Partial Identification

Francis DiTraglia

Oxford Economics Summer School

Roadmap for this Lecture

The Limits of Causal Inference

- (Y_0, Y_1) never observed for the same person; can't learn their joint distribution.
- Quantities like $\text{Var}(Y_1 - Y_0)$ or $\mathbb{P}(Y_1 - Y_0 > 0)$ are **not identifiable.**

Partial Identification

- Even if we can't pin θ down *exactly*, we may be able to **rule out** many values.

Outline

1. Simplest example of partial identification.
2. Bounds on ATE while allowing for selection bias.
3. Bound the distribution of treatment effects.

Simple Example: Reverse Regression Bounds

Population Linear Regression

- α and β are intercept and slope from population linear regression of Y on X
- Thus we can write $Y = \alpha + \beta X + U$ where we define

$$\beta \equiv \frac{\text{Cov}(X, Y)}{\text{Var}(X)}, \quad \alpha \equiv \mathbb{E}[Y] - \beta \mathbb{E}[X], \quad U \equiv Y - \alpha - \beta X$$

- By **construction** we have $\mathbb{E}(XU) = \mathbb{E}(U) = 0$.

Point Identification

- If we could observe the whole population from which our sample was drawn, could we uniquely determine the parameters of interest?
- Suppose we observe the joint distribution of (X, Y)
- This is enough information to calculate (α, β) explicitly: they are **point identified**.

Classical Measurement Error

- Suppose we observe $(Y, \widetilde{X})$ rather than (Y, X), where $\widetilde{X} = X + W$
- W is **classical measurement error**: $\text{Cov}(W, X) = \text{Cov}(W, U) = \mathbb{E}(W) = 0$
- Are α and β still point identified?

The Good News

$$\mathbb{E}(\widetilde{X}) = \mathbb{E}(X + W) = \mathbb{E}(X)$$

$$\begin{aligned}
\text{Cov}(\widetilde{X}, Y) &= \text{Cov}(X + W, Y) = \text{Cov}(X, Y) + \text{Cov}(W, Y) \\
&= \text{Cov}(X, Y) + \text{Cov}(W, \alpha + \beta X + U) \\
&= \text{Cov}(X, Y) + \text{Cov}(W, U) + \beta \text{Cov}(W, X) \\
&= \text{Cov}(X, Y)
\end{aligned}$$

Are α and β still point identified?

The Bad News

- Because $\text{Var}(W)$ is not point identified, neither are α and β.

$$\text{Var}(\widetilde{X}) = \text{Var}(X + W) = \text{Var}(X) + \text{Var}(W) \geq \text{Var}(X)$$

$$\beta \equiv \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X}) - \text{Var}(W)}, \quad \alpha \equiv \mathbb{E}[Y] - \beta\mathbb{E}[X] = \mathbb{E}[Y] - \beta\mathbb{E}[\widetilde{X}].$$

Partial Identification

- We can still **bound** β and hence α: the so-called **reverse regression bounds**

A Lower Bound for β

- Since $\text{Cov}(X, Y) = \text{Cov}(\widetilde{X}, Y)$,

$$\frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X})} = \frac{\text{Cov}(X, Y)}{\text{Var}(X) + \text{Var}(W)} = \frac{\text{Cov}(X, Y)/\text{Var}(X)}{1 + \text{Var}(W)/\text{Var}(X)} = \frac{\beta}{1 + \text{Var}(W)/\text{Var}(X)}.$$

- Since $\text{Var}(W)/\text{Var}(X)$ is non-negative, $\text{Cov}(\widetilde{X}, Y)/\text{Var}(\widetilde{X})$ has same *sign* as β and

$$\left| \frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X})} \right| \leq |\beta|.$$

An Upper Bound for β

- Run the **reverse regression** $\widetilde{X}$ on Y

$$\frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(Y)} = \frac{\text{Cov}(X, Y)}{\beta^2 \text{Var}(X) + \text{Var}(U)} = \frac{\beta \text{Var}(X)}{\beta^2 \text{Var}(X) + \text{Var}(U)}.$$

- Take the reciprocal:

$$\frac{\text{Var}(Y)}{\text{Cov}(\widetilde{X}, Y)} = \beta + \frac{\text{Var}(U)}{\beta \text{Var}(X)} = \beta \left[1 + \frac{\text{Var}(U)}{\beta^2 \text{Var}(X)}\right].$$

- Factor in brackets *greater* than one, so $\text{Var}(Y)/\text{Cov}(\widetilde{X}, Y)$ has same sign as β and

$$\left|\frac{\text{Var}(Y)}{\text{Cov}(\widetilde{X}, Y)}\right| \geq |\beta|.$$

Reverse Regression Bounds

Terminology

- A bound is **sharp** if it cannot be improved, under our assumptions.
- A bound is **tight** if it is short enough to be useful in a practical example.

Assumptions

- $Y = \alpha + \beta X + U$ where $\mathbb{E}(XU) = \mathbb{E}(U) = 0$.
- Observe $(\widetilde{X}, Y)$
- $\widetilde{X} = X + W$ with $\mathbb{E}(W) = \text{Cov}(W, X) = \text{Cov}(W, U) = 0$

Sharp Bounds for β

- β lies between $\dfrac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X})}$ and $\dfrac{\text{Var}(Y)}{\text{Cov}(\widetilde{X}, Y)}$

How tight are the reverse regression bounds?

- Let r denote the correlation between $\widetilde{X}$ and Y. Then:

$$r^2 \equiv \frac{\text{Cov}(\widetilde{X}, Y)^2}{\text{Var}(\widetilde{X})\text{Var}(Y)} = \frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X})} \cdot \frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(Y)}.$$

- Re-arranging, it follows that:

$$r^2 \cdot \frac{\text{Var}(Y)}{\text{Cov}(\widetilde{X}, Y)} = \frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X})}.$$

- All else equal, bounds for β are *tighter* when $\widetilde{X}$ and Y are strongly correlated:

$$\text{Width} = \left| \frac{\text{Var}(Y)}{\text{Cov}(\widetilde{X}, Y)} - \frac{\text{Cov}(\widetilde{X}, Y)}{\text{Var}(\widetilde{X})} \right| = (1 - r^2) \left| \frac{\text{Var}(Y)}{\text{Cov}(\widetilde{X}, Y)} \right|.$$

```r
library(tidyverse)
library(broom) # for tidy()
set.seed(1066)

n <- 5000
X <- rnorm(n)
U <- rnorm(n)
W <- rnorm(n)

alpha <- 0.5
beta <- 1
Y <- alpha + beta * X + U
Xtilde <- X + W
```

```r
c(forward = cov(Xtilde, Y) / var(Xtilde),
  truth = beta,
  reverse = var(Y) / cov(Xtilde, Y)) |>  round(2)
```

```
## forward   truth reverse
##    0.51    1.00    1.95
```

```r
# The regression we can't run in practice!
lm(Y ~ X) |>  tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    0.489    0.0140      34.8 9.56e-238
## 2 X              1.02     0.0138      73.9 0
```

```r
# Reduce the correlation between X and Y, hence Xtilde and Y
Y <- alpha + beta * X + 3 * U

c(forward = cov(Xtilde, Y) / var(Xtilde),
  truth = beta,
  reverse = var(Y) / cov(Xtilde, Y)) |> round(2)
```

```
## forward   truth reverse
##    0.52    1.00    9.31
```

```r
# The regression we can't run in practice!
lm(Y ~ X) |> tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    0.466    0.0421      11.1 3.95e- 28
## 2 X              1.07     0.0414      25.7 7.45e-137
```

Review of Potential Outcomes Framework

- See https://expl.ai/QHUAVRV and https://expl.ai/DWVNRZU for more details.
- Binary **Treatment** $D \in \{0, 1\}$
- **Observed Outcome** Y depends on **Potential Outcomes** (Y_0, Y_1) via

$$Y = (1 - D)Y_0 + DY_1 = Y_0 + D(Y_1 - Y_0)$$

- Only one of (Y_0, Y_1) is observed for any given person at any given time.
- The unobserved potential outcome is a **counterfactual**, i.e. a **what if?**
- **Average Treatment Effect:** $\text{ATE} \equiv \mathbb{E}(Y_1 - Y_0)$.
- **Treatment on the Treated:** $\text{TOT} \equiv \mathbb{E}(Y_1 - Y_0 | D = 1)$.

Example: Y is Wage, D is Attend University

Counterfactuals

- $D = 1 \implies Y_0$ is the wage you *would have earned* if you *hadn't* attended.
- $D = 0 \implies Y_1$ is the wage you *would have earned* if you *had* attended.

Treatment Effects

- ATE $= \mathbb{E}(Y_1 - Y_0)$ is the average effect of *forcing* a randomly-chosen person to attend university.
- TOT $= \mathbb{E}(Y_1 - Y_0 | D = 1)$ is the average effect of attending university *for the sort of people who choose to attend*.

Problem: Selection Bias

- We don't force randomly-chosen people to attend university!
- People who choose to attend are likely different in *many ways*

Selection Bias

Naïve Comparison of Means

$$
\begin{aligned}
\mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0) &= \mathbb{E}(Y_1|D=1) - \mathbb{E}(Y_0|D=0) \\
&= \mathbb{E}(Y_1|D=1) - \mathbb{E}(Y_0|D=0) + \mathbb{E}(Y_0|D=1) - \mathbb{E}(Y_0|D=1) \\
&= \underbrace{\mathbb{E}(Y_1 - Y_0|D=1)}_{\text{TOT}} + \underbrace{[\mathbb{E}(Y_0|D=1) - \mathbb{E}(Y_0|D=0)]}_{\text{Selection Bias}}
\end{aligned}
$$

How does selection matter?

1. TOT is probably different from ATE: selection on gains.
2. Average value of Y_0 (“outside option”) probably varies with D.

How to solve the problem of selection bias?

Randomized Controlled Trial

- $D \perp\!\!\!\perp (Y_0, Y_1) \implies \mathbb{E}(Y_0|D) = \mathbb{E}(Y_0)$ and $\mathbb{E}(Y_1|D) = \mathbb{E}(Y_1)$ (video)
- Hence: TOT = ATE and Selection Bias = 0.

Other Approaches

- Selection-on-observables (chapter 4, video 1, video 2, slides, more slides)
- Instrumental Variables (chapter 5, tomorrow's lecture)
- Regression Discontinuity (chapter 7, slides)
- Difference-in-differences (chapter 8, slides)

Partial Identification

Bound the ATE *without* using the above approaches while *allowing* for selection bias.

Bounding the ATE when Y and D are Binary

- Example: $Y = 1$ if you earn a PhD, $D = 1$ if you attend an Ivy League University
- We know that D is *not* randomly assigned, and expect selection bias.

Starting point

- Assume that (Y, D) are observed.
- Since Y is binary we know that $-1 \leq \text{ATE} \leq 1$ without observing any data!

$$0 \leq Y_0 \leq 1 \quad \text{and} \quad 0 \leq Y_1 \leq 1 \implies 0 \leq \mathbb{E}(Y_0) \leq 1 \quad \text{and} \quad 0 \leq \mathbb{E}(Y_1) \leq 1$$

Shorthand

$$\begin{aligned}
P_{11} &\equiv \mathbb{P}(Y = 1|D = 1) = \mathbb{E}[Y|D = 1] = \mathbb{E}[Y_1|D = 1] \\
P_{10} &\equiv \mathbb{P}(Y = 1|D = 0) = \mathbb{E}[Y|D = 0] = \mathbb{E}[Y_0|D = 0] \\
p &\equiv \mathbb{P}(D = 1) = \mathbb{E}(D).
\end{aligned}$$

Assumption-Free Bounds: Improving on $-1 \leq \text{ATE} \leq 1$

Y and D Are Observed

- $\implies P_{11} \equiv \mathbb{E}[Y_1|D=1]$, $P_{10} \equiv \mathbb{E}[Y_0|D=0]$, and $p \equiv \mathbb{E}(D)$ are observed

Iterated Expectations

$$\mathbb{E}[Y_1] = \mathbb{E}_D\left[\mathbb{E}\left(Y_1|D\right)\right] = P_{11}p + \mathbb{E}[Y_1|D=0](1-p)$$

$$\mathbb{E}[Y_0] = \mathbb{E}_D\left[\mathbb{E}(Y_0|D)\right] = \mathbb{E}[Y_0|D=1]p + P_{10}(1-p).$$

Bound the Unobserved Quantities

- $\mathbb{E}[Y_1|D=0]$ and $\mathbb{E}[Y_0|D=1]$ are between 0 and 1

$$pP_{11} \leq \mathbb{E}[Y_1] \leq pP_{11} + (1-p)$$

$$(1-p)P_{10} \leq \mathbb{E}[Y_0] \leq p + (1-p)P_{10}$$

Assumption-Free Bounds: Width Equals 1

Previous Slide

$$
\begin{aligned}
pP_{11} \leq \mathbb{E}[Y_1] &\leq pP_{11} + (1-p) \\
(1-p)P_{10} \leq \mathbb{E}[Y_0] &\leq p + (1-p)P_{10}
\end{aligned}
$$

Combine These

$$
pP_{11} - (1-p)P_{10} - p \leq \mathbb{E}[Y_1 - Y_0] \leq pP_{11} - (1-p)P_{10} + (1-p).
$$

Written More Compactly

$$
q \leq \text{ATE} \leq (q+1), \quad q \equiv [pP_{11} - (1-p)P_{10} - p]
$$

- Half as wide as $-1 \leq \text{ATE} \leq 1$ but **always includes zero**

Add Assumptions, Tighten the Bounds (Details in Lecture Notes)

Monotone Treatment Selection (MTS)

- Suppose we know direction of self-selection into treatment, e.g. *positive*:

$$\mathbb{E}(Y_1|D=0) \leq \mathbb{E}(Y_1|D=1) \quad \text{and} \quad \mathbb{E}(Y_0|D=0) \leq \mathbb{E}(Y_0|D=1).$$

- Positive MTS gives an improved *upper bound* for the ATE:

$$q \leq \text{ATE} \leq P_{11} - P_{10} \leq (q+1), \quad q \equiv [pP_{11} - (1-p)P_{10} - p]$$

Monotone Treatment Response (MTR)

- Suppose we know the direction of the **causal effect**: e.g. *positive effect*: $Y_1 > Y_0$.
- Positive MTR gives an improved *lower* bound for the ATE, namely zero:

$$0 \leq \text{ATE} \leq (q+1)$$

A Comparison of Bounds

- Preceding bounds are *sharp* under their respective assumptions. How *tight* are they?
- Example: suppose that 8% of Ivy League graduates earn a PhD versus 1.5% of the general public and that 0.2% of people attend an Ivy League institution.

$$(P_{11} = 0.08,\ P_{10} = 0.015,\ p = 0.002) \implies q \equiv [pP_{11} - (1-p)P_{10} - p] \approx -0.017$$

$$\begin{aligned}
\text{No Asumptions:} \quad & [q, q+1] \approx [-0.017, 0.983] \\
\text{Positive MTS:} \quad & [q, P_{11} - P_{10}] \approx [-0.017, 0.065] \\
\text{Positive MTR:} \quad & [0, q+1] \approx [0, 0.983] \\
\text{Positive MTS + MTR:} \quad & [0, P_{11} - P_{10}] = [0, 0.065].
\end{aligned}$$

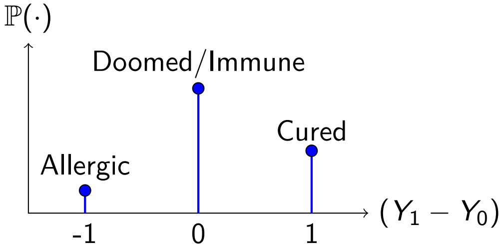
- Here positive MTR has little effect; positive MTS makes a dramatic difference!

Bounding the Distribution of Treatment Effects

- Randomly assign $D \implies$ ATE point identified: no selection bias!
- (Y_0, Y_1) never observed for same person; can't learn joint distribution.
- Anything that depends on this joint distribution is *not point identified*.
- Examples: $\text{Var}(Y_1 - Y_0)$, $\mathbb{P}(Y_1 - Y_0 > 0)$
- Can we *partially identify* the distribution of treatment effect $(Y_1 - Y_0)$?
- Start with binary Y case; then consider the general case.

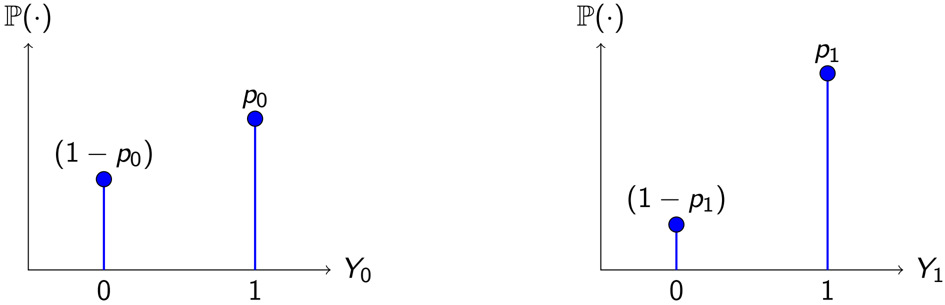
Unobserved: Joint Distribution of (Y_0, Y_1), Distribution of $(Y_1 - Y_0)$

		Y_1	
		0	1
Y_0	0	$\mathbb{P}$(Doomed)	$\mathbb{P}$(Cured)
	1	$\mathbb{P}$(Allergic)	$\mathbb{P}$(Immune)



- Dangerous disease, and dangerous treatment.
- Treatment helps some people (the "Cured"), harms others (the "Allergic").
- Treatment has no effect on other people (the "Doomed" and "Immune")
- **Are more people helped than harmed?**

Observed: Marginal Distributions of Y_0 and Y_1



- Assume (Y, D) come from a randomized, double-blind, placebo-controlled trial.
- p_0 is the share of untreated who recover; p_1 is the share of treated who recover.
- The ATE is $p_1 - p_0$
- Try to bound what we can't observe using what we can observe.

From Joint (Unobserved) to Marginals (Observed)

Recall: $p_0 \equiv \mathbb{P}(Y_0 = 1)$ and $p_1 \equiv \mathbb{P}(Y_1 = 1)$.

$$
\begin{array}{|cc|cc|cc}
\hline
 & & \multicolumn{2}{c|}{Y_1} & & \\
 & & 0 & 1 & & \\
\hline
\multirow{2}{*}{Y_0} & 0 & \mathbb{P}(\text{Doomed}) & \mathbb{P}(\text{Cured}) & \Longrightarrow & (1 - p_0) = \mathbb{P}(D) + \mathbb{P}(C) \\
 & 1 & \mathbb{P}(\text{Allergic}) & \mathbb{P}(\text{Immune}) & \Longrightarrow & p_0 = \mathbb{P}(A) + \mathbb{P}(I) \\
\hline
 & & \Downarrow & \Downarrow & & \\
 & & (1 - p_1) = \mathbb{P}(D) + \mathbb{P}(A) & p_1 = \mathbb{P}(C) + \mathbb{P}(I) & &
\end{array}
$$

Shorthand: $\alpha \equiv \mathbb{P}(\text{Allergic})$

Previous Slide

$$
\begin{aligned}
(1 - p_0) &= \mathbb{P}(\text{Doomed}) + \mathbb{P}(\text{Cured}) \\
p_0 &= \mathbb{P}(\text{Allergic}) + \mathbb{P}(\text{Immune}) \\
(1 - p_1) &= \mathbb{P}(\text{Doomed}) + \mathbb{P}(\text{Allergic}) \\
p_1 &= \mathbb{P}(\text{Cured}) + \mathbb{P}(\text{Immune})
\end{aligned}
$$

Rearranging

$$
\begin{aligned}
\mathbb{P}(\text{Immune}) &= p_0 - \alpha \\
\mathbb{P}(\text{Doomed}) &= (1 - p_1) - \alpha \\
\mathbb{P}(\text{Cured}) &= (p_1 - p_0) + \alpha
\end{aligned}
$$

- Everything is written in terms of observables (p_0, p_1) and α!

Bounding $\alpha \equiv \mathbb{P}(\text{Allergic})$

Previous Slide

- $\mathbb{P}(\text{Immune}) = p_0 - \alpha,\ \mathbb{P}(\text{Doomed}) = (1 - p_1) - \alpha,\ \mathbb{P}(\text{Cured}) = (p_1 - p_0) + \alpha$

Probabilities are between 0 and 1

- Apply Immune, Doomed, and Cured to bound α:

$$0 \leq (p_1 - p_0) + \alpha \leq 1, \quad 0 \leq (1 - p_1) - \alpha \leq 1, \quad 0 \leq p_0 - \alpha \leq 1.$$

Simplify

- Rearrange the preceding, and combine with $0 \leq \alpha \leq 1$

$$\max\{-\text{ATE}, 0\} \leq \alpha \leq \min\{p_0, (1 - p_1)\}, \quad \text{ATE} = (p_1 - p_0).$$

(Pointwise) Sharp Bounds for Distribution of Treatment Effects

Previous Slide

- $\mathbb{P}(\text{Immune}) = p_0 - \alpha,\ \mathbb{P}(\text{Doomed}) = (1 - p_1) - \alpha,\ \mathbb{P}(\text{Cured}) = (p_1 - p_0) + \alpha$
- $\max\{-(p_1 - p_0), 0\} \leq \alpha \leq \{p_0, (1 - p_1)\}$

Shorthand

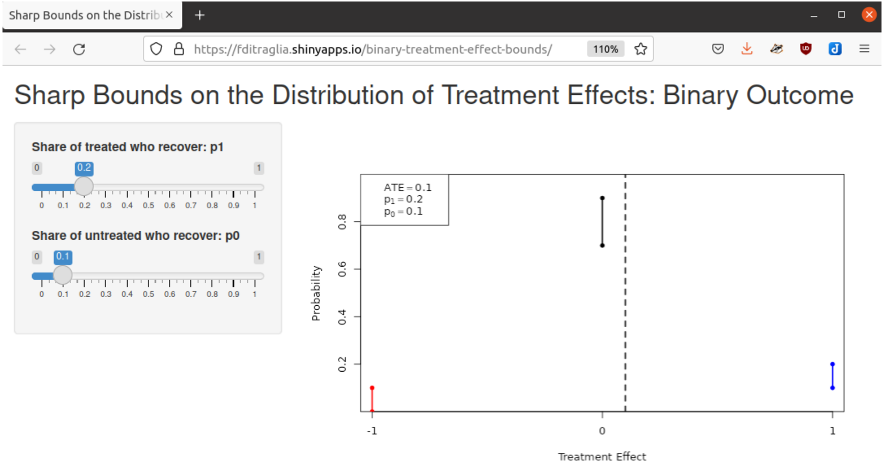
- $\underline{\alpha} \equiv \max\{-(p_1 - p_0), 0\}, \quad \overline{\alpha} \equiv \min\{p_0, (1 - p_1)\}$

Combine

- Recall that $\alpha \equiv \mathbb{P}(\text{Allergic}) = \mathbb{P}(Y_1 - Y_0 = -1)$

$$
\begin{aligned}
\underline{\alpha} \leq \mathbb{P}(Y_1 - Y_0 = -1) &\leq \overline{\alpha} \\
(1 - p_1) + p_0 - 2\overline{\alpha} \leq \mathbb{P}(Y_1 - Y_0 = 0) &\leq (1 - p_1) + p_0 - 2\underline{\alpha} \\
(p_1 - p_0) + \underline{\alpha} \leq \mathbb{P}(Y_1 - Y_0 = 1) &\leq (p_1 - p_0) + \overline{\alpha}
\end{aligned}
$$

https://fditraglia.shinyapps.io/binary-treatment-effect-bounds/



The General Case: Fan & Park (2010)

- Above we assumed that (Y_0, Y_1) were both binary.
- We asked which joint distributions were **not ruled out** based on the marginals.
- Pointwise sharp bounds for $\mathbb{P}(Y_1 - Y_0 = -1)$, $\mathbb{P}(Y_1 - Y_0 = 0)$ and $\mathbb{P}(Y_1 - Y_0 = 1)$.
- Special case of a general result: Fan and Park (2010).
- Same basic idea, but math is harder when (Y_0, Y_1) may not be binary.
- **This is a result you may actually use in practice!**
- Explain their result without proving it.

Fan & Park (2010) Bounds

Observables

- $F_0(y) \equiv \mathbb{P}(Y_0 \le y)$ and $F_1(y) \equiv \mathbb{P}(Y_1 \le y)$

Goal

- Sharp bounds for $F(\delta) \equiv \mathbb{P}(Y_1 - Y_0 \le \delta)$

Notation

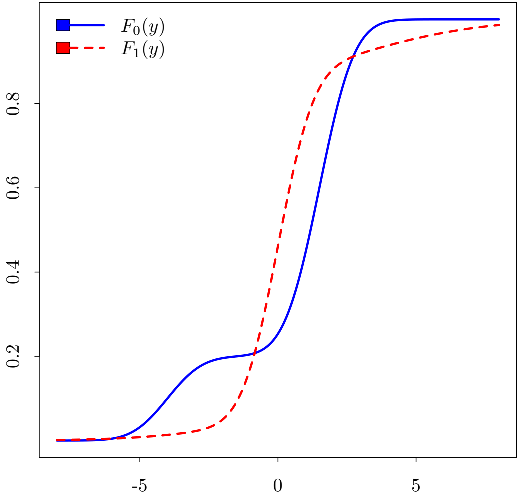
$$\underline{F}(\delta) \equiv \sup_y F_1(y) - F_0(y - \delta)$$

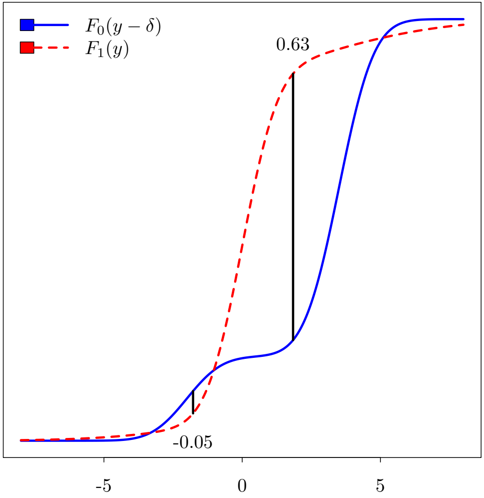
$$\overline{F}(\delta) \equiv 1 + \left[\inf_y F_1(y) - F_0(y - \delta)\right]$$

Theorem

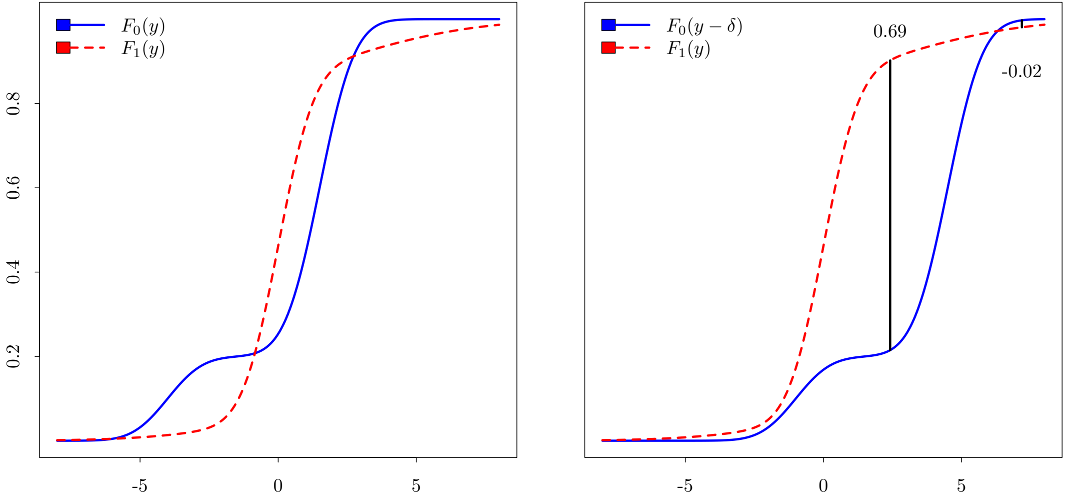
- For any δ, $0 \le \underline{F}(\delta) \le F(\delta) \le \overline{F}(\delta) \le 1$. These bounds are (pointwise) sharp.

Left: $\delta = 0$, Right: $\delta = 2$

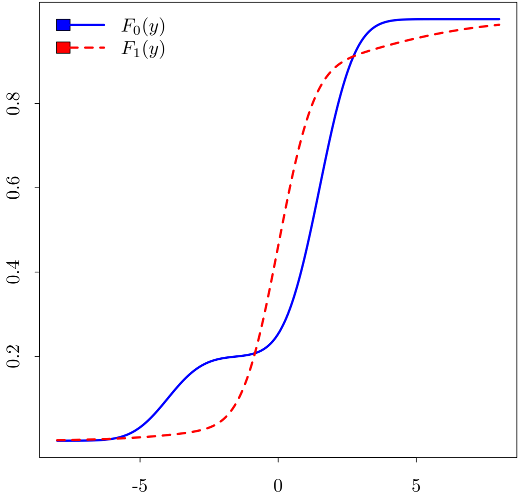


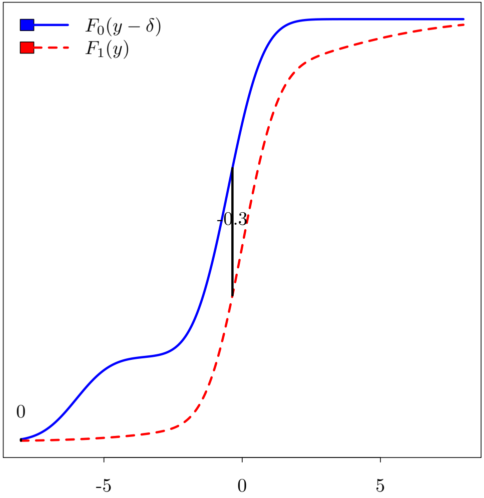


Left: $\delta = 0$, Right: $\delta = 3$



Left: $\delta = 0$, Right: $\delta = -2$





All the bounds!