Bounding Program Benefits When Participation is Misreported

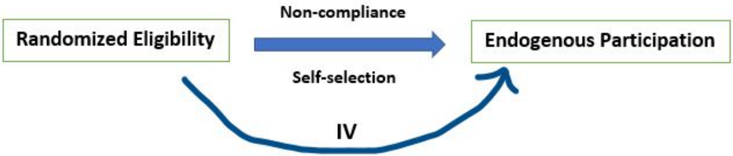
Denni Tommasi[⋆] and Lina Zhang[+]

[⋆]University of Bologna and IZA,
[+]University of Amsterdam and Tinbergen Institute

EEA-ESEM 2022

Introduction

Aim: Program benefits under endogenous participation



Problem: participation is endogenously misreported

- Stigma of welfare program, privacy concern, social bads

→ This paper: measure program benefits on "those who really take it up"

Model setup

- $$Y = DY_1 + (1-D)Y_0,$$
  $$D = \sum_{k=0}^{K} 1[Z = z_k] D_k,$$
  $$T = DT_1 + (1-D)T_0$$

- Z binary, discrete and multiple discrete IV(s)
  - Mogstad-Torgovitsky-Walters (2020) – more than 50% papers in top journals "use multiple IVs"
- $D \in \{0,1\}$ true treatment
- $T \in \{0,1\}$ misreported treatment
- $(T_0, T_1) \in \{0,1\}^2$ misclassification and $(Y_1, Y_0) \not\perp (T_1, T_0)$

Target Estimands

- **Local average treatment effect** (Imbens-Angrist, 1994)

$$LATE_k = \alpha_k^* = E[Y_1 - Y_0 | C_k]$$
$$\text{Compliers} = C_k = \{D_k = 1, D_{k-1} = 0\}$$

- **IV estimand**

$$\alpha^* = \frac{\text{Cov}(Y, g(Z))}{\text{Cov}(D, g(Z))} = \sum_{k=1}^{K} \gamma_k^* \alpha_k^*,$$

where known fun $g : \Omega_Z \mapsto \mathbb{R}$ and weight $\gamma_k^* \geq 0$ and $\sum_k \gamma_k^* = 1$

Bias due to misclassification

Identifiable estimand $$\alpha = \frac{\text{Cov}(Y, g(Z))}{\text{Cov}(T, g(Z))}$$

		false positive w^p					
Rel. Bias $\frac{\alpha}{\alpha^*} - 1$		0	0.05	0.10	0.20	0.30	0.40
false negative w^n	0	0	0.05	0.11	0.25	0.43	0.67
	0.05	0.05	0.11	0.18	0.33	0.54	0.82
	0.10	0.11	0.18	0.25	0.43	0.67	1.00
	0.20	0.25	0.33	0.43	0.67	1.00	1.50
	0.30	0.43	0.54	0.67	1.00	1.50	2.33
	0.40	0.67	0.82	1.00	1.50	2.33	4.00

Note: $w^n = \Pr(T = 0|D = 1)$; $w^p = \Pr(T = 1|D = 0)$

- Misreporting **inflates** treatment effect: $|\alpha^*| < |\alpha|$
- **Severe** bias even with infrequent errors

Contribution

1. **Partial identification** of LATE and IV estimand with discrete IV(s)
2. **External information of misreporting rates** to tighten bounds
3. **Re-examine benefits of the 401(k) pension plan on savings**
    - improve comparable bound in the literature by 36%

→ STATA package "`ivbounds`" (Lin-Tommasi-Zhang, 2021)

Literature

- Exogenous treatment & exogenous misclassification
  - Point id. (e.g., Mahajan (2006), Lewbel (2007), and Hu (2008))
  - Partial id. (e.g., Klepper (1988), Bollinger (1996))
- Endogenous treatment & exogenous misclassification
  - Point id. (e.g., Battistin-Nadai-Sianesi (2014), Yanagi (2017), DiTraglia-GarciaJimeno (2018), Calvi-Lewbel-Tommasi (2021))
  - Partial id. (e.g., Calvi-Lewbel-Tommasi (2021))
- **Endogenous treatment & endogenous misclassification**
  - Point id. (e.g., Nguimkeu-Denteh-Tchernis (2018))
  - Partial id. (e.g., **Ura (2018)**)
- **External information/administrative data**
  - e.g., Dushi-Iams (2010), Kreider-Gundersen-Jolliffe (2012), Meyer-Mittag-Goerge (2018), Meyer-Mittag (2019)

Assumptions

Assumption 1. Imbens and Angrist (1994)

Valid IV and Monotonicity

Note: monotonicity under multiple IVs $\Rightarrow$ homogeneous treatment choice across individuals

Assumption 2. Treatment misclassification

- $Z \perp (T_1, T_0)$
- (T **is better than pure guess on** D) For $d = \{0, 1\}$,

$$Pr(T = 0|C_k, D = 1) < 0.5, \quad Pr(T = 1|C_k, D = 0) < 0.5$$

Bias in α

Theorem. Naive IV estimand

Let Assumptions 1-2 hold:

$$\alpha = \frac{\text{Cov}(Y, g(Z))}{\text{Cov}(T, g(Z))} = \sum_{k=1}^{K} \gamma_k \alpha_k^*,$$

Corollary. Bias of α

Let Assumptions 1-2 hold:

$$\alpha^* = \xi\alpha, \quad \text{where } \xi = \sum_{k=1}^{K} \gamma_k^* \xi_k,$$

denote

$$\xi_k = 1 - Pr(T = 0|C_k, D = 1) - Pr(T = 1|C_k, D = 0)$$
$$\xi = 1 - w^n - w^p \in [0,\ 1].$$

Bounding Probability of Compliers

- **Why $Pr(C_k)$??**

$$\text{LATE} = \alpha_k^* = \frac{ITT_k}{Pr(C_k)}$$

where

$$\begin{aligned} ITT_k =& E[Y|Z = z_k] - E[Y|Z = z_{k-1}] \quad \text{identifiable} \\ Pr(C_k) =& E[D|Z = z_k] - E[D|Z = z_{k-1}] \quad \text{unknown.} \end{aligned}$$

- **Solution**: Total variation distance

$$TV_k = \frac{1}{2} \int \left| f_{(Y,T)|Z=z_k}(x) - f_{(Y,T)|Z=z_{k-1}}(x) \right| dx.$$

  – distributional "ITT" effect of IV(s) on observables (Y, T)

Bounding Probability of Compliers

Lemma (Ura, 2018)

Use subpopulation $Z = z_k$ and $Z = z_{k-1}$,

$$TV_k \leq Pr(C_k) \leq 1.$$

Lemma 1. Multiple and multi-valued IV(s)

Under Assumptions 1-2, for $\forall k = 1, 2, ..., K$,

$$TV_k \leq Pr(C_k) \leq 1 - \sum_{k' \neq k} TV_{k'}.$$

- We gain identification power by using multiple total variation distances

Bounding LATE $\alpha_k^* = E[Y_1 - Y_0 | C_k] = ITT_k / Pr(C_k)$

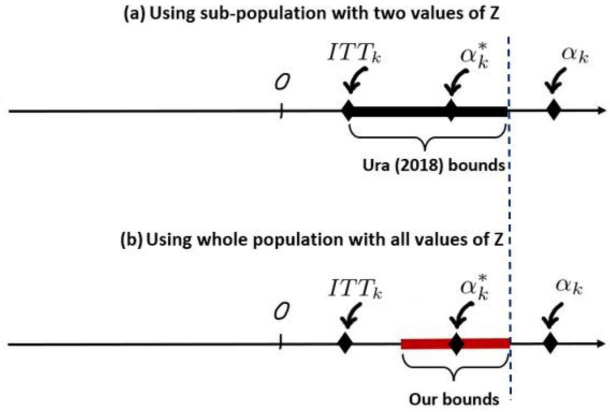
Theorem

(1) **Bound of LATE for C_k:**

$$\alpha_k^* \in \begin{cases} \left[\frac{ITT_k}{1-\sum_{k' \neq k} TV_{k'}}, \frac{ITT_k}{TV_k}\right], & \text{if } ITT_k > 0, \\ \{0\}, & \text{if } ITT_k = 0, \\ \left[\frac{ITT_k}{TV_k}, \frac{ITT_k}{1-\sum_{k' \neq k} TV_{k'}}\right], & \text{if } ITT_k < 0; \end{cases} \tag{1}$$

(2) Set in (1) is **sharp** if $TV_k > 0$ and $TV_{k'} = 0$ for $\forall k' \neq k$

Bounding LATE $\alpha_k^* = E[Y_1 - Y_0 | C_k]$



- Our bounds $\subseteq$ two-value IV bounds (Ura, 2018) $\subseteq [ITT_k,\ \alpha_k]$, where

$$\alpha_k = \frac{E[Y|Z = z_k] - E[Y|Z = z_{k-1}]}{E[T|Z = z_k] - E[T|Z = z_{k-1}]}$$

Partial identification of α^*

Strategy 1 & 2 + no external info.

Strategy 1. $\alpha^* = \sum_{k=1}^{K} \gamma_k^* \alpha_k^*$

Because $\min_k \{\alpha_k^*\} \leq \alpha^* \leq \max_k \{\alpha_k^*\}$

$$\alpha^* \in \bigcup_k \left\{ \text{bounds of } \alpha_k^* \right\}.$$

Strategy 2. $\xi = \sum_{k=1}^{K} \gamma_k^* \xi_k$, where $\xi_k = \frac{E[T|Z=z_k]-E[T|Z=z_{k-1}]}{Pr(C_k)}$

Because $\min_k \{\xi_k\} \leq \xi \leq \max_k \{\xi_k\}$ and $\alpha^* = \xi\alpha$

$$\alpha^* \in \alpha \times \bigcup_k \left\{ \text{bounds of } \xi_k \right\}.$$

- Strategy 2 is better than 1, if **less heterogeneous** in ξ_k across k.

Partial identification of α^*

Strategy 3 + external info.

- External information
  - Administrative records, small validation studies, or repeated measures

Strategy 3. $\alpha^* = \xi\alpha$

Suppose $\xi \in [\underline{\xi}, \overline{\xi}] \in [0, 1]$ with **known** $\underline{\xi}$ and $\overline{\xi}$.

(1) If $\alpha \geq 0$, then $0 < \underline{\xi}\alpha \leq \alpha^* \leq \overline{\xi}\alpha$.

(2) If $\alpha \leq 0$, then $\overline{\xi}\alpha \leq \alpha^* \leq \underline{\xi}\alpha < 0$.

- Strategy 3 is at least the same or better than Strategy 2
- Point identification if $\xi = 1 - w^n - w^p$ is known

Numerical Illustration

Table: Identified Sets of LATE ($w^n = 0.1, w^p = 0.05$)

IV strength	α_1^*	[ITT$_1$, α_1]	two-value IV bound	our bound: single proxy	our bound: multi proxy
low	5	[0.68, 6.54]	[0.68, 5.21]	[1.91, 5.21]	[1.93, 5.13]
high		[1.42, 6.31]	[1.42, 5.19]	[3.74, 5.19]	[3.77, 5.12]

IV strength	α_2^*	[ITT$_2$, α_2]	two-value IV bound	our bound: single proxy	our bound: multi proxy
low	5	[1.70, 6.02]	[1.70, 5.16]	[2.48, 5.16]	[2.49, 5.09]
high		[2.67, 5.76]	[2.67, 5.15]	[3.72, 5.15]	[3.76, 5.08]

- our bound $\subseteq$ two-value IV bound $\subseteq [ITT_k, \alpha_k]$
- Stronger IV strength $\Longrightarrow$ narrower bounds

Numerical Illustration

Table: Bounds of α^* ($\alpha^* = 5, w^n \approx 0.1, w^p \approx 0.05$)

IV strength	α	S1	S2	S3 $\xi \in [1 - 2w^n, 1 - w^n]$	S3 $\xi = 1 - w^n - w^p$
low	6.2	[1.91, 5.21]	[1.84, 5.37]	[4.80, 5.34]	5.10
high	5.9	[3.72, 5.19]	[3.61, 5.31]	[4.71, 5.30]	5.00

- Biased point identification (red) with information of ξ
- Inference
  - Testing moment inequalities (Chernozhukov-Chetverikov-Kato, 2019)
  - Intersecting bounds and bias correction (Chernozhukov-Lee-Rosen, 2013)

Conclusion

- Measure the program benefits when participation is **misclassified**
- Our method has several applications
  - leading identification strategy
  - robustness check
  - sensitivity analysis

Empirical example

- **Benefit of 401(k) pension plan on savings?**
  - **Aim**: increase savings via tax deduction
  - **IVs**: firm eligibility + duration of exposure to the plan (from 1981)
  - **Endogenous** participation frequently **misreported**
  - **In SIPP**:

    w^n =17% of participants self-report as non-participants

    w^p =10% of non-participants self-report as participants

Table: Empirical Results (Panel A: Binary instrument)

Naive α		**Bounds LATE α^***			
2SLS Abadie (2003)	$\frac{cov(Y,Z)}{cov(T,Z)}$	Ura (2018)	Strategy 1 $\equiv$ 2	Strategy 3	
				$\xi \in [1-2w^n, 1-w^n]$	$\xi = 1-w^n-w^p$
9.4 (5.3, 13.5)	16.3 (6.0, 27.6)	(4.4, 28.3)	(4.4, 28.3)	(4.7, 21.2)	11.9 (5.2, 18.6)

Note: 95% CI is in parentheses.

- Compared Ura's, our bounds in Strategy 3 is $1 - \frac{(21.2-4.7)-(28.3-4.4)}{28.3-4.4} = \textbf{36\%}$ narrower in width

Table: Empirical Results (Panel B: Discrete instrument)

	Naive α	Bounds WLATE α^*			
		Strategy 1	Strategy 2	Strategy 3	
				$\xi \in [1-2w^n, 1-w^n]$	$\xi = 1-w^n-w^p$
Stratum 1	21.8 (16.3, 27.3)	(2.5, 42.4)	(2.9, 29.4)	(11.2, 23.0)	15.9 (12.2, 20.2)
Stratum 2	23.1 (19.2, 27.0)	(2.3, 70.1)	(4.6, 28.2)	(12.7, 22.4)	16.9 (14.0, 19.7)
Stratum 3	54.5 (44.3, 64.8)	(19.2, 120.9)	(15.5, 68.2)	(29.6, 53.2)	39.8 (32.8, 46.7)

Note: 95% CI is in parentheses. In Panel B, stratify samples based on Pr$(T=1|X)$.

- Compare two point estimates: naive α (red) is **37%** larger than that in Strategy 3 (blue)

Inference

Testing for moment inequalities (Chernozhukov-Chetverikov-Kato (2019))

Theorem: CI of LATEs

Denote $\mathcal{C}_k(\beta)$ as the confidence interval of α_k.

(i) (Size) $\mathcal{C}_k(\beta)$ controls the asymptotic size uniformly over $\mathcal{P}_0$

(ii) (Power) For any $\alpha_k \notin \Theta_k$, Pr $[\alpha_k \notin \mathcal{C}_k(\beta)] \to 1$.

Corollary. CI of α^*

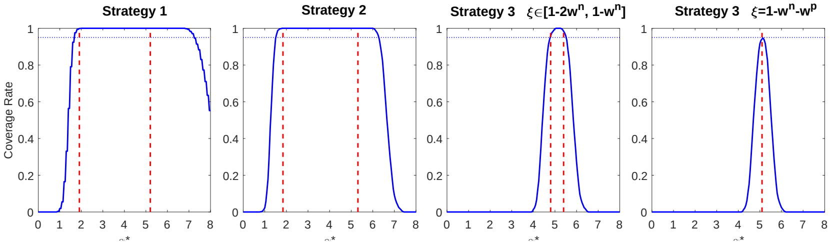
Denote $\mathcal{C}(\beta)$ as the confidence interval of α^*. For all three strategies,

$$\liminf_{n\to\infty} \inf_{\mathbf{P}\in\mathcal{P}_0,\ \alpha^*\in\Theta(\mathbf{P})} Pr\left[\alpha^{IV} \in \mathcal{C}(\beta)\right] \geq 1-\beta,$$

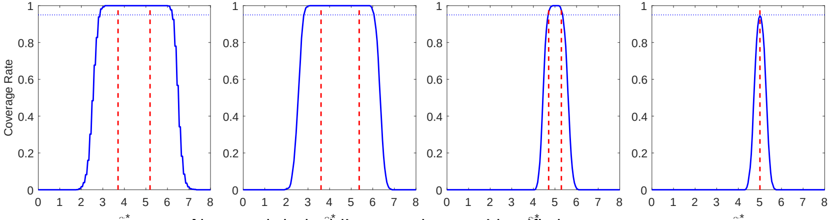
with significance level β.

low IV strength, $\alpha^* = 5$, $\alpha \approx 6$

Figure: Coverage Rates of the 95% Confidence Intervals



high IV strength, $\alpha^* = 5$, $\alpha \approx 6$



Note: red dashed lines are the true identified set.