Estimating the Effect of a Mis-measured, Endogenous, Binary Regressor

Francis J. DiTraglia

Camilo García-Jimeno

University of Pennsylvania

June 18th, 2017

Additively Separable Model

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

- y – Outcome of interest
- m – 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

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

This Paper:

- Does a binary instrument z suffice to identify $\beta(\mathbf{x})$?
- What assumptions are required for z and "surrogate" T?
- How should we carry out inference in this model?

Example: Job Training Partnership Act (JPTA)

Heckman et al. (2000, QJE)

Randomized offer of job training, but about 30% of those *not* offered also obtain training and about 40% of those offered training don't attend. Estimate causal effect of *training* rather than *offer* of training.

- y – Log wage
- T^* – True training attendence
- T – Self-reported training attendance
- $\mathbf{x}$ – Individual characteristics
- z – Offer of job training

Related Literature

Continuous Treatment

Lewbel (1997, 2012), Schennach (2004, 2007), Chen et al. (2005), Hu & Schennach (2008), Song (2015), Hu et al. (2015)...

Binary, Exogenous Treatment

Aigner (1973), Bollinger (1996), Kane et al. (1999), Black et al. (2000), Frazis & Loewenstein (2003), Mahajan (2006), Lewbel (2007), Hu (2008)

Binary, Endogenous Treatment

Mahajan (2006), Shiu (2015), Ura (2015), Denteh et al. (2016)

Baseline Assumptions

Additively Separable Model

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

Valid & Relevant Instrument

$$\mathbb{E}[\varepsilon|\mathbf{x}, z] = 0, \quad \mathbb{E}\left[T^*|\mathbf{x}, z = 0\right] \neq \mathbb{E}\left[T^*|\mathbf{x}, z = 1\right]$$

Non-differential Measurement Error

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

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

(iii) $\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x}) \neq 1$ (T is relevant for T^*)

Observable Moments: $y = c + \beta T^* + \varepsilon$

	$z = 1$	$z = 2$	$\dots$	$z = K$
$T = 0$	$\bar{y}_{01}$ p_{01}	$\bar{y}_{02}$ p_{02}	$\dots$	$\bar{y}_{0K}$ p_{0K}
$T = 1$	$\bar{y}_{11}$ p_{11}	$\bar{y}_{12}$ p_{12}	$\dots$	$\bar{y}_{1K}$ p_{1K}

$$\bar{y}_{tk} = \mathbb{E}[y \mid T = t, z = z_k], \quad p_{tk} = q_k p_k$$

$$q_k = \mathbb{P}(z = z_k), \quad p_k = \mathbb{P}(T = 1 | z = z_k)$$

System of Equations given $E[\varepsilon|z] = 0$

Let $m^*_{tk} = \mathbb{E}[\varepsilon| T^* = t, z = k]$

$$(1 - p_k)\bar{y}_{0k} = \alpha_1(p_k - \alpha_0)\left(\frac{\beta}{1 - \alpha_0 - \alpha_1}\right) + (1 - \alpha_0)c - (p_k - \alpha_0)(c + m^*_{1k})$$

$$p_k\bar{y}_{1k} = (1 - \alpha_1)(p_k - \alpha_0)\left(\frac{\beta}{1 - \alpha_0 - \alpha_1}\right) + \alpha_0 c + (p_k - \alpha_0)(c + m^*_{1k})$$

Theorem

$2K$ equations in $K + 4$ unknowns, but β is unidentified from conditional means of y regardless of how many values, K, the instrument takes on.

Intuition

Using $E[\varepsilon|z] = 0$ to eliminate m^*_{0k} from the system "entangles" the equations such that each pair only provides one restriction.

First Moment Condition

Assumptions

- $\mathbb{E}[\varepsilon|z] = 0$
- $\mathbb{E}[\varepsilon|T^*, T, z] = \mathbb{E}[\varepsilon|T^*, z]$
- $\alpha_0 = \mathbb{P}(T = 1|T^* = 0, z)$
- $\alpha_1 = \mathbb{P}(T = 0|T^* = 1, z)$

Moment Condition

$$\mathrm{Cov}(y, z) - \left(\frac{\beta}{1 - \alpha_0 - \alpha_1}\right)\mathrm{Cov}(T, z) = 0$$

MC # 1 identifies $\beta/(1 - \alpha_0 - \alpha_1)$

Second Moment Condition

Additional Assumptions

- $\mathbb{E}[\varepsilon^2 | z] = \mathbb{E}[\varepsilon^2]$
- $\mathbb{E}[\varepsilon^2 | T^*, T, z] = \mathbb{E}[\varepsilon^2 | T^*, z]$

Moment Condition

$$\mathrm{Cov}(y^2, z) - \frac{\beta}{1-\alpha_0-\alpha_1}\left\{2\mathrm{Cov}(yT, z) - \beta\mathrm{Cov}(T, z)\left(\frac{1+\alpha_0-\alpha_1}{1-\alpha_0-\alpha_1}\right)\right\} = 0$$

Given MC #1, MC #2 identifies $(\alpha_1 - \alpha_0)$

Third Moment Condition

Additional Assumptions

- $\mathbb{E}[\varepsilon^3|z] = \mathbb{E}[\varepsilon^3]$
- $\mathbb{E}[\varepsilon^3|T^*, T, z] = \mathbb{E}[\varepsilon^3|T^*, z]$

Moment Condition

$$\mathrm{Cov}(y^3, z) - \left(\frac{\beta}{1-\alpha_0-\alpha_1}\right)\left\{ \beta^2\left[1 + \frac{6\alpha_0(1-\alpha_1)}{(1-\alpha_0-\alpha_1)^2}\right]\mathrm{Cov}(T, z) \right.$$
$$\left. -3\beta\left[\frac{1-(\alpha_1-\alpha_0)}{1-\alpha_0-\alpha_1}\right]\mathrm{Cov}(yT, z) + 3\mathrm{Cov}(y^2T, z)\right\} = 0$$

Theorem

Model is identified if $\beta \neq 0$ and $\alpha_0 + \alpha_1 < 1$. If $\beta = 0$, reduced form identifies β. If $\alpha_0 + \alpha_1 > 1$, β is identified up to sign.

GMM Estimator in Simple Special Case: $\alpha_0 = 0$

$$u(\boldsymbol{\theta}) = y - c - \frac{\beta}{1-\alpha_1} T$$

$$v(\boldsymbol{\theta}) = y^2 - \sigma_{\varepsilon\varepsilon} - c^2 - \frac{\beta}{1-\alpha_1} 2yT + \frac{\beta^2}{1-\alpha_1}$$

$$\mathbb{E}\left[g_1(\mathbf{x}, \boldsymbol{\theta})\right] = \mathbb{E}\begin{bmatrix} u(\boldsymbol{\theta}) \\ v(\boldsymbol{\theta}) \end{bmatrix} = \mathbf{0}, \quad \mathbb{E}\left[g_2(\mathbf{x}, \boldsymbol{\theta})\right] = \mathbb{E}\begin{bmatrix} u(\boldsymbol{\theta})z \\ v(\boldsymbol{\theta})z \end{bmatrix} = \mathbf{0}$$

$$\beta = \frac{2\text{Cov}(yT, z)}{\text{Cov}(T, z)} - \frac{\text{Cov}(y^2, z)}{\text{Cov}(y, z)}$$

Simulation DGP: $y = \beta T^* + \varepsilon$

Errors

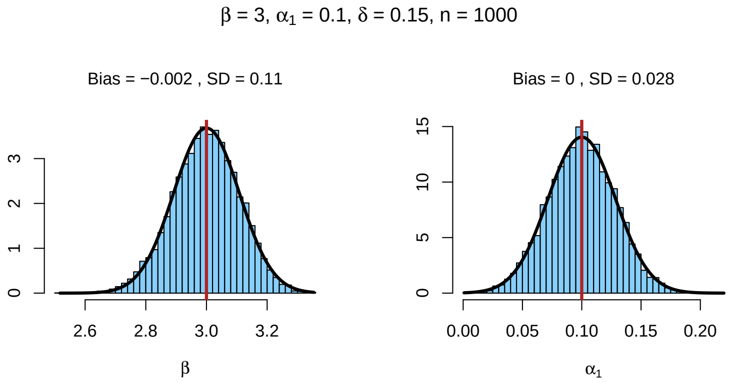
$(\varepsilon, \eta) \sim$ jointly normal, mean 0, variance 1, correlation 0.5.

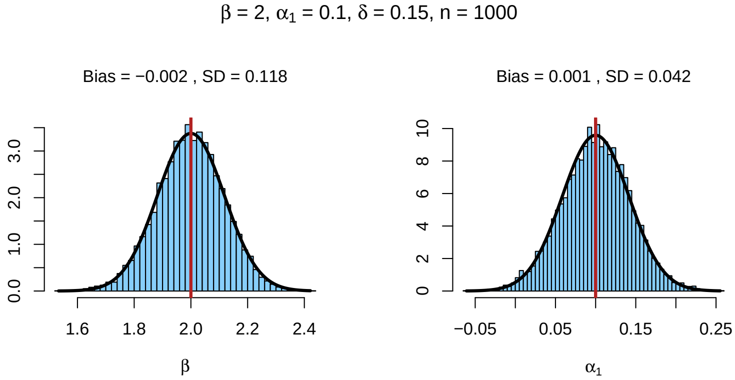
First-Stage

- Half of subjects have $z = 1$, the rest have $z = 0$.
- $T^* = \mathbf{1}\{\gamma_0 + \gamma_1 z + \eta > 0\}$
- $\delta = \mathbb{P}(T^* = 0 | z = 1) = \mathbb{P}(T^* = 1 | z = 0)$

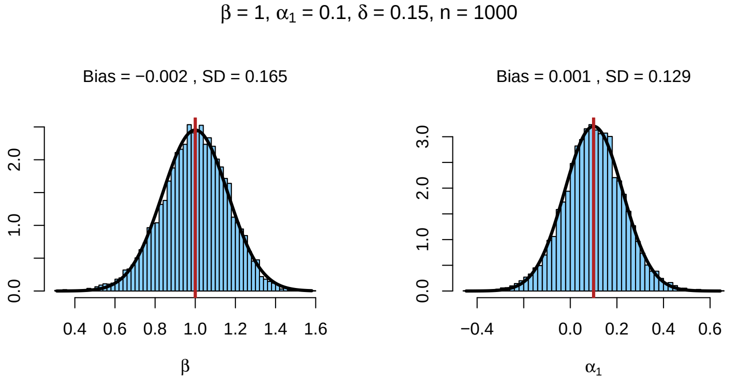
Mis-classification

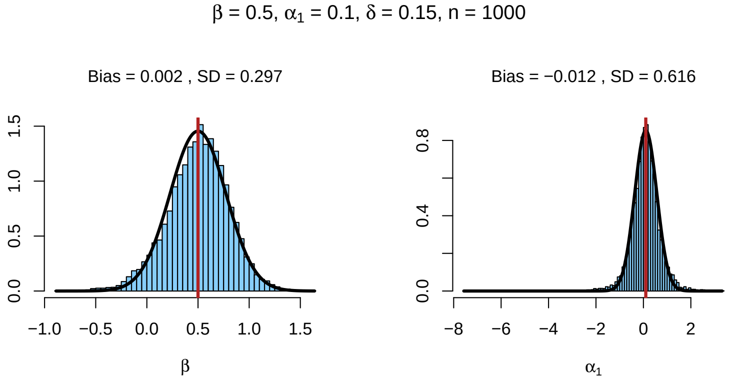
- Set $\alpha_0 = 0$
- $T | T^* = 1 \sim \text{Bernoulli}(1 - \alpha_1)$


β = 3, α1 = 0.1, δ = 0.15, n = 1000
Bias = −0.002 , SD = 0.11
Bias = 0 , SD = 0.028
0
1
2
3
2.6
2.8
3.0
3.2
β
0
5
10
15
0.00
0.05
0.10
0.15
0.20
α1



β = 2, α1 = 0.1, δ = 0.15, n = 1000
Bias = −0.002 , SD = 0.118
0.0
1.0
2.0
3.0
1.6
1.8
2.0
2.2
2.4
β
Bias = 0.001 , SD = 0.042
0
2
4
6
8
10
−0.05
0.05
0.15
0.25
α1


β = 1, α1 = 0.1, δ = 0.15, n = 1000
Bias = −0.002 , SD = 0.165
Bias = 0.001 , SD = 0.129
0.0
1.0
2.0
3.0
0.4
0.6
0.8
1.0
1.2
1.4
1.6
−0.4
0.0
0.2
0.4
0.6
β
α1



β = 0.5, α1 = 0.1, δ = 0.15, n = 1000
Bias = 0.002 , SD = 0.297
Bias = −0.012 , SD = 0.616
0.0
0.5
1.0
1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
β
0.0
0.4
0.8
−8
−6
−4
−2
0
2
α1

Identification Failure when $\beta = 0$

Simple Special Case: $\alpha_0 = 0$

$$u(\boldsymbol{\theta}) = y - c - \frac{\beta}{1-\alpha_1} T$$

$$v(\boldsymbol{\theta}) = y^2 - \sigma_{\varepsilon\varepsilon} - c^2 - \frac{\beta}{1-\alpha_1} 2yT + \frac{\beta^2}{1-\alpha_1}$$

$$\mathbb{E}\left[g_1(\mathbf{x}, \boldsymbol{\theta})\right] = \mathbb{E}\begin{bmatrix} u(\boldsymbol{\theta}) \\ v(\boldsymbol{\theta}) \end{bmatrix} = \mathbf{0}, \quad \mathbb{E}\left[g_2(\mathbf{x}, \boldsymbol{\theta})\right] = \mathbb{E}\begin{bmatrix} u(\boldsymbol{\theta})z \\ v(\boldsymbol{\theta})z \end{bmatrix} = \mathbf{0}$$

- β small $\Rightarrow$ moment equalities uninformative about α_1
- $(c, \sigma_{\varepsilon\varepsilon})$ are identified at any hypothesized pair (α_1, β)

Auxiliary Moment Inequalities

General Case $\alpha_0 \neq 0$

$$\alpha_0(z) = \alpha_0,\ \alpha_1(z) = \alpha_1$$

$$\implies \quad p_k^* = \frac{p_k - \alpha_0}{1 - \alpha_0 - \alpha_1}, \quad 1 - p_k^* = \frac{1 - p_k - \alpha_1}{1 - \alpha_0 - \alpha_1}$$

$$\alpha_0 + \alpha_1 < 1 \iff \text{Cor}(T, T^*) > 0 \iff (1 - \alpha_0 - \alpha_1) > 0$$

Implications

- $\alpha_0 < \min_k\{p_k\}, \quad \alpha_1 < \min_k\{1 - p_k\}$
- β is between β_{RF} and β_{IV}
- β_{IV} *inflated* but has correct sign

Even Tighter Bounds for α_0, α_1 from Conditional Variances

Assume

$\mathbb{E}[\varepsilon^2 | T^*, T, z] = \mathbb{E}[\varepsilon^2 | T^*, z]$

Observables

$\sigma_{tk}^2 = \text{Var}(y | T = t, z = k)$

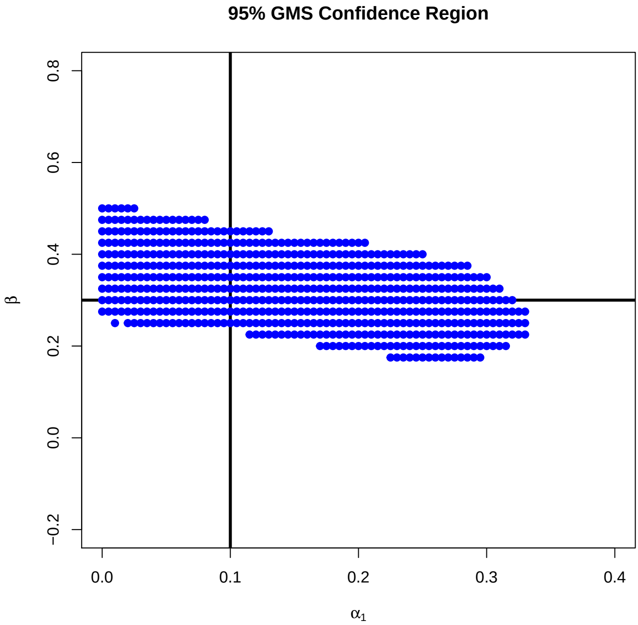
Constrain Unobservables

$s_{tk}^{*2} = Var(u | T^* = t, z_k) > 0$

$$(p_k - \alpha_0)\left[(1-\alpha_0)p_k\sigma_{1k}^2 - \alpha_0(1-p_k)\sigma_{0k}^2\right] > \alpha_0(1-\alpha_0)p_k(1-p_k)(\bar{y}_{1k} - \bar{y}_{0k})^2$$
$$(1 - p_k - \alpha_1)\left[(1-\alpha_1)(1-p_k)\sigma_{0k}^2 - \alpha_1 p_k\sigma_{1k}^2\right] > \alpha_1(1-\alpha_1)p_k(1-p_k)(\bar{y}_{1k} - \bar{y}_{0k})^2$$

Identification-Robust Joint Inference for $(\alpha_0, \alpha_1, \beta)$

- Auxiliary moment inequalities to bound (α_0, α_1)
- Joint CS for $(\alpha_0, \alpha_1, \beta)$ by inverting Anderson-Rubin Test
- Marginal inference for β by projection.
- Generalized Moment Selection (Andrews & Soares, 2010) for tighter confidence sets.
- Results are preliminary (not exploiting full set of inequalities) but this approach seems to work extremely well.



95% GMS Confidence Region
0.8
0.6
0.4
0.2
0.0
−0.2
β
0.0
0.1
0.2
0.3
0.4
α1

Conclusion

- Endogenous, mis-measured binary treatment.
- Important in applied work but no solution in the literature.
- Usual (1st moment) IV assumption fails to identify β
- Higher moment / independence restrictions identify β
- Identification-Robust Inference incorportating additional inequality moment conditions.