Statistical Modeling, Causal Inference, and Social Science

Statistical controversy on estimating racial bias in the criminal justice system

Posted on July 6, 2020 9:02 AM by Andrew

1. Background

A bunch of people have asked me to comment on these two research articles:

Administrative Records Mask Racially Biased Policing, by Dean Knox, Will Lowe, and Jonathan Mummolo:

> *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.*

Deconstructing Claims of Post-Treatment Bias in Observational Studies of Discrimination, by Johann Gaebler, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel, and Jennifer Hill:

> *In studies of discrimination, researchers often seek to estimate a causal effect of race or gender on outcomes. For example, in the criminal*

> *justice context, one might ask whether arrested individuals would have been subsequently charged or convicted had they been a different race. It has long been known that such counterfactual questions face measurement challenges related to omitted-variable bias, and conceptual challenges related to the definition of causal estimands for largely immutable characteristics. Another concern, raised most recently in Knox et al. [2020], is post-treatment bias. The authors argue that many studies of discrimination condition on intermediate outcomes, like being arrested, which themselves may be the product of discrimination, corrupting statistical estimates. Here we show that the Knox et al. critique is itself flawed, suffering from a mathematical error. Within their causal framework, we prove that a primary quantity of interest in discrimination studies is nonparametrically identifiable under a standard ignorability condition that is common in causal inference with observational data. More generally, though, we argue that it is often problematic to conceptualize discrimination in terms of a causal effect of protected attributes on decisions. We present an alternative perspective that avoids the common statistical difficulties, and which closely relates to long-standing legal and economic theories of disparate impact. We illustrate these ideas both with synthetic data and by analyzing the charging decisions of a prosecutor's office in a large city in the United States.*

I heard about these papers a couple years ago but didn't think too hard about them. Then recently a bunch of different people contacted me and asked for my opinion. Apparently there's been some discussion on twitter. The Knox et al. paper was officially published in the journal so maybe that was what set off the latest round of controversy.

Also, racially biased policing is in the news. Not so many people are defending the police—even the people saying the police aren't racially biased are making that claim based on evidence that police mistreat white people too—but I suppose that much of the current debate revolves around the idea that the police enforce inequality. So whatever the statistics are regarding police mistreatment of suspects, these are part of a larger issue of the accountability of police use of force. We've been hearing a lot about police unions, which raises other political questions, such as what is it like for police officers who disagree with the positions taken by their leadership.

Anyway, that's all part of the background for how this current academic dispute has

become such a topic of discussion.

2. Disclaimer

I have no financial interest in this issue, nor do I have any direct academic stake. Some colleagues and I did some research twenty years ago on racial disparities in police stops, but neither of the new articles at hand has any criticism of what we did, I guess in part because our analysis was more descriptive than causal.

I do have a personal interest, though, as Sharad Goel and Jennifer Hill are friends and collaborators of mine. So my take on these papers is kind of overdetermined: Jennifer is my go-to expert on causal inference (although we have never fought more than over the causal inference chapters in our book), and I have a lot of respect for Sharad too as a careful social science researcher. I'll give you my own read on these articles in a moment, but I thought I should let you know where I'm coming from.

Just to be clear: I'm not saying that I expect to agree with Gaebler et al. because Sharad and Jennifer are my friends—I disagree with my friends all the time!—; rather, I'm saying that, coming into this one, my expectation is that, when it comes to causal inference, Jennifer knows what she's talking about and Sharad has a clear sense of what's going on in this particular application area.

3. My read of the two articles

Suppose you're studying racial discrimination, and you compare outcomes for whites to outcomes for blacks. You'll want to adjust for other variables: mere differences between the two groups does not demonstrate discrimination. In addition, you won't be working with the entire population; you'll be working with the subset who are in some situation. For example, Knox et al. look at the results of police-civilian encounters: these are restricted to the subset of civilians who are in this encounter in the first place.

Knox et al. make the argument that analyses using administrative data are implicitly conditioning on a post-treatment variable because they subset on whether you were stopped or not. To use the words in their title, adjusting for or conditioning on administrative records can mask racially biased policing. Knox et al. point out that this bias can be viewed as an example of conditioning on intermediate outcomes. It's a well known principle in causal inference that you can't give regression coefficients a direct causal interpretation if you're conditioning on intermediate outcomes; see, for example, section 19.6, "Do not adjust for post-treatment variables," of Regression and Other Stories, but this is standard advice—my point in citing ourselves here is not to

claim any priority but rather to say that this is standard stuff.

Knox et al. are right that conditioning on post-treatment variables is a common mistake. R. A. Fisher made that mistake once! Causal inference is subtle. In general, there's a tension between the principles of adjusting for more things and the principle of not adjusting away the effect that you're interested in studying. And, as Knox et al. point out, there's an additional challenge for criminal justice research because of missing data: "police administrative records lack information on civilians police observe but do not investigate." This is an important point, and we should always be aware of the limitations of our data.

From a substantive point of view, the message that I take from Knox et al. is to be careful with what might seem to be kosher regression analyses purporting to estimate the extent of discrimination. I would not want to read Knox et al. as saying that regression methods can't work here. You have to be aware of what you are conditioning on, but if you interpret the results carefully, you should be able to estimate a causal effect of one stage in the process. The concern is that it can be easy for the most important effects to be hidden in the data setup.

Gaebler et al. discuss similar issues. However, they disagree with Knox et al. on technical grounds. Gaebler et al. argue that there exist situations in which you may be able to estimate causal effects of discrimination or perception of race just fine, even when conditioning on variables that are affected by race. It's just that these won't capture all aspects of discrimination in the system; they will capture the discrimination inherent solely at that point in the process (for example, when a prosecutor makes a decision about charging).

For a simple example, suppose you have stone-cold evidence of racial bias in sentencing. That's still conditional on who gets arrested, charged, and prosecuted. So it doesn't capture all potential sources of racial bias, not by a long shot. Or, maybe you find no racial bias in sentencing. That doesn't mean that total racial bias is zero; it just means that you don't find anything at that stage in the process.

I see Gaebler et al. as being in agreement with Knox et al. on this key substantive point. The distinction is that Knox et al. are saying that in principle you can't estimate effects of discrimination using the basic causal regression, whereas Gaebler et al. are saying that it can be, and often is, possible to estimate these effects, even though these effects are not the whole story.

Gaebler et al. give an example where the estimand corresponds to what one would

measure in a randomized controlled trial where the stated race of arrested individuals was randomly masked on the police narratives that prosecutors use when making their decisions. With this setup, it is possible to estimate causal racial bias in a particular part of the system.

For another example, suppose you did a study of racial discrimination and you included, as a predictor, the neighborhood where people were living. And you found some amount of discrimination X, a difference between what happens to whites and to blacks, conditional on neighborhood. That could be a causal effect, but, even if X = 0, that would not mean that there is no racial discrimination. If there is discrimination by neighborhood, this can have disparate impact.

4. From the study of discrimination to the study of disparate outcomes

One thing I especially like about the Gaebler et al. article is that they move beyond the question of racial discrimination to the more relevant, I think, issue of disparate outcomes: "much of the literature has framed discrimination in terms of causal effects of race on behavior, but other conceptions of discrimination, such as disparate impact, are equally important for assessing and reforming practices."

Again, suppose that all discrimination were explainable not by race but by what neighborhood you live in. People are segregated geographically, so discrimination by neighborhood really would be a form of racial discrimination, but this could be set up so that the causal effect of race itself (for example, in police or prosecutor decisions) would be zero. But from the Knox et al. perspective, you could say that this control variable (the ethnic and racial composition of your neighborhood) is an intermediate outcome. I prefer Gaebler et al.'s framing, but ultimately I think both articles are coming to the same point here.

And here's how Gaebler et al. end it:

> *The conclusions of discrimination studies are generally limited to specific decisions that happen within a long chain of potentially discriminatory actions. Quantifying discrimination at any one point (e.g., charging decisions) does not reflect specific or cumulative discrimination at other stages—for example, arrests. Looking forward, we hope our work offers a path toward quantifying disparities, and provokes further interest in the subtle conceptual and methodological issues at the heart of discrimination studies.*

I agree. Lots more can be said on this topic, and I recommend Gaebler et al. as a

starting point.

5. So why the controversy?

Given that, from my perspective, the two papers have such similar messages, why the controversy? Why the twitter war?

I guess I can see where the authors of both papers are coming from. From the perspective of Gaebler et al., the distinction is clear. Knox et al. have made a mathematically false statement leading to a confusion about the identifiability of causal effects within the justice system. Knox et al. don't seem to agree regarding the mathematical error, but in any case they take the position that their result is correct in practice, labeling Gaebler et al.'s results as "silly logic puzzles about knife-edge scenarios" that "aren't useful to applied researchers."

The concern I have about Knox et al. is that they make a claim that's too strong. They say that they "show that when there is any racial discrimination in the decision to detain civilians . . . then estimates of the effect of civilian race on subsequent police behavior are biased absent additional data and/or strong and untestable assumptions," that "the observed difference in means fails to recover any known causal quantity without additional, and highly implausible, assumptions," and that this estimation strategy "[cannot] be rehabilitated by simply redefining the quantity of interest" or by adding additional covariates.

In some sense, sure, you can't make any causal inference without strong assumptions. Even in a textbook randomized controlled experiment, all you're doing is estimating the average causal effect for the people in the study (or a larger population for which your participants can be considered a random sample). When your data are observational, you need even more assumptions. And don't get me started on instrumental variables, regression discontinuity, etc.: all these methods can be useful, but they don't come without lots of modeling.

But that can't be what Knox et al. are trying to say. The statement, "estimates of the effect of X on Y are biased absent additional data and/or strong and untestable assumptions," is true for *any* X and Y. It's a good point to make, quite possibly worth a paper in the APSR, but nothing in particular to do with criminal justice. The relevant point to make, and hence the point I will extract from Knox et al., is that, in this particular example of arrests, the problems of selection is, in the words of Morrissey, really serious. But you can use standard methods of causal inference to estimate causal effects here, as long as you're careful in interpreting the results and don't take

the estimate of discrimination in one part of the system as representing the entirety of racial bias in the whole process.

My own resolution to this is to take Knox et al.'s message of the difficulty of causal inference under selection, as applied to the important problem of estimating disparities in the criminal justice system as a caution to be careful in your causal thinking: Define your causal effects carefully, and recognize their limitations (as in the above example where the causal effect of race in arrest or sentencing decision, conditional on neighborhood, might be of interest, but at the same time we realize this particular causal effect would only capture one of many sources of racial discrimination). I like Gaebler et al.'s framing of the problem in terms of the specificity of their definition of the treatment variable. I think the authors of both papers would agree that overinterpretation of naive regressions has been a problem, hence the value of this work.

6. tl;dr summary

From both papers I draw the same substantive conclusion, which is it that simple, or even not-so-simple regressions of outcome on race and pre-treatment predictors can give misleading results if you're trying to understand the possibility of racial discrimination in the criminal justice system without thinking carefully about these issues.

There are also some technical disputes. It's my impression that Gaebler et al. are correct on these issues, but, now that I have a sense of the statistical issues here, I'm not so interested in the theorem, or the explanation of error in the theorem, or the error in the explanation of the error in the theorem, or the corrected explanation of the error in the theorem. My read of this particular dispute is that Knox et al. were trying to prove something that is not quite correct. The correct statement is that, even when standard regression-based inferences allow you to estimate a causal effect of race at some stage in the criminal justice process, this causal effect is conditional on everything that came before, and so a focus on any particular causal effect will not catch other biases in the system. The incorrect statement is that you can't estimate causal effects in standard regression-based inferences. These local causal effects don't tell the fully story, but they're still causal effects, and that's the technical point that Gaebler et al. are making.

My tl;dr is different from that of political scientist Ethan Bueno de Mesquita, who wrote this:

tl;dr

@jonmummolo and co-authors are right. You should read their papers, not this dreck.

I disagree with the "dreck" comment.

Don't get me wrong. I have no *general* problem with labeling papers as dreck; I've many times called published papers "crappy," and this blog is no stranger to pottymouth. It's just that I like the Gaebler et al. paper. I disagree with the above-quoted assessment of its quality. I can't really say more unless I hear Bueno de Mesquita's reasons for giving in the "dreck" label. He and others should feel free to respond in the comments.

This entry was posted in Political Science, Sociology **by** Andrew**. Bookmark the** permalink [https://statmodeling.stat.columbia.edu/2020/07/06/statistical-controversy-on-racial-bias-in-the-criminal-justice-system/] **.**

---

66 THOUGHTS ON "STATISTICAL CONTROVERSY ON ESTIMATING RACIAL BIAS IN THE CRIMINAL JUSTICE SYSTEM"

ZB
on July 6, 2020 9:16 AM at 9:16 am said:

Andrew, as a long-time reader, I'm a taken aback that you would endorse a paper that is, in no uncertain terms, saying "if you assume away selection problems, you can estimate causal effects." That's both trivially true and monumentally bad research advice! You've castigated papers for saying less egregious things.

> **Jake T** on July 6, 2020 1:30 PM at 1:30 pm said:
>
> Your comparison seems incomplete. In fact, I could propose a pithy summarization of Gaebler et al. as, "It is possible to carefully measure causal effects at any given stage of a multi-stage process." KLM, in that framing, argue that such careful analysis is impossible, and your conclusions will always be false. Since the universe is turtles all the way down, the absurd extension is that causal influence is impossible!
>
> In particular, KLM are _not_ arguing about omitted variables or selection effects at the second stage. That would have been a valid issue, though well understood. They are arguing that one must consider omitted variable bias at the first stage, even while studying and making claims only about the second. They certainly haven't provided sufficient evidence to justify this extraordinary claim – as far as I can tell, they haven't even corrected their invalid proof from the paper.