MTE with Misspecification*

Julián Martínez-Iriarte †
Department of Economics
UC Santa Cruz

Pietro Emilio Spini ‡
Department of Economics
UC San Diego

Click here for the latest version

This draft: November 2021
First draft: June 2021

Abstract

This paper studies the implication of a fraction of the population not responding to the instrument when selecting into treatment. We show that, in general, the presence of non-responders biases the Marginal Treatment Effect (MTE) curve and many of its functionals. Yet, we show that, when the propensity score is fully supported on the unit interval, it is still possible to restore identification of the MTE curve and its functionals with an appropriate re-weighting.

Keywords: Marginal Treatment Effects, Misspecification, Weak Instruments.

*We thank Vitor Possebom, Yixiao Sun, Kaspar Wuthrich, and seminar participants at UC San Diego for helpful comments. All remaining errors are ours.

†Email: jmart425@ucsc.edu
‡Email: pspini@ucsd.edu

1 Introduction

Marginal treatment effects (MTEs) have unified the identification theory of several policy parameters. While the MTE framework is essentially non-parametric,[1] it is required that the recipient's participation into treatment follows a (generalized) Roy model. This is often referred to as additive separability: an "additive" comparison of costs and benefits determines selection. On the other hand, identification of the MTE is achieved via the local instrumental variable (LIV) approach (Heckman and Vytlacil (2001, 2005)). An excellent survey is provided by Mogstad and Torgovitsky (2018). An early effort to analyze MTE under misspecification can be found in the appendix of the seminal paper by Heckman and Vytlacil (2001). They consider a case where the additive separability in the selection equation does not hold. The most serious consequence is that the LIV approach does not identify the MTE curve.

In this paper we analyze a different type of misspecification. We model a situation in which, under additive separability, a proportion of the population does not take into account the instrumental variable when deciding whether to take up treatment or not. We refer to them as non-responders. To analyze the resulting bias, we define a pseudo-MTE curve which results from the LIV approach. Under no misspecification, the pseudo-MTE curve would coincide with the MTE curve. The resulting bias can be interpreted as a location-scale change of the MTE curve, parameterized by the proportion of non-responders and their propensity score.

We have two main results. The first one shows that the ability to recover the conditional average treatment effect (CATE) for the subpopulation of responders depends on the proportion of non-responders only through the support of the responders' propensity score. Indeed, when the support of the propensity score is the unit interval, it is possible to identify the CATE *without* having to recover the true MTE curve in the first place. In a nutshell, ignoring misspecification and integrating under the pseudo-MTE curve over the support of observed propensity score yields the correct CATE for the subpopulation of responders.

While the previous identification result for the CATE is independent of the proportion of non-responders, this is not true of the MTE curve and other parameters derived from it such as LATE and MPRTE. However, in our second result, we show how to recover the MTE curve for responders by undoing the location-scale change induced by the presence of non-responders. The correction is based on an estimate of the support of the propensity score and requires only observable data. It gives an estimator of the policy parameter of interest that is simple to implement. Cases where the propensity score is fully supported are relevant in practice. For a recent example, see the survey approach of Briggs, Caplin, Leth-Petersen, Tonetti, and Violante (2020) the probability of having a child is supported on the full unit interval.

Recently, Acerenza, Ban, and Kedagni (2021) and Possebom (2021) focus on the effect of measurement error in treatment status on the MTE curve. We complement such results by noting that a simple change to our setup can cover the case of misclassification. In a setting where treatment

[1] Linearity is sometimes assumed to facilitate estimation. See, e.g., Appendix B in Heckman, Urzua, and Vytlacil (2006)

status is misclassified, the observed outcome is generated with the true treatment status. In our setting of misclassification, the observed outcome can be regarded as a mixture of responders and non-responders. The proportion of non-responders is analogous to the proportion of misreporters. Indeed, our results also hold if instead of having a fraction of non-responders, we have a fraction of misreporters.

Another consequence of the presence of non-responders in the sample is that the effect of the instrumental variable on the propensity score is attenuated. Motivated by this, we model a situation where the proportion of non-responders approaches 1, analogous to the setting of weak instruments of Staiger and Stock (1997). Thus, we can derive weak-instrument-like asymptotic distributions for the parameters derived from the MTE curve.

The rest of the paper is organized as follows: section 2 introduces the model; section 3 contains the main identification results; section 4 provides bounds for the case where the propensity score is not fully supported in the unit interval; section 5 traces the connection to the weak IV literature; and section 6 concludes. While this paper only deals with identification, we expect to extend our results to cover estimation and inference.

2 Misspecification and MTE

In this section we introduce our model for misspecification in the MTE framework (Bjorklund and Moffitt (1987), Heckman and Vytlacil (2001, 2005)). We analyze the consequences of misspecification from the identification point of view.

2.1 The Model

We start with a general non-separable potential outcome model

$$\begin{aligned} Y(0) &= h_0(X, U_0), \\ Y(1) &= h_1(X, U_1), \\ Y &= D^* Y(1) + (1 - D^*) Y(0), \end{aligned}$$

where D^* is the observed treatment status, X are observable covariates with support denoted by $\mathcal{X}$, and $\{Y(0), Y(1)\}, Y$ are potential and observed outcomes, respectively. The functions h_0 and h_1 are unknown.

We model misspecification as a situation where there are two types of individuals: responders and non-responders. Responders select into treatment taking into account the incentives in Z. Their selection equation is given by $D = \mathbb{1}\{\mu(X, Z) \geq V\}$. On the other hand, non-responders do not react to incentives in Z at all. Their selection equation is given by $\tilde{D} = \mathbb{1}\{\tilde{\mu}(X) \geq \tilde{V}\}$. Notice how Z is not featured in $\tilde{\mu}(\cdot)$. For the non-responders, Z fails the relevance condition of the standard MTE model.

Let S be the latent status of an individual: $S = 1$ for a responder and $S = 0$ for a non-responder. The observed treatment status D^* is given by:

$$D^* = S \cdot D + (1 - S) \cdot \tilde{D}. \tag{1}$$

We allow for the proportion of non-responders may vary with X. To this end, we define $\delta_X = \Pr(S = 0|X) = \Pr(D^* = \tilde{D}|X)$. Thus, for every subpopulation with characteristics $X = x$ there is a proportion $\delta_x = \Pr(S = 0|X = x) \in [0, 1)$ of non-responders. We consider values where $\sup_{x \in \mathcal{X}} \delta_x < 1$ to avoid a situation where no-one responds to the instrumental variable.

Remark 1. *We observe Y according to $Y = D^*Y(1) + (1 - D^*)Y(0)$, which is given by the actual choice D^*. If, instead, we have $Y = DY(1) + (1 - D)Y(0)$, then we can interpret D^* as a misclassified treatment status. In this case, all individuals decide according to $D = \mathbb{1}\{\mu(X, Z) \geq V\}$, but a fraction of them reports according to $\tilde{D} = \mathbb{1}\{\tilde{\mu}(X) \geq \tilde{V}\}$ See Acerenza, Ban, and Kedagni (2021) and Possebom (2021) for recent studies on MTE under misclassification.*

The econometrician observes a cross section of (Y_i, D_i^*, X_i, Z_i). When $\delta_X = 0$ almost surely, then $D^* = D$ and we are in the familiar MTE framework of Heckman and Vytlacil (2001, 2005). Otherwise, if $\delta_X \neq 0$ almost surely, for an observation of D_i^*, we do not know whether we are observing the treatment status of a non-responder or of a responder. That is, it is unknown if we are observing D_i or $\tilde{D}_i$.

Assumption 1. ***Type Independence.*** $S \perp Z \| X$.

Assumption 1 states that once we control for X, the latent status of a individuals does not vary with the instrumental variable Z.

Assumption 2. ***Relevance and Exogeneity***

1. *$\mu(X, Z)$ is a nondegenerate random variable conditional on X.*
2. *$(U_0, U_1, V, \tilde{V})$ are independent of Z conditional on X.*

Note that, for the subpopulation of non-responders, the instrument is valid but totally irrelevant. The larger the value of δ_x, the "weaker" the instrument Z, since most participants with $X = x$ are non-responders. With the exception of the requirement that $\tilde{V} \perp Z \| X$, these are the same conditions of Heckman and Vytlacil (2001, 2005). Our additional requirement covers the subpopulation of non-responders: neither the "cost" of treatment $\tilde{V}$ nor the "benefit" $\tilde{\mu}(X)$ depend on Z when conditioned on X.

Example 1. *To fix ideas, we can think of a two part cost of providing the incentive. A fixed cost associated to targeting a particular subpopulation with covariates $X = x$ and the cost of the incentive itself. If Z is a voucher, there could be administrative costs associated to making it available to subpopulation $X = x$. For non-responders who do not redeem the voucher, the cost of the incentive is zero. Such a scenario would satisfy Assumption 2.*

The misclassification structure of Equation (1) allows to define three different propensity scores. An observed/identified one which is based on the observables (D^*, X, Z), and two latent/unobserved propensity scores: one for the reponders and one for the non-responders. Formally, they are given by

$$\begin{aligned} P^*(X,Z) &:= \Pr(D^* = 1|X,Z) && \textbf{(Observed)} \\ P(X,Z) &:= \Pr(D = 1|S = 1,X,Z) && \textbf{(Responders)} \\ \tilde{P}(X) &:= \Pr(\tilde{D} = 1|S = 0,X) && \textbf{(Non-responders)} \end{aligned}$$

The next result takes (mainly) advantage of Assumption 1 to derive a useful linear relation between them.

Lemma 1. *Under Assumptions 1 and 2.2 we can relate the different propensity scores by*

$$P^*(X,Z) = (1-\delta_X)\cdot P(X,Z) + \delta_X \cdot \tilde{P}(X). \tag{2}$$

Proof. Starting with the model in (1) we can write

$$\begin{aligned} \Pr(D^* = 1|X,Z) = \Pr(S = 1|X,Z)\cdot \Pr(D = 1|S = 1,X,Z) \\ + \Pr(S = 0|X,Z)\cdot \Pr(\tilde{D} = 1|S = 0,X,Z). \end{aligned}$$

Assumption 1 simplifies the mixing probabilities to $\Pr(S = 1|X) = 1-\delta_X$ and $\Pr(S = 0|X) = \delta_X$. We obtain

$$\Pr(D^* = 1|X,Z) = (1-\delta_X)\cdot \Pr(D = 1|S = 1,X,Z) + \delta_X \cdot \Pr(\tilde{D} = 1|S = 0,X,Z).$$

To see that $Pr(\tilde{D} = 1|S = 0,X,Z) = Pr(\tilde{D} = 1|S = 0,X)$, we note that By Assumptions 1 and 2.2:

$$\Pr(\tilde{D} = 1|S = 0,X,Z) = \Pr(\tilde{\mu}(X) \geq \tilde{V}|S = 0,X,Z) = \Pr(\tilde{\mu}(X) \geq \tilde{V}|X) = \Pr(\tilde{D} = 1|S = 0,X).$$

Therefore

$$\begin{aligned} \Pr(D^* = 1|X,Z) &= (1-\delta_X)\cdot \Pr(D = 1|S = 1,X,Z) + \delta_X \cdot \Pr(\tilde{D} = 1|S = 0,X) \\ &= (1-\delta_X)\cdot P(X,Z) + \delta_X \cdot \tilde{P}(X). \end{aligned}$$

□

For a fixed $X = x$, the result in Lemma 1 shows that the observed propensity (still random through Z) is a linear transformation of the propensity score for the responders. If, additionally, we take two different values of Z, for example z and z', we can remove the contribution of $\tilde{P}(X)$,

which is invariant with respect to z and obtain[2]

$$P^*(x,z) - P^*(x,z') = (1-\delta_x) \cdot \left[P(x,z) - P(x,z')\right] \tag{3}$$

Equation (3) says that the changes on the observed propensity score induced by varying Z are proportional to the changes on the true propensity score induced by varying Z. Thus, if we knew δ_x, we could recover the change in the propensity score for the responders. When Z is continuous, we can take a limiting version of this argument, *e.g.*, as $z' \to z$, to obtain

$$\frac{\partial P^*(x,z)}{\partial z} = (1-\delta_x) \cdot \frac{\partial P(x,z)}{\partial z}. \tag{4}$$

Both the discrete (equation (3)), and the continuous (equation(4)) change in the propensity score play a role in the relationship between the MTE curve (defined below) and certain parameters of interest.

2.2 The MTE for Responders

For the subpopulation of responders, the standard MTE framework holds. This motivates us to define an MTE curve for this subpopulation. In doing so, we are implicitly assuming that this is our object of interest. The reason for this is that many times we can also control the instrumental variable Z. Thus, to asses the effects of manipulations of Z we look at the MTE curve for responders.

Let $\mathcal{P}_x$ and $\mathcal{P}_x^*$ denote the support of $P(x,Z) := \Pr(D=1|X=x,Z)$ and $P^*(x,Z) := \Pr(D^*=1|X=x,Z)$ respectively. For the subpopulation of responders, we rewrite the selection equation as $D = \mathbb{1}\{P(X,Z) \geq U_D\}$ where $U_D \sim U_{(0,1)}$.[3] Thus, we define the MTE curve for responders as

$$\text{MTE}(u,x) := \mathbb{E}\left[Y(1) - Y(0) | S=1, U_D = u, X = x\right].$$

By the LIV approach we have the following equivalence result:[4]

$$\text{MTE}(u,x) = \frac{\partial \mathbb{E}\left[Y|S=1, P(X,Z)=u, X=x\right]}{\partial u} \text{ for } u \in \mathcal{P}_x. \tag{5}$$

Since we do not observe $P(X,Z)$, this is *not* an identification result in our setting. In a similar fashion, we *define* the following pseudo-MTE curve:

$$\text{MTE}^*(u,x;\delta_x) := \frac{\partial \mathbb{E}\left[Y|P^*(X,Z)=u, X=x\right]}{\partial u} \text{ for } u \in \mathcal{P}_x^*. \tag{6}$$

We emphasize that the pseudo-MTE curve is indexed by δ_x because it depends implicitly on

[2] We write $P^*(x,z)$ for $\Pr(D^*=1|X=x,Z=z)$, and $P(x,z)$ for $\Pr(D=1|S=1,X=x,Z=z)$.

[3] This follows from $D = \mathbb{1}\{F_{V|S,X,Z}(\mu(X,Z)|1,X,Z) \geq F_{V|S,X,Z}(V|1,X,Z)\}$. Noting that by assumptions 2.(2) and 1, we have $D = \mathbb{1}\{P(X,Z) \geq F_{V|S,X}(V|1,X)\}$. Finally, we take $U_D := F_{V|S,X}(V|1,X)$.

[4] See Heckman and Vytlacil (2001) for sufficient conditions.

the proportion of the nonresponders. From the data only, we can only compute $\text{MTE}^*(u,x;\delta_x)$, not $\text{MTE}(u,x)$. The pseudo-MTE curve is the curve that would be mistakenly taken to be the MTE curve. Indeed, in the absence of non-responders, $\text{MTE}^*(u,x;0) = \text{MTE}(u,x)$. If non-responders are present in the $X = x$ subpopulation, that is if $\delta_x > 0$, the observed $\text{MTE}^*(u,x;\delta_x)$ does not identify $\text{MTE}(u,x)$. In another words, the LIV approach is biased. We can now fully characterize the bias induced by δ_x on the MTE curve.

Lemma 2. *Under Assumptions 1 and 2, we can write*

$$MTE(v,x) = (1-\delta_x)MTE^*\left((1-\delta_x)v + \delta_x\tilde{P}(x), x; \delta_x\right) \text{ for } v \in \mathcal{P}_x. \tag{7}$$

Proof. Using (2), for $u \in \mathcal{P}_x^*$, we can write

$$\begin{aligned}\mathbb{E}\left[Y|P^*(X,Z) = u, X = x\right] &= \mathbb{E}\left[Y|(1-\delta_x)\cdot P(X,Z) + \delta_x \cdot \tilde{P}(X) = u, X = x\right] \\ &= \mathbb{E}\left[Y\middle|P(X,Z) = \frac{u-\delta_x\tilde{P}(x)}{1-\delta_x}, X = x\right]\end{aligned}$$

Differentiating with respect to u, we obtain

$$\text{MTE}^*(u,x;\delta_x) = \frac{1}{1-\delta_x}\text{MTE}\left(\frac{u-\delta_x\tilde{P}(x)}{1-\delta_x}, x\right) \text{ for } u \in \mathcal{P}_x^*. \tag{8}$$

since $\frac{u-\delta_x\tilde{P}(x)}{1-\delta} \in \mathcal{P}_x$ by (2). Alternatively, we can write

$$\text{MTE}(v,x) = (1-\delta_x)\text{MTE}^*\left((1-\delta_x)v + \delta_x\tilde{P}(x), x; \delta_x\right) \text{ for } v \in \mathcal{P}_x.$$

□

Lemma 2 shows that the bias is in the form of both location and scale. Equation (8), which is equivalent to Equation (7),[5] shows that MTE^* is obtained by changing the location from u to $u - \delta_x\tilde{P}(x)$, and rescaling by $(1-\delta_x)^{-1}$. Thus, as in a location-scale family of densities, we can regard MTE^* as a family of curves, defined over $\mathcal{P}_x^*$, which is indexed by δ_x and $\tilde{P}(x)$.

3 Automatic and explicit de-biasing

We now introduce our two main results. We show that, for any subpopulation $X = x$ where the instrument is strong enough to induce a propensity score supported on the full unit interval $[0,1]$, the associated $CATE(x)$ can be identified for responders. This is true even if the $MTE^*(u,x,\delta_x)$ curve is biased for $MTE(u,x)$. We note that the identified $CATE(x)$ parameters corresponds to the subpopulation of responders.

Assumption 3. ***Full Support.*** *The support of $P(x,Z)$ is $\mathcal{P}_x = [0,1]$ for every x in a subset $\mathcal{X}_B \subseteq \mathcal{X}$.*

[5] Note the changes in the domain of integration between (7) and (8).

Assumption 3 says that the incentive in the instrument Z is strong enough to induce any individual in the $X = x$ subpopulation into or out of treatment. Perhaps surprisingly, the CATE(x), can be recovered only by resorting to the full support assumption. That is, to correctly compute the CATE(x) we do not need to recover the true MTE curve for responders.

Theorem 1. *Let Assumptions 1, 2, and 3 hold. Then, for any $x \in \mathcal{X}_B$:*

$$CATE(x) = \int_{\inf \mathcal{P}_x^*}^{\sup \mathcal{P}_x^*} MTE^*(u, x; \delta_x) du.$$

Proof. The Conditional Average Treatment Effect, CATE(x), could be computed using the true MTE curve (if it was observed) as

$$\text{CATE}(x) = \int_0^1 \text{MTE}(u, x) du.$$

Given that $\mathcal{P}_x = [0, 1]$, then $\mathcal{P}_x^* := [\underline{p_x^*}, \overline{p_x^*}]$ where $\underline{p_x^*} := \inf \mathcal{P}_x^* = \delta_x \tilde{P}(x)$ and $\overline{p_x^*} := \sup \mathcal{P}_x^*(1 - \delta_x) + \delta_x \tilde{P}(x)]$. Consider the integrating the pseudo-MTE curve over the support of the observed propensity score:

$$\int_{\delta_x \tilde{P}(x)}^{(1-\delta_x)+\delta_x \tilde{P}(x)} \text{MTE}^*(u, x; \delta_x) du.$$

Using (8), we have

$$\begin{aligned}
\int_{\delta_x \tilde{P}(x)}^{(1-\delta_x)+\delta_x \tilde{P}(x)} \text{MTE}^*(u, x; \delta_x) du &= \int_{\delta_x \tilde{P}(x)}^{(1-\delta_x)+\delta_x \tilde{P}(x)} \frac{1}{1-\delta_x} \text{MTE}\left(\frac{u - \delta_x \tilde{P}(x)}{1-\delta_x}, x\right) du \\
&= \int_0^1 \text{MTE}(u, x) du \\
&= \text{CATE}(x)
\end{aligned}$$

where we have done the change of variables

$$v = \frac{u - \delta_x \tilde{P}(x)}{1 - \delta_x}.$$

□

Remark 2. *The result of Theorem 1 states that by integrating the observed (and biased) marginal treatment effect curve over the support of the observed (and biased) propensity score leads to the CATE(x) provided that the propensity score for responders has full support. Thus, under the type of misspecification described in (1), CATE(x) is robust to $\delta_x \neq 0$.*

Remark 3. *This result also hold in a setting of misclassification and was our original motivation. That is, in a setting where instead of $Y = D^* Y(1) + (1 - D^*) Y(0)$, we have $Y = DY(1) + (1 - D) Y(0)$ and we*

interpret D^* as a misclassified treatment status.

Unfortunately, the automatic "de-biasing" in Theorem 1 does not hold for the other policy parameters that can be obtained via the MTE curve. On the other hand, we show that the full support assumption can be used to identify δ_x which allows an explicit "de-biasing" procedure. Given that $\mathcal{P}_x^* := [\underline{p_x^*}, \overline{p_x^*}] = [\delta_x \tilde{P}(x), (1-\delta_x) + \delta_x \tilde{P}(x)]$ we can actually identify both δ_x and $\tilde{P}(x)$. It follows then from Lemma 2 that we can recover the MTE(u, x) curve.

Proposition 1. *Let Assumptions 1, 2, and 3 hold. Then δ_x is identified for any $x \in \mathcal{X}_B$ through:*

$$\delta_x = 1 - (\overline{p_x^*} - \underline{p_x^*})$$

Proof. According to Equation (2), the range of the observed propensity score is given by $\mathcal{P}_x^* = [\delta_x \tilde{P}(x), (1-\delta_x) + \delta_x \tilde{P}(x)]$. For each x, the observed propensity score $P^*(\cdot)$ can be viewed as an affine function of $P(\cdot)$. This affine function is parameterized by δ_x and $\tilde{P}_x$. For the endpoints $\underline{p}_x$ and $\overline{p}_x$ of the true propensity score, we have the mappings:

$$\underline{p_x} \mapsto (1-\delta_x)\underline{p_x} + \delta_x \tilde{P}(x)$$
$$\overline{p_x} \mapsto (1-\delta_x)\overline{p_x} + \delta_x \tilde{P}(x)$$

The images of this collection of mapping are observed. They are the endpoints of the observed propensity score $P^*(Z, x)$. If the original endpoints of the true $P(\cdot)$ are known to be $\underline{p_x} = 0$ and $\overline{p_x} = 1$, like stated in Assumption 3, the mapping above can be recovered by the following system of two equations in two unknowns: $\tilde{P}(x)$ and δ_x.

$$\underline{p_x^*} = \delta_x \tilde{P}(x)$$
$$\overline{p_x^*} = (1-\delta_x) + \delta_x \tilde{P}(x)$$

which implies that

$$\delta_x = 1 - (\overline{p_x^*} - \underline{p_x^*})$$
$$\tilde{P}(x) = \underline{p_x^*} \cdot \frac{1}{\delta_x}$$

□

The intuition for this result is simple. Because the original propensity score $P(Z, x)$, for any fixed x, is supported on the unit interval, the observed support $\mathcal{P}_x^* = [\underline{p_x^*}, \overline{p_x^*}]$ will contain enough information to identify δ_x. This is summarized Figure 1.

Having identified δ_x, then we use Equation (8) to identify the MTE curve.

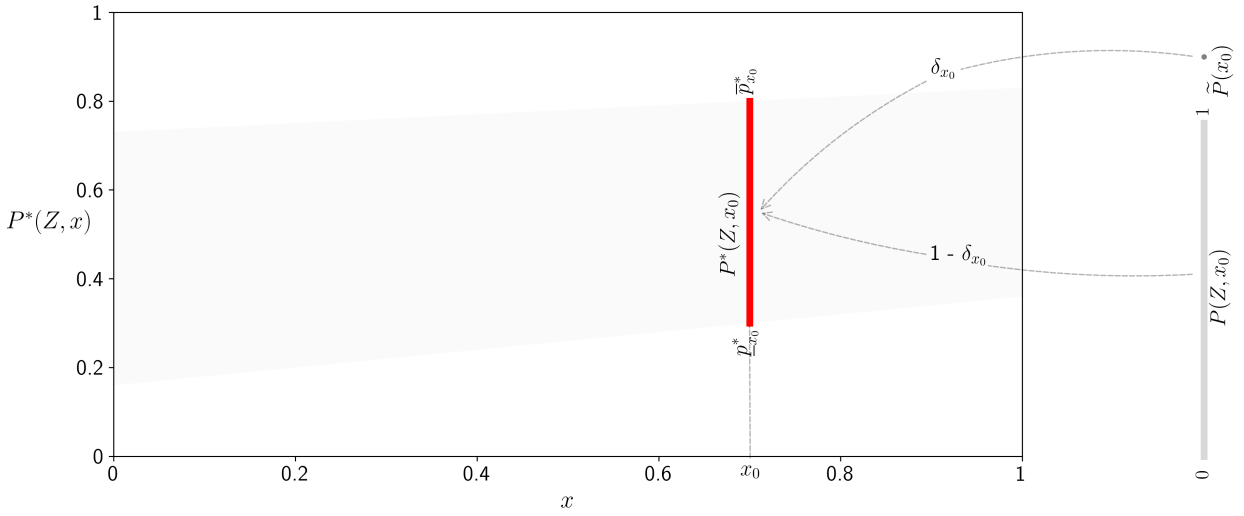

Figure 1: Identifying δ_x: The figure shows the link between the non-responders propensity score, the proportion of non-responders and the observed propensity score. Because the non-responders propensity score does not vary with the instrument Z and $supp(P(Z,x)) = [0,1]$ the δ_x can be recovered from observing the discrepancy from the observed support $P^*(Z,x)$ and $[0,1]$. The picture shows one of those points, x_0.

Corollary 1. *Let Assumptions 1, 2, and 3 hold. Then, the MTE curve is identified:*

$$MTE(v,x) = (\overline{p^*_x} - \underline{p^*_x})MTE^*\left((\overline{p^*_x} - \underline{p^*_x})v + \underline{p^*_x}, x; 1 - (\overline{p^*_x} - \underline{p^*_x})\right) \text{ for } v \in \mathcal{P}_x = [0,1].$$

where $\underline{p^*_x} = \inf \mathcal{P}^*_x$ *and* $\overline{p^*_x} = \sup \mathcal{P}^*_x$.

This corollary provides the correct "de-biasing" to be performed on the observed MTE curve to match the true MTE curve. However, it is possible to recover parameters that are based on the MTE curve *without* having to recover the MTE curve in the first place. We provide two examples.

Example 2 (LATE)**.** *Consider the LATE, for* $P(x,z') < P(x,z)$ *with* $z, z' \in \mathcal{Z}$*, which can be obtained from MTE curve as*

$$LATE(x, P(x,z), P(x,z')) = \frac{1}{P(x,z) - P(x,z')} \int_{P(x,z')}^{P(x,z)} MTE(u,x)du.$$

Under misspecification, for the same $z, z' \in \mathcal{Z}$*, we have*

$$\begin{aligned} LATE^*(x, P^*(x,z), P^*(x,z')) &= \frac{1}{P^*(x,z) - P^*(x,z')} \int_{P^*(x,z')}^{P^*(x,z)} MTE^*(u,x;\delta_x)du \\ &= \frac{(1-\delta_x)^{-1}}{P(x,z) - P(x,z')} \int_{(1-\delta_x)P(x,z') + \delta_x \tilde{P}(x)}^{(1-\delta_x)P(x,z) + \delta_x \tilde{P}(x)} \frac{1}{1-\delta_x} \\ &\times MTE\left(\frac{u - \delta_x \tilde{P}(x)}{1-\delta_x}, x\right) du. \end{aligned}$$

Note that to go from MTE* to MTE we used Lemma 2. We did not use Corollary 1. Defining the change of variables $\tilde{u} = \frac{u-\delta_x \tilde{P}(x)}{1-\delta_x}$, *we get* $(1-\delta_x)d\tilde{u} = du$. *We then write*

$$\begin{aligned} LATE^*(x, P^*(x,z), P^*(x,z')) &= \frac{(1-\delta_x)^{-1}}{P(x,z) - P(x,z')} \int_{(1-\delta_x)P(x,z')+\delta_x \tilde{P}(x)}^{(1-\delta_x)P(x,z)+\delta_x \tilde{P}(x)} \frac{1}{1-\delta_x} \\ &\quad \times MTE\left(\frac{u-\delta_x \tilde{P}(x)}{1-\delta_x}, x\right) du \\ &= \frac{(1-\delta_x)^{-1}}{P(x,z) - P(x,z')} \int_{P(x,z')}^{P(x,z)} MTE(u,x) du \\ &= \frac{1}{1-\delta_x} LATE(x, P(x,z), P(x,z')). \end{aligned}$$

Now, since $\delta_x = 1 - (\overline{p_x^*} - \underline{p_x^*})$ *by Proposition 1, the explicit de-biasing is achieved by*

$$(\overline{p_x^*} - \underline{p_x^*}) LATE^*(x, P^*(x,z), P^*(x,z')) = LATE(x, P(x,z), P(x,z')).$$

The left hand side can be computed from the data.

Example 3 (MPRTE). *The marginal policy relevant treatment effect (MPRTE) is an average of the* $MTE(u,x)$ *along the margin of indifference: when* $U_D = P(X,Z)$. *It is given by*

$$MPRTE(x) = \int_{\mathcal{Z}} MTE(P(x,z),x) \frac{\partial P(x,z)}{\partial z} \left(E\left[\frac{\partial [P(x,Z)]}{\partial z} \right] \right)^{-1} f_{Z|X}(z|x) dz$$

Then, using Equations (4) and (7) we get

$$\begin{aligned} MPRTE^*(x) &= \int_{\mathcal{Z}} MTE^*(P^*(x,z), x; \delta_x) \frac{\partial P^*(x,z)}{\partial z} \left(E\left[\frac{\partial [P^*(x,Z)]}{\partial z} \right] \right)^{-1} f_{Z|X}(z|x) dz \\ &= \int_{\mathcal{Z}} \frac{1}{1-\delta_x} MTE(P(x,z),x) \frac{\partial P(x,z)}{\partial z} \left(E\left[\frac{\partial [P(X,Z)]}{\partial z} \right] \right)^{-1} f_{Z|X}(z|x) dz \\ &= \frac{1}{1-\delta_x} MPRTE(x). \end{aligned}$$

Thus, again, by Proposition 1, we obtain

$$(\overline{p_x^*} - \underline{p_x^*}) MPRTE^*(x) = MPRTE(x).$$

In the previous examples, proceeding as if there were no misspecification, yields biased parameters. Thus, the automatic "de-biasing" in CATE is the exception rather than the rule.

4 Bounds under limited support

Instead of assuming full support, now we allow for limited support of the propensity score $P(x,Z)$, but we still require that it is an interval.

Assumption 4. ***Limited Support.*** *The support of* $P(x,Z)$ *is* $\mathcal{P}_x = [\underline{p_x}, \overline{p_x}] \subset [0,1]$.

Under Assumption 4, and using (2), we have that the observed support of $P^*(X,Z)$ is

$$[\underline{p_x^*}, \overline{p_x^*}] = [(1-\delta_x)\underline{p_x} + \delta_x \tilde{P}(x), (1-\delta_x)\overline{p_x} + \delta_x \tilde{P}(x)].$$

Taking the difference we obtain that $\overline{p_x^*} - \underline{p_x^*} = (1-\delta_x)(\overline{p_x} - \underline{p_x})$. Since $\overline{p_x} - \underline{p_x} \leq 1$, then $\overline{p_x^*} - \underline{p_x^*} \leq (1-\delta_x)$, so that a lower bound for δ_x is $\delta_x \geq 1 - (\overline{p_x^*} - \underline{p_x^*})$.

In general, it is not possible to provide an upper bound for δ_x. This is similar to the case of misclassification. Following that literature (see Assumption 4 in Acerenza, Ban, and Kedagni (2021), and references therein), we assume it is known that for some $\overline{\delta}_x$: $\delta_x \leq \overline{\delta}_x < 1$. Thus, we can write $1 - (\overline{p_x^*} - \underline{p_x^*}) \leq \delta_x \leq \overline{\delta}_x$. The correction factor in Examples 2 and 3 is $(1-\delta_x)$. Now, it bounded by $1 - \overline{\delta}_x \leq 1 - \delta_x \leq \overline{p_x^*} - \underline{p_x^*}$. Thus, we can bound both LATE and MPRTE using this:

$$\begin{aligned}(1-\overline{\delta})\text{LATE}^*(x, P^*(x,z), P^*(x,z')) &\leq \text{LATE}(x, P(x,z), P(x,z')) \\ &\leq (\overline{p_x^*} - \underline{p_x^*})\text{LATE}^*(x, P^*(x,z), P^*(x,z')),\end{aligned}$$

and

$$(1-\overline{\delta})\text{MPRTE}^*(x) \leq \text{MPRTE}(x) \leq (\overline{p_x^*} - \underline{p_x^*})\text{MPRTE}^*(x).$$

Naturally, if $\overline{\delta}_x$ is not known, we can only provide upper bounds.

Again, we stress that is not necessary to bound the MTE curve in the first place. Such a bound can be complicated to obtain since, by Lemma 2, δ_x enters in three different ways in the observed MTE curve.

5 Misspecification as a weak instrument

We can frame our model as the triangular scheme of Staiger and Stock (1997) and consider a sequence $\{\delta_{x,n}\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} \delta_{x,n} = 1$ at a certain rate as $n \to \infty$. Thus, as $n \to \infty$, the instrument becomes irrelevant in the model. A possible indicator of the presence of a large value of $\delta_{x,n}$ can be the average derivative of the observed propensity score. This equals an attenuated version of the average derivative of the true propensity score. For a given value of $\delta_{x,n}$, by equation (4), we have

$$E\left[\frac{\partial P^*(x,Z)}{\partial z}\right] = (1-\delta_{x,n})E\left[\frac{\partial P(x,Z)}{\partial z}\right]$$

Thus a "small" value can be an indication that $\delta_{x,n}$ is close to 1. This is similar to a first stage regression in the linear model. We take the derivative with respect to z to get rid of the propensity score that does not respond to Z. We average, because this likely to be a non-linear expression. Thus, $(1-\delta_{x,n})$ can be thought of as the counterpart of $C/\sqrt{T}$ in the notation of Staiger and

Stock (1997). Indeed, define

$$Cov_x(Z, D^*) := E[ZD^*|X = x] - E[Z|X = x]E[D^*|X = x].$$

We have

$$\begin{aligned} E[ZD^*|X = x] &= E[ZSD|X = x] + E[Z(1 - S)\tilde{D}|X = x] \\ &= E[ZSD|X = x] + E[Z|X = x]E[(1 - S)\tilde{D}|X = x] \end{aligned}$$

and

$$E[D^*|X = x] = E[SD|X = x] + E[(1 - S)\tilde{D}|X = x]$$

Thus,

$$\begin{aligned} Cov_x(Z, D^*) &= E[ZSD|X = x] - E[Z|X = x]E[SD|X = x] \\ &+ E[Z|X = x]E[(1 - S)\tilde{D}|X = x] - E[Z|X = x]E[(1 - S)\tilde{D}|X = x] \\ &= Cov_x(Z, SD) \end{aligned}$$

which is the covariance between the instrument and treatment status for the responders with $X = x$. To see the role of the rate at which $\delta_{x,n}$ converges to 1, suppose for a second that we know the functional form of $P^*(x, Z)$, and we estimate the average derivative using a sample mean:

$$\hat{E}\left[\frac{\partial P^*(x, Z)}{\partial z}\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial P^*(x, Z_i)}{\partial z} = (1 - \delta_{x,n})\frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(x, Z_i)}{\partial z}$$

Then

$$\hat{E}\left[\frac{\partial P^*(x, Z)}{\partial z}\right] - E\left[\frac{\partial P^*(x, Z)}{\partial z}\right] = (1 - \delta_{x,n})\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(x, Z_i)}{\partial z} - E\left[\frac{\partial P(x, Z)}{\partial z}\right]\right)$$

In order to investigate possible discontinuities in the limiting distributions, we follow Hahn and Kuersteiner (2002), and we let $(1 - \delta_{x,n}) = n^{\nu_x}$, for $\nu_x < 0$. We obtain

$$\hat{E}\left[\frac{\partial P^*(X, Z)}{\partial z}\right] - E\left[\frac{\partial P^*(X, Z)}{\partial z}\right] = O_p(n^{\nu_x - 1/2}).$$

Then, we obtain a degenerate limit:

$$\sqrt{n}\left(\hat{E}\left[\frac{\partial P^*(X, Z)}{\partial z}\right] - E\left[\frac{\partial P^*(X, Z)}{\partial z}\right]\right) = o_p(1)$$

Now consider the MPRTE. Recall that, by Example 3, under the full support guaranteed by

Assumption 3,

$$n^{\nu_x}\text{MPRTE}^*(x) = \text{MPRTE}(x).$$

Assume that, if $\delta_x = 0$, there exists $\widehat{\text{MPRTE}}(x)$, a $\sqrt{n}$-consistent estimator of $\text{MPRTE}(x)$ such that

$$\widehat{\text{MPRTE}}^*(x) - \text{MPRTE}^*(x) = n^{-\nu_x}\left(\widehat{\text{MPRTE}}(x) - \text{MPRTE}(x)\right).$$

Thus, if $\nu_x = -1/2$, then $\widehat{\text{MPRTE}}^*(x)$ does not converge in probability. In future work, we will use these results to construct confidence intervals for the parameters of interest.

6 Conclusion

In this paper we use the MTE framework to model a proportion of individuals who do not respond to the incentives of the instrumental variable. We show that in the special case where the observed propensity score is fully supported on the unit interval, i) the CATE is automatically identified regardless of the non-responders, and ii) we can identify the proportion of non-responders and use it to recover the MTE curve, and we can recover any parameter associated with it. We show that for some parameters, such as LATE and MPRTE, it is even possible to bypass the recovery of the MTE curve, and directly recover these parameters. Moreover, if the propensity has limited support, we find bounds for the LATE, the MPRTE, and the MTE curve. When we let the proportion of non-responders approach 1 at a certain rate, the framework resembles that of weak instruments. In future research we hope to leverage the results in this literature to construct valid confidence intervals for the MTE curve and related parameters.

References

Acerenza, S., K. Ban, and D. Kedagni (2021): "Marginal Treatment Effects with Misclassified Treatment," Working Paper.

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

Briggs, J., A. Caplin, S. Leth-Petersen, C. Tonetti, and G. Violante (2020): "Estimating Marginal Treatment Effects with Survey Instruments," Working Paper.

Hahn, J., and G. Kuersteiner (2002): "Discontinuities of weak instrument limiting distributions," *Economics Letters*, 75, 325–331.

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

Heckman, J. J., and E. Vytlacil (2001): "Local Instrumental Variables," in *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, ed. by C. Hsiao, K. Morimune, and J. Powell, pp. 1–46. Cambrigde University Press.

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

Mogstad, M., and A. Torgovitsky (2018): "Identification and Extrapolation of Causal Effects with Instrumental Variables," *Annual Review of Economics*, 10, 577–613.

Possebom, V. (2021): "Crime and Mismeasured Punishment: Marginal Treatment Effect with Misclassification," Working Paper.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557–586.