

Learning-Enabled Estimation:

Tight Characterizations under Sample Selection Biases



Vikram Kher
Yale



Jane H. Lee
Yale



Anay Mehrotra
Stanford



Manolis Zampetakis
Yale

Roadmap

1. Motivation
2. Preliminaries
3. Main results
4. Implications for identification
5. Final takeaways



<https://xkcd.com/2618/>

Sample selection bias is ubiquitous

Data are often filtered before we see them

In many empirical settings, we observe outcomes only after individuals or measurements pass through a selection mechanism.

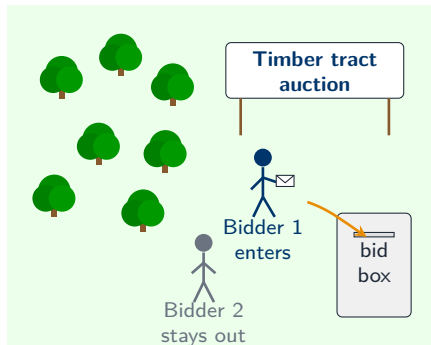
- ▶ **Labor markets and wages.** Wages are only observed for people who participate in the labor market. [Roy, 1951; Heckman, 1979]
- ▶ **Clinical studies.** Outcomes may be missing because of patient dropout, noncompliance, or censoring. [Little and Rubin, 2019]
- ▶ **Auctions and platforms.** Bids are only available from agents who choose to participate in the auction. [Samuelson, 1985; Levin and Smith, 1994; Gentry and Li, 2014]
- ▶ **Physical-science measurement pipelines.** Detection thresholds and quality filters mean that objects are recorded only if signals are sufficiently detectable. [Teerikorpi, 1997; Palarea-Albaladejo and Martín-Fernández, 2013; Webb-Robertson et al., 2015]

Selection bias in bidding: endogenous entry

Why selection arises

In many auctions, participation is *endogenous*: potential bidders decide whether to incur the cost of preparing and submitting a bid.

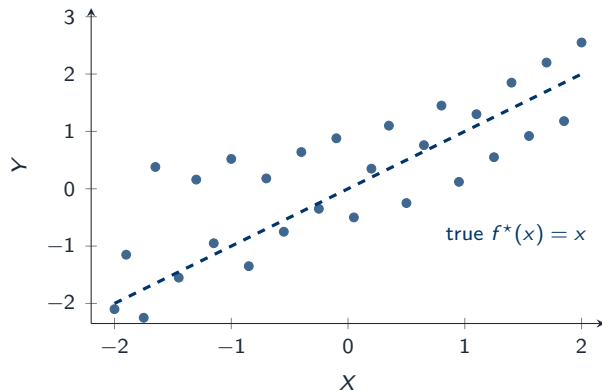
- ▶ **Entry costs matter.** Gathering information, estimating costs, and preparing a bid can be costly.
- ▶ **Observed bids come from a selected subset.** We only see bids from firms whose expected surplus exceeds their entry cost.



Examples

Timber auctions (Athey, Levin, Seira [2011]), online auctions (Bajari and Hortacsu [2003]), among others.

Effects of sample selection



No selection

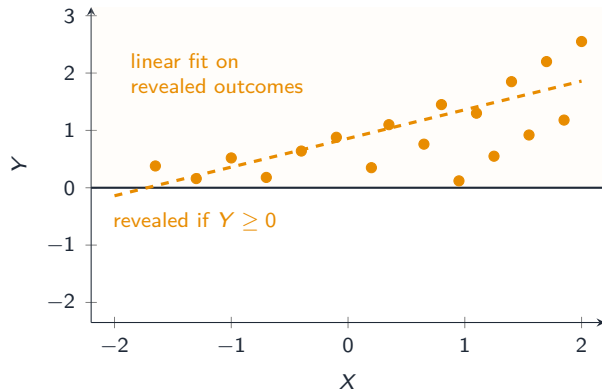
The economist sees uncensored samples from

$$Y = X + \xi, \quad X \sim \text{Unif}[-2, 2],$$

$$\xi \sim \mathcal{N}(0, 1), \quad D = 1.$$

With no sample selection, standard regression techniques work.

Effects of sample selection



With Selection

The population is still

$$X \sim \text{Unif}[-2, 2], \quad Y = X + \xi,$$

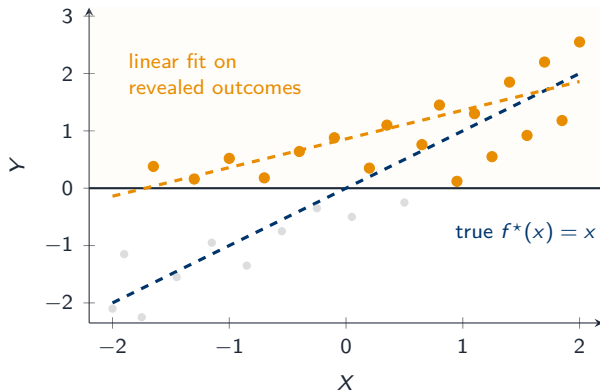
$$\xi \sim \mathcal{N}(0, 1),$$

but the outcome is revealed only when

$$D = \mathbf{1}\{Y \geq 0\}.$$

The economist only sees the orange points, so naive regression does not target f^* .

The problem: selection changes downstream estimation



With Selection

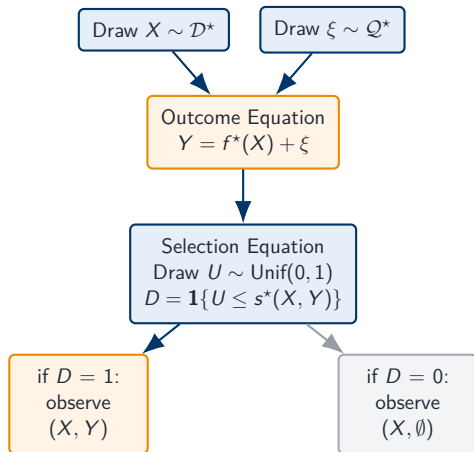
Consider the regression setting below:

$$Y = X + \xi, \quad X \sim \text{Unif}[-2, 2],$$

$$\xi \sim \mathcal{N}(0, 1), \quad D = \mathbf{1}\{Y \geq 0\}.$$

Standard OLS will not recover the underlying parameter.

Preliminaries: sample-selection model



Concept classes

An instance of our sample selection problem is described by a tuple $(f^*, s^*, \mathcal{D}^*, \mathcal{Q}^*)$, where $f^* \in \mathbb{F}$, $s^* \in \mathbb{S}$, $\mathcal{D}^* \in \mathbb{D}$, and $\mathcal{Q}^* \in \mathbb{Q}$.

- ▶ $\mathbb{F}$: regression functions.
- ▶ $\mathbb{S}$: selection filters.
- ▶ $\mathbb{D}, \mathbb{Q}$: covariate and noise distributions.

Goal: find a minimal set of assumptions under which f^* is recoverable by only observing the censored dataset

Prior classic models and results

Model	Main structural assumptions
Heckman [1979]	Linear outcome $Y = x^\top \beta + \xi$ and latent selection $D = \mathbf{1}\{z^\top \gamma + \xi' > 0\}$; (ξ, ξ') jointly normal.
Lee [1982, 1983]	Extends Heckman's model by allowing more flexible selection rules using a known transformation or distribution.
Ahn and Powell [1993]	Linear outcome; nonparametric selection mechanism; selection bias summarized by propensity score; requires rank/nonsingularity conditions.
Das et al. [2003]	Nonparametric f and selection; requires index restriction, positivity, rank, and normalization conditions.

For surveys see Vella [1998] and Puhani [2000].

Prior work gives sufficient conditions for identification. But, we will show these conditions can be overly restrictive.

Main Results

What does the censored data reveal?

Every record contains X

The population covariate distribution $\mathcal{D}^*$ is observable.

When $D = 1$:

We observe (X, Y) and the corresponding density of the event $\{D = 1, Y \in dy\}$ is

$$k^*(x, y) := s^*(x, y) \mathcal{Q}^*(y - f^*(x)).$$

When $D = 0$:

We observe $(X, \emptyset)$ and it holds that

$$\mathbb{P}(D = 0 \mid X = x) = 1 - \int k^*(x, t) dt.$$

The observable kernel

The censored data reveal the regression function, noise density, and selection rule only through the observable product kernel $k^*(x, y)$.

First main result: exact identification criterion

Condition 1: identification

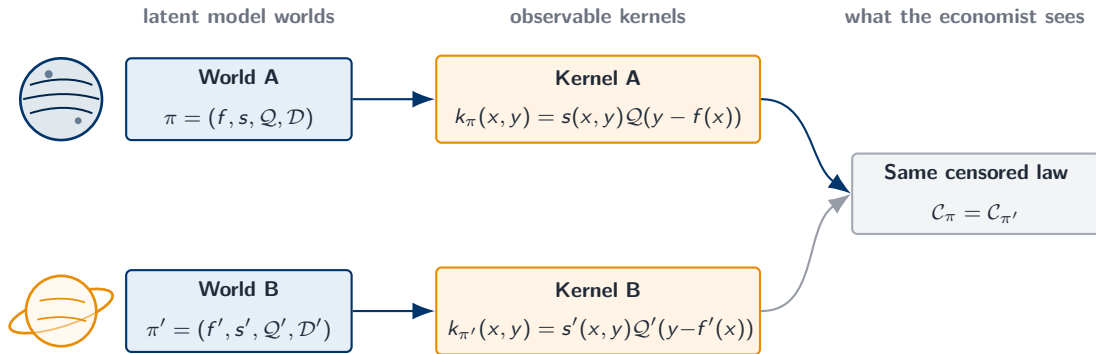
For every **distinct** $f, f' \in \mathbb{F}$, every $s, s' \in \mathbb{S}$, every $Q, Q' \in \mathbb{Q}$, and every $\mathcal{D} \in \mathbb{D}$, there is some $(x, y) \in \text{supp}(\mathcal{D}) \times \mathbb{R}$ such that

$$s(x, y) Q(y - f(x)) \neq s'(x, y) Q'(y - f'(x)).$$

Theorem 1

Assuming densities exist, every $f^* \in \mathbb{F}$ is identifiable from censored data *if and only if* Condition 1 holds.

Failure of condition 1 $\implies f^*$ is not identifiable



Suppose $f \neq f'$. Since X is always observed, $\mathcal{D}$ is known. Thus, we may assume that $\mathcal{D} = \mathcal{D}'$. The remaining question is whether two different regressors can produce the same observable kernel.

From identification to estimation

Identification goal

Find minimal assumptions on $(\mathbb{F}, \mathbb{S}, \mathbb{D}, \mathbb{Q})$ such that there exists a deterministic map

$$\psi : \{\text{realizable censored laws}\} \rightarrow \mathbb{F}$$

satisfying $\psi(\mathcal{C}_{f,s,\mathcal{D},\mathcal{Q}}) = f$ for every $(f, s, \mathcal{D}, \mathcal{Q}) \in \mathbb{F} \times \mathbb{S} \times \mathbb{D} \times \mathbb{Q}$.

Estimation goal

Find minimal assumptions on $(\mathbb{F}, \mathbb{S}, \mathbb{D}, \mathbb{Q})$ such that, given n i.i.d. samples $S \sim \mathcal{C}_{f,s,\mathcal{D},\mathcal{Q}}^n$, there is an algorithm outputting $f_S \in \mathbb{F}$ with

$$\mathbb{E}_{X \sim \mathcal{D}}[|f_S(X) - f(X)|] \leq \varepsilon(n), \quad \lim_{n \rightarrow \infty} \varepsilon(n) = 0.$$

Second main result: estimation condition

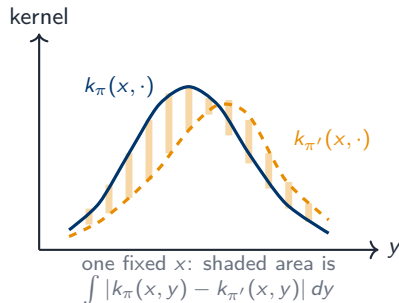
Two distance definitions

Regressor separation:

$$\Delta_{\mathcal{D}}(f, f') := \mathbb{E}_{X \sim \mathcal{D}} [|f(X) - f'(X)|].$$

Observable-kernel separation:

$$d_{\mathcal{D}}(\pi, \pi') := \iint |k_{\pi}(x, y) - k_{\pi'}(x, y)| d\mathcal{D}(x) dy.$$

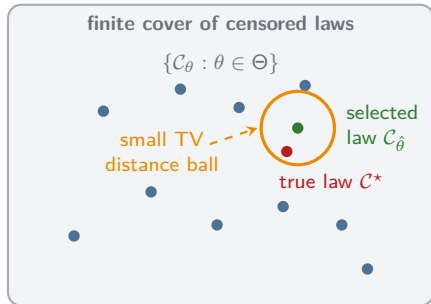


Condition 2: estimation

The concept classes satisfy Condition 2 with rate $R(\epsilon)$ if, for every $\mathcal{D} \in \mathbb{D}$, every $\pi = (f, s, \mathcal{Q})$, $\pi' = (f', s', \mathcal{Q}')$, and every $\epsilon \geq 0$,

$$\Delta_{\mathcal{D}}(f, f') \geq \epsilon \implies d_{\mathcal{D}}(\pi, \pi') \geq R(\epsilon).$$

High-level estimator outline



Proof chain

1. **Construct a finite cover.** Build $\Theta \subseteq \mathbb{F} \times \mathbb{S} \times \mathbb{D} \times \mathbb{Q}$ so that some $\theta \in \Theta$ induces a censored law close to C^* .
2. **Use Yatracos [1985] density selection.** From samples, select $\hat{\theta} \in \Theta$ such that

$$d_{\text{TV}}(C_{\hat{\theta}}, C^*) \leq 3 \min_{\theta \in \Theta} d_{\text{TV}}(C_\theta, C^*) + O\left(\sqrt{\frac{\log(|\Theta|/\delta)}{n}}\right).$$

3. **Convert TV error to kernel discrepancy.** Let

$$\hat{\pi} = (\hat{f}, \hat{s}, \hat{Q}), \quad \pi^* = (f^*, s^*, Q^*).$$

Then, we show $d_{\mathcal{D}^*}(\hat{\pi}, \pi^*) \leq 4 d_{\text{TV}}(C_{\hat{\theta}}, C^*)$.

4. **Apply Condition 2.** If the right-hand side is at most $R(\varepsilon)$, then

$$\mathbb{E}_{X \sim \mathcal{D}^*} \left[|\hat{f}(X) - f^*(X)| \right] \leq \varepsilon.$$

Finite-sample guarantee

Theorem 2 (informal)

Assume Condition 2 with rate R , regularity/covering assumptions, and an overlap assumption of $\mathbb{E}[s^*(X, Y)] \geq \eta > 0$. Then, with probability at least $1 - \delta$, the estimator outputs $\hat{f}$ with

$$\mathbb{E}_{X \sim \mathcal{D}^*} \left| \hat{f}(X) - f^*(X) \right| \leq \varepsilon$$

using

$$n = O\left(\frac{1}{R(\varepsilon)^2} \cdot \log \frac{M(\eta R(\varepsilon))}{\delta}\right)$$

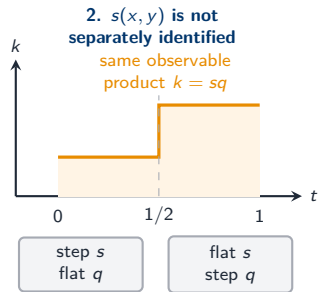
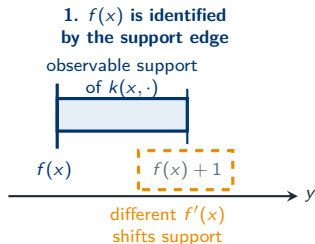
samples, where $M(\cdot)$ is a product of covering numbers.

Implications for Identification

f^* identifiable even when s^* is not

Setup

Consider the usual sample selection model with one-sided noise: $\text{supp}(q) \subseteq [0, 1]$, $q(t) > 0$ on $(0, 1)$, and $s(x, y) \geq \eta > 0$. Fix x and write $t = y - f(x)$. The observed kernel is $k(x, y) = s(x, y)q(y - f(x))$.



Why this matters: many existing approaches (e.g., double ML) typically first learn a selection rule/correction, then estimate f^* . Here, s^* is not identifiable at all, but f^* still is.

Identification with outcome-dependent survival sets

Setup

Let $X \in \{x_0, x_1\}$ and

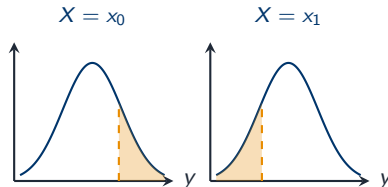
$$Y = f^*(X) + \xi, \quad \xi \sim \mathcal{N}(0, 1).$$

For the illustration, normalize

$$f^*(x_0) = f^*(x_1) = 0.$$

Fix a threshold $a > 0$. Selection depends on which part of the outcome distribution survives:

$$D = 1 \iff \begin{cases} Y \geq a, & X = x_0, \\ Y \leq -a, & X = x_1. \end{cases}$$



Why this matters:

- ▶ Outcome-dependent filters arise commonly in physical-science measurement pipelines.
- ▶ Acceptance region determined by detection thresholds, trigger rules, saturation windows.
- ▶ Not covered by Das et al. [2003] as the selection bias is not a function of the propensity score.

Identification with bounded-friction participation costs

Setup

Let

$$Y = f^*(X) + \xi, \quad \xi \sim \mathcal{N}(0, 1).$$

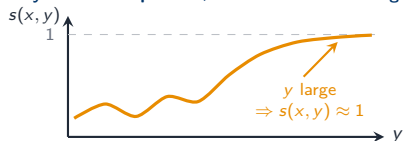
A bidder enters with probability

$$s(x, y) = G(\rho(x)y - h(x, y)),$$

where $\rho(x) \geq \rho_L > 0$, $|h(x, y)| \leq C$, and $G(t) \rightarrow 1$ as $t \rightarrow \infty$.

- ▶ Y : latent value.
- ▶ $\rho(x)$: positive responsiveness to value.
- ▶ $h(x, y)$: bounded participation costs.

Entry can be complicated, but saturates in the right tail



Why this matters:

- ▶ Entry costs and participation frictions occur commonly in procurement, natural resource, and platform auctions.
- ▶ These frictions may vary with bidder characteristics, and latent value.
- ▶ Not covered by Das et al. [2003] scalar-index conditions or Gentry and Li [2014] monotonicity conditions.

Takeaways and open directions

- ▶ **Identification:** We characterize exactly when f^* is identifiable from censored data under a very general model.
- ▶ **Estimation:** A quantitative strengthening of the identification condition gives sample-efficient estimation.
- ▶ **Implications:**
 - Unified identification proofs for classic models.
 - f^* can be identifiable even when s^* is not.
 - Identifiability in new economically meaningful settings.
- ▶ **Open directions:**
 - Efficient density-selection oracles for rich $\mathbb{F}, \mathbb{S}, \mathbb{D}, \mathbb{Q}$ (e.g., diffusion models)
 - Efficient algorithms for structured selectors in auctions and labor-market models.
 - Heteroskedastic extensions to the sample selection model.
 - Explore empirical performance of estimation method.