Causal Data Science for Business Decision Making

Sample Selection Bias

Paul Hünermund

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

CEMS

PIM

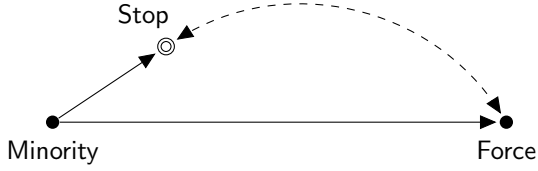
Administrative Records Mask Racially Biased Policing

DEAN KNOX *Princeton University*
WILL LOWE *Hertie School of Governance*
JONATHAN MUMMOLO *Princeton University*

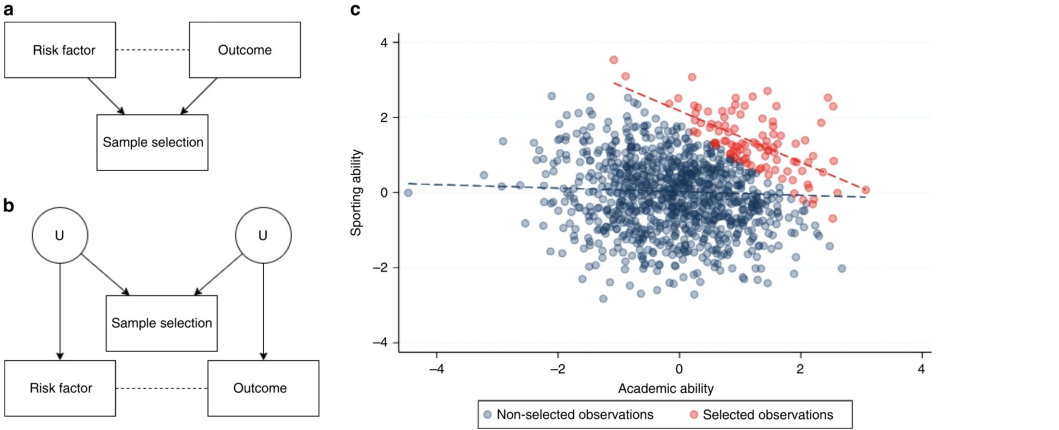
Researchers often lack the necessary data to credibly estimate racial discrimination in policing. In particular, police administrative records lack information on civilians police observe but do not investigate. In this article, we show that if police racially discriminate when choosing whom to investigate, analyses using administrative records to estimate racial discrimination in police behavior are statistically biased, and many quantities of interest are unidentified—even among investigated individuals—absent strong and untestable assumptions. Using principal stratification in a causal mediation framework, we derive the exact form of the statistical bias that results from traditional estimation. We develop a bias-correction procedure and nonparametric sharp bounds for race effects, replicate published findings, and show the traditional estimator can severely underestimate levels of racially biased policing or mask discrimination entirely. We conclude by outlining a general and feasible design for future studies that is robust to this inferential snare.

The Problem of Selection Bias

- Fryer (2019) studied the extent to which racial minorities (blacks, Hispanics) in the U.S. experience police violence more frequently (while controlling for context and civilian behavior) and found relatively small effects
- Knox et al. (2020) criticize this and similar papers that try to measure the degree of racial-bias in policing with the help of administrative records
  - Problem: An individual only appears in the data, if it was stopped by the police
  - If there is a racial bias in policing, stopping can be the result of minority status
  - There are unobserved confounders, such as officers' suspicion, between the selection variable and outcome
  - In a reanalysis they find effects that are at least four times larger



Sample Selection and Collider Bias



a A directed acyclic graph (DAG) illustrating a scenario in which collider bias would distort the estimate of the causal effect of the risk factor on the outcome. Directed arrows indicate causal effects and dotted lines indicate induced associations. Note that the risk factor and the outcome can be associated with sample selection indirectly (e.g. through unmeasured confounding variables), as shown in **b**. The type of collider bias induced in graph (**b**) is sometimes referred to as M-bias. To illustrate the example in **a**, consider academic ability and sporting ability to each influence selection into a prestigious school. As shown in **c**, these traits are negligibly correlated in the general population (blue dotted line), but because they are selected for enrolment they become strongly correlated when analysing only the selected individuals (red dotted line).

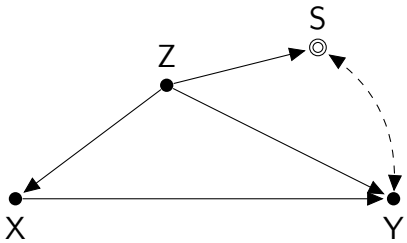
Two Special Cases

- Traditionally, economists knew two techniques for dealing with selection bias
  1. Selection propensity scores
  2. (Heckman) selection models
- The selection propensity score is the probability of being included in the sample, S, given the parent nodes of S: $e(PA) = P(s|PA)$
  - In the policing example this would be $P(stop|minority)$
- If we are willing to make certain functional form assumptions about e
  - e.g., that $e(PA)$ is monotonically increasing (stopping probability increases with minority status)
  - Then by conditioning on $e(PA)$ in the analysis, we can control for selection bias (Angrist, 1997)

Heckman Selection Model

- Heckman (1976, 1979) was interested in studying the hourly wage of women
- His sample included 2,253 working women interviewed in 1967
- Problem: Wages are not observed for women who choose not to work
  - If the wages they are able to obtain on the market are too low, some women might decide to stay at home
  - Working women are not a representative sample of the population (keep in mind, this is the 60s)

Heckman Selection Model (II)



$$s_i \leftarrow \quad \mathbb{1}[Z_i'\delta - \eta_i > 0]$$

$$y_i \leftarrow \begin{cases} x_i\beta + Z_i'\gamma + \varepsilon_i & \text{if } s_i = 1, \\ \text{unobserved} & \text{if } s_i = 0. \end{cases}$$

X: hours worked

Y: wages

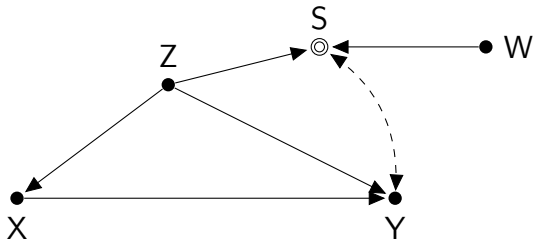
Z: other socio-economic factors

S: Sample selection indicator

with $Corr(\eta, \varepsilon) \neq 0$, i.e., there are unobservables that jointly affect whether a women chooses to work and her market wage (e.g., financial situation)

Heckman Selection Model (III)

- By making specific distributional assumption, such as joint normality of η and ε, we can recover from selection bias
- The standard version of the Heckman selection model, where Z_i is the same in both equations, relies heavily on distributional assumptions
- To avoid this, you should generally have at least one observed variable W that affects S but not Y in the model ("exclusion restriction", similar to an instrument)



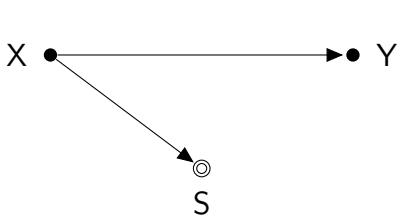
Selection Bias in Causal Diagrams

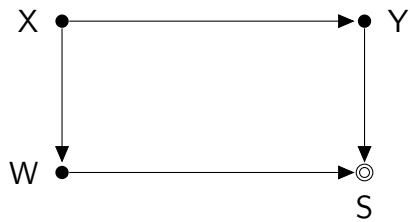
Task: Use the rules of do-calculus to transform $Q = P(y|do(x))$ into an expression that only contains probabilities conditional on $S = 1$

- There is a principled solution for dealing with selection bias in DAGs based on do-calculus (Bareinboim and Pearl, 2012; Bareinboim et al., 2014; Bareinboim and Tian, 2015)
- Compared to standard approaches in econometrics, these results do not rely on
  - functional-form assumptions about the selection propensity score $P(s|PA)$ (Heckman, 1979)
  - or, ignorability of the selection mechanism (Angrist, 1997)
- In addition, there is the possibility to combine biased and unbiased data in order to increase identifying power (Bareinboim et al., 2014; Correa et al., 2017)
  - In many applied settings, unbiased records of covariates are available from secondary data sources (e.g., census data)

Selection Diagram

- We can model the selection mechanism by incorporating a variable S in the DAG, which denotes whether we observe an observation ($S = 1$) or not ($S = 0$)
  - S receives an incoming arrow from those variables in the model that cause the sample selection
  - We denote a DAG that has been augmented with a selection node by G_S





Recovering from Selection Bias

- How can we recover the causal effect from a selected sample without invoking functional form assumptions such as linearity and normality (as in the famous Heckman selection model Heckman, 1979)?
- In order to recover $P(y|do(x))$, we need to translate it into a do-free expression that also conditions on $S = 1$, because that's all we can observe
- Bareinboim and Pearl (2012), Bareinboim et al. (2014) and Bareinboim and Tian (2015) give this problem a full formal treatment and derive graphical criteria and algorithms for recovering causal effects from selected samples

Recovering from Selection Bias (II)

Theorem 1 Bareinboim and Tian (2015)

The conditional distribution $P(y|t)$ is recoverable from G_S (as $P(y|t, S = 1)$) if and only if $(Y \perp\!\!\!\perp S|T)$.

Corollary 1 Bareinboim and Tian (2015)

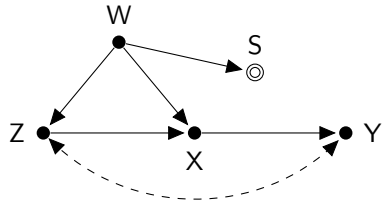
The causal effect $Q = P(y|do(x))$ is recoverable from selection-biased data if using the rules of the do-calculus, Q is reducible to an expression in which no do-operator appears, and recoverability is determined by the previous Theorem.

Recovering from Selection Bias (III)

- In the previous selection diagram with $X \rightarrow Y$ and $X \rightarrow S$, the causal effect is recoverable because $P(y|do(x)) = P(y|x)$ and S is d-separated from Y
  - Thus, $P(y|do(x)) = P(y|x, S = 1)$, which can be estimated from selection-biased data
- An immediate consequence of Theorem 1 is that if S is directly connected to Y, as in the racial policing case, the causal effect will not be recoverable (without strengthening assumptions)
- Note that corollary 1 is a sufficient but not necessary condition for recoverability
  - We can use do-calculus to reduce $P(y|do(x))$ to a do-free expression that conditions on $S = 1$ in cases when corollary 1 is not applicable
  - Bareinboim and Tian (2015) develop an algorithm which automatizes this step

Do-Calculus Example: Selection Bias

Take the following DAG augmented with selection node S:



By the first rule of do-calculus, since $(S, W \perp\!\!\!\perp Y)$ in $G_{\overline{X}}$ (the resulting graph when all incoming arrows in X are deleted),

$$P(y|do(x)) = P(y|do(x), w, S = 1),$$

$$= \sum_z P(y|do(x), z, w, S = 1)P(z|do(x), w, S = 1),$$

where the second line follows from the law of total probability.

Do-Calculus Example: Selection Bias (II)

Applying rule 2, since $(Y \perp\!\!\!\perp X | W, Z)$ in $G_{\underline{X}}$ (the resulting graph when all arrows emitted by X are deleted), we can eliminate the do-operator in the first term

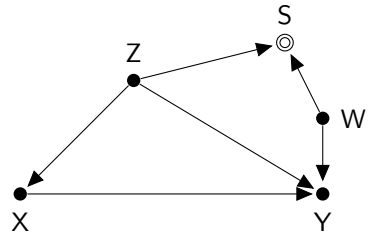
$$= \sum_z P(y|x, z, w, S = 1) P(z|do(x), w, S = 1).$$

Finally, because $(Z \perp\!\!\!\perp X | W)$ in $G_{\overline{X}}$, it follows from rule 3 that

$$= \sum_z P(y|x, z, w, S = 1) P(z|w, S = 1).$$

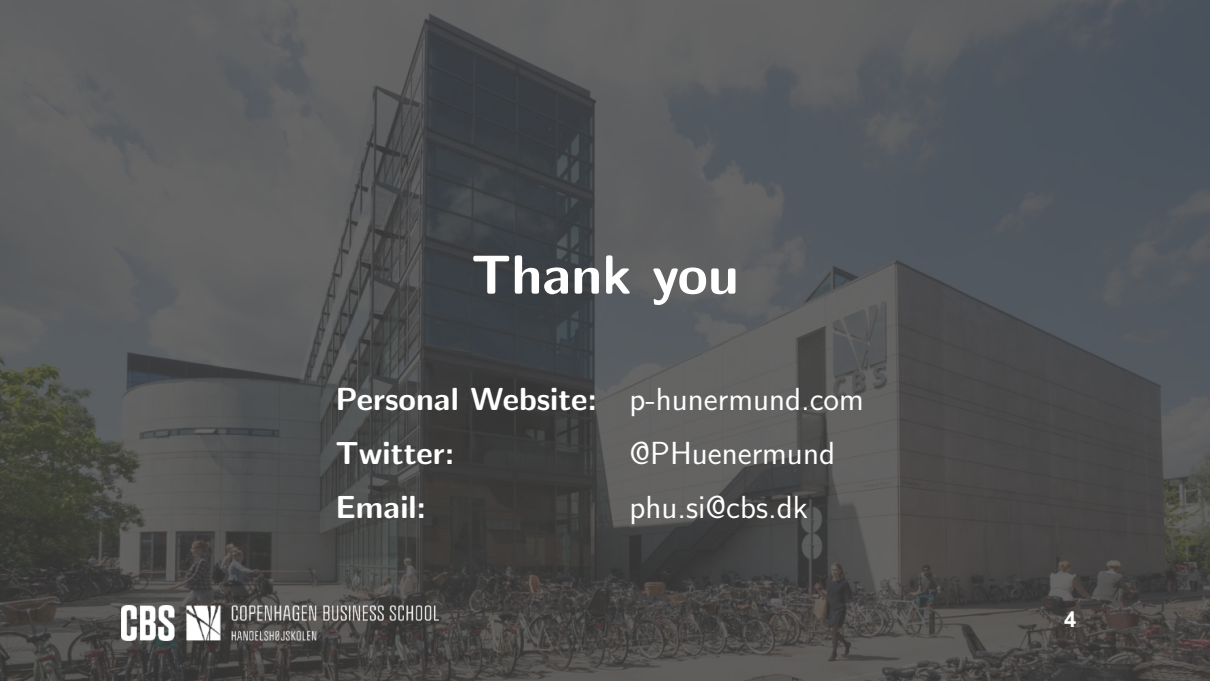
Combining Biased and Unbiased Data

- Sometimes we can get at unbiased measurements of covariates, e.g., from census data
- Take the graph on the right. Conditioning on the set $\{Z, W\}$ closes all backdoor paths and d-separates Y from S. Thus



$$P(y|do(x)) = \sum_{z,w} P(y|x,z,w)P(z,w)$$
$$= \sum_{z,w} P(y|x,z,w,S=1)P(z,w)$$

- $P(Z=z, W=w)$ is not recoverable according to Theorem 1. But if we can get unbiased measurements of Z and Y, this expression is estimable
- Bareinboim et al. (2014) provide more general criteria for this idea



Thank you

Personal Website: p-hunermund.com

Twitter: @PHuenermund

Email: phu.si@cbs.dk

References I

Joshua D. Angrist. Conditional independence in sample selection models. *Economics Letters*, 54:103–112, 1997.

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 100–108, 2012.

Elias Bareinboim and Jin Tian. Recovering causal effects from selection bias. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Juan D. Correa, Jin Tian, and Elias Bareinboim. Generalized adjustment under confounding and selection biases. Technical Report R-29-L, 2017.

Roland G. Fryer. An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, 127(3), 2019. doi: 10.1086/701423.

James Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, 5:475–492, 1976.

James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

Dean Knox, Will Lowe, and Jonathan Mummolo. Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637, 2020.