DISCUSSION PAPER SERIES

Identifying Marginal Treatment Effects in the Presence of Sample Selection

Otávio Bartalotti
Désiré Kédagni
Vitor Possebom

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

Identifying Marginal Treatment Effects in the Presence of Sample Selection*

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 derive uniformly sharp bounds on this parameter under three increasingly restrictive sets of assumptions. The first result imposes standard MTE assumptions with an unrestricted sample selection mechanism. The second set of conditions imposes monotonicity of the sample selection variable with respect to treatment, considerably shrinking the identified set. Finally, we incorporate a stochastic dominance assumption which tightens the lower bound for the MTE. Our analysis extends to discrete instruments. The results rely on a mixture reformulation of the problem where the mixture weights are identified, extending Lee's (2009) trimming procedure to the MTE context. We propose estimators for the bounds derived and use data made available by Deb, Munkin, and Trivedi (2006) to empirically illustrate the usefulness of our approach.

Corresponding author:
Otávio Bartalotti
260 Heady Hall
Department of Economics
Iowa State University
Ames, IA, 50011
USA

E-mail: bartalot@iastate.edu

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 effect of a job training program on wages, she has to consider the individuals' decision to enroll in the training program as well as their decision to participate in the labor market. In the health sciences, the same identification challenges appear when analyzing the effect of a drug on well-being as the outcome of interest — health status — is observed only for those who survive. Moreover, in randomized control trials (RCTs), non-compliance and endogenous attrition in treated and control groups lead to the same identification concerns. This double selection problem is also present when analyzing the effect of attending college on wages, the effect of an educational intervention on short- and long-term outcomes, and the effect of procedural laws on litigation outcomes.[1]

In this paper, we derive novel uniformly sharp bounds on the marginal treatment effect (MTE) for individuals who would self-select into the sample regardless of their treatment status (MTE^{OO}). To do so, we propose identification strategies under increasingly restrictive sets of assumptions that extend the MTE identification to scenarios with endogenous sample selection. Furthermore, the choice of treatment is allowed to be endogenous, and can be related to the sample selection mechanism. To address both identification challenges described above, 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 (employment status), and the treatment choice (training program participation) is observed for all individuals in the data being analyzed.

The MTE and MTE^{OO} provide important and intuitive measures of the treatment effect

---

[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). 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). RCTs with attrition are illustrated by DeMel, McKenzie, and Woodruff (2013) and Angelucci, Karlan, and Zinman (2015).

and its heterogeneity across the population. For example, consider a job training program. Training influences both the likelihood of employment and wages, which are observed only for individuals who are employed. The MTE reflects the returns to training for individuals with different levels of the (latent) cost of attending the program. As a consequence, this parameter sheds light on the heterogeneity of the training program's effects, i.e., to understand who would benefit from taking extra training. This knowledge can be used to design policies focusing on targeting of the program, affordability and services offered. Common parameters evaluated in the literature — such as the average treatment effect (ATE), the average treatment effect on the treated (ATT), the average treatment effect on the untreated (ATU), and the local average treatment effect ($LATE$) — may not adequately capture this heterogeneity. For example, they could be positive even when most people are affected adversely by a policy, masking its effects.[2] Similarly, the MTE^{OO} reflects the training program's effects at the intensive margin, i.e., to the group of individuals that are more attached to the labor force, and would be employed even if they had not attended the training program.

Our strategy to partially identify the MTE^{OO} relies on standard assumptions regarding selection into treatment. Those are the same conditions usually imposed for identification of the MTE when there is no sample selection, i.e., there is an exogenous instrument excluded from the outcome determination, treatment choice is monotone in the instrument, and the propensity score is continuous.[3] We add the requirement that the instrument is also excluded from the sample selection mechanism. In the job training literature, a similar assumption is frequently imposed when analyzing RCTs with imperfect compliance, as the interest frequently lies on the effect on labor earnings and employment.

In this paper's first contribution, the partial identification result for the MTE^{OO} leaves sample selection unrestricted, providing very general uniformly sharp bounds on that parameter. Our second result tightens the bounds around the MTE^{OO} by exploiting a "monotonicity in selection" assumption. This condition is similar to the usual LATE monotonicity assump-

---

[2] The empirical relevance of the MTE is recently illustrated by Carneiro, Heckman, and Vytlacil (2011), Kline and Walters (2016), Arnold, Dobbie, and Yang (2018), Cornelissen et al. (2018), Bhuller et al. (2019), Humphries et al. (2019) and Mountjoy (2019) in different contexts.

[3] See Bjorklund and Moffitt (1987), Heckman and Vytlacil (1999, 2001a, 2005) and Andresen (2018) for a detailed discussion.

tion and requires individuals to be at least as likely to be observed in the sample if they are treated. In the job training program example, this additional assumption imposes that the treatment can induce workers to join the labor force, but not the opposite.

The identified set is further reduced by imposing a stochastic dominance assumption to our third and final set of assumptions. This condition mandates that the subpopulation who self-selects into the sample regardless of the treatment status has higher potential outcomes if treated than the subpopulation who self-selects into the sample only when treated. Intuitively, workers with high attachment to the labor force (that would be employed regardless of receiving job training), would earn higher wages after the extra training than those that would choose to participate in the labor force only if they participated in the training program.

Identification relies on a reformulation of the potential outcomes' conditional probabilities 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 ATE and 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.

The results can be used to construct bounds for any treatment effect parameter that can be written as a weighted average of the MTE^{OO}. We derive new weights to obtain sharp bounds on the ATE, the ATT, the ATU, any LATE (Imbens and Angrist, 1994) and any policy-relevant treatment effect (PRTE, Heckman and Vytlacil, 2001b) within the always-observed subpopulation. Differently from the weights for the case without sample selection (Heckman and Vytlacil, 2005, Carneiro and Lee, 2009, and Carneiro, Heckman, and Vytlacil, 2011), the new weights must be integrated over the distribution of the latent heterogeneity for the always-observed subpopulation instead of its unconditional distribution. Furthermore, we show these weights are identified under the "monotonicity in selection" assumption if the support of the propensity score is the full unit interval.

Moreover, we propose and discuss nonparametric and parametric estimators for the bounds

around the MTE^{OO}. The usefulness and feasibility of the parametric procedure is illustrated in an empirical example using the data set organized by Deb, Munkin, and Trivedi (2006). We find that the effect of insurance plan choice on ambulatory expenditures depends negatively on the agents' relative cost of choosing managed care plans over fee for service plans. This corroborates the results found by Deb, Munkin, and Trivedi (2006), indicating that agents endogenously self-select into managed care plans.

The main identification results are extended to the case where the researcher only has access to multi-valued discrete instruments. In this context, we derive nonparametric sharp bounds on a weighted average of the MTE. This result is conceptually similar to Chen and Flores (2015), who provide an outer set for the LATE parameter when the instrument is binary.[4]

This paper is organized as follows. Section 2 presents the structural model and sample selection mechanism considered, followed by a discussion of the identifying assumptions. In Section 3, we provide the identification results for the MTE bounds in the case of a continuous instrument under each set of assumptions while Section 4 discusses general aspects of estimation for the bounds. Section 5 illustrates an estimation procedure and the identifying power of each set of assumptions using data made available by Deb, Munkin, and Trivedi (2006). Section 6 presents identification results for the case of discrete instruments. Section 7 concludes. The proofs, sharp testable implications of the model, extensions to DMTE, numerical example, proposed estimation methods' details, Monte Carlo simulations for the estimator's performance, economic models illustrating our identifying assumptions, and an alternative specification for our empirical example are presented in the appendix.

---

[4]An additional extension is presented in Appendix H, where we derive sharp bounds around a more general object of interest: the distributional marginal treatment effect (DMTE), which captures the effect of the treatment on the outcome's distribution for individuals at the margin for participation. The DMTE can then be used to derive bounds on the quantile version of the marginal treatment effect and many other parameters.

2 Analytical Framework

Following 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., random assignment of cash incentives for participation in a job training program) with support given by $\mathcal{Z} \subset \mathbb{R}^{d_z}$, D is the treatment status indicator (job training program enrollment). The variable Y^* is the possibly censored realized outcome variable (wages) with support $\mathcal{Y} \subset \mathbb{R}$, while Y_0^* and Y_1^* are the possibly censored 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 uncensored observed outcome (labor earnings), and V represents unobserved individual characteristics (cognitive or social costs of attending the job training program). The researcher observes only the vector (Y, D, S, Z), while Y_1^*, Y_0^*, S_1, S_0 and V are latent variables.[5] This model is a generalization of that considered in Heckman and Vytlacil (1999, 2001a, 2005) to the sample selection setting.

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 receives treatment when its idiosyncratic cost V is less than or equal to a threshold $P(Z)$. This assumption is equivalent to imposing monotonicity of the treatment in the instrument Z (Imbens and Angrist, 1994) as shown by Vytlacil (2002). This setup is similar

---

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

to the one proposed by Heckman and Vytlacil (2005) and leads to the definition of the MTE,

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

for any $p \in [0, 1]$.

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), even in the simpler case of the ATE, point identification is no longer possible even if the treatment is randomly assigned, leading him to derive 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 define four latent groups based on the potential sample selection indicators. The subpopulations are defined as: always-observed ($S_0 = 1, S_1 = 1$), observed-only-when-treated ($S_0 = 0, S_1 = 1$), observed-only-when-untreated ($S_0 = 1, S_1 = 0$), and never-observed ($S_0 = 0, S_1 = 0$).[6] Those subgroups are summarized in Table 1.

Table 1: Subgroups based on the sample selection status

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 always-observed subpopulation ($S_0 = 1, S_1 = 1$). Importantly, this subpopulation is the only group whose censored potential outcomes are observed in both treatment arms. For the other three subpopulations, treatment effect parameters are not point-identified or bounded in a

---

[6]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 which the four latent groups would be called strata.

non-trivial way without further parametric assumptions, since at least one of the potential outcomes (Y_0^* or Y_1^*) is never observed.[7] Since our focus is on a fully non-parametric identification strategy, we do not discuss parametric identification of unconditional treatment effect parameters or for treatment effect parameters associated with the latent groups ON, NO and NN.

Our target parameter is the MTE function for the subpopulation who is always observed ($MTE^{OO} \colon [0,1] \to \mathbb{R}$):

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

for any $p \in [0,1]$. Note that this parameter captures the intensive margin of the treatment effect.[8] For example, when evaluating the effect of a job training program on wages, this parameter captures the effect of the training program on the wage of a worker who is employed regardless of her treatment status.[9] In medical applications where selection is due to the death of a patient, this parameter 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, this parameter is the effect on test scores for the subpopulation who does not drop out of school in any case. Moreover, the definition of the MTE^{OO} does not depend on the instrument being used, implying that our target parameter is policy-invariant. This feature is an advantage in comparison with the $LATE^{OO}$ (Chen and Flores, 2015), which is not policy-invariant. Although the MTE^{OO} function shares the policy invariance property with the usual MTE function, it has one important drawback: the definition of MTE^{OO}

---

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

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

[9]This conditional parameter is interesting for a policy maker and can offer meaningful information regarding treatment and its targeting. For example, the effect of a training program on individuals' wages reflects their productivity and the overall economic welfare. If a training program only affects overall earnings by attracting more participants to the labor market without any effect on productivity, the program's targeting and overall benefit to society should consider this aspect.

conditions on a latent group $(S_0 = 1, S_1 = 1)$ to simultaneously address the selection-into-treatment and sample selection problems.

Analogously to Lee (2009), identification of MTE^{OO} is complex because sample selection is nonrandom and possibly impacted by the treatment. To address this issue, we consider three sets of increasingly restrictive assumptions that allow partial identification of the target parameter, shrinking the identified sets as the assumptions are strengthened.[10]

Assumptions 1-5 are sufficient to partially identify the MTE^{OO} function.

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 an interval $\mathcal{P} := [\underline{p}, \overline{p}] \subseteq [0,1]$.[11]*

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 $p \in \mathcal{P}$.*

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

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

Assumption 1 is a modification of the IV independence assumption to account for sample selection. Instead of assuming that the instrument is independent of the latent heterogeneity and of the potential outcomes only (Heckman and Vytlacil, 2005), we also assume independence of the potential sample selection indicators. Intuitively, we rely on changes in Z shifting

---

[10]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).

[11]The assumption that $\mathcal{P}$ is a interval is made for notational simplicity. All the proofs can be easily extended to case when $\mathcal{P}$ is a set with a non-empty interior.

treatment status and, hence, sample participation to identify the marginal treatment effect bounds. Such an assumption is common in the empirical literature. A researcher interested in analyzing the impact of a job training program on wages and employment status, usually imposes a similar restriction on the marginal distributions of labor earnings and employment.

Assumption 2 is important for our bounding strategy of the MTE across values of the latent variable, V. In Section 6, we relax this assumption to allow for discrete instruments, and we show how our methodology can be used to bound the LATE, instead of the MTE.

Assumptions 3-5 are technical assumptions to ensure that the objects of interest are well-defined and are common in the literature about MTE (Heckman, Urzua, and Vytlacil, 2006). 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 the calculation of the MTE bounds 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 partial 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 $\mathbb{P}[D = 1|Z = z]$ (Heckman and Vytlacil, 2005, p. 677). Indeed, $\mathbb{P}[D = 1|Z = z] = \mathbb{P}[V \leq P(z)|Z = z] = \mathbb{P}[V \leq P(z)] = P(z)$, where the second equality holds under Assumption 1 and the last holds under Assumption 5. This first set of restrictions partially identifies 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 in 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 sample selection literature (Lee, 2009, Chen and

Flores, 2015).[12] To obtain some intuition on the mechanisms behind this assumption, consider the job training program example. An individual is employed when her job search skills $\vartheta(D)$, a function of training take-up, are above a threshold U_S so that $S = \mathbb{1}\{\vartheta(D) \geq U_S\}$. Additionally, suppose that attending the job training program does not decrease someone's job search skills, i.e., $\vartheta(1) \geq \vartheta(0)$, making it more likely that program's trainees would be observed in the data. In such a case, Assumption 6 holds. However, if attending the job training program raises the agents' reservation wages or if the lost labor market experience is very costly in terms of job finding, 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 Proposition 2. 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

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

In other words, the share of the population for which the outcome is observed rises with p. We discuss further testable implications in Section 3, and formally characterize sharp testable implications arising from those assumptions in Appendix B. In the job training example, this testable implication means that the likelihood of employment increases with the probability of attending the training program.

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 treated for the always-observed subpopulation first-order stochastically dominates the distribu-*

---

[12] 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.

tion of the same random variable for the observed-only-when-treated subpopulation, i.e.,

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

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.[13] Naturally, the plausibility of Assumption 7 depends on the empirical context. In some cases this stochastic dominance assumption could be hard to interpret and motivate empirically. However, this challenge can be confronted constructively in a layered policy analysis (Manski, 2011), since we can offer a menu of estimates based on different assumptions, allowing the researcher to understand the continuum of information that we can gather about a specific economic parameter, as advocated by Tamer (2010). In Appendix D, we provide a simple economic model related to our job training example to illustrate that Assumption 7 may hold under plausible economic restrictions.

Remark 1. *Point identification of MTE^{OO} is achieved if, in addition to assumptions 1-5, we assume that the always-observed and never-observed subpopulations are the only existing groups, i.e., $S_0 = S_1$. This "no selection effect" assumption imposes that the treatment has no impact on sample selection. This set of assumptions has a testable implication: $\frac{\partial \mathbb{P}[S=1|P(Z)=p]}{\partial p} = 0$ for all $p \in \mathcal{P}$. In the job training program context, "no selection" implies that workers' employment status would not be affected by program take-up and the likelihood of employment does not depend on the probability of attending the training program.*

---

[13]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.

3 Identification Results

This section presents the main results of this paper, the identification for $MTE^{OO}(p)$ under the three different sets of assumptions described in Section 2. As stepping stones, Subsection 3.1 shows identification of the conditional joint distribution of $(Y_d^*, S_d = 1)|V$ 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 monotone sample selection (Subsection 3.4) and stochastic dominance (Subsection 3.5). For completeness, we also show that the MTE^{OO} is point-identified under the "no selection effect" assumption (Remark 1) at the end of Subsection 3.5. Finally, in Subsection 3.6, 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 the MTE^{OO}, we point-identify the conditional joint distribution of each potential outcome and sample selection for different levels of individual heterogeneity, $(Y_d^*, S_d = 1)|V$ 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

$$
\begin{aligned}
\mathbb{P}\left[Y \in A, S = 1, D = 1 | P(Z) = p\right] &= \mathbb{P}\left[Y_1^* \in A, S_1 = 1, V \le p | P(Z) = p\right] \\
&= \mathbb{P}\left[Y_1^* \in A, S_1 = 1, V \le p\right] \\
&= \mathbb{P}\left[Y_1^* \in A, S_1 = 1 | V \le p\right] \mathbb{P}\left[V \le p\right] \\
&= \left(\int_0^p \mathbb{P}\left[Y_1^* \in A, S_1 = 1 | V = v\right] \frac{f_V(v)}{\mathbb{P}\left[V \le p\right]} dv\right) \cdot \mathbb{P}\left[V \le p\right] \\
&= \int_0^p \mathbb{P}\left[Y_1^* \in A, S_1 = 1 | V = v\right] dv,
\end{aligned}
$$

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 differentiating each side with respect to p, we point-identify the conditional distribution of $(Y_1^*, S_1 = 1)$ given $V = p$ as follows:

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

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 since equations (3.1) and (3.2) reflect probabilities, they 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, for people with observable characteristics (Z) that indicate a higher likelihood of being treated, the share of treated (untreated) individuals that self-select into the sample increases (decreases) for any range of the outcome.[14]

Similarly to the local IV 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 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

---

[14] A formal characterization of the sharp testable implications implied by our model is given in Appendix B.

14

levels. We can also identify the MTE on the observed outcome,

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

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 considering 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 observed 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:

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

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

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

Similarly, the conditional distribution of $Y_0^* | S_0 = 1, V = p$ is the mixture of $Y_d^* | V = p$

for two latent groups, the always-observed and the observed-only-when-untreated group:

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

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), but, differently from them, the trimmed distribution is conditional on a specific value for the latent heterogeneity variable. This type of trimming approach, used in the current and above papers, relies on results derived in a more general mixture model by Horowitz and Manski (1995).

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

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

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

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

Similarly, the conditional distribution of $Y_0^*\middle| S_0 = 1, V = p$ can be written as the mixture of $Y_d^*\middle| V = p$ 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 ($\mathbb{E}\left[Y_0^*|\,S_0=1,S_1=1,V=p\right]$) lies within the interval $[LB_0(p),UB_0(p)]$, where

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

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

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

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

Remark 2. *The issue central to identification of the target parameter is what can be learned about the mixture weights $(\alpha(p),\beta(p))$, and $\mathbb{E}\left[Y_d^*|\,S_d=1,V=p\right]$. 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 $\mathbb{E}\left[Y_1^*|\,S_0=1,S_1=1,V=p\right]$ 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 the 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, $\mathbb{P}\left[S_0 = 1, S_1 = 1\right]$ 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:

$$\mathbb{P}\left[S_0 = 1, S_1 = 1|V = p\right] \in \left[\max\left\{\mathbb{P}\left[S_0 = 1|V = p\right] + \mathbb{P}\left[S_1 = 1|V = p\right] - 1, 0\right\},\right.$$

$$\left. \min\left\{\mathbb{P}\left[S_0 = 1|V = p\right], \mathbb{P}\left[S_1 = 1|V = p\right]\right\}\right]. \qquad (3.12)$$

Combining this information with Equations (3.5) and (3.6), leads to Lemma 1.

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

$$\mathbb{P}\left[S_0 = 1, S_1 = 1|V = p\right]$$

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

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

$$=: \Upsilon\left(p\right).$$

These bounds are sharp.

Note that, since $\Upsilon\left(p\right)$ 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 $\mathbb{P}\left[S_0 = 1, S_1 = 1|V = p\right]$ take any particular value, $\upsilon \in \Upsilon\left(p\right)$. Define,

$$\alpha\left(p, \upsilon\right) := \mathbb{P}\left[S_0 = 1|S_1 = 1, V = p\right] = \frac{\upsilon}{\mathbb{P}\left[S_1 = 1|V = p\right]} = \frac{\upsilon}{\frac{\partial\mathbb{P}\left[S = 1, D = 1|P(Z) = p\right]}{\partial p}},$$

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

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:

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

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

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

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

Combining the bounds around $\mathbb{E}\left[Y_1^* | S_0 = 1, S_1 = 1, V = p\right]$ and $\mathbb{E}\left[Y_0^* | S_0 = 1, S_1 = 1, V = p\right]$, 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 $\mathbb{P}\left[S_0 = 1, S_1 = 1 | V = p\right] = \upsilon$.

To bound the target parameter, we find worst- and best-case scenarios 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, which need only to be evaluated at the end point 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, i.e.,*

$$\underline{\Delta}_1(p) \le MTE^{OO}(p) \le \overline{\Delta}_1(p)$$

for any $p \in \mathcal{P}$, where $\underline{\Delta}_1 \colon \mathcal{P} \to \mathbb{R}$ and $\overline{\Delta}_1 \colon \mathcal{P} \to \mathbb{R}$ are given by

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

for any $p \in \mathcal{P}$, the conditional distribution of $\tilde{Y}_d$ is given by

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

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

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

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

Moreover, these bounds are uniformly sharp.

Remark 3. *The definition of uniform sharpness (Firpo and Ridder, 2019) used in this article states that, for any function $\delta\colon \mathcal{P} \to \mathbb{R}$ such that $\delta\left(p\right) \in \left[\underline{\Delta}_1\left(p\right), \overline{\Delta}_1\left(p\right)\right]$ for any $p \in \mathcal{P}$, it is possible to construct random variables $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ that satisfy all the restrictions imposed on the data by the underlying assumptions, induce the joint distribution on the data (Y, S, D, Z) and achieve the candidate target parameter $\delta\left(p\right) = \mathbb{E}\left[\tilde{Y}_1^* - \tilde{Y}_0^* \middle| \tilde{S}_0 = 1, \tilde{S}_1 = 1, \tilde{V} = p\right]$ for every $p \in \mathcal{P}$.*

Remark 4. *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 ATE for the always-observed. Then, the bounds derived in Proposition 1 simplify to those derived by Imai (2008, Proposition 1).*

3.4 Bounds under the Monotonicity Assumption

In this subsection, we introduce monotonicity of sample selection in the treatment (Assumption 6), which can considerably shrink 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, 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 no longer need to rely on the partial identification results in Lemma 1. Specifically, we have that

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

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

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. 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 expected 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 in the interval $[LB_1(p), UB_1(p)]$, given in Equations (3.8)-(3.9). Combining the bounds and identification results in equations (3.1), (3.2), (3.5), (3.13) and (3.14), the following proposition holds:

Proposition 2. *Under Assumptions 1-6, the MTE is partially identified for the always-observed, i.e.,*

$$\underline{\Delta}_2\left(p\right) \leq MTE^{OO}\left(p\right) \leq \overline{\Delta}_2\left(p\right)$$

for any $p \in \mathcal{P}$, where $\underline{\Delta}_2 \colon \mathcal{P} \to \mathbb{R}$ and $\overline{\Delta}_2 \colon \mathcal{P} \to \mathbb{R}$ are given by

$$\underline{\Delta}_2\left(p\right) := \mathbb{E}\left[\tilde{Y}_1 | S = 1, D = 1, P(Z) = p, \tilde{Y}_1 \leq F_{\tilde{Y}_1 | S=1, D=1, P(Z)=p}^{-1}\left(\alpha(p)\right)\right]$$

22

$$- \mathbb{E}\left[\tilde{Y}_0 | S = 1, D = 0, P(Z) = p\right],$$

$$\overline{\Delta}_2\left(p\right) := \mathbb{E}\left[\tilde{Y}_1 | S = 1, D = 1, P(Z) = p, \tilde{Y}_1 > F^{-1}_{\tilde{Y}_1 | S=1, D=1, P(Z)=p}\left(1 - \alpha(p)\right)\right]$$

$$- \mathbb{E}\left[\tilde{Y}_0 | S = 1, D = 0, P(Z) = p\right]$$

for any $p \in \mathcal{P}$, 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

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

Moreover, these bounds are uniformly sharp.

Remark 5. *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.*[15] *This improvement shows the identifying power of Assumption 6.*

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 in Proposition 2 simplify to the ones proposed by Lee (2009, Proposition 1a).*

3.5 Bounds under the Monotonicity and Dominance Assumptions

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

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

---

[15]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.

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

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

The upper bound remains unchanged. Naturally, that leads to tighter identified sets relative to the ones in Proposition 2, which are presented in the following proposition.

Proposition 3. *Under Assumptions 1-7, the MTE is partially identified for the always-observed, i.e.,*

$$\underline{\Delta}_3\left(p\right) \leq MTE^{OO}\left(p\right) \leq \overline{\Delta}_3\left(p\right)$$

for any $p \in \mathcal{P}$, where $\underline{\Delta}_3\left(p\right) : \mathcal{P} \to \mathbb{R}$ and $\overline{\Delta}_3\left(p\right) : \mathcal{P} \to \mathbb{R}$ are given by

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

for any $p \in \mathcal{P}$, 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 uniformly sharp.

Remark 7. *The lower bound proposed here is strictly greater than the one in Proposition 2 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 8. *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 in Proposition 3 simplify to those proposed in Imai (2008, Equation (8)).*

For completeness in the identification discussion, note that point identification of MTE^{OO} is achieved if, in addition to assumptions 1-5, we assume that the always-observed and never-observed subpopulations are the only existing groups, i.e., $S_0 = S_1$. This "no selection effect" assumption imposes that the treatment has no impact on sample selection. In the job training program context, this implies that workers' employment would not be affected by the program.

Under Assumptions 1-5 and "no selection effect", the distributions of $Y_d^*|S_1 = 1, V$ for $d \in \{0,1\}$ are exclusively composed of always-observed individuals ($\alpha(p) = \beta(p) = 1$). Then,

$$\mathbb{P}[Y_d^* \in A|S_0 = 1, S_1 = 1, V = p] = \frac{\mathbb{P}[Y_d^* \in A, S_d = 1|V = p]}{\mathbb{P}[S_0 = 1|V = p]} \text{ for } d = \{0,1\},$$

where point-identification follows from equations (3.1), (3.2), (3.5) and (3.6). The expectations $\mathbb{E}[Y_d^*|S_0 = 1, S_1 = 1, V = p]$ are also point-identified. Then, $MTE^{OO}(p) = \frac{\frac{\partial \mathbb{E}[YS|P(Z)=p]}{\partial p}}{\frac{\partial \mathbb{E}[SD|P(Z)=p]}{\partial p}}$.

Alternatively, point-identification of the unconditional MTE is achieved if potential sample selection status and outcomes are independent given the unobserved characteristics, $(S_0, S_1) \perp\!\!\!\perp (Y_0^*, Y_1^*)|V$. In this case, the distributions $\mathbb{P}[Y_d^* \leq y|V = p]$ are point-identified from Equations (3.1) and

(3.2), and the unconditional MTE is point-identified as:

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

3.6 Empirical Relevance of Bounds for the $\underline{\text{MTE}^{\text{OO}}}$

The partial identification results for the MTE^{OO} are relevant for a vast array of empirical objectives. First, bounds for the MTE^{OO} can illuminate the treatment effect's heterogeneity, allowing researchers to better understand who would benefit from a specific treatment. This feature 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 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} because

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

where $t \in \{1, 2, 3\}$, $\underline{\Delta}_t$ and $\overline{\Delta}_t$ are described in Propositions 1-3, and $\omega(\cdot)$ is a known or identifiable weighting function. Those bounds are sharp, as summarized in Proposition 4.[16]

Proposition 4. *Let $\omega_{TE} \colon \mathcal{P} \to \mathbb{R}$ be a known or identifiable weighting function and define the treatment effect parameter $TE := \int_{\mathcal{P}} MTE^{OO} \left(p \right) \cdot \omega_{TE} \left(p \right) \, dp$. Under Assumptions 1-6, the treatment effect parameter TE is partially identified:*

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

Moreover, these bounds are sharp.

Under Assumptions 1-6, Table 2 shows the most relevant treatment effect parameters that

---

[16]We focus on the case under the monotonicity restriction (Assumptions 1-6) for brevity. Similar results hold under our other identifying sets of assumptions.

are partially identified using Proposition 4. The weights in Table 3 are derived in Appendix A.6 and are identified if the support of the propensity score is the full unit interval, i.e., $\mathcal{P} = [0, 1]$.

Table 2: Treatment Effects as Weighted Integrals of the MTE^{OO}

$$ATE^{OO} = \mathbb{E}\left[Y_1^* - Y_0^* \,|\, S_0 = 1, S_1 = 1\right] = \int_0^1 MTE^{OO}(p) \cdot \omega_{ATE}(p)\, \mathrm{d}p$$

$$ATT^{OO} = \mathbb{E}\left[Y_1^* - Y_0^* \,|\, D = 1, S_0 = 1, S_1 = 1\right] = \int_0^1 MTE^{OO}(p) \cdot \omega_{ATT}(p)\, \mathrm{d}p$$

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

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

$$PRTE^{OO} = \frac{\mathbb{E}\left[Y_a^* - Y_{a'}^* \,|\, S_0 = 1, S_1 = 1\right]}{\int_0^1 \left(F_{P_{a'}}(p) - F_{P_a}(p)\right) \cdot f_{V|S_0=1,S_1=1}(p)\, \mathrm{d}p} = \int_0^1 MTE^{OO}(p) \cdot \omega_{PRTE}(p, a, a')\, \mathrm{d}p$$

for two policies a and a' that affect only Z (See notational details in Appendix A.6.)

Importantly, note that, differently from the weights for the case without sample selection (Heckman and Vytlacil, 2005, Carneiro and Lee, 2009, and Carneiro, Heckman, and Vytlacil, 2011), these weights must be integrated over the distribution of the latent heterogeneity for the always-observed subpopulation instead of its unconditional distribution.

4 Estimation

This section describes the general estimation steps for the bounds proposed in Proposition 2, while details on two proposed methods are provided in Appendix E and G. For brevity, we focus on the bounds identified under monotonicity of sample selection in the treatment (Assumptions 1-6), as it is the most relevant (and feasible) case empirically. Estimators for the bounds in Proposition 1 and Proposition 3 are natural extensions. Appendix F presents a Monte Carlo Simulation that evaluates the small sample properties of the estimator.

To estimate the bounds in Proposition 2, it is useful to focus on the building blocks that are the foundation for the identification results. In particular, we need the CDFs:

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

$$\omega_{ATE}\left(p\right) = \frac{\dfrac{\partial \mathbb{P}\left[S = 1, D = 0\mid P\left(Z\right) = p\right]}{\partial p}}{\displaystyle\int_{0}^{1} \dfrac{\partial \mathbb{P}\left[S = 1, D = 0\mid P\left(Z\right) = p\right]}{\partial p}\,\mathrm{d}p}$$

$$\omega_{ATT}\left(p\right) = \frac{\left(\int_{p}^{1} f_{P(Z)}\left(u\right)\,\mathrm{d}u\right) \cdot \dfrac{\partial \mathbb{P}\left[S = 1, D = 0\mid P\left(Z\right) = p\right]}{\partial p}}{\displaystyle\int_{0}^{1} \left(\int_{p}^{1} f_{P(Z)}\left(u\right)\,\mathrm{d}u\right) \cdot \dfrac{\partial \mathbb{P}\left[S = 1, D = 0\mid P\left(Z\right) = p\right]}{\partial p}\,\mathrm{d}p}$$

$$\omega_{ATU}\left(p\right) = \frac{\left(\int_{0}^{p} f_{P(Z)}\left(u\right)\,\mathrm{d}u\right) \cdot \dfrac{\partial \mathbb{P}\left[S = 1, D = 0\mid P\left(Z\right) = p\right]}{\partial p}}{\displaystyle\int_{0}^{1} \left(\int_{0}^{p} f_{P(Z)}\left(u\right)\,\mathrm{d}u\right) \cdot \dfrac{\partial \mathbb{P}\left[S = 1, D = 0\mid P\left(Z\right) = p\right]}{\partial p}\,\mathrm{d}p}$$

$$\omega_{LATE}\left(p\right) = \frac{\dfrac{\partial \mathbb{P}\left[S = 1, D = 0\mid P\left(Z\right) = p\right]}{\partial p}}{\displaystyle\int_{\underline{p}}^{\overline{p}} \dfrac{\partial \mathbb{P}\left[S = 1, D = 0\mid P\left(Z\right) = p\right]}{\partial p}\,\mathrm{d}p}$$

$$\omega_{PRTE}\left(p, a, a'\right) = \frac{\left(F_{P_{a'}}\left(p\right) - F_{P_{a}}\left(p\right)\right) \cdot \dfrac{\partial \mathbb{P}\left[S = 1, D = 0\mid P\left(Z\right) = p\right]}{\partial p}}{\displaystyle\int_{0}^{1} \left(F_{P_{a'}}\left(p\right) - F_{P_{a}}\left(p\right)\right) \cdot \dfrac{\partial \mathbb{P}\left[S = 1, D = 0\mid P\left(Z\right) = p\right]}{\partial p}\,\mathrm{d}p}$$

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

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

Consequently, we need to estimate:

$$\Gamma_1\left(p, y\right) := \frac{\partial \mathbb{P}\left[Y \leq y, S = 1, D = 1\mid P(Z) = p\right]}{\partial p}, \qquad \pi_1\left(p\right) := \frac{\partial \mathbb{P}\left[S = 1, D = 1\mid P(Z) = p\right]}{\partial p},$$

$$\Gamma_0\left(p, y\right) := -\frac{\partial \mathbb{P}\left[Y \leq y, S = 1, D = 0\mid P(Z) = p\right]}{\partial p}, \qquad \pi_0\left(p\right) := -\frac{\partial \mathbb{P}\left[S = 1, D = 0\mid P(Z) = p\right]}{\partial p}.$$

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

Each of these components can be estimated by standard approaches, and multiple procedures might be valid depending on the assumptions and model structure the researcher is willing to impose. For example, one could resort to nonparametric methods to estimate $P(Z)$, $\pi_0(p)$, $\pi_1(p)$, $\Gamma_0(p,y)$, and $\Gamma_1(p,y)$, avoiding functional form choices as discussed in Appendix E. While appealing, nonparametric estimation can be quite challenging in practice especially when covariates are added, as illustrated in the empirical application. Alternatively, we could estimate the parameters based on parametric functions for $P(Z)$, $\mathbb{P}\left[S=1, D=d \mid P(Z)=p\right]$, and a partition of the outcome's support $\gamma_d(p,k) = \mathbb{P}\left[y_{k-1} \leq Y < y_k, S=1, D=d \mid P(Z)=p\right]$ for $d = \{0,1\}$, $p \in [0,1]$, $k \in \{1, \ldots, K\}$. The choice of estimator should be guided by the nature of the problem being studied and the data available to the researcher.

With estimates $\hat{P}(Z)$, $\hat{\pi}_0(p)$, $\hat{\pi}_1(p)$, $\hat{\Gamma}_d(p,y)$, and $\hat{\gamma}_d(p,y)$ at hand, we can estimate $\alpha(p)$, by its sample analog $\hat{\alpha}(p) := \dfrac{\hat{\pi}_0(p)}{\hat{\pi}_1(p)}$. Finally, the estimated bounds $LB_2(p)$, $UB_2(p)$, can be obtained as

$$\widehat{LB}_2(p) := \sum_{k=2}^{K_N} \overline{y}_k \cdot \mathbb{1}\left\{\hat{F}_1(p, y_k) \leq \hat{\alpha}(p)\right\} \cdot \frac{\hat{f}_1(p,k)}{\hat{\alpha}(p)} \tag{4.1}$$

$$\widehat{UB}_2(p) := \sum_{k=2}^{K_N} \overline{y}_k \cdot \mathbb{1}\left\{1 - \hat{F}_1(p, y_k) < \hat{\alpha}(p)\right\} \cdot \frac{\hat{f}_1(p,k)}{\hat{\alpha}(p)}, \tag{4.2}$$

where $\hat{f}_1(p,k) := \frac{\hat{\gamma}_1(p,k)}{\hat{\pi}_1(p)}$, $\hat{F}_1(p, y_k) := \sum_{j=2}^{k} \hat{f}_1(p,j)$ and $\overline{y}_k$ is the center point of each bin $[y_{k-1}, y_k]$ for any $k \in \{2, \ldots, K_N\}$. Moreover, we can estimate $\mathbb{E}\left[\tilde{Y}_0 \mid S=1, D=0, P(Z)=p\right]$ in Proposition 2 using $\hat{\Xi}_{OO,0}(p) := \sum_{k=2}^{K_N} \overline{y}_k \cdot \hat{f}_0(p,k)$, where $\hat{f}_0(p,k) := \frac{\hat{\gamma}_0(p,k)}{\hat{\pi}_0(p)}$. Naturally, the estimated MTE^{OO} bounds can then be obtained by $\hat{\underline{\Delta}}_2(p) := \widehat{LB}_2(p) - \hat{\Xi}_{OO,0}(p)$ and $\hat{\overline{\Delta}}_2(p) := \widehat{UB}_2(p) - \hat{\Xi}_{OO,0}(p)$.

We summarize the estimation procedure in the following steps:

STEP 1. Estimate $\mathbb{P}(D=1 \mid Z=z)$ 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.

29

STEP 4. Implement $\widehat{LB}_2(p)$, $\widehat{UB}_2(p)$ and $\hat{\Xi}_{OO,0}(p)$.

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

5 Empirical Illustration: Managed Care and Ambulatory Expenditures

To illustrate the empirical usefulness of our partial identification strategy in a concrete application, we analyze the impact of insurance plan choice on ambulatory expenditures using the data set made available by Deb, Munkin, and Trivedi (2006) through the Journal of Applied Econometrics' Data Archive.[17] We simplify their analyzes in two important dimensions. To enforce a binary treatment, we follow Papadoulos and Silva (2012) and combine Health Maintenance Organizations (HMO) and Preferred Provider Organization (PPO) in one treatment category (managed care) while the control group contains all individuals who choose fee-for-service (FFS) plans. Second, while Deb, Munkin, and Trivedi (2006) analyze ambulatory and hospital expenditures, we focus only on the former since the large share of zero hospital expenditures (Deb, Munkin, and Trivedi, 2006, Table 1) imply that bounds around the MTE^{OO} would be very wide as explained in Appendix C.

In this application, the treatment variable D is equal to one if the person chooses a managed care plan and equal to zero if the person chooses an FFS plan. Y_0^* and Y_1^* represent the potential ambulatorial expenditures in dollars, that is only observed if the person seeks care. S_0 and S_1 represent the potential indicators for seeking care, i.e., for spending a positive amount of money on ambulatory services. The latent heterogeneity V in our treatment choice model (Equation (2.1)) can be interpreted as the individual's relative cost of choosing a managed care plan over FFS plan. Importantly, our treatment effect of interest $\left(MTE^{OO}(p) \coloneqq \mathbb{E}\left[Y_1^* - Y_0^* \mid V = p, S_0 = 1, S_1 = 1\right]\right)$ captures the intensive margin of the impact of managed care plans on ambulatory expenditures, that is, the increased intensity in use of services by those individuals that would seek ambulatory care in both insurance scenarios.

We use data from the 1996-2001 waves of the Medical Expenditure Panel Survey (MEPS) made available by Deb, Munkin, and Trivedi (2006). The sample is restricted to employed

---

individuals who bought a private insurance plan and whose age is between 21 and 64 years. In this data set, we observe 24 covariates.[18] Moreover, our instruments are spouse's age (Z_1) and spouse's insurance plan type in the previous year (Z_2).[19] The discussion about instruments' validity follows Deb, Munkin, and Trivedi (2006). Regarding instrument relevance, the argument is that insurance plans cover an entire family, so spouse's age and lagged choice of the spouse's insurance plan should be a determinant of plan choice. The instruments' independence relies on the assertion that *conditional on own age and other individual characteristics*, spouse's age should not affect personal medical expenditures directly. Moreover, since lagged spouse's choice was pre-determined, it should not impact personal medical expenditures directly either. Therefore, conditioning on covariates is crucial for the instrument's credibility and can be more clearly dealt by specifying parametric functions for the probabilities underlying the DGP.

We calculate bounds for $MTE^{OO}(p, x)$ based on parametric estimates for the functions $\mathbb{P}[S = 1, D = d | X = x, P(Z) = p]$, and $\mathbb{P}[y_{k-1} \leq Y < y_k, S = 1, D = d | X = x, P(Z) = p]$ for $d = \{0, 1\}$, $p \in [0, 1]$, $k \in \{1, \ldots, K\}$ and covariate values x. These probabilities are modeled as logit functions that depend on a linear index of the covariates and a quadratic function of the propensity score, using 20 grid points for the outcome variable. The propensity score is also estimated using a logit model that depends linearly on covariate and instrumental variables. To enforce the common support assumption, we trim the top and bottom 1% of the overlapping estimated propensity score distribution. After estimating those probabilities, we estimate $\alpha(p, x)$, $\beta(p, x)$ and the bounds around $MTE^{OO}(p, x)$ for each covariate value x. Then, following Deb, Munkin, and Trivedi (2006), we assume that the covariates are independent of (S_0, S_1, V) and average the bounds for $MTE^{OO}(p, x)$ across the sample using observed covariates values. By doing so, we recover the bounds for the unconditional

---

[18]The covariate variables are family size, age, squared age, years of schooling, income, female indicator, interaction term between age and female, African-American indicator, Hispanic indicator, marriage indicator, three geographic region dummies, metropolitan area indicator, three subjective health dummies, physical limitation indicator, number of chronic conditions, injury indicator and four year dummies. The descriptive statistics of all variables can be found in Table 1 in Deb, Munkin, and Trivedi (2006).

[19]Differently from Deb, Munkin, and Trivedi (2006), we also combine the spouse's insurance plan choice in a binary variable that is equal to one if the spouse chose a managed care plan and equal to zero if the spouse chose a FFS plan.

$MTE^{OO}(p)$.[20] For details on this parametric approach, see Appendix G.

We estimate bounds around the MTE^{OO} function under three sets of assumptions: (i) no restrictions on the sample selection mechanism (Assumptions 1-5), (ii) "monotonicity of sample selection in the treatment" (Assumptions 1-6), and (iii) "monotonicity of sample selection in the treatment" and "stochastic dominance" (Assumptions 1-7). We are interested in understanding how the different sets of assumptions impact the identified set for MTE^{OO} and interpreting the heterogeneity captured by the MTE.

First, we analyze the MTE^{OO} bounds based only on Assumptions 1-5 (Subsection 3.3). Subfigures 1(a) and 1(b) show the estimated proportion of the always-observed subpopulation within the observed-when-treated and observed-when-untreated groups, that is, $\alpha\left(p, v^{\ell}\right)$ and $\beta\left(p, v^{\ell}\right)$ in Proposition 1. Importantly, in some regions of the support those proportions are equal to one or zero, implying that MTE^{OO} is point-identified or not-identified, respectively. Subfigure 2(a) shows the estimated bounds. For most of the propensity score's support, the bounds without restrictions on the sample selection mechanism are very wide or identification is lost. The estimated bounds do not rule out the possibility of homogeneous treatment effects, i.e., we can still place a constant function inside the bounds in Subfigure 2(a).

In order to tighten those bounds, we consider restrictions on the sample selection mechanism. The "monotonicity of sample selection in the treatment" condition (Assumption 6) imposes that agents who would spend a positive amount of money on ambulatory services if allocated to a FFS plan would also have positive ambulatory expenditures if allocated to a managed care plan. This assumption's direction is in accordance with the results described by Deb, Munkin, and Trivedi (2006), who found that individuals enrolled in HMOs and PPOs are more likely to seek care than FFS enrollees. Note that, under Assumption 6, $\beta(p)$ is always equal to 1. Subfigure 1(c) shows that, under Assumptions 1-6, the estimated proportion of
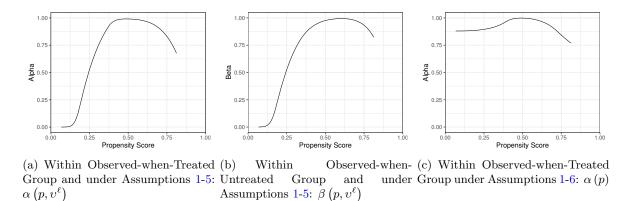
---

[20]By averaging with respect to the observed density of the covariates, we compute bounds around a summary measure of the conditional MTE for the always-observed subgroup: $SCMTE^{OO}(p) := \int \mathbb{E}\left[Y_1^* - Y_0^* \mid V = p, S_0 = 1, S_1 = 1, X = x'\right] \mathrm{d}F_X(x')$. If $X \perp\!\!\!\perp (V = p, S_0 = 1, S_1 = 1)$ holds, then $SCMTE^{OO}(p) = MTE^{OO}(p)$, implying that the summary bounds are valid for the unconditional MTE function for the always-observed subgroup. Importantly, Deb, Munkin, and Trivedi (2006) assumed that the covariates are fully exogenous, implying that $X \perp\!\!\!\perp (V = p, S_0 = 1, S_1 = 1)$ holds. Carneiro and Lee (2009) assume a similar exogeneity assumption. Alternatively, we can analyze the conditional MTE^{OO} function for pre-specified values of the covariates. These results are available upon request.

the always-observed subpopulation within the observed-when-treated group $(\alpha\,(p))$ is strictly positive everywhere in this example. Consequently, the MTE^{OO} function is at least partially identified for the entire support of the propensity score, as can be seen in the bounds reported Subfigure 2(b), illustrating the identifying power of the "monotonicity of sample selection in the treatment" assumption. As a result of tighter bounds, we can rule out the possibility of homogeneous treatment effects, i.e., we cannot fit a constant function inside the bounds in Subfigure 2(b). Interestingly, the upper bound under monotonicity is decreasing. Hence, if the MTE^{OO} function followed the pattern from the upper bound, our results would suggest that the agents who are more likely to enroll in a managed care plan are the ones who incur larger additional ambulatory expenses due to their choice of insurance coverage. This interpretation is compatible with individuals taking into account their potential expenditures when selecting their insurance plans, at least among those for which those expenditures will be positive regardless of their plan (always-observed), and provides extra support for the selectivity result found by Deb, Munkin, and Trivedi (2006).

In order to further tighten the bounds around the MTE^{OO} function, we impose the stochastic dominance assumption (Assumption 7). To interpret this assumption, recall that the "always-observed" (OO) individuals are those for whom ambulatory expenditures would be positive regardless of their insurance plans, while the "observed-only-when-treated" (NO) population encompasses people who would incur expenses only if enrolled in managed care plans. Assumption 7 compares the (counterfactual) expenditures that would take place if all employees were enrolled in a managed care plan (Y_1^*) between those two groups. Formally, it says that for any particular level of expenditures, say 1000\$, the share of individuals that spend less than 1000\$ will be larger for the NO group compared to the OO type. Alternatively, it implies that the average expenditures among the lowest 25% (or any particular quantile) of spenders, will be smaller for the NO than for the OO subgroup. Intuitively, if everyone had a managed care plan, typical patients who would go to ambulatories only if insured by a managed care plan spend less in services than those that would go regardless of their plan. In Appendix D.2, we provide a simple theoretical framework in which this assumption holds.[21]

---

[21]As pointed out by a referee, this assumption is hard to interpret and motivate empirically. However, in a

Importantly, Assumption 7 has no impact on the information about $\alpha(p)$. Consequently, it only impacts the results by increasing the lower bound for the MTE^{OO}, which is much tighter as can be seen in Subfigure 2(c), illustrating the identifying power of the stochastic dominance assumption. If we consider Assumption 7 plausible, we find that the lower bound is positive for most values of the propensity score. This finding reinforces the results in Deb, Munkin, and Trivedi (2006), who obtained a positive effect of PPO choice on ambulatory expenditures and a zero effect of HMO choice after controlling for selection. Furthermore, our results also suggest that the choice of managed care plan reduces ambulatory expenditures for individuals who face high latent costs.

(a) Within Observed-when-Treated Group and under Assumptions 1-5: $\alpha\left(p, v^{\ell}\right)$

(b) Within Observed-when-Untreated Group and under Assumptions 1-5: $\beta\left(p, v^{\ell}\right)$

(c) Within Observed-when-Treated Group under Assumptions 1-6: $\alpha\left(p\right)$
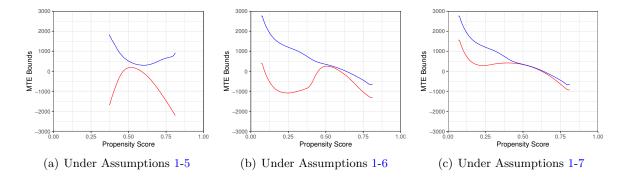
Notes: Solid lines are the parametric estimated proportions of the always-observed group.

Figure 1: Unconditional Proportion of the Always-Observed Group

6   Extension: Identification with discrete instruments

This section extends Proposition 2 to the case with multi-valued discrete instruments, sharply bounding many LATE parameters for the always-observed subpopulation. We focus on the case under the monotonicity restriction (Assumption 1-6). Similar results hold under our other identifying sets of assumptions.

In many applications, the only instruments available are discrete, e.g., treatment eligibility, number of children in the household and quarter of birth. In this section, we provide sharp

---

layered policy analysis (Manski, 2011), we offer a menu of estimates based on different assumptions, that may or may not be plausible according to each researcher's own expertise and beliefs (Tamer, 2010).

(a) Under Assumptions 1-5	(b) Under Assumptions 1-6	(c) Under Assumptions 1-7

Notes: The solid red (blue) line is the estimated lower (upper) bound of the MTE^{OO}.

Figure 2: Unconditional Bounds around the MTE^{OO}

identification results when the instrument is multi-valued discrete, implying the support of the propensity score is finite. The results here can be seen as an extension of Chen and Flores (2015), who provide an outer set for the LATE parameter when the instrument is binary.

Assumption 8. *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 8 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 can only 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), one could obtain averages of the MTE for more specific ranges of the unobservable characteristic V.[22]

The identification argument is similar to the one presented for the continuous instrument case in Subsection 3.4. Under Assumption 1, we have $p_\ell = \mathbb{P}[V \le P(z_\ell)]$ and $\mathbb{P}[P(z_{\ell-1}) < V \le P(z_\ell)] = p_\ell - p_{\ell-1}$. 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 $\mathbb{P}[Y \in A, S = 1, D = 1|P = p_\ell] = \mathbb{P}[Y_1^* \in A, S_1 = 1, V \le p_\ell]$. Therefore,

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

---

[22]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.

$$-\mathbb{P}\left[Y \in A, S = 1, D = 1 | P = p_{\ell-1}\right],$$

which implies that

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

Similarly, we have

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

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

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

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

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

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

To find bounds around $\mathbb{E}[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1, p_{\ell-1} < V \le p_\ell]$, we follow the steps in Subsection 3.4 to derive the following proposition:

Proposition 5. *Under Assumptions 1, 3-6 and 8, the LATE parameters are partially iden-*

tified for the always-observed, i.e.,

$$\underline{\Delta}_{LATE}(\ell) \leq LATE^{OO}(\ell) \leq \overline{\Delta}_{LATE}(\ell)$$

for any $\ell \in \{2, \ldots, K\}$, where $LATE^{OO}\colon \{2, \ldots, K\} \to \mathbb{R}$, $\underline{\Delta}_{LATE}\colon \{2, \ldots, K\} \to \mathbb{R}$ and $\overline{\Delta}_{LATE}\colon \{2, \ldots, K\} \to \mathbb{R}$ are given by

$$LATE^{OO}(\ell) \coloneqq E[Y_1^* - Y_0^*|S_0 = 1, S_1 = 1, p_{\ell-1} < V \leq p_\ell],$$

$$\underline{\Delta}_{LATE}(\ell) \coloneqq \mathbb{E}\left[Y_1^*|S_1 = 1, p_{\ell-1} < V \leq p_\ell, Y_1^* \leq F_{Y_1^*|S_1=1,p_{\ell-1}<V\leq p_\ell}^{-1}(\tilde{\alpha}(p_{\ell-1}, p_\ell))\right]$$

$$- \mathbb{E}[Y_0^*|S_0 = 1, S_1 = 1, p_{\ell-1} < V \leq p_\ell], \text{ and}$$

$$\overline{\Delta}_{LATE}(\ell) \coloneqq \mathbb{E}\left[Y_1^*|S_1 = 1, p_{\ell-1} < V \leq p_\ell, Y_1^* > F_{Y_1^*|S_1=1,p_{\ell-1}<V\leq p_\ell}^{-1}(1 - \tilde{\alpha}(p_{\ell-1}, p_\ell))\right]$$

$$- \mathbb{E}[Y_0^*|S_0 = 1, S_1 = 1, p_{\ell-1} < V \leq p_\ell].$$

Moreover, these bounds are uniformly sharp.

As mentioned above, the quantity for which we derive bounds in Proposition 5 is an 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 Proposition 5 provides sharp bounds on the MTE for the always-observed. This result is related to previous work in which discrete instruments are used to identify MTE in the absence of sample selection. For example, Brinch, Mogstad, and Wiswall (2017) leverages additional functional structure for identification, while Mogstad, Santos, and Torgovitsky (2018) provided partial identification results for the MTE. An extension of their results to the current framework is an interesting topic 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

three increasingly restrictive sets of assumptions. First, we impose standard MTE assumptions without any restrictions to the sample selection mechanism. The second case, which is our main result, imposes monotonicity of the sample selection variable with respect to the treatment, considerably shrinking the identified set. Finally, we consider a strong stochastic dominance assumption which tightens the lower bound for the MTE.

All the results rely on the insight that observed individuals in the sample are a mixture 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 (un)treated. The mixture weights can be identified, leading to a trimming procedure that partially identifies the target parameter, extending Imai (2008), Lee (2009) and Chen and Flores (2015) results to the context of MTE. Moreover, we derive testable implications of our identifying assumptions, and provide extensions to bound LATE parameters with multi-valued discrete instruments. A feasible estimator is proposed and implemented in an empirical illustration analyzing the impacts of managed health care options on health related expenditures, following Deb, Munkin, and Trivedi (2006) and highlighting the practical relevance of the results.

Acknowledgment

ric Society for helpful discussions, and Seung Jin Cho for excellent research assistance.

References

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

Andresen, Martin Eckhoff. 2018. "Exploring Marginal Treatment Effects: Flexible Estimation using Stata." The Stata Journal 18 (1):pp. 118–158.

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.

Arnold, David, Will Dobbie, and Crystal S Yang. 2018. "Racial Bias in Bail Decisions*." The Quarterly Journal of Economics 133 (4):1885–1932. URL https://doi.org/10.1093/qje/qjy012.

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.

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, 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.

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.

———. 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.

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, 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.

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. 2013. "Inference on Counterfactual Distributions." Econometrica 81 (6):2205–2268. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA10582.

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.

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.

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.

Firpo, Sergio and Geert Ridder. 2019. "Partial Identification of the Treatment Effect Distribution and its Functionals." Journal of Econometrics 213:pp. 210–234.

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

Galichon, Alfred and Marc Henry. 2011. "Set identification in models with multiple equilibria." The Review of Economic Studies 78 (4):1264–1298.

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

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, 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.

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.

Kline, Patrick and Christopher R. Walters. 2016. "Evaluating Public Programs with Close Substitutes: The Case of Head Start*." The Quarterly Journal of Economics 131 (4):1795–1848. URL https://doi.org/10.1093/qje/qjw027.

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.

Laffers, Lukas and Giovanni Mellace. 2017. "A Note on Testing Instrument Validity for the Identification of LATE." Empirical Economics 53 (3):pp. 1281–1286.

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

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

———. 2011. "Policy Analysis with Incredible Certitude." The Economic Journal 121 (554):F261–F289.

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.

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.

Mountjoy, Jack. 2019. "Community Colleges and Upward Mobility ." Available at: http://bit.ly/39FnCmo.

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

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

Papadoulos, George and J.M.C. Santos Silva. 2012. "Identification issues in some double-index models for non-negative data." Economics Letters 17:pp. 365–367.

Robinson, P. M. 1988. "Root-N-Consistent Semiparametric Regression." Econometrica 56 (4):931–954. URL http://www.jstor.org/stable/1912705.

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.

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.

Supporting Information
(Online Appendix)

A   Proofs

A.1   Proof of Lemma 1

The validity of the bounds is proven in the main text. It remains to show that the bounds are uniformly sharp. Given the restrictions that Assumptions 1-5 impose on the data, we need to find a joint distribution on $(\tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ that satisfies these restrictions and achieves any value $\upsilon \colon [0,1] \to [0,1]$ such that $\upsilon(p) \in \Upsilon(p)$ for any $p \in \mathcal{P} = [\underline{p}, \overline{p}]$.

Define

$$\upsilon(p) = \begin{cases} \overline{\upsilon}(p) \in \Upsilon(p) & \text{if } p \in [\underline{p}, \overline{p}] \\ \epsilon_0 \in [0,1] & \text{if } p < \underline{p} \\ \epsilon_1 \in [0,1] & \text{if } p > \overline{p} \end{cases}$$

We need to define the joint density (mass) function of $(\tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$. To do so, we will define the density function $f_{\tilde{V}}$, define the mass function $\tilde{\pi}_{(\tilde{S}_0, \tilde{S}_1)|\tilde{V}}$ and use the density function of Z — f_Z — to define $f_{(\tilde{S}_0, \tilde{S}_1), \tilde{V}, Z} = \tilde{\pi}_{(\tilde{S}_0, \tilde{S}_1)|\tilde{V}} \cdot f_{\tilde{V}} \cdot f_Z$. Note that, by construction, Assumption 1 holds. With this goal in mind, fix $(s_0, s_1, p, z) \in \{0,1\}^2 \times \mathbb{R}^2$ arbitrarily. Define $f_{\tilde{V}}(p) = \mathbb{1}\{p \in [0,1]\}$, ensuring that Assumption 5 holds by construction. For brevity, denote the strata by $\underline{OO}$ = always observed, $\underline{NO}$ = observed only when treated, $\underline{ON}$ = observed only when untreated and $\underline{NN}$ = never observed, and the probability of the stratum k conditional on $\tilde{V} = p$ by $\tilde{\pi}_{k|p}$. We now propose the following DGP:

$$\mathbb{P}\left(\tilde{S}_1 = s_1 | \tilde{V} = p\right) = \begin{cases} \frac{\partial \mathbb{P}(S = s_1, D = 1 | P(Z) = p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\ \frac{\mathbb{P}\left(S = s_1, D = 1 | P(Z) = \underline{p}\right)}{\underline{p}} & \text{if } p < \underline{p} \\ \upsilon(p) \cdot \mathbb{1}\{s_1 = 1\} + (1 - \upsilon(p)) \cdot \mathbb{1}\{s_1 = 0\} & \text{if } p > \overline{p} \end{cases}$$

and

$$
\mathbb{P}\left(\tilde{S}_0 = s_0 | \tilde{V} = p\right) = \begin{cases} -\frac{\partial \mathbb{P}(S=s_0, D=0 | P(Z)=p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\ \upsilon(p) \cdot \mathbb{1}\{s_0 = 1\} + (1 - \upsilon(p)) \cdot \mathbb{1}\{s_0 = 0\} & \text{if } p < \underline{p} \\ \frac{\mathbb{P}(S=s_0, D=0 | P(Z)=\overline{p})}{1-\overline{p}} & \text{if } p > \overline{p} \end{cases}.
$$

Define

$$
\tilde{\pi}_{OO|p} = \upsilon(p)
$$
$$
\tilde{\pi}_{NO|p} = \mathbb{P}\left(\tilde{S}_1 = 1 | \tilde{V} = p\right) - \upsilon(p)
$$
$$
\tilde{\pi}_{ON|p} = \mathbb{P}\left(\tilde{S}_0 = 1 | \tilde{V} = p\right) - \upsilon(p)
$$
$$
\tilde{\pi}_{NN|p} = 1 - \tilde{\pi}_{OO|p} - \tilde{\pi}_{NO|p} - \tilde{\pi}_{ON|p}.
$$

We are going to show that the proposed joint distribution satisfies all restrictions on the unconditional joint distribution of (S_0, S_1) and the marginals of S_0 and S_1. We define the correspondence G between the unobservables (S_0, S_1) and the observables (S, D):

$$
G\{(0,0)\} = \{(0,0), (0,1)\}, \; G\{(0,1)\} = \{(0,0), (1,1)\},
$$
$$
G\{(1,0)\} = \{(1,0), (0,1)\}, \; G\{(1,1)\} = \{(1,0), (1,1)\}.
$$

By Galichon and Henry (2011, Theorem 1), we have that all restrictions on the unconditional joint distribution of (S_0, S_1) and the marginals of S_0 and S_1 are given by: for all $A \subset \{(0,0), (0,1), (1,0), (1,1)\}$,

$$
\begin{aligned}
\mathbb{P}((S, D) \in A | Z = z) & \leq \mathbb{P}(G(S_0, S_1) \cap A \neq \emptyset | Z = z) \\
& = \mathbb{P}(G(S_0, S_1) \cap A \neq \emptyset), \; \forall z \in \mathcal{Z},
\end{aligned}
$$

that is:

for singletons,

$$\mathbb{P}(S = 1, D = 1|Z = z) \leq \pi_{NO} + \pi_{OO}, \tag{A.1}$$

$$\mathbb{P}(S = 0, D = 1|Z = z) \leq \pi_{NN} + \pi_{ON}, \tag{A.2}$$

$$\mathbb{P}(S = 1, D = 0|Z = z) \leq \pi_{OO} + \pi_{ON}, \tag{A.3}$$

$$\mathbb{P}(S = 0, D = 0|Z = z) \leq \pi_{NN} + \pi_{NO}; \tag{A.4}$$

for pairs,

$$\mathbb{P}(S = 1, D = 1|Z = z) + \mathbb{P}(S = 1, D = 0|Z = z) \leq \pi_{OO} + \pi_{NO} + \pi_{ON}, \tag{A.5}$$

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

$$\mathbb{P}(S = 1, D = 1|Z = z) + \mathbb{P}(S = 0, D = 0|Z = z) \leq \pi_{OO} + \pi_{NN} + \pi_{NO}, \tag{A.7}$$

$$\mathbb{P}(S = 1, D = 0|Z = z) + \mathbb{P}(S = 0, D = 1|Z = z) \leq \pi_{NN} + \pi_{OO} + \pi_{ON}, \tag{A.8}$$

$$\mathbb{P}(S = 1, D = 0|Z = z) + \mathbb{P}(S = 0, D = 0|Z = z) \leq 1, \tag{A.9}$$

$$\mathbb{P}(S = 0, D = 1|Z = z) + \mathbb{P}(S = 0, D = 0|Z = z) \leq \pi_{NN} + \pi_{NO} + \pi_{ON}; \tag{A.10}$$

for triplets,

$$\mathbb{P}(S = 1, D = 1|Z = z) + \mathbb{P}(S = 1, D = 0|Z = z) + \mathbb{P}(S = 0, D = 1|Z = z) \leq 1 \tag{A.11}$$

$$\mathbb{P}(S = 1, D = 1|Z = z) + \mathbb{P}(S = 1, D = 0|Z = z) + \mathbb{P}(S = 0, D = 0|Z = z) \leq 1 \tag{A.12}$$

$$\mathbb{P}(S = 1, D = 1|Z = z) + \mathbb{P}(S = 0, D = 1|Z = z) + \mathbb{P}(S = 0, D = 0|Z = z) \leq 1 \tag{A.13}$$

$$\mathbb{P}(S = 1, D = 0|Z = z) + \mathbb{P}(S = 0, D = 1|Z = z) + \mathbb{P}(S = 0, D = 0|Z = z) \leq 1 \tag{A.14}$$

Restrictions (A.6), (A.9), (A.11)-(A.14) are redundant. We only need to check that $\tilde{\pi}$ satisfies (A.1)-(A.4), (A.5), (A.7), (A.8), and (A.10). We have

$$\begin{aligned}
\tilde{\pi}_{OO} + \tilde{\pi}_{NO} &= \int_0^1 \tilde{\pi}_{OO|p} + \tilde{\pi}_{NO|p} \, dp = \int_0^1 \mathbb{P}\left(\tilde{S}_1 = 1|\tilde{V} = p\right) dp, \\
&= \mathbb{P}(S = 1, D = 1|P(Z) = \bar{p}) + \epsilon_1(1 - \bar{p}),
\end{aligned}$$

$$\geq \quad \mathbb{P}(S = 1, D = 1 | P(Z) = \bar{p}) = \sup_{p \in \mathcal{P}} \left\{ \mathbb{P}(S = 1, D = 1 | P(Z) = p) \right\},$$

$$= \quad \sup_{P(z) \in \mathcal{P}} \left\{ \mathbb{P}(S = 1, D = 1 | P(Z) = P(z)) \right\} = \sup_{z \in \mathcal{Z}} \left\{ \mathbb{P}(S = 1, D = 1 | Z = z) \right\}$$

because $\mathbb{P}(S = 1, D = 1 | P(Z) = p)$ is increasing in p. Hence, condition (A.1) is satisfied. Similarly, we can show that

$$
\begin{aligned}
\tilde{\pi}_{NN} + \tilde{\pi}_{ON} \quad &= \quad \int_0^1 \mathbb{P}\left(\tilde{S}_1 = 0 | \tilde{V} = p \right) dp, \\
&= \quad \mathbb{P}(S = 0, D = 1 | P(Z) = \bar{p}) + (1 - \epsilon_1)(1 - \bar{p}), \\
&\geq \quad \sup_{z \in \mathcal{Z}} \left\{ \mathbb{P}(S = 0, D = 1 | Z = z) \right\}.
\end{aligned}
$$

Therefore, $\tilde{\pi}$ satisfies condition (A.2). Moreover, we have:

$$
\begin{aligned}
\tilde{\pi}_{OO} + \tilde{\pi}_{ON} \quad &= \quad \int_0^1 \tilde{\pi}_{OO|p} + \tilde{\pi}_{ON|p} dp = \int_0^1 \mathbb{P}\left(\tilde{S}_0 = 1 | \tilde{V} = p \right) dp, \\
&= \quad \mathbb{P}(S = 1, D = 0 | P(Z) = \underline{p}) + \epsilon_0 \underline{p}, \\
&\geq \quad \mathbb{P}(S = 1, D = 0 | P(Z) = \underline{p}) = \sup_{p \in \mathcal{P}} \left\{ \mathbb{P}(S = 1, D = 0 | P(Z) = p) \right\}, \\
&= \quad \sup_{P(z) \in \mathcal{P}} \left\{ \mathbb{P}(S = 1, D = 0 | P(Z) = P(z)) \right\} = \sup_{z \in \mathcal{Z}} \left\{ \mathbb{P}(S = 1, D = 0 | Z = z) \right\}
\end{aligned}
$$

because $\mathbb{P}(S = 1, D = 0 | P(Z) = p)$ is decreasing in p. Thus, condition (A.3) is satisfied. Similarly, we have

$$
\begin{aligned}
\tilde{\pi}_{NN} + \tilde{\pi}_{NO} \quad &= \quad \int_0^1 \mathbb{P}\left(\tilde{S}_0 = 0 | \tilde{V} = p \right) dp, \\
&= \quad \mathbb{P}(S = 0, D = 0 | P(Z) = \underline{p}) + (1 - \epsilon_0) \underline{p}, \\
&\geq \quad \sup_{z \in \mathcal{Z}} \left\{ \mathbb{P}(S = 0, D = 0 | Z = z) \right\}.
\end{aligned}
$$

Hence, condition (A.4) is verified.

Condition (A.5) is equivalent to $\pi_{NN} \leq \mathbb{P}(S = 0, D = 1 | Z = z) + \mathbb{P}(S = 0, D = 0 | Z = z)$. By definition,

$$\tilde{\pi}_{NN|p} \quad = \quad 1 - \mathbb{P}(\tilde{S}_1 = 1 | \tilde{V} = p) - \mathbb{P}(\tilde{S}_0 = 1 | \tilde{V} = p) + \upsilon(p),$$

$$= \begin{cases} \mathbb{P}(\tilde{S}_1 = 0|\tilde{V} = p) - \mathbb{P}(\tilde{S}_0 = 1|\tilde{V} = p) + \upsilon(p) \\ \mathbb{P}(\tilde{S}_0 = 0|\tilde{V} = p) - \mathbb{P}(\tilde{S}_1 = 1|\tilde{V} = p) + \upsilon(p) \end{cases}$$

$$\leq \begin{cases} \mathbb{P}(\tilde{S}_1 = 0|\tilde{V} = p) \\ \mathbb{P}(\tilde{S}_0 = 0|\tilde{V} = p) \end{cases}$$

since $\upsilon(p) \leq \min\left\{\mathbb{P}(\tilde{S}_1 = 1|\tilde{V} = p), \mathbb{P}(\tilde{S}_0 = 1|\tilde{V} = p)\right\}$. Thus,

$$
\begin{aligned}
\tilde{\pi}_{NN|p} &\leq \min\left\{\mathbb{P}(\tilde{S}_1 = 0|\tilde{V} = p), \mathbb{P}(\tilde{S}_0 = 0|\tilde{V} = p)\right\}, \\
&\leq \mathbb{P}(\tilde{S}_1 = 0|\tilde{V} = p)\lambda(z) + \mathbb{P}(\tilde{S}_0 = 0|\tilde{V} = p)(1 - \lambda(z)) \quad \text{for all} \quad \lambda(z) \in [0, 1], \\
&\leq \mathbb{P}(\tilde{S}_1 = 0|\tilde{V} = p)\mathbb{1}\{p \leq P(z)\} + \mathbb{P}(\tilde{S}_0 = 0|\tilde{V} = p)\mathbb{1}\{p > P(z)\} \quad \text{for} \quad \lambda(z) = \mathbb{1}\{p \leq P(z)\},
\end{aligned}
$$

which implies

$$
\begin{aligned}
\int_0^1 \tilde{\pi}_{NN|p}dp &\leq \int_0^1 \mathbb{P}(\tilde{S}_1 = 0|\tilde{V} = p)\mathbb{1}\{p \leq P(z)\}dp + \int_0^1 \mathbb{P}(\tilde{S}_0 = 0|\tilde{V} = p)\mathbb{1}\{p > P(z)\}dp, \\
&= \mathbb{P}(S = 0, D = 1|P(Z) = P(z)) + \mathbb{P}(S = 0, D = 0|P(Z) = P(z)), \\
&= \mathbb{P}(S = 0, D = 1|Z = z) + \mathbb{P}(S = 0, D = 0|Z = z).
\end{aligned}
$$

Hence, $\tilde{\pi}_{NN} \leq \mathbb{P}(S = 0, D = 1|Z = z) + \mathbb{P}(S = 0, D = 0|Z = z)$ for all $z \in \mathcal{Z}$ and Condition (A.5) holds.

Condition (A.10) is equivalent to $\pi_{OO} \leq \mathbb{P}(S = 1, D = 1|Z = z) + \mathbb{P}(S = 1, D = 0|Z = z)$. We have $\tilde{\pi}_{OO|p} = \upsilon(p) \leq \min\left\{\mathbb{P}(\tilde{S}_1 = 1|\tilde{V} = p), \mathbb{P}(\tilde{S}_0 = 1|\tilde{V} = p)\right\}$. Using a reasoning similar to the previous derivation, we obtain $\tilde{\pi}_{OO} \leq \mathbb{P}(S = 1, D = 1|Z = z) + \mathbb{P}(S = 1, D = 0|Z = z)$. Condition (A.7) is equivalent to $\pi_{ON} \leq \mathbb{P}(S = 1, D = 0|Z = z) + \mathbb{P}(S = 0, D = 1|Z = z)$, while condition (A.8) is equivalent to $\pi_{NO} \leq \mathbb{P}(S = 0, D = 0|Z = z) + \mathbb{P}(S = 1, D = 1|Z = z)$. We are going to show that (A.8) is satisfied, and similar derivation holds for (A.7). By definition,

$$
\begin{aligned}
\tilde{\pi}_{NO|p} &= \mathbb{P}(\tilde{S}_1 = 1|\tilde{V} = p) - \upsilon(p), \\
&\leq \begin{cases} \mathbb{P}(\tilde{S}_1 = 1|\tilde{V} = p) \quad \text{since} \quad \upsilon(p) \geq 0 \\ \mathbb{P}(\tilde{S}_0 = 0|\tilde{V} = p) \quad \text{since} \quad \upsilon(p) \geq \mathbb{P}(\tilde{S}_1 = 1|\tilde{V} = p) + \mathbb{P}(\tilde{S}_0 = 1|\tilde{V} = p) - 1 \end{cases}
\end{aligned}
$$

$$\leq \ \min\left\{\mathbb{P}(\tilde{S}_1 = 1|\tilde{V} = p), \mathbb{P}(\tilde{S}_0 = 0|\tilde{V} = p)\right\},$$

$$\leq \ \mathbb{P}(\tilde{S}_1 = 1|\tilde{V} = p)\mathbb{1}\{p \leq P(z)\} + \mathbb{P}(\tilde{S}_0 = 0|\tilde{V} = p)\mathbb{1}\{p > P(z)\}.$$

Therefore, by taking the integral of each side over [0,1], the result follows as for $\tilde{\pi}_{NO}$:

$$\tilde{\pi}_{NO} \leq \mathbb{P}(S = 1, D = 1|Z = z) + \mathbb{P}(S = 0, D = 0|Z = z).$$

A.2 Proof of Proposition 1

The validity of the bounds is proven in the main text. It remains to show that the bounds are uniformly sharp. Given the restrictions that Assumptions 1-5 impose on the data (i.e., equations (3.1) and (3.2)), 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 restrictions, induces the joint distribution on the data (Y, S, D, Z), and achieves any value $\delta \in \left[\underline{\Delta}_1, \overline{\Delta}_1\right]$. To do so, assume that Y^* is absolutely continuous and has a strictly positive density.

First, we show that the lower bound $\underline{\Delta}_1 \colon \mathcal{P} \to \mathbb{R}$ is attainable. We need to define the joint density (mass) function of $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$. To do so, we will define the density functions $f_{\tilde{Y}_0^*|\tilde{S}_0,\tilde{S}_1,\tilde{V}}$, $f_{\tilde{Y}_1^*|\tilde{S}_0,\tilde{S}_1,\tilde{V}}$ and $f_{\tilde{V}}$, define the mass function $\tilde{\pi}_{(\tilde{S}_0,\tilde{S}_1)|\tilde{V}}$ and use the density function of Z — f_Z — to define $f_{\tilde{Y}_0^*,\tilde{Y}_1^*,(\tilde{S}_0,\tilde{S}_1),\tilde{V},Z} = f_{\tilde{Y}_0^*|\tilde{S}_0,\tilde{S}_1,\tilde{V}} \cdot f_{\tilde{Y}_1^*|\tilde{S}_0,\tilde{S}_1,\tilde{V}} \cdot \tilde{\pi}_{(\tilde{S}_0,\tilde{S}_1)|\tilde{V}} \cdot f_{\tilde{V}} \cdot f_Z$. Note that, by construction, Assumption 1 holds. With this goal in mind, fix $(s_0, s_1, y_0, y_1, p, z) \in \{0,1\}^2 \times \mathbb{R}^4$ arbitrarily. Define $f_{\tilde{V}}(p) = \mathbb{1}\{p \in [0,1]\}$, ensuring that Assumption 5 holds by construction. Define

$$v(p) = \begin{cases} \overline{v}(p) = \operatorname{argmin}_{v \in \Upsilon(p)}\{LB_1(p,v) - UB_0(p,v)\} & \text{if } p \in [\underline{p}, \overline{p}] \\ \epsilon_0 \in [0,1] & \text{if } p < \underline{p} \\ \epsilon_1 \in [0,1] & \text{if } p > \overline{p} \end{cases}$$

where $LB_1(p,v)$ and $UB_0(p,v)$ are defined in Subsection 3.3. For brevity, denote the strata by $\underline{OO}$ = always observed, $\underline{NO}$ = observed only when treated, $\underline{ON}$ = observed only when untreated and $\underline{NN}$ = never observed, and the probability of the stratum k conditional on

$\tilde{V} = p$ by $\tilde{\pi}_{k|p}$. Define

$$\mathbb{P}\left(\tilde{S}_1 = s_1 | \tilde{V} = p\right) = \begin{cases} \frac{\partial \mathbb{P}(S=s_1, D=1 | P(Z)=p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\ \frac{\mathbb{P}(S=s_1, D=1 | P(Z)=\underline{p})}{\underline{p}} & \text{if } p < \underline{p} \\ \upsilon(p) \cdot \mathbb{1}\{s_1 = 1\} + (1 - \upsilon(p)) \cdot \mathbb{1}\{s_1 = 0\} & \text{if } p > \overline{p} \end{cases}$$

and

$$\mathbb{P}\left(\tilde{S}_0 = s_0 | \tilde{V} = p\right) = \begin{cases} -\frac{\partial \mathbb{P}(S=s_0, D=0 | P(Z)=p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\ \upsilon(p) \cdot \mathbb{1}\{s_0 = 1\} + (1 - \upsilon(p)) \cdot \mathbb{1}\{s_0 = 0\} & \text{if } p < \underline{p} \\ \frac{\mathbb{P}(S=s_0, D=0 | P(Z)=\overline{p})}{1-\overline{p}} & \text{if } p > \overline{p} \end{cases}$$

The probabilities $\tilde{\pi}_{k|p}$ are given by:

$$\tilde{\pi}_{OO|p} = \upsilon(p)$$

$$\tilde{\pi}_{NO|p} = \mathbb{P}\left(\tilde{S}_1 = 1 | \tilde{V} = p\right) - \upsilon(p)$$

$$\tilde{\pi}_{ON|p} = \mathbb{P}\left(\tilde{S}_0 = 1 | \tilde{V} = p\right) - \upsilon(p)$$

$$\tilde{\pi}_{NN|p} = 1 - \tilde{\pi}_{OO|p} - \tilde{\pi}_{NO|p} - \tilde{\pi}_{ON|p}.$$

Note that the above quantities are positive according to Lemma 1 and add up to 1 by construction.

Define

$$\mathbb{P}\left(\tilde{Y}_1^* \le y_1, \tilde{S}_1 = s_1 | \tilde{V} = p\right) = \begin{cases} \frac{\partial \mathbb{P}(Y \le y_1, S=s_1, D=1 | P(Z)=p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\ \frac{\mathbb{P}(Y \le y_1, S=s_1, D=1 | P(Z)=\underline{p})}{\underline{p}} & \text{if } p < \underline{p} \\ (\upsilon(p) \cdot \mathbb{1}\{s_1 = 1\} + (1 - \upsilon(p)) \cdot \mathbb{1}\{s_1 = 0\}) \, \mathbb{P}(Y \le y_1) & \text{if } p > \overline{p} \end{cases}$$

and

$$\mathbb{P}\left(\tilde{Y}_0^* \leq y_0, \tilde{S}_0 = s_0 | \tilde{V} = p\right) = \begin{cases} -\frac{\partial \mathbb{P}(Y \leq y_0, S = s_0, D = 0 | P(Z) = p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\ (\upsilon(p) \cdot \mathbb{1}\{s_0 = 1\} + (1 - \upsilon(p)) \cdot \mathbb{1}\{s_0 = 0\}) \mathbb{P}(Y \leq y_0) & \text{if } p < \underline{p} \\ \frac{\mathbb{P}(Y \leq y_0, S = s_0, D = 0 | P(Z) = \overline{p})}{1 - \overline{p}} & \text{if } p > \overline{p} \end{cases}$$

Now, we define $f_{\tilde{Y}_0^* | \tilde{S}_0, \tilde{S}_1, \tilde{V}}(y_0 | k, p) = \frac{\partial \mathbb{P}(\tilde{Y}_0^* \leq y_0 | k, \tilde{V} = p)}{\partial y_0}$ and $f_{\tilde{Y}_1^* | \tilde{S}_0, \tilde{S}_1, \tilde{V}}(y_1 | k, p) = \frac{\partial \mathbb{P}(\tilde{Y}_1^* \leq y_1 | k, \tilde{V} = p)}{\partial y_1}$
for any $k \in \{OO, NO, ON, NN\}$, where we only need to define $\frac{\partial \mathbb{P}(\tilde{Y}_0^* \leq y_0 | k, \tilde{V} = p)}{\partial y_0}$ and $\frac{\partial \mathbb{P}(\tilde{Y}_1^* \leq y_1 | k, \tilde{V} = p)}{\partial y_1}$.
Note that the data restriction imposed by equation (3.1) is satisfied by $(\tilde{Y}_1^*, \tilde{S}_1, \tilde{V}, Z)$, imply-
ing that $F_{\tilde{Y}_1, \tilde{S}_1 | \tilde{V}}$ is a proper C.D.F.. Note also that the data restriction imposed by equation
(3.2) is satisfied by $(\tilde{Y}_0^*, \tilde{S}_0, \tilde{V}, Z)$, implying that $F_{\tilde{Y}_0, \tilde{S}_0 | \tilde{V}}$ is a proper C.D.F.. Suppose that
$\tilde{Y}_0 \sim F_{\tilde{Y}_0^* | \tilde{S}_0 = 1, \tilde{V} = p}$ and $\tilde{Y}_1 \sim F_{\tilde{Y}_1^* | \tilde{S}_1 = 1, \tilde{V} = p}$. Define

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

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

$$\mathbb{P}(\tilde{Y}_1^* \leq y_1 | k, \tilde{V} = p) = \frac{\mathbb{P}(\tilde{Y}_1^* \leq y_1, \tilde{S}_1 = 1 | \tilde{V} = p)}{\mathbb{P}(\tilde{S}_1 = 1 | \tilde{V} = p)}, \quad k \in \{ON, NN\}$$

$$\mathbb{P}(\tilde{Y}_0^* \leq y_0 | OO, \tilde{V} = p) = \mathbb{P}\left(\tilde{Y}_0 \leq y_0 | \tilde{Y}_0 > F_{\tilde{Y}_0}^{-1}\left(\frac{\tilde{\pi}_{ON|p}}{\tilde{\pi}_{OO|p} + \tilde{\pi}_{ON|p}}\right)\right),$$

$$\mathbb{P}(\tilde{Y}_0^* \leq y_0 | ON, \tilde{V} = p) = \mathbb{P}\left(\tilde{Y}_0 \leq y_0 | \tilde{Y}_0 \leq F_{\tilde{Y}_0}^{-1}\left(\frac{\tilde{\pi}_{ON|p}}{\tilde{\pi}_{OO|p} + \tilde{\pi}_{ON|p}}\right)\right),$$

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

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

$$\underline{\Delta}_1(p) = \mathbb{E}\left[\tilde{Y}_1^* - \tilde{Y}_0^* | OO, \tilde{V} = p\right]$$

for any $p \in \mathcal{P}$. Finally, we show that the joint distribution of $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ induces the
joint distribution on the data (Y, S, D, Z). Define $\tilde{D} = \mathbb{1}\left\{\tilde{V} \leq P(Z)\right\}$. For any $(y, z) \in \mathbb{R} \times \mathcal{Z}$,

$$\mathbb{P}\left(\tilde{Y} \leq y, \tilde{S} = 1, \tilde{D} = 1 \Big| Z = z\right)$$

$$= \mathbb{P}\left(\tilde{Y}_1^* \leq y, \tilde{S}_1 = 1, \tilde{V} \leq P(z) \middle| Z = z\right)$$

$$= \mathbb{P}\left(\tilde{Y}_1^* \leq y, \tilde{S}_1 = 1, \tilde{V} \leq P(z)\right)$$

$$= \int_0^{P(z)} \mathbb{P}\left(\tilde{Y}_1^* \leq y, \tilde{S}_1 = 1 \middle| \tilde{V} = v\right) \mathrm{d}v$$

$$= \int_0^{\underline{p}} \frac{\mathbb{P}(Y \leq y, S = 1, D = 1|P(Z) = \underline{p})}{\underline{p}} \mathrm{d}v + \int_{\underline{p}}^{P(z)} \frac{\partial \mathbb{P}(Y \leq y, S = 1, D = 1|P(Z) = v)}{\partial p} \mathrm{d}v$$

$$= \mathbb{P}\left(Y \leq y, S = 1, D = 1 \middle| P(Z) = P(z)\right)$$

$$= \mathbb{P}\left(Y \leq y, S = 1, D = 1 \middle| Z = z\right)$$

where the last equality is a testable implication of the model (see Appendix B). Analogously,

$$\mathbb{P}\left(\tilde{Y} \leq y, \tilde{S} = 1, \tilde{D} = 0 \middle| Z = z\right) = \mathbb{P}\left(Y \leq y, S = 1, D = 0 \middle| Z = z\right),$$

$$\mathbb{P}\left(\tilde{S} = 0, \tilde{D} = 1 \middle| Z = z\right) = \mathbb{P}\left(S = 0, D = 1 \middle| Z = z\right),$$

$$\mathbb{P}\left(\tilde{S} = 0, \tilde{D} = 0 \middle| Z = z\right) = \mathbb{P}\left(S = 0, D = 0 \middle| Z = z\right).$$

Similar reasoning holds for the upper bound, $\overline{\Delta}_1$. To attain any function $\delta \in \left(\underline{\Delta}_1, \overline{\Delta}_1\right)$, we can use convex combinations of the joint distributions that achieve the lower and upper bounds, where the weights of the convex combination may depend on the value $p \in \mathcal{P}$.

A.3   Proof of Proposition 2

The validity of the bounds is proven in the main text. It remains to show that the bounds are uniformly sharp. Given the restrictions that Assumptions 1-6 impose on the data (i.e., equations (3.1) and (3.2)), 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 restrictions, induces the joint distribution on the data (Y, S, D, Z), and achieves any value $\delta \in \left[\underline{\Delta}_2, \overline{\Delta}_2\right]$. Assume that Y^* is absolutely continuous with strictly positive density.

First, we show that the lower bound $\underline{\Delta}_2 \colon \mathcal{P} \to \mathbb{R}$ is attainable. We need to define the joint density (mass) function of $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$. To do so, define the density functions $f_{\tilde{Y}_0^*|\tilde{S}_0,\tilde{S}_1,\tilde{V}}$, $f_{\tilde{Y}_1^*|\tilde{S}_0,\tilde{S}_1,\tilde{V}}$ and $f_{\tilde{V}}$, define the mass function $\tilde{\pi}_{(\tilde{S}_0,\tilde{S}_1)|\tilde{V}}$ and use the density function of Z — f_Z — to obtain $f_{\tilde{Y}_0^*,\tilde{Y}_1^*,(\tilde{S}_0,\tilde{S}_1),\tilde{V},Z} = f_{\tilde{Y}_0^*|\tilde{S}_0,\tilde{S}_1,\tilde{V}} \cdot f_{\tilde{Y}_1^*|\tilde{S}_0,\tilde{S}_1,\tilde{V}} \cdot \tilde{\pi}_{(\tilde{S}_0,\tilde{S}_1)|\tilde{V}} \cdot f_{\tilde{V}} \cdot f_Z.$

Then, by construction, Assumption 1 holds. Fix $(s_0, s_1, y_0, y_1, p, z) \in \{0,1\}^2 \times \mathbb{R}^4$ arbitrarily. Define $f_{\tilde{V}}(p) = \mathbb{1}\{p \in [0,1]\}$, ensuring Assumption 5 holds by construction. For brevity, denote the strata by $\underline{OO}$ = always observed, $\underline{NO}$ = observed only when treated and $\underline{NN}$ = never observed, and the probability of the stratum k conditional on $\tilde{V} = p$ by $\tilde{\pi}_{k|p}$. Define

$$\mathbb{P}\left(\tilde{Y}_1^* \leq y_1, \tilde{S}_1 = 1 | \tilde{V} = p\right) = \begin{cases} \frac{\partial \mathbb{P}(Y \leq y_1, S=1, D=1|P(Z)=p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\[2mm] \frac{\mathbb{P}(Y \leq y_1, S=1, D=1|P(Z)=\underline{p})}{\underline{p}} & \text{if } p < \underline{p} \\[2mm] \frac{\mathbb{P}(Y \leq y_1, S=1, D=0|P(Z)=\overline{p})}{1-\overline{p}} & \text{if } p > \overline{p} \end{cases}$$

$$\mathbb{P}\left(\tilde{S}_1 = 0 | \tilde{V} = p\right) = \begin{cases} \frac{\partial \mathbb{P}(S=0, D=1|P(Z)=p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\[2mm] \frac{\mathbb{P}(S=0, D=1|P(Z)=\underline{p})}{\underline{p}} & \text{if } p < \underline{p} \\[2mm] \frac{\mathbb{P}(S=0, D=0|P(Z)=\overline{p})}{1-\overline{p}} & \text{if } p > \overline{p} \end{cases}$$

$$\mathbb{P}\left(\tilde{Y}_0^* \leq y_0, \tilde{S}_0 = 1 | \tilde{V} = p\right) = \begin{cases} -\frac{\partial \mathbb{P}(Y \leq y_0, S=1, D=0|P(Z)=p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\[2mm] \epsilon \frac{\mathbb{P}(Y \leq y_0, S=1, D=1|P(Z)=\underline{p})}{\underline{p}} & \text{if } p < \underline{p} \\[2mm] \frac{\mathbb{P}(Y \leq y_0, S=1, D=0|P(Z)=\overline{p})}{1-\overline{p}} & \text{if } p > \overline{p} \end{cases}$$

$$\mathbb{P}\left(\tilde{S}_0 = 0 | \tilde{V} = p\right) = \begin{cases} -\frac{\partial \mathbb{P}(S=0, D=0|P(Z)=p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\[2mm] 1 - \epsilon \frac{\mathbb{P}(S=1, D=1|P(Z)=\underline{p})}{\underline{p}} & \text{if } p < \underline{p} \\[2mm] \frac{\mathbb{P}(S=0, D=0|P(Z)=\overline{p})}{1-\overline{p}} & \text{if } p > \overline{p} \end{cases}$$

where $\epsilon \in (0,1)$.

The probabilities $\tilde{\pi}_{k|p}$ are given by:

$$\tilde{\pi}_{OO|p} = \mathbb{P}(\tilde{S}_0 = 1 | \tilde{V} = p)$$
$$\tilde{\pi}_{NO|p} = \mathbb{P}(\tilde{S}_1 = 1 | \tilde{V} = p) - \mathbb{P}(\tilde{S}_0 = 1 | \tilde{V} = p)$$

$$\tilde{\pi}_{NN|p} = \mathbb{P}(\tilde{S}_1 = 0|\tilde{V} = p).$$

Under Assumptions 1-6, the above quantities are positive, and add to one. For $p \in \mathcal{P}$,

$$\begin{aligned}
\tilde{\pi}_{OO|p} + \tilde{\pi}_{NO|p} &= \mathbb{P}(\tilde{S}_1 = 1|\tilde{V} = p) \\
&= \frac{\partial \mathbb{P}(S = 1, D = 1|P(Z) = p)}{\partial p},
\end{aligned}$$

implying that

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

We also have

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

Therefore, $\tilde{\pi}_{OO|p,z} + \tilde{\pi}_{NO|p,z} + \tilde{\pi}_{NN|p,z} = 1$. The same result holds for $p < \underline{p}$ and $p > \overline{p}$ from the fact that $\mathbb{P}(D = 1|P(Z) = \underline{p}) = \underline{p}$ and $\mathbb{P}(D = 0|P(Z) = \overline{p}) = 1 - \overline{p}$.

Now, define $f_{\tilde{Y}_0^*|\tilde{S}_0,\tilde{S}_1,\tilde{V}}(y_0|\,k,p) = \frac{\partial \mathbb{P}(\tilde{Y}_0^* \le y_0|k,\tilde{V}=p)}{\partial y_0}$ and $f_{\tilde{Y}_1^*|\tilde{S}_0,\tilde{S}_1,\tilde{V}}(y_1|\,k,p) = \frac{\partial \mathbb{P}(\tilde{Y}_1^* \le y_1|k,\tilde{V}=p)}{\partial y_1}$ for any $k \in \{OO, NO, NN\}$, where we only need to characterize $\frac{\partial \mathbb{P}(\tilde{Y}_0^* \le y_0|k,\tilde{V}=p)}{\partial y_0}$ and $\frac{\partial \mathbb{P}(\tilde{Y}_1^* \le y_1|k,\tilde{V}=p)}{\partial y_1}$. The data restriction imposed by equation (3.1) is satisfied by $(\tilde{Y}_1^*, \tilde{S}_1, \tilde{V}, Z)$, implying that $F_{\tilde{Y}_1,\tilde{S}_1|\tilde{V}}$ is a proper CDF. Similarly, the data restriction imposed by equation (3.2) is satisfied by $(\tilde{Y}_0^*, \tilde{S}_0, \tilde{V}, Z)$, and $F_{\tilde{Y}_0,\tilde{S}_0|\tilde{V}}$ is a proper CDF. Suppose that $\tilde{Y}_1 \sim F_{\tilde{Y}_1^*|\tilde{S}_1=1,\tilde{V}=p}$. Define

$$\mathbb{P}(\tilde{Y}_1^* \le y_1|OO, \tilde{V} = p) = \mathbb{P}\left(\tilde{Y}_1 \le y_1 | \tilde{Y}_1 \le F_{\tilde{Y}_1}^{-1}\left(\frac{\tilde{\pi}_{OO|p}}{\tilde{\pi}_{OO|p} + \tilde{\pi}_{NO|p}}\right)\right), \qquad (A.15)$$

$$\mathbb{P}(\tilde{Y}_1^* \le y_1|NO, \tilde{V} = p) = \mathbb{P}\left(\tilde{Y}_1 \le y_1 | \tilde{Y}_1 > F_{\tilde{Y}_1}^{-1}\left(\frac{\tilde{\pi}_{OO|p}}{\tilde{\pi}_{OO|p} + \tilde{\pi}_{NO|p}}\right)\right), \qquad (A.16)$$

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

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

Notice that the lower bound in Proposition 2 is attained by the distributions of $\tilde{Y}_0^* \big| OO, \tilde{V}$ and $\tilde{Y}_1^* \big| OO, \tilde{V}$ because

$$\underline{\Delta}_2(p) = \mathbb{E}\left[\tilde{Y}_1^* - \tilde{Y}_0^* | OO, \tilde{V} = p\right] \text{ for any } p \in \mathcal{P}.$$

Finally, the joint distribution of $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ induces the joint distribution on the data (Y, S, D, Z) by construction. The proof is identical to the proof in Appendix A.2.

Similar reasoning holds for the upper bound, $\overline{\Delta}_2$. To attain any function $\delta \in (\underline{\Delta}_2, \overline{\Delta}_2)$, we can use convex combinations of the joint distributions that achieve the lower and upper bounds, where the weights of the convex combination may depend on the value $p \in \mathcal{P}$.

A.4   Proof of Proposition 3

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 (3.1) and (3.2)), 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, satisfy the stochastic dominance assumption, induce the joint distribution on the data (Y, S, D, Z), and achieve any value $\delta \in [\underline{\Delta}_3, \overline{\Delta}_3]$.

The proof of Proposition 3 is very similar to the one in Appendix A.3. We only have to modify equations (A.15) and (A.16) to:

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

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

These 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.5 Proof of Proposition 4

The validity of the bounds is proven in the main text. We first prove that the bounds around TE under Assumptions (1)-(6) are sharp. Pick any treatment effect parameter $TE := \int_{\mathcal{P}} MTE^{OO}(p) \cdot \omega(p) \, dp$, where the weighting function $\omega \colon \mathcal{P} \to \mathbb{R}$ is known or identified. Given the restrictions that Assumptions (1)-(6) impose on the data (i.e., equations (3.1) and (3.2)), 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 restrictions, induces the joint distribution on the data (Y, S, D, Z), and achieves any value $\delta \in \left[\int_{\mathcal{P}} (\underline{\Delta}_2(p) \cdot \omega(p)) \, dp, \int_{\mathcal{P}} (\underline{\Delta}_2(p) \cdot \omega(p)) \, dp \right]$. To do so, assume that Y^* is absolutely continuous and has a strictly positive density.

First, we show that the lower bound $\underline{\Delta}_2 \colon \mathcal{P} \to \mathbb{R}$ is attainable. The joint density (mass) function of $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ is defined as the same distribution used for Proposition 2.

Now, observe that the lower bound $\int_{\mathcal{P}} (\underline{\Delta}_2(p) \cdot \omega(p)) \, dp$ in Proposition 4 is attained by the distributions of $\tilde{Y}_0^* \big| OO, \tilde{V}$ and $\tilde{Y}_1^* \big| OO, \tilde{V}$ because

$$\underline{\Delta}_2(p) = \mathbb{E}\left[\tilde{Y}_1^* - \tilde{Y}_0^* | OO, \tilde{V} = p \right] \text{ for any } p \in \mathcal{P}, \text{ and}$$

$$\int_{\mathcal{P}} (\underline{\Delta}_2(p) \cdot \omega(p)) \, dp = \int_{\mathcal{P}} \left(\mathbb{E}\left[\tilde{Y}_1^* - \tilde{Y}_0^* | OO, \tilde{V} = p \right] \cdot \omega(p) \right) \, dp.$$

Finally, the joint distribution of $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ induces the joint distribution on the data (Y, S, D, Z) by construction. The proof is identical to the proof in Appendix A.3.

Similar reasoning holds for the upper bound, $\int_{\mathcal{P}} (\overline{\Delta}_2(p) \cdot \omega(p)) \, dp$. To attain any value $\delta \in \left(\int_{\mathcal{P}} (\underline{\Delta}_2(p) \cdot \omega(p)) \, dp, \int_{\mathcal{P}} (\overline{\Delta}_2(p) \cdot \omega(p)) \, dp \right)$, we can use convex combinations of the joint distributions that achieve the lower and upper bounds, where the weights of the convex combination does not depend on the value $p \in \mathcal{P}$.

The proofs that the bounds around TE under Assumptions 1-5 and 1-7 are similar to the proof above and they use the distributions described in Appendices A.2 and A.4.

A.6 Deriving Identifiable Weights for ATE^{OO}, ATT^{OO}, ATU^{OO}, $LATE^{OO}$ and $PRTE^{OO}$ (Tables 2 and 3)

In this section, we refer to the Law of Iterated Expectations using its acronym, LIE.

A.6.1 ATE within the always-observed population

Observe that

$$ATE^{OO} := \mathbb{E}\left[Y_1^* - Y_0^* \mid S_0 = 1, S_1 = 1\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[Y_1^* - Y_0^* \mid V, S_0 = 1, S_1 = 1\right] \mid S_0 = 1, S_1 = 1\right] \text{ (LIE)}$$

$$= \int_0^1 MTE^{OO}(p) \cdot f_{V \mid S_0=1, S_1=1}(p) \; \mathrm{d}p \text{ (Equation (2.2) and the expectation operator)}$$

$$= \int_0^1 MTE^{OO}(p) \cdot \frac{\mathbb{P}\left[S_0 = 1, S_1 = 1 \mid V = p\right]}{\mathbb{P}\left[S_0 = 1, S_1 = 1\right]} \cdot f_V(p) \; \mathrm{d}p \text{ (Bayes' rule)}$$

$$= \int_0^1 MTE^{OO}(p) \cdot \frac{\mathbb{P}\left[S_0 = 1, S_1 = 1 \mid V = p\right]}{\mathbb{P}\left[S_0 = 1, S_1 = 1\right]} \; \mathrm{d}p \text{ (Assumption 5)}$$

$$= \int_0^1 MTE^{OO}(p) \cdot \frac{\mathbb{P}\left[S_0 = 1, S_1 = 1 \mid V = p\right]}{\int_0^1 \mathbb{P}\left[S_0 = 1, S_1 = 1 \mid V = p\right] \; \mathrm{d}p} \; \mathrm{d}p \text{ (Assumption 5)}$$

$$= \int_0^1 MTE^{OO}(p) \cdot \left[\frac{\dfrac{\partial \mathbb{P}\left[S = 1, D = 0 \mid P(Z) = p\right]}{\partial p}}{\displaystyle\int_0^1 \frac{\partial \mathbb{P}\left[S = 1, D = 0 \mid P(Z) = p\right]}{\partial p} \; \mathrm{d}p}\right] \; \mathrm{d}p \text{ (Equation (3.13))},$$

implying that $\omega_{ATE}(p) = \dfrac{\dfrac{\partial \mathbb{P}\left[S = 1, D = 0 \mid P(Z) = p\right]}{\partial p}}{\displaystyle\int_0^1 \frac{\partial \mathbb{P}\left[S = 1, D = 0 \mid P(Z) = p\right]}{\partial p} \; \mathrm{d}p}.$

A.6.2 ATT within the always-observed population

Observe that

$$ATT^{OO} := \mathbb{E}\left[Y_1^* - Y_0^* \mid D = 1, S_0 = 1, S_1 = 1\right]$$

$$= \mathbb{E}\left[Y_1^* - Y_0^* \mid V \leq P(Z), S_0 = 1, S_1 = 1\right] \text{ (Equation (2.1))}$$

$$= \frac{\mathbb{E}\left[\mathbb{1}\{V \leq P(Z)\} \cdot (Y_1^* - Y_0^*) \mid S_0 = 1, S_1 = 1\right]}{\mathbb{P}\left[V \leq P(Z) \mid S_0 = 1, S_1 = 1\right]}$$

$$= \frac{\mathbb{E}\left[\mathbb{1}\{V \leq P(Z)\} \cdot (Y_1^* - Y_0^*) \mid S_0 = 1, S_1 = 1\right]}{\dfrac{\mathbb{P}\left[V \leq P(Z), S_0 = 1, S_1 = 1\right]}{\mathbb{P}\left[S_0 = 1, S_1 = 1\right]}} \text{ (by the definition of a conditional expectation)}$$

$$= \frac{\mathbb{E}\left[\mathbb{1}\{V \leq P(Z)\} \cdot \mathbb{E}\left[Y_1^* - Y_0^* \mid V, P(Z), S_0 = 1, S_1 = 1\right] \mid S_0 = 1, S_1 = 1\right]}{\dfrac{\mathbb{E}\left[\mathbb{1}\{V \leq P(Z)\} \cdot \mathbb{P}\left[S_0 = 1, S_1 = 1 \mid V, P(Z)\right]\right]}{\mathbb{P}\left[S_0 = 1, S_1 = 1\right]}} \text{ (LIE)}$$

$$= \frac{\int_0^1 \int_p^1 \mathbb{E}\left[Y_1^* - Y_0^* \mid V = p, P(Z) = u, S_0 = 1, S_1 = 1\right] \cdot f_{P(Z),V\mid S_0=1,S_1=1}(u, p) \ \mathrm{d}u \ \mathrm{d}p}{\dfrac{\int_0^1 \int_p^1 \mathbb{P}\left[S_0 = 1, S_1 = 1 \mid V = p, P(Z) = u\right] \cdot f_{P(Z),V}(u, p) \ \mathrm{d}u \ \mathrm{d}p}{\mathbb{P}\left[S_0 = 1, S_1 = 1\right]}}$$

$$= \frac{\int_0^1 \int_p^1 \mathbb{E}\left[Y_1^* - Y_0^* \mid V = p, S_0 = 1, S_1 = 1\right] \cdot f_{P(Z)}(u) \cdot f_{V\mid S_0=1,S_1=1}(p) \ \mathrm{d}u \ \mathrm{d}p}{\dfrac{\int_0^1 \int_p^1 \mathbb{P}\left[S_0 = 1, S_1 = 1 \mid V = p\right] \cdot f_{P(Z)}(u) \cdot f_V(p) \ \mathrm{d}u \ \mathrm{d}p}{\mathbb{P}\left[S_0 = 1, S_1 = 1\right]}} \text{ (Assumption 1)}$$

$$= \frac{\int_0^1 MTE^{OO}(p) \cdot \left(\int_p^1 f_{P(Z)}(u) \ \mathrm{d}u\right) \cdot f_{V\mid S_0=1,S_1=1}(p) \ \mathrm{d}p}{\dfrac{\int_0^1 \left(\int_p^1 f_{P(Z)}(u) \ \mathrm{d}u\right) \cdot \mathbb{P}\left[S_0 = 1, S_1 = 1 \mid V = p\right] \ \mathrm{d}p}{\mathbb{P}\left[S_0 = 1, S_1 = 1\right]}}$$

by Assumption 5, definition (2.2) and the linearity of the integral operator,

$$= \frac{\int_0^1 MTE^{OO}(p) \cdot \left(\int_p^1 f_{P(Z)}(u) \, \mathrm{d}u\right) \cdot \frac{\mathbb{P}[S_0 = 1, S_1 = 1 | V = p]}{\mathbb{P}[S_0 = 1, S_1 = 1]} \, \mathrm{d}p}{\frac{\int_0^1 \left(\int_p^1 f_{P(Z)}(u) \, \mathrm{d}u\right) \cdot \mathbb{P}[S_0 = 1, S_1 = 1 | V = p] \, \mathrm{d}p}{\mathbb{P}[S_0 = 1, S_1 = 1]}}$$

by the argument outlined in Appendix A.6.1,

$$= \int_0^1 MTE^{OO}(p) \cdot \frac{\left(\int_p^1 f_{P(Z)}(u) \, \mathrm{d}u\right) \cdot \mathbb{P}[S_0 = 1, S_1 = 1 | V = p]}{\int_0^1 \left(\int_p^1 f_{P(Z)}(u) \, \mathrm{d}u\right) \cdot \mathbb{P}[S_0 = 1, S_1 = 1 | V = p] \, \mathrm{d}p} \, \mathrm{d}p$$

$$= \int_0^1 MTE^{OO}(p) \cdot \left[\frac{\left(\int_p^1 f_{P(Z)}(u) \, \mathrm{d}u\right) \cdot \frac{\partial \mathbb{P}[S = 1, D = 0 | P(Z) = p]}{\partial p}}{\int_0^1 \left(\int_p^1 f_{P(Z)}(u) \, \mathrm{d}u\right) \cdot \frac{\partial \mathbb{P}[S = 1, D = 0 | P(Z) = p]}{\partial p} \, \mathrm{d}p} \right] \, \mathrm{d}p$$

by Equation (3.13),

implying that $\omega_{ATT}(p) = \dfrac{\left(\int_p^1 f_{P(Z)}(u) \, \mathrm{d}u\right) \cdot \dfrac{\partial \mathbb{P}[S = 1, D = 0 | P(Z) = p]}{\partial p}}{\int_0^1 \left(\int_p^1 f_{P(Z)}(u) \, \mathrm{d}u\right) \cdot \dfrac{\partial \mathbb{P}[S = 1, D = 0 | P(Z) = p]}{\partial p} \, \mathrm{d}p}.$

A.6.3   ATU within the always-observed population

Analogously to Appendix A.6.2, we have that

$$ATU^{OO} := \mathbb{E}[Y_1^* - Y_0^* | D = 0, S_0 = 1, S_1 = 1]$$

$$= \int_0^1 MTE^{OO}(p) \cdot \left[\frac{\left(\int_0^p f_{P(Z)}(u) \, \mathrm{d}u\right) \cdot \frac{\partial \mathbb{P}[S = 1, D = 0 | P(Z) = p]}{\partial p}}{\int_0^1 \left(\int_0^p f_{P(Z)}(u) \, \mathrm{d}u\right) \cdot \frac{\partial \mathbb{P}[S = 1, D = 0 | P(Z) = p]}{\partial p} \, \mathrm{d}p} \right] \, \mathrm{d}p$$

implying that $\omega_{ATU}(p) = \dfrac{\left(\int_0^p f_{P(Z)}(u)\,\mathrm{d}u\right) \cdot \dfrac{\partial \mathbb{P}\left[S = 1, D = 0 \mid P(Z) = p\right]}{\partial p}}{\displaystyle\int_0^1 \left(\int_0^p f_{P(Z)}(u)\,\mathrm{d}u\right) \cdot \dfrac{\partial \mathbb{P}\left[S = 1, D = 0 \mid P(Z) = p\right]}{\partial p}\,\mathrm{d}p}.$

A.6.4   LATE within the always-observed population

Observe that, for any $(\underline{p}, \overline{p}) \in [0,1]^2$,

$$LATE^{OO}(\underline{p}, \overline{p}) = \mathbb{E}\left[Y_1^* - Y_0^* \mid V \in \left[\underline{p}, \overline{p}\right], S_0 = 1, S_1 = 1\right],$$

$$= \int_{\underline{p}}^{\overline{p}} \mathbb{E}\left[Y_1^* - Y_0^* \mid V = p, S_0 = 1, S_1 = 1\right] f_{V \mid S_0 = 1, S_1 = 1}(p)\,\mathrm{d}p.$$

The proof is similar to that of ATE^{OO}, except that the integral is over $\left[\underline{p}, \overline{p}\right]$ instead of $[0,1]$.

Consequently, we have that $\omega_{LATE}(p) = \dfrac{\dfrac{\partial \mathbb{P}\left[S = 1, D = 0 \mid P(Z) = p\right]}{\partial p}}{\displaystyle\int_{\underline{p}}^{\overline{p}} \dfrac{\partial \mathbb{P}\left[S = 1, D = 0 \mid P(Z) = p\right]}{\partial p}\,\mathrm{d}p}.$

A.6.5   PRTE within the always-observed population

First, we need to define a policy. A policy a is a reassignment of the treatment based on $D_a = \mathbb{1}\{V \le P_a\}$, where the random variable P_a is distributed according to the density f_{P_a} and to the cumulative distribution function F_{P_a}. The new random variable P_a is still assumed to be independent of $(V, Y_0^*, Y_1^*, S_0, S_1)$. Moreover, we define the possibly censored outcome under policy a as $Y^* := D_a Y_1^* + (1 - D_a) Y_0^*$.

Now, observe that, for a policy a,

$\mathbb{E}\left[Y_a^* \mid S_0 = 1, S_1 = 1\right]$

$\quad = \mathbb{E}\left[\mathbb{E}\left[Y_a^* \mid P_a, S_0 = 1, S_1 = 1\right] \mid S_0 = 1, S_1 = 1\right]$ (LIE)

$\quad = \int_0^1 \mathbb{E}\left[Y_a^* \mid P_a = u, S_0 = 1, S_1 = 1\right] \cdot f_{P_a \mid S_0 = 1, S_1 = 1}(u)\,\mathrm{d}u$

$\quad = \int_0^1 \mathbb{E}\left[Y_a^* \mid P_a = u, S_0 = 1, S_1 = 1\right] \cdot f_{P_a}(u)\,\mathrm{d}u$

$\quad$ because $P_a \perp\!\!\!\perp (V, Y_0^*, Y_1^*, S_0, S_1)$,

$$= \int_0^1 \mathbb{E}\left[D_a Y_1^* + (1 - D_a) Y_0^* \,|\, P_a = u, S_0 = 1, S_1 = 1\right] \cdot f_{P_a}(u) \, \mathrm{d}u \text{ (by definition)}$$

$$= \int_0^1 \mathbb{E}\left[\mathbb{1}\left\{V \le u\right\} \cdot Y_1^* \,|\, P_a = u, S_0 = 1, S_1 = 1\right] \cdot f_{P_a}(u) \, \mathrm{d}u$$

$$\quad + \int_0^1 \mathbb{E}\left[\mathbb{1}\left\{u < V\right\} \cdot Y_0^* \,|\, P_a = u, S_0 = 1, S_1 = 1\right] \cdot f_{P_a}(u) \, \mathrm{d}u$$

$$= \int_0^1 \mathbb{E}\left[\mathbb{1}\left\{V \le u\right\} \cdot Y_1^* \,|\, S_0 = 1, S_1 = 1\right] \cdot f_{P_a}(u) \, \mathrm{d}u$$

$$\quad + \int_0^1 \mathbb{E}\left[\mathbb{1}\left\{u < V\right\} \cdot Y_0^* \,|\, S_0 = 1, S_1 = 1\right] \cdot f_{P_a}(u) \, \mathrm{d}u$$

because $P_a \perp\!\!\!\perp (V, Y_0^*, Y_1^*, S_0, S_1)$,

$$= \int_0^1 \mathbb{E}\left[\mathbb{1}\left\{V \le u\right\} \cdot \mathbb{E}\left[Y_1^* \,|\, V, S_0 = 1, S_1 = 1\right] \,|\, S_0 = 1, S_1 = 1\right] \cdot f_{P_a}(u) \, \mathrm{d}u$$

$$\quad + \int_0^1 \mathbb{E}\left[\mathbb{1}\left\{u < V\right\} \cdot \mathbb{E}\left[Y_0^* \,|\, V, S_0 = 1, S_1 = 1\right] \,|\, S_0 = 1, S_1 = 1\right] \cdot f_{P_a}(u) \, \mathrm{d}u \text{ (LIE)}$$

$$= \int_0^1 \left[\int_0^u \mathbb{E}\left[Y_1^* \,|\, V = p, S_0 = 1, S_1 = 1\right] \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p\right] \cdot f_{P_a}(u) \, \mathrm{d}u$$

$$\quad + \int_0^1 \left[\int_u^1 \mathbb{E}\left[Y_0^* \,|\, V = p, S_0 = 1, S_1 = 1\right] \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p\right] \cdot f_{P_a}(u) \, \mathrm{d}u$$

$$= \int_0^1 \mathbb{E}\left[Y_1^* \,|\, V = p, S_0 = 1, S_1 = 1\right] \cdot \left(\int_p^1 f_{P_a}(u) \, \mathrm{d}u\right) \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p$$

$$\quad + \int_0^1 \mathbb{E}\left[Y_0^* \,|\, V = p, S_0 = 1, S_1 = 1\right] \cdot \left(\int_0^p f_{P_a}(u) \, \mathrm{d}u\right) \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p$$

by Assumption 4 and Fubini's Theorem,

$$= \int_0^1 \mathbb{E}\left[Y_1^* \,|\, V = p, S_0 = 1, S_1 = 1\right] \cdot (1 - F_{P_a}(p)) \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p$$

$$\quad + \int_0^1 \mathbb{E}\left[Y_0^* \,|\, V = p, S_0 = 1, S_1 = 1\right] \cdot F_{P_a}(p) \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p \text{ (by definition)}$$

$$= \int_0^1 \mathbb{E}\left[Y_1^* \,|\, V = p, S_0 = 1, S_1 = 1\right] \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p$$

$$\quad - \int_0^1 MTE^{OO}(p) \cdot F_{P_a}(p) \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p \text{ (Equation (2.2))}.$$

Analogously, we have that, for a policy a',

$$\mathbb{E}\left[Y_{a'}^* \,|\, S_0 = 1, S_1 = 1\right]$$

$$= \int_0^1 \mathbb{E}\left[Y_1^* \,|\, V = p, S_0 = 1, S_1 = 1\right] \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p$$

61

$$- \int_0^1 MTE^{OO}(p) \cdot F_{P_{a'}}(p) \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p.$$

Combining the last two results, we have that

$$PRTE^{OO} := \frac{\mathbb{E}\left[Y_a^* - Y_{a'}^* \,|\, S_0 = 1, S_1 = 1\right]}{\int_0^1 \left(F_{P_{a'}}(p) - F_{P_a}(p)\right) \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p}$$

$$= \frac{\int_0^1 MTE^{OO}(p) \cdot \left(F_{P_{a'}}(p) - F_{P_a}(p)\right) \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p}{\int_0^1 \left(F_{P_{a'}}(p) - F_{P_a}(p)\right) \cdot f_{V|S_0=1,S_1=1}(p) \, \mathrm{d}p}$$

$$= \frac{\int_0^1 MTE^{OO}(p) \cdot \left(F_{P_{a'}}(p) - F_{P_a}(p)\right) \cdot \mathbb{P}\left[S_0 = 1, S_1 = 1 \,|\, V = p\right] \, \mathrm{d}p}{\int_0^1 \left(F_{P_{a'}}(p) - F_{P_a}(p)\right) \cdot \mathbb{P}\left[S_0 = 1, S_1 = 1 \,|\, V = p\right] \, \mathrm{d}p}$$

by the argument outlined in Appendix A.6.1,

$$= \int_0^1 MTE^{OO}(p) \cdot \left[\frac{\left(F_{P_{a'}}(p) - F_{P_a}(p)\right) \cdot \dfrac{\partial \mathbb{P}\left[S = 1, D = 0 \,|\, P(Z) = p\right]}{\partial p}}{\int_0^1 \left(F_{P_{a'}}(p) - F_{P_a}(p)\right) \cdot \dfrac{\partial \mathbb{P}\left[S = 1, D = 0 \,|\, P(Z) = p\right]}{\partial p} \, \mathrm{d}p} \right] \mathrm{d}p$$

by Equation (3.13),

implying that $\omega_{PRTE}(p, a, a') = \dfrac{\left(F_{P_{a'}}(p) - F_{P_a}(p)\right) \cdot \dfrac{\partial \mathbb{P}\left[S = 1, D = 0 \,|\, P(Z) = p\right]}{\partial p}}{\displaystyle\int_0^1 \left(F_{P_{a'}}(p) - F_{P_a}(p)\right) \cdot \dfrac{\partial \mathbb{P}\left[S = 1, D = 0 \,|\, P(Z) = p\right]}{\partial p} \, \mathrm{d}p}.$

A.7   Proof of Proposition 5

This proof is similar to that of Proposition 2. It is given for completeness. The validity of the bounds is proven in the main text. It remains to show that the bounds are uniformly sharp. Given the restrictions that Assumptions 1, 5, 6 and 8 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 any value $\delta \in \left[\underline{\Delta}_{LATE}, \overline{\Delta}_{LATE}\right]$. To do so, assume that Y^* is absolutely continuous and has a strictly positive density.

First, we show that the lower bound $\underline{\Delta}_{LATE} \colon \{2, \ldots, K\} \to \mathbb{R}$ is attainable. We need to define the joint density (mass) function of $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$. To do so, we will define the density functions $f_{\tilde{Y}_0^* | \tilde{S}_0, \tilde{S}_1, \tilde{V}}$, $f_{\tilde{Y}_1^* | \tilde{S}_0, \tilde{S}_1, \tilde{V}}$ and $f_{\tilde{V}}$, define the mass function $\tilde{\pi}_{(\tilde{S}_0, \tilde{S}_1) | \tilde{V}}$ and use the density function of Z — f_Z — to define $f_{\tilde{Y}_0^*, \tilde{Y}_1^*, (\tilde{S}_0, \tilde{S}_1), \tilde{V}, Z} = f_{\tilde{Y}_0^* | \tilde{S}_0, \tilde{S}_1, \tilde{V}} \cdot f_{\tilde{Y}_1^* | \tilde{S}_0, \tilde{S}_1, \tilde{V}} \cdot \tilde{\pi}_{(\tilde{S}_0, \tilde{S}_1) | \tilde{V}} \cdot f_{\tilde{V}} \cdot f_Z$. Note that, by construction, Assumption 1 holds. Fix $(y_0, y_1, p, z) \in \mathbb{R}^4$ arbitrarily. Define $f_{\tilde{V}}(p) = \mathbb{1}\{p \in [0, 1]\}$, ensuring that Assumption 5 holds by construction.

For brevity, denote the strata by $\underline{OO}$ = always observed, $\underline{NO}$ = observed only when treated and $\underline{NN}$ = never observed, and the probability of the stratum k conditional on $\tilde{V} = p$ by $\tilde{\pi}_{k|p}$. The probabilities $\tilde{\pi}_{k|p}$ are given by

$$
\begin{aligned}
\tilde{\pi}_{OO|p} &= \sum_{\ell=2}^{K} \mathbb{1}\{p_{\ell-1} < p \le p_\ell\} \cdot \left(-\frac{\mathbb{P}(S=1, D=0|P=p_\ell) - \mathbb{P}(S=1, D=0|P=p_{\ell-1})}{p_\ell - p_{\ell-1}} \right) \\
&\quad + \mathbb{1}\{p \le p_1\} \cdot \frac{\mathbb{P}(S=1, D=1|P=p_1)}{p_1} + \mathbb{1}\{p_K < p\} \cdot \frac{\mathbb{P}(S=1, D=0|P=p_K)}{1 - p_K}, \\
\tilde{\pi}_{NO|p} &= \sum_{\ell=2}^{K} \mathbb{1}\{p_{\ell-1} < p \le p_\ell\} \cdot \left(\frac{\mathbb{P}(S=1|P=p_\ell) - \mathbb{P}(S=1|P=p_{\ell-1})}{p_\ell - p_{\ell-1}} \right), \\
\tilde{\pi}_{NN|p} &= \left(\sum_{\ell=2}^{K} \mathbb{1}\{p_{\ell-1} < p \le p_\ell\} \cdot \frac{\mathbb{P}(S=0, D=1|P=p_\ell) - \mathbb{P}(S=0, D=1|P=p_\ell - 1)}{p_\ell - p_{\ell-1}} \right) \\
&\quad + \mathbb{1}\{p \le p_1\} \cdot \left(1 - \frac{\mathbb{P}(S=1, D=1|P=p_1)}{p_1} \right) \\
&\quad + \mathbb{1}\{p_K < p\} \cdot \left(1 - \frac{\mathbb{P}(S=1, D=0|P=p_K)}{1 - p_K} \right).
\end{aligned}
$$

Under Assumptions 1, 5, and 6, the above quantities are positive, and sum up to one. Indeed, for any $p \in [p_{\ell-1}, p_\ell]$ such that $\ell \in \{2, \ldots, K\}$.

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

implying that

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

$$= \frac{p_\ell - p_{\ell-1}}{p_\ell - p_{\ell-1}} = 1.$$

Now, define $f_{\tilde{Y}_0^*|\tilde{S}_0,\tilde{S}_1,\tilde{V}}(y_0|\,k,p) = \frac{\partial \mathbb{P}(\tilde{Y}_0^* \leq y_0|k,\tilde{V}=p)}{\partial y_0}$ and $f_{\tilde{Y}_1^*|\tilde{S}_0,\tilde{S}_1,\tilde{V}}(y_1|\,k,p) = \frac{\partial \mathbb{P}(\tilde{Y}_1^* \leq y_1|k,\tilde{V}=p)}{\partial y_1}$ for any $k \in \{OO, NO, NN\}$. Define

$$\mathbb{P}(\tilde{Y}_1^* \leq y_1, \tilde{S}_1 = 1|\tilde{V} = p) = \sum_{\ell=2}^{K} \mathbb{1}\{p_{\ell-1} < p \leq p_\ell\}$$
$$\cdot \left(\frac{\mathbb{P}(Y \leq y_1, S=1, D=1|P=p_\ell) - \mathbb{P}(Y \leq y_1, S=1, D=1|P=p_{\ell-1})}{p_\ell - p_{\ell-1}} \right)$$
$$+ \mathbb{1}\{p \leq p_1\} \cdot \frac{\mathbb{P}(Y \leq y_1, S=1, D=1|P=p_1)}{p_1}$$
$$+ \mathbb{1}\{p_K < p\} \cdot \frac{\mathbb{P}(Y \leq y_1, S=1, D=0|P=p_K)}{1-p_K},$$

$$\mathbb{P}(\tilde{Y}_0^* \leq y_0, \tilde{S}_0 = 1|\tilde{V} = p) = \sum_{\ell=2}^{K} \mathbb{1}\{p_{\ell-1} < p \leq p_\ell\}$$
$$\cdot \left(-\frac{\mathbb{P}(Y \leq y_0, S=1, D=0|P=p_\ell) - \mathbb{P}(Y \leq y_0, S=1, D=0|P=p_{\ell-1})}{p_\ell - p_{\ell-1}} \right)$$
$$+ \mathbb{1}\{p \leq p_1\} \cdot \frac{\mathbb{P}(Y \leq y_0, S=1, D=1|P=p_1)}{p_1}$$
$$+ \mathbb{1}\{p_K < p\} \cdot \frac{\mathbb{P}(Y \leq y_0, S=1, D=0|P=p_K)}{1-p_K}, \text{ and}$$
$$\mathbb{P}(\tilde{Y}_0^* \leq y_0|S_0 = 1, S_1 = 1, \tilde{V} = p) = \mathbb{P}(\tilde{Y}_0^* \leq y_0|\tilde{S}_0 = 1, \tilde{V} = p).$$

Define $\tilde{Y}_1 \sim F_{\tilde{Y}_1^*|\tilde{S}_1=1,\tilde{V}=p}$. Define also

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

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

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

$$\mathbb{P}(\tilde{Y}_0^* \leq y_0|k, \tilde{V} = p) = \mathbb{P}(\tilde{Y}_0^* \leq y_0, \tilde{S}_0 = 1|\tilde{V} = p) \text{ for } k \in \{NO, NN\}.$$

64

Notice that the lower bound in Proposition 5 is attained by the distributions of $\tilde{Y}_0^* \big| OO, \tilde{V}$ and $\tilde{Y}_0^* \big| OO, \tilde{V}$ because

$$\underline{\Delta}_{LATE}\left(\ell\right) = \mathbb{E}\left[\tilde{Y}_1^* - \tilde{Y}_0^* \big| OO, p_{\ell-1} < V \le p_\ell\right]$$

for any $\ell \in \{2, \ldots, K\}$.

Finally, we show that the joint distribution of $(\tilde{Y}_0^*, \tilde{Y}_1^*, \tilde{S}_0, \tilde{S}_1, \tilde{V}, Z)$ induces the joint distribution on the data (Y, S, D, Z). Let $\tilde{D} = \mathbb{1}\left\{\tilde{V} \le P\left(Z\right)\right\}$. For any $(y, \ell) \in \mathbb{R} \times \{2, \ldots, K\}$,

$$\mathbb{P}\left(\tilde{Y} \le y, \tilde{S} = 1, \tilde{D} = 1 \Big| Z = z_\ell\right) = \mathbb{P}\left(\tilde{Y}_1^* \le y, \tilde{S}_1 = 1, \tilde{V} \le p_\ell \Big| Z = z_\ell\right)$$

$$= \mathbb{P}\left(\tilde{Y}_1^* \le y, \tilde{S}_1 = 1, \tilde{V} \le p_\ell\right) = \int_0^{p_\ell} \mathbb{P}\left(\tilde{Y}_1^* \le y, \tilde{S}_1 = 1 \Big| \tilde{V} = v\right) dv$$

$$= \int_0^{p_\ell} \left[\begin{array}{c} \sum_{k=2}^\ell \mathbb{1}\left\{p_{k-1} < v \le p_k\right\} \cdot \frac{\mathbb{P}(Y \le y_1, S=1, D=1 | P=p_k) - \mathbb{P}(Y \le y_1, S=1, D=1 | P=p_{k-1})}{p_k - p_{k-1}} \\ + \mathbb{1}\left\{v \le p_1\right\} \cdot \frac{\mathbb{P}(Y \le y_1, S = 1, D = 1 | P = p_1)}{p_1} \end{array} \right] dv$$

$$= \mathbb{P}\left(Y \le y, S = 1, D = 1 | P = p_\ell\right) = \mathbb{P}\left(Y \le y, S = 1, D = 1 | Z = z_\ell\right)$$

and, analogously,

$$\mathbb{P}\left(\tilde{Y} \le y, \tilde{S} = 1, \tilde{D} = 0 \Big| Z = z_\ell\right) = \mathbb{P}\left(Y \le y, S = 1, D = 0 | Z = z_\ell\right),$$

$$\mathbb{P}\left(\tilde{S} = 0, \tilde{D} = 1 \Big| Z = z_\ell\right) = \mathbb{P}\left(S = 0, D = 1 | Z = z_\ell\right),$$

$$\mathbb{P}\left(\tilde{S} = 0, \tilde{D} = 0 \Big| Z = z_\ell\right) = \mathbb{P}\left(S = 0, D = 0 | Z = z_\ell\right).$$

Similar reasoning holds for the upper bound, $\overline{\Delta}_{LATE}$. To attain any function $\delta \in \left(\underline{\Delta}_{LATE}, \overline{\Delta}_{LATE}\right)$, we can use convex combinations of the joint distributions that achieve the lower and upper bounds, where the weights of the convex combination may depend on the value $\ell \in \{2, \ldots, K\}$.

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.

Proposition 6. *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

$$0 \le \frac{\mathbb{P}(Y \in A, S = 1, D = 1|Z = 1) - \mathbb{P}(Y \in A, S = 1, D = 1|Z = 0)}{P(1) - P(0)} \le 1,$$

$$0 \le -\frac{(\mathbb{P}(Y \in A, S = 1, D = 0|Z = 1) - \mathbb{P}(Y \in A, S = 1, D = 0|Z = 0))}{P(1) - P(0)} \le 1,$$

$$0 \le \frac{\mathbb{P}(S = 1|Z = 1) - \mathbb{P}(S = 1|Z = 0)}{P(1) - P(0)} \le 1.$$

These latter 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), Huber and Mellace (2015) or Laffers and Mellace (2017).

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.

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

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

$$\mathbb{P}\left(\tilde{Y}_1^* \leq y_1, \tilde{S}_1 = 1 \middle| \tilde{V} = p\right) = \begin{cases} \frac{\partial \mathbb{P}(Y \leq y_1, S=1, D=1 | P(Z)=p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\[2mm] \frac{\mathbb{P}\left(Y \leq y_1, S=1, D=1 | P(Z)=\underline{p}\right)}{\underline{p}} & \text{if } p < \underline{p} \\[2mm] \frac{\mathbb{P}(Y \leq y_1, S=1, D=0 | P(Z)=\overline{p})}{1-\overline{p}} & \text{if } p > \overline{p} \end{cases}$$

$$\mathbb{P}\left(\tilde{S}_1 = 0 \middle| \tilde{V} = p\right) = \begin{cases} \frac{\partial \mathbb{P}(S=0, D=1 | P(Z)=p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\[2mm] \frac{\mathbb{P}\left(S=0, D=1 | P(Z)=\underline{p}\right)}{\underline{p}} & \text{if } p < \underline{p} \\[2mm] \frac{\mathbb{P}(S=0, D=0 | P(Z)=\overline{p})}{1-\overline{p}} & \text{if } p > \overline{p} \end{cases}$$

$$\mathbb{P}\left(\tilde{Y}_0^* \leq y_0, \tilde{S}_0 = 1 \middle| \tilde{V} = p\right) = \begin{cases} -\frac{\partial \mathbb{P}(Y \leq y_0, S=1, D=0 | P(Z)=p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\[2mm] \epsilon \frac{\mathbb{P}\left(Y \leq y_0, S=1, D=1 | P(Z)=\underline{p}\right)}{\underline{p}} & \text{if } p < \underline{p} \\[2mm] \frac{\mathbb{P}(Y \leq y_0, S=1, D=0 | P(Z)=\overline{p})}{1-\overline{p}} & \text{if } p > \overline{p} \end{cases}$$

and

$$\mathbb{P}\left(\tilde{S}_0 = 0 \middle| \tilde{V} = p\right) = \begin{cases} -\frac{\partial \mathbb{P}(S=0, D=0 | P(Z)=p)}{\partial p} & \text{if } p \in [\underline{p}, \overline{p}] \\[2mm] 1 - \epsilon \frac{\mathbb{P}\left(S=1, D=1 | P(Z)=\underline{p}\right)}{\underline{p}} & \text{if } p < \underline{p} \\[2mm] \frac{\mathbb{P}(S=0, D=0 | P(Z)=\overline{p})}{1-\overline{p}} & \text{if } p > \overline{p} \end{cases}$$

where $\epsilon \in (0,1)$.

Define the distribution on the strata:

$$\tilde{\pi}_{OO|p} = \mathbb{P}(\tilde{S}_0 = 1 | \tilde{V} = p)$$

$$\tilde{\pi}_{NO|p} = \mathbb{P}(\tilde{S}_1 = 1 | \tilde{V} = p) - \mathbb{P}(\tilde{S}_0 = 1 | \tilde{V} = p)$$

$$\tilde{\pi}_{NN|p} = \mathbb{P}(\tilde{S}_1 = 0 | \tilde{V} = p).$$

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

one. Define

$$\mathbb{P}(\tilde{Y}_1^* \le y_1|k, \tilde{V} = p, Z = z) = \mathbb{P}(\tilde{Y}_1^* \le y_1|\tilde{S}_1 = 1, \tilde{V} = p), \quad k \in \{OO, NO, NN\}$$

and

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

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

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

Finally, define

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

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

$$
\begin{aligned}
\mathbb{P}(\tilde{Y} \le y, \tilde{S} = 1, \tilde{D} = 1|Z = z) &= \mathbb{P}(\tilde{Y}_1^* \le y, \tilde{S}_1 = 1, \tilde{V} \le P(z)|Z = z), \\
&= \mathbb{P}(\tilde{Y}_1^* \le y, \tilde{S}_1 = 1|\tilde{V} \le P(z), Z = z)\mathbb{P}(\tilde{V} \le P(z)|Z = z), \\
&= \mathbb{P}(\tilde{Y}_1^* \le y, \tilde{S}_1 = 1|\tilde{V} \le P(z), Z = z)P(z), \\
&= \left[\int_0^{P(z)} \mathbb{P}(\tilde{Y}_1^* \le y, \tilde{S}_1 = 1|\tilde{V} = v, Z = z)\frac{f_{\tilde{V}|Z=z}(v)}{P(z)}dv\right] \cdot P(z), \\
&= \int_0^{P(z)} \mathbb{P}(\tilde{Y}_1^* \le y, \tilde{S}_1 = 1|\tilde{V} = v, Z = z)dv,
\end{aligned}
$$

69

$$
\begin{aligned}
&= \int_0^{P(z)} \mathbb{P}(\tilde{Y}_1^* \le y, \tilde{S}_1 = 1 | \tilde{V} = v) dv, \\
&= \int_0^{\underline{p}} \frac{\mathbb{P}\left(Y \le y_1, S = 1, D = 1 | P(Z) = \underline{p}\right)}{\underline{p}} dv \\
&\quad + \int_{\underline{p}}^{P(z)} \frac{\partial \mathbb{P}\left(Y \le y_1, S = 1, D = 1 | P(Z) = p\right)}{\partial p} dv, \\
&= \mathbb{P}(Y \le y, S = 1, D = 1 | P(Z) = P(z)) \\
&= \mathbb{P}(Y \le y, S = 1, D = 1 | Z = z).
\end{aligned}
$$

Similarly,

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

Finally, by construction, Assumptions 1, 5 and 6 hold. $\qquad\square$

C   Numerical Illustration

In this appendix, 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. In Appendix C.1, we present the bounds for MTE^{OO} based on Propositions 1 and 2. The illustration provides insights on the functioning of the bounds as well as the mechanisms driving how informative those bounds are. In Appendix C.2, we discuss what is identified by the *local instrumental variable* (LIV) estimand (Heckman and Vytlacil, 1999) when directly applied to the selected sample and explain its limitations. Appendix C.3 details the data-generating process used in this numerical illustration.

C.1   Illustrating the Bounds around MTE^{OO}

Consider the following data generating process (DGP):

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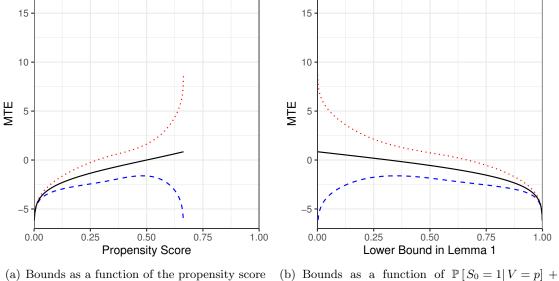
we set

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

where $(\theta, \epsilon_S, Z, \xi)' \sim N(0, I)$, $T = \mathbb{1}\{\xi \ge 0\}$, 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. Intuitively, there are two sets of individuals that might face different returns to treatment due to, for example, their gender or race. From a technical standpoint, this choice guarantees reasonable overlap for treated and untreated groups in the observed population over the support of 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 3, and Proposition 2 $(\underline{\Delta}_2, \overline{\Delta}_2)$ in Figure 4 for parameters $\delta_0 = 0.1$, $\delta_1 = 0.4$, $\beta_{0,0} = \beta_{0,1} = \beta_{1,0} = 1$ and $\beta_{1,1} = 5$.[23]

As can be seen on Subfigure 3(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 0.664, the lower bound on the proportion of the always-observed becomes exactly zero, the

---

[23]Note that these bounds can be computed using numeric integration. See Appendix C.3 for more details.

(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 3: Numerical Bounds based on Proposition 1

MTE^{OO} is not identified and the bounds diverge. Nevertheless, the sign of MTE^{OO} is identified for propensity scores smaller than 0.28.
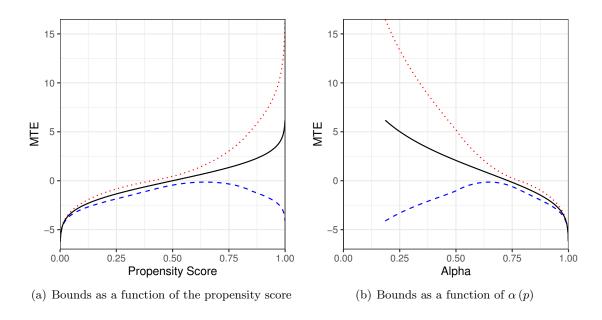
Subfigure 3(b) plots the identified interval for MTE^{OO} against v^{ℓ} on the horizontal axis, emphasizing the important role of the lower bound on the proportion of the always-observed. As v^{ℓ} increases, the expectation of the observed outcomes conditional on $V = p$ becomes heavily composed by the always-observed group, leading to point identification of the MTE^{OO} when it reaches one.

Figure 4 plots the MTE^{OO} and its bounds based on Proposition 2, i.e., with the addition of the monotonicity assumption. The bounds presented on Subfigure 4(a) are in general informative. As discussed 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} identified set is very tight. When p is close to 1, the proportion of the always-observed decreases and the identified set around the MTE^{OO} expands. The sign of the MTE^{OO} is identified for

$p < 0.409$, illustrating that Assumption 6 allow us to identify the sign of MTE^{OO} in more cases than in Figure 3.

Subfigure 4(b) plots the same curves with $\alpha(p)$ on the horizontal axis, emphasizing the trimming proportion's importance. 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, under Assumption 6, $\alpha(p)$ never reaches zero, allowing us to non-trivially bound the MTE^{OO} for all values of the propensity score.
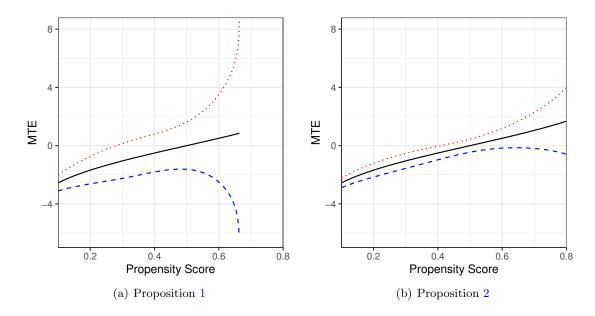
Figure 5 is a zoomed version of Subfigures 3(a) and 4(a). Note that the bounds based on Proposition 2 are much tighter than the bounds based on Proposition 1, especially for larger values of p. This is expected as the difference in trimming proportions in Proposition 1 and Proposition 2 increases with the propensity score for this DGP.[24]

(a) Bounds as a function of the propensity score   (b) Bounds as a function of $\alpha(p)$

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 1,000,000 simulated points for each value of the propensity score.

Figure 4: Numerical Bounds based on Proposition 2

---

[24]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 Proposition 3.

(a) Proposition 1	(b) Proposition 2

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 1,000,000 simulated points for each value of the propensity score.

Figure 5: Comparing Proposition 1 and Proposition 2

C.2 Understanding the LIV Estimand in the Selected Sample

A researcher may believe that sample selection is not a problem and decide to apply the *local instrumental variable* (LIV) estimator (Heckman and Vytlacil, 1999) directly to the selected sample ($S = 1$). Similarly to the issues generated by sample selection when trying to identify ATE or $LATE$ parameters (Heckman, 1979; Lee, 2009; Chen and Flores, 2015), the LIV estimand applied directly to the selected sample will not identify our target parameter (MTE^{OO}). The LIV estimand in this case can be described as:

$$
\begin{aligned}
LIV_{S=1}\left(p\right) &:= \frac{\partial \mathbb{E}\left[Y\middle| P\left(Z\right)=p, S=1\right]}{\partial p} \\
&= \frac{\partial}{\partial p}\left\{\frac{\mathbb{E}\left[Y \cdot S\middle| P\left(Z\right)=p\right]}{\mathbb{E}\left[S\middle| P\left(Z\right)=p\right]}\right\} \\
&= \frac{1}{\mathbb{E}\left[S\middle| P\left(Z\right)=p\right]} \cdot \frac{\partial \mathbb{E}\left[Y \cdot S\middle| P\left(Z\right)=p\right]}{\partial p} - \frac{\mathbb{E}\left[Y \cdot S\middle| P\left(Z\right)=p\right]}{\left(\mathbb{E}\left[S\middle| P\left(Z\right)=p\right]\right)^{2}} \cdot \frac{\partial \mathbb{E}\left[S\middle| P\left(Z\right)=p\right]}{\partial p} \\
&= \frac{\mathbb{E}\left[Y_{1}-Y_{0}\middle| V=p\right]}{\mathbb{E}\left[S\middle| P\left(Z\right)=p\right]} - \frac{\mathbb{E}\left[Y \cdot S\middle| P\left(Z\right)=p\right]}{\left(\mathbb{E}\left[S\middle| P\left(Z\right)=p\right]\right)^{2}} \cdot \mathbb{E}\left[S_{1}-S_{0}\middle| V=p\right] \quad \text{(C.3)}
\end{aligned}
$$

for any $p \in \mathcal{P}$ such that $\mathbb{E}\left[S|\,P\left(Z\right) = p\right] \neq 0$. As Equation (C.3) illustrates, the LIV estimand captures a linear combination of the marginal treatment effect on the non-censored outcome variable ($\mathbb{E}\left[Y_1 - Y_0|\,V = p\right]$) and of the marginal treatment effect on the sample selection ($\mathbb{E}\left[S_1 - S_0|\,V = p\right]$). Since the weights on this linear combination can be negative, the LIV estimand does not identify an interpretable treatment effect parameter when applied directly to the selected sample. For instance, observe that in the job training example and in the health insurance example (Section 5), the weight on the marginal treatment effect on the sample selection $\left(-\dfrac{\mathbb{E}\left[Y \cdot S|\,P\left(Z\right) = p\right]}{\left(\mathbb{E}\left[S|\,P\left(Z\right) = p\right]\right)^2}\right)$ must be negative, since Y and S are non-negative. Note also that, even when there is no differential sample selection ($\mathbb{E}\left[S_1 - S_0|\,V = p\right] = 0$), the LIV estimand still does not have a clear interpretation. In this case, it will capture the marginal treatment effect on the non-censored outcome variable magnified by the probability of being selected into the sample.

Figure 6 illustrates the differences between the LIV estimand and our target parameter — marginal treatment effect for the always-observed subpopulation (MTE^{OO}). In Subfigure 6(a), the dashed line is the MTE^{OO} function explained in Appendix C.1, while the solid line is the LIV estimand described in Equation (C.3). Note that the target parameter is monotone and highly heterogeneous, assuming large positive and negative values. However, the LIV estimand is non-monotone and only assumes small positive values, hiding important aspects related to the heterogeneity captured by the MTE^{OO}. As a consequence, the LIV estimand's bias $\left(LIV_{S=1}\left(p\right) - MTE^{OO}\left(p\right)\right)$ is large as illustrated in Subfigure 6(b). To have a better understanding of the relevance of the LIV estimand's bias, we can compare this estimand against the bounds based on Proposition 2. These bounds are depicted in Subfigure 6(a) as dotted lines. Note that LIV estimand's bias is so large that this estimand is outside the bounds based on Proposition 2 for most values of the propensity score.

Since the LIV estimand applied directly to the selected sample does not capture an intepretable treatment effect parameter and would be biased as a measure for the MTE^{OO}, we do not recommend using it when there is sample selection. When faced with selection-into-treatment and sample selection, the researcher should use a methodology that addresses

(a) LIV Estimand and the MTE^{OO} (b) LIV Estimand's Bias

Notes: In Subfigure 6(a), the solid line is the LIV estimand (Equation (C.3)) computed by numerical integration using 600,000 simulated points while the dashed line is the true value of the MTE^{OO}. The dotted lines are the upper and lower bounds around the MTE^{OO} from Subfigure 4(a). In Subfigure 6(b), the solid line is the LIV estimand's bias in comparison with the target parameter (MTE^{OO}).

Figure 6: Understanding the LIV Estimand in the Selected Sample

both challenges.

C.3 Details on the Numerical Illustration

This appendix lists the relevant densities, expectations and objects of interest implied by the DGP used in Appendices C and F. We have that

$$\mathbb{P}\left[S_0 = 1, S_1 = 1 \middle| V = p\right] = \Phi\left(\delta_0\sqrt{2} - \Phi^{-1}(p)\right),$$

$$\mathbb{P}\left[S_0 = 1 \middle| V = p\right] = \Phi\left(\delta_0\sqrt{2} - \Phi^{-1}(p)\right),$$

$$\mathbb{P}\left[S_1 = 1 \middle| V = p\right] = \Phi\left(\delta_0\sqrt{2} + \delta_1\sqrt{2} - \Phi^{-1}(p)\right),$$

$$\alpha\left(p, v^\ell\right) = \max\left\{1 + \frac{\Phi\left(\delta_0\sqrt{2} - \Phi^{-1}(p)\right) - 1}{\Phi\left(\delta_0\sqrt{2} + \delta_1\sqrt{2} - \Phi^{-1}(p)\right)}, 0\right\},$$

$$\beta\left(p, v^\ell\right) = \max\left\{1 + \frac{\Phi\left(\delta_0\sqrt{2} + \delta_1\sqrt{2} - \Phi^{-1}(p)\right) - 1}{\Phi\left(\delta_0\sqrt{2} - \Phi^{-1}(p)\right)}, 0\right\},$$

$$\alpha\left(p\right)=\frac{\Phi\left(\delta_0\sqrt{2}-\Phi^{-1}\left(p\right)\right)}{\Phi\left(\delta_0\sqrt{2}+\delta_1\sqrt{2}-\Phi^{-1}\left(p\right)\right)},$$

$$E\left[Y_1^*-Y_0^*\,|\,T=1,S_0=1,S_1=1,V=p\right]=\left(\beta_{1,1}-\beta_{0,1}\right)\cdot\Phi^{-1}(p),$$

$$E\left[Y_1^*-Y_0^*\,|\,T=0,S_0=1,S_1=1,V=p\right]=\left(\beta_{0,0}-\beta_{1,0}\right)\cdot\Phi^{-1}(p),$$

$$E\left[Y_1^*-Y_0^*\,|\,S_0=1,S_1=1,V=p\right]=\left(\beta_{1,1}-\beta_{1,0}-\beta_{0,1}+\beta_{0,0}\right)\cdot\frac{\Phi^{-1}(p)}{2},$$

$$P\left[Y_0^*\le y\,|\,T=1,S_0=1,V=p\right]=\Phi\left(y-\beta_{0,1}\Phi^{-1}\left(p\right)\right),$$

$$P\left[Y_1^*\le y\,|\,T=1,S_1=1,V=p\right]=\Phi\left(y-\beta_{1,1}\Phi^{-1}\left(p\right)\right),$$

$$P\left[Y_0^*\le y\,|\,T=0,S_0=1,V=p\right]=\Phi\left(y+\beta_{0,0}\Phi^{-1}\left(p\right)\right),$$

$$P\left[Y_1^*\le y\,|\,T=0,S_1=1,V=p\right]=\Phi\left(y+\beta_{1,0}\Phi^{-1}\left(p\right)\right),$$

$$P\left[Y_0^*\le y\,|\,S_0=1,V=p\right]=\frac{1}{2}\Phi\left(y-\beta_{0,1}\Phi^{-1}\left(p\right)\right)+\frac{1}{2}\Phi\left(y+\beta_{0,0}\Phi^{-1}\left(p\right)\right),$$

$$P\left[Y_1^*\le y\,|\,S_1=1,V=p\right]=\frac{1}{2}\Phi\left(y-\beta_{1,1}\Phi^{-1}\left(p\right)\right)+\frac{1}{2}\Phi\left(y+\beta_{1,0}\Phi^{-1}\left(p\right)\right).$$

D Two Economic Models Satisfying Assumptions 1-7

D.1 Model 1: Job Training Program

To better understand the intuition behind the stochastic dominance assumption, we provide a simple economic model of a worker deciding to enroll in a job training program that satisfies assumptions 1-7.

Let Y_0^* be the untreated wage whose support satisfy $\mathcal{Y}_0^*=[0,c_0\cdot\bar{y})$, where $c_0\in(1,+\infty)$ and $\bar{y}\in\mathbb{R}_{++}$. Let $\tilde{V}\sim F_{\tilde{V}}$ be the training program's effect on wages whose support satisfy $\mathcal{V}=[0,\bar{v}]$, where $\bar{v}\in\mathbb{R}_{++}$, and whose cumulative distribution function is strictly increasing in its support. To simplify our argument, assume that $\tilde{V}\perp\!\!\!\perp Y_0^*$. Define $Y_1^*:=Y_0^*+\tilde{V}$ as the treated wage. Let Z be the training program's cash allowance whose support is given by $\mathcal{Z}=\mathbb{R}_+$ and whose cumulative distribution function is continuous. To further simplify our argument, assume that the support of the joint distribution of $\tilde{V}$ and Z satisfy $\text{supp}\left(\tilde{V},Z\right)=\mathcal{Z}\times\mathcal{V}$. The worker enrolls in the training program if its total benefits are larger than its non-random and homogeneous cost, $c_D\in\mathbb{R}_{++}$. Formally, define $D:=\mathbb{1}\left\{Z+\tilde{V}\ge c_D\right\}=$

$\mathbb{1}\left\{F_{-\tilde{V}}\left(Z-c_D\right) \geq F_{-\tilde{V}}\left(-\tilde{V}\right)\right\} = \mathbb{1}\left\{P\left(Z\right) \geq V\right\}$, where $P\left(Z\right) = F_{-\tilde{V}}\left(Z-c_D\right)$ and $V = F_{-\tilde{V}}\left(-\tilde{V}\right)$. Moreover, the worker is hired if her productivity is high enough, i.e., $S_d := \mathbb{1}\left\{Y_d^* \geq \overline{y}\right\}$ for any $d \in \{0,1\}$. Finally, assume that the cash allowance is randomly chosen so that Assumption 1 holds by design.

Now, we check whether Assumptions 2-7 hold in this economic model. First, observe that Assumption 6 holds because $\mathcal{V} = \mathbb{R}_+$ and $S_1 = \mathbb{1}\left\{Y_0^* + \tilde{V} \geq \overline{y}\right\}$. Assumption 2 holds because $F_{\tilde{V}}$ is a strictly increasing function and Z is a absolutely continuous random variable. For Assumption 3, note that $\mathrm{supp}\left(\tilde{V}, Z\right) = [0, \overline{v}] \times \mathbb{R}_+$ and $c_D > 0$ imply $\mathbb{P}\left[Z + \tilde{V} \geq c_D\right] > 0$ and $\mathbb{P}\left[Z + \tilde{V} < c_D\right] > 0$, and because $S_1 \geq S_0$, $\mathcal{Y}_0^* = [0, c_0 \cdot \overline{y})$, $c_0 > 1$ and $\tilde{V} \perp\!\!\!\perp Y_0^*$ imply that $\mathbb{P}\left[S_0 = 1, S_1 = 1 | V = p\right] = \mathbb{P}\left[S_0 = 1 | V = p\right] = \mathbb{P}\left[Y_0^* \geq \overline{y} | V = p\right] = \mathbb{P}\left[Y_0^* \geq \overline{y}\right] > 0$ for any $p \in \mathcal{P}$. Assumption 4 holds because $\sup|\mathcal{Y}_0^*| = c_0 \cdot \overline{y} < +\infty$ and $\sup|\mathcal{Y}_1^*| = c_0 \cdot \overline{y} + \overline{v} < +\infty$, while Assumption 5 holds by definition.

Finally, we discuss whether the stochastic dominance assumption (Assumption 7) holds. Since $\tilde{V} \perp\!\!\!\perp Y_0^*$, we have, for any $y \in \mathbb{R}_+$ and $p \in \mathcal{P}$, that

$$\begin{aligned}
\mathbb{P}\left[Y_1^* \leq y | V = p, S_0 = 1, S_1 = 1\right] &= \mathbb{P}\left[Y_1^* \leq y \middle| F_{-\tilde{V}}\left(-\tilde{V}\right) = p, Y_0^* \geq \overline{y}\right] \\
&= \mathbb{P}\left[Y_0^* - F_{-\tilde{V}}^{-1}\left(p\right) \leq y \middle| F_{-\tilde{V}}\left(-\tilde{V}\right) = p, Y_0^* \geq \overline{y}\right] \\
&= \mathbb{P}\left[Y_0^* \leq y + F_{-\tilde{V}}^{-1}\left(p\right) \middle| Y_0^* \geq \overline{y}\right]
\end{aligned}$$

and

$$\mathbb{P}\left[Y_1^* \leq y | V = p, S_0 = 0, S_1 = 1\right] = \mathbb{P}\left[Y_0^* \leq y + F_{-\tilde{V}}^{-1}\left(p\right) \middle| \overline{y} + F_{-\tilde{V}}^{-1}\left(p\right) \leq Y_0^* < \overline{y}\right].$$

As a consequence, we have that

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

2. if $y > \overline{y}$ and $y + F_{-\tilde{V}}^{-1}\left(p\right) \leq \overline{y}$, then $\mathbb{P}\left[Y_1^* \leq y | V = p, S_0 = 1, S_1 = 1\right] = 0$ and $\mathbb{P}\left[Y_1^* \leq y | V = p, S_0 = 0, S_1 = 1\right] \in (0, 1]$; and

3. if $y > \bar{y}$ and $y + F_{-\tilde{V}}^{-1}(p) > \bar{y}$, then $\mathbb{P}\left[Y_1^* \leq y | V = p, S_0 = 1, S_1 = 1\right] \in (0, 1]$ and

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

We can, then, conclude that Assumption 7 holds in this simple economic model.

D.2 Model 2: Managed Care and Ambulatory Expenditures

Suppose that an individual's decision to choose a managed care plan (HMO or PPO) is based on a cost-benefit analysis, such that she decides to enroll in a managed care plan if its cost V is less than its benefit $P(Z)$, i.e., $D = \mathbb{1}\{V < P(Z)\}$. The individual's potential ambulatory expenditure when she enrolls in a managed care plan is Y_1^*, and her potential ambulatory expenditure when she enrolls in a fee-for-service plan is Y_0^*, where $Y_d^* \geq 0$ with $\mathbb{P}(Y_d^* = 0) > 0$ for all d. We observe the individual's ambulatory expenditure when $Y^* = Y_1^* D + Y_0^*(1 - D)$ is positive, i.e., $S = \mathbb{1}\{Y^* > 0\}$, and the observed ambulatory expenditure is $Y = Y^* S$. Suppose now that $Y_1^* = Y_0^* + \varepsilon$, where $\varepsilon \perp\!\!\!\perp Y_0^* | V$, and $\varepsilon | V = p \sim \mathcal{U}_{[0, 1/p]}$. We are going to show that Assumptions 6 and 7 hold in this framework. By definition, $S_0 = \mathbb{1}\{Y_0^* > 0\}$ and $S_1 = \mathbb{1}\{Y_1^* > 0\} = \mathbb{1}\{Y_0^* + \varepsilon > 0\}$. We can see that $\{S_0 = 1\}$ implies $\{S_1 = 1\}$ since $\varepsilon \geq 0$. Therefore, Assumption 6 holds. We now show that

$$\mathbb{P}(Y_1^* \leq y | V = p, S_0 = 1, S_1 = 1) \leq \mathbb{P}(Y_1^* \leq y | V = p, S_0 = 0, S_1 = 1).$$

We have

$$
\begin{aligned}
\mathbb{P}(Y_1^* \leq y | V = p, S_0 = 0, S_1 = 1) &= \mathbb{P}(Y_1^* \leq y | V = p, Y_0^* = 0, Y_1^* > 0), \\
&= \mathbb{P}(\varepsilon \leq y | V = p, \varepsilon > 0) \text{ since } \varepsilon \perp\!\!\!\perp Y_0^* | V, \\
&= \begin{cases} p.y & \text{if } y \geq 0 \\ 0 & \text{if } y < 0 \end{cases}
\end{aligned}
$$

and

$$
\mathbb{P}(Y_1^* \leq y | V = p, S_0 = 1, S_1 = 1) = \int_{y_0 > 0} \mathbb{P}(Y_1^* \leq y | V = p, Y_0^* > 0, Y_0^* = y_0, Y_1^* > 0) \cdot
$$
$$
f_{Y_0^* | V = p, Y_0^* > 0, Y_1^* > 0}(y_0) dy_0,
$$

$$= \int_{y_0>0} \mathbb{P}(\varepsilon \leq y - y_0|V=p, \varepsilon > -y_0) \cdot$$

$$f_{Y_0^*|V=p,Y_0^*>0,Y_1^*>0}(y_0)dy_0 \ \text{ since } \ \varepsilon \perp\!\!\!\perp Y_0^*|V,$$

$$= \int_{y_0>0} \mathbb{P}(\varepsilon \leq y - y_0|V=p) \cdot f_{Y_0^*|V=p,Y_0^*>0,Y_1^*>0}(y_0)dy_0$$

and

$$\mathbb{P}(\varepsilon \leq y - y_0|V=p) \ = \ \begin{cases} p(y-y_0) \ \text{ if } \ y > y_0 \\ 0 \ \text{ if } \ y \leq y_0 \end{cases}$$

$$\leq \ \begin{cases} p.y \ \text{ if } \ y \geq 0 \text{ and } y_0 > 0 \\ 0 \ \text{ if } \ y < 0 \end{cases}$$

Therefore,

$$\mathbb{P}(Y_1^* \leq y|V=p, S_0 = 1, S_1 = 1) \ \leq \ \begin{cases} \int_{y_0>0} p.y f_{Y_0^*|V=p,Y_0^*>0,Y_1^*>0}(y_0)dy_0 \ \text{ if } \ y \geq 0 \\ 0 \ \text{ if } \ y < 0 \end{cases}$$

$$= \ \begin{cases} p.y \ \text{ if } \ y \geq 0 \\ 0 \ \text{ if } \ y < 0 \end{cases}$$

$$= \ \mathbb{P}(Y_1^* \leq y|V=p, S_0 = 0, S_1 = 1).$$

Hence, Assumption 7 holds.

E   Estimation Details

This appendix presents the details on how to estimate the bounds proposed in Proposition 2, building upon the discussion on Section 4 on the main text. For brevity, we focus on the bounds identified under monotonicity of sample selection in the treatment (Assumptions 1-6), as it is the most relevant (and feasible) case empirically. Estimators for the bounds proposed in Propositions 1 and 3 are natural extensions of the estimator discussed here.

The following subsections (E.1-E.4) present the details on a suggested non(semi)parametric estimation approach that could be implemented with a large and informative dataset.[25] Nat-

---

[25] Appendix F presents a Monte Carlo Simulation that evaluates the small sample properties of this estimator.

urally, parametric alternatives are available as long as the researcher is willing to specify functional forms for the probabilities described below, imposing some structure to the model. This can be an useful alternative which greatly simplifies estimation, especially when conditioning on covariates. For concreteness, we present the details of the particular parametric estimation procedure implemented in the empirical application (Section 5) at Appendix G.

Recall from Section 4 that we need estimates for:

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

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:

$$\Gamma_1(p, y) := \frac{\partial \mathbb{P}[Y \leq y, S = 1, D = 1| P(Z) = p]}{\partial p}, \quad \pi_1(p) := \frac{\partial \mathbb{P}[S = 1, D = 1| P(Z) = p]}{\partial p},$$

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

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

E.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, since treatment status D is observed for all individuals, and are summarized here. We suggest to 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 is always between 0 and 1.

E.2 Estimating $\pi_1(p)$, $\pi_0(p)$ and $\alpha(p)$

To estimate $\pi_1(p)$ and $\pi_0(p)$, consider the local polynomial estimators (Fan and Gijbels, 1996):

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

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. When implementing these estimators, we recommend using recent developments in Calonico, Cattaneo, and Farrell (2018, 2019a) for optimal coverage error bandwidth and kernel selection methods that are nonparametric robust bias-corrected (RBC). We suggest implementing the RBC version of these estimators with optimal bandwidth choice as conveniently implemented in the software R, using the package *nprobust* (Calonico, Cattaneo, and Farrell, 2019b).

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

E.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](#)),

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

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

for any $k \in \{2, \ldots, K_n\}$. Similarly to the estimation of $\hat{\pi}_0$ and $\hat{\pi}_1$, we implement the optimal bandwidth selector and RBC procedures mentioned previously.

For simplicity, let $\hat{f}_d(p,k) = \frac{\hat{\gamma}_d(p,k)}{\hat{\pi}_d(p)}$ for any $k \in \{2, \ldots, K_n\}$. Natural estimators for $F_1(p,y) := \frac{\Gamma_1(p,y)}{\pi_1(p)}$ and $F_0(p,y) := \frac{\Gamma_0(p,y)}{\pi_0(p)}$ are given by $\hat{F}_1(p,y_k) := \sum_{j=2}^{k} \hat{f}_1(p,j)$ and $\hat{F}_0(p,y_k) := \sum_{j=2}^{k} \hat{f}_0(p,j)$.

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 among both treated and untreated individuals. One can mitigate the challenges to feasibility of the estimator by choosing wider bins $[y_{k-1} \le Y_i \le y_k]$, 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.

E.4   Estimating $MTE^{OO}(p)$ Bounds

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

$$\widehat{LB}_2(p) := \sum_{k=2}^{K_n} \overline{y}_k \cdot \mathbb{1}\left\{ \hat{F}_1(p, y_k) \le \hat{\alpha}(p) \right\} \cdot \frac{\hat{f}_1(p,k)}{\hat{\alpha}(p)} \tag{E.1}$$

$$\widehat{UB}_2(p) := \sum_{k=2}^{K_n} \overline{y}_k \cdot \mathbb{1}\left\{ 1 - \hat{F}_1(p, y_k) < \hat{\alpha}(p) \right\} \cdot \frac{\hat{f}_1(p,k)}{\hat{\alpha}(p)}, \tag{E.2}$$

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 Proposition [2](#) using $\hat{\Xi}_{OO,0}(p) := \sum_{k=2}^{K_n} \overline{y}_k \cdot \hat{f}_0(p,k)$.

Note that the estimators for $\widehat{LB}_2(p), \widehat{UB}_2(p)$ and $\hat{\Xi}_{OO,0}(p)$ rely on the estimates for den-

sities and trimming points obtained in previous steps, not using on the original data for the observed outcomes.

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

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.

F   Monte Carlo Simulation

In this appendix, we use the DGP described in Appendix C 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 nonparametric estimation procedure described in the Appendix E, 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: $\hat{\alpha}(p)$, $\hat{\Xi}_{OO,0}(p)$, $\widehat{LB}_2(p)$, $\widehat{UB}_2(p)$, $\underline{\hat{\Delta}}_2(p)$ and $\overline{\hat{\Delta}}_2(p)$. Table D.1 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 D.2 reports the average bias, while Table D.3 presents the mean squared error (MSE) 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 (high) values of the propensity score for the (un)treated group. As a consequence, 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 D.1: True Value of the Estimands

	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: $\delta_0 = 0.75$, $\delta_1 = 1.5$, $\beta_{00} = \beta_{01} = \beta_{10} = 0.1$, $\beta_{11} = 0.2$						
$\alpha(p)$	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 $\Xi_{OO,0}(p) := \mathbb{E}[Y_0^* | S_0 = 1, S_1 = 1, V = p]$. The true values of the estimands are computed by numerical integration using 100,000 simulated points for each value of the propensity score.

Table D.2: Average Bias: $\mathbb{E}\left[\hat{\theta}\right] - \theta_0$

	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: $\delta_0 = 0.75, \delta_1 = 1.5, \beta_{00} = \beta_{01} = \beta_{10} = 0.1, \beta_{11} = 0.2$								
$\hat{\alpha}(p)$	-0.02 (0.07)	0.01 (0.03)	0.01 (0.03)	-0.01 (0.03)	-0.01 (0.03)	-0.02 (0.02)	-0.02 (0.02)	0.00 (0.03)	0.01 (0.05)
$\hat{\Xi}_{OO,0}(p)$	-0.02 (0.32)	-0.02 (0.14)	-0.01 (0.08)	0 (0.07)	-0.01 (0.06)	-0.01 (0.06)	-0.01 (0.07)	-0.01 (0.08)	-0.01 (0.18)
$\widehat{LB}_1(p)$	0.04 (0.22)	0.06 (0.15)	-0.07 (0.08)	-0.10 (0.09)	-0.12 (0.07)	-0.15 (0.08)	-0.16 (0.09)	-0.14 (0.17)	-0.07 (0.41)
$\widehat{UB}_1(p)$	0.14 (0.18)	0.04 (0.08)	-0.02 (0.09)	0.04 (0.16)	0.09 (0.10)	0.13 (0.10)	0.14 (0.11)	0.13 (0.19)	0.15 (0.52)
$\hat{\underline{\Delta}}_2(p)$	0.07 (0.38)	0.08 (0.2)	-0.07 (0.11)	-0.09 (0.11)	-0.11 (0.10)	-0.14 (0.10)	-0.15 (0.11)	-0.14 (0.19)	-0.06 (0.45)
$\hat{\overline{\Delta}}_2(p)$	0.17 (0.36)	0.05 (0.16)	-0.01 (0.12)	0.04 (0.18)	0.10 (0.12)	0.14 (0.12)	0.15 (0.13)	0.14 (0.21)	0.16 (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 (0.09)	0.04 (0.05)	0.01 (0.03)	-0.01 (0.03)	-0.01 (0.03)	-0.01 (0.02)	-0.01 (0.02)	0.01 (0.02)	0.02 (0.03)
$\hat{\Xi}_{OO,0}(p)$	0.01 (0.29)	0.00 (0.13)	0.00 (0.07)	-0.01 (0.07)	-0.01 (0.07)	-0.01 (0.08)	-0.01 (0.09)	-0.01 (0.11)	0.00 (0.26)
$\widehat{LB}_1(p)$	0.07 (0.25)	0.02 (0.13)	-0.06 (0.08)	-0.09 (0.08)	-0.11 (0.09)	-0.14 (0.09)	-0.15 (0.11)	-0.12 (0.25)	0.48 (0.76)
$\widehat{UB}_1(p)$	0.08 (0.22)	-0.02 (0.17)	0.11 (0.10)	0.15 (0.09)	0.18 (0.09)	0.22 (0.10)	0.26 (0.14)	0.35 (0.28)	0.67 (1.01)
$\hat{\underline{\Delta}}_2(p)$	0.06 (0.39)	0.02 (0.18)	-0.06 (0.11)	-0.09 (0.10)	-0.10 (0.11)	-0.13 (0.11)	-0.14 (0.13)	-0.11 (0.26)	0.48 (0.81)
$\hat{\overline{\Delta}}_2(p)$	0.07 (0.36)	-0.03 (0.2)	0.12 (0.12)	0.16 (0.11)	0.19 (0.12)	0.23 (0.12)	0.28 (0.16)	0.36 (0.3)	0.67 (1.04)

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

Table D.3: 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)
$\hat{\widehat{\Delta}}_2(p)$	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)
$\hat{\widehat{\Delta}}_2(p)$	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 $\Xi_{OO,0}(p) := \mathbb{E}[Y_0^* | S_0 = 1, S_1 = 1, V = p]$. The results are based on 1,000 Monte Carlo repetitions. The sample size in each generate data set is $n = 10,000$.

G Empirical Illustration: Additional Details and Results

In this appendix, we present the parametric specification and estimation details for the results in Section 5.

As mentioned in the main text, to properly account for covariates, we calculate bounds for $MTE^{OO}(p, x)$ based on parametric estimates for the functions $\mathbb{P}[D = 1 | X = x, Z = z]$, $\mathbb{P}[S = 1, D = d | X = x, P(Z) = p]$, and $\mathbb{P}[y_{k-1} \leq Y < y_k, S = 1, D = d | X = x, P(Z) = p]$ for $d = \{0, 1\}$, $p \in [0, 1]$, $k \in \{1, \ldots, K\}$ and covariate value x observed in the sample.

In particular, the probabilities above are modeled as logit functions that depend on a linear index of the covariates, instruments (for the propensity score), and a quadratic function of the propensity score (for the last two equations), while using $K = 20$ grid points (y_k) for the outcome variable in the last case. Let $\Lambda(\cdot)$ and $\lambda(\cdot)$ be the logistic distribution's CDF and PDF respectively, then:

$$\mathbb{P}[D = 1 | X = x, Z = z] = \Lambda(x\psi_X^d + z\psi_Z)$$

$$\mathbb{P}[S = 1, D = d | X = x, P(Z) = p] = \Lambda(\delta_1^d p + \delta_2^d p^2 + x\delta_X^d),$$

$$\mathbb{P}[y_{k-1} \leq Y < y_k, S = 1, D = d | X = x, P(Z) = p] = \Lambda(\theta_1^d p + \theta_2^d p^2 + x\theta_X^d).$$

The coefficients in the index function can then be estimated using a standard logit estimator, where the estimated $\hat{p}_i = \Lambda(x_i\hat{\psi}_X^d + z_i\hat{\psi}_Z)$ replace p_i in the estimation for the last two terms. To enforce the common support assumption, we trim the bottom 1% and top 1% of the overlapping estimated propensity score distribution.

Hence, the estimates for the partial derivatives of these functions, $\pi_d(p, x)$ and $\gamma_d(p, x, k)$, can be obtained by plugging in the relevant estimates:

$$\hat{\pi}_d(p, x) = \lambda(\hat{\delta}_1^d p + \hat{\delta}_2^d p^2 + x\hat{\delta}_X^d)(\hat{\delta}_1^d + 2\hat{\delta}_2^d p)(-1)^{1-d},$$

$$\hat{\gamma}_d(p, x, k) = \lambda(\hat{\theta}_1^d p + \hat{\theta}_2^d p^2 + x\hat{\theta}_X^d)(\hat{\theta}_1^d + 2\hat{\theta}_2^d p)(-1)^{1-d}.$$

The values for $\alpha(p, x)$ and $\beta(p, x)$ depend on the assumptions being considered. For

the case in which the sample selection mechanism is unrestricted we can obtain $\alpha(p, v^{\ell}, x)$ and $\beta(p, v^{\ell}, x)$ by plugging in $\hat{\pi}_d(p, x)$ in the formulas in Lemma 1. In the leading case of "monotonicity of sample selection in the treatment" (Assumptions 1-6), $\hat{\alpha}(p, x) = \dfrac{\hat{\pi}_0(p, x)}{\hat{\pi}_1(p, x)}$.

After estimating $\alpha(p, x)$, $\beta(p, x)$, the bounds around $MTE^{OO}(p, x)$ for each covariate value x, can be obtained. Below we focus on the "monotonicity of sample selection in the treatment" case, but the other situations can be handled analogously. For simplicity, let $\hat{f}_d(p, x, k) = \frac{\hat{\gamma}_d(p, x, k)}{\hat{\pi}_d(p, x)}$. Then, estimates of the bounds in Proposition 2 are given by,

$$\widehat{LB}_2(p, x) := \sum_{k=2}^{K_n} \overline{y}_k \cdot \mathbb{1}\left\{\hat{F}_1(p, x, k) \leq \hat{\alpha}(p, x)\right\} \cdot \frac{\hat{f}_1(p, x, k)}{\hat{\alpha}(p, x)} \tag{G.1}$$

$$\widehat{UB}_2(p, x) := \sum_{k=2}^{K_n} \overline{y}_k \cdot \mathbb{1}\left\{1 - \hat{F}_1(p, x, k) < \hat{\alpha}(p, x)\right\} \cdot \frac{\hat{f}_1(p, x, k)}{\hat{\alpha}(p, x)}, \tag{G.2}$$

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

Finally, the $MTE^{OO}(p, x)$ estimated bounds are given by $\hat{\underline{\Delta}}_2(p) := \widehat{LB}_2(p) - \hat{\Xi}_{OO,0}(p)$ and $\hat{\overline{\Delta}}_2(p) := \widehat{UB}_2(p) - \hat{\Xi}_{OO,0}(p)$.

To summarize the treatment effects, we average the bounds for $MTE^{OO}(p, x)$ across the sample using observed covariates values. By averaging with respect to the observed density of the covariates, we compute bounds around the summary measure of the conditional MTE for the always-observed subgroup:

$$SCMTE^{OO}(p) := \int \mathbb{E}\left[Y_1^* - Y_0^* | V = p, S_0 = 1, S_1 = 1, X = x'\right] \, dF_X(x').$$

If $X \perp\!\!\!\perp (V = p, S_0 = 1, S_1 = 1)$ holds (i.e., the covariates are exogenous), then

$$SCMTE^{OO}(p) = \int \mathbb{E}\left[Y_1^* - Y_0^* | V = p, S_0 = 1, S_1 = 1, X = x'\right] \, dF_{X|V=p, S_0=1, S_1=1}(x')$$

$$= \mathbb{E}\left[Y_1^* - Y_0^* | V = p, S_0 = 1, S_1 = 1\right]$$

$$= MTE^{OO}(p),$$

implying that the summary bounds are valid for the unconditional MTE function for the always-observed subgroup. Importantly, Deb, Munkin, and Trivedi (2006) assumed that the covariates are fully exogenous, implying that $X \perp\!\!\!\perp (V = p, S_0 = 1, S_1 = 1)$ holds. Carneiro and Lee (2009) used a similar exogeneity assumption.

H   Extension: Bounds for the distributional marginal treatment effect (DMTE)

In this appendix, we extend Proposition 2 to uniformly and sharply bound the distributional marginal treatment effect for the always-observed sub-population. As discussed by Carneiro and Lee (2009), policy-makers may care about distributional effects of a policy instead of the average effects, as individuals could respond differently to it depending on their position in the outcome distribution. For example, people at the bottom of the income distribution may not have the same response to a policy that increases college accessibility as people at the top of the income distribution. The distributional marginal treatment effect for the always-observed for any set A and any $p \in [0, 1]$ is defined as

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

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

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

Proposition 7. *For any set A, bounds on the $DMTE^{OO}(A; p)$ under Assumptions 1-6 are given by*

$$\underline{\Delta}_2^A(p) \leq DMTE^{OO}(A; p) \leq \overline{\Delta}_2^A(p)$$

for any $p \in \mathcal{P}$, where $\underline{\Delta}_2^A := \mathcal{P} \to \mathbb{R}$ and $\overline{\Delta}_2^A := \mathcal{P} \to \mathbb{R}$ are given by

$$\underline{\Delta}_2^A(p) := \max\left\{0, \frac{\mathbb{P}\left[Y_1^* \in A | S_1 = 1, V = p\right] - (1 - \alpha(p))}{\alpha(p)}\right\} - \mathbb{P}\left[Y_0^* \in A | S_0 = 1, S_1 = 1, V = p\right],$$

$$\overline{\Delta}_2^A(p) := \min\left\{1, \frac{\mathbb{P}\left[Y_1^* \in A | S_1 = 1, V = p\right]}{\alpha(p)}\right\} - \mathbb{P}\left[Y_0^* \in A | S_0 = 1, S_1 = 1, V = p\right]$$

for any $p \in \mathcal{P}$. Moreover, these bounds are sharp.