

# Generalized method of moments with incomplete data

Grigory Franguridi (USC CESR)

Hyungsik Roger Moon (USC Econ)

CIPHER 2026

# Outline

- **Incomplete data problem**
- **Examples and general framework**
- **Identification:** characterization of the identified set via a criterion function
- **Estimation**
- **Hypothesis testing and confidence regions** via bootstrap of Fang & Santos (2019)
- **Monte Carlo simulation:** simple linear regression with missing covariates

### When some data are missing, what can we do?

1. **ignore**: may lead to severe biases
2. **model the missingness**  
e.g., missing at random, selection on observables (reweighting), nonignorability, etc.  
For some models / assumptions, the parameter of interest can only be **bounded** (partial identification)
3. **leave missingness unrestricted** and derive the **worst-case bounds (this paper)**

## Literature

### **Bounds (partial identification) with incomplete (missing) data**

Rubin (1976), Manski (1989, 2005), **Horowitz & Manski (1995, 1998, 2000)**, Horowitz, Manski, Ponomareva, & Stoye (2003), **Lee (2009)**, Molinari (2020)

### **Panels with attrition** (our primary application)

Kish & Hess (1959), Hausman & Wise (1979), Moffit, Fitzgerald, & Gottschalk (1999), Hirano, Imbens, Ridder, & Rubin (2011), Wooldridge (2002), Bhattacharya (2008), Semykina & Wooldridge (2010), Deng, Hillygus, Reiter, Si, & Zheng (2013), Franguridi, Hahn, Honnhout, Kapteyn, & Ridder (2025, 2026), Franguridi & Liu (2025), Franguridi & Kosenkova (2026)

**This paper** provides a **full methodology** for

- **identification**
- **estimation** (bounds)
- **inference** (testing and confidence intervals/regions)

in a general class of models (GMM) under **incomplete data**

It can be considered an extension of Horowitz & Manski (1995) to GMM

## Example I (easy): proportion

**Goal:** estimate the **prevalence of dementia**

$$\theta = \mathbb{E}[Y_i],$$

where  $Y_i$  is a binary indicator of dementia for individual  $i$

**Data:** survey of dementia status with nonresponse:  $Y_i$  is **missing** if  $S_i = 0$

- $p = P(S_i = 1)$  (probability of responding to survey; observed)
- $\theta_1 = \mathbb{E}[Y_i | S_i = 1]$  (prevalence of dementia among respondents; observed)
- $\theta_0 = \mathbb{E}[Y_i | S_i = 0]$  (prevalence of dementia among nonrespondents; **unobserved**)

$$\text{Key identity: } \theta = p\theta_1 + (1 - p)\theta_0$$

Nonresponse is...	Restriction on $\theta_0$	Bounds on $\theta$
Completely random	$\theta_0 = \theta_1$	$\theta = \theta_1$ (point identification)
Monotone	$\theta_0 \geq \theta_1$	$\theta \in [\theta_1, p\theta_1 + (1 - p)]$ (“Lee bounds”)
<b>Unrestricted</b>	$\theta_0 \in [0, 1]$ (unrestricted)	$\theta \in [p\theta_1, p\theta_1 + (1 - p)]$ (“Horowitz-Manski bounds”)

**We focus on unrestricted nonresponse** (incompleteness, missingness)

The interval  $\Theta_I := [p\theta_1, p\theta_1 + (1 - p)]$  is called the **identified set** for  $\theta$

## Example II (harder): linear regression with one covariate

### Model:

$$Y_i = \theta_0 X_i + \varepsilon_i$$

$Y_i$  is always observed

$X_i$  is observed when  $S_i = 1$  and **missing** when  $S_i = 0$ ; denote  $p = P(S_i = 1)$

If there were no missing data, we would use the **moment condition (OLS)**

$$\mathbb{E}_\pi[\phi(X, Y, \theta_0)] := \mathbb{E}_\pi[X(Y - \theta_0 X)] = 0, \quad (1)$$

but the joint distribution  $\pi$  of  $(X, Y)$  is **not identified** due to missing data on  $X$

**Identified set for  $\theta_0$ :**

$$\Theta_I = \{\theta \in \mathbb{R} : \ell(\theta) \leq 0 \leq u(\theta)\}$$

where **lower/upper bounds on the moment (1)** are

$$\ell(\theta) = p \mathbb{E}[\phi(X, Y, \theta) | S = 1] + (1 - p) \mathbb{E}[\min_x \phi(x, Y, \theta) | S = 0]$$

$$u(\theta) = p \mathbb{E}[\phi(X, Y, \theta) | S = 1] + (1 - p) \mathbb{E}[\max_x \phi(x, Y, \theta) | S = 0]$$

## Example II (harder): linear regression with one covariate

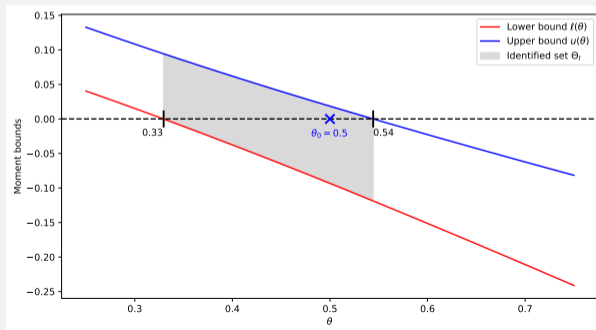
### Model:

$$Y_i = \theta_0 X_i + \varepsilon_i$$

$X_i$  is binary(0.5),  $\varepsilon_i$  is uniform $[-0.5, 0.5]$ ,  $\theta_0 = 0.5$

$Y_i$  is always observed

$X_i$  is **missing** at random (but the researcher does not know it) with probability  $P(S_i = 0) = 0.3$



## Example III (hard): multiple linear regression

### Model:

$$Y_i = \mathbf{X}_i' \boldsymbol{\theta}_0 + \varepsilon_i$$

$Y_i \in \mathbb{R}$  is always observed,  $\mathbf{X}_i \in \mathbb{R}^d$  is **missing** when  $S_i = 0$

If there were no missing data, we would use **moment condition (OLS)**

$$\mathbb{E}_{\pi}[\mathbf{X}(Y - \mathbf{X}'\boldsymbol{\theta}_0)] = \mathbf{0}_d \quad (2)$$

The problem is that the joint distribution  $\pi$  of  $(\mathbf{X}, Y)$  is **not identified** due to missing data on  $\mathbf{X}$

### Identified set for $\boldsymbol{\theta}$ :

$$\Theta_I = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \text{there exists a data-compatible distribution } \pi \text{ such that (2) holds} \right\}$$

Notation	Description	Observed (identified)?
$p = P(S = 1)$	Probability of complete observation	Yes
$f^0$	Distribution of $Y \mid S = 0$	Yes
$\pi^1$	Distribution of $(\mathbf{X}, Y) \mid S = 1$	Yes
$\pi^0$	Distribution of $(\mathbf{X}, Y) \mid S = 0$	<b>No</b> (but second marginal equals $f^0$ )
$\pi = p\pi^1 + (1 - p)\pi^0$	Distribution of $(\mathbf{X}, Y)$	<b>No</b> (but restricted)

### Example III (hard): multiple linear regression

The set of data-compatible distributions (“identified set of distributions”)  $\Pi$ :

$$\Pi = \{ \pi = p\pi^1 + (1-p)\pi^0 : \pi^0 \text{ has second marginal } f^0 \}$$

Moment function:

$$\phi(\mathbf{X}, Y, \boldsymbol{\theta}) := \mathbf{X}(Y - \mathbf{X}'\boldsymbol{\theta})$$

With this notation, the identified set for  $\boldsymbol{\theta}$  is

$$\Theta_I = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \text{there exists } \pi \in \Pi \text{ such that } \mathbb{E}_\pi [\phi(\mathbf{X}, Y, \boldsymbol{\theta})] = 0 \right\}$$

**This characterization is inconvenient:** how do we search over the set of distributions  $\pi$ ?

Instead, we will obtain a characterization of the form:

$$\Theta_I = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : Q(\boldsymbol{\theta}) = 0 \right\}$$

for a relatively simple function  $Q(\boldsymbol{\theta})$

## Example IV (hard): panel regression under attrition

Outcomes  $Y_{it}$ , covariates  $\mathbf{X}_{it}$ ; stack into  $\mathbf{Z}_{it} = (Y_{it}, \mathbf{X}_{it})$ ,  $t = 1, 2$

### Data:

Period 1:  $\mathbf{Z}_{i1}$  is observed for all units  $i = 1, \dots, n$

Period 2:  $\mathbf{Z}_{i2}$  is **only observed for units who stay in the sample** (notation:  $S_i = 1$ )

**Model:** linear regression with fixed effects

$$Y_{it} = \alpha_i + f_t + \mathbf{X}'_{it}\boldsymbol{\theta} + \varepsilon_{it}$$

Difference out  $\alpha_i$  and  $f_i$  to obtain the **moment condition for complete data** (dropping subscript  $i$ )

$$\mathbb{E}_{\pi}[\boldsymbol{\phi}(\mathbf{Z}_1, \mathbf{Z}_2, \boldsymbol{\theta})] := \mathbb{E}_{\pi}[(\mathbf{X}_2 - \mathbf{X}_1)(Y_2 - Y_1 - (\mathbf{X}_1 - \mathbf{X}_2)'\boldsymbol{\theta})] = \mathbf{0}$$

Again,  $\pi$  is the underidentified distribution of  $(\mathbf{Z}_1, \mathbf{Z}_2)$

## Example IV (hard): panel regression under attrition

How to handle attrition?

### 1. **with auxiliary data** in the form of a refreshment sample:

- **with “nonignorable” attrition**: model is **point identified**  
Hirano, Imbens, Ridder, & Rubin (2003)  
Hoonhout & Ridder (2019)  
Franguridi, Hahn, Hoonhout, Kapteyn, & Ridder (2025, 2026)  
Franguridi & Kosenkova (2026)
- **without restrictions on attrition**: model is **partially identified**  
Franguridi, Liu (2025)

### 2. **without auxiliary data**

- **with “monotone” attrition**: model is **partially identified**  
“Lee bounds”: Lee (2009), Semenova (2023)
- **without restrictions on attrition**: model is **partially identified**  
**this paper**

## General framework: GMM with incomplete data

Stack  $\mathbf{Z} = (\mathbf{X}, Y)$ , where  $Y$  is outcome,  $\mathbf{X}$  are covariates

**Missingness:** we can split  $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$  such that

- $\mathbf{Z}_1$  is always observed
- $\mathbf{Z}_2$  only observed if  $S = 1$
- **no restrictions on missingness**, i.e., on the relation between  $S$  and  $(\mathbf{Z}_1, \mathbf{Z}_2)$

**Data:** random sample from  $\mathbf{W} = (\mathbf{Z}_1, S\mathbf{Z}_2, S)$

**Model:**

$$\mathbb{E}_{\pi}[\phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta_0)] = 0,$$

where  $\phi$  is a known function and  $\pi$  is the (unobserved) distribution of  $(\mathbf{Z}_1, \mathbf{Z}_2)$

**All examples above are special cases**, e.g.,

- **cross-sectional regression with missing covariates:**  $\mathbf{Z}_1 = Y$  (outcome),  $\mathbf{Z}_2 = \mathbf{X}$  (covariates)
- **panel regression with attrition:**  $\mathbf{Z}_1 = (\mathbf{X}_1, Y_1)$  (period-1 variables),  $\mathbf{Z}_2 = (\mathbf{X}_2, Y_2)$  (period-2 variables)

**Goal:** bounds and confidence regions for  $\theta_0$

## Identification

The parameter value  $\theta_0$  is **consistent with the data** (i.e., in the identified set  $\Theta_I$ )



There exists  $\pi \in \Pi$  such that  $\mathbb{E}_\pi[\phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta_0)] = \mathbf{0}$



There exists  $\pi \in \Pi$  such that  $\mathbb{E}_\pi[\mathbf{u}'\phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta_0)] = 0$  for all  $\mathbf{u} \in \mathbb{B}$  (unit ball in  $\mathbb{R}^d$ )



$$\min_{\pi \in \Pi} \underbrace{\max_{\mathbf{u} \in \mathbb{B}} \mathbb{E}_\pi[\mathbf{u}'\phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta_0)]}_{\geq 0} = 0$$



$$Q(\theta_0) := \min_{\mathbf{u} \in \mathbb{B}} \underbrace{\max_{\pi \in \Pi} \mathbb{E}_\pi[\mathbf{u}'\phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta_0)]}_{\psi_{\theta_0}(\mathbf{u})} = 0 \quad (\text{characterization via a criterion function})$$

## Identification

The function  $\psi_{\theta}(\mathbf{u})$  **simplifies to a population mean**:

$$\begin{aligned}\psi_{\theta}(\mathbf{u}) &= \max_{\pi \in \Pi} \mathbb{E}_{\pi} [\mathbf{u}' \phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta)] \\ &= \max_{\pi \in \Pi} \mathbb{E}_{\pi} [S \mathbf{u}' \phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta) + (1 - S) \mathbf{u}' \phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta)] \\ &= \mathbb{E}[S \mathbf{u}' \phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta)] + \max_{\pi \in \Pi} \mathbb{E}_{\pi} [(1 - S) \mathbf{u}' \phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta)] \\ &= \mathbb{E} \left[ S \mathbf{u}' \phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta) + (1 - S) \max_{\mathbf{z}_2} \mathbf{u}' \phi(\mathbf{Z}_1, \mathbf{z}_2, \theta) \right]\end{aligned}$$

Hence **the identified set is**

$$\Theta_I = \{\theta \in \mathbb{R}^d : Q(\theta) = 0\}$$

where the criterion function  $Q(\theta)$  is the min of a **population mean**:

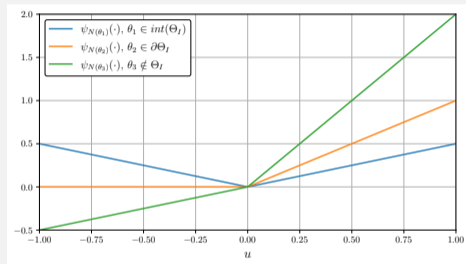
$$Q(\theta) = \min_{\mathbf{u} \in \mathbb{B}} \mathbb{E} \left[ S \mathbf{u}' \phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta) + (1 - S) \max_{\mathbf{z}_2} \mathbf{u}' \phi(\mathbf{Z}_1, \mathbf{z}_2, \theta) \right]$$

$Q(\theta)$  is also the **(negative) distance between the model (the parameter value  $\theta$ ) and the data**

## Identification with one moment condition

Suppose there is only **one moment condition** ( $\phi$  is a scalar)

Then  $\mathbf{u}$  is a scalar and the unit ball  $\mathbb{B} = [-1, 1]$



**Blue line:**  $\theta$  is **inside** the identified set  $\Theta_I$

**Orange line:**  $\theta$  is **on the boundary** of the identified set  $\Theta_I$

**Green line:**  $\theta$  is **outside** of the identified set  $\Theta_I$

## Estimation

Recall the identified set

$$\Theta_I = \{\boldsymbol{\theta} \in \mathbb{R}^d : Q(\boldsymbol{\theta}) = 0\}$$

where

$$Q(\boldsymbol{\theta}) = \min_{\mathbf{u} \in \mathbb{B}} \mathbb{E} \left[ S \mathbf{u}' \phi(\mathbf{Z}_1, \mathbf{Z}_2, \boldsymbol{\theta}) + (1 - S) \max_{\mathbf{z}_2} \mathbf{u}' \phi(\mathbf{Z}_1, \mathbf{z}_2, \boldsymbol{\theta}) \right]$$

Then a **natural estimator** is

$$\hat{\Theta}_I = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \left| \hat{Q}(\boldsymbol{\theta}) \right| \leq \eta_n \right\}$$

where  $\eta_n \downarrow 0$  is a **tuning parameter** and

$$\hat{Q}(\boldsymbol{\theta}) = \min_{\mathbf{u} \in \mathbb{B}} \frac{1}{n} \sum_{i=1}^n \left[ S_i \mathbf{u}' \phi(\mathbf{Z}_{1i}, \mathbf{Z}_{2i}, \boldsymbol{\theta}) + (1 - S) \max_{\mathbf{z}_2} \mathbf{u}' \phi(\mathbf{Z}_{1i}, \mathbf{z}_2, \boldsymbol{\theta}) \right]$$

The minimization over  $u \in \mathbb{B}$  is a **constrained convex program**  $\Rightarrow$  projected gradient descent

## Inference (testing and confidence regions)

**Goal:** test the hypothesis  $H_0 : \theta = \theta_0$

The collection of  $\theta_0$  not rejected by the test yields a **confidence interval/region** for  $\theta$

**Test statistic:**  $\hat{T}(\theta_0) := \sqrt{n}\hat{Q}(\theta_0)$  (the scaled negative distance between the model and the data)

**Decision rule:** reject when  $|\hat{T}(\theta_0)| > c_\alpha$  for an appropriately chosen critical value  $c_\alpha$

**How to pick  $c_\alpha$ ?**

- asymptotic distribution of  $\hat{Q}(\theta_0)$  **is not available in closed form and is hard to simulate**
- **standard (nonparametric) bootstrap is invalid** when  $\theta_0$  is on the boundary of the identified set
- **solution: bootstrap for directionally differentiable functionals of Fang & Santos (2019)**

## Inference: algorithm

**Testing**  $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$  at the nominal size  $\alpha \in (0, 1)$

(**confidence region** for  $\boldsymbol{\theta}$  is the collection of all  $\boldsymbol{\theta}_0$  not rejected by the test)

Pick: (i) **tuning parameter**  $\varepsilon_n \downarrow 0$  such that  $\sqrt{n}\varepsilon_n \rightarrow \infty$  (ii) number of bootstrap samples  $B$

1. Compute the test statistic  $\hat{T}(\boldsymbol{\theta}_0) = \sqrt{n} \min_{\mathbf{u} \in \mathbb{B}} \hat{\psi}(\mathbf{u})$ .
2. Compute the  $\varepsilon_n$ -enlarged argmin set  $\hat{\mathbf{U}}_n = \left\{ \mathbf{u} \in \mathbb{B} : \hat{\psi}(\mathbf{u}) \leq \min_{\mathbf{v} \in \mathbb{B}} \hat{\psi}(\mathbf{v}) + \varepsilon_n \right\}$ .
3. For each  $b = 1, \dots, B$ :
  - 3.1 Draw a random sample  $\mathbf{W}_i^b, i = 1, \dots, n$ , with replacement from the data  $\mathbf{W}_i = (S_i, \mathbf{Z}_{1i}, S_i \mathbf{Z}_{2i}), i = 1, \dots, n$ .
  - 3.2 Compute  $\hat{\psi}^{*b}(\mathbf{u})$  on the sample  $\mathbf{W}_i^b, i = 1, \dots, n$ .
  - 3.3 Compute

$$\hat{T}^{*b} = \min_{\mathbf{u} \in \hat{\mathbf{U}}_n} \sqrt{n} \left( \hat{\psi}^{*b}(\mathbf{u}) - \hat{\psi}(\mathbf{u}) \right).$$

4. Denote by  $\hat{c}_\alpha^*$  the  $(1 - \alpha)$ -quantile of  $|\hat{T}^{*1}|, \dots, |\hat{T}^{*B}|$ .
5. Reject  $H_0$  if  $|\hat{T}(\boldsymbol{\theta}_0)| > \hat{c}_\alpha^*$ .

## Inference on a subvector

Suppose **we are interested only in a scalar component**  $\theta_1$  of  $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2)$  and  $\Theta = \Theta_1 \times \Theta_2$  (e.g., a coefficient  $\theta_1$  on the treatment variable in a TWFE linear regression)

**The identified set** for  $\theta_1$  is

$$\Theta_{I,1} = \left\{ \theta_1 \in \Theta_1 : \max_{\boldsymbol{\theta}_2 \in \Theta_2} Q(\theta_1, \boldsymbol{\theta}_2) = 0 \right\},$$

cf. Romano & Shaikh (2008) and Bugni, Canay, & Shi (2017).

For the null  $H_0 : \theta_1 = \theta_{1,0}$ , this suggests the **“profiled” test statistic**

$$\hat{Q}_1(\theta_{1,0}) = \max_{\boldsymbol{\theta}_2 \in \Theta_2} \hat{Q}(\theta_{1,0}, \boldsymbol{\theta}_2)$$

Our theory for the full-vector case can be adapted to the subvector case:

- **estimation:**  $\hat{\Theta}_{I,1} = \left\{ \theta_1 \in \Theta_1 : \left| \hat{Q}_1(\theta_1) \right| \leq \eta_n \right\}$
- **inference** via bootstrap of Fang & Santos (2019)

## Monte Carlo simulation

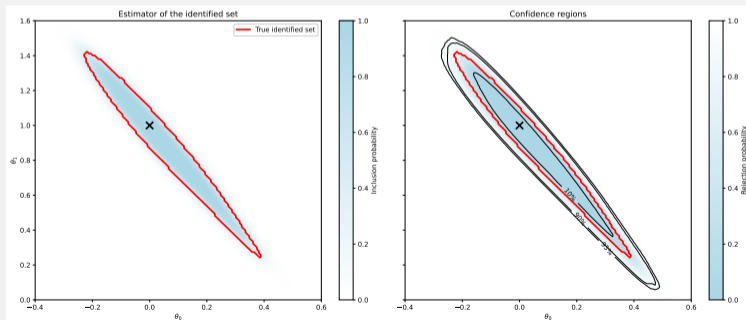
**Model:** simple cross-sectional linear regression

$$Y_i = \theta_0 + \theta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where some observations of  $X_i$  are **missing**,  $X_i \sim U[0, 1]$ ,  $\varepsilon \sim U[-1, 1]$ ,  $\varepsilon_i \perp X_i$

Selection is completely random with probability  $p = P(S_i = 1) = 0.9$

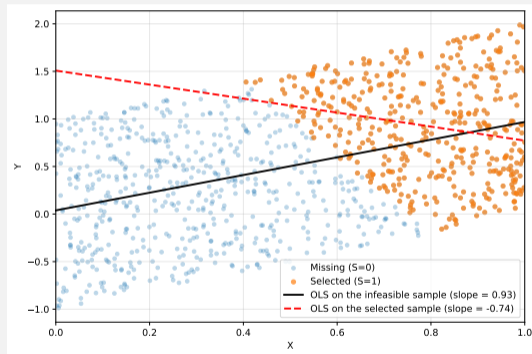
Sample size  $n = 1000$  and the tuning parameters  $\eta_n = \varepsilon_n = 0.1 \log n / \sqrt{n} \approx 0.022$



## Monte Carlo simulation

Now suppose selection is  $S_i = 1(\underline{s} < 3X_i + Y_i < \bar{s})$  for some  $\underline{s}, \bar{s}$

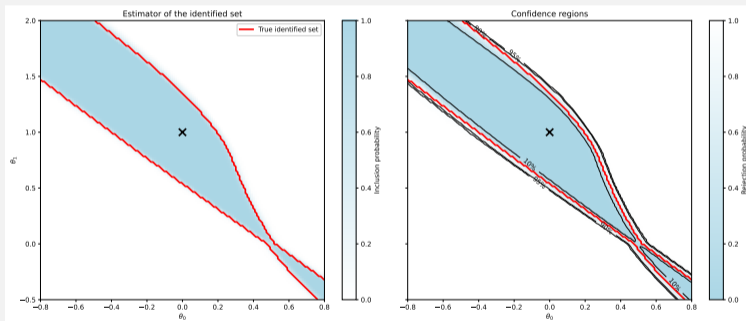
“**Berkson’s paradox**”: selection can lead to swap in the sign of the slope coefficient



## Monte Carlo simulation

Now suppose selection is  $S_i = 1(\underline{s} < 3X_i + Y_i < \bar{s})$  for some  $\underline{s}, \bar{s}$

“**Berkson’s paradox**”: selection can lead to swap in the sign of the slope coefficient



## Conclusion

### Generalized method of moments with incomplete (missing) data:

$$\mathbb{E}_{\pi}[\phi(\mathbf{Z}_1, \mathbf{Z}_2, \theta_0)] = \mathbf{0},$$

where  $\pi$  is the distribution of  $(\mathbf{Z}_1, \mathbf{Z}_2)$  and some observations of  $\mathbf{Z}_2$  may be missing

### This paper:

- characterizes the identified set via a simple criterion function
- suggests a Hausdorff-consistent estimator
- shows how to conduct tests / construct confidence regions via the bootstrap of Fang & Santos (2019)
- demonstrates that the procedure performs well in simulations

### To do:

- formal results on subvector inference
- empirical application