A Note on Posttreatment Selection in Studying Racial Discrimination in Policing

Qingyuan Zhao, Luke Keele, Dylan Small and Marshall Joffe

Presented by Shreya Prakash
STAT 572

Roadmap

- Problem Background
- Proposal
- Methods
- Results
- Extensions
- Conclusions

Roadmap

- Problem Background
- Proposal
- Methods
- Results
- Extensions
- Conclusions

Problem Motivation

- Policing in America has been characterized by immense racial disparities and high-profile incidents of using excessive use of force against minorities
- Previous research on policing has produced **contradictory** and **misleading** results
- Example: Fryer 2019 claims there's racial bias in sublethal force but not lethal force but Johnson et al. 2019 concludes no anti-minority bias in lethal force

Problem Motivation

- Previous methods often use administrative data on police stops/detainments, resulting in *posttreatment selection bias*
- Fail to *formalize assumptions* and implicitly make causal quantities without explicitly stating estimand of interest

Problem Motivation

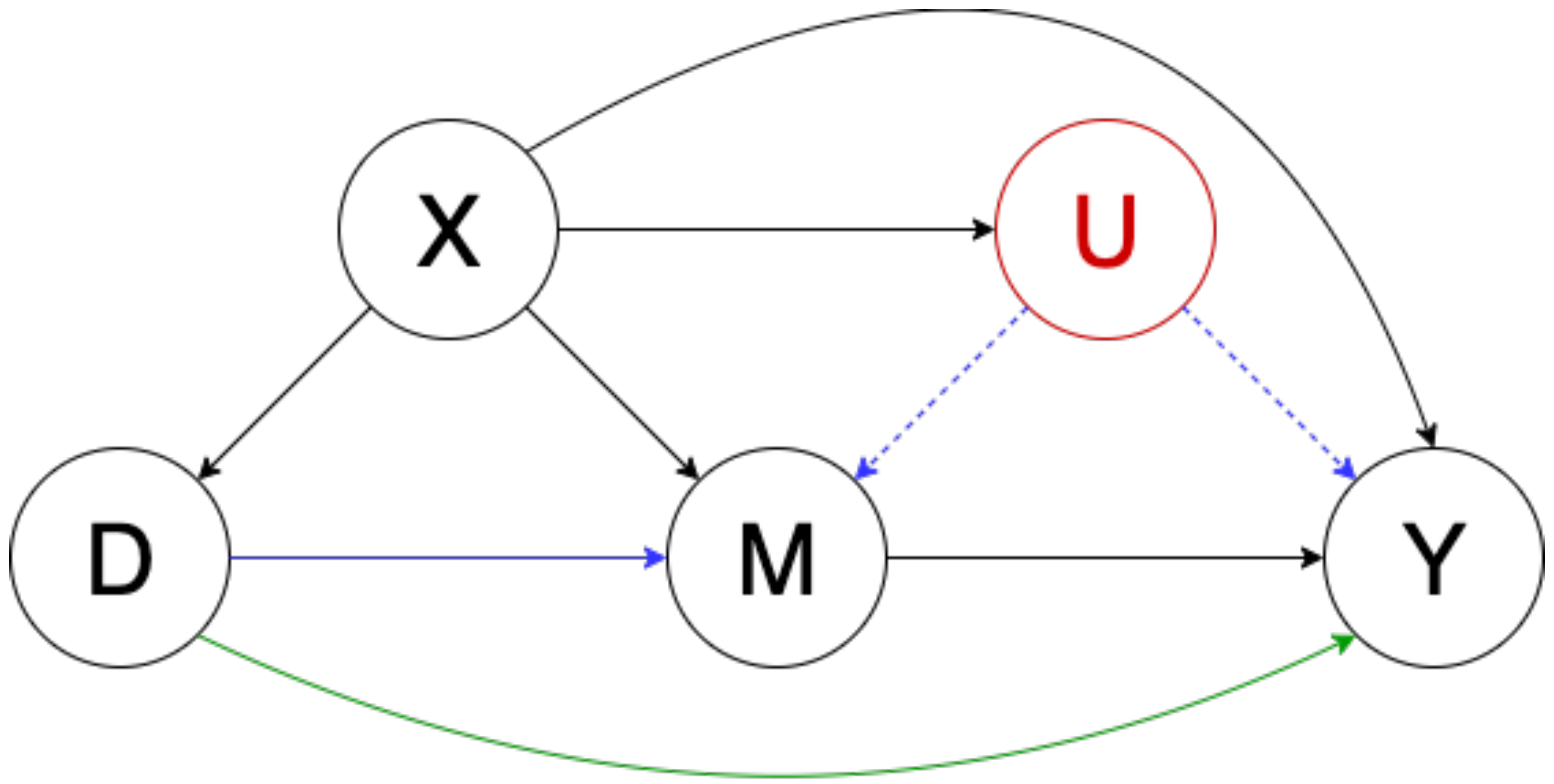
- Knox, Lowe, Mummolo (KLM)
  1. Makes causal quantities of interest explicit
  2. Addresses posttreatment selection bias
- Aim to *comment* and *improve* upon the methods used by KLM

Background: Posttreatment Selection Bias

- Posttreatment selection bias here is the bias that results from studying racial discrimination using *records that are the product of racial discrimination*

Background: Posttreatment Selection Bias

- D is treatment race
- Use M_i to the indicate the mediator, a police detainment or a stop of civilian
- Y_i characterizes the outcome, police use of force
- X_i is the collection of covariates that describe the aspects of the encounter
- U_i represents the unobservable subjective aspects of the encounter like the officer's suspicion or sense of threat



Background: Causal Problem Setup

- Define the **counterfactual** as an encounter with comparable person who is participating in comparable behavior but is of a different race
- Police detainment is **mediator**: the process through which race causes a police officer's use of force in an encounter
- Introduce potential outcomes for M_i and Y_i: $M_i(d), Y_i(d), Y_i(d, m)$

Variable Reminders:
D: Race
M: Police Stops/Detainments
Y: Police use of Force
X: Covariates
U: Unobserved Variables

Previous Literature (KLM)

- Partial identification of local causal estimands

$$ATE_{M=1} = E[Y(1) - Y(0) \mid M = 1]$$

$$ATT_{M=1} = E[Y(1) - Y(0) \mid M = 1, D = 1]$$

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Previous Literature (KLM)

- Key assumptions used:

    1. Mandatory Reporting: $Y(0,0) = Y(1,0) = 0$ and all police stops recorded

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Previous Literature (KLM)

- Key assumptions used:

1. Mandatory Reporting: $Y(0,0) = Y(1,0) = 0$ and all police stops recorded

2. Treatment Ignorability: $D \perp Y(d,1) \,|\, M(d), X$ and $M(d) \perp D \,|\, X$

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Previous Literature (KLM)

- Key assumptions used:

1. Mandatory Reporting: $Y(0,0) = Y(1,0) = 0$ and all police stops recorded

2. Treatment Ignorability: $D \perp Y(d,1) \,|\, M(d), X$ and $M(d) \perp D \,|\, X$

3. Mediator Monotonicity : $M(1) \geq M(0)$, but reverse is never true

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Previous Literature (KLM)

- Key assumptions used:

1. Mandatory Reporting: $Y(0,0) = Y(1,0) = 0$ and all police stops recorded
2. Treatment Ignorability: $D \perp Y(d,1) \,|\, M(d), X$ and $M(d) \perp D \,|\, X$
3. Mediator Monotonicity : $M(1) \geq M(0)$, but reverse is never true
4. Relative Nonseverity of Racial Stops:
   $E[Y(d,m) \,|\, D = d', M(1) = 1, M(0) = 1, X = x] \geq E[Y(d,m) \,|\, D = d', M(1) = 1, M(0) = 0, X = x]$

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Previous Literature (KLM)

- Key assumptions used:

1. Mandatory Reporting: $Y(0,0) = Y(1,0) = 0$ and all police stops recorded

2. Treatment Ignorability: $D \perp Y(d,1) \,|\, M(d), X$ and $M(d) \perp D \,|\, X$

3. Mediator Monotonicity : $M(1) \geq M(0)$, but reverse is never true

4. Relative Nonseverity of Racial Stops:
   $E[Y(d,m) \,|\, D = d', M(1) = 1, M(0) = 1, X = x] \geq E[Y(d,m) \,|\, D = d', M(1) = 1, M(0) = 0, X = x]$

- Identification of $ATE = E[Y(1) - Y(0)]$

Roadmap

- Problem Background
- Proposal
- Methods
- Results
- Extensions
- Conclusions

Proposal

- Shows that KLM's local causal estimands cannot give any information about the global estimands

- Introduces the global causal risk ratio (CRR) which helps identify the global causal effect

$$CRR(x) = \frac{E[Y(1) \mid X = x]}{E[Y(0) \mid X = x]}$$

Roadmap

- Problem Background
- Proposal
- Methods
- Results
- Extensions
- Conclusions

Local vs Global Estimands Setup

- Global estimand helps understand when an unreported <u>white</u> encounter would have **escalated to a stop** if the individual was a <u>minority</u>

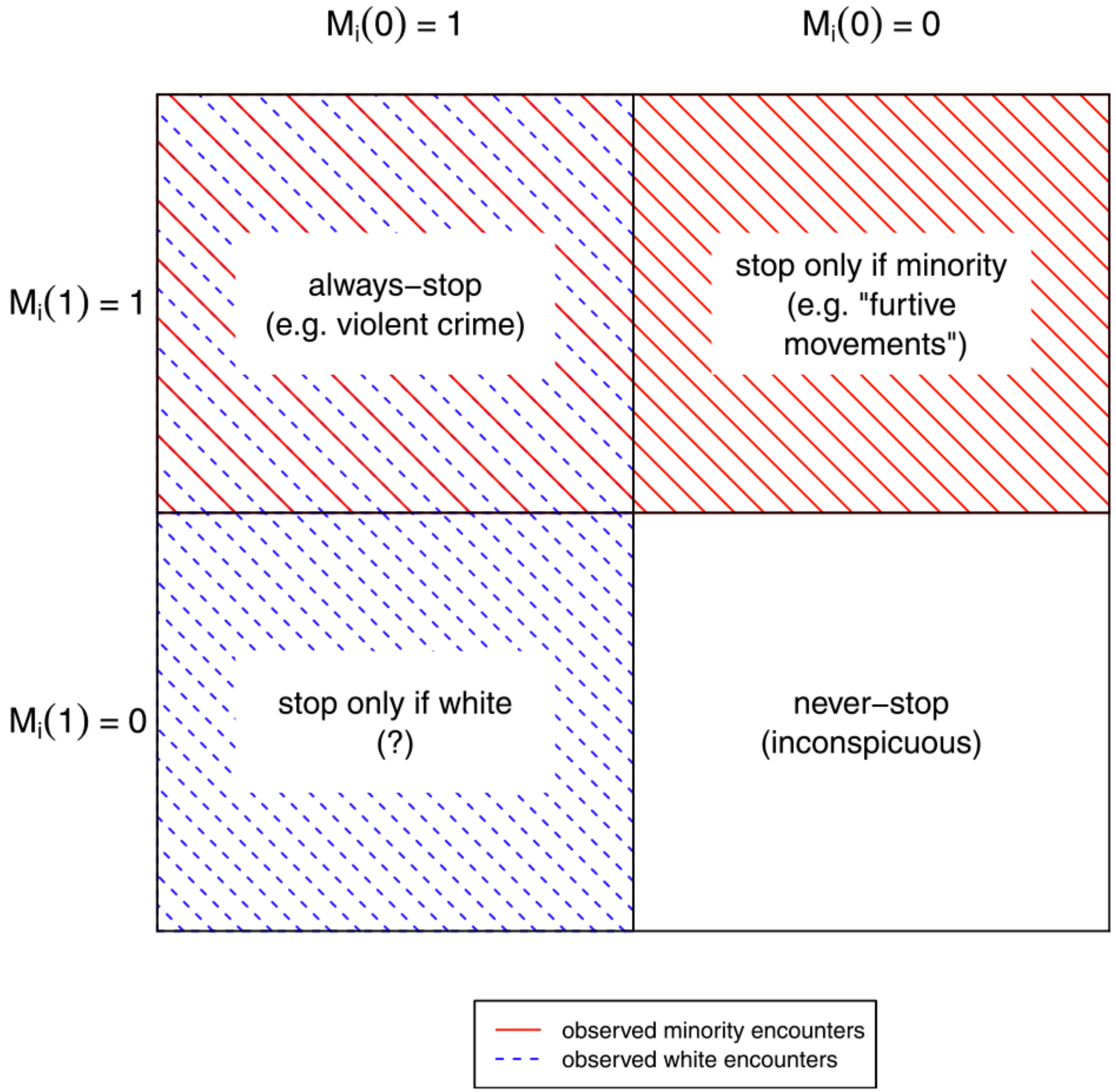
- In mediation analysis often break down the ATE into the pure indirect effect (PIE) and pure direct effect (PDE):

$$ATE = E[Y(1) - Y(0)] = \underbrace{E[Y(1,M(1)) - Y(1,M(0))]}_{\text{PIE}} + \underbrace{E[Y(1,M(0)) + Y(0,M(0))]}_{\text{PDE}}$$

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Local vs Global Estimands Setup

- Introduce principal stratum:



Local vs Global Estimands Assumptions

Using the assumptions:

1. Variables (D,M,X,Y) are generated from a nonparametric structural equation model (SEM):

$$X = f_X(\epsilon_X), D = f_D(X, \epsilon_D), M = f_M(X, D, \epsilon_M), Y = f_Y(X, D, M, \epsilon_Y)$$

where $\epsilon_X, \epsilon_D, \epsilon_M, \epsilon_Y$ are mutually independent

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Local vs Global Estimands Assumptions

Using the assumptions:

1. Variables (D,M,X,Y) are generated from a nonparametric structural equation model (SEM)

2. Mandatory Reporting

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Local vs Global Estimands Assumptions

Using the assumptions:

1. Variables (D,M,X,Y) are generated from a nonparametric structural equation model (SEM)

2. Mandatory Reporting

Show that

$$PIE = \beta_M \times E[(Y(1,1)]$$

$$PDE = \beta_Y \times E[M(0)]$$

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Local vs Global Estimands

- Shows that $ATE_{M=1}$ and $ATT_{M=1}$ may have a different sign even if the PDE and PIE have the same sign

- Local effects do not give any intuition about the global effects

Global Causal Risk Ratio Motivation

- KLM notes that ATE estimation requires estimating the magnitude of $P(M = 1)$

$$ATE = E[Y \mid D = 1, M = 1]P(M = 1 \mid D = 1) - E[Y \mid D = 0, M = 1]P(M = 1 \mid D = 0)$$

- Can be avoided by using a ratio

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Global Causal Risk Ratio

- Introduce the causal risk ratio (CRR) at the covariate level x,

$$CRR(x) = \frac{E[Y(1) \mid X = x]}{E[Y(0) \mid X = x]}$$

- Compare to the naive risk ratio which has posttreatment selection bias

$$NaiveRR(x) = \frac{E[Y \mid D = 1, M = 1, X = x]}{E[Y \mid D = 0, M = 1, X = x]}$$

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Identification of Causal Risk Ratio

- Assumes treatment ignorability to identify the CRR

$$CRR(x) = \frac{E[Y(1) \mid X = x]}{E[Y(0) \mid X = x]} = \underbrace{\frac{E[Y \mid D = 1, M = 1, X = x]}{E[Y \mid D = 0, M = 1, X = x]}}_{\text{naive risk ratio}} \times \underbrace{\frac{\dfrac{P[D = 1 \mid M = 1, X = x]}{P[D = 0 \mid M = 1, X = x]}}{\dfrac{P[D = 1 \mid X = x]}{P[D = 0 \mid X = x]}}}_{\text{bias factor}}$$

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Identification of Causal Risk Ratio

$$CRR(x) = \frac{E[Y(1) \mid X = x]}{E[Y(0) \mid X = x]} = \underbrace{\frac{E[Y \mid D = 1, M = 1, X = x]}{E[Y \mid D = 0, M = 1, X = x]}}_{\text{naive risk ratio}} \times \underbrace{\frac{\frac{P[D = 1 \mid M = 1, X = x]}{P[D = 0 \mid M = 1, X = x]}}{\frac{P[D = 1 \mid X = x]}{P[D = 0 \mid X = x]}}}_{\text{bias factor}}$$

1. Naive Risk Ratio:

$$\frac{E[Y \mid D = 1, M = 1, X = x]}{E[Y \mid D = 0, M = 1, X = x]}$$

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Identification of Causal Risk Ratio

$$CRR(x) = \frac{E[Y(1) \mid X = x]}{E[Y(0) \mid X = x]} = \underbrace{\frac{E[Y \mid D = 1, M = 1, X = x]}{E[Y \mid D = 0, M = 1, X = x]}}_{\text{naive risk ratio}} \times \underbrace{\frac{\frac{P[D = 1 \mid M = 1, X = x]}{P[D = 0 \mid M = 1, X = x]}}{\frac{P[D = 1 \mid X = x]}{P[D = 0 \mid X = x]}}}_{\text{bias factor}}$$

2. Bias factor numerator:

$$\frac{P[D = 1 \mid M = 1, X = x]}{P[D = 0 \mid M = 1, X = x]}$$

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Identification of Causal Risk Ratio

$$CRR(x) = \frac{E[Y(1) \mid X = x]}{E[Y(0) \mid X = x]} = \underbrace{\frac{E[Y \mid D = 1, M = 1, X = x]}{E[Y \mid D = 0, M = 1, X = x]}}_{\text{naive risk ratio}} \times \underbrace{\frac{\frac{P[D = 1 \mid M = 1, X = x]}{P[D = 0 \mid M = 1, X = x]}}{\frac{P[D = 1 \mid X = x]}{P[D = 0 \mid X = x]}}}_{\text{bias factor}}$$

3. Bias factor denominator:

$$\frac{P[D = 1 \mid X = x]}{P[D = 0 \mid X = x]}$$

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Identification of Causal Risk Ratio

$$CRR(x) = \frac{E[Y(1) \mid X = x]}{E[Y(0) \mid X = x]} = \underbrace{\frac{E[Y \mid D = 1, M = 1, X = x]}{E[Y \mid D = 0, M = 1, X = x]}}_{\text{naive risk ratio}} \times \underbrace{\frac{\frac{P[D = 1 \mid M = 1, X = x]}{P[D = 0 \mid M = 1, X = x]}}{\frac{P[D = 1 \mid X = x]}{P[D = 0 \mid X = x]}}}_{\text{bias factor}}$$

4. Bias factor :

$$\frac{\frac{P[D = 1 \mid M = 1, X = x]}{P[D = 0 \mid M = 1, X = x]}}{\frac{P[D = 1 \mid X = x]}{P[D = 0 \mid X = x]}}$$

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Global Causal Risk Ratio

- If assume stochastic mediator monotonicity:

$$E[M(1) \mid X = x] \geq E[M(0) \mid X = x]$$

- Naive risk ratio provides a **lower bound** for the causal risk ratio

Variable Reminders:
D: Race
M: Police Stops/Detaiments
Y: Police use of force
X: covariates
U: unobserved variables

Problems

1. Need two data sources for estimates
2. Conditional on covariates X

Problems

1. Need two data sources for estimates
2. Conditional on covariates X

Problems

1. Need two data sources for estimates

2. Conditional on covariates X

Roadmap

- Problem Background
- Proposal
- Methods
- Results
- Extensions
- Conclusions

Analysis of NYPD Stop and Frisk Data

- Use 2 different supplementary data source to estimate the bias factor:

    1. Current Population Survey (CPS)

    2. Police-Public Contact Survey (PPCS)

Estimates of the Causal Effect of Minority Race (Black) on Police Violence		
External Dataset	**Estimated Risk Ratio**	**95% CI**
None (Naive Estimator)	**1.291**	(1.284-1.299)
CPS	13.566	(12.812-14.375)
PPCS	32.300	(31.289-33.402)
PPCS (MV Stop)	29.549	(26.726-32.903)
PPCS (Other Stop)	29.241	(23.446-37.201)
PPCS (Large Metro)	16.688	(15.237-18.180)
PPCS*	31.131	(28.203-34.736)
PPCS* (Large Metro)	19.873	(14.147-28.607)

$\geq 10 \times$ **Naive**

Experiment Results

- Conduct stratified analysis by age and gender
  - <u>Female</u> minorities likely to have **smaller** risk ratio than <u>male</u> minorities
- Used census data to get risk ratio by precinct
  - Found that CRR is **much larger** than NaiveRR in most cases

Roadmap

- Problem Background
- Proposal
- Methods
- Results
- Extensions
- Conclusions

Extensions: Local vs Global Estimands

- If one assumes that $D = 1$ is in fact the minority group, then we have that $\beta_Y < 0, \beta_M < 0$ implies that $ATE_{M=1} < 0$

- If one assumes mediator monotonicity, then we have that the sign of the local estimands $ATE_{M=1}, ATT_{M=1}$ are consistent with the sign of the global estimand.

Extensions: Other Global Estimands

- Can better understand the gravity of the anti-minority discrimination using the following estimand:

$$P(Y(D=1)=1 \mid Y(D=0)=0)$$

- Can get non-parametric sharp bounds on this quantity using algorithm presented by Duarte, Finkelstein, Knox, et al. (2021)

Roadmap

- Problem Background
- Proposal
- Methods
- Results
- Extensions
- Conclusions

Conclusions

- A local causal estimator for ATE conditioned on the administrative data does not tell you anything about the global ATE

Conclusions

- A local causal estimator for ATE conditioned on the administrative data does not tell you anything about the global ATE

- Introduced a global causal risk ratio and used in practice the NYPD Stop and Frisk dataset

Conclusions

- A local causal estimator for ATE conditioned on the administrative data does not tell you anything about the global ATE

- Introduced a global causal risk ratio and used in practice the NYPD Stop and Frisk dataset

- Some limitations that came from their estimator is that it is <u>*hard to find data*</u> to properly estimate the bias term and hard to <u>*condition on multiple confounders*</u> in practice

Conclusions

- A local causal estimator for ATE conditioned on the administrative data does not tell you anything about the global ATE
- Introduced a global causal risk ratio and used in practice the NYPD Stop and Frisk dataset
- Some limitations that came from their estimator is that it is *hard to find data* to properly estimate the bias term and hard to *condition on multiple confounders* in practice
- Additional data can help further research

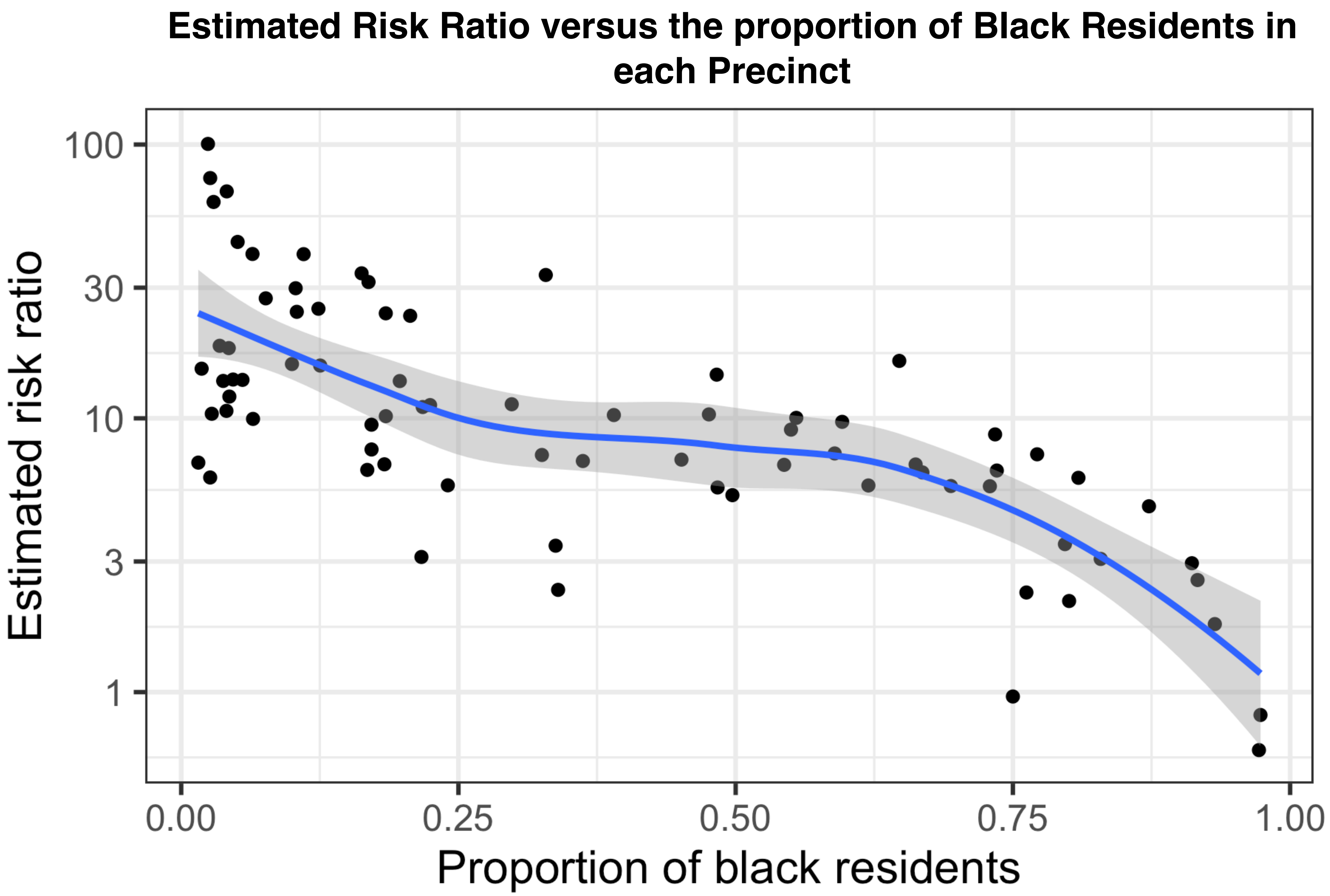
Comments and Critiques

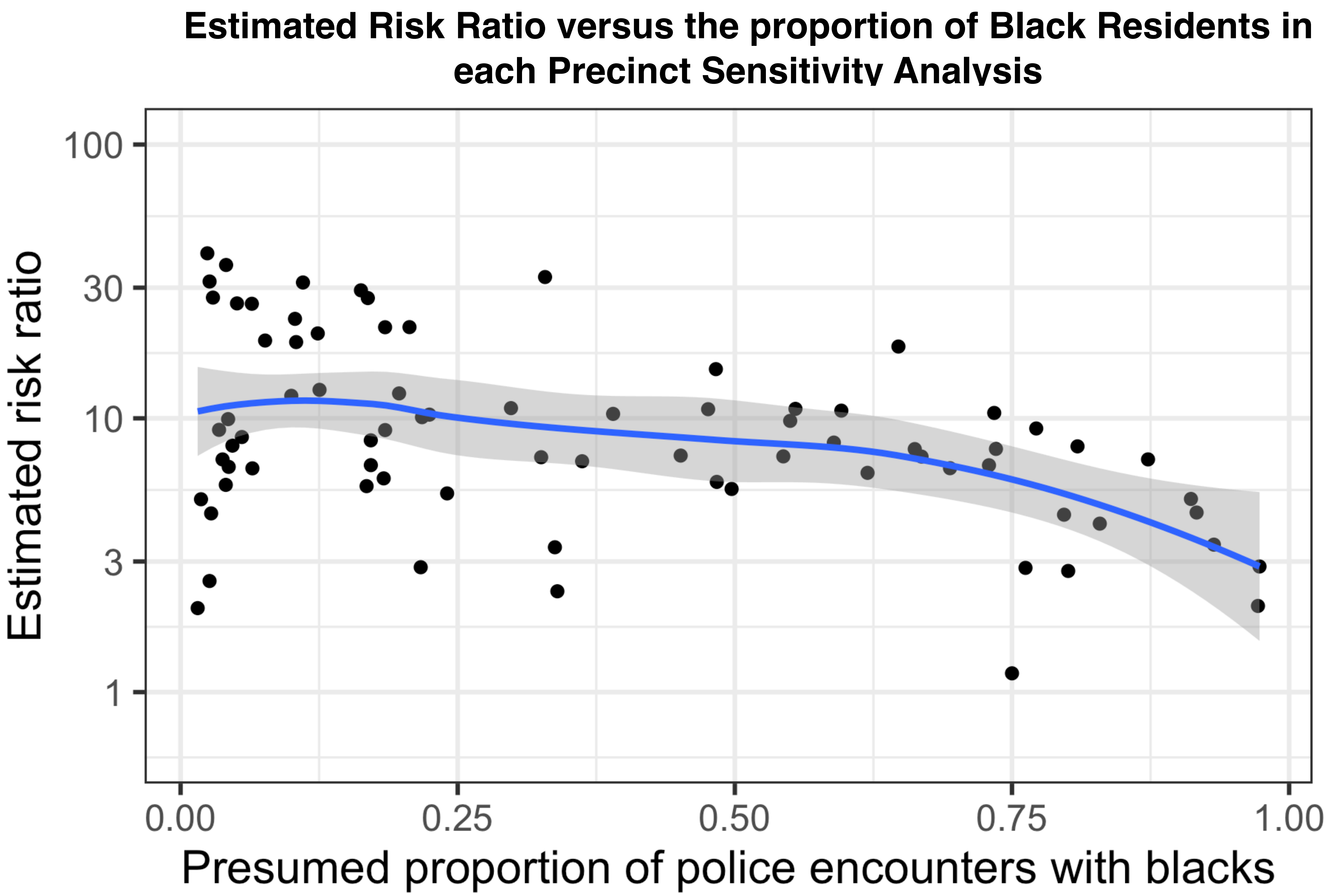
- Zhao et. al paper extends KLM method to tell more about the global effect
- Mandatory Reporting assumption will almost always be wrong
- Officer’s race perception may affect analysis

Thanks! Questions?

Appendix

Sensitivity Analysis





Local and Global Estimator Counterexamples

$$ATE = ATT = PIE + PDE$$

if $\beta_M \geq 0$ and $\beta_Y \geq 0$, then

$$ATE = ATT = \underbrace{\beta_M}_{\geq 0} \underbrace{E[Y(1,1)]}_{\geq 0} + \underbrace{\beta_Y}_{\geq 0} \underbrace{E[M(0)]}_{\geq 0} \geq 0$$

if $\beta_M \leq 0$ and $\beta_Y \leq 0$, then

$$ATE = ATT = \underbrace{\beta_M}_{\leq 0} \underbrace{E[Y(1,1)]}_{\geq 0} + \underbrace{\beta_Y}_{\leq 0} \underbrace{E[M(0)]}_{\geq 0} \leq 0$$

Local and Global Estimator Counterexamples

(i) When $\beta_M = \beta_Y = 0.01, P(S = al) = 0.1, P(S = ma) = 0.05, E[Y(0,1)] = 0.1$ and $P(D = 1) = 0.01$, we have that $ATE_{M=1} = -0.003884$

(ii) When $\beta_M = \beta_Y = -0.01, P(S = al) = 0.1, P(S = ma) = 0.05, E[Y(0,1)] = 0.1$ and $P(D = 1) = 0.99$, we have that $ATE_{M=1} = 0.002514$

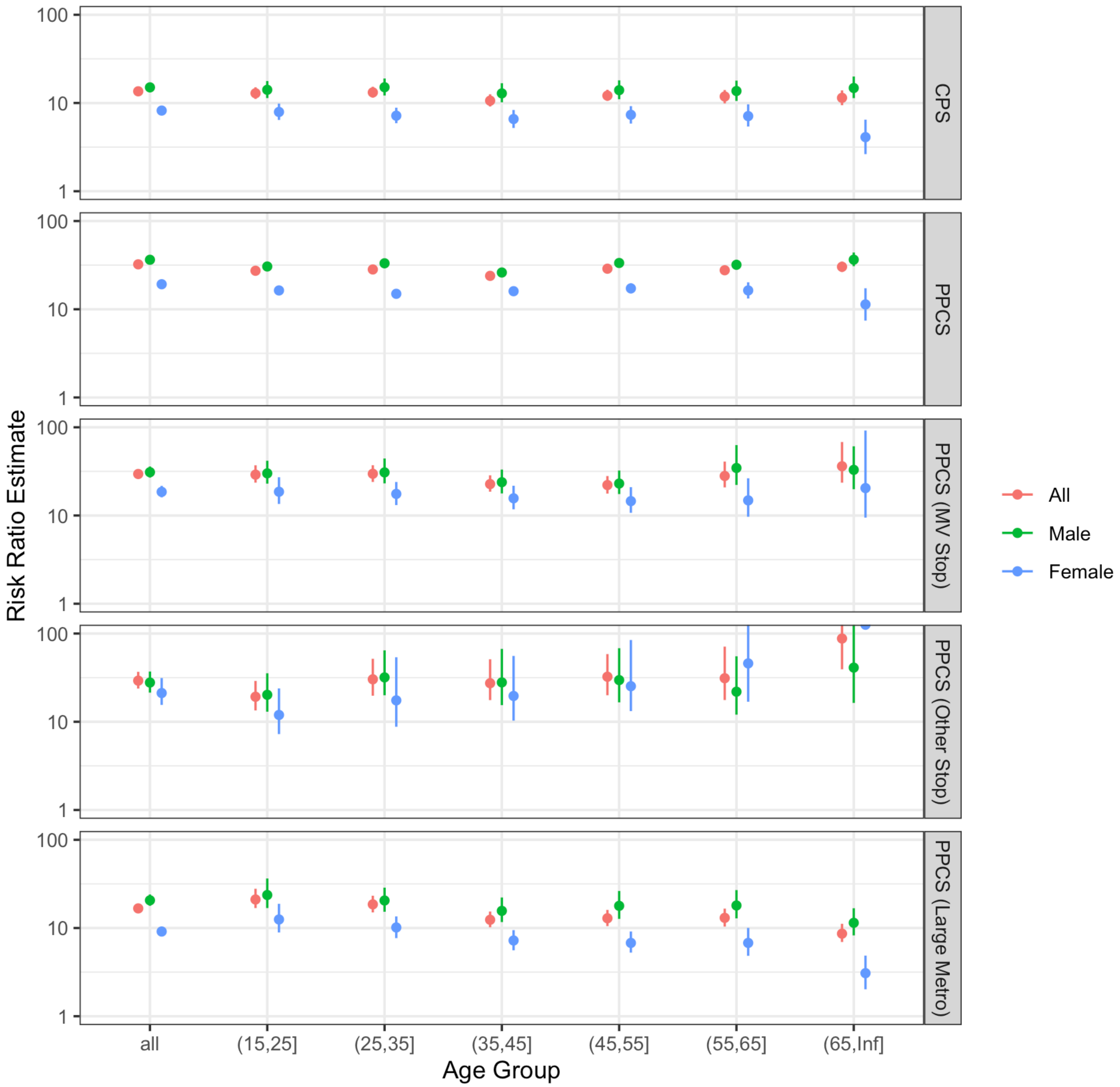
(iii) When $\beta_M = \beta_Y = -0.01, P(S = al) = 0.1, P(S = ma) = 0.05, E[Y(0,1)] = 0.1$ and $P(D = 1) = 0.01$, we have that $ATE_{M=1} = 0.0026$

Derivation of CRR

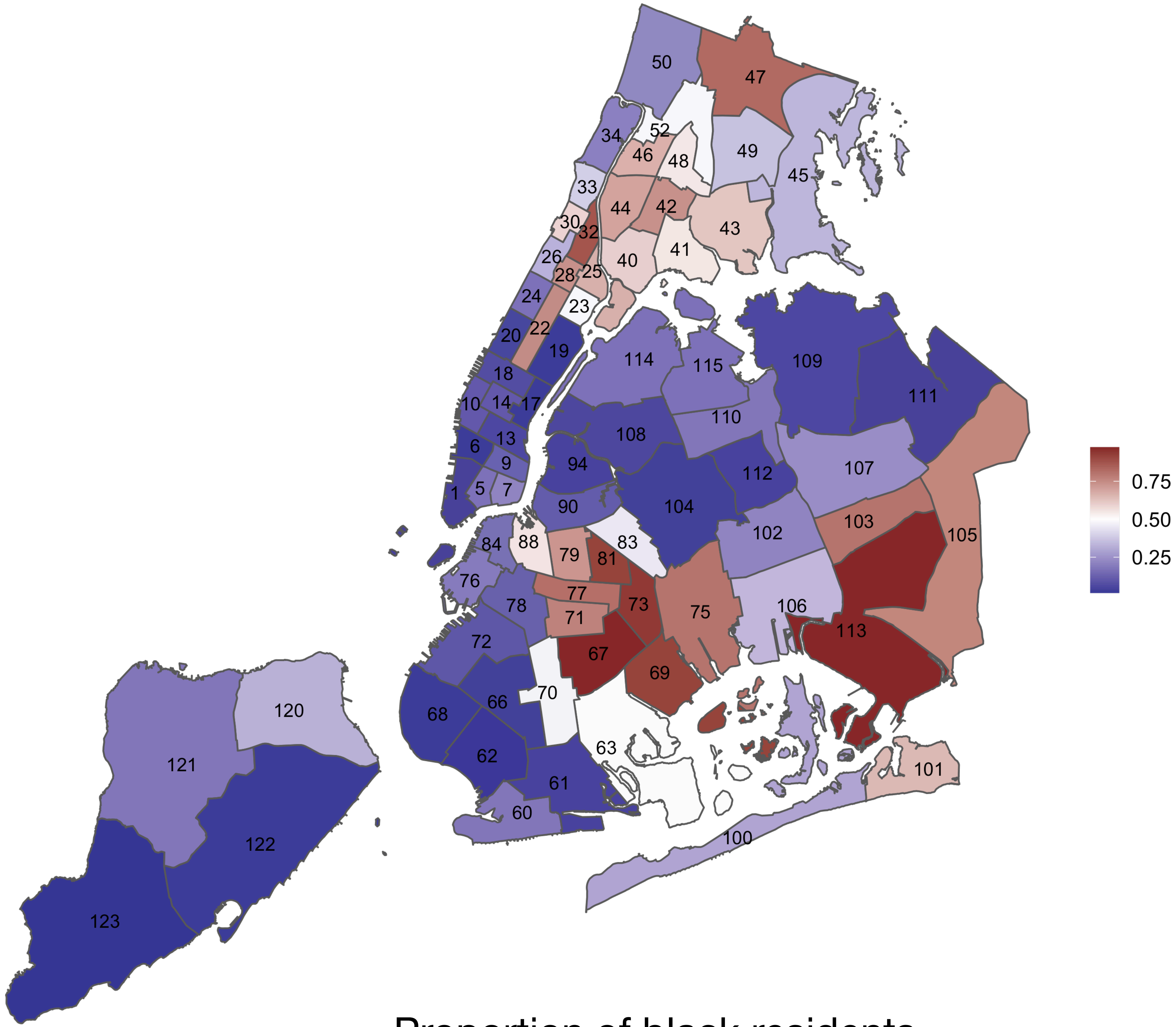
$$
\begin{aligned}
E[Y(d)|X] &= E[E[Y(d)|M(d)=1, X=x]|X=x] \\
&= E[Y(d)|M(d)=1, X=x]P(M(d)=1|X=x) \\
&= E[Y(d,1)|M(d)=1, X=x]P(M(d)=1|X=x) \\
&= E[Y(d,1)|M(d)=1, D=d, X=x]P(M(d)=1|X=x) \quad D \perp Y(d,1)|M(d), X \quad \text{(i.e conditional treatment ignorability)} \\
&= E[Y|M=1, D=d, X=x]P(M(d)=1|X=x) \quad \text{SUTVA/consistency} \\
&= E[Y|M=1, D=d, X=x]P(M(d)=1|D=d, X=x) \quad D \perp M(d) \\
&= E[Y|M=1|D=d, X=x]P(M=1|D=d, X=x) \quad d=0,1
\end{aligned}
$$

$$
\begin{aligned}
E[Y(1)|X=x] &= E[Y|M=1|D=1, X=x]P(M=1|D=1, X=x) \\
P(M=1|D=1, X=x) &= \frac{P(D=1|M=1, X=x)P(X=x, M=1)}{P(M=1|D=1, X=x)P(X=x)} \\
E[Y(0)|X=x] &= E[Y|M=1|D=0, X=x]P(M=1|D=0, X=x) \\
P(M=1|D=0, X=x) &= \frac{P(D=0|M=1, X=x)P(X=x, M=1)}{P(M=1|D=0, X=x)P(X=x)} \\
\frac{E[Y(1)|X=x]}{E[Y(0)|X=x]} &= \frac{E[Y|M=1|D=1, X=x]}{E[Y|M=1|D=0, X=x]} \times \frac{\frac{P(D=1|M=1,X=x)P(X=x,M=1)}{P(M=1|D=1,X=x)P(X=x)}}{\frac{P(D=0|M=1,X=x)P(X=x,M=1)}{P(M=1|D=0,X=x)P(X=x)}} \\
&= \frac{E[Y|M=1|D=1, X=x]}{E[Y|M=1|D=0, X=x]} \times \frac{\frac{P(D=1|M=1,X=x)}{P(M=1|D=1,X=x)}}{\frac{P(D=0|M=1,X=x)}{P(M=1|D=0,X=x}}
\end{aligned}
$$

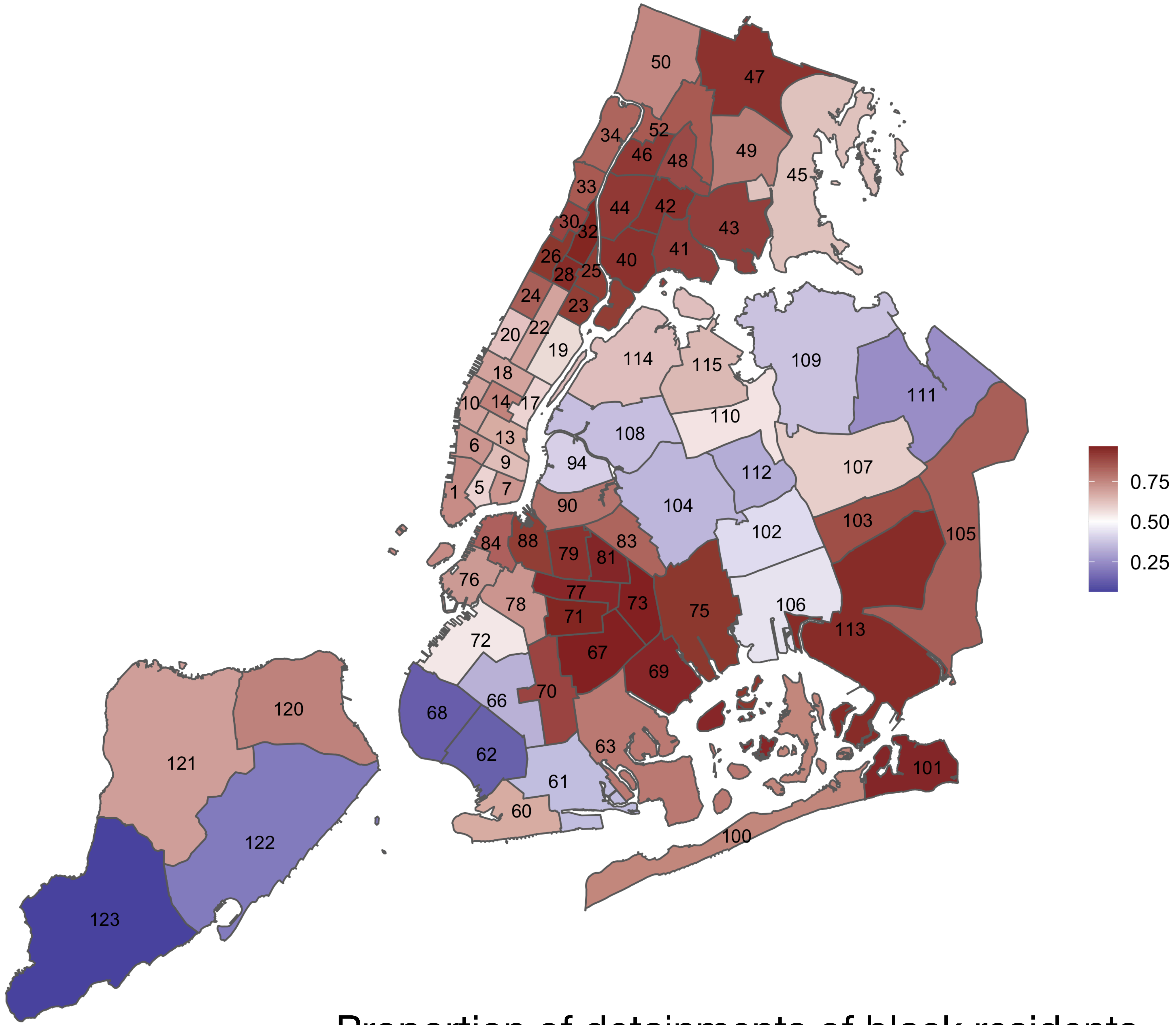
Gender and Age Stratified Analysis



Precinct Analysis: Racial Distributions



Proportion of black residents



Proportion of detainments of black residents

Precinct Analysis