we need to find joint distributions on $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ that satisfy these restrictions, induce the joint distribution on the data (Y, S, D, Z), and achieve any value $\delta \in \left[\underline{\Delta}_2, \overline{\Delta}_2\right]$.

Assume that Y^* is absolutely continuous and has a strictly positive density. Fix $p \in [0,1]$ arbitrarily. For brevity, denote the strata by OO = always observed, NO = observed only when treated, and NN = never observed, and the probability of the stratum k conditional on $(\tilde{V} = p, Z = z)$ by $\tilde{\pi}_{k|p,z}$. The probabilities $\tilde{\pi}_{k|p,z}$ are given by:

$$\begin{aligned}
\tilde{\pi}_{OO|p,z} &= -\frac{\partial \mathbb{P}(S=1, D=0|P(Z)=p)}{\partial p} \\
\tilde{\pi}_{NO|p,z} &= \frac{\partial \mathbb{P}(S=1|P(Z)=p)}{\partial p}, \\
\tilde{\pi}_{NN|p,z} &= \frac{\partial \mathbb{P}(S=0, D=1|P(Z)=p)}{\partial p}.
\end{aligned}$$

Under Assumptions 1-6, the above quantities are positive. We show that they sum up to one. Indeed,

$$\begin{aligned}
\tilde{\pi}_{OO|p,z} + \tilde{\pi}_{NO|p,z} &= -\frac{\partial \mathbb{P}(S=1, D=0|P(Z)=p)}{\partial p} + \frac{\partial \mathbb{P}(S=1|P(Z)=p)}{\partial p}, \\
&= \frac{\partial \mathbb{P}(S=1, D=1|P(Z)=p)}{\partial p}.
\end{aligned}$$

Then

$$\begin{aligned}\tilde{\pi}_{OO|p,z}+\tilde{\pi}_{NO|p,z}+\tilde{\pi}_{NN|p,z} &= \frac{\partial \mathbb{P}(S=0,D=1|P(Z)=p)}{\partial p}+\frac{\partial \mathbb{P}(S=1,D=1|P(Z)=p)}{\partial p},\\ &= \frac{\partial \mathbb{P}(D=1|P(Z)=p)}{\partial p}\end{aligned}$$

We have

$$\begin{aligned}\mathbb{P}(D=1|P(Z)=p) &= \mathbb{P}(V\leq p|P(Z)=p) \quad \text{by definition},\\ &= \mathbb{P}(V\leq p) \quad \text{under Assumption 1},\\ &= p \quad \text{under Assumption 5}.\end{aligned}$$

Therefore, $\tilde{\pi}_{OO|p,z}+\tilde{\pi}_{NO|p,z}+\tilde{\pi}_{NN|p,z}=1$.

Now, we will show that the the lower bound $\underline{\Delta}_1$ is attainable. Define

$$F_{\tilde{Y}_1^*|\tilde{S}_1=1,\tilde{V}=p,Z=z}(y_1)=\frac{\frac{\partial \mathbb{P}(Y\leq y_1,S=1,D=1|P(Z)=p)}{\partial p}}{\frac{\partial \mathbb{P}(S=1,D=1|P(Z)=p)}{\partial p}}$$

and note that this function has to be a C.D.F. under the identifying assumptions as it is a mixture of two distributions $F_{Y_1^*|OO,V=p}$ and $F_{Y_1^*|NO,V=p}$. Similarly, define

$$F_{\tilde{Y}_0^*|OO,\tilde{V}=p,Z=z}(y_0)=\frac{\frac{\partial \mathbb{P}(Y\leq y_0,S=1,D=0|P(Z)=p)}{\partial p}}{\frac{\partial \mathbb{P}(S=1,D=0|P(Z)=p)}{\partial p}}$$

and note that this function has to be a C.D.F. under the identifying assumptions as it is a mixture of two distributions $F_{Y_0^*|OO,V=p}$ and $F_{Y_0^*|ON,V=p}$.

Suppose that $\tilde{Y}_1\sim F_{\tilde{Y}_1^*|\tilde{S}_1=1,\tilde{V}=p,Z=z}$. Define $\tilde{V}=F_{\tilde{Y}_1}(\tilde{Y}_1)$ and

$$\mathbb{P}(\tilde{Y}_1^*\leq y_1|OO,\tilde{V}=p,Z=z)=\mathbb{P}\left(\tilde{Y}_1\leq y_1|\tilde{Y}_1\leq F_{\tilde{Y}_1}^{-1}\left(\frac{\tilde{\pi}_{OO|p,z}}{\tilde{\pi}_{OO|p,z}+\tilde{\pi}_{NO|p,z}}\right)\right), \tag{A.1}$$

$$\mathbb{P}(\tilde{Y}_1^*\leq y_1|NO,\tilde{V}=p,Z=z)=\mathbb{P}\left(\tilde{Y}_1\leq y_1|\tilde{Y}_1> F_{\tilde{Y}}^{-1}\left(\frac{\tilde{\pi}_{OO|p,z}}{\tilde{\pi}_{OO|p,z}+\tilde{\pi}_{NO|p,z}}\right)\right), \tag{A.2}$$

$$\mathbb{P}(\tilde{Y}_1^* \leq y_1 | NN, \tilde{V} = p, Z = z) = \frac{\frac{\partial \mathbb{P}(Y \leq y_1, S=1, D=1 | P(Z)=p)}{\partial p}}{\frac{\partial \mathbb{P}(S=1, D=1 | P(Z)=p)}{\partial p}},$$

$$\mathbb{P}(\tilde{Y}_0^* \leq y_0 | k, \tilde{V} = p, Z = z) = \frac{\frac{\partial \mathbb{P}(Y \leq y_0, S=1, D=0 | P(Z)=p)}{\partial p}}{\frac{\partial \mathbb{P}(S=1, D=0 | P(Z)=p)}{\partial p}}, \quad k \in \{NO, NN\}.$$

Finally, define the joint density (mass) function on $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$:

$$f_{\tilde{Y}_0^*, \tilde{Y}_1^*, (\tilde{S}_0, \tilde{S}_1), \tilde{V}, Z}(y_0, y_1, k, p, z) = \frac{\partial \mathbb{P}(\tilde{Y}_0^* \leq y_0 | k, \tilde{V} = p, Z = z)}{\partial y_0} \cdot \frac{\partial \mathbb{P}(\tilde{Y}_1^* \leq y_1 | k, \tilde{V} = p, Z = z)}{\partial y_1} \cdot \tilde{\pi}_{k|p,z} \cdot f_Z(z)$$

for $k \in \{OO, NO, NN\}$, where $f_Z(z)$ is the density function of Z.

Notice that the lower bound in Proposition 1 is attained by the distributions of $\tilde{Y}_0^* \Big| OO, \tilde{V}, Z$ and $\tilde{Y}_1^* \Big| OO, \tilde{V}, Z$, i.e.,

$$\underline{\Delta}_2 = \mathbb{E}(\tilde{Y}_1^* - \tilde{Y}_0^* | OO, \tilde{V} = p, Z = z).$$

Similar reasoning holds for the upper bound, $\overline{\Delta}_2$. To attain any value $\delta \in \left(\underline{\Delta}_2, \overline{\Delta}_2\right)$, we can use convex combinations of the joint distributions that attain the lower and upper bounds. Moreover, note that the joint distribution of $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ satisfy the restrictions imposed by Assumptions 1-6 and induce the joint distribution on the data (Y, S, D, Z) by construction.

A.3 Proof of Theorem 2

The validity of the bounds is proven in the main text. It remains to show that the bounds are sharp. Given the restrictions that Assumptions 1-7 impose on the data (i.e., equations (2.3), (3.3) and (3.4)), we need to find joint distributions on $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ that satisfy these restriction, satisfy the stochastic dominance assumption, induce the joint distribution on the data (Y, S, D, Z), and achieve any value $\delta \in \left[\underline{\Delta}_3, \overline{\Delta}_3\right]$.

The proof of Theorem 2 is very similar to the one in Appendix A.1. We only have to

modify equations (A.1) and (A.2) to:

$$\mathbb{P}(\tilde{Y}_1^* \leq y_1 | OO, \tilde{V} = p, Z = z) = \mathbb{P}\left(\tilde{Y}_1 \leq y_1\right),$$

$$\mathbb{P}(\tilde{Y}_1^* \leq y_1 | NO, \tilde{V} = p, Z = z) = \mathbb{P}\left(\tilde{Y}_1 \leq y_1\right).$$

This changes ensure that the joint distribution of $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ satisfy the Stochastic Dominance Assumption by construction.

A.4 Proof of Proposition 3

The validity of the bounds is proven in the main text. The joint distributions on $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ described in proofs A.1, A.2 and A.3 can also be used to prove that the bounds in Proposition 3 are sharp.

A.5 Proof of Theorem 5

This proof is similar to that of Theorem 1. It is given for completeness. The validity of the bounds is proven in the main text. It remains to show that the bounds are sharp. Given the restrictions that Assumptions 1, 5, 6 and 9 impose on the data, we need to find a joint distribution on $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ that satisfies these assumptions, induces the joint distribution on the data (Y, S, D, Z), and achieves the lower bound, and similarly for the upper bound.

Assume that Y is absolutely continuous and has a strictly positive density. Denote $\tilde{\pi}_{k|p,z}$ the probability of the stratum k given $(\tilde{V} = p, Z = z)$. For a given $p \in (p_1, p_K)$, there exists a unique $\ell \in \{2, \ldots, K\}$ such that $p_{\ell-1} < p \leq p_\ell$. Define the distribution on the strata:

$$\begin{aligned}
\tilde{\pi}_{OO|p,z} &= -\frac{\mathbb{P}(S=1, D=0|P=p_\ell) - \mathbb{P}(S=1, D=0|P=p_{\ell-1})}{p_\ell - p_{\ell-1}}, \\
\tilde{\pi}_{NO|p,z} &= \frac{\mathbb{P}(S=1|P=p_\ell) - \mathbb{P}(S=1|P=p_{\ell-1})}{p_\ell - p_{\ell-1}}, \\
\tilde{\pi}_{NN|p,z} &= \frac{\mathbb{P}(S=0, D=1|P=p_\ell) - \mathbb{P}(S=0, D=1|P=p_\ell - 1)}{p_\ell - p_{\ell-1}}.
\end{aligned}$$

Under Assumptions 1, 5, and 6, the above quantities are positive. We will now show that they sum up to one. Indeed,

$$\tilde{\pi}_{OO|p,z} + \tilde{\pi}_{NO|p,z} = \frac{\mathbb{P}(S=1, D=1|P=p_\ell) - \mathbb{P}(S=1, D=1|P=p_{\ell-1})}{p_\ell - p_{\ell-1}}.$$

Then

$$\begin{aligned}\tilde{\pi}_{OO|p,z} + \tilde{\pi}_{NO|p,z} + \tilde{\pi}_{NN|p,z} &= \frac{\mathbb{P}(D=1|P=p_\ell) - \mathbb{P}(D=1|P=p_{\ell-1})}{p_\ell - p_{\ell-1}}, \\ &= \frac{p_\ell - p_{\ell-1}}{p_\ell - p_{\ell-1}} = 1\end{aligned}$$

Define

$$\mathbb{P}(\tilde{Y}_1^* \leq y_1, \tilde{S}_1 = 1|\tilde{V}=p) = \frac{\mathbb{P}(Y \leq y_1, S=1, D=1|P=p_\ell) - \mathbb{P}(Y \leq y_1, S=1, D=1|P=p_{\ell-1})}{p_\ell - p_{\ell-1}},$$

$$\mathbb{P}(\tilde{Y}_0^* \leq y_0, \tilde{S}_0 = 1|\tilde{V}=p) = -\frac{\mathbb{P}(Y \leq y_0, S=1, D=0|P=p_\ell) - \mathbb{P}(Y \leq y_0, S=1, D=0|P=p_{\ell-1})}{p_\ell - p_{\ell-1}},$$

and

$$\mathbb{P}(\tilde{Y}_d^* \leq y_d|\tilde{S}_d=1, \tilde{V}=p) = \frac{\mathbb{P}(\tilde{Y}_d^* \leq y_d, \tilde{S}_d=1|p_{\ell-1} < \tilde{V} \leq p_\ell)}{\mathbb{P}(\tilde{S}_d=1|p_{\ell-1} < \tilde{V} \leq p_\ell)} \text{ for } d \in \{0,1\},$$

where $\mathbb{P}(\tilde{S}_d=1|\tilde{V}=p) = \lim_{y_d \to \infty} \mathbb{P}(\tilde{Y}_d^* \leq y_d, \tilde{S}_d=1|p_{\ell-1} < \tilde{V} \leq p_\ell)$, and

$$\mathbb{P}(\tilde{Y}_0^* \leq y_0|\tilde{S}_0=1, \tilde{V}=p) = \mathbb{P}(\tilde{Y}_0^* \leq y_0|S_0=1, S_1=1, \tilde{V}=p).$$

Suppose that $\tilde{Y}_1 \sim F_{\tilde{Y}_1^*|S_1=1,\tilde{V}=p}$. Define $\tilde{V} = F_{\tilde{Y}_1}(\tilde{Y}_1)$ and

$$\mathbb{P}(\tilde{Y}_1^* \leq y_1|S_0=1, S_1=1, \tilde{V}=p, Z=z) = \mathbb{P}\left(\tilde{Y}_1 \leq y_1|\tilde{Y}_1 \leq F_{\tilde{Y}_1}^{-1}\left(\frac{\tilde{\pi}_{OO|p,z}}{\tilde{\pi}_{OO|p,z} + \tilde{\pi}_{NO|p,z}}\right)\right),$$

$$\mathbb{P}(\tilde{Y}_1^* \leq y_1 | NE, \tilde{V} = p, Z = z) \quad = \quad \mathbb{P}\left(\tilde{Y}_1 \leq y_1 | \tilde{Y}_1 > F_{\tilde{Y}}^{-1}\left(\frac{\tilde{\pi}_{OO|p,z}}{\tilde{\pi}_{OO|p,z} + \tilde{\pi}_{NO|p,z}}\right)\right),$$

$$\mathbb{P}(\tilde{Y}_1^* \leq y_1 | NN, \tilde{V} = p, Z = z) = \frac{\mathbb{P}(Y \leq y_1, S = 1, D = 1 | P = p_\ell) - \mathbb{P}(Y \leq y_1, S = 1, D = 1 | P = p_{\ell-1})}{p_\ell - p_{\ell-1}},$$

$$\mathbb{P}(\tilde{Y}_0^* \leq y_0 | k, \tilde{V} = p, Z = z) = -\frac{\mathbb{P}(Y \leq y_0, S = 1, D = 0 | P = p_\ell) - \mathbb{P}(Y \leq y_0, S = 1, D = 0 | P = p_{\ell-1})}{p_\ell - p_{\ell-1}}, \quad k \in \{NO, NN\}.$$

Finally, define the joint density (mass) function on $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$:

$$\begin{aligned} f_{\tilde{Y}_0^*, \tilde{Y}_1^*, (\tilde{S}_0, \tilde{S}_1), \tilde{V}, Z}(y_0, y_1, k, p, z) \quad = \quad & \frac{\partial \mathbb{P}(\tilde{Y}_0^* \leq y_0 | k, \tilde{V} = p, Z = z)}{\partial y_0} * \\ & \frac{\partial \mathbb{P}(\tilde{Y}_1^* \leq y_1 | k, \tilde{V} = p, Z = z)}{\partial y_1} * \tilde{\pi}_{k|z} * f_Z(z), \\ & k \in \{OO, NO, NN\}, \end{aligned}$$

where $f_Z(z)$ is the probability mass function of Z.

Notice that the lower bound in Theorem 5 is

$$LB \quad = \quad \mathbb{E}\left[\tilde{Y}_1 | \tilde{Y}_1 \leq F_{\tilde{Y}_1}^{-1}\left(\frac{\tilde{\pi}_{OO|p,z}}{\tilde{\pi}_{OO|p,z} + \tilde{\pi}_{NO|p,z}}\right)\right] - \mathbb{E}[\tilde{Y}_0^* | S_0 = 1, S_1 = 1, \tilde{V} = p],$$

and is attained by the proposed joint distribution.

Similar reasoning holds for the upper bound, which is

$$UB \quad = \quad \mathbb{E}\left[\tilde{Y}_1 | \tilde{Y}_1 > F_{\tilde{Y}_1}^{-1}\left(1 - \frac{\tilde{\pi}_{OO|p,z}}{\tilde{\pi}_{OO|p,z} + \tilde{\pi}_{NO|p,z}}\right)\right] - \mathbb{E}[\tilde{Y}_0^* | S_0 = 1, S_1 = 1, \tilde{V} = p].$$

B Sharp testable implications for Assumptions 1, 5 and 6

Suppose that Z contains at least one continuous instrument. Whenever Assumptions 1, 5 and 6 hold, inequalities (3.3), (3.4) and (2.3) must hold, i.e.,

$$0 \leq \frac{\partial \mathbb{E}[\mathbb{1}\{Y \in A\} SD | P(Z) = p]}{\partial p} \leq 1, \tag{B.1}$$

$$0 \leq -\frac{\partial \mathbb{E}[\mathbb{1}\{Y \in A\} S(1-D) | P(Z) = p]}{\partial p} \leq 1, \tag{B.2}$$

$$0 \leq \frac{\partial \mathbb{P}(S = 1 | P(Z) = p)}{\partial p} \leq 1 \tag{B.3}$$

for all borel sets $A \subset \mathbb{R}$ and $p \in (0, 1)$, where the last inequality holds because

$$\mathbb{P}(NO | V = p) = \frac{\partial \mathbb{P}(S = 1 | P(Z) = p)}{\partial p}.$$

In addition to the inequalities above, the following equalities must hold:

$$\mathbb{P}(Y \in A, S = 1, D = 1 | Z = z) = \mathbb{P}(Y \in A, S = 1, D = 1 | P(Z) = P(z)), \tag{B.4}$$

$$\mathbb{P}(Y \in A, S = 1, D = 0 | Z = z) = \mathbb{P}(Y \in A, S = 1, D = 0 | P(Z) = P(z)), \tag{B.5}$$

$$\mathbb{P}(S = 0, D = 1 | Z = z) = \mathbb{P}(S = 0, D = 1 | P(Z) = P(z)), \tag{B.6}$$

$$\mathbb{P}(S = 0, D = 0 | Z = z) = \mathbb{P}(S = 0, D = 0 | P(Z) = P(z)). \tag{B.7}$$

These equalities hold trivially when $P(z)$ is strictly monotone in z.

Theorem 3. *Consider the model (2.1).*

(i) If Assumptions 1, 5 and 6 hold, then inequalities (B.1) to (B.3) and equalities (B.4) to (B.7) hold.

(ii) If inequalities (B.1) to (B.3) and equalities (B.4) to (B.7) hold, then there exists a vector $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{V}, \tilde{S}_0, \tilde{S}_1, , Z)$ that satisfies model (2.1) and Assumptions 1, 5 and 6.

These testable implications are identical to those in Heckman and Vytlacil (2005) when there is no sample selection, i.e., $S = 1$ almost surely. If the instrument Z is binary, these

testable implications become

$$
\begin{aligned}
&0 \leq \mathbb{P}(Y \in A, S=1, D=1|Z=1) - \mathbb{P}(Y \in A, S=1, D=1|Z=0) \leq 1, \\
&0 \leq -(\mathbb{P}(Y \in A, S=1, D=0|Z=1) - \mathbb{P}(Y \in A, S=1, D=0|Z=0)) \leq 1, \\
&0 \leq \mathbb{P}(S=1|Z=1) - \mathbb{P}(S=1|Z=0) \leq 1.
\end{aligned}
$$

These inequalities generalize those in Balke and Pearl (1997) and Heckman and Vytlacil (2005) to the sample selection case, and can therefore be tested using the procedures proposed by Machado, Shaikh, and Vytlacil (2018), Mourifié and Wan (2017), Kitagawa (2015) or Huber and Mellace (2015).

Proof. (i) Inequalities (B.1) to (B.3) have been shown in the main text. It remains to show equalities (B.4) to B.7). We show (B.4) and the proofs for the other equalities can be obtained similarly.

$$
\begin{aligned}
\mathbb{P}(Y \in A, S=1, D=1|Z=z) &= \mathbb{P}(Y_1^* \in A, S_1=1, V \leq P(z)|Z=z), \\
&= \mathbb{P}(Y_1^* \in A, S_1=1, V \leq P(z)), \\
&= \mathbb{P}(Y_1^* \in A, S_1=1, V \leq P(z)|P(Z)=P(z)), \\
&= \mathbb{P}(Y_1^* \in A, S_1=1, V \leq P(Z)|P(Z)=P(z)), \\
&= \mathbb{P}(Y_1^* \in A, S_1=1, D=1|P(Z)=P(z)), \\
&= \mathbb{P}(Y \in A, S=1, D=1|P(Z)=P(z)),
\end{aligned}
$$

where the second and third equalities hold under Assumption 1.

(ii) Define $P(z) = \mathbb{P}(D=1|Z=z)$, and $\tilde{\pi}_{k|p,z}$ the probability of the stratum k given $(\tilde{V}=p, Z=z)$. Define the distribution on the strata:

$$
\begin{aligned}
\tilde{\pi}_{OO|p,z} &= -\frac{\partial \mathbb{P}(S=1, D=0|P(Z)=p)}{\partial p}, \\
\tilde{\pi}_{NO|p,z} &= \frac{\partial \mathbb{P}(S=1|P(Z)=p)}{\partial p},
\end{aligned}
$$

$$\tilde{\pi}_{NN|p,z} = \frac{\partial \mathbb{P}(S=0, D=1|P(Z)=p)}{\partial p}.$$

Inequalities (B.1) to (B.3) imply that the above quantities are positive and they sum up to one. Define

$$F_{\tilde{Y}_1^*|\tilde{S}_1=1,\tilde{V}=p,Z=z}(y_1) = \frac{\frac{\partial \mathbb{P}(Y\leq y_1,S=1,D=1|P(Z)=p)}{\partial p}}{\frac{\partial \mathbb{P}(S=1,D=1|P(Z)=p)}{\partial p}}$$

This function has to be a c.d.f. under the identifying assumptions as it is a mixture of two distributions $F_{Y_1^*|S_0=1,S_1=1,V=p}$ and $F_{Y_1^*|NO,V=p}$. Similarly,

$$F_{\tilde{Y}_0^*|S_0=1,S_1=1,\tilde{V}=p,Z=z}(y_0) = \frac{\frac{\partial \mathbb{P}(Y\leq y_0,S=1,D=0|P(Z)=p)}{\partial p}}{\frac{\partial \mathbb{P}(S=1,D=0|P(Z)=p)}{\partial p}}$$

Define

$$\mathbb{P}(\tilde{Y}_1^* \leq y_1|S_0=1,S_1=1,\tilde{V}=p,Z=z) = \mathbb{P}(\tilde{Y}_1^* \leq y_1|\tilde{S}_1=1,\tilde{V}=p),$$

$$\mathbb{P}(\tilde{Y}_1^* \leq y_1|NO,\tilde{V}=p,Z=z) = \mathbb{P}(\tilde{Y}_1^* \leq y_1|\tilde{S}_1=1,\tilde{V}=p),$$

$$\mathbb{P}(\tilde{Y}_1^* \leq y_1|NN,\tilde{V}=p,Z=z) = \frac{\frac{\partial \mathbb{P}(Y\leq y_1,S=1,D=1|P(Z)=p)}{\partial p}}{\frac{\partial \mathbb{P}(S=1,D=1|P(Z)=p)}{\partial p}},$$

$$\mathbb{P}(\tilde{Y}_0^* \leq y_0|k,\tilde{V}=p,Z=z) = \frac{\frac{\partial \mathbb{P}(Y\leq y_0,S=1,D=0|P(Z)=p)}{\partial p}}{\frac{\partial \mathbb{P}(S=1,D=0|P(Z)=p)}{\partial p}}, \quad k \in \{NO, NN\}.$$

Define the joint conditional distribution of $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1)$ given $(\tilde{V}=p, Z=z)$:

$$\begin{aligned}\mathbb{P}(\tilde{Y}_0^* \leq y_0, \tilde{Y}_1^* \leq y_1, (\tilde{S}_0,\tilde{S}_1)=k, |\tilde{V}=p,Z=z) = {} & \mathbb{P}(\tilde{Y}_0^* \leq y_0|k,\tilde{V}=p,Z=z) * \\ & \mathbb{P}(\tilde{Y}_1^* \leq y_1|k,\tilde{V}=p,Z=z) * \tilde{\pi}_{k|z}, \\ & k \in \{OO, NO, NN\},\end{aligned}$$

$$\mathbb{P}(\tilde{V} \leq p | Z = z) \quad = \quad p.$$

Finally, define

$$\begin{cases} \tilde{Y}^* & = & \tilde{Y}_1^* \tilde{D} + \tilde{Y}_0^* (1 - \tilde{D}) \\ \tilde{D} & = & \mathbb{1}\left\{ \tilde{V} \leq P(Z) \right\} \\ \tilde{S} & = & \tilde{S}_1 \tilde{D} + \tilde{S}_0 (1 - \tilde{D}) \\ \tilde{Y} & = & \tilde{Y}^* \tilde{S} \end{cases} \tag{B.8}$$

We can show that $(\tilde{Y}, \tilde{S}, \tilde{D}, Z)$ has the same joint distribution as (Y, S, D, Z).

$$\begin{aligned} \mathbb{P}(\tilde{Y} \leq y, \tilde{S} = 1, \tilde{D} = 1 | Z = z) &= \mathbb{P}(\tilde{Y}_1^* \leq y, \tilde{S}_1 = 1, \tilde{V} \leq P(z) | Z = z), \\ &= \mathbb{P}(\tilde{Y}_1^* \leq y, \tilde{S}_1 = 1 | \tilde{V} \leq P(z), Z = z) \mathbb{P}(\tilde{V} \leq P(z) | Z = z), \\ &= \mathbb{P}(\tilde{Y}_1^* \leq y, \tilde{S}_1 = 1 | \tilde{V} \leq P(z), Z = z) P(z), \\ &= \int_0^{P(z)} \mathbb{P}(\tilde{Y}_1^* \leq y, \tilde{S}_1 = 1 | \tilde{V} = v, Z = z) \frac{f_{\tilde{V}|Z=z}(v)}{P(z)} dv * P(z), \\ &= \int_0^{P(z)} \mathbb{P}(\tilde{Y}_1^* \leq y, \tilde{S}_1 = 1 | \tilde{V} = v, Z = z) dv, \\ &= \int_0^{P(z)} \mathbb{P}(\tilde{Y}_1^* \leq y | \tilde{S}_1 = 1, \tilde{V} = v, Z = z) \mathbb{P}(\tilde{S}_1 = 1 | \tilde{V} = v, Z = z) dv, \\ &= \int_0^{P(z)} \frac{\partial \mathbb{P}(Y \leq y, S = 1, D = 1 | P(Z) = v)}{\partial v} dv, \\ &= \mathbb{P}(Y \leq y, S = 1, D = 1 | P(Z) = P(z)) \\ &= \mathbb{P}(Y \leq y, S = 1, D = 1 | Z = z). \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbb{P}(\tilde{Y} \leq y, \tilde{S} = 1, \tilde{D} = 0 | Z = z) &= \mathbb{P}(Y \leq y, S = 1, D = 1 | Z = z), \\ \mathbb{P}(\tilde{S} = 0, \tilde{D} = 1 | Z = z) &= \mathbb{P}(S = 1, D = 1 | Z = z), \\ \mathbb{P}(\tilde{S} = 0, \tilde{D} = 0 | Z = z) &= \mathbb{P}(S = 0, D = 0 | Z = z). \end{aligned}$$

Finally, by construction, Assumptions 1, 5 and 6 hold. □

C Details on the numerical illustration

This brief appendix lists the relevant densities, expectations and objects of interest implied by the DGP used in Section 4 and Appendix D. We have that

$$\begin{aligned}
\mathbb{P}\left[S_0 = 1, S_1 = 1 | V = p\right] &= \Phi\left(\delta_0\sqrt{2} - \Phi^{-1}(p)\right), \\
\mathbb{P}\left[S_0 = 1 | V = p\right] &= \Phi\left(\delta_0\sqrt{2} - \Phi^{-1}(p)\right), \\
\mathbb{P}\left[S_1 = 1 | V = p\right] &= \Phi\left(\delta_0\sqrt{2} + \delta_1\sqrt{2} - \Phi^{-1}(p)\right), \\
\alpha\left(p, \upsilon^\ell\right) &= \max\left\{1 + \frac{\Phi\left(\delta_0\sqrt{2} - \Phi^{-1}(p)\right) - 1}{\Phi\left(\delta_0\sqrt{2} + \delta_1\sqrt{2} - \Phi^{-1}(p)\right)}, 0\right\}, \\
\beta\left(p, \upsilon^\ell\right) &= \max\left\{1 + \frac{\Phi\left(\delta_0\sqrt{2} + \delta_1\sqrt{2} - \Phi^{-1}(p)\right) - 1}{\Phi\left(\delta_0\sqrt{2} - \Phi^{-1}(p)\right)}, 0\right\}, \\
\alpha(p) &= \frac{\Phi\left(\delta_0\sqrt{2} - \Phi^{-1}(p)\right)}{\Phi\left(\delta_0\sqrt{2} + \delta_1\sqrt{2} - \Phi^{-1}(p)\right)}, \\
E\left[Y_1^* - Y_0^* | T = 1, S_0 = 1, S_1 = 1, V = p\right] &= (\beta_{1,1} - \beta_{0,1}) \cdot \Phi^{-1}(p), \\
E\left[Y_1^* - Y_0^* | T = 0, S_0 = 1, S_1 = 1, V = p\right] &= (\beta_{0,0} - \beta_{1,0}) \cdot \Phi^{-1}(p), \\
E\left[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1, V = p\right] &= (\beta_{1,1} - \beta_{1,0} - \beta_{0,1} + \beta_{0,0}) \cdot \frac{\Phi^{-1}(p)}{2}, \\
P\left[Y_0^* \leq y | T = 1, S_0 = 1, V = p\right] &= \Phi\left(y - \beta_{0,1}\Phi^{-1}(p)\right), \\
P\left[Y_1^* \leq y | T = 1, S_1 = 1, V = p\right] &= \Phi\left(y - \beta_{1,1}\Phi^{-1}(p)\right), \\
P\left[Y_0^* \leq y | T = 0, S_0 = 1, V = p\right] &= \Phi\left(y + \beta_{0,0}\Phi^{-1}(p)\right), \\
P\left[Y_1^* \leq y | T = 0, S_1 = 1, V = p\right] &= \Phi\left(y + \beta_{1,0}\Phi^{-1}(p)\right), \\
P\left[Y_0^* \leq y | S_0 = 1, V = p\right] &= \frac{1}{2}\Phi\left(y - \beta_{0,1}\Phi^{-1}(p)\right) + \frac{1}{2}\Phi\left(y + \beta_{0,0}\Phi^{-1}(p)\right), \\
P\left[Y_1^* \leq y | S_1 = 1, V = p\right] &= \frac{1}{2}\Phi\left(y - \beta_{1,1}\Phi^{-1}(p)\right) + \frac{1}{2}\Phi\left(y + \beta_{1,0}\Phi^{-1}(p)\right).
\end{aligned}$$

D Monte Carlo Simulation

In this appendix, we use the DGP described in Section 4 to produce Monte Carlo simulations using the estimator proposed in the main text. We analyze two sets of parameters: (i) $\delta_0 = 0.75$, $\delta_1 = 1.5$, $\beta_{00} = \beta_{01} = \beta_{10} = 0.1$, $\beta_{11} = 0.2$ and (ii) $\delta_0 = 0.2$, $\delta_1 = 2.0$,

$\beta_{00} = \beta_{01} = \beta_{10} = 0.1$, $\beta_{11} = 0.2$. The first set of parameters is ideal for the proposed estimator in the sense that the effective sample size is large since the trimming proportion $\alpha(p)$ is never small and the sample selection problem is not severe. The second set of parameters intentionally decreases the trimming proportion $\alpha(p)$, reducing the effective sample size and worsening the sample selection problem. We find that our estimator performs adequately in both DGPs when the sample size is equal to $n = 10,000$.

Based on the procedure describe in the Section 5, we need to specify the propensity score estimator, the grid points for the observed outcome variable ($\{y_1, \ldots, y_{K_n}\}$) and the evaluation points for the unobserved characteristic V. We estimate the propensity score with a logit estimator whose index is linear in the instrument Z, implying that the propensity score estimator is misspecified. For the grid points ($\{y_1, \ldots, y_{K_n}\}$), we choose the sample percentiles 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, implying that $K_n = 11$. For the evaluation points of the the unobserved characteristic V, we choose $p \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

In this simulation, we focus on the performance of six estimators: $\hat{\alpha}(p)$, $\hat{\Xi}_{OO,0}(p)$, $\widehat{LB}_1(p)$, $\widehat{UB}_1(p)$, $\hat{\underline{\Delta}}_2(p)$ and $\hat{\overline{\Delta}}_2(p)$. Table 4 reports the true value of their estimands for the first and second sets of parameters in Panel A and B, respectively. The important distinction between Panels A and B is the value of $\alpha(p)$.

Table 5 reports the average bias of our estimators, while Table 6 presents the mean squared error (MSE) of our estimators after normalizing it by the sample size ($n = 10,000$). For the first set of parameters (Panel A), the estimators' average bias and MSE is smaller for intermediate values of the propensity score. In this DGP, the treatment is determined by $D = \mathbb{1}\{V \leq \Phi(Z)\}$, implying that the data becomes sparser at low values of the propensity score for the treated group and at high values of the propensity score for the untreated group. As a consequence, the estimator's performance is worse when the propensity score is either small or large. Moreover, when the propensity score is large, the sample selection problem reduces the effective sample size, worsening the estimator's performance. Despite those challenges, the estimator's average bias and MSE are reasonably small for both set of parameters.

Table 4: True Value of the Estimands

	p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5	p = 0.6	p = 0.7	p = 0.8	p = 0.9
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: $\delta_0 = 0.75$, $\delta_1 = 1.5$, $\beta_{00} = \beta_{01} = \beta_{10} = 0.1$, $\beta_{11} = 0.2$									
$\alpha(p)$	0.99	0.97	0.94	0.91	0.86	0.79	0.71	0.59	0.42
$\Xi_{OO,0}(p)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$LB_1(p)$	-0.09	-0.11	-0.15	-0.2	-0.27	-0.35	-0.46	-0.62	-0.88
$UB_1(p)$	-0.04	0.03	0.10	0.17	0.26	0.37	0.51	0.70	1.00
$\underline{\Delta}_2(p)$	-0.09	-0.11	-0.15	-0.2	-0.27	-0.35	-0.46	-0.62	-0.88
$\overline{\Delta}_2(p)$	-0.04	0.03	0.10	0.17	0.26	0.37	0.51	0.70	1.00
$MTE^{OO}(p)$	-0.06	-0.04	-0.03	-0.01	0.00	0.01	0.03	0.04	0.06
Panel B: $\delta_0 = 0.2$, $\delta_1 = 2.0$, $\beta_{00} = \beta_{01} = \beta_{10} = 0.1$, $\beta_{11} = 0.2$									
$\alpha(p)$	0.94	0.87	0.79	0.7	0.61	0.51	0.41	0.29	0.16
$\Xi_{OO,0}(p)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$LB_1(p)$	-0.19	-0.29	-0.39	-0.5	-0.63	-0.77	-0.93	-1.14	-1.47
$UB_1(p)$	0.06	0.2	0.34	0.48	0.63	0.79	0.98	1.23	1.60
$\underline{\Delta}_2(p)$	-0.19	-0.29	-0.39	-0.5	-0.63	-0.77	-0.93	-1.14	-1.47
$\overline{\Delta}_2(p)$	0.06	0.2	0.34	0.48	0.63	0.79	0.98	1.23	1.60
$MTE^{OO}(p)$	-0.06	-0.04	-0.03	-0.01	0.00	0.01	0.03	0.04	0.06

Note: We define $\Xi_{OO,0}(p) \coloneqq \mathbb{E}[Y_0^* \,|\, S_0 = 1, S_1 = 1, V = p]$. The true values of the estimands are computed by numerical integration using 100,000 simulated points for each value of the propensity score.

Table 5: Average Bias: $\mathbb{E}\left[\hat{\theta}\right] - \theta_0$

	p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5	p = 0.6	p = 0.7	p = 0.8	p = 0.9
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: $\delta_0 = 0.75$, $\delta_1 = 1.5$, $\beta_{00} = \beta_{01} = \beta_{10} = 0.1$, $\beta_{11} = 0.2$									
$\hat{\alpha}(p)$	-0.02	0.01	0.01	-0.01	-0.01	-0.02	-0.02	0.00	0.01
	(0.07)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.05)
$\hat{\Xi}_{OO,0}(p)$	-0.02	-0.02	-0.01	0	-0.01	-0.01	-0.01	-0.01	-0.01
	(0.32)	(0.14)	(0.08)	(0.07)	(0.06)	(0.06)	(0.07)	(0.08)	(0.18)
$\widehat{LB}_1(p)$	0.04	0.06	-0.07	-0.10	-0.12	-0.15	-0.16	-0.14	-0.07
	(0.22)	(0.15)	(0.08)	(0.09)	(0.07)	(0.08)	(0.09)	(0.17)	(0.41)
$\widehat{UB}_1(p)$	0.14	0.04	-0.02	0.04	0.09	0.13	0.14	0.13	0.15
	(0.18)	(0.08)	(0.09)	(0.16)	(0.10)	(0.10)	(0.11)	(0.19)	(0.52)
$\hat{\underline{\Delta}}_2(p)$	0.07	0.08	-0.07	-0.09	-0.11	-0.14	-0.15	-0.14	-0.06
	(0.38)	(0.2)	(0.11)	(0.11)	(0.10)	(0.10)	(0.11)	(0.19)	(0.45)
$\hat{\overline{\Delta}}_2(p)$	0.17	0.05	-0.01	0.04	0.10	0.14	0.15	0.14	0.16
	(0.36)	(0.16)	(0.12)	(0.18)	(0.12)	(0.12)	(0.13)	(0.21)	(0.54)
Panel B: $\delta_0 = 0.2$, $\delta_1 = 2.0$, $\beta_{00} = \beta_{01} = \beta_{10} = 0.1$, $\beta_{11} = 0.2$									
$\hat{\alpha}(p)$	0.00	0.04	0.01	-0.01	-0.01	-0.01	-0.01	0.01	0.02
	(0.09)	(0.05)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)
$\hat{\Xi}_{OO,0}(p)$	0.01	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	0.00
	(0.29)	(0.13)	(0.07)	(0.07)	(0.07)	(0.08)	(0.09)	(0.11)	(0.26)
$\widehat{LB}_1(p)$	0.07	0.02	-0.06	-0.09	-0.11	-0.14	-0.15	-0.12	0.48
	(0.25)	(0.13)	(0.08)	(0.08)	(0.09)	(0.09)	(0.11)	(0.25)	(0.76)
$\widehat{UB}_1(p)$	0.08	-0.02	0.11	0.15	0.18	0.22	0.26	0.35	0.67
	(0.22)	(0.17)	(0.10)	(0.09)	(0.09)	(0.10)	(0.14)	(0.28)	(1.01)
$\hat{\underline{\Delta}}_2(p)$	0.06	0.02	-0.06	-0.09	-0.10	-0.13	-0.14	-0.11	0.48
	(0.39)	(0.18)	(0.11)	(0.10)	(0.11)	(0.11)	(0.13)	(0.26)	(0.81)
$\hat{\overline{\Delta}}_2(p)$	0.07	-0.03	0.12	0.16	0.19	0.23	0.28	0.36	0.67
	(0.36)	(0.2)	(0.12)	(0.11)	(0.12)	(0.12)	(0.16)	(0.3)	(1.04)

Note: We define θ_0 as the true population value of the estimand, $\hat{\theta}$ as the estimator of θ_0 and $\Xi_{OO,0}(p) := \mathbb{E}[Y_0^* | S_0 = 1, S_1 = 1, V = p]$. The results are based on 1,000 Monte Carlo repetitions.

Table 6: Mean Squared Error: $\mathbb{E}\left[n \cdot \left(\hat{\theta} - \theta_0\right)^2\right]$

	p = 0.1 (1)	p = 0.2 (2)	p = 0.3 (3)	p = 0.4 (4)	p = 0.5 (5)	p = 0.6 (6)	p = 0.7 (7)	p = 0.8 (8)	p = 0.9 (9)
Panel A: $\delta_0 = 0.75$, $\delta_1 = 1.5$, $\beta_{00} = \beta_{01} = \beta_{10} = 0.1$, $\beta_{11} = 0.2$									
$\hat{\alpha}(p)$	46.54	10.07	8.72	7.02	7.69	8.89	9.52	7.39	27.1
	(153.91)	(17.2)	(11.09)	(9.97)	(10.81)	(10.62)	(11.11)	(10.32)	(41.41)
$\hat{\Xi}_{OO,0}(p)$	1005.74	185.55	57.02	44.87	41.4	40.3	43.53	60.12	311.93
	(1351.09)	(259.3)	(83.15)	(61.98)	(61.76)	(60.61)	(60.6)	(79.85)	(479.29)
$\widehat{LB}_1(p)$	500.89	275.72	121.81	168.21	194.74	283.95	343.43	502.21	1723.3
	(673.6)	(221.19)	(155.71)	(193.96)	(182.51)	(241.64)	(312.75)	(651.88)	(2463.16)
$\widehat{UB}_1(p)$	536.17	81.04	91.79	282.78	187.07	259.52	321.22	548.54	2956.32
	(1047.25)	(195.52)	(170.81)	(211.73)	(250.02)	(248.53)	(321.19)	(694.02)	(4377.82)
$\hat{\underline{\Delta}}_2(p)$	1517.37	463.07	157.49	203.19	219.32	292.6	347.73	533.7	2075.69
	(2219.03)	(557.22)	(208.91)	(258.66)	(244.93)	(302.02)	(377.33)	(684.14)	(2894.41)
$\hat{\overline{\Delta}}_2(p)$	1539.34	289.92	141.24	328.43	238.21	321.69	392.97	635.88	3180.7
	(2464.02)	(507.75)	(242.44)	(311.5)	(318.06)	(335.83)	(411.57)	(830.93)	(4596.81)
Panel B: $\delta_0 = 0.2$, $\delta_1 = 2.0$, $\beta_{00} = \beta_{01} = \beta_{10} = 0.1$, $\beta_{11} = 0.2$									
$\hat{\alpha}(p)$	87.38	47.18	10.25	8.52	7.62	5.71	4.27	4.73	14.29
	(203.68)	(53.35)	(15)	(12.27)	(11.64)	(8.22)	(6.3)	(6.59)	(20.35)
$\hat{\Xi}_{OO,0}(p)$	832.73	157.01	55.6	48.6	54.85	59.32	74.1	115.4	666.71
	(1192.08)	(226.59)	(81.12)	(71.15)	(80.75)	(86.91)	(119.01)	(171.63)	(979.21)
$\widehat{LB}_1(p)$	667.96	177.78	96.89	147.49	200.67	264.66	350.65	761.64	8061.64
	(740.58)	(332.85)	(123.89)	(170.56)	(217.53)	(269.91)	(416.65)	(1136.95)	(8254.23)
$\widehat{UB}_1(p)$	558.54	279.09	223.32	315.69	413.11	579.47	878.94	2019.95	14733.92
	(1225.8)	(242.89)	(242.07)	(288.42)	(366.06)	(490.23)	(779.26)	(2477.32)	(34135.5)
$\hat{\underline{\Delta}}_2(p)$	1538.87	326.63	150.56	181.28	229.65	285.19	370.49	818	8804.48
	(2203.9)	(504.4)	(213.22)	(231.91)	(286.43)	(347.38)	(478.26)	(1159.35)	(9722.85)
$\hat{\overline{\Delta}}_2(p)$	1337.34	409.26	281.66	379.82	491.03	674.47	1008.06	2186	15299.42
	(2259.27)	(517.54)	(331.17)	(387.45)	(476.41)	(620.89)	(972.28)	(2637.36)	(34933.81)

Note: We define θ_0 as the true population value of the estimand, $\hat{\theta}$ as the estimator of θ_0 and $\Xi_{OO,0}(p) \coloneqq \mathbb{E}[Y_0^* | S_0 = 1, S_1 = 1, V = p]$. The results are based on 1,000 Monte Carlo repetitions. The sample size in each generate data set is $n = 10,000$.