

# Estimating the Effect of a Mis-measured, Endogenous, Binary Treatment

Francis J. DiTraglia

Camilo Garcia-Jimeno

University of Pennsylvania

November 19th, 2015

## What is the causal effect of $T^*$ ?

$$y_i = h(T_i^*, \mathbf{x}_i) + \varepsilon_i$$

- ▶  $y$  – Outcome of interest
- ▶  $h$  – Unknown function that *does not depend on*  $i$
- ▶  $T^*$  – Unobserved, endogenous binary treatment
- ▶  $T$  – Observed, mis-measured surrogate for  $T^*$
- ▶  $\mathbf{x}$  – Exogenous covariates
- ▶  $\varepsilon$  – Mean-zero error term
- ▶  $z$  – Discrete instrumental variable

# Example 1: 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

## Example 2: Schooling and Test Scores

Burde & Linden (2013, AEJ Applied )

RCT in Afghanistan: 32 villages divided into 11 clusters. Randomly choose 6 and build a school in each village of these clusters.

- ▶  $y$  – Score on math and language test
- ▶  $T^*$  – True school attendance
- ▶  $T$  – Self-reported school attendance
- ▶  $\mathbf{x}$  – Household characteristics
- ▶  $z$  – School built in village

## Example 3: 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 attendance
- ▶  $T$  – Self-reported training attendance
- ▶  $\mathbf{x}$  – Individual characteristics
- ▶  $z$  – Offer of job training

## Non-classical Measurement Error: Binary $T^*$

- ▶ Many applications of linear model have *binary* treatment
- ▶ Binary  $T^* \implies \mathbb{E}[T^* w] \leq 0$
- ▶ Misclassification Probabilities:

$$\alpha_0 = \mathbb{P}(T = 1 | T^* = 0)$$

$$\alpha_1 = \mathbb{P}(T = 0 | T^* = 1)$$

- ▶ Non-Differential Measurement Error:  $T \perp (z, u) | T^*$
- ▶  $\sigma_{T^*}^2 \not\leq \sigma_T^2$  so work with  $\alpha_0, \alpha_1$  rather than  $\kappa$
- ▶ *Four-dimensional* Problem. . .

## Results for a Mis-classified Binary Regressor

Aigner (1973), Bollinger (1996)...

- ▶ Even if  $\rho_{T^*u} = 0$ , OLS is biased and inconsistent: typically attenuated towards zero *but could flip signs!*

Kane et al. (1999), Black et al. (2000), Frazis et al. (2003)...

- ▶  $\rho_{zu} = 0 \implies$  IV solves endogenous regressor problem if there is no mis-classification
- ▶  $\rho_{T^*u} = 0$  and  $\rho_{zu} = 0 \implies$  non-linear GMM estimator can solve the mis-classification problem

## OLS and IV Probability Limits: Binary $T^*$

$$\text{plim} \left( \hat{\beta}_{OLS} \right) = \frac{\sigma_{T^*}^2}{\sigma_T^2} \left[ \beta (1 - \alpha_0 - \alpha_1) + \frac{\sigma_{T^*u}}{\sigma_{T^*}^2} \right]$$

$$\text{plim} \left( \hat{\beta}_{IV} \right) = \frac{\beta}{1 - \alpha_0 - \alpha_1} + \frac{\sigma_{zu}}{\sigma_{zT}}$$

$$\sigma_{T^*}^2 = \frac{(p - \alpha_0)(1 - p - \alpha_1)}{(1 - \alpha_0 - \alpha_1)^2}$$

Where  $p = \mathbb{P}(T = 1)$



## What About Endogenous, Mis-measured $T^*$ , Valid $z$ ?

$$y = \beta T^* + u$$

$$u = c + \varepsilon$$

- ▶ No results in the literature for this case
- ▶ Important setting in applied work: e.g. RCTs
- ▶ Discrete Instrument:  $z \in \{z_1, \dots, z_K\}$
- ▶ Non-parametric First Stage:  $p_k^* = \mathbb{P}(T^* = 1 | z = z_k)$
- ▶ What does  $E[\varepsilon | z] = 0$  buy us in this case?

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

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

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

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

# Unobservable Moments: $y = \beta T^* + u$

	$z = 1$	$z = 1$	$\dots$	$z = K$
$T^* = 0$	$m_{01}^*$ $p_{01}^*$	$m_{02}^*$ $p_{02}^*$	$\dots$	$m_{0K}^*$ $p_{0K}^*$
$T^* = 1$	$m_{11}^*$ $p_{11}^*$	$m_{12}^*$ $p_{12}^*$	$\dots$	$m_{1K}^*$ $p_{1K}^*$

$$m_{tk}^* = \mathbb{E}[u | T^* = t, z = z_k], \quad p_{tk}^* = q_k p_k^*$$

$$p_k^* = \mathbb{P}(T^* = 1 | z = z_k) = \frac{p_k - \alpha_0}{1 - \alpha_0 - \alpha_1}$$

## Possible Assumptions On $m_{tk}^*$

Joint Exogeneity:  $\mathbb{E}[\varepsilon | T^*, z] = 0$

$$\implies m_{tk}^* = c \quad \text{for all } t, k$$

Exogenous Treatment:  $\mathbb{E}[\varepsilon | T^*] = 0$

$$\implies \frac{1}{\mathbb{P}(T^* = t)} \sum_k p_{tk}^* m_{tk}^* = c \quad \text{for all } t$$

Exogenous Instrument:  $\mathbb{E}[\varepsilon | z] = 0$

$$\implies (1 - p_k^*) m_{0k}^* + p_k^* m_{1k}^* = c \quad \text{for all } k$$

# Mahajan (2006, Econometrica)

## Regression Model

$$y = \mathbb{E}[y|T^*] + \nu$$

$$\mathbb{E}[\nu|T^*] = 0 \text{ by construction}$$

## Causal Model

$$y = c + \beta T^* + \varepsilon$$

$$\mathbb{E}[\varepsilon|T^*] \neq 0$$

## Main Result (Correct) – Exogenous Treatment

Relevant binary instrument  $z$  ( $p_1^* \neq p_2^*$ ) identifies  $\alpha_0, \alpha_1$  and  $\mathbb{E}[y|T^*]$  provided that  $\mathbb{E}[\nu|T^*, T, z] = 0$ .

## Mahajan (2006, Econometrica)

### Regression Model

$$y = \mathbb{E}[y|T^*] + \nu$$

$$\mathbb{E}[\nu|T^*] = 0 \text{ by construction}$$

### Causal Model

$$y = c + \beta T^* + \varepsilon$$

$$\mathbb{E}[\varepsilon|T^*] \neq 0$$

### Additional Result (Incorrect) – Endogenous Treatment

$$\mathbb{E}[\varepsilon|z] = 0, p_1^* \neq p_2^*, \mathbb{E}[\varepsilon|T, T^*, z] = \mathbb{E}[\varepsilon|T^*] \implies \beta \text{ identified.}$$

# Mahajan's Argument

## Regression Model

$$y = \mathbb{E}[y | T^*] + \nu$$

$$\mathbb{E}[\nu | T^*] = 0 \text{ by construction}$$

## Causal Model

$$y = c + \beta T^* + \varepsilon$$

$$\mathbb{E}[\varepsilon | T^*] \neq 0$$

## Ingredients

1. If  $p_1^* \neq p_2^*$ ,  $\mathbb{E}[\varepsilon | z] = 0$  then, since  $\beta_{IV} = \beta / (1 - \alpha_0 - \alpha_1)$ , knowledge of  $\alpha_0, \alpha_1$  is sufficient to recover  $\beta$ . (Correct)
2. If  $p_1^* \neq p_2^*$ ,  $\mathbb{E}[\nu | T^*, T, z] = 0$ ,  $\alpha_0, \alpha_1$  are identified. (Correct)

How to satisfy both 1 and 2 while allowing  $\mathbb{E}[\varepsilon | T^*] \neq 0$ ?

3. Assume that  $\mathbb{E}[\varepsilon | T^*, T, z] = \mathbb{E}[\varepsilon | T^*]$   
(i.e.  $m_{01}^* = m_{02}^*$  and  $m_{11}^* = m_{12}^*$ )

# Mahajan's Argument

## Proposition

If  $\mathbb{E}[\varepsilon | T^*] \neq 0$  then  $\mathbb{E}[\varepsilon | T^*, T, z] = \mathbb{E}[\varepsilon | T^*]$  combined with  $\mathbb{E}[\varepsilon | z] = 0$  implies  $p_1^* = p_2^*$ , i.e.  $z$  is irrelevant for  $T^*$ .

## Proof

Recall that  $\mathbb{E}[\varepsilon | z] = 0$  implies

$$(1 - p_1^*)m_{01}^* + p_1^*m_{11}^* = c$$

$$(1 - p_2^*)m_{02}^* + p_2^*m_{12}^* = c$$

while Mahajan's assumption implies  $m_{01}^* = m_{02}^*$  and  $m_{11}^* = m_{12}^*$ .

Therefore either  $m_{01}^* = m_{02}^* = m_{11}^* = m_{12}^* = c$ , which is ruled out by  $E[\varepsilon | T^*] = 0$ , or  $p_1^* = p_2^*$ .



## Moment Conditions Imposing $\mathbb{E}[\varepsilon|z] = 0$

One pair of equations for each  $k = 1, \dots, K$

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

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

where  $\hat{y}_{0k} = (1 - p_k)\bar{y}_{0k}$  and  $\hat{y}_{1k} = p_k\bar{y}_{1k}$

2K Equations in  $K + 4$  Unknowns

## Proposition: $\beta$ is Unidentified Regardless of $K$

### Proof Sketch

- (1) Show that  $\mathcal{W} = \beta/(1 - \alpha_0 - \alpha_1)$  is identified.
- (2) Show that  $\mathcal{Q} = c + \beta(1 - \alpha_0)/(1 - \alpha_0 - \alpha_1)$  is identified.
- (3) (1) + (2)  $\implies (\mathcal{Q}, \mathcal{W})$  are *fixed*
- (4) Use (3) to rewrite equations in terms of  $(\mathcal{Q}, \mathcal{W})$ .
- (5) Discover that there is only *one* equation per  $k$ ! Rearranging:

$$m_{1k}^* = \frac{\mathcal{W}(\hat{y}_{0k} - \alpha_1 \mathcal{Q}) - \beta(\mathcal{Q} - \beta - \mathcal{W}\alpha_1) + \mathcal{W}^2(1 - p_k)\alpha_1}{\mathcal{W}(1 - p_k - \alpha_1) - \beta}$$

## Special Case of Prev Proof: $\alpha_0 = 0$

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

$$\hat{y}_{1k} = p_k \beta + p_k m_{1k}^*$$

# Identification by Conditional Variances?

## New Assumption

Homoskedastic errors w.r.t. the *instrument*:  $E[\varepsilon^2|z] = E[\varepsilon^2]$

## Not Crazy!

Holds in an RCT or a *true* natural experiment.

## New Moment Conditions

For each pair  $(k, \ell)$

$$\begin{aligned}s_k^2 - s_\ell^2 &= \mathcal{W}^2 [p_k(1 - p_k) - p_\ell(1 - p_\ell) + (\alpha_0 - \alpha_1)(p_k - p_\ell)] \\ &\quad + 2\mathcal{W} [(p_k - \alpha_0)(m_{1k}^* - c) - (p_\ell - \alpha_0)(m_{1\ell}^* - c)]\end{aligned}$$

Where  $s_k^2 = \text{Var}(y|z = z_k)$ , and  $\mathcal{W}$  is the Wald IV estimator.

## *Proposition:* $(\alpha_0 - \alpha_1)$ is Identified

Define

$$\widetilde{\mathcal{W}}_{k\ell} = \frac{\mathbb{E}[yT|z_k] - \mathbb{E}[yT|z_\ell]}{p_k - p_\ell}$$

Show that:

$$\begin{aligned} (p_k - \alpha_0)(m_{1k}^* - c) - (p_\ell - \alpha_0)(m_{1\ell}^* - c) = \\ (p_k - p_\ell) \left[ \widetilde{\mathcal{W}}_{k\ell} - \mathbb{E}[y] - \mathcal{W} \{ (1 - p) + (\alpha_0 - \alpha_1) \} \right] \end{aligned}$$

Substituting and rearranging:

$$\alpha_0 - \alpha_1 = (2p - 1 - p_k - p_\ell) + \frac{2(\widetilde{\mathcal{W}}_{k\ell} - \mathbb{E}[y])}{\mathcal{W}} - \frac{s_k^2 - s_\ell^2}{(p_k - p_\ell)\mathcal{W}^2}$$

## What Good is $(\alpha_0 - \alpha_1)$ ?

- ▶ Test a necessary condition for *no mis-classification*:  $\alpha_0 = \alpha_1$
- ▶ Simple, tighter partial identification bounds for  $\beta$
- ▶ In some settings, one of the mis-classification probabilities is known to be zero  $\implies \beta$  point identified

# Identification from Third Moments

## Simulation Study

$$y = \beta T^* + \varepsilon$$

$$T^* = \mathbf{1}\{\gamma_0 + \gamma_1 z + \eta > 0\}$$

$$\gamma_0 = \Phi^{-1}(\delta), \gamma_1 = \Phi^{-1}(1 - \delta) - \Phi(\delta) \text{ so that } \delta$$

E.g. if  $\delta = 0.1$  then 10% of those *not* offered treatment get it anyway, and 10% of those offered treatment don't take it up.

If  $T^* = 0$  then  $T = 0$  (E.g. Birthweight and smoking)

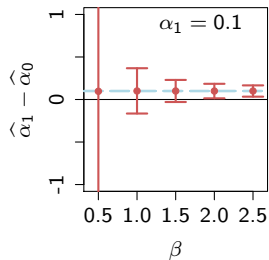
$$T|T^* = 1 \sim \text{Bernoulli}(?)$$

$$\begin{bmatrix} \varepsilon \\ \eta \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}\right)$$

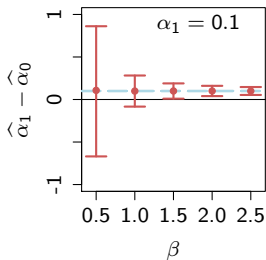


Sampling Distribution of  $\hat{\alpha}_1 - \hat{\alpha}_0$

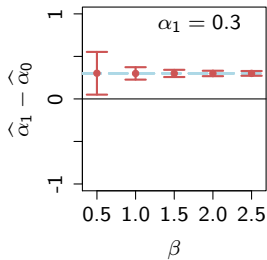
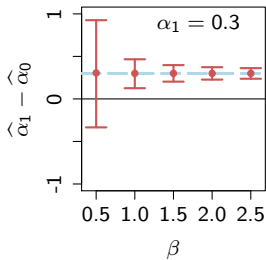
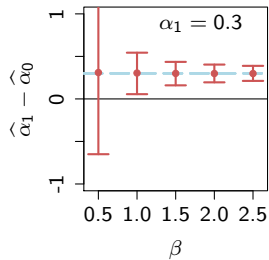
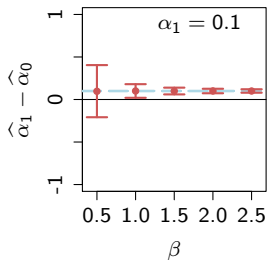
(a)  $N = 500, \delta = 0.1$



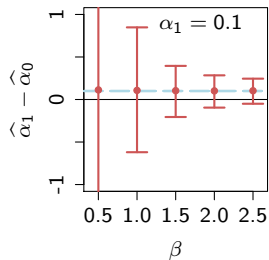
(b)  $N = 1000, \delta = 0.1$



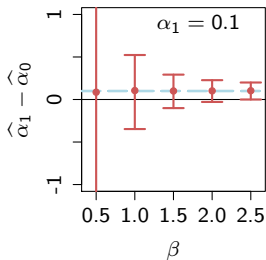
(c)  $N = 5000, \delta = 0.1$



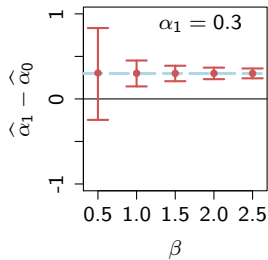
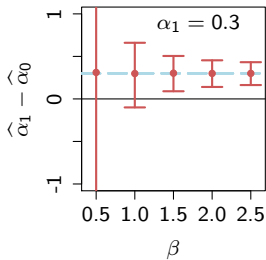
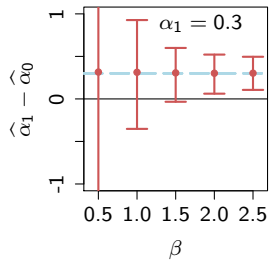
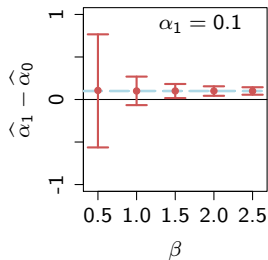
(a)  $N = 500, \delta = 0.3$



(b)  $N = 1000, \delta = 0.3$



(c)  $N = 5000, \delta = 0.3$

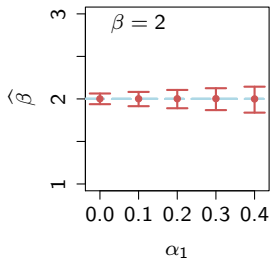
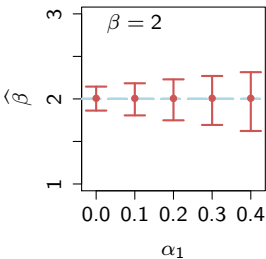
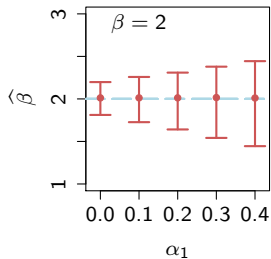
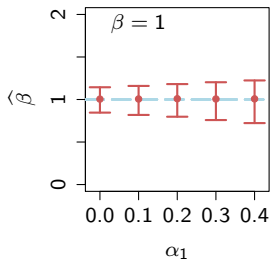
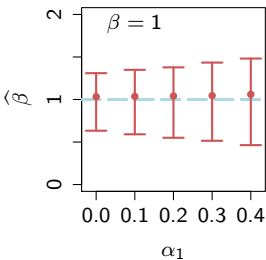
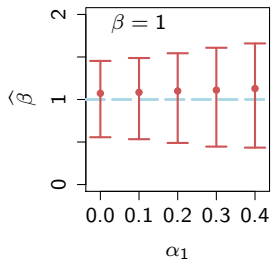


Sampling Distribution of  $\hat{\beta} = (1 - \hat{\alpha}_0 - \hat{\alpha}_1)\hat{\beta}_{IV}$

(a)  $N = 500, \delta = 0.1$

(b)  $N = 1000, \delta = 0.1$

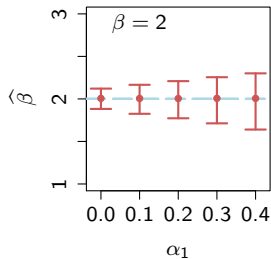
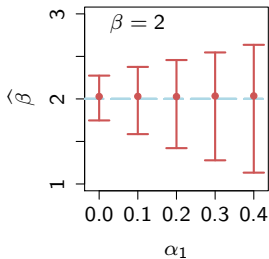
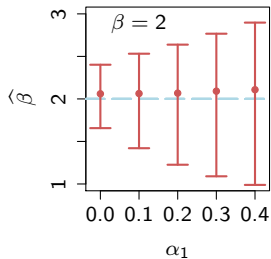
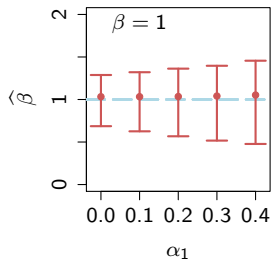
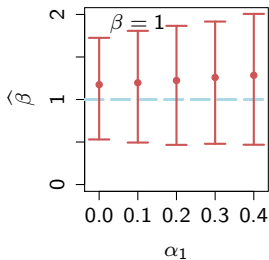
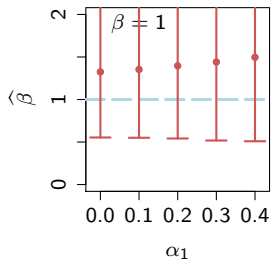
(c)  $N = 5000, \delta = 0.1$



(a)  $N = 500, \delta = 0.3$

(b)  $N = 1000, \delta = 0.3$

(c)  $N = 5000, \delta = 0.3$



# Empirical Illustration: Schooling and Test Scores