Simulation: Inference About Policy Relevant Treatment Effects [*]

Julian Budde[†]

August 28, 2023

Term Paper (Econometrics Topics Course Summer 2023)

Abstract

Mogstad, Santos, and Torgovitsky (2018) propose an identification framework for extrapolating from identified IV-like estimands to a broad set of policy relevant treatment effect parameters. I perform simulation studies using the estimators for the identified set proposed in the paper.

[*]Econometrics Topics Course, Summer 2023, University of Bonn.

[†]University of Bonn

Contents

1 Introduction

2 The Identification Problem

2.1 Model Setup

Mogstad, Santos, and Torgovitsky (2018) use an IV model based on a selection equation, an approach summarized for example in Heckman and Vytlacil (2007a) and Heckman and Vytlacil (2007b). They key to the model is the selection equation which determines treatment status as a function of observed covariates (including the instrument) and unobserved heterogeneity in the likelihoood to select into treatment. The key selection problem usually is that this unobserved heterogeneity is correlated with treatment effects or the level of potential outcomes. The instrumental variable solves this problem by shifting people into or out of treatment in a way uncorrelated to their unobserved heterogeneity.

They key model is formulated as follows: We consider a model with binary treatment $D \in \{0,1\}$. For convenience I drop all the subscripts throughout the paper. The **outcome equation** relates potential outcomes (the outcomes observed if individuals where exogenously assigned their value $D = d$) and treatment to *observed* outcomes Y by

$$Y = Y_1 D + Y_0(1-D). \tag{1}$$

Treatment D itself is determined by the *choice equation*, which relates treatment status to observed covariates (in particular the instrument denoted by Z) and unobserved heterogeneity U:

$$D = I\{p(Z) - U \geq 0\}. \tag{2}$$

U is modeled to follow a standard Uniform distribution, although this is not a restriction because for any continuous U we can redefine the selection equation by applying the CDf F_U on both sides of the inequality. With this normalization, $p(Z) = P(D = 1|Z = 1)$ and thus is the *propensity score*, the probability to take up treatment conditional on observed covariates. U can be understood as the "resistance" to treatment: conditional on the propensity score (i.e. observables), individuals with a sufficiently high realization of U will never take up treatment.

While the model can be formulated to include both exogenous covariates X and "outside" instruments Z_0 (so $Z = (Z_0, X)$), in what follows I focus on the case without any covariates. Thus all the following statements will not include any condiitoning on X.

IV Model: In addition to the outcome equation and choice equation, the IV model requires three further assumptions

- I.1 $U \perp Z_0$
- I.2 $E[Y_d|Z,U] = E[Y_d|U]$ and $E[Y_d^2] < \infty$ for $d \in \{0,1\}$.
- I.3 U has a uniform distribution on $[0,1]$ conditional on Z.

The first two assumptions guarantee exogeneity of Z_0 (exogenous shift in the choice probability and no direct effect on potential outcomes). The first assumption in combination with the additive separability of the choice equation, is equivalent to the monotonicity assumption in Angrist, Imbens, and Rubin (1996) that allows identificaiton of the LATE among instrument-compliers, a result proven by Vytlacil (2002).

For example, a binary IV $Z \in \{0,1\}$ with propensity score $p(0) = \underline{u} < p(1) = \overline{u}$ allows to identify $LATE(\underline{u}, \overline{u})$. Intuitively, individuals with realization of U in the interval $[\underline{u}, \overline{u}]$ are those for which the instrument realization randomly shifts them between treatment states (the compliers). Those with realizations smaller than $\underline{u}$ always take up treatment (the always-taker), while those with realizations larger than $\overline{u}$ never take up treatment. The next section introduces the identification or extrapolation problem.

2.2 Extrapolation

While the Imbens and Angrist (1994) result shows that we can identify a LATE in this model (or multiple LATE if Z takes on several values), these might not necessarily the parameters of interest. The key insight of Mogstad, Santos, and Torgovitsky (2018) is that many target parameter of interests *as well as* identified parameters like the LATE or IV slope coefficients are functions of the same underlying **marginal treatment response** (MTR) functions. The MTR functions are denoted m_0, m_1 and defined as

$$m_d(u) = E[Y_d | U = u] \tag{3}$$

For some target parameters, which will be denoted by β^*, writing them in terms of MTR functions is immediate. For example, $LATE(a,b)$ averages the difference $m_1(u) - m_0(u)$ over the range $u \in [a,b]$.

More generally, target parameters can be written in the form

$$\beta^* = E\left[\int_0^1 m_0(u,X)\omega_0^*(u,Z)d\mu^*(u)\right] + E\left[\int_0^1 m_1(u,X)\omega_1^*(u,Z)d\mu^*(u)\right] \tag{4}$$

where ω_d^* are identified weighting functions depending on the target parameter (e.g. 1 and -1 for the ATE).

A central result in the paper (Proposition 1) is that also all **IV-like estimands** of the form $E[s(D,Z)Y]$ are weighted averages of MTR functions:

$$\beta_s = E\left[\int_0^1 m_0(u,X)\omega_{0s}(u,Z)du\right] + E\left[\int_0^1 m_1(u,X)\omega_{1s}(u,Z)du\right] \tag{5}$$

where the ewights are $\omega_{0s} \equiv s(0,z)I[u > p(z)]$, $\omega_{1s} \equiv s(1,z)I[u \leq p(z)]$.

Introduce some further notation:

- S: Set of IV-like specifications implying identified parameters β_s.
- $\mathscr{M}$: Space of possible MTR functions, potentially including some a priori restrictions.
- $\mathscr{M}_{\mathscr{S}} \subseteq \mathscr{M}$: Sub-space of MTR functions *consistent* with identified estimands β_s for all $s \in S$.

Then the *identified set* for β^* denoted by $\mathscr{B}_S^*$ is the set of $b \in \mathbb{R}$ that is generated by some $m \equiv (m_0, m_1) \in \mathscr{M}_S$.

Proposition 2 in the paper establishes that for a convex $\mathscr{M}$ the identified is of the form $\mathscr{B}_S^* = [\underline{\beta}^*, \overleftarrow{\beta}^*] \subseteq \mathbb{R}$. Further, these bounds are the solution to an optimiuation problem over $m \in \mathscr{M}_S$ that can be recast as a linear program. In this program, the objective is to make the target parameter as small (or large) as possible while satisfying the constraint that at the optimal solution the chosen MTR functions imply the identified estimands (implicit in $m \in \mathscr{M}_S$).

Sharp identified set: Proposition 3 in the paper establihes that if we use "enough" IV-like specifications to identify $\mathscr{B}_S^*$, then this is the smallest set consistent with conditional means $E[Y|Z=z,D=d]$ and the model assumptions. For example, for a binary instrument we need to use all cross moments of the form $E[I\{Z=z\}I\{D=d\}Y]$. Intuitively, if we think about the numerator of the Wald estimand $E[Y|Z=1]-E[Y|Z=0]$ this differences out $E[Y_1]$ for the always-taker and $E[Y_0]$ for the never-taker, which allows to identify the (scaled) average treatment effect for the complier subpopulation. However, these moments itself constraint the admissable MTR functions so for extrapolation we want to use estimands that contain this information.

2.3 Implementation

In practice we need to consider a finite-dimensional parameter space $\mathscr{M}_{fd} \subseteq \mathscr{M}$. For example we can model $m_d(u,x)$ as a finite number of basis functions:

$$m_d(u) = \sum_{k=1}^{K_d} \theta_{dj} b_{dj}(u).$$

For the simulation exercise here, the setting is however a lot easier. Proposition 4 in the paper establishes that for a Z with discrete support and target weights on the MTR functions that are piecewise constant over u, a finite-dimensional space of MTR functions recovers the exact solution. In particular, we can use constant splines as the basis functions, defined over a partition of u where all relevant weights (target and identified parameters) are constant.

It is useful to define linear maps Γ and Γ^* that takes as argument some $m \in \mathscr{M}$ and return a parameter β. In particular, define for identified estimands β_s

$$\Gamma_s(m) = E\left[\int_0^1 m_0(u,X)\omega_{0s}(u,Z)\,du\right] + E\left[\int_0^1 m_1(u,X)\omega_{1s}(u,Z)\,du\right] \tag{6}$$

and for the target parameter β^*

$$\Gamma^*(m) = E\left[\int_0^1 m_0(u,X)\omega_0^*(u,Z)\,du\right] + E\left[\int_0^1 m_1(u,X)\omega_1^*(u,Z)\,du\right]. \tag{7}$$

In both cases, ω_{ds} are the relevant weights implied by the IV-like specification.

Then our MTR space constrained by the identified estimands becomes

$$\mathscr{M}_S \equiv \{m \in \mathscr{M} : \Gamma_s(m) = \beta_s \text{ for all } s \in S\}. \tag{8}$$

and the identified set is given by

$$\mathscr{B}_S^* \equiv \{b \in \mathscr{R} : b = \Gamma^*(m) \text{ for some } m \in \mathscr{M}_S\}. \tag{9}$$

Because these maps are linear and we use a linear combination of basis functions to approcximate $\mathscr{M}$ we can restate the problem as

$$\bar{\beta}_{\text{fd}}^\star \equiv \sup_{(\theta_0,\theta_1)\in\Theta} \sum_{k=1}^{K_0} \theta_{0k}\Gamma_0^\star(b_{0k}) + \sum_{k=1}^{K_1} \theta_{1k}\Gamma_1^\star(b_{1k})$$

s.t. $\sum_{k=1}^{K_0} \theta_{0k}\Gamma_{0s}(b_{0k}) + \sum_{k=1}^{K_1} \theta_{1k}\Gamma_{1s}(b_{1k}) = \beta_s \quad$ for all $s \in \mathscr{S}$.

In the case I consider in this study with discrete Z and a target parameter (LATE) that has constant weights, Proposition 4 applies and we can get exact results using constant splines as basis functions. In particular, these have to use knots corresponding to the points at which either the target parameter or the propensity score changes. For example, with a propensity score of $p = (0.35, 0.6, 0.7)$ and a target parameter $LATE(0.35, 0.9)$ we use a partition of u with $u_{part} = [0, 0.35, 0.6, 0.7, 0.9, 1]$, implying five basis functions for each of the intervals. As argued in the paper, for the case of constant splines (as well as Bernstein polynomials) the integrals in the linear maps $\Gamma_d^*(b_{dj})$ and $\Gamma_{ds}(b_{dj})$ can be solved analytically. For example, considering $d = 1$ and some IV-like specification $s(0,z)$, for Γ_{ds} (focusing on a single constant spline) we get:

$$\begin{aligned}
\Gamma_{0s}(b_{0j}) &= E_Z\left[\int_0^1 m_0(u)\, w_{0s}(u,Z)\, du\right] \\
&= E_Z\left[\int_0^1 I\{u \in [\underline{u}_j, \overline{u}_j]\}\theta_{0j}\, s(0,Z)\, I\{p(Z) < u\}\, du\right] \\
&= \theta_{0j} E_Z\left[s(0,Z)\int_0^1 I\{u \in [\underline{u}_j, \overline{u}_j]\} I\{p(Z) < \underline{u}_j\}\, du\right] \\
&= \theta_{0j} E_Z\left[s(0,Z)\, I\{p(Z) < \underline{u}_j\}(\overline{u}_j - \underline{u}_j)\right] \\
&= \theta_{0j}(\overline{u}_j - \underline{u}_j) \sum_{z \in \mathscr{Z}} f_Z(z)\, s(0,z)\, I\{p(z) < \underline{u}_j\}.
\end{aligned} \tag{10}$$

The second line follows from the definition of the weights and taking a constant spline as the basis functions, where θ_{0j} is the coefficient on the basis function b_{dj} corresponding to some element of the u-partition ranging from $[\underline{u}_j, \overline{u}_j]$. This θ_{0j}, as written earlier, will be one of the choice varibales in the linear program. The third line uses the fact that we use a partition such that weights (and thus also the propensity scores) are constant over a given element of the partition. Therefore, $p(Z) < u \iff p(Z) < \underline{u}_j$[1]. The last lines then pull out constants and write out the expectation using that Z has discrete support $mathcalZ$. An equivalent result holds for $d = 1$:

$$\Gamma_{1s}(b_{1j}) = \theta_{1j}(\overline{u}_j - \underline{u}_j) \sum_{z \in \mathscr{Z}} f_Z(z)\, s(1,z)\, I\{p(z) > \underline{u}_j\}. \tag{11}$$

3 The Estimation Problem

When observing only a random sample we cannot exactly satisfy the constraint that the optimizer *exactly* implies identified estimands β_s. Both the identified estimands and the weights on the constant splines will be estimated. Instead, Mogstad, Santos, and Torgovitsky (2018) propose to solve the following problem (stated for the upper bound):

$$\hat{\overline{\beta}}^* = \sup_{m \in \mathscr{M}} \hat{\Gamma}^*(m) \text{ s.t. } \sum_{s \in S} |\hat{\Gamma}_s(m) - \hat{\beta}_s| \leq \inf_{m' \in \mathscr{M}} \sum_{s \in S} |\hat{\Gamma}_s(m') - \hat{\beta}_s| + \kappa_n.$$

A few things to note:

- The upper bound makes the target estimand as large as possible for some $m \in \mathscr{M}$, but note that the linear map $\Gamma*(m)$ needs to be estimated (the weights on the MTR functions are functions of the data as will be clear

[1] $p(Z) = \underline{u}_j$ might hold with equality but then this only holds exactly for that point so the interval also evaluates to zero.

below).

- The constraint is reformulated:
  - All admissable $m \in \mathcal{M}$ have to come as close to the estimated identified estimands $\hat{\beta}_s$ as the MTR functions that are closest to satisfying it plus some tolerance κ_n.
  - The tolerance κ_n has to shrink with the sample size. If κ_n is too large the bounds will be too wide, while a very small κ_n will introduce a lot of noise (e.g. think $\kappa_n = 0$ which leaves the minimizer on the RHS as the only solution).

This implies that we now have to solve a first-step linear program that finds the minimizer to the problem on the RHS of the constraint. [Briefly explain how this is done using tricks for the absolute value.] I explore the choice of κ_n in the simulations below but generally find $\frac{1}{N}$ or $\frac{1}{N^2}$ to result in similar estimates with MSE considerably lower than $\frac{1}{\sqrt{N}}$ or $\frac{1}{N^{\frac{1}{4}}}$.

Mogstad, Santos, and Torgovitsky (2018) propose the following plug-in estimators

$$\hat{\Gamma}_{ds}(b_{dk}) \equiv \frac{1}{n}\sum_{i=1}^{n}\int_0^1 b_{dk}(u, X_i)\,\hat{\omega}_{ds}(u, Z_i)\,d\mu^\star(u),$$
$$\text{where } \hat{\omega}_{0s}(u,z) \equiv \hat{s}(0,z)\mathbb{1}[u > \hat{p}(z)]$$
$$\text{and } \hat{\omega}_{1s}(u,z) \equiv \hat{s}(1,z)\mathbb{1}[u \leq \hat{p}(z)],$$

where $\hat{s}$ is an estimator of s, and $\hat{p}$ is an estimator of the propensity score. An estimator of $\hat{\Gamma}^\star_d(b_{dk})$ can be constructed similarly as

$$\hat{\Gamma}^\star(b_{dk}) \equiv \frac{1}{n}\sum_{i=1}^{n}\int_0^1 b_{dk}(u, X_i)\,\hat{\omega}^\star_d(u, Z_i)\,d\mu^\star(u),$$

where $\hat{\omega}^\star_d$ is an estimator of $\omega^\star_d$, the form of which will depend on the form of the target parameter. As pointed out in the paper, these estimators simplify considerably with constant spline basis functions for some parameters because the integrals can actually be solved analytically [TODO do this in the simulation/code; add formulas for this; this could be potential bug]. β_s can be estimated based on

$$\hat{\beta}_s \equiv \frac{1}{n}\sum_{i=1}^{n}\hat{s}(D_i, Z_i)\,Y_i.$$

Appendix Proposition S3 establishes the consistency of their procedure, in particular that $\hat{\underline{\beta}}^* \to_p \underline{\beta}^*$ and $\hat{\overline{\beta}}^* \to_p \overline{\beta}^*$.

In our case, the estimators of the linear map simplify to

$$\hat{\Gamma}_{1s}(b_{1j}) = \theta_{1j}(\overline{u}_j - \underline{u}_j)\sum_{i=1}^{n}\hat{s}(1, Z_i)I\{\hat{p}(Z_i) > \underline{u}_j\}. \tag{12}$$

and similarly for $d == 0$. Here, we replace the expectation by sample moments and $s(1,z)$ and $p(z)$ by their estimated counterparts.

3.1 Implementation

The estimation procedure I implement closely follows the suggestions in Mogstad, Santos, and Torgovitsky (2018). In particular I solve the estimation problem in three steps:

Before solving the two linear programs I first estimate all the required objects. These include

- The identified estimands: $\beta_s = E[s(D,Z)Y]$ for $s \in S$.
- The "weights" on the θ_{dj} choice variables implied by the target and identified estimands via the linear maps $\Gamma*$ and Γ_{ds} (see equation 12).
- The propensity score $p(z)$.

Step 1 — LP for the Constraint: The first task is to solve the minimization problem in the RHS of the constraint in equation 3. This solves for the θ_{jd} that imply the minimal deviation from the estimates of the identified estimands β_s in terms of absolute loss. As suggested in the paper, to model absolute loss in the objective function it is possible to introduce dummy variables that have to satisfy two constraints corresponding to the negative and positive part of each $|Ga\hat{m}ma_s(m')\hat{\beta}_s|$ [2].

Step 2 — LP for the Bounds: With the solution from Step 1 at hand and chosen tolerance level κ_n, we can now solve the two linear programs corresponding to (3). These are standard linear programs with the objective of maximizing (minimizing) the implied target estimand. Now the constraint includes absolute values (the deviations from estimated target parameters in the LHS of the constraint in (3)), which we can again address by introducing dummy variables and additional constraints. The resulting optimal solutions constitute the upper and lower bounds.

I solve both linear programs using AMPL, specifically the AMPL Python API. AMPL licenses are free for researchers and graduate students. The solver I used was "HiGHS" or "gurobi". All estimation, data handling and analysis steps are performed in Python. I discuss some practical issues relating to the tolerance level and the u-partition after introducing the DGP and main simulation analysis below.

4 Simulation Setting

I use the main data generating process (DGP) used in the numerical example by Mogstad, Santos, and Torgovitsky, 2018. They have a discrete instrument with the following specifications:

- Support of Z: $Z \in \{0,1,2\}$;
- Density of Z: $f_Z(0) = 0.5$, $f_Z(1) = 0.4$, $f_Z(2) = 0.1$;
- Propensity score: $P(d = 1|Z = 0) \equiv p(0) = 0.35$,.

Note the setup has no covariates X. Following Imbens and Angrist (1994) three local average treatment effects (LATE) are point-identified: LATE(0.35, 0.6), LATE(0.6,0.7), and LATE(0.35, 0.7). This will show up in the identification results below, which cover point-identification as a special case.

While the paper assumes a binary outcome I directly simulate (potential) outcomes using the underlying MTR

[2] For example, with a program of the form $\min|X|$ introducing a dummy X', and objective $\min X'$ with two constraints $X \leq X'$ and $-X \leq X'$ mimicks the desired program. In our case the objective will generally be of the form $\min \sum_{s\in S}|X_s|$ which can be modeled by introducing $|S|$ variables and $2|S|$ constraints. I mainly used the lpSolve manual 5.1 as a reference.

functions, such that $Y_d = m_d(u)$ which imply outcomes Y. These are of course no longer binary, but all essential moments remain the same so consistency should not be affected [3].

I study a range of different targets of the form $LATE(0.35, \overline{u})$, where $0.35 \leq \overline{u} <= 1$.

I focus on the sharp, non-parametric bounds depicted in Figure 5 of Mogstad, Santos, and Torgovitsky, 2018. These are constructed using all cross-moments of D, Z with the data Y, i.e. IV-like estimands of the form

$$\beta_s = E[I\{D = d, Z = z\}Y]$$

for $d \in \{0,1\}$ and $z \in \{0,1,2\}$. Proposition 3 in the paper establishes that this set of identified estimands delivers the sharpest bounds that are consistent with the conditional means of Y and the assumptions of the model.

Throughout the analysis I impose the restriction that the MTR functions m are bounded between $0,1$ (formulated as $\theta_{jd} \in [0,1]$), a restriction imposed in the paper and immediate for a binary outcome.

In the following section I report a Figure similar to Figure 8 in the paper, which reports bounds for LATE$(0.35, \overline{u})$ for a range of values, although with a different parametric assumption on the MTRs. In addition to the identification result (which in the plot I call the "true bounds"), I also report estimation results. For a grid of $\overline{u}$ values, I estimate the bounds and plot their means and distributions. Thus, the resulting plot consists of a separate simulation for each value of $\overline{u}$[4]. Throughout, for each simulation I use a sample size of $N = 10000$, $R = 1000$ replications and a tolerance for the identification constraint equivalent to $\frac{1}{N}$.

Before turning to the main results in the next section I want to discuss some practical issues I encountered during implementation. First, the choice of the **tolerance criterion** κ_n seems to be non-trivial. In Appendix Figure A.3 I report results for varying tolerances ($\kappa_n \in \{\frac{1}{\sqrt{n}}, \frac{3}{n}, \frac{1}{n}, \frac{1}{n^2}\}$). A higher κ_n should result in (weakly) wider bounds beause more MTR pairs satisfy the constraint. Choosing κ_n too lower might result in a very noisy estimator; in the extreme case $\kappa_n = 0$ by construction we only have a single solution (assuming there is a unique minmizer which seems highly likely given the sampling). As expected I observe that bounds are significantly wider for $\kappa_n = \frac{1}{\sqrt{n}}$ and narrower when choosing on of the smaller tolerances. Choosing a $\kappa_n < \frac{1}{\sqrt{n}}$ also significatnly reduced bias. Beyond a κ_n of the order $\frac{1}{n}$ there does not seem to be much of a change, however.

Second, I found that estimation results are highly unstable when the **partition of** u includes values very close to each other. When I first ran the simulation below which uses a target $LATE(0.35, \overline{u})$ I ran into this issue: Because $p(0) = 0.35$ for a large enough N, $\hat{p}(1)$ will be very close to 0.35, thus resulting in a partition that includes 0.35 as well as a number very close to it. The result is a bi-modal distrbution of the estimator for the lower bound, as shown in Appendix Figure C.5. Similar issues arise for the upper bound when $\overline{u}$ comes close to $p(1)$ or $p(2)$. The two modes for the lower bound exactly correspond to situations where we estimate $\hat{p}(0) < 0.35$ (smaller mode) or $\hat{p}(0) > 0.35$ (larger mode). A simple fix to this issue I rely on in the following results is to remove one of the close knots in the partition. This seems also the relevant usecase in practice: When a researcher estimates the (to them unknown) propensity score to be $\hat{p}(0)$ close to 0.35 and wants to extrapolate, $LATE(0.35, 0.9)$ is not a natural choice of the

[3] If we want to use a binary Y consistent with the DGP one way would be to draw outcomes from $\{0,1\}$ with probability corresponding to $m_{d_i}(u_i)$ for each $i = 1, \ldots, n$. This might be preferred if we want to do simulations for an actual DGP corresponding to some observed data, because otherwise we understate the variance Y.

[4] This would not be the correct simulation design if we were interested in studying the joint distribution of estimators for different targets, but should be sufficient for studying individual consistency.

target paramter, but instead $LATE(\hat{p}(0) \approx 0.35, 0.9)$.

5 Results

The figure below reports identification and estimation results for each $LATE(0.35, \overline{u})$.

Identification: Turning to identification first, a few things are noteworthy. First, as alluded to earlier we achieve point identification for $LATE(0.35, 0.6)$ and $LATE(0.35, 0.7)$ as indicated by the upper and lower bound coinciding at these points. In that sense, the procedure covers point identification as a special case. Second, as argued in the paper, extrapolation to parameters "further" away from what is point-identified results in wider bounds. For example, for $\overline{u} \geq 0.7$ bounds are monotonically increasing in $\overline{u}$. In cases of multiple point-identified estimands this relationship is obviously not monotone in $\overline{u}$ for all values, as shown for example by the range $\overline{u} \in [0.6, 0.7]$ in between the two point-identified estimands. Third, in stark contrast to the Figure 8 reported in the paper, which imposes parametric assumptions on the underlying MTR functions, the sharp non-parametric bounds reported here quickly become completely uninformative as $\overline{u} < 0.6$ with bounds $[-1, 1]$ for $\overline{u} \leq 0.45$. This highlights the importance of assumptions on the ability to extrapolate. Interestingly, the sharp non-parametric bounds here are generally more narrow for $\overline{u} \geq 0.7$ than $\overline{u} \leq 0.6$, implying that the data alone in the form of the conditional moments $E[Y|D, Z]$ contain more information about these parameters. Imposing parametric assumptions in Figure 8 of the paper makes these bounds approximately equally wide, implying that the relative importance of parametric assumptions strongly varies by the target, i.e. $\overline{u}$.

Estimation: In terms of estimation, I first finde that the average bounds in Figure 1 are for the most part too wide. This holds in particular for allmost all $\overline{u}$ for the estimator of the lower bound, while the upper bound overestimates the true bound for all values $\overline{u} \in [0.45, 0.75]$. For any larger values, the estimator for the upper bound has a mean indistinguishable from the true bound. Generally, for the given sample size the distributions of the estimators are fairly tightly centered around their respective means. For example, almost none of the individuals simulation runs for any of the lower bound estimators come cloe to the true estimator. While the Mogstad, Santos, and Torgovitsky (2018) only prove consistency and there is no reason to believe that the estimator is unbiased in finite samples, the quality of the result does not change when increasing the sample size. If this was only a finite sample bias we would at least hope that the bias decreases when increasing N. Appendix Figure B.4 reports simulation results in similar form for $N \in \{25000, 50000, 100000\}$. While the standard deviation of the estimator decreases as expected, estimated means are largely unchanged. If true, this would call into question the consistency of the estimator in this setting[5].

Figure 2 additionally plots the distribution of the lower and upper bound estimators for a fixed $\overline{u} = 0.95$. As noted, before this figure shows that the lower bound estimator considerabl underestimates the true bound, a feature also present for a large sample of $N = 50,000$. The upper bound estimator, however, for this target parameter is tightly centered around the true mean. To get a sense of the distribution, the plot overlays corresponding normal denities with matched means and variances. The upper bound estimator clearly has a non-standard asymmetric distribution, with a median to the right of the mean. For the lower bound estimator the distribution actually looks somewhat closer to a normal, although this is hard to judge as $R = 500$ probably introduces noticeable simulation noise.

[5]While technically the estimator might still converge at some point, this seems unlikely. Also, for these sample sizes we would in any way hope for a lot smaller finite sample bias.

Figure 1: Simulation Results for Sharp Non-Parametric Bounds

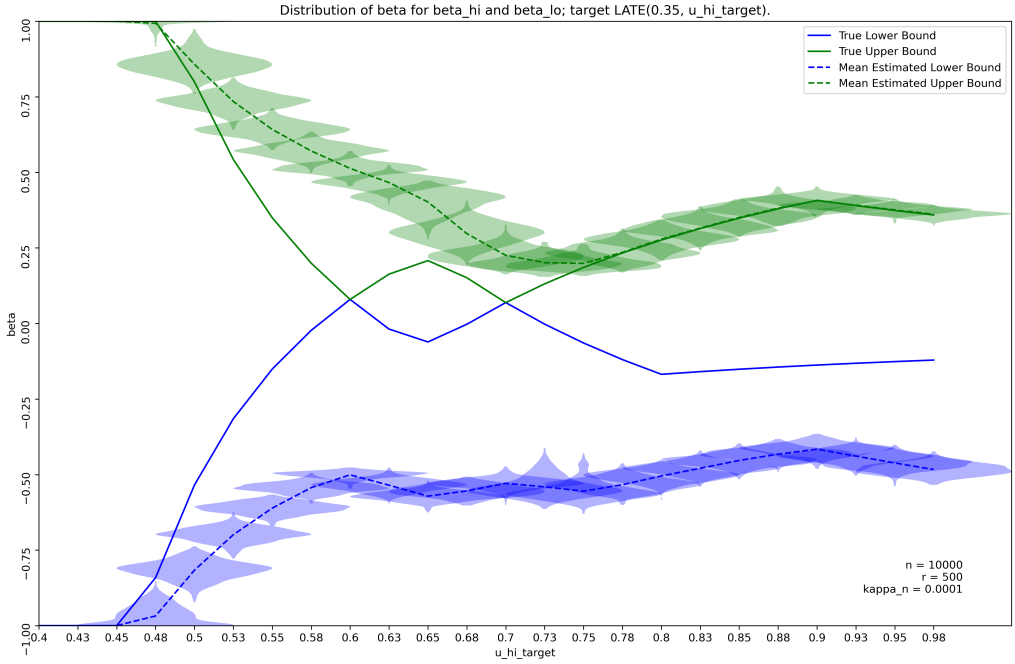
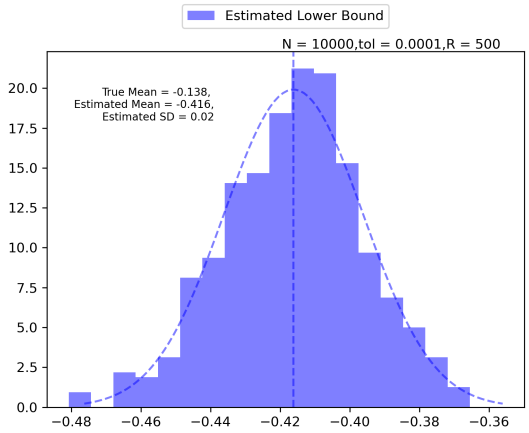
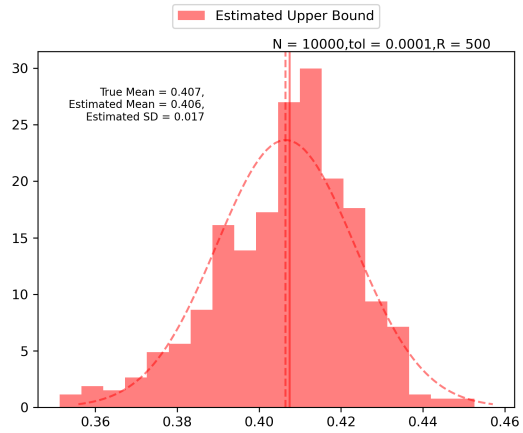


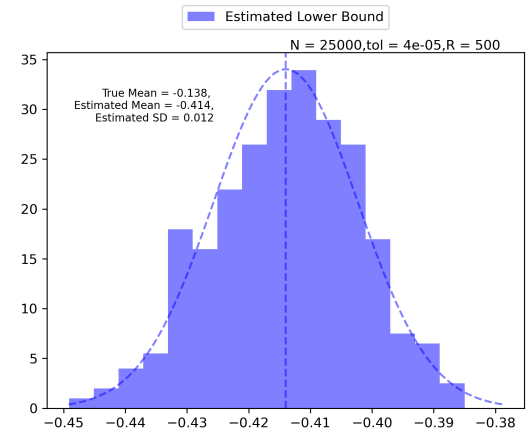
Figure 2: Distribution of Estimator for $LATE(0.35, 0.9)$ by N



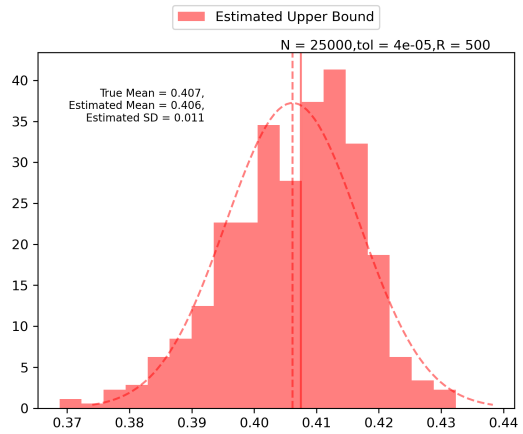
(a) $N = 10,000$



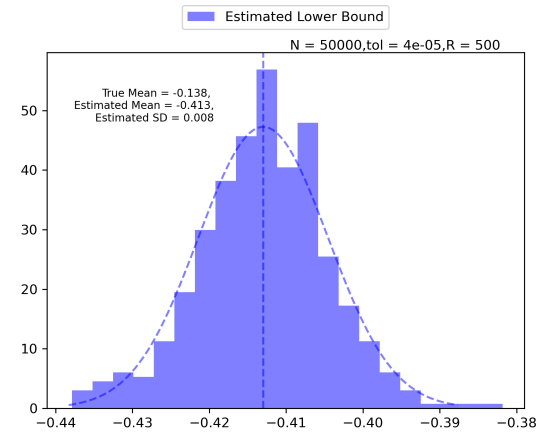
(b) $N = 10,000$



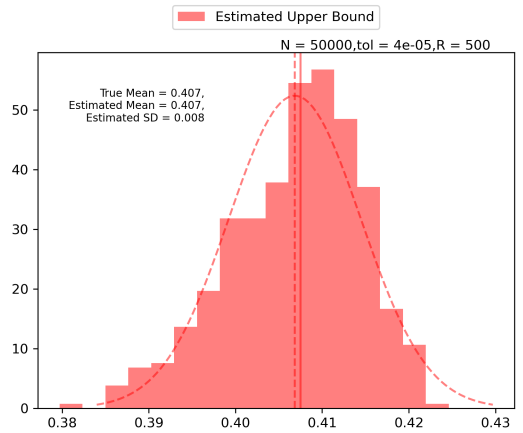
(c) $N = 25,000$



(d) $N = 25,000$



(e) $N = 50,000$



(f) $N = 50,000$

Notes: This figure compares simulation results for the main specificaiton, a fixed target parameter $LATE(0.35, 0.9)$ and different sample sizes N. The solid line indicates the true upper bound, the true lower bound is given by -0.13. The dashed line correspond to the mean of the estimator. Additionally, the plot shows a normal distrbution with the mean and variance equal to the moments of the estimator distributions.

6 Conclusion

While the (identificaiton) approach proposed by Mogstad, Santos, and Torgovitsky (2018) offers a framework for applied researchers enabling them to coherently think about and identify relevant target estimands, I think it would be beneficial if more was known on the behavior of these estimators by the means of simulation results. Simulation are neither reported in Mogstad, Santos, and Torgovitsky (2018) or their working paper version, nor the *ivmte* R-package implementation described in Shea and Torgovitsky (2021). While the results here should be viewed very cautiously, as they are probably subject to implementation error, this is also one of the simplest possible settings in terms of DGP and assumptions[6]. Yet, already some problems showed up that make the estimator hard to implement. In settings applied researchers want to use, say with Z taking on many more values, other covariates X, or while imposing restrictions on the MTRs, additional issues might show up that might be worth investigating.

Thus, future analysis could focus on the behavior of the estimator in these types of settings. Beyond that, Mogstad, Santos, and Torgovitsky (2018) propose a bootstrap method for constructing confidence intervals for the identified sets in their working paper version. However, they already note that no accepted procedure for these types of problems exist. While a lot more computationally demanding, it would be interesting to have simulation results on these or think about a different approaches to quantify uncertainty about the estimator[7]. Lastly, applied reseachers are concerned with problems about weak instrments (Andrews, Stock, and Sun, 2019). It would be interesting to study how these translate into the estimation of the bounds. From the point of identificaiton of course, there is no problem with weak instrments — as long as the propensity score varies over Z, the instruments identify something, only bounds for extrapolation will be wide. Still the estimator may suffer from a particularly bad finite sample bias or the bootstrap approach to CI construction might fail[8].

References

Andrews, Isaiah, James H Stock, and Liyang Sun (2019). "Weak instruments in instrumental variables regression: Theory and practice". In: *Annual Review of Economics* 11, pp. 727–753.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin (1996). "Identification of causal effects using instrumental variables". In: *Journal of the American statistical Association* 91.434, pp. 444–455.

Heckman, James J and Edward Vytlacil (2007a). "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation". In: *Handbook of econometrics* 6, pp. 4779–4874.

– (2007b). "Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments". In: *Handbook of econometrics* 6, pp. 4875–5143.

Imbens, Guido W and Joshua D Angrist (1994). "Identification and estimation of local average treatment effects". In: *Econometrica: journal of the Econometric Society*, pp. 467–475.

[6] I also tried to replicate the analysis using the *ivmte* package by Shea and Torgovitsky (2021). However, for the same DGP, similar restrictions and using a saturated regressions for the identified moments I found the resulting distrbutions of the bounds to be even more biased. It is hard to tell, however, how exactly the estimator is implemented, as the package does not exactly follow the original Mogstad, Santos, and Torgovitsky (2018) paper.

[7] Maybe it would be worthiwhile recasting this as a Bayesian problem, although this might give up the attractive distinction between model uncertainty (in the form of the identified set) and sampling unertainty (the frequentist CI for that set).

[8] Mogstad, Santos, and Torgovitsky (2018) briefly mention the weak IV in footnote 3 and suggest to instead dropping the first-stage scaling by $\frac{1}{Cov(D,Z_0)}$ from the IV-like specification, which they argue imposes the same constraints on the MTR functions.

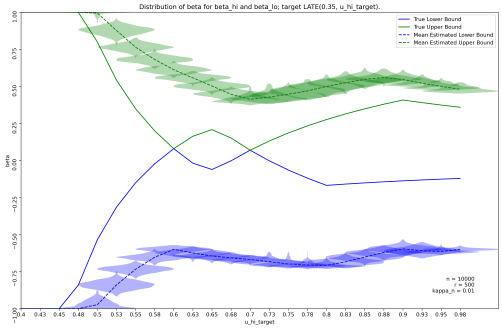
Mogstad, Magne, Andres Santos, and Alexander Torgovitsky (2018). “Using instrumental variables for inference about policy relevant treatment parameters”. In: *Econometrica* 86.5, pp. 1589–1619.

Shea, Joshua and Alexander Torgovitsky (2021). “ivmte: An R package for implementing marginal treatment effect methods”. In: *University of Chicago, Becker Friedman Institute for Economics Working Paper* 2020-01.

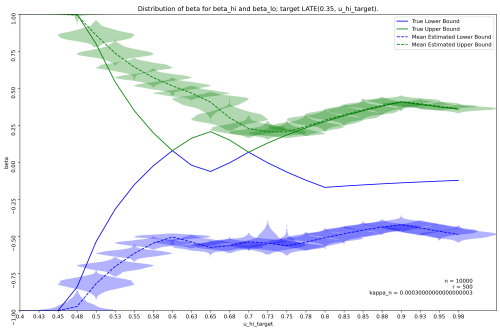
Vytlacil, Edward (2002). “Independence, monotonicity, and latent index models: An equivalence result”. In: *Econometrica* 70.1, pp. 331–341.

A Choice of Tolerance

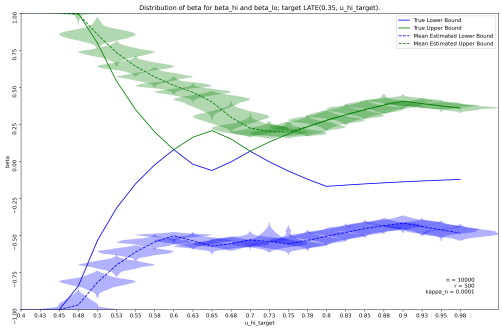
Figure A.3: Simulation Results by Choice of κ_n



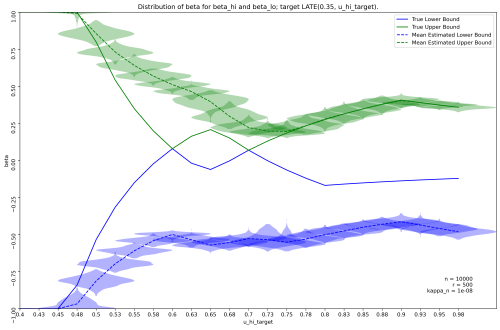
(a) $\kappa_n = \frac{1}{\sqrt{n}}$



(b) $\kappa_n = \frac{3}{n}$



(c) $\kappa_n = \frac{1}{n}$

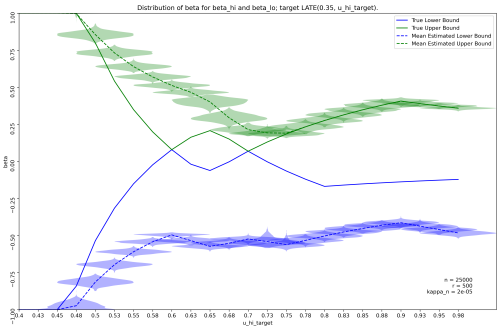


(d) $\kappa_n = \frac{1}{n^2}$

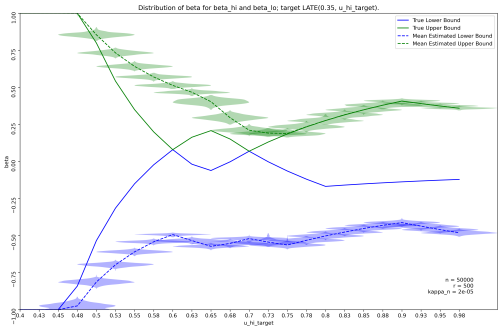
Notes: This figure compares simulation results for the main specificaiton described in Section 4 for different tolerances levels κ_n.

B Results for Larger Samples

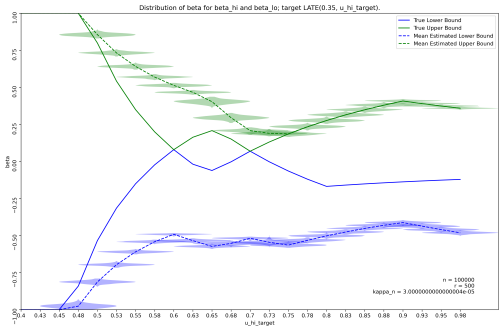
Figure B.4: Simulation Results for Larger Sample Sizes



(a) $N = 25,000$



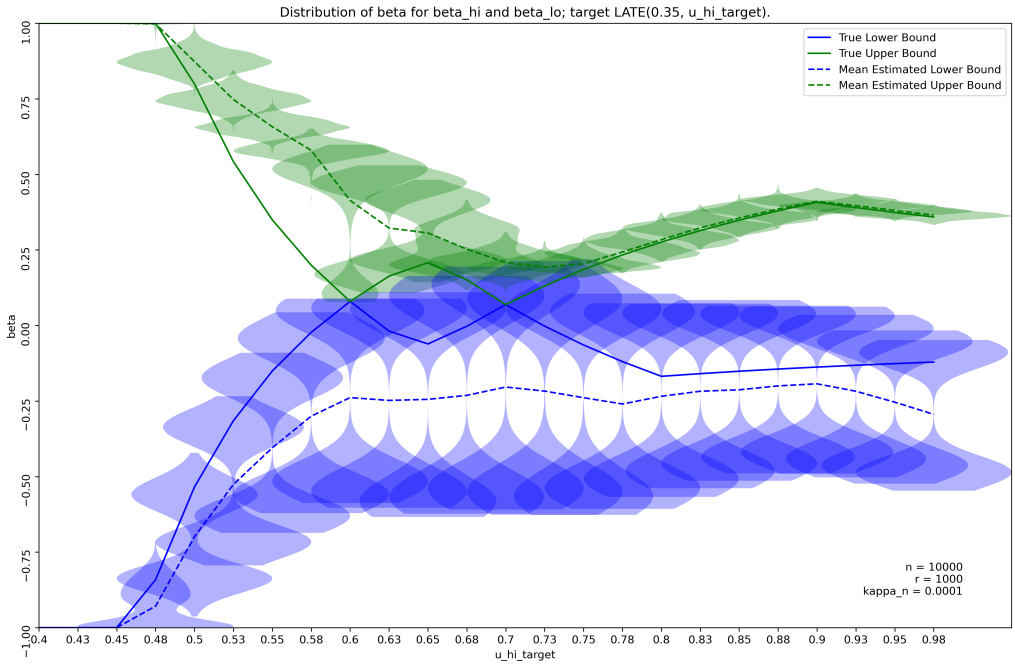
(b) $N = 50,000$



(c) $N = 100,000$

Notes: This figure compares simulation results for the main specificaiton described in Section 4 different samples sizes N.

Figure C.5: Simulation Results: Bi-Modal Distribution



Notes: This figure shows results for the main specification when we use a partition including both estimated $\hat{p}(0)$ close to 0.35, and 0.35 in the partition for u.

C Unstable Estimator