Mis-Classified, Binary, Endogenous Regressors: Identification and Inference

Francis J. DiTraglia

Camilo García-Jimeno

University of Pennsylvania

January 7th, 2018

Additively Separable Model

$$y = h(T^*, \mathbf{x}) + \varepsilon$$

- y – Outcome of interest
- h – Known or unknown function
- T^* – Unobserved, endogenous binary regressor
- T – Observed, mis-measured binary surrogate for T^*
- $\mathbf{x}$ – Exogenous covariates
- ε – Mean-zero error term

What is the Effect of T^*?

Re-write the Model

$$
\begin{aligned}
y &= c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon \\
\beta(\mathbf{x}) &= h(1,\mathbf{x}) - h(0,\mathbf{x}) \\
c(\mathbf{x}) &= h(0,\mathbf{x})
\end{aligned}
$$

This Paper:

- Does a discrete instrument z (typically binary) identify $\beta(\mathbf{x})$?
- What assumptions are required for z and the surrogate T?
- How to carry out inference for a mis-classified regressor?

Example: Smoking and Birthweight (SNAP Trial)

Coleman et al. (N Engl J Med, 2012)

RCT with 1050 pregnant smokers in England: 521 given nicotine patches, the rest given placebo patches.

- y – Birthweight
- T^* – True smoking behavior
- T – Self-reported smoking behavior
- $\mathbf{x}$ – Mother characteristics
- z – Indicator of nicotine patch

“Baseline” Assumptions I – Model & Instrument

Additively Separable Model

$y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon, \quad \mathbb{E}[\varepsilon] = 0$

Valid & Relevant Instrument: $z \in \{0, 1\}$

- $\mathbb{P}(T^* = 1|\mathbf{x}, z = 1) \neq \mathbb{P}(T^* = 1|\mathbf{x}, z = 0)$
- $\mathbb{E}[\varepsilon|\mathbf{x}, z] = 0$
- $0 < \mathbb{P}(z = 1|\mathbf{x}) < 1$

If T^* were observed, these conditions would identify β.

"Baseline" Assumptions II – Measurement Error

Notation: Mis-classification Rates

"$\uparrow$" $\alpha_0(\mathbf{x}, z) \equiv \mathbb{P}\left(T = 1 \middle| T^* = 0, \mathbf{x}, z\right)$

"$\downarrow$" $\alpha_1(\mathbf{x}, z) \equiv \mathbb{P}\left(T = 0 \middle| T^* = 1, \mathbf{x}, z\right)$

Mis-classification unaffected by z

$$\alpha_0(\mathbf{x}, z) = \alpha_0(\mathbf{x}), \quad \alpha_1(\mathbf{x}, z) = \alpha_1(\mathbf{x})$$

Extent of Mis-classification

$\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x}) < 1$ (T is positively correlated with T^*)

Non-differential Mis-classification

$$\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon|\mathbf{x}, z, T^*]$$

Identification Results from the Literature

Mahajan (2006) Theorem 1, Frazis & Loewenstein (2003)

$\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*] = 0$, plus "Baseline" $\implies \beta(\mathbf{x})$ identified

Requires (T^*, z) jointly exogenous.

Mahajan (2006) A.2

$\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon|\mathbf{x}, T^*]$, plus "Baseline" $\implies \beta(\mathbf{x})$ identified

Allows T^* endogenous, but we prove this claim is false.

Open Question

Do the baseline assumptions identify $\beta(\mathbf{x})$ when T^* is endogenous?

Sharp Identified Set under Baseline Assumptions

Theorem

Under the baseline assumptions, the sharp identified set for $\beta(\mathbf{x})$ is never a singleton, regardless of how many (discrete) values the instrument z takes on.

Point identification from slightly stronger assumptions?

Point Identification: 1st Ingredient

Reparameterization

$$\theta_1(\mathbf{x}) = \beta(\mathbf{x}) / \left[1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right]$$

$$\theta_2(\mathbf{x}) = \left[\theta_1(\mathbf{x})\right]^2 \left[1 + \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right]$$

$$\theta_3(\mathbf{x}) = \left[\theta_1(\mathbf{x})\right]^3 \left[\left\{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right\}^2 + 6\alpha_0(\mathbf{x})\left\{1 - \alpha_1(\mathbf{x})\right\}\right]$$

$$\boxed{\beta(\mathbf{x}) = 0 \iff \theta_1(\mathbf{x}) = \theta_2(\mathbf{x}) = \theta_3(\mathbf{x}) = 0}$$

Lemma

Baseline Assumptions $\implies$ $\text{Cov}(y, z|\mathbf{x}) = \theta_1(\mathbf{x})\text{Cov}(z, T|\mathbf{x})$.

Point Identification: 2nd Ingredient

Assumption (II)

$\mathbb{E}[\varepsilon^2|\mathbf{x}, z] = \mathbb{E}[\varepsilon^2|\mathbf{x}]$

Lemma

(Baseline) + (II) $\implies$

$$\mathrm{Cov}(y^2, z|\mathbf{x}) = 2\mathrm{Cov}(yT, z|\mathbf{x})\theta_1(\mathbf{x}) - \mathrm{Cov}(T, z|\mathbf{x})\theta_2(\mathbf{x})$$

Corollary

(Baseline) + (II) + $[\beta(\mathbf{x}) \neq 0] \implies [\alpha_1(\mathbf{x}) - \alpha_0(\mathbf{x})]$ is identified.

Hence, $\beta(\mathbf{x})$ is identified if mis-classification is one-sided.

Point Identification: 1st Ingredient

Assumption (III)

(i) $\mathbb{E}[\varepsilon^2|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon^2|\mathbf{x}, z, T^*]$

(ii) $\mathbb{E}[\varepsilon^3|\mathbf{x}, z] = \mathbb{E}[\varepsilon^3|\mathbf{x}]$

Lemma

(Baseline) + (II) + (III) $\implies$

$$\mathrm{Cov}(y^3, z|\mathbf{x}) = 3\mathrm{Cov}(y^2 T, z|\mathbf{x})\theta_1(\mathbf{x}) - 3\mathrm{Cov}(yT, z|\mathbf{x})\theta_2(\mathbf{x}) + \mathrm{Cov}(T, z|\mathbf{x})\theta_3(\mathbf{x})$$

Point Identification Result

Theorem

(Baseline) + (II) + (III) $\implies \beta(\mathbf{x})$ is point identified. If $\beta(\mathbf{x}) \neq 0$, then $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ are likewise point identified.

Proof Sketch

1. $\beta(\mathbf{x}) = 0 \iff \theta_1(\mathbf{x}) = 0$ so suppose this is not the case.
2. Lemmas: full-rank linear system in $\theta_1(\mathbf{x}), \theta_2(\mathbf{x}), \theta_3(\mathbf{x})$ & observables.
3. Non-linear eqs. relating $\theta_1(\mathbf{x}), \theta_2(\mathbf{x}), \theta_3(\mathbf{x})$ to $\beta(\mathbf{x})$ and $\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x})$. Show that solution exists and is unique.

Sufficient Conditions for (II) and (III)

(i) T is conditionally independent of (ε, z) given $(T^*, \mathbf{x})$

(ii) z is conditionally independent of ε given $\mathbf{x}$

Just-Identified System of Moment Equalities

Suppress dependence on **x** to simplify the notation from here on...

Collect Lemmas from Above:

$$\begin{aligned}
\text{Cov}(y,z) - \text{Cov}(T,z)\theta_1 &= 0 \\
\text{Cov}(y^2,z) - 2\text{Cov}(yT,z)\theta_1 + \text{Cov}(T,z)\theta_2 &= 0 \\
\text{Cov}(y^3,z) - 3\text{Cov}(y^2T,z)\theta_1 + 3\text{Cov}(yT,z)\theta_2 - \text{Cov}(T,z)\theta_3 &= 0
\end{aligned}$$

Notation: Observed Data Vector

$$\mathbf{w}_i' = (T_i, y_i, y_iT_i, y_i^2, y_i^2T_i, y_i^3)$$

Just-Identified System of Moment Equalities

$$\boxed{\mathbb{E}\left[\left(\boldsymbol{\Psi}'(\boldsymbol{\theta})\mathbf{w}_i - \boldsymbol{\kappa}\right) \otimes \begin{pmatrix} 1 \\ z_i \end{pmatrix}\right] = \mathbf{0}}$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\psi}_1 & \boldsymbol{\psi}_2 & \boldsymbol{\psi}_3 \end{bmatrix} \qquad \boldsymbol{\kappa} = (\kappa_1, \kappa_2, \kappa_3)' \equiv \text{ “Intercepts”}$$

$$\boldsymbol{\psi}_1' = \begin{bmatrix} -\theta_1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \theta_1 = \beta/(1-\alpha_0-\alpha_1)$$

$$\boldsymbol{\psi}_2' = \begin{bmatrix} \theta_2 & 0 & -2\theta_1 & 1 & 0 & 0 \end{bmatrix} \qquad \theta_2 = \theta_1^2\left[1+\alpha_0-\alpha_1\right]$$

$$\boldsymbol{\psi}_3' = \begin{bmatrix} -\theta_3 & 0 & 3\theta_2 & 0 & -3\theta_1 & 1 \end{bmatrix} \qquad \theta_3 = \theta_1^3\left[(1-\alpha_0-\alpha_1)^2 + 6\alpha_0(1-\alpha_1)\right]$$

Weak Identification Problem

Moment conditions are uninformative about (α_0, α_1) when β is small.

Non-standard Inference Problem

- β small $\Rightarrow$ moment equalities uninformative about (α_0, α_1)
- (α_0, α_1) could be on the boundary of the parameter space
- Partial identification bounds remain informative even if β is small or zero
- Same problem for other estimators from the literature but hasn't been pointed out...

Our Approach

Identification-robust inference combining equality and inequality moment conditions based on generalized moment selection (GMS)

Inference With Moment Equalities and Inequalities

Moment Conditions

$$\mathbb{E}\left[m_j(\mathbf{w}_i, \vartheta_0)\right] \geq 0, \quad j = 1, \cdots, J$$

$$\mathbb{E}\left[m_j(\mathbf{w}_i, \vartheta_0)\right] = 0, \quad j = J+1, \cdots, J+K$$

Test Statistic

$$T_n(\vartheta) = \sum_{j=1}^{J} \left[\frac{\sqrt{n}\ \bar{m}_{n,j}(\vartheta)}{\widehat{\sigma}_{n,j}(\vartheta)}\right]_-^2 + \sum_{j=J+1}^{J+K} \left[\frac{\sqrt{n}\ \bar{m}_{n,j}(\vartheta)}{\widehat{\sigma}_{n,j}(\vartheta)}\right]^2$$

$$[x]_- = \min\{x, 0\}$$

$$\bar{m}_{n,j}(\vartheta) = n^{-1} \sum_{i=1}^{n} m_j(\mathbf{w}_i, \vartheta)$$

$$\widehat{\sigma}^2_{n,j}(\vartheta) = \text{consistent est. of AVAR} \left[\sqrt{n}\ \bar{m}_{n,j}(\vartheta)\right]$$

Moment Inequalities: Part I

$\alpha_0(\mathbf{x}) \leq p_k \leq 1 - \alpha_1$ becomes $\mathbb{E}\left[m^I_{1k}(\mathbf{w}_i, \boldsymbol{\vartheta})\right] \geq \mathbf{0}$ for all k where

$$m^I_{1k}(\mathbf{w}_i, \boldsymbol{\vartheta}) \equiv \left[\begin{array}{l} \mathbf{1}(z_i = k)(T - \alpha_0) \\ \mathbf{1}(z_i = k)(1 - T_i - \alpha_1) \end{array}\right]$$

Moment Inequalities: Part II

For all k, we have $\mathbb{E}[m^l_{2k}\left(\mathbf{w}_i, \vartheta, \mathbf{q}_k\right)] \geq 0$ where

$$m^l_{2k}\left(\mathbf{w}_i, \vartheta, \mathbf{q}_k\right) \equiv \left[\begin{array}{r} y_i \mathbf{1}\left(z_i = k\right)\left\{(T_i - \alpha_0) - \mathbf{1}(y_i \leq \underline{q}_{0k})(1 - T_i)\left(\frac{1-\alpha_0-\alpha_1}{\alpha_1}\right)\right\} \\ -y_i \mathbf{1}(z_i = k)\left\{(T_i - \alpha_0) - \mathbf{1}(y_i > \overline{q}_{0k})(1 - T_i)\left(\frac{1-\alpha_0-\alpha_1}{\alpha_1}\right)\right\} \\ y_i \mathbf{1}\left(z_i = k\right)\left\{(T_i - \alpha_0) - \mathbf{1}(y_i \leq \underline{q}_{1k}) T_i \left(\frac{1-\alpha_0-\alpha_1}{1-\alpha_1}\right)\right\} \\ -y_i \mathbf{1}(z_i = k)\left\{(T_i - \alpha_0) - \mathbf{1}(y_i > \overline{q}_{1k}) T_i \left(\frac{1-\alpha_0-\alpha_1}{1-\alpha_1}\right)\right\} \end{array}\right]$$

and $\mathbf{q}_k \equiv (\underline{q}_{0k},\, \overline{q}_{0k},\, \underline{q}_{1k},\, \overline{q}_{1k})'$ defined by $\mathbb{E}[h^l_k(\mathbf{w}_i, \vartheta, \mathbf{q}_k)] = 0$ with

$$h^l_k(\mathbf{w}_i, \vartheta, \mathbf{q}_k) = \left[\begin{array}{l} \mathbf{1}(y_i \leq \underline{q}_{0k})\mathbf{1}(z_i = k)(1 - T_i) - \left(\frac{\alpha_1}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(T_i - \alpha_0) \\ \mathbf{1}(y_i \leq \overline{q}_{0k})\mathbf{1}(z_i = k)(1 - T_i) - \left(\frac{1-\alpha_0}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(1 - T_i - \alpha_1) \\ \mathbf{1}(y_i \leq \underline{q}_{1k})\mathbf{1}(z_i = k) T_i - \left(\frac{1-\alpha_1}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(T_i - \alpha_0) \\ \mathbf{1}(y_i \leq \overline{q}_{1k})\mathbf{1}(z_i = k) T_i - \left(\frac{\alpha_0}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(1 - T_i - \alpha_1) \end{array}\right]$$

Inference via Generalized Moment Selection

Andrews & Soares (2010)

Moment Selection Step

If $\dfrac{\sqrt{n}\,\bar{m}_{n,j}(\vartheta_0)}{\widehat{\sigma}_{n,j}(\vartheta_0)} > \sqrt{\log n}$ then drop inequality j

Critical Value

- $\sqrt{n}\,\bar{m}_n(\vartheta_0) \to_d$ normal limit with covariance matrix $\Sigma(\vartheta_0)$
- Use this to bootstrap the limit distribution of the test statistic.

Theoretical Guarantees

Uniformly valid test of $H_0\colon \vartheta = \vartheta_0$ regardless of whether ϑ_0 is identified.

Not asymptotically conservative.

Drawback

Joint test for the whole parameter vector but we're only interested in β

Bonferroni-Based Inference Procedure

Leverage Special Structure of Model

- β only enters MCs through $\theta_1 = \beta/(1 - \alpha_0 - \alpha_1)$
- Inference for θ_1 is standard if z is a strong IV.
- $(\boldsymbol{\kappa}, \mathbf{q})$ strongly identified under null for (α_0, α_1)

Procedure

1. Concentrate out $(\theta_1, \boldsymbol{\kappa}, \boldsymbol{q}) \implies$ joint GMS test for (α_0, α_1)
2. Invert $\implies (1 - \delta_1) \times 100\%$ confidence set for (α_0, α_1)
3. Project $\implies$ CI for $(1 - \alpha_0 - \alpha_1)$
4. Construct standard $(1 - \delta_2) \times 100\%$ IV CI for θ_1
5. Bonferroni $\implies (1 - \delta - \delta_2) \times 100\%$ CI for β

Conclusion

Summary

- Endogenous, mis-classified binary treatment.
- Usual (1st moment) IV assumption fails to identify β
- Derive sharp identified set.
- Stronger assumptions point identify β
- Identification-Robust Inference incorportating equality and inequality moment conditions.

Extensions / Future Work

- Arbitrary discrete T^*
- Endogenous Mis-classification: "returns to lying"