Administrative Records Mask Racially Biased Policing

Dean Knox, Will Lowe and Jonathan Mummolo[1]
Princeton University

January 28, 2021

[1]We thank Michael Pomirchy for research assistance.

How do we measure racial bias in policing?

Eric Garner, 2014

Racial bias in policing

- Causal question.

Racial bias in policing

- Causal question. Focus on traditional omitted variable bias.

Racial bias in policing

- Causal question. Focus on traditional omitted variable bias.
- Overlooked:

Racial bias in policing

- Causal question. Focus on traditional omitted variable bias.
- Overlooked:

  Sample selection bias due to post-treatment conditioning

Intuition for the problem

- Goal: estimate effect of civilian race on police use of force

Intuition for the problem

- Goal: estimate effect of civilian race on police use of force
- Suppose perfect as-if experimental conditions: a set of police encounters (i.e. sightings) identical but for race of civilians

Intuition for the problem

- Goal: estimate effect of civilian race on police use of force
- Suppose perfect as-if experimental conditions: a set of police encounters (i.e. sightings) identical but for race of civilians
- Suppose racial bias leads police to stop white civilians if engaged in serious crime (e.g. bank robbery), stop black civilians regardless of behavior

Intuition for the problem

- Goal: estimate effect of civilian race on police use of force
- Suppose perfect as-if experimental conditions: a set of police encounters (i.e. sightings) identical but for race of civilians
- Suppose racial bias leads police to stop white civilians if engaged in serious crime (e.g. bank robbery), stop black civilians regardless of behavior
- Now throw away all data on civilians police observe but do not stop (e.g. the NYPD "Stop, Question and Frisk" (SQF) database).

Intuition for the problem

- Goal: estimate effect of civilian race on police use of force
- Suppose perfect as-if experimental conditions: a set of police encounters (i.e. sightings) identical but for race of civilians
- Suppose racial bias leads police to stop white civilians if engaged in serious crime (e.g. bank robbery), stop black civilians regardless of behavior
- Now throw away all data on civilians police observe but do not stop (e.g. the NYPD "Stop, Question and Frisk" (SQF) database). We've just ruined our experiment!

Intuition for the problem

- Goal: estimate effect of civilian race on police use of force
- Suppose perfect as-if experimental conditions: a set of police encounters (i.e. sightings) identical but for race of civilians
- Suppose racial bias leads police to stop white civilians if engaged in serious crime (e.g. bank robbery), stop black civilians regardless of behavior
- Now throw away all data on civilians police observe but do not stop (e.g. the NYPD "Stop, Question and Frisk" (SQF) database). We've just ruined our experiment!
- Comparing white bank robbers to black civilians committing no crime.

Intuition for the problem

- Goal: estimate effect of civilian race on police use of force
- Suppose perfect as-if experimental conditions: a set of police encounters (i.e. sightings) identical but for race of civilians
- Suppose racial bias leads police to stop white civilians if engaged in serious crime (e.g. bank robbery), stop black civilians regardless of behavior
- Now throw away all data on civilians police observe but do not stop (e.g. the NYPD "Stop, Question and Frisk" (SQF) database). We've just ruined our experiment!
- Comparing white bank robbers to black civilians committing no crime. If we then found no disparity in rates of force against black/white civilians, that should be alarming!

Intuition for the problem

- Goal: estimate effect of civilian race on police use of force
- Suppose perfect as-if experimental conditions: a set of police encounters (i.e. sightings) identical but for race of civilians
- Suppose racial bias leads police to stop white civilians if engaged in serious crime (e.g. bank robbery), stop black civilians regardless of behavior
- Now throw away all data on civilians police observe but do not stop (e.g. the NYPD "Stop, Question and Frisk" (SQF) database). We've just ruined our experiment!
- Comparing white bank robbers to black civilians committing no crime. If we then found no disparity in rates of force against black/white civilians, that should be alarming!
- Current literature reads this result as "no evidence of racial bias in the use of force"

Estimating racial bias with police data

1. Racial bias likely affects who police choose to investigate $\rightarrow$ which encounters appear in police data

Estimating racial bias with police data

1. Racial bias likely affects who police choose to investigate $\rightarrow$ which encounters appear in police data
2. Police administrative data are inherently **post-treatment**

Estimating racial bias with police data

1. Racial bias likely affects who police choose to investigate $\rightarrow$ which encounters appear in police data
2. Police administrative data are inherently **post-treatment**
3. Results statistically biased; bias often can't be "controlled away"

Estimating racial bias with police data

1. Racial bias likely affects who police choose to investigate $\rightarrow$ which encounters appear in police data

2. Police administrative data are inherently **post-treatment**

3. Results statistically biased; bias often can't be "controlled away"

4. Bias has a precise form, can derive **informative** bounds on the true causal effect of civilian race on police behavior

Estimating racial bias with police data

1. Racial bias likely affects who police choose to investigate $\rightarrow$ which encounters appear in police data
2. Police administrative data are inherently **post-treatment**
3. Results statistically biased; bias often can't be "controlled away"
4. Bias has a precise form, can derive **informative** bounds on the true causal effect of civilian race on police behavior
5. Prior work ignoring this feature substantially underestimates racial bias in use of force (Fryer, 2019)

Estimating racial bias with police data

1. Racial bias likely affects who police choose to investigate $\rightarrow$ which encounters appear in police data
2. Police administrative data are inherently **post-treatment**
3. Results statistically biased; bias often can't be "controlled away"
4. Bias has a precise form, can derive **informative** bounds on the true causal effect of civilian race on police behavior
5. Prior work ignoring this feature substantially underestimates racial bias in use of force (Fryer, 2019)
6. New research designs to avoid this pitfall

Defining the Statistical Problem

A causal mediation framework

- Unit of analysis: police-civilian *encounters*:

A causal mediation framework

- Unit of analysis: police-civilian *encounters*:
  - "Encounter" = sighting of individual by police officer

A causal mediation framework

- Unit of analysis: police-civilian *encounters*:
  - "Encounter" = sighting of individual by police officer
  - Counterfactual: substitution of individual of differing race into police-civilian encounter, holding circumstance and civilian behavior fixed

A causal mediation framework

- Unit of analysis: police-civilian *encounters*:
  - "Encounter" = sighting of individual by police officer
  - Counterfactual: substitution of individual of differing race into police-civilian encounter, holding circumstance and civilian behavior fixed
- Treatment (civilian is racial minority) $D_i \in \{0,1\}$

A causal mediation framework

- Unit of analysis: police-civilian *encounters*:
  - "Encounter" = sighting of individual by police officer
  - Counterfactual: substitution of individual of differing race into police-civilian encounter, holding circumstance and civilian behavior fixed
- Treatment (civilian is racial minority) $D_i \in \{0,1\}$
- Outcome (use of force) $Y_i \in \{0,1\}$

A causal mediation framework

- Unit of analysis: police-civilian *encounters*:
  - "Encounter" = sighting of individual by police officer
  - Counterfactual: substitution of individual of differing race into police-civilian encounter, holding circumstance and civilian behavior fixed
- Treatment (civilian is racial minority) $D_i \in \{0,1\}$
- Outcome (use of force) $Y_i \in \{0,1\}$
- Mediator (being stopped by police) $M_i \in \{0,1\}$

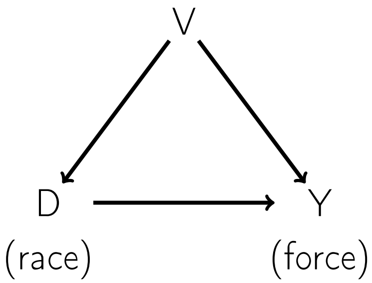
A causal mediation framework

- Unit of analysis: police-civilian *encounters*:
  - "Encounter" = sighting of individual by police officer
  - Counterfactual: substitution of individual of differing race into police-civilian encounter, holding circumstance and civilian behavior fixed
- Treatment (civilian is racial minority) $D_i \in \{0,1\}$
- Outcome (use of force) $Y_i \in \{0,1\}$
- Mediator (being stopped by police) $M_i \in \{0,1\}$
- Racial bias in police stops ($D_i \rightarrow M_i$)
  (e.g. Gelman, Fagan & Kiss 2007; Glaser 2014; Lerman & Weaver 2014; Goel, Rao & Shroff 2016)

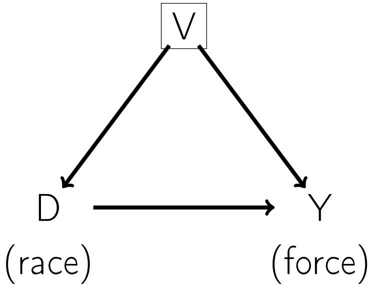
Existing theory of race and police-civilian encounters

$$\underset{\text{(race)}}{D} \longrightarrow \underset{\text{(force)}}{Y}$$

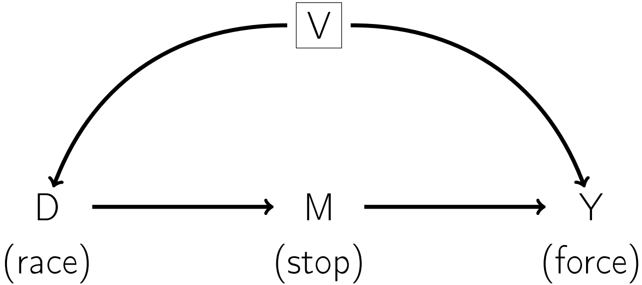
Existing theory of race and police-civilian encounters



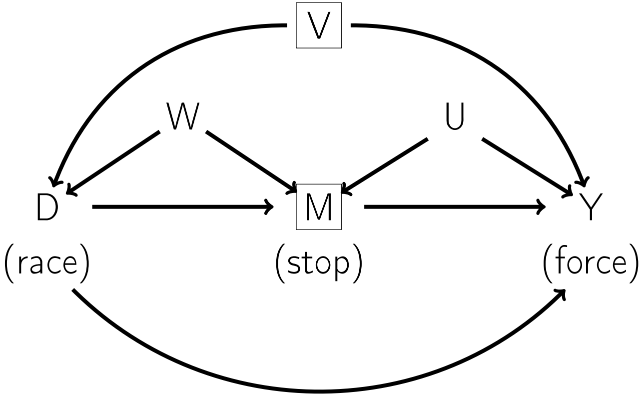
Existing theory of race and police-civilian encounters



A more complete theory



A more complete theory



Potential outcomes with mediation

Normally we consider $Y(d) \in \{Y(1), Y(0)\}$; potential force given race (treatment)

Potential outcomes with mediation

Normally we consider $Y(d) \in \{Y(1), Y(0)\}$; potential force given race (treatment)

Instead consider $Y(d, M(d))$

Potential outcomes with mediation

Normally we consider $Y(d) \in \{Y(1), Y(0)\}$; potential force given race (treatment)

Instead consider $Y(d, M(d))$

$Y(1, 1) =$ potential use of force if minority civilian stopped

Potential outcomes with mediation

Normally we consider $Y(d) \in \{Y(1), Y(0)\}$; potential force given race (treatment)

Instead consider $Y(d, M(d))$

$Y(1,1) =$ potential use of force if minority civilian stopped
$Y(1,0) =$ potential use of force if minority civilian *not* stopped

Potential outcomes with mediation

Normally we consider $Y(d) \in \{Y(1), Y(0)\}$; potential force given race (treatment)

Instead consider $Y(d, M(d))$

$Y(1,1)$ = potential use of force if minority civilian stopped
$Y(1,0)$ = potential use of force if minority civilian *not* stopped
$Y(0,1)$ = potential use of force if white civilian stopped

Potential outcomes with mediation

Normally we consider $Y(d) \in \{Y(1), Y(0)\}$; potential force given race (treatment)

Instead consider $Y(d, M(d))$

$Y(1,1) =$ potential use of force if minority civilian stopped
$Y(1,0) =$ potential use of force if minority civilian *not* stopped
$Y(0,1) =$ potential use of force if white civilian stopped
$Y(0,0) =$ potential use of force if white civilian *not* stopped

Formalizing the Missing Data Problem

Solution: principal stratification

- If $D \to M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$		
$M_i(1) = 0$		

Types of police-civilian encounters

- If $D \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	
$M_i(1) = 0$		

Types of police-civilian encounters

- If $D \to M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	
$M_i(1) = 0$		"never stop" (inconspicuous)

Types of police-civilian encounters

- If $D \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$		"never stop" (inconspicuous)

Types of police-civilian encounters

- If $D \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$	stop if white ?	"never stop" (inconspicuous)

Types of police-civilian encounters

- If $D \to M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$	stop if white ?	"never stop" (inconspicuous)

What do we get to see in police data?

Types of police-civilian encounters

- If $D \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$	stop if white ?	"never stop" (inconspicuous)

For black civilians . . .

Types of police-civilian encounters

- If $D \rightarrow M$, four types of police-civilian encounters:

	$M_i(0) = 1$	$M_i(0) = 0$
$M_i(1) = 1$	"always stop" (serious crime)	stop if black (jaywalking)
$M_i(1) = 0$	stop if white ?	"never stop" (inconspicuous)

For white civilians ...

Encounters (sightings) belong to one of four principal strata

always-stop	anti-black racial-stop	anti-white racial-stop	never-stop
$M_i(1) = M_i(0) = 1$	$M_i(1) = 1, M_i(0) = 0$	$M_i(1) = 0, M_i(0) = 1$	$M_i(1) = 0, M_i(0) = 0$

Within each, civilian is treated (black) or not (white)

$$\begin{array}{cccc}
\text{always-stop} & \text{anti-black racial-stop} & \text{anti-white racial-stop} & \text{never-stop} \\
M_i(1) = M_i(0) = 1 & M_i(1) = 1, M_i(0) = 0 & M_i(1) = 0, M_i(0) = 1 & M_i(1) = 0, M_i(0) = 0 \\
/ \quad \backslash & / \quad \backslash & / \quad \backslash & / \quad \backslash \\
D_i = 0 \quad D_i = 1 & D_i = 0 \quad D_i = 1 & D_i = 0 \quad D_i = 1 & D_i = 0 \quad D_i = 1
\end{array}$$

Four potential outcomes we may need to estimate

always-stop
$M_i(1) = M_i(0) = 1$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
$Y_i(1,0)$	$Y_i(1,0)$
$Y_i(0,1)$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$

anti-black racial-stop
$M_i(1) = 1, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
$Y_i(1,0)$	$Y_i(1,0)$
$Y_i(0,1)$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$

anti-white racial-stop
$M_i(1) = 0, M_i(0) = 1$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
$Y_i(1,0)$	$Y_i(1,0)$
$Y_i(0,1)$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$

never-stop
$M_i(1) = 0, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
$Y_i(1,0)$	$Y_i(1,0)$
$Y_i(0,1)$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$

Very few potential outcomes appear in police data

always-stop $M_i(1) = M_i(0) = 1$		anti-black racial-stop $M_i(1) = 1, M_i(0) = 0$		anti-white racial-stop $M_i(1) = 0, M_i(0) = 1$		never-stop $M_i(1) = 0, M_i(0) = 0$	
$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$	$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$	$Y_i(1,1)$	$Y_i(1,1)$	$Y_i(1,1)$	$Y_i(1,1)$
$Y_i(1,0)$	$Y_i(1,0)$	$Y_i(1,0)$	$Y_i(1,0)$	$Y_i(1,0)$	$Y_i(1,0)$	$Y_i(1,0)$	$Y_i(1,0)$
$\boldsymbol{Y_i(0,1)}$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$\boldsymbol{Y_i(0,1)}$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$	$Y_i(0,0)$	$Y_i(0,0)$	$Y_i(0,0)$	$Y_i(0,0)$	$Y_i(0,0)$	$Y_i(0,0)$

Causal Quantities of Interest

Which causal effect?

- Prior work does not name specific causal estimands

Which causal effect?

- Prior work does not name specific causal estimands
- Without naming an estimand, we can't consider identifying assumptions or evaluate validity of an analysis

Which causal effect?

- Prior work does not name specific causal estimands
- Without naming an estimand, we can't consider identifying assumptions or evaluate validity of an analysis
- There are many causal effects:

Which causal effect?

- Prior work does not name specific causal estimands
- Without naming an estimand, we can't consider identifying assumptions or evaluate validity of an analysis
- There are many causal effects:
  - Average Treatment Effect (ATE) in the population

Which causal effect?

- Prior work does not name specific causal estimands
- Without naming an estimand, we can't consider identifying assumptions or evaluate validity of an analysis
- There are many causal effects:
  - Average Treatment Effect (ATE) in the population
  - Effect among those who interact with police ($ATE_{M=1}$)

Which causal effect?

- Prior work does not name specific causal estimands
- Without naming an estimand, we can't consider identifying assumptions or evaluate validity of an analysis
- There are many causal effects:
  - Average Treatment Effect (ATE) in the population
  - Effect among those who interact with police ($ATE_{M=1}$)
  - Effect among minorities who interact with police ($ATT_{M=1}$)

Causal estimands

i	Stratum	D_i	M_i	$M_i(0)$	$M_i(1)$	$Y_i(1,1)$	$Y_i(1,0)$	$Y_i(0,1)$	$Y_i(0,0)$	ATE	$ATE_{M=1}$	$ATT_{M=1}$	$CDE_{M=1}$
1	Always-Stop	1	1	1	1	1	0	1	0	0	0	0	0
2	Always-Stop	0	1	1	1	1	0	1	0	0	0		0
3	Racial Stop	1	1	0	1	1	0	1	0	1	1	1	0
4	Never-Stop	0	0	0	0	1	0	0	0	0			

Average Treatment Effect

$$ATE = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))]$$

i	Stratum	D_i	M_i	$M_i(0)$	$M_i(1)$	$Y_i(1,1)$	$Y_i(1,0)$	$Y_i(0,1)$	$Y_i(0,0)$	ATE	$ATE_{M=1}$	$ATT_{M=1}$	$CDE_{M=1}$
1	Always-Stop	1	1	1	1	1	0	1	0	0	0	0	0
2	Always-Stop	0	1	1	1	1	0	1	0	0	0		0
3	Racial Stop	1	1	0	1	1	0	1	0	1	1	1	0
4	Never-Stop	0	0	0	0	1	0	0	0	0			

Average Treatment Effect Among the Stopped

$$ATE_{M=1} = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0)) | M_i = 1]$$

i	Stratum	D_i	M_i	$M_i(0)$	$M_i(1)$	$Y_i(1,1)$	$Y_i(1,0)$	$Y_i(0,1)$	$Y_i(0,0)$	ATE	$ATE_{M=1}$	$ATT_{M=1}$	$CDE_{M=1}$
1	Always-Stop	1	1	1	1	1	0	1	0	0	0	0	0
2	Always-Stop	0	1	1	1	1	0	1	0	0	0		0
3	Racial Stop	1	1	0	1	1	0	1	0	1	1	1	0
4	Never-Stop	0	0	0	0	1	0	0	0	0			

Average Treatment Effect Among the Treated and Stopped

$$ATT_{M=1} = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0)) | D_i = 1, M_i = 1]$$

i	Stratum	D_i	M_i	$M_i(0)$	$M_i(1)$	$Y_i(1,1)$	$Y_i(1,0)$	$Y_i(0,1)$	$Y_i(0,0)$	ATE	$ATE_{M=1}$	$ATT_{M=1}$	$CDE_{M=1}$
1	Always-Stop	1	1	1	1	1	0	1	0	0	0	0	0
2	Always-Stop	0	1	1	1	1	0	1	0	0	0		0
3	Racial Stop	1	1	0	1	1	0	1	0	1	1	1	0
4	Never-Stop	0	0	0	0	1	0	0	0	0			

Can anything causal be estimated with these data?

- To estimate causal quantities, need additional assumptions

Can anything causal be estimated with these data?

- To estimate causal quantities, need additional assumptions
- Goal: minimal, non-parametric, plausible

Assumptions

Assumption 1: Mandatory reporting

$$Y_i(d, 0) = 0$$

- If encounter not in the data, no force was applied

Assumption 1: Mandatory reporting

$$Y_i(d, 0) = 0$$

- If encounter not in the data, no force was applied
- Highly plausible for lethal/severe force

Assumption 1: Mandatory reporting

$$Y_i(d, 0) = 0$$

- If encounter not in the data, no force was applied
- Highly plausible for lethal/severe force
- Increasingly plausible for sub-lethal force given civilian oversight boards, cell phone cameras

Assumption 1: Mandatory reporting

always-stop	anti-black racial-stop	anti-white racial-stop	never-stop
$M_i(1) = M_i(0) = 1$	$M_i(1) = 1, M_i(0) = 0$	$M_i(1) = 0, M_i(0) = 1$	$M_i(1) = 0, M_i(0) = 0$

always-stop ($M_i(1) = M_i(0) = 1$)

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$
$Y_i(1,0)$	$Y_i(1,0)$
$\boldsymbol{Y_i(0,1)}$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$

anti-black racial-stop ($M_i(1) = 1, M_i(0) = 0$)

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$
$Y_i(1,0)$	$Y_i(1,0)$
$Y_i(0,1)$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$

anti-white racial-stop ($M_i(1) = 0, M_i(0) = 1$)

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
$Y_i(1,0)$	$Y_i(1,0)$
$\boldsymbol{Y_i(0,1)}$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$

never-stop ($M_i(1) = 0, M_i(0) = 0$)

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
$Y_i(1,0)$	$Y_i(1,0)$
$Y_i(0,1)$	$Y_i(0,1)$
$Y_i(0,0)$	$Y_i(0,0)$

Assumption 1: Mandatory reporting

always-stop

$M_i(1) = M_i(0) = 1$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$
0	0
$Y_i(0,1)$	$Y_i(0,1)$
0	0

anti-black racial-stop

$M_i(1) = 1, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$
0	0
$Y_i(0,1)$	$Y_i(0,1)$
0	0

anti-white racial-stop

$M_i(1) = 0, M_i(0) = 1$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
0	0
$Y_i(0,1)$	$Y_i(0,1)$
0	0

never-stop

$M_i(1) = 0, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
0	0
$Y_i(0,1)$	$Y_i(0,1)$
0	0

Assumption 2: Mediator monotonicity

$$M_i(1) \geq M_i(0)$$

- No anti-white bias in stopping

Assumption 2: Mediator monotonicity

always-stop		anti-black racial-stop		anti-white racial-stop		never-stop	
$M_i(1) = M_i(0) = 1$		$M_i(1) = 1, M_i(0) = 0$		$M_i(1) = 0, M_i(0) = 1$		$M_i(1) = 0, M_i(0) = 0$	
$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$	$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$	$Y_i(1,1)$	$Y_i(1,1)$	$Y_i(1,1)$	$Y_i(1,1)$
0	0	0	0	0	0	0	0
$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$
0	0	0	0	0	0	0	0

Assumption 2: Mediator monotonicity

always-stop

$M_i(1) = M_i(0) = 1$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$
0	0
$Y_i(0,1)$	$Y_i(0,1)$
0	0

anti-black
racial-stop

$M_i(1) = 1, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$
0	0
$Y_i(0,1)$	$Y_i(0,1)$
0	0

never-stop

$M_i(1) = 0, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
0	0
$Y_i(0,1)$	$Y_i(0,1)$
0	0

Assumption 3: Relative non-severity of racial stops

$$E[Y_i(d,m)|D_i = d', M_i(1) = 1, M_i(0) = 1] \geq \mathbb{E}[Y_i(d,m)|D_i = d', M_i(1) = 1, M_i(0) = 0]$$

- Level of force applied in always-stop encounters (serious crimes) $\geq$ level applied in racial stop encounters *on average*

Assumption 3: Relative non-severity of racial stops

always-stop
$M_i(1) = M_i(0) = 1$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$
0	0
$Y_i(0,1)$	$Y_i(0,1)$
0	0

anti-black
racial-stop
$M_i(1) = 1, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$
0	0
$Y_i(0,1)$	$Y_i(0,1)$
0	0

never-stop
$M_i(1) = 0, M_i(0) = 0$

$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$Y_i(1,1)$
0	0
$Y_i(0,1)$	$Y_i(0,1)$
0	0

Assumption 3: Relative non-severity of racial stops

always-stop $M_i(1) = M_i(0) = 1$			anti-black racial-stop $M_i(1) = 1, M_i(0) = 0$		never-stop $M_i(1) = 0, M_i(0) = 0$	
$D_i = 0$	$D_i = 1$		$D_i = 0$	$D_i = 1$	$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$	$\geq$	$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$	$Y_i(1,1)$	$Y_i(1,1)$
0	0		0	0	0	0
$Y_i(0,1)$	$Y_i(0,1)$		$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$	$Y_i(0,1)$
0	0		0	0	0	0

Assumption 4: Treatment ignorability

$$
\begin{aligned}
M_i(d) &\perp\!\!\!\perp D_i \\
Y_i(d, M(d)) &\perp\!\!\!\perp D_i | M_i(0) = m', M_i(1) = m''
\end{aligned}
$$

- No omitted variables with respect to mediator or outcome

Assumption 4: Treatment ignorability

$$\begin{aligned} M_i(d) &\perp\!\!\!\perp D_i \\ Y_i(d, M(d)) &\perp\!\!\!\perp D_i | M_i(0) = m', M_i(1) = m'' \end{aligned}$$

- No omitted variables with respect to mediator or outcome
- More plausible in recent years (data on lat/lon, time, officer and suspect features, etc.)

Assumption 4: Treatment Ignorability

always-stop			anti-black racial-stop			never-stop	
$M_i(1) = M_i(0) = 1$			$M_i(1) = 1, M_i(0) = 0$			$M_i(1) = 0, M_i(0) = 0$	
$D_i = 0$	$D_i = 1$		$D_i = 0$	$D_i = 1$		$D_i = 0$	$D_i = 1$
$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$	$\geq$	$Y_i(1,1)$	$\boldsymbol{Y_i(1,1)}$		$Y_i(1,1)$	$Y_i(1,1)$
0	0		0	0		0	0
$Y_i(0,1)$	$Y_i(0,1)$		$Y_i(0,1)$	$Y_i(0,1)$		$Y_i(0,1)$	$Y_i(0,1)$
0	0		0	0		0	0

Assumption 4: Treatment Ignorability

always-stop		anti-black racial-stop	never-stop
$M_i(1) = M_i(0) = 1$		$M_i(1) = 1, M_i(0) = 0$	$M_i(1) = 0, M_i(0) = 0$
$\boldsymbol{Y_i}(1,1)$		$\boldsymbol{Y_i}(1,1)$	$Y_i(1,1)$
0	$\geq$	0	0
$\boldsymbol{Y_i}(0,1)$		$\boldsymbol{Y_i}(0,1)$	$Y_i(0,1)$
0		0	0

Given assumptions 1-4, can we recover a causal quantity?

- Consider the naïve estimator:

$$\hat{\Delta} = \hat{\mathbb{E}}[Y_i|D_i = 1, M_i = 1] - \hat{\mathbb{E}}[Y_i|D_i = 0, M_i = 1]$$

Given assumptions 1-4, can we recover a causal quantity?

- Consider the naïve estimator:

  $$\hat{\Delta} = \hat{\mathbb{E}}[Y_i | D_i = 1, M_i = 1] - \hat{\mathbb{E}}[Y_i | D_i = 0, M_i = 1]$$

- Target the $ATE_{M=1}$ and $ATT_{M=1}$

Bias in the naïve estimator for $ATE_{M=1}$

Under Assumptions 1-4:

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - ATE_{M=1} \\
&\quad = \big(\mathbb{E}[Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\quad - \mathbb{E}[Y_i(1,1) - Y_i(0,0) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\quad \big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1) \Pr(D_i = 1 | M_i = 1) \\
&\qquad - \big(\mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\qquad - \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\qquad \big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)
\end{aligned}
$$

Bias in the naïve estimator for $ATE_{M=1}$

Under Assumptions 1-4:

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - ATE_{M=1} \\
&\quad = \Big(\mathbb{E}[Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\quad - \mathbb{E}[Y_i(1,1) - Y_i(0,0) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\quad \Big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1) \Pr(D_i = 1 | M_i = 1) \\
&\qquad - \Big(\mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\qquad - \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\qquad \Big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)
\end{aligned}
$$

Bias in the naïve estimator for $ATE_{M=1}$

Under Assumptions 1-4:

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - ATE_{M=1} \\
&\quad = \Big(\mathbb{E}[Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\quad - \mathbb{E}[Y_i(1,1) - Y_i(0,0) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\quad \Big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1) \Pr(D_i = 1 | M_i = 1) \\
&\qquad - \Big(\mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\qquad - \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\qquad \Big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)
\end{aligned}
$$

Bias in the naïve estimator for $ATE_{M=1}$

Under Assumptions 1-4:

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - ATE_{M=1} \\
&\quad = \big(\mathbb{E}[Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\quad - \mathbb{E}[Y_i(1,1) - Y_i(0,0) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\quad \big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1) \Pr(D_i = 1 | M_i = 1) \\
&\qquad - \big(\mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\qquad - \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\qquad \big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)
\end{aligned}
$$

Biased even without omitted variables. More

Bias in the naïve estimator for $ATE_{M=1}$

Under Assumptions 1-4:

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - ATE_{M=1} \\
&\quad = \Big(\mathbb{E}[Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\quad - \mathbb{E}[Y_i(1,1) - Y_i(0,0) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\Big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1) \Pr(D_i = 1 | M_i = 1) \\
&\qquad - \Big(\mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\qquad - \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\quad\Big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)
\end{aligned}
$$

Biased even without omitted variables. Bias is always nonpositive.

More

Bias in the naïve estimator for $ATE_{M=1}$

Under Assumptions 1-4:

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - ATE_{M=1} \\
&\quad = \Big(\mathbb{E}[Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\quad - \mathbb{E}[Y_i(1,1) - Y_i(0,0) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\quad \Big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1) \Pr(D_i = 1 | M_i = 1) \\
&\qquad - \Big(\mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\qquad - \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\qquad \Big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)
\end{aligned}
$$

Bias remains unless there are no racial stops. More

Bias in the naïve estimator for $ATE_{M=1}$

Under Assumptions 1-4:

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - ATE_{M=1} \\
&\quad = \big(\mathbb{E}[Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\quad - \mathbb{E}[Y_i(1,1) - Y_i(0,0) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\quad \big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1) \Pr(D_i = 1 | M_i = 1) \\
&\qquad - \big(\mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad\qquad - \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad\qquad \big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)
\end{aligned}
$$

Biased even if goal is to estimate effect *among the stopped*. More

Bias in the naïve estimator for $ATT_{M=1}$

What about $ATT_{M=1}$, the total effect among stopped black civilians?

Bias in the naïve estimator for $ATT_{M=1}$

What about $ATT_{M=1}$, the total effect among stopped black civilians?

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - ATT_{M=1} \\
&\quad = -\mathbb{E}[Y_i(0,1)|M_i(1)=1, M_i(0)=1]\Pr(M_i(0)=0|M_i(1)=1)
\end{aligned}
$$

Bias in the naïve estimator for $ATT_{M=1}$

What about $ATT_{M=1}$, the total effect among stopped black civilians?

$$\begin{aligned}&\mathbb{E}[\hat{\Delta}] - ATT_{M=1} \\ &\quad = -\mathbb{E}[Y_i(0,1)|M_i(1)=1, M_i(0)=1]\Pr(M_i(0)=0|M_i(1)=1)\end{aligned}$$

Again, bias remains unless there are no racial stops

Bias in the naïve estimator for $ATT_{M=1}$

What about $ATT_{M=1}$, the total effect among stopped black civilians?

$$\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - ATT_{M=1} \\
&\quad = -\mathbb{E}[Y_i(0,1)|M_i(1)=1, M_i(0)=1]\Pr(M_i(0)=0|M_i(1)=1)
\end{aligned}$$

Again, bias remains unless there are no racial stops, or no use of force against whites

Bias in the naïve estimator for $ATT_{M=1}$

What about $ATT_{M=1}$, the total effect among stopped black civilians?

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - ATT_{M=1} \\
&\quad = -\mathbb{E}[Y_i(0,1)|M_i(1)=1, M_i(0)=1]\Pr(M_i(0)=0|M_i(1)=1)
\end{aligned}
$$

Again, bias remains unless there are no racial stops, or no use of force against whites (empirically falsified).

What have we learned?

Under assumptions 1-4, which are implicit in prior work, the naïve estimator:

What have we learned?

Under assumptions 1-4, which are implicit in prior work, the naïve estimator:

- is biased for the $ATE_{M=1}$

What have we learned?

Under assumptions 1-4, which are implicit in prior work, the naïve estimator:

- is biased for the $ATE_{M=1}$
- is biased for the $ATT_{M=1}$

What have we learned?

Under assumptions 1-4, which are implicit in prior work, the naïve estimator:

- is biased for the $ATE_{M=1}$
- is biased for the $ATT_{M=1}$
- *unless* we assume no racial stops

What have we learned?

Under assumptions 1-4, which are implicit in prior work, the naïve estimator:

- is biased for the $ATE_{M=1}$
- is biased for the $ATT_{M=1}$
- *unless* we assume no racial stops
- *even with a perfect set of pre-treatment control variables*

What have we learned?

Under assumptions 1-4, which are implicit in prior work, the naïve estimator:

- is biased for the $ATE_{M=1}$
- is biased for the $ATT_{M=1}$
- *unless* we assume no racial stops
- *even with a perfect set of pre-treatment control variables*

Can we estimate racial bias with police administrative data?

Bounds and bias correction

- Using precise form of bias, we can construct nonparametric sharp bounds on true effects

Bounding the true $ATE_{M=1}$

Bounding the true $ATE_{M=1}$

- Bias can be re-written in terms of all things that can be directly estimated from data except two:

Bounding the true $ATE_{M=1}$

- Bias can be re-written in terms of all things that can be directly estimated from data except two:

  1. $\rho = \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)$: share of minority stops due to race

Bounding the true $ATE_{M=1}$

- Bias can be re-written in terms of all things that can be directly estimated from data except two:

  1. $\rho = \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)$: share of minority stops due to race (unknown)

Bounding the true $ATE_{M=1}$

- Bias can be re-written in terms of all things that can be directly estimated from data except two:

  1. $\rho = \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)$: share of minority stops due to race (unknown)
  2. $\theta = \mathbb{E}[Y(1,1) | D_i = 1, M_i(1) = 1, M_i(0) = 0]$, violence rate among racially stopped minorities

Bounding the true $ATE_{M=1}$

- Bias can be re-written in terms of all things that can be directly estimated from data except two:

1. $\rho = \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)$: share of minority stops due to race (unknown)

2. $\theta = \mathbb{E}[Y(1,1) | D_i = 1, M_i(1) = 1, M_i(0) = 0]$, violence rate among racially stopped minorities

    - If we knew the joint distribution $\Pr(Y(1,1), M_i(0) = 0 | D_i = 1, M_i(1) = 1) = \Pr(A, B)$, we could then back out $\theta = P(A|B)$, the conditional probability

Bounding the true $ATE_{M=1}$

- Bias can be re-written in terms of all things that can be directly estimated from data except two:

  1. $\rho = \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)$: share of minority stops due to race (unknown)
  2. $\theta = \mathbb{E}[Y(1,1) | D_i = 1, M_i(1) = 1, M_i(0) = 0]$, violence rate among racially stopped minorities

     - If we knew the joint distribution $\Pr(Y(1,1), M_i(0) = 0 | D_i = 1, M_i(1) = 1) = \Pr(A, B)$, we could then back out $\theta = P(A|B)$, the conditional probability
     - $\theta = P(A|B) = \frac{\Pr(A,B)}{\Pr(B)} = \frac{\Pr(A,B)}{\rho}$

Bounding the true $ATE_{M=1}$

- Bias can be re-written in terms of all things that can be directly estimated from data except two:

  1. $\rho = \Pr(M_i(0) = 0 | D_i = 1, M_i = 1)$: share of minority stops due to race (unknown)

  2. $\theta = \mathbb{E}[Y(1,1) | D_i = 1, M_i(1) = 1, M_i(0) = 0]$, violence rate among racially stopped minorities

     - If we knew the joint distribution $\Pr(Y(1,1), M_i(0) = 0 | D_i = 1, M_i(1) = 1) = \Pr(A, B)$, we could then back out $\theta = P(A|B)$, the conditional probability

     - $\theta = P(A|B) = \frac{\Pr(A,B)}{\Pr(B)} = \frac{\Pr(A,B)}{\rho}$

     - We don't, but we can place Fréchet bounds on $\Pr(A, B)$

Maurice Fréchet

Fréchet Inequalities

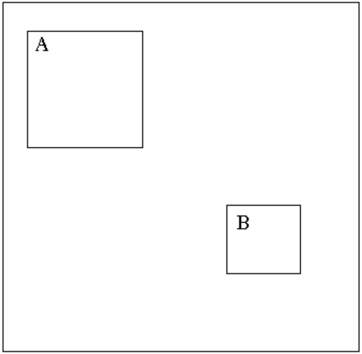
Given two marginal distributions $\Pr(A)$ and $\Pr(B)$, the joint distribution $\Pr(A, B)$ is bounded by:

$$max\{0, Pr(A) + Pr(B) - 1\} \leq P(A, B) \leq min\{Pr(A), Pr(B)\}$$

Fréchet Inequalities

Given two marginal distributions $\Pr(A)$ and $\Pr(B)$, the joint distribution (A, B) is bounded by:

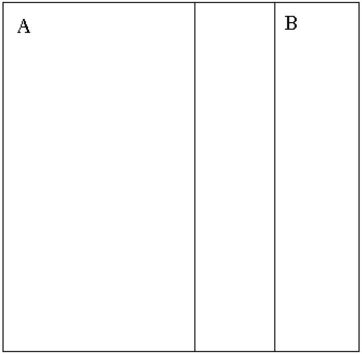
$$max\{0, Pr(A) + Pr(B) - 1\} \leq P(A, B) \leq min\{Pr(A), Pr(B)\}$$



Fréchet Inequalities

Given two marginal distributions $\mathrm{Pr}(A)$ and $\mathrm{Pr}(B)$, the joint distribution (A, B) is bounded by:

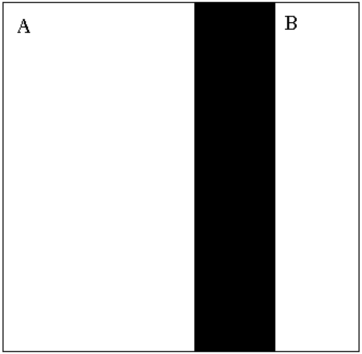
$$max\{0, Pr(A) + Pr(B) - 1\} \leq P(A, B) \leq min\{Pr(A), Pr(B)\}$$



Fréchet Inequalities

Given two marginal distributions $\Pr(A)$ and $\Pr(B)$, the joint distribution (A, B) is bounded by:

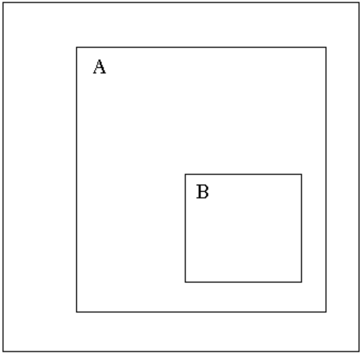
$$\textit{max}\{0, Pr(A) + Pr(B) - 1\} \leq P(A, B) \leq \textit{min}\{Pr(A), Pr(B)\}$$



Fréchet Inequalities

Given two marginal distributions $\Pr(A)$ and $\Pr(B)$, the joint distribution (A, B) is bounded by:

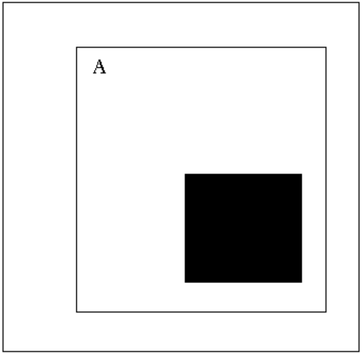
$$max\{0, Pr(A) + Pr(B) - 1\} \leq P(A, B) \leq min\{Pr(A), Pr(B)\}$$



Fréchet Inequalities

Given two marginal distributions $\mathrm{Pr}(A)$ and $\mathrm{Pr}(B)$, the joint distribution (A, B) is bounded by:

$$max\{0, Pr(A) + Pr(B) - 1\} \leq P(A, B) \leq min\{Pr(A), Pr(B)\}$$



Deriving sharp bounds for true $ATE_{M=1}$

- If we can bound $\Pr(A, B)$ we can also bound $\theta = \frac{\Pr(A,B)}{\rho}$ and plug the bounds back into the bias term

Deriving sharp bounds for true $ATE_{M=1}$

- If we can bound $\Pr(A, B)$ we can also bound $\theta = \frac{\Pr(A,B)}{\rho}$ and plug the bounds back into the bias term

- If $ATE_{M=1} = \mathbb{E}[\hat{\Delta}] + bias$, then subbing in Fréchet bounds for θ into the bias term $\implies$

$$\mathbb{E}[\hat{\Delta}] + \underline{bias}_{LB} \leq ATE_{M=1} \leq \mathbb{E}[\hat{\Delta}] + \overline{bias}^{UB}$$

Sharp nonparametric bounds

Given ρ, bounds for the true $ATE_{M=1}$ are given by:

$$\mathbb{E}[\hat{\Delta}] + \rho\ \mathbb{E}[Y_i|D_i = 0, M_i = 1]\ (1 - \Pr(D_i = 0|M_i = 1))$$
$$\leq\ ATE_{M=1}\ \leq$$
$$\mathbb{E}[\hat{\Delta}] + \frac{\rho}{1-\rho}\left(\mathbb{E}[Y_i|D_i = 1, M_i = 1] - K\right)\Pr(D_i = 0|M_i = 1)$$
$$+\rho\ \mathbb{E}[Y_i|D_i = 0, M_i = 1]\ (1 - \Pr(D_i = 0|M_i = 1)).$$

where

$$K = \max\left\{0, 1 + \frac{1}{\rho}\mathbb{E}[Y_i|D_i = 1, M_i = 1] - \frac{1}{\rho}\right\}.$$

Sharp nonparametric bounds

Given ρ, bounds for the true $ATE_{M=1}$ are given by:

$$\mathbb{E}[\hat{\Delta}] + \rho \, \mathbb{E}[Y_i | D_i = 0, M_i = 1] \, (1 - \Pr(D_i = 0 | M_i = 1))$$
$$\leq \; ATE_{M=1} \; \leq$$
$$\mathbb{E}[\hat{\Delta}] + \frac{\rho}{1-\rho} \left(\mathbb{E}[Y_i | D_i = 1, M_i = 1] - K \right) \Pr(D_i = 0 | M_i = 1)$$
$$+ \rho \, \mathbb{E}[Y_i | D_i = 0, M_i = 1] \, (1 - \Pr(D_i = 0 | M_i = 1)).$$

where

$$K = \max \left\{ 0, 1 + \frac{1}{\rho} \mathbb{E}[Y_i | D_i = 1, M_i = 1] - \frac{1}{\rho} \right\}.$$

The $ATT_{M=1}$ must similarly satisfy:

$$ATT_{M=1} = \mathbb{E}[\hat{\Delta}] + \rho \, \mathbb{E}[Y_i | D_i = 0, M_i = 1]$$

Does this matter in practice?

Replication and Extension: Fryer (2019)

Fryer (2019)

- Police-civilian interactions (e.g. Stop and Frisk, arrest records, summaries of shootings)

Fryer (2019)

- Police-civilian interactions (e.g. Stop and Frisk, arrest records, summaries of shootings)
- Logistic regressions of force measures on race dummies, controls for circumstance, suspect features, officer features

Fryer (2019)

- Police-civilian interactions (e.g. Stop and Frisk, arrest records, summaries of shootings)
- Logistic regressions of force measures on race dummies, controls for circumstance, suspect features, officer features
- Conclusions:
    - Some racial bias in sub-lethal force

Fryer (2019)

- Police-civilian interactions (e.g. Stop and Frisk, arrest records, summaries of shootings)
- Logistic regressions of force measures on race dummies, controls for circumstance, suspect features, officer features
- Conclusions:
  - Some racial bias in sub-lethal force
  - No bias in lethal force

Fryer (2019)

- Police-civilian interactions (e.g. Stop and Frisk, arrest records, summaries of shootings)
- Logistic regressions of force measures on race dummies, controls for circumstance, suspect features, officer features
- Conclusions:
  - Some racial bias in sub-lethal force
  - No bias in lethal force
- Problem: No data on those police observe but do not stop

Concern over post-treatment bias

VOL. 2 NO. ISSUE *RACIAL DIFFER*

encounter recounted by police. A second type of bias is that officers may be more likely to charge black suspects with crimes such as resisting arrest or attempted assault on a public safety officer rather than misdemeanors, relative to whites, for identical behavior. This type of bias is an important limitation of Fryer (*forthcoming*) because it implies that the counterfactuals coded from arrest data may themselves contain bias. It is unclear how to estimate the extent of such bias or how to address it statistically.

Replication of Fryer (2019)

- Replicate two analyses of sub-lethal force using NYPD's "Stop, Question and Frisk" (SQF) data (2003-2013), $N \approx 5$ million

Replication of Fryer (2019)

- Replicate two analyses of sub-lethal force using NYPD's "Stop, Question and Frisk" (SQF) data (2003-2013), $N \approx 5$ million
- Stipulate to regression model (logit form, assume no omitted variables)

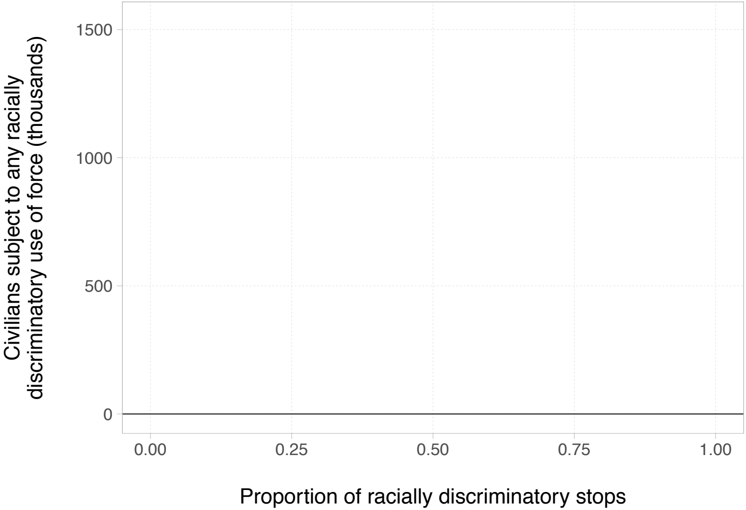
Replication of Fryer (2019)

- Replicate two analyses of sub-lethal force using NYPD's "Stop, Question and Frisk" (SQF) data (2003-2013), $N \approx 5$ million
- Stipulate to regression model (logit form, assume no omitted variables)
  - Use of any force (at least laying hands on civilian; binary)

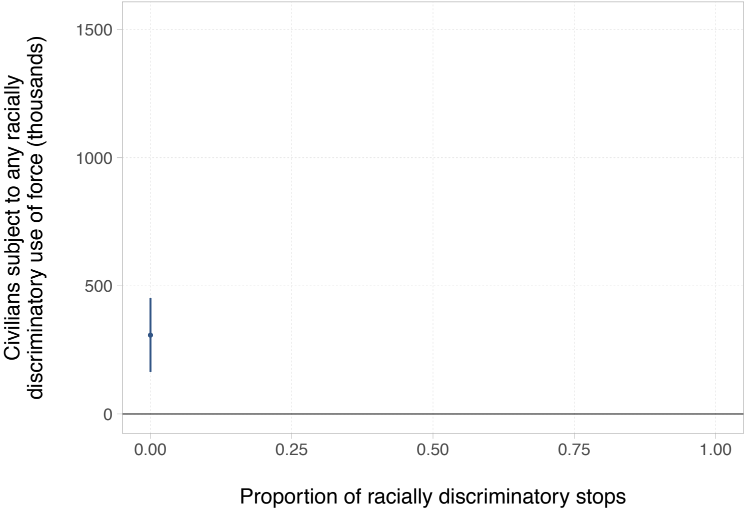
Replication of Fryer (2019)

- Replicate two analyses of sub-lethal force using NYPD's "Stop, Question and Frisk" (SQF) data (2003-2013), $N \approx 5$ million
- Stipulate to regression model (logit form, assume no omitted variables)
  - Use of any force (at least laying hands on civilian; binary)
  - Force thresholds (e.g. at least handcuffs) with seven categories: laying hands; push to wall; handcuffs; draw weapon; push to ground; point weapon; baton/pepper spray

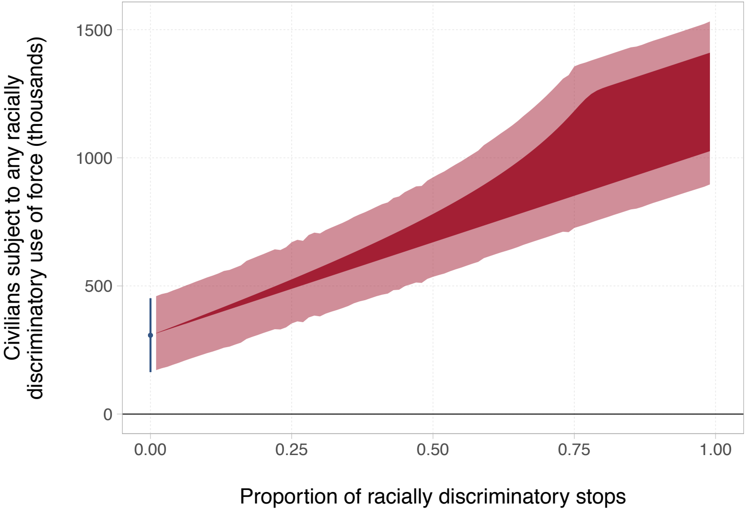
Bounds on race effects, black vs. white



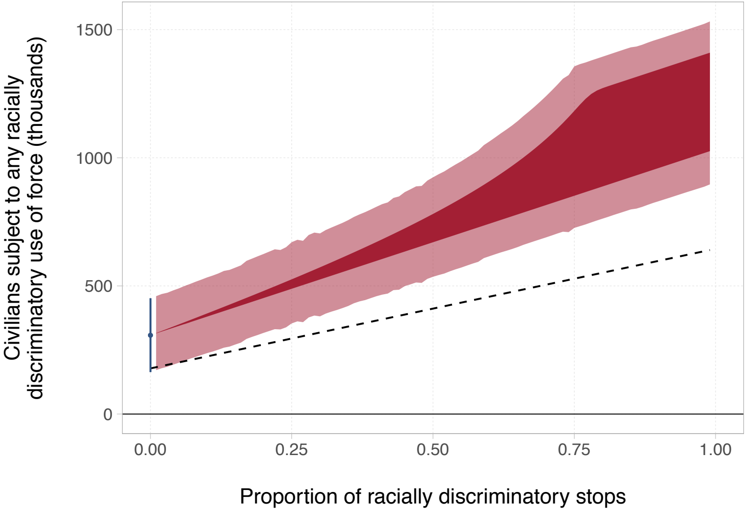
Bounds on race effects, black vs. white



Bounds on race effects, black vs. white



Bounds on race effects, black vs. white



What is ρ?

What is the share of minority stops that would not have happened if civilians had been white?

What is ρ?

What is the share of minority stops that would not have happened if civilians had been white?

- Can be anywhere in $[0, 1)$. If $\rho = 0$, bias disappears.

What is ρ?

What is the share of minority stops that would not have happened if civilians had been white?

- Can be anywhere in $[0, 1)$. If $\rho = 0$, bias disappears.
- Two prior studies estimate this using data on "Stop, Question and Frisk" in NYC

What is ρ?

What is the share of minority stops that would not have happened if civilians had been white?

- Can be anywhere in $[0, 1)$. If $\rho = 0$, bias disappears.
- Two prior studies estimate this using data on "Stop, Question and Frisk" in NYC
- Gelman, Fagan & Kiss (2007) and Goel, Rao and Schroff (2016)

What is ρ?

What is the share of minority stops that would not have happened if civilians had been white?

- Can be anywhere in $[0, 1)$. If $\rho = 0$, bias disappears.
- Two prior studies estimate this using data on "Stop, Question and Frisk" in NYC
- Gelman, Fagan & Kiss (2007) and Goel, Rao and Schroff (2016)
- Studies take totally different approaches

What is ρ?

What is the share of minority stops that would not have happened if civilians had been white?

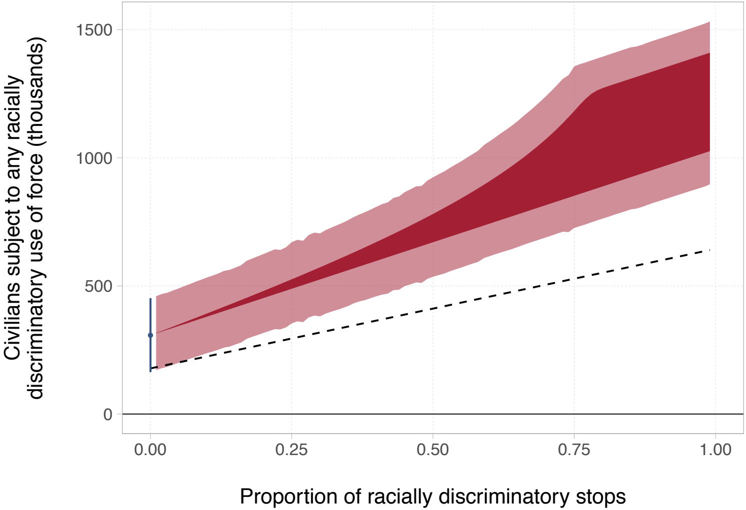
- Can be anywhere in $[0, 1)$. If $\rho = 0$, bias disappears.
- Two prior studies estimate this using data on "Stop, Question and Frisk" in NYC
- Gelman, Fagan & Kiss (2007) and Goel, Rao and Schroff (2016)
- Studies take totally different approaches
- Results imply ρ is at least .32 or .34, respectively

What is ρ?

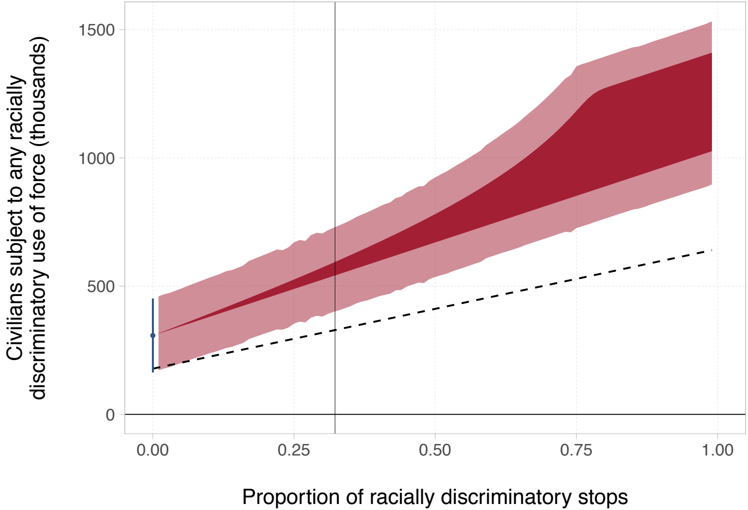
What is the share of minority stops that would not have happened if civilians had been white?

- Can be anywhere in $[0, 1)$. If $\rho = 0$, bias disappears.
- Two prior studies estimate this using data on "Stop, Question and Frisk" in NYC
- Gelman, Fagan & Kiss (2007) and Goel, Rao and Schroff (2016)
- Studies take totally different approaches
- Results imply ρ is at least .32 or .34, respectively
- We use $\rho = .32$ to be conservative

Bounds on race effects, black vs. white



Bounds on race effects, black vs. white



Bounds for force thresholds, black vs. white

	TE_S for encounters with black civilians (vs. white)			
	No covariates		Full specification	
Minimum force	bounds	naïve	bounds	naïve
Use of hands	**(112.66, 124.59)**	**61.69**	**(86.99, 96.74)**	**23.53**
	(84.6, 151.84)	*(32.89, 90.63)*	*(81.7, 102.15)*	*(16.41, 30.61)*
Push to wall	**(24.15, 27.75)**	**4.2**	**(26.48, 30.21)**	**6.67**
	(15.5, 37.35)	*(-5.29, 14.02)*	*(24.29, 32.38)*	*(3.73, 9.52)*
Use of handcuffs	**(14.6, 16.92)**	**1.32**	**(16.56, 19.02)**	**3.9**
	(9.45, 22.61)	*(-4.83, 7.53)*	*(15.05, 20.55)*	*(1.87, 5.88)*
Draw weapon	**(4.52, 5.14)**	**1.26**	**(4.71, 5.35)**	**1.46**
	(3.13, 6.67)	*(-0.33, 2.83)*	*(4.22, 5.86)*	*(0.79, 2.13)*
Push to ground	**(4.04, 4.58)**	**1.22**	**(4.11, 4.66)**	**1.26**
	(2.79, 5.97)	*(-0.21, 2.66)*	*(3.68, 5.09)*	*(0.68, 1.82)*
Point weapon	**(1.49, 1.7)**	**0.36**	**(1.64, 1.86)**	**0.55**
	(0.96, 2.29)	*(-0.29, 1)*	*(1.37, 2.13)*	*(0.18, 0.91)*
Baton or pepper spray	**(0.17, 0.19)**	**0.08**	**(0.17, 0.19)**	**0.07**
	(0.1, 0.26)	*(-0.01, 0.15)*	*(0.12, 0.24)*	*(-0.01, 0.14)*

Bounds for force thresholds, black vs. white

	TE_S for encounters with black civilians (vs. white)			
	No covariates		Full specification	
Minimum force	bounds	naïve	bounds	naïve
Use of hands	**(112.66, 124.59)**	**61.69**	**(86.99, 96.74)**	**23.53**
	(84.6, 151.84)	*(32.89, 90.63)*	*(81.7, 102.15)*	*(16.41, 30.61)*
Push to wall	**(24.15, 27.75)**	**4.2**	**(26.48, 30.21)**	**6.67**
	(15.5, 37.35)	*(-5.29, 14.02)*	*(24.29, 32.38)*	*(3.73, 9.52)*
Use of handcuffs	**(14.6, 16.92)**	**1.32**	**(16.56, 19.02)**	**3.9**
	(9.45, 22.61)	*(-4.83, 7.53)*	*(15.05, 20.55)*	*(1.87, 5.88)*
Draw weapon	**(4.52, 5.14)**	**1.26**	**(4.71, 5.35)**	**1.46**
	(3.13, 6.67)	*(-0.33, 2.83)*	*(4.22, 5.86)*	*(0.79, 2.13)*
Push to ground	**(4.04, 4.58)**	**1.22**	**(4.11, 4.66)**	**1.26**
	(2.79, 5.97)	*(-0.21, 2.66)*	*(3.68, 5.09)*	*(0.68, 1.82)*
Point weapon	**(1.49, 1.7)**	**0.36**	**(1.64, 1.86)**	**0.55**
	(0.96, 2.29)	*(-0.29, 1)*	*(1.37, 2.13)*	*(0.18, 0.91)*
Baton or pepper spray	**(0.17, 0.19)**	**0.08**	**(0.17, 0.19)**	**0.07**
	(0.1, 0.26)	*(-0.01, 0.15)*	*(0.12, 0.24)*	*(-0.01, 0.14)*

Bounds for force thresholds, black vs. white

	TE_S for encounters with black civilians (vs. white)			
	No covariates		Full specification	
Minimum force	bounds	naïve	bounds	naïve
Use of hands	**(112.66, 124.59)**	**61.69**	**(86.99, 96.74)**	**23.53**
	(84.6, 151.84)	*(32.89, 90.63)*	*(81.7, 102.15)*	*(16.41, 30.61)*
Push to wall	**(24.15, 27.75)**	**4.2**	**(26.48, 30.21)**	**6.67**
	(15.5, 37.35)	*(-5.29, 14.02)*	*(24.29, 32.38)*	*(3.73, 9.52)*
Use of handcuffs	**(14.6, 16.92)**	**1.32**	**(16.56, 19.02)**	**3.9**
	(9.45, 22.61)	*(-4.83, 7.53)*	*(15.05, 20.55)*	*(1.87, 5.88)*
Draw weapon	**(4.52, 5.14)**	**1.26**	**(4.71, 5.35)**	**1.46**
	(3.13, 6.67)	*(-0.33, 2.83)*	*(4.22, 5.86)*	*(0.79, 2.13)*
Push to ground	**(4.04, 4.58)**	**1.22**	**(4.11, 4.66)**	**1.26**
	(2.79, 5.97)	*(-0.21, 2.66)*	*(3.68, 5.09)*	*(0.68, 1.82)*
Point weapon	**(1.49, 1.7)**	**0.36**	**(1.64, 1.86)**	**0.55**
	(0.96, 2.29)	*(-0.29, 1)*	*(1.37, 2.13)*	*(0.18, 0.91)*
Baton or pepper spray	**(0.17, 0.19)**	**0.08**	**(0.17, 0.19)**	**0.07**
	(0.1, 0.26)	*(-0.01, 0.15)*	*(0.12, 0.24)*	*(-0.01, 0.14)*

Bounds for force thresholds, black vs. white

	TE$_S$ for encounters with black civilians (vs. white)			
	No covariates		Full specification	
Minimum force	bounds	naïve	bounds	naïve
Use of hands	**(112.66, 124.59)**	**61.69**	**(86.99, 96.74)**	**23.53**
	(84.6, 151.84)	*(32.89, 90.63)*	*(81.7, 102.15)*	*(16.41, 30.61)*
Push to wall	**(24.15, 27.75)**	**4.2**	**(26.48, 30.21)**	**6.67**
	(15.5, 37.35)	*(-5.29, 14.02)*	*(24.29, 32.38)*	*(3.73, 9.52)*
Use of handcuffs	**(14.6, 16.92)**	**1.32**	**(16.56, 19.02)**	**3.9**
	(9.45, 22.61)	*(-4.83, 7.53)*	*(15.05, 20.55)*	*(1.87, 5.88)*
Draw weapon	**(4.52, 5.14)**	**1.26**	**(4.71, 5.35)**	**1.46**
	(3.13, 6.67)	*(-0.33, 2.83)*	*(4.22, 5.86)*	*(0.79, 2.13)*
Push to ground	**(4.04, 4.58)**	**1.22**	**(4.11, 4.66)**	**1.26**
	(2.79, 5.97)	*(-0.21, 2.66)*	*(3.68, 5.09)*	*(0.68, 1.82)*
Point weapon	**(1.49, 1.7)**	**0.36**	**(1.64, 1.86)**	**0.55**
	(0.96, 2.29)	*(-0.29, 1)*	*(1.37, 2.13)*	*(0.18, 0.91)*
Baton or pepper spray	**(0.17, 0.19)**	**0.08**	**(0.17, 0.19)**	**0.07**
	(0.1, 0.26)	*(-0.01, 0.15)*	*(0.12, 0.24)*	*(-0.01, 0.14)*

How can we do better?

- Only partially identified.

How can we do better?

- Only partially identified. Can't get the population ATE.

How can we do better?

- Only partially identified. Can't get the population ATE.
- Only way to do better: improved research design

Option 1:

- Identify situations with **race-blind contact with police** (e.g. rules for DUI stops; traffic stops and night; traffic accidents?)

Option 2: Gather data on the non-stopped

- Need data on those police observe but do not stop

Option 2: Gather data on the non-stopped

- Need data on those police observe but do not stop

- Answer: traffic cameras.

Option 2: Gather data on the non-stopped

- Need data on those police observe but do not stop
- Answer: traffic cameras.
  Passive data collection on non-stop encounters.

Option 2: Gather data on the non-stopped

- Need data on those police observe but do not stop
- Answer: traffic cameras.
  Passive data collection on non-stop encounters.
- Link cars to DMV records, ticket/arrest data

Option 2: Gather data on the non-stopped

- Need data on those police observe but do not stop

- Answer: traffic cameras.
  Passive data collection on non-stop encounters.

- Link cars to DMV records, ticket/arrest data

- Still have to contend with omitted variables

Option 2: Gather data on the non-stopped

- Need data on those police observe but do not stop
- Answer: traffic cameras.
  Passive data collection on non-stop encounters.
- Link cars to DMV records, ticket/arrest data
- Still have to contend with omitted variables
- But, plausible to measure most (all?) observable covariates available to officer when making stop

Option 2: Gather data on the non-stopped

- Need data on those police observe but do not stop
- Answer: traffic cameras.
  Passive data collection on non-stop encounters.
- Link cars to DMV records, ticket/arrest data
- Still have to contend with omitted variables
- But, plausible to measure most (all?) observable covariates available to officer when making stop
- No need to condition on being stopped during analysis

Option 2: Gather data on the non-stopped

- Need data on those police observe but do not stop
- Answer: traffic cameras.
  Passive data collection on non-stop encounters.
- Link cars to DMV records, ticket/arrest data
- Still have to contend with omitted variables
- But, plausible to measure most (all?) observable covariates available to officer when making stop
- No need to condition on being stopped during analysis
- Post-treatment conditioning avoided by design

Police data mask racially biased policing

- Lots of new/big data on policing $\rightarrow$ raft of studies estimating racial bias

Police data mask racially biased policing

- Lots of new/big data on policing $\rightarrow$ raft of studies estimating racial bias
- At present, inadequate theory: insufficient attention to role of race *throughout entire process* risks severely understating racial violence

Police data mask racially biased policing

- Lots of new/big data on policing $\rightarrow$ raft of studies estimating racial bias
- At present, inadequate theory: insufficient attention to role of race *throughout entire process* risks severely understating racial violence
- Risk confusing/misleading the public and policymakers

Thanks!

Please send feedback to:

dcknox@princeton.edu

wlowe@princeton.edu

jmummolo@princeton.edu