Econometrics Topics Course: Simulation

Julian Budde

August 24, 2023

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 **heckman2007econometric1** and **heckman2007econometric2**. 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 **angrist1996identification** that allows identificaiton of the LATE among instrument-compliers, a result proven by **vytlacil2002independence**.

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

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

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:

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

Introduce some further notation:

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

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

Proposition 2 in the paper establishes that for a convex $\mathcal{M}$ the identified is of the form $\mathcal{B}^*_{\mathcal{S}} = [\underline{\beta}^*, \overleftarrow{\beta}^*] \subseteq \mathbb{R}$. Further, these bounds are the solution to an optimiuation problem over $m \in \mathcal{M}_{\mathcal{S}}$ 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 $m \in \mathcal{M}_{\mathcal{S}}$).

Sharp identified set: Proposition 3 in the paper establihes that if we use "enough" IV-like specifications to identify $\mathcal{B}^*_{\mathcal{S}}$, 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 treatment effect (or reduced form) 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 $\mathcal{M}_{fd} \subseteq \mathcal{M}$. For example we can model $m_d(u,x)$ as a finite number of basis functions:

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

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.

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):

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

A few things to note:

- The upper bound makes the target estimand as large as possible for some $m \in \mathcal{M}$, 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 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 $\hat{\underline{\beta}}^{*} \rightarrow_p \underline{\beta}^{*}$ and $\hat{\overline{\beta}}^{*} \rightarrow_p \overline{\beta}^{*}$.

3.1 Implementation

[Describe here how this simplifies in our setup with constant splines.]

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.

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.

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.

Throughout I use a sample size of $N = 10000$, $R = 1000$ replications and a tolerance for the identification constraint equivalent to $\frac{1}{N}$.

5 Results

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

Identification:

- We achieve point identification for $\overline{u} \in \{0.6, 0.7\}$.
- Generally, as argued in the paper, extrapolation to parameters further away from what is point-identified results in wider bounds.
- This is particularly striking for smaller values of $\overline{u}$, where for $\overline{u} < 0.45$ bounds become completely uninformative ranging from -1 to 1.
- This is in stark contrast to the paper, and highlights the importance of shape restrictions (or in the case of the paper parametric assumptions) on the MTR functions.
- Bounds for larger values are a lot smaller, highlighting that in some sense the data is more informative about individuals with a high resistance to treatment, i.e. high u (Todo: Understand how so? "Informative"?).

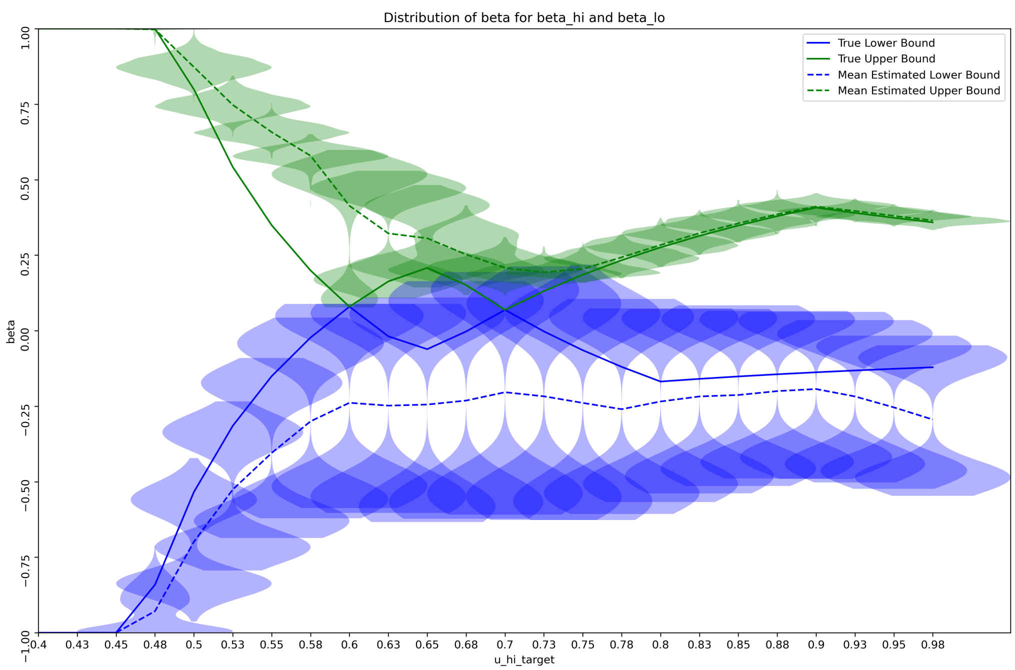
Estimation:

- Estimated bounds are on average wider than the true bounds.
- Most notably, for many of the target parameters the distribution is **bimodal**.
- This is particulary true for the lower bounds which has a bimodal distribution for *all* values of $\overline{u}$.
- While the for values up to 0.7 one of the modes seems to be close to the true bound, this is not the case for larger values of $\overline{u}$.
- For the upper bounds distributoins are only bimodal in the range 0.6 to 0.7 (roughly or exactly?)
- For all targets > 0.7 the upper bound estimator is tightly centered.

Questions

- Mean coverage might not be informative if for all estimators one bound covers while the other does not with roles reversing.

Figure 1: Simulation Results for Sharp Non-Parametric Bounds



6 Potential Future Analysis

- Weak IV
  - Need to think about how this would be relevant; really weak IV is a failure of the regular asymptotic approximation to the finite sample distribution thereby making inference unreliable. This is not really an identification problem; i.e. for a discrete IV as long as not all propensity scores are the same we can always point identify some LATE. However, weak IV might show up in estimation of the bounds.
  - Case of weak IV/many weak IV
  - Weak IV for IV slope –> Use reduced form/cross-moments instead of IV-slope coefficient; thereby avoiding weak IV problem. Is this a trade-off between estimation and inference? I.e. identification bounds become wider but inference becomes more reliable?

A Choice of Tolerance and General Simulation Results

Table A1: True Lower Bound: -0.138 True Upper Bound: 0.407 (N = 1000, R = 1000)

	$\hat{\beta}_{lo}$	$\hat{\beta}_{hi}$	sd_{lo}	sd_{hi}	tolerance	MSE_{lo}	MSE_{hi}
0	-0.208877	0.414150	0.223488	0.056918	0.000001	0.055002	0.003284
1	-0.220794	0.432195	0.223509	0.056564	0.001000	0.056848	0.003810
2	-0.516540	0.636670	0.223578	0.042400	0.031623	0.193446	0.054320
3	-0.908906	0.998436	0.087718	0.007537	0.177828	0.602330	0.349271

Table A2: True Lower Bound: -0.138 True Upper Bound: 0.407 (N = 10000, R = 1000)

	$\hat{\beta}_{lo}$	$\hat{\beta}_{hi}$	sd_{lo}	sd_{hi}	tolerance	MSE_{lo}	MSE_{hi}
0	-0.194711	0.409557	0.224812	0.017092	0.000000	0.053782	0.000296
1	-0.183867	0.410482	0.225333	0.016916	0.000100	0.052899	0.000295
2	-0.364951	0.539501	0.226147	0.017154	0.010000	0.102749	0.017720
3	-0.736040	0.867981	0.207569	0.017359	0.100000	0.400999	0.212351

Table A3: True Lower Bound: -0.138 True Upper Bound: 0.407 (N = 25000, R = 1000)

	$\hat{\beta}_{lo}$	$\hat{\beta}_{hi}$	sd_{lo}	sd_{hi}	tolerance	MSE_{lo}	MSE_{hi}
0	-0.179065	0.408356	0.225880	0.010979	0.000000	0.052726	0.000121
1	-0.185695	0.408570	0.225684	0.011040	0.000040	0.053229	0.000123
2	-0.297524	0.518655	0.225895	0.010378	0.006325	0.076547	0.012465
3	-0.679197	0.791070	0.224184	0.013218	0.079527	0.343390	0.147307

Table A4: True Lower Bound: -0.448 True Upper Bound: 0.734 (N = 1000, R = 1000)

	$\hat{\beta}_{lo}$	$\hat{\beta}_{hi}$	sd_{lo}	sd_{hi}	tolerance	MSE_{lo}	MSE_{hi}
0	-0.208877	0.414150	0.223488	0.056918	0.000001	0.107130	0.105307
1	-0.220794	0.432195	0.223509	0.056564	0.001000	0.101582	0.094062
2	-0.516540	0.636670	0.223578	0.042400	0.031623	0.054684	0.011199
3	-0.908906	0.998436	0.087718	0.007537	0.177828	0.220122	0.070180

Table A5: True Lower Bound: -0.448 True Upper Bound: 0.734 (N = 10000, R = 1000)

	$\hat{\beta}_{lo}$	$\hat{\beta}_{hi}$	sd_{lo}	sd_{hi}	tolerance	MSE_{lo}	MSE_{hi}
0	-0.194711	0.409557	0.224812	0.017092	0.000000	0.114700	0.105315
1	-0.183867	0.410482	0.225333	0.016916	0.000100	0.120545	0.104710
2	-0.364951	0.539501	0.226147	0.017154	0.010000	0.058041	0.037980
3	-0.736040	0.867981	0.207569	0.017359	0.100000	0.126047	0.018352

Table A6: True Lower Bound: -0.448 True Upper Bound: 0.734 (N = 25000, R = 1000)

	$\hat{\beta}_{lo}$	$\hat{\beta}_{hi}$	sd_{lo}	sd_{hi}	tolerance	MSE_{lo}	MSE_{hi}
0	-0.179065	0.408356	0.225880	0.010979	0.000000	0.123352	0.105923
1	-0.185695	0.408570	0.225684	0.011040	0.000040	0.119741	0.105785
2	-0.297524	0.518655	0.225895	0.010378	0.006325	0.073674	0.046321
3	-0.679197	0.791070	0.224184	0.013218	0.079527	0.103707	0.003474

Table A7: True Lower Bound: -0.69 True Upper Bound: 0.674 (N = 1000, R = 1000)

	$\hat{\beta}_{lo}$	$\hat{\beta}_{hi}$	sd_{lo}	sd_{hi}	tolerance	MSE_{lo}	MSE_{hi}
0	-0.208877	0.414150	0.223488	0.056918	0.000001	0.281207	0.070691
1	-0.220794	0.432195	0.223509	0.056564	0.001000	0.269897	0.061603
2	-0.516540	0.636670	0.223578	0.042400	0.031623	0.079997	0.003181
3	-0.908906	0.998436	0.087718	0.007537	0.177828	0.055714	0.105404

Table A8: True Lower Bound: -0.69 True Upper Bound: 0.674 (N = 10000, R = 1000)

	$\hat{\beta}_{lo}$	$\hat{\beta}_{hi}$	sd_{lo}	sd_{hi}	tolerance	MSE_{lo}	MSE_{hi}
0	-0.194711	0.409557	0.224812	0.017092	0.000000	0.295627	0.070150
1	-0.183867	0.410482	0.225333	0.016916	0.000100	0.306716	0.069656
2	-0.364951	0.539501	0.226147	0.017154	0.010000	0.156652	0.018348
3	-0.736040	0.867981	0.207569	0.017359	0.100000	0.045226	0.037983

Table A9: True Lower Bound: -0.69 True Upper Bound: 0.674 (N = 25000, R = 1000)

	$\hat{\beta}_{lo}$	$\hat{\beta}_{hi}$	sd_{lo}	sd_{hi}	tolerance	MSE$_{lo}$	MSE$_{hi}$
0	-0.179065	0.408356	0.225880	0.010979	0.000000	0.311844	0.070615
1	-0.185695	0.408570	0.225684	0.011040	0.000040	0.305027	0.070503
2	-0.297524	0.518655	0.225895	0.010378	0.006325	0.204887	0.024197
3	-0.679197	0.791070	0.224184	0.013218	0.079527	0.050370	0.013912