Toward a General Causal Framework for the Study of Racial Bias in Policing

Online Appendix

Contents

A Outcome Selection

Assuming ignorability of civilian race,

$$\begin{aligned}
\text{ATE} &= \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))] \\
&= \Pr(Y_i = 1|D_i = 1) - \Pr(Y_i = 1|D_i = 0) \\
&= \frac{\Pr(D_i = 1|Y_i = 1)\Pr(Y_i = 1)}{\Pr(D_i = 1)} - \frac{\Pr(D_i = 0|Y_i = 1)\Pr(Y_i = 1)}{\Pr(D_i = 0)} \\
&= 1 - \frac{\Pr(D_i = 1|Y_i = 0)\Pr(Y_i = 0)}{\Pr(D_i = 1)} - 1 + \frac{\Pr(D_i = 0|Y_i = 0)\Pr(Y_i = 0)}{\Pr(D_i = 0)} \\
&= -\frac{\Pr(D_i = 1|Y_i = 0)\Pr(Y_i = 0)}{\Pr(D_i = 1)} + \frac{\Pr(D_i = 0|Y_i = 0)\Pr(Y_i = 0)}{\Pr(D_i = 0)} \qquad (1) \\
&= -\frac{\Pr(D_i = 1|Y_i = 0)\Pr(Y_i = 0)}{\Pr(D_i = 1|Y_i = 0)\Pr(Y_i = 0) + \Pr(D_i = 1|Y_i = 1)\Pr(Y_i = 1)} \\
&\quad + \frac{\Pr(D_i = 0|Y_i = 0)\Pr(Y_i = 0)}{\Pr(D_i = 0|Y_i = 0)\Pr(Y_i = 0) + \Pr(D_i = 0|Y_i = 1)\Pr(Y_i = 1)} \qquad (2)
\end{aligned}$$

In studies that select on the outcome, analysts typically have no information about $\Pr(Y_i = 1)$, how frequently officers engage in the behavior of interest (e.g., what proportion of encounters result in a shooting). Rather, analysts only have data to estimate $\Pr(D_i = 1|Y_i = 1)$ and $\Pr(D_i = 0|Y_i = 1)$. In this case, bounds on the ATE follow by substituting extreme values for the missing information, $\Pr(Y_i = y)$ and $\Pr(D_i = d|Y_i = 0)$.

For example, one extreme possibility is as follows: almost all encounters are unobserved non-shootings ($\Pr(Y_i = 0)$ approaching one and $\Pr(Y_i = 1)$ approaching zero), and all of these non-shooting encounters are with white civilians ($\Pr(D_i = 1|Y_i = 0) = 0$, meaning $\Pr(D_i = 0|Y_i = 0) = 1$). In this scenario—which analysts cannot rule out using the available data—the racial bias in shootings approaches ATE $= +1$, the highest possible value. Similarly, analysts cannot rule out the reverse, that all of the unobserved non-shooting encounters are with *minority* civilians, so that $\Pr(D_i = 1|Y_i = 0) = 0$ and $\Pr(D_i = 0|Y_i = 0) = 1$, which would imply the ATE approaches -1. Thus, despite having data on all shootings, researchers know nothing more about the quantity of interest than they did before beginning the study—and any conclusions to the contrary are based entirely on assumptions that the data cannot support.

If the shooting rate, $\Pr(Y_i = 1)$, is known, then these bounds can be narrowed somewhat. In this case, plugging in extreme values for $\Pr(D_i = 0|Y_i = 0)$ and $\Pr(D_i = 1 = 0)$ in Equation 2 reveals that the range of possible bias is

$$-\frac{\Pr(Y_i = 0)}{\Pr(Y_i = 0) + \Pr(D_i = 1|Y_i = 1)\Pr(Y_i = 1)} \le \text{ATE} \le \frac{\Pr(Y_i = 0)}{\Pr(Y_i = 0) + \Pr(D_i = 0|Y_i = 1)\Pr(Y_i = 1)}$$

B Proportion Test

In the proportion test, analysts compare the stops made by minority officers to the stops made by white officers. In particular, analysts compare the proportion of each officer group's stops that are of minority civilians, as opposed to white civilians. The basic logic of this approach is to assess whether both officer groups take the same actions (e.g., stopping civilians) when facing identical pools of civilian behavior. This is a necessary, but not sufficient, condition for both groups to be unbiased: if the two groups of officers behave differently, then at least one must be biased in some direction. However, the converse is not true: if both groups behave identically, it could be that both are equally biased. Thus, the proportion test offers an asymmetric test of officer bias.

This test proceeds by estimating

$$\begin{aligned}\Delta = &\big(\Pr(D_i = 1|X_i = 1, M_i = 1) - \Pr(D_i = 0|X_i = 1, M_i = 1)\big) \\ &- \big(\Pr(D_i = 1|X_i = 0, M_i = 1) - \Pr(D_i = 0|X_i = 0, M_i = 1)\big),\end{aligned}$$

which can be rewritten as

$$\begin{aligned}&= \frac{\Pr(D_i = 1, M_i = 1|X_i = 1) - \Pr(D_i = 0, M_i = 1|X_i = 1)}{\Pr(M_i = 1|X_i = 1)} \\ &\quad - \frac{\Pr(D_i = 1, M_i = 1|X_i = 0) - \Pr(D_i = 0, M_i = 1|X_i = 0)}{\Pr(M_i = 1|X_i = 0)} \\ &= \frac{\Pr(M_i = 1|X_i = 1, D_i = 1)\Pr(D_i = 1|X_i = 1) - \Pr(M_i = 1|X_i = 1, D_i = 0)\Pr(D_i = 0|X_i = 1)}{\Pr(M_i = 1|X_i = 1)} \\ &\quad - \frac{\Pr(M_i = 1|X_i = 0, D_i = 1)\Pr(D_i = 1|X_i = 0) - \Pr(M_i = 1|X_i = 0, D_i = 0)\Pr(D_i = 0|X_i = 0)}{\Pr(M_i = 1|X_i = 0)}\end{aligned}$$

invoking the ignorability of civilian race,

$$\begin{aligned}&= \frac{\mathbb{E}[M_i(1) - M_i(0)|X_i = 1]\Pr(D_i = 1|X_i = 1)}{\Pr(M_i = 1|X_i = 1)} \\ &\quad - \frac{\Pr(M_i = 1|X_i = 1, D_i = 0)\big(\Pr(D_i = 0|X_i = 1) - \Pr(D_i = 1|X_i = 1)\big)}{\Pr(M_i = 1|X_i = 1)} \\ &\quad - \frac{\mathbb{E}[M_i(1) - M_i(0)|X_i = 0]\Pr(D_i = 1|X_i = 0)}{\Pr(M_i = 1|X_i = 0)} \\ &\quad + \frac{\Pr(M_i = 1|X_i = 0, D_i = 0)\big(\Pr(D_i = 0|X_i = 0) - \Pr(D_i = 1|X_i = 0)\big)}{\Pr(M_i = 1|X_i = 0)}\end{aligned}$$

and finally, the common pool assumption gives $\Pr(D_i = d|X_i = 0) = \Pr(D_i = d|X_i = 1) = \Pr(D_i = d)$, yielding

$$= \frac{\mathbb{E}[M_i(1) - M_i(0)|X_i = 1]\Pr(D_i = 1) - \mathbb{E}[M_i(0)|X_i = 1]\big(\Pr(D_i = 0) - \Pr(D_i = 1)\big)}{\Pr(M_i = 1|X_i = 1)} - \frac{\mathbb{E}[M_i(1) - M_i(0)|X_i = 0]\Pr(D_i = 1) - \mathbb{E}[M_i(0)|X_i = 0]\big(\Pr(D_i = 0) - \Pr(D_i = 1)\big)}{\Pr(M_i = 1|X_i = 0)}.$$

This shows that the desired quantity of interest, the difference in differences ($\mathbb{E}[M_i(1) - M_i(0)|X_i = 1] - \mathbb{E}[M_i(1) - M_i(0)|X_i = 0]$), is not identified by the proportion test: a number of additional assumptions are required to connect the two. Specifically, to draw inferences about the difference in differences, analysts must first assume that overall stopping rates are equal across officer groups, or that $\Pr(M_i = 1|X_i = 0, D_i = d) = \Pr(M_i = 1|X_i = 1, D_i = d)$. In other words, one officer group cannot patrol more actively or be more stringent in enforcement; among other things, this ensures that the denominators are comparable. Then, analysts would need to further assume that both groups treat white civilians equally, so that the second and fourth terms cancel, and the remaining terms contain the desired difference in differences (multiplied by a scaling factor). To be clear, we do not advocate these highly implausible assumptions. Rather, we enumerate them to convey the difficulty in directly interpreting the results of the proportion test in terms of a substantively useful quantity of interest. However, as we discuss in the main text, the proportion test remains a useful test that can reject the null hypothesis that no officer group is biased, and it offers analysts the ability to examine this question when no other statistical test is applicable.