Testing for Discrimination and the Risk of Error

Boris Babic

California Institute of Technology

1 Introduction

The Stanford Open-Policing Project, launched in 2017, currently contains information on more than 60 million police stops across 20 US states conducted between 2011-2015 (Pierson et al., 2017).[1] The dataset was initially compiled through a series of public records requests and for each stop, we have information about the location, time and date, the infraction, the driver's age, gender, and race (black, hispanic, or white), whether a search of their vehicle was conducted, and whether contraband was found. With such a rich dataset we should be able to gain important insights about the pervasiveness of racial profiling in the United States.

This raises an important methodological question, however; one that I think philosophers ought to consider (to put it differently, one that I think should not be entrusted to economists and statisticians alone): how should we test for racial profiling? In other words, what sorts of patterns in this (or any other) data would be indicative of wrongful police discrimination? To answer this question we need a test for discrimination – or at least some device that allows us to make inferences as to whether, and the extent to which, it is happening. Ideally, we want this test (or other device) to be grounded in ethical theory but we also need a test that can be operationalized and applied in empirical research. In this project, I try to offer one such test. It's really an advertisement for a sketch of an outline of a test, to paraphrase Yablo. I also apply this test in three different ways to approximately 300,000 police stops in Connecticut. I show that all three ways suggest substantial racial profiling against blacks and hispanics in Connecticut and I discuss the limitations of each.

Elizabeth Anderson uses the phrase "empirically informed philosophy" to describe work reflecting relevant empirical research. I like this notion and here I try to take it one step further. Let's call it empirically engaged philosophy (though not experimental philosophy, in the sense this expression is traditionally understood). We will not just be relying on related empirical work – we will do that work ourselves. In some cases we will construct our models

[1] https://openpolicing.stanford.edu/data/.

and in other cases we will apply models of others. Either way, we are making the inferences. This forces us to take responsibility for the empirical conclusions as well.

The paper proceeds as follows. In section 2, I briefly describe the particular subset of data I will rely on, and I offer a brief overview of how lawyers, economists, and criminologists have defined discrimination. In section 3, I try to give a conception of discrimination of my own, which relies on considering how individuals of different races are viewed by government institutions. Impermissible discrimination exists to the extent we can conclude that state actors are deemed to hold people of different races on different moral footing. I then show that a sufficient condition for finding evidence of discrimination under this principle is to find evidence that state actors have different epistemic risk profiles with respect to people of different races. Finally, I show that a sufficient condition for finding a difference in epistemic risk profiles in the specific context of police stops is to show that officers have different search thresholds for people of different races. In sections 4-6, I apply three different tests aimed at estimating the search thresholds. First, I evaluate the marginal effect of race in both a Frequentist and a hierarchical Bayesian logistic regression with the probability of search as the dependent variable. Next, I consider and apply the famous outcome test proposed by Becker (1957) and developed by Knowles et al. (2001). Finally, I consider and apply the threshold test proposed by Simoiu et al. (2017). As I go along, I explain the limitations of each. We will see that all three tests are subject to some form of omitted variable bias when it comes to making inferences about the officer's epistemic risk profile.

2 Background

2.1 Data

Throughout this paper, I will focus on police stops from Connecticut in particular. This dataset was initially compiled by the Connecticut Data Collaborative and a raw version may be found on the group's website.[2] After some processing, the set I rely on includes 310,969 police stops, occurring between 2013-2015. I focus on Connecticut for convenience. The data is complete and the processing time for running the Bayesian Monte Carlo models is manageable. To implement the Bayesian models on the full set of 60 million stops one needs to adopt faster converging algorithms. One such algorithm for Simoiu et al. (2017)'s threshold test in particular is proposed in Pierson et al. (2017).

2.2 Beyond the impact/intent dichotomy

There are roughly two notions of discrimination referred to by various names across the relevant disciplines. This will be an imperfect grouping but it may be helpful still. The difference between the two concepts rests on the subjective mental state of the police officer. First, there is intentional discrimination – i.e., discrimination occurring due to racial bias,

[2]http://ctrp3.ctdata.org/rawdata/.

animus, prejudice, or bigotry. In legal scholarship, this is appropriately known as discriminatory purpose. If, for example, we can show that a landlord refuses to rent apartments to blacks because of their animus toward them as a group, then we have shown a purpose to discriminate on behalf of the landlord. Such discriminatory purpose is in violation of the Equal Protection Clause of the Fifth and Fourteenth Amendments and a pattern of police practice exhibiting it would be unlawful. In the employment context, this kind of discrimination is sometimes referred to as disparate treatment, following the language that has emerged out of the case law interpreting Title VII of the Civil Rights Act of 1964 (Griggs v. Duke Power Co., 401 US 424 (1971)). Economists, following Becker (1957), have given this behavior the unfortunate name 'taste-based' discrimination. It is taste-based because discrimination exists in the utility function, which is where, for economists, all matters of morals, preferences, and aesthetic value exist. To have a taste for discrimination is to have a preference for discriminating.

The second kind of discrimination is discrimination that is unintentional but results in a selection procedure that disproportionately affects a particular group. In legal scholarship this is known as discriminatory effect. In the employment context, it is known as disparate impact. And in the economic context, it has been dubbed statistical discrimination (Arrow, 1973) or business purpose (Borooah, 2001). For example, suppose a landlord imposes a high minimum income for anyone wishing to apply for residence in their apartment building because they sincerely believe that by doing so they will avoid tenants who are more likely to default on their rent. In a socioeconomically unequal society, this policy has the effect that the proportion of black tenants in this building will likely be lower than the proportion of black residents in the community at large. However, the landlord defends the policy on the ground that because the average income of blacks is lower they are more likely to default. Therefore, her policy is the result of a facially-neutral business judgment. Similarly, the landlord may turn away black applicants by simply assuming that they are more likely to command a lower income on the basis of her prior experience. Here she is discriminating because she believes that this is a statistically sound policy though not due to racial animus. Clearly non-intentional discrimination can come in different forms, and different degrees of moral turpitude, but the subjective intent, for what it is worth, remains different.

I suspect this notion of intent is likely to raise suspicion. And I think it should. In the second case above, is it really true that the landlord does not intend to discriminate if she knows that will be a consequence of her actions? Perhaps we may argue that there is a difference between intentional and merely foreseen harms and apply the doctrine of double effect (Quinn, 1989). But the fact remains that in most cases of interest, all we have to go in is observational data of behavior and we have no choice but to use it to make inferences about subjective mental states. What I will propose instead is a standard that lends itself to statistical estimation. Roughly, the standard says that if we can reliably impute to a government agency a certain perspective, then they are engaging in impermissible discrimination. Of course, 'reliably impute' is a relative term. And it should be. We will never have conclusive evidence of discrimination from empirical observation of outcomes alone. We can, however, have better or worse evidence. We will now make this perspective explicit.

3 Moral worth and epistemic risk

Suppose we found a racist guidebook and some witnesses speaking to intent with respect to a particular police department. With this kind of evidence, we would not need large scale observational data. That would be adequate to justify filing a lawsuit against the offending department or an internal investigation of its affairs. The reason we ordinarily need to make inferences about intent is because such evidence is unavailable and disparate impact in and of itself is not illegal. In Washington v. Davis, 426 U.S. 229 (1976), for example, the Supreme Court found that evidence of discriminatory impact in government hiring was insufficient to establish a violation of the Equal Protection Clause of the Fourteenth Amendment. The employment context is different because in 1991 Congress amended Title VII of the Civil Rights Act to explicitly include a disparate impact test as evidence of employment discrimination. This is the basis on which the plaintiffs in Ricci v. DeStefano, 557 U.S. 557 (2009), prevailed against New Haven for throwing out the results of a standardized test for fear of a disparate impact lawsuit. As a result, the Supreme Court did not consider their claim under the Equal Protection Clause. Indeed, in his concurrence, Justice Scalia argued that the disparate impact provisions of Title VII of the Civil Rights Act are unconstitutional under the Equal Protection Clause.

But my point is not just about the prevailing legal regime. It is typically assumed that disparate impact on its own is not necessarily morally objectionable. Consider a tax law that disproportionately affects white people because as a group they tend to be financially better off. Rather, it is *unjustified* disparate impact that is of concern. This is, I think, what the legal scholarship on discriminatory intent and purpose is getting at. *Intent* is not central to the analysis. What matters is justification. But when is disparate impact unjustified? I'll try to give one answer to this question. I propose a standard grounded in the risk of error – or, as I have called it elsewhere, epistemic risk (Babic, 2018).

The moral principle I appeal to is roughly the following: that persons of different groups should have equal moral worth under the law. This is very general. If we can fairly *impute* to a government agency the position that they are treating citizens unequally under the law, then they are violating this moral principle. In other words, a system of legal enforcement which is such that people of different groups are valued differentially is impermissibly discriminatory. The moral force of such a norm stems, appropriately, from considerations of equality – the same considerations that give rise to equal protection law to begin with. While most theories of equality do not require equal treatment they do require at a minimum the notion that differences in treatment are justifiable. In the tax law case, differences in treatment would be justified by the better-off group's ability to contribute more to the public purse relative to their financial means. This demand for justification only makes sense in the background of a norm where the moral worth of citizens ought to be equal, ceteris peribus.

This principle of equal moral worth is also often expressed in the dictum to treat like cases alike, often traced to Book 5 of Aristotle's Nicomachean Ethics. Lon Fuller associated the principle with legality itself, and saw it as a part of the internal morality of the rule

of law.[3] Similarly one can find expression of the dictum in Dworkin's defense of the law's integrity[4] and at least implicitly in Raz's analysis of coherence.[5]

The dictum to treat like cases alike, however, has little moral force in and of itself. Rather, what matters is whether differences in treatment can be justified. And the intuition for why this matters is that if differences in treatment cannot be justified then the differential across groups implies that individuals from different groups are valued differently under the law and the difference is due to their group membership. In short, the relevant norm is simple and, I hope, not too controversial:

> **Equal moral worth principle.** If a system of police enforcement exhibits disparate impact in such a way that we may reasonably impute to the relevant agency (department) the view that people from different groups are valued differently by the agency then such disparate impact is to that extent unjustified and morally wrongful. (It should probably be legally impermissible too, but I do not want to get bogged down by prevailing legal standards.)

This principle is simply a variant of equal concern principles arising in theories of distributive justice. Indeed, it is inspired by Anderson (1999)'s notion of democratic equality. Democratic equality is equally concerned with the expressive demands of equal respect as it is with resource distribution. This is to be contrasted with luck egalitarian theories (such as Dworkin (2000)) that theorists in the discrimination literature sometimes (imprudently) draw upon in order to articulate a yardstick for discrimination (e.g., Thacher (2002)). The central aim in democratic equality is equal standing in one's community of peers. It imposes on the state an unconditional obligation to respect their dignity or autonomy. Channeling Kant, Anderson writes: "every individual has a worth or dignity that is not conditional upon anyone's desires or preferences" (Anderson, 1999). This makes clear how a preference, or taste, for discrimination on behalf of the police officer would violate such a principle.

What remains to be seen is how if at all we can use the EMW principle to evaluate and construct statistical tests of discrimination from observational data involving police stops. Here is one way to proceed.

> **Epistemic risk criterion.** When an officer decides whether or not to conduct a search, they risk two types of mistakes. They could erroneously search a car without contraband (a Type I or false positive error, so to speak) or they could fail to search a car containing contraband (a Type II or false negative error). The police officer must balance these costs of error in deciding whether or not to conduct a search. In other words, they must adopt a specific subjective attitude toward epistemic risk. If we learn that this balance is made differently across race, this implies that the police officer's assessment of the cost to a person's well-being for being wrongfully searched varies by race. If this is true, then individuals from different races are not being treated equally as moral agents. That is, the harm caused to them by a wrongful search is valued differently according to their race.

[3] Fuller (1964).
[4] Dworkin (1986).
[5] Raz (1995), Chapter 13.

> If this is something we can impute to a police department then we have evidence of unjustified disparate impact – i.e., evidence of wrongful discrimination under the EMW principle.

This approach extends what I think is the guiding moral principle behind theories of equality which require false positive parity though it does not necessarily require it. What it requires is more specific – a difference in the way error costs are weighted. In what follows, we will see how the prevailing tests in the statistical literature fare with respect to the normative guideline we have thus established.

4 A Lockean discrimination test

When a police officer stops a driver, they observe some evidence regarding their culpability. This evidence informs their credence regarding the hypothesis that there are drugs in the car. But an officer observes more than this. They also observe things like the driver's race, age, and gender. Features that they should not be conditioning on. After all the evidence is in, they decide whether or not to conduct a search. In other words, the officer has some implicit, unobserved credal threshold t above which they conduct a search. Consider the following test of discrimination. I call it the Lockean test by loose analogy to the threshold analysis of belief and assent Locke gives in book IV, ch. XV of *An Essay Concerning Human Understanding*.

> **Lockean discrimination test**. Let t_i be an officer's search threshold for group i and t_j the search threshold for group j. If $t_j < t_i$ then group j is being discriminated against.

I will now show that if we find evidence of discrimination under the Lockean threshold test, then we have evidence of discrimination under the epistemic risk criterion. In other words, evidence of discrimination under the Lockean test is evidence of wrongful discrimination per the EMW principle.

Consider the following setup. The officer, our decision-maker, makes a stop and must decide whether or not to conduct a search. The parameter space consists of two possible true states of the world (no drugs in car, drugs in car), given by, $\Omega := \{\theta_0, \theta_1\}$. The officer can make two types of decisions, do not search the car, and search the car, given by the decision space, $D := \{s_0, s_1\}$. We can think about s_1 as corresponding to rejecting $H_0 : \theta = \theta_0$ and s_0 corresponding to rejecting $H_1 : \theta = \theta_1$. There is a loss function, $L : \Omega \times D \to \mathbb{R}^+$, given as follows: If θ_0 is true and she does not search (true negative) suppose her loss is 0. If θ_0 is true but she does conduct a search (false positive) suppose her loss is ℓ_0. If θ_1 is true and she does not conduct a search (false negative) suppose her loss is ℓ_1. Finally if θ_1 is true and she does not conduct a search (true negative) suppose her loss is 0. Therefore, the agent's loss $L(\theta_i, s_j)$ is given by the following table:

	No Search	Search
No Drugs	0	ℓ_0
Drugs	ℓ_1	0

Table 1: Loss function

We do not want to assume that $\ell_1 = \ell_0$. Indeed, that they can differ will be central to my argument. Therefore, the associated loss is generalized $0-1$ loss. This form for the loss function is forced by the nature of the problem as one of making a binary choice with respect to whether or not to search the vehicle in a state of uncertainty characterized by two mutually exclusive and exhaustive simple hypotheses. Once the officer decides to conduct a search, if H_0 is true then the loss is the same regardless of just how high above the threshold their probability is. So we cannot use a strictly monotonically decreasing loss, for example. By the same token it means the loss is not proper but such is the nature of the officer's decision problem.

Let $f(\mathbf{x})$ be the joint sampling distribution of the feature vector (the evidence the officer observes) for a sample $\mathbf{x} \in \mathcal{X}$. $f(x_i)$ and $f(x_j)$ will often not be independent for many i and j but we will assume that they follow the same parametric distribution. For example, we might have, as we will see X_1 (race), X_2 (age) and X_3 (gender). We also assume the officer has a coherent prior distribution over Ω. Let $\Delta(\Omega)$ be the set of all σ-additive probability measures on Ω and assume the officer has a prior probability for true unknown θ given by $\pi(\theta) \in \Delta(\Omega)$. $\pi(\theta_0)$ and $\pi(\theta_1) = 1-\pi(\theta_0)$ are their prior probabilities for the two hypotheses. For any one stop, θ follows a categorical distribution.

The officer will observe some evidence $\mathbf{x}$ after making the stop and compute a posterior probability $\pi(\theta|\mathbf{x}) \propto \pi(\theta)f(\mathbf{x}|\theta)$. After processing the evidence and obtaining a posterior probability she must decide whether to search the car, s_1, or not search the car, s_0. Therefore, let $\delta : \mathcal{X} \to D$ be the officer's decision rule for whether or not to conduct a search. A rational Bayesian agent (this much we must assume) should choose the act with the relatively higher posterior probability. That is, choose s_1 if $\pi(\theta_1|\mathbf{x}) > \pi(\theta_0|\mathbf{x})$. This assumes however that $\ell_0 = \ell_1$. Given the duality between loss and prior, as Lindley (1985) and Robert (2007) put it, what a rational Bayesian agent actually does is to maximize expected utility by minimizing posterior expected loss. It is only in the special case where the two costs of error are equal that the posterior expected loss is minimized by choosing the act with the relatively higher posterior probability.

To balance the relevant consequences in identifying a decision procedure we need to pay attention to the relative costs of the different types of error. There are two ways to do this. We can do it from the officer's prior perspective. That is, we can consider a non-biased decision procedure that the officer settles on in advance. Or we can do it from a posterior perspective. That is, we can ask what posterior probability does either type of decision require. This is the usual presentation in Bayesian texts. But it is an unusual way to divide things since these two approaches should yield the same answer. And indeed they do. However we approach the matter, the decision procedure depends on, and only on, the ratio of the cost of false positive to false negative mistakes. This is the rational officer's

search threshold t. Let us see why this is true.

4.1 Prior frame of mind

The Type I and Type II error rates are functions of δ and we can refer to them as $\alpha(\delta)$ and $\beta(\delta)$, respectively. In other words,

$$\alpha(\delta) = Pr(s_1|\theta_0)$$
$$\beta(\delta) = Pr(s_0|\theta_1)$$

The agent's total loss incurred due to false positive error is $\ell_0\alpha(\delta)$. The probability of this loss is given by $\pi(\theta = \theta_0)$. The agent's total loss due to false negative error is given by $\ell_1\beta(\delta)$ and the probability of this loss is given by $\pi(\theta = \theta_1)$. Now consider the officer's prior expected loss, given by

$$\mathrm{E}_{\pi(\theta)}[\ell(s_i, \theta_j)] = \pi(\theta_0)\ell_0\alpha(\delta) + \pi(\theta_1)\ell_1\beta(\delta) \tag{1}$$

To compute these terms we must divide the sample space into the region of possible data under which the officer would conduct a search, call it S_1 (i.e., if $\mathbf{x} \in S_1$ then the officer takes action s_1), and the region of possible data under which the officer would not conduct a search, call it S_0 (i.e., if $\mathbf{x} \in S_0$ then the officer takes action s_0). $S_1 \cup S_0 = \mathcal{X}$.

The term $\alpha(\delta)$ may then be computed as the sum of all cases where $\mathbf{x} \in S_1$ but the distribution of the evidence is governed by the true parameter θ_0. Likewise, $\beta(\delta)$ is the sum of all cases where $\mathbf{x} \in S_0$ but the distribution of the evidence is governed by θ_1. Notice, then, that the officer's prior expected loss is simply a linear combination of false positive and false negative error rates weighted by their severity. As DeGroot and Schervish (2012) point out, we can rewrite this expression as follows.

$$\mathrm{E}_{\pi(\theta)}[\ell(s_i, \theta_j)] = \pi(\theta_0)\ell_0 \sum_{\mathbf{x}\in S_1} f(\mathbf{x}|\theta_0) + \pi(\theta_1)\ell_1 \sum_{\mathbf{x}\in S_0} f(\mathbf{x}|\theta_1) \tag{2}$$
$$= \pi(\theta_1)\ell_1 + \sum_{\mathbf{x}\in S_1} [\pi(\theta_0)\ell_0 f(\mathbf{x}|\theta_0) - \pi(\theta_1)\ell_1 f(\mathbf{x})|\theta_1] \tag{3}$$

where $f(\cdot|\theta)$ is the probability distribution of the data under θ. In other words, we want the region that includes every point x for which $\pi(\theta_0)\ell_0 f(\mathbf{x}|\theta_0) - \pi(\theta_1)\ell_1 f(\mathbf{x}|\theta_1) < 0$ because every such point will decrease the overall sum. Therefore, the search procedure δ^* that minimizes the prior expected loss is to conduct a search when $\pi(\theta_0)\ell_0 f(\mathbf{x}|\theta_0) > \pi(\theta_1)\ell_1 f(\mathbf{x}|\theta_1)$.

As a result, we will conduct a search whenever the probability of the evidence under the hypothesis that there is no contraband in the car, weighted by its prior probability and the cost of searching a car without contraband in it, is less than the probability of the evidence under the hypothesis that there is contraband in the car, weighted by its prior probability and the cost of failing to search a car with drugs in it. Rearranging and expressing the

statistic as a function of the true unknown state θ, we get a weighted likelihood ratio test. That is, search the car if,

$$\frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} > \frac{\pi(\theta_1)\ell_1}{\pi(\theta_0)\ell_0} \tag{4}$$

From our Bayesian perspective, we recognize that this is equivalent to,

$$\frac{\pi(\theta_0|\mathbf{x})}{\pi(\theta_1|\mathbf{x})} > \frac{\ell_1}{\ell_0} \tag{5}$$

In other words, the officer conducts a search if the posterior odds exceed a ratio of error costs. Even simpler still, we can say the following: conduct a search if,

$$\pi(\theta_1|\mathbf{x}) > \frac{\ell_0}{\ell_0 + \ell_1} \tag{6}$$

This will be our final expression. Let $t = \ell_0/(\ell_0 + \ell_1)$. This is the rational officer's search threshold. Since t is equal to $(\ell_0/\ell_1) \times \ell_1/(\ell_0 + \ell_1)$, t is a function of the ratio $\ell_0 : \ell_1$. In other words, to determine the search threshold t, for a rational Bayesian officer who makes a decision on the basis of their posterior probability, it is sufficient to know the relative cost this officer assigns to falsely searching a car without drugs in it against failing to search a car with drugs in it.

In sum: the rational officer searches a vehicle if the posterior probability exceeds a search threshold determined by the relative cost of different types of error. Now our desired goal is immediate: we have said that balancing the cost of errors differently violates the equal moral worth principle. But we have now seen that if the search threshold varies by race this implies that the errors are in fact balanced differently. So, a difference in thresholds implies a difference in epistemic risk profile implies a difference in moral worth.

4.2 Posterior frame of mind

Above, we derived the Lockean threshold test by starting from the supposition that the rational officer seeks to minimize prior expected loss. Before we move on, let us see that we can do the same thing by starting directly from the posterior distribution. Consider the agent's posterior expected loss, given by $\mathrm{E}_{\theta|\mathbf{x}}\ell(\theta, s)$. This is the average loss under the possible values of θ weighted by their now posterior probabilities.

$$\mathrm{E}_{\theta|\mathbf{x}}[\ell(\theta, s)] = \int_{\Omega} \ell(\theta, s)\pi(\theta|\mathbf{x})d\theta \tag{7}$$

Since $\Omega = \theta_0 \cup \theta_1$ we can rewrite this as,

$$\mathrm{E}_{\theta|\mathbf{x}}[\ell(\theta, s)] = \pi(\theta_0|\mathbf{x})\ell_0 + \pi(\theta_1|\mathbf{x})\ell_1 \tag{8}$$

This suggests that we should conduct a search if,

$$\pi(\theta_1|\mathbf{x})\ell_1 < \pi(\theta_0|\mathbf{x})\ell_0 \tag{9}$$

In other words, we should conduct a search if,

$$\frac{\pi(\theta_0|\mathbf{x})}{\pi(\theta_1|\mathbf{x})} > \frac{\ell_1}{\ell_0} \tag{10}$$

Which is again equivalent to conducting a search if,

$$\pi(\theta_1|\mathbf{x}) > \frac{\ell_0}{\ell_0 + \ell_1} \tag{11}$$

Letting $t = \ell_0/(\ell_0 + \ell_1)$ as before, we have again reached a threshold test whose level is determined by the officer's relative attitude to risk of error.

5 Linear models and marginal effects

The natural place to start if we're going to look for evidence of racial profiling is to perform some regressions. Let's do this and see what conclusions, if any, we can draw regarding discrimination as operationalized by the EMP principle and the epistemic risk criterion. The idea in multivariate modeling is to construct some sort of linear model of police decision-making after a stop has occurred and estimate the effect of race on the probability of a search. If the marginal effect of race, holding other things equal, is substantial, then this is evidence of a difference in the search thresholds. The problem with this model, as is often case in linear modeling, is omitted variable bias. For example, even though race has a significant effect in a model that includes race, age and gender, it may not be informative in a model that includes, in addition, the type of car being driven, whether or not the driver's windows are tinted, and so forth. However, we will see that a version of omitted variable bias affects other prevailing tests as well, so it will be instructive to pursue this model further and see what conclusions we can make.

For each stop in the Connecticut data, we have evidence of the driver's race, age, and gender. These are the 'objective' features, so to speak, that a police officer observes and that we have access to. However, it is unreasonable to suppose that a police officer discerns a driver's precise age, so we will discretize the age variable into several categories, 16-25, 26-39, 40-49, and 50+. There are undoubtedly other relevant features that an officer observes but unfortunately we do not have data on these features. First, we perform an ordinary logistic regression. Then we apply a hierarchical Bayesian model. Both results suggest discrimination against blacks and hispanics in Connecticut.

5.1 Logistic regression

We will model the probability that a driver gets searched given their race, age, and gender and evaluate the marginal contribution of race. Let Y be a binomially distributed random variable that takes the value 1 with probability θ if a stopped driver is searched and the value 0 with probability $1-\theta$ if a stopped driver is not searched. Y follows a binomial distribution

with mean θ. Let $\mathbf{X}_i$ be a vector of random variables representing a driver's race, age, and gender and $\mathbf{X}$ the corresponding matrix for our data. We want a model that uses $\mathbf{X}_i$ and outputs $Pr(Y_i = 1)$.

For a binary response as ours, regression by least squares could work. That is, we could use $\boldsymbol{\beta}\mathbf{X}$ to estimate $Pr(Y = 1)$ where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$ are the ordinary regression coefficients for race, age, and gender, respectively. However, by fitting a straight line to a 0/1 variable we cannot guarantee that some of the estimates will not be less than zero or greater than 1. Since we want our estimates to be probabilities, it is preferable (and customary) to apply a sigmoid link to the response, such as the logistic function. As a result, we will use the following model.

$$Pr(Y = 1|\mathbf{X}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}\mathbf{X})}{1 + \exp(\beta_0 + \boldsymbol{\beta}\mathbf{X})} \tag{12}$$

Equivalently,

$$\log\left(\frac{Pr(Y = 1|\mathbf{X})}{1 - Pr(Y = 1|\mathbf{X})}\right) = \beta_0 + \beta\mathbf{X} \tag{13}$$

Since the left-hand side is the log odds of getting searched, the logistic regression model is said to have a logit that is linear in the predictor variables. As a result, after specifying some values for $\mathbf{X}$ within the model and applying the sigmoid link to the raw response y, i.e., $e^y/(1 + e^y)$, we get the estimated probability that a person with a particular set of characteristics will be searched. The following table displays the results of this analysis applied to the Connecticut data.

Coefficient	Estimate	Standard Error	p-value
Intercept	-4.44	0.04	$< 2 \times 10^{-16}$
Race = black	0.83	0.03	$< 2 \times 10^{-16}$
Race = hispanic	0.68	0.04	$< 2 \times 10^{-16}$
Age = 26-39	-0.50	0.03	$< 2 \times 10^{-16}$
Age = 40-49	-1.30	0.04	$< 2 \times 10^{-16}$
Age = 50+	-1.90	0.06	$< 2 \times 10^{-16}$
Gender = Male	1.03	0.04	$< 2 \times 10^{-16}$

Table 2: Logistic Regression Summary

Note that the p-values for each parameter are less than .0000000000000005 so there is little ambiguity regarding their statistical significance. Further, most of the effect sizes are reasonably substantial. For example, a young (age 16-25) white male has a 3% chance of getting searched whereas a young black male has a 7% chance of getting searched. The relative risk ratio among young men is 2.2. Therefore, among young men, the risk of getting searched doubles if one is black. It is likewise approximately double if one is hispanic.

More generally, we can look at the effect of being black or hispanic on one's probability of getting searched across all demographic categories. This is often done by comparing the

probability of, say, a black person getting searched to a null (intercept-only) model, but this would be misleading because the null model is essentially giving us the probability that a young white woman is getting searched. Rather, we want to predict the likelihood of getting searched if one is black for each different variable and average across them in order to isolate the marginal effect of being black across all demographic categories. We do this by comparing what the model predicts on average, if we set everyone's race to white, against the predictions if we set everyone's race to black. The average probability of getting searched for a white person, holding all other covariates constant, is 1.3% whereas the average probability of getting searched for a black person, holding all other variables constant, is 3%. The relative risk ratio is 2.23 across all demographics. While the percentage chance of being searched decreases (since we are now including women and older people) the relative risk ratio remains similar – i.e., the chance of being searched doubles if one is black.

5.2 Hierarchical Bayesian model

But perhaps you are skeptical of the preceding Frequentist analysis. We can also take a Bayesian approach. This approach is conceptually quite different. The model I will develop is a hierarchical Bayesian logistic regression model. We assume again that the probability of being searched, our response, follows a binomial distribution with unknown mean θ_i for $i \in [1, n]$. However, θ_i is now a random variable about which we want to make inferences and it can be expressed in terms of the data and the parameter coefficients because the logit of θ_i is given by $\beta_0 + \boldsymbol{\beta}\mathbf{X}_i$ and the parameters about which a Bayesian agent has prior beliefs are each of the β coefficients – i.e., the intercept and each of the remaining (six) levels corresponding to our demographic categories. To initiate the model we have to specify a prior distribution for each of the parameters. Based on the maximum likelihood estimates of the coefficients, above, the parameters are not too far from zero so to keep things simple we will assume a standard normal prior distribution. In sum, the hierarchical Bayesian model is given as follows:

$$
\begin{aligned}
&y_i|\theta_i \overset{iid}{\sim} \text{Bin}(1, \theta_i) \text{ for } i \in [1, n] \\
&\text{logit}(\theta_i) = \beta_0 + \boldsymbol{\beta}\boldsymbol{X}_i \text{ where } \boldsymbol{\beta} = (\beta_1, ...\beta_6)^T \\
&\beta_0 \overset{iid}{\sim} \text{N}(0, 1) \\
&\beta_j \overset{iid}{\sim} \text{N}(0, 1) \text{ for } j \in [1, 6]
\end{aligned}
$$

Given the above information, the likelihood, or sampling distribution, for each i may be written in terms of $\boldsymbol{\beta}$ as follows,

$$
\pi(y_i|\boldsymbol{\beta}, \boldsymbol{X}_i, n_i) \propto \{\text{logit}^{-1}(\beta_0 + \boldsymbol{\beta}\boldsymbol{X}_i)\}^{y_i}\{1 - \text{logit}^{-1}(\beta_0 + \boldsymbol{\beta}\boldsymbol{X}_i)\}^{n_i - y_i} \tag{14}
$$

And as before,

$$
\begin{aligned}
&\Pr(y_i = 1) = \text{logit}^{-1}(X_i\boldsymbol{\beta}) \\
&\Pr(y_i = 0) = 1 - \Pr(y_i = 1)
\end{aligned}
$$

Our model, therefore, is characterized by β_0 and $\boldsymbol{\beta} = (\beta_1, ...\beta_6)^T$ whose joint posterior distribution is given by,

$$\pi(\beta_0, \boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{X}) \propto \pi(\beta_0, \beta_1, ...\beta_6) \prod_{i=1}^{k} \pi(y_i|\beta_0, \beta_1, ..., \beta_6, n_i, x_i) \tag{15}$$

Computing the posterior distribution of each parameter mechanically is not feasible. For example, consider the prior only. The joint distribution of all seven parameters is given by

$$\pi(\beta_0, \boldsymbol{\beta}) = \prod_{i=1}^{7} (1/2\pi)^{-1/2} e^{-\beta_i^2/2} \tag{16}$$

We also have a joint sampling distribution for more than a quarter million data points without a nice closed form expression. Meanwhile, the normal prior density is not conjugate to the binomial likelihood. Finally, to get the posterior estimate of each parameter, we would have to integrate out the remaining nuisance parameters. That is,

$$\pi(\beta_i|\boldsymbol{x}) = \int_{...} \int_{j \neq i} \pi(\beta_0, \boldsymbol{\beta}|\boldsymbol{x}) \partial \beta_{j \neq i} \tag{17}$$

As a result, I estimate the posterior distributions of each regression coefficient through Hamiltonian Monte Carlo sampling, a form of Markov Chain Monte Carlo (Neal, 1995; Duane and Roweth, 1987; Mackay, 2003). In particular, I use the No-U-Turn sampler (NUTS) implemented in Stan, an open-source modeling language for Bayesian inference (Carpenter et al., 2016). In implementing a Stan sampler we need to pick a number of chains and the number of iterations for each chain. I took 8000 samples: 4 sequential chains, each with 2000 iterations. The idea behind MCMC sampling for Bayesian inference is that if the relevant assumptions are met the Monte Carlo algorithm constructs the Markov chain that converges to a stationary distribution. By devising a Markov chain whose stationary distribution is our desired posterior distribution we can run it to get draws that are approximately from the posterior once the chain has converged.

In practice, this means that we want to discard some initial draws as "burn-in" before the chain has converged to its stationary distribution. I discarded the first 1000 draws from each chain. This appears to be sufficient for convergence of the posterior to a stationary distribution, as indicated by the trace plots of post-warm-up iterations associated with each parameter of interest. A trace plot is a time series plot of the parameter that we monitor as the Markov chain proceeds. Trace plots provide a graphical assessment of the behavior of the Monte Carlo sampler with respect to each fitted parameter. The plots, Figure (1), suggest the sampler is mixing well.

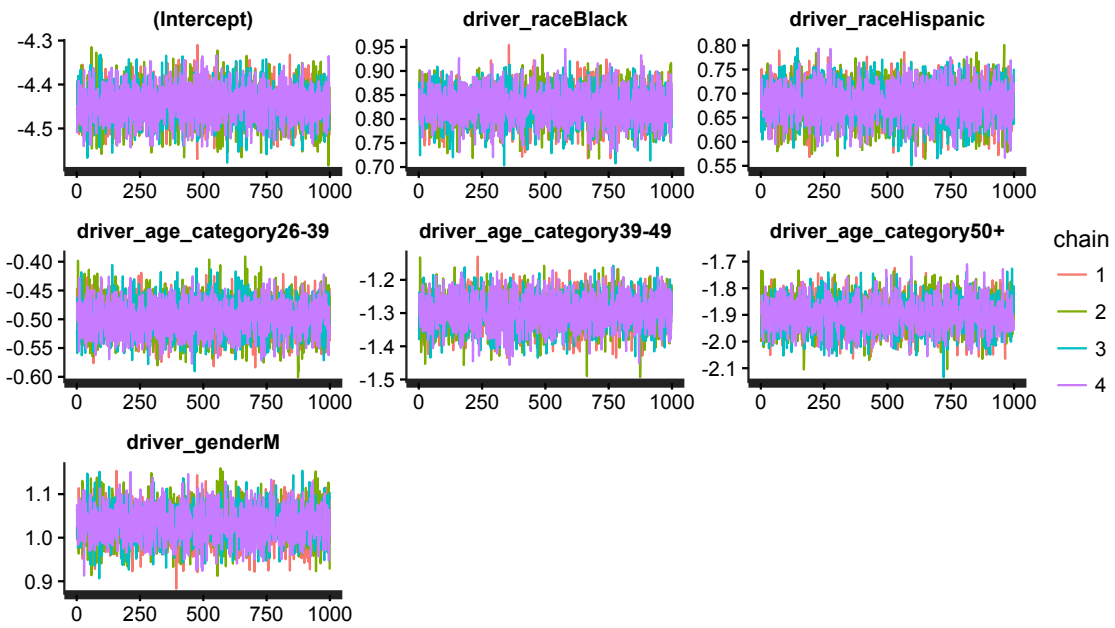

Figure 1: Trace plots of MCMC samples for each regression parameter

Further, if the Gelman-Rubin convergence statistic, $\hat{R}$, for each parameter are near their nominal value of 1, as they are here, this suggests that the assumption of ergodicity is satisfied (Gelman et al., 2013). $\hat{R}$ compares the variance of each chain to the pooled variance and provides an estimate of how much variance could be reduced by increasing the length of the chains. The statistic approaches 1 as the chains run longer. This is important because we need our chain to be ergodic. If this condition is met, then with probability 1 the mean of the sampled values approaches the mean of the stationary distribution in the limit. This is like a strong law of large numbers for Markov chains. It allows us to ignore the dependence between draws of the Markov chain when we calculate quantities of interest (for us this will be the posterior means) from the draws.

Table (2) summarizes the mean of the posterior distributions of each parameter along with their 95% credible interval (which is to be distinguished from the ordinary Frequentist confidence interval). Notice the difference in the summary statistics. Instead of a maximum likelihood estimate and a p-value computed using the z-statistic, we now have a posterior mean and its credible interval to be interpreted, literally, as we are 95% confident that the true value of the parameter is within this very tight range.

Coefficient	Mean	2.5%	97.5%
Intercept	-4.4	-4.5	-4.4
Race = black	0.8	0.8	0.9
Race = hispanic	0.7	0.6	0.8
Age = 26-39	-0.5	-0.6	-0.4
Age = 40-49	-1.3	-1.4	-1.2
Age = 50+	-1.90	-2	-1.8
Gender = Male	1	1	1.1

Table 3: Marginal posterior coefficient estimates

Figure (2) displays the information from Table (2) in graphical form. Each posterior mean is displayed with its tight 95% credibility interval. This makes it easy to see which variables increase one's risk of getting searched, which ones decrease it, and their relative effect.

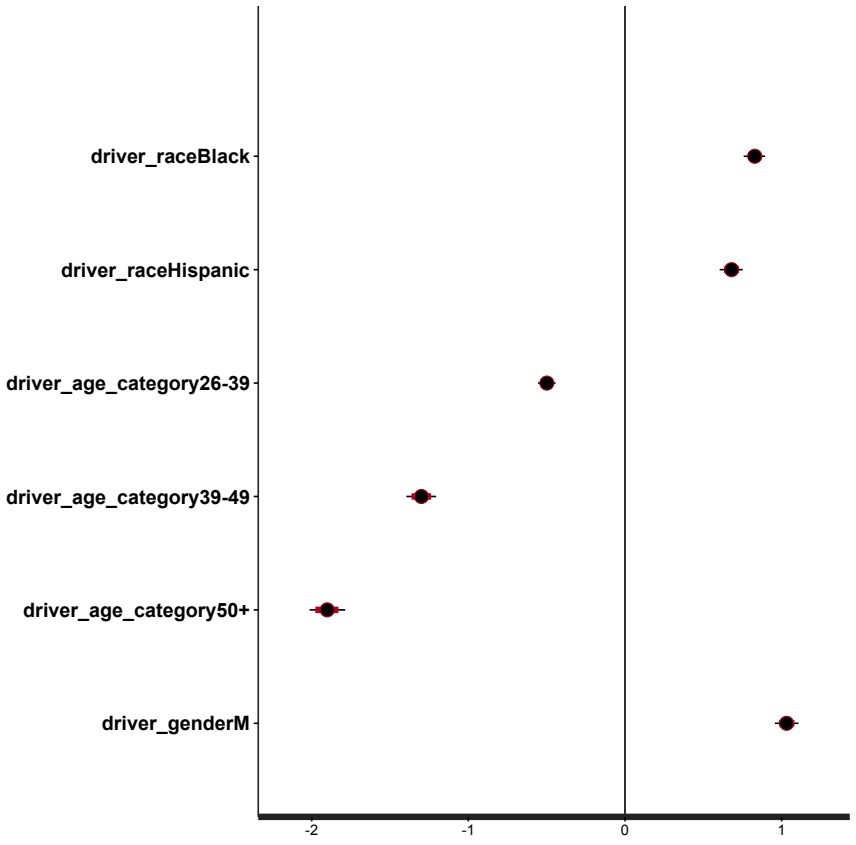

Figure 2: Posterior marginal distributions of regression coefficients

Overall the Bayesian analysis yields similar conclusions. Notice the similarity between the MLE estimates of the regression parameters (Table 1) and their Bayesian posterior means (Table 2).

Notice, also, the extent to which this approach is *not* Bayesian in the philosopher's sense. If we found that the trace plots are not mixing well, or that the $\hat{R}$ values suggest the ergodicity assumption is not satisfied, this would suggest that our model was incorrectly specified. But if the model is incorrectly specified, we do not have a Bayesian means by which to navigate across the model space (Gelman and Shalizi, 2013). Rather, we would re-group and set up a different model. Which model we move to would depend in part on our subjective diagnosis of what went wrong. This implies, for what it's worth, there is little hope of dynamic coherence across the model space.

5.3 Inferring discrimination

This analysis likely confirms some suspicions we might have had. Being black or hispanic in Connecticut is quite a strong indicator of getting searched assuming one is stopped. If we had to bet on who gets searched, one would want to bet on young black men. This is true both from the maximum likelihood analysis and the Bayesian analysis. The problem is that a logistic regression, Bayesian or otherwise, is equivocal with respect to whether wrongful discrimination as have defined it is occurring.

If young black men are especially likely to exhibit non-race related characteristics of culpability then we would expect race to be significant in the model even if the police do not in fact treat drivers from different races differently in violation of the EMW principle. For example, a training manual issued by the Illinois State Police includes, as relevant factors, characteristics as tinted windows, cell phones (it's a manual from the 90's!), leased cars, religious paraphernalia (for some reason), and attorney business cards (Knowles et al., 2001). We could, in theory, evaluate the role of race in regression involving all such indicators of criminality. The problem, however, is not only that we don't have data including all such variables, but even if we did have a much richer dataset it could always be possible that other relevant features have been left out.

Before we move on, consider one way to avoid the problem of omitted variable bias. This will be informative as it comes up again in the assessment of other tests for discrimination. Suppose we assume that drivers and police officers are both rational actors, the first group aiming to minimize the probability of being searched and the second group aiming to maximize the likelihood of finding drugs. For example, police might find that tinted windows are indicators of contraband and search more cars with tinted windows. Drivers will then respond by avoiding cars with tinted windows in order to reduce their risk of being searched. Eventually, tinted windows will lose their probative value. The same analysis will be true for other observational cues of criminality, like religious paraphernalia or attorney business cards. If we make these assumptions, then under idealized circumstances all that is left for the police officer to go on is those features that the drivers can't, or wouldn't, change – i.e., endogenous variables like the driver's race, age, and gender, which are precisely the variables included in our model. If this is true, then by definition there are no omitted variables under equilibrium and the regression analysis is informative of discrimination under the epistemic risk criterion and therefore the EMW principle. Since the probability of being searched, given that one is black is greater than the probability of being searched, given that one is

white, other things being equal, then under idealized conditions the officer's threshold for searching blacks cannot be equal to their threshold for searching whites.

6 Search and hit rates

To avoid problems of omitted variable bias, economists and statisticians turn to analysis of outcomes. Rather than performing a regression, we might start, naively, by simply comparing the proportion of blacks who are searched against the general search rate in the population or the white search rate. This is common when expert witnesses are called to testify in cases involving racial discrimination.[6] If we take this approach, we find that approximately 1% of white drivers and 3% of black drivers are searched. The search rate is approximately three times higher among black drivers. This is consistent with the risk ratio we found in our logistic regressions. But we have not avoided the problem since, again, it might be that black drivers are more likely to drive cars with tinted windows and keep attorney business cards lying around, indicating to the police they should perform a search. If this is the case, then a differential in search rates would not be evidence of discrimination under the EMW principle.

As a result, an alternative is to look at the hit rates – i.e., compare the proportion of searched cars of black drivers in which drugs are in fact found to the proportion of searched cars of white drivers in which drugs are found. What would this tell us? Note that the false positive rate is equal to 1−hit rate, so the hit test is essentially a test that requires false positive parity. The hit rate test was originally proposed by Becker (1957). It has since been applied to examine disparate treatment of minorities in a variety of contexts, including bail decisions (Ayres, 2001), commercial mortgage lending (Becker, 1993), organ transplants (Ayres, 2005), and social determinants of academic citations in law reviews (Ayres and Vars, 2000).

Applying this test to the Connecticut data, we find that 38% of searched white drivers have contraband in their cars compared to 28% of searched black drivers. The lower hit rate among black drivers would suggest that they are being targeted by the police due to their taste for discrimination. But false positive disparity is compatible with the absence of discrimination if the underlying latent indicators of criminality are differently distributed across the groups (Borsboom et al., 2008). This is the problem with the hit rate test.

It was proposed by Becker as a way of detecting racial prejudice. Prejudice, in the form of a taste for discrimination, is sufficient to violate the EMW principle. But how do we go from false positive disparity to an inference about prejudice? Here's how Anwar and Fang (2006) put it: If racial prejudice is the reason for racial profiling, then the success rate against the marginal minority driver will be lower than the success rate against the marginal white driver (pg. 128). This has led to the problem of infra-marginality.

First, what is the normative significance of the marginal driver – i.e., "the last minority driver deemed suspicious enough to be searched" (Anwar and Fang, 2006). The significance

[6] Wilkins v. Maryland State Police, MJG 93-468 (D.Md.), 1996.

is that the marginal search rate is economists' speak for the search threshold. If this threshold is different across groups, then drivers from the two different groups are being treated differently. From our perspective, this is unjustifiable discrimination. In other words, Becker proposed the hit rate test as a proxy for the Lockean threshold test. As a result, this test is only useful to us insofar as we can derive from it information about marginal success rates (thresholds).

But when we computed the statistics, above, for the hit rates what we actually computed was the average success rates. And this is really all that we can observe. We can never directly observe search success rates against the marginal driver. Equivalently, in order to identify the marginal driver (the minimally suspicious driver to be searched), we would need to have information on every variable a police officer might use in determining a driver's degree of suspicion. But as we said in the marginal effects discussion, this is not something we can be assured of having due to omitted variable bias.

As a result, since we can only observe the average success rate and not the marginal success rate we can't be sure that a difference in hit rates constitutes a violation of the EMW principle. Simoiu et al. (2017) give a nice illustration of this. Suppose we have a police department that applies the same search threshold to two groups, A and B. By hypothesis, then, the marginal search rates are equivalent and they are not discriminating. However, suppose further that the underlying distributions of apparent criminality between the two groups are different. In particular, the distribution of group A has higher variance than the distribution of group B but their mean is the same. If this is true, it will be harder to distinguish guilty from non-guilty drivers among group B. As a result, we can have a situation where the search threshold remains the same and yet searches of drivers from group B will be less successful on average. Equivalently, their false positive rate will be higher even though, by hypothesis, they are not being treated differently. The opposite is also true. We can have a case where the true underlying marginal search rates are different, and the groups are being treated differently in violation of the EMW principle, but the difference does not show up as a difference in false positive rates.

In a seminal paper, Knowles et al. (2001) (KPT for short) develop a theoretical model regarding driving behavior and show that in equilibrium the infra-marginality problem does not arise. KPT assume that the police officer wants to maximize the number of successful searches (minus some cost to them of performing a search). For a racially prejudiced officer this cost will be different as compared to a non-racially prejudiced officer. Meanwhile, the driver wants to avoid getting caught. They most prefer carrying drugs without getting searched, and least prefer getting caught with drugs. Not carrying drugs adds zero to their utility. This means that as the probability of getting searched increases, a driver is less likely to carry drugs. This is a zero-sum game, akin to matching pennies, whose Nash equilibrium will be in mixed strategies for both parties. Given KPT's setup, they should randomize. If police are not discriminating, all drivers, if they are searched at all, must in equilibrium carry drugs with equal probability regardless of their race and other characteristics. As a result, there is no difference in equilibrium between the marginal driver and the average driver. Therefore, if we observe a difference in hit rates the difference is due to unjustified discrimination.

There are several problems with this. First, in KPT's model, driver characteristics are either exogenous (drivers cannot choose them) or they are endogenous and wash out in equilibrium. For example, they first set up the model by supposing that drivers do not get to choose whether to tint their windows. They then relax this assumption and show that even if they could choose whether or not do so, the informativeness of tinted windows would wash out in equilibrium, in the same way that we described previously. The same is true of any other characteristic observable by the officer that can be modeled endogenously. Therefore, in equilibrium, there can be no such thing as omitted variable bias. But the concern remains that KPT's model defines away the problem rather than solving it. If you are convinced that drivers and police are rational actors in approximately equilibrium conditions, then the marginal effects analysis I started with appears to be equally appropriate.

7 Direct threshold tests

The problem with the hit rate test was that we could not make an inferential leap from the average rate to the marginal rate, where the significance of the marginal rate was that it corresponds to the officer's search threshold. But what if we construct a model that estimates the search threshold directly? Instead of using statistical tests as proxies for the theoretical Lockean threshold test, we can implement the Lockean threshold test itself. This is what Simoiu et al. (2017) recently proposed.

They articulate a novel approach which treats the threshold, t_{race}, as a latent random variable to be inferred from the data and then use a hierarchical Bayesian model (similar to the model we used above) to make inferences about t_{race}. Given our conception of unjustified discrimination this is the most direct way to find it. Evidence of a difference in thresholds is, as we have seen, equivalent to evidence of a difference in epistemic risk profiles.

Their model proceeds as follows. Given the race of the driver and the department of the officer, the officer observes a signal p, which is to be interpreted as the likelihood of culpability (i.e., the chance that there are drugs in the car). In other words, the signal comes from a race and location specific distribution. We then assume that the officer, in the manner of Lewis (1980)'s Principal Principle, matches their credence to the chance as indicated by the signal. A search is then conducted if the signal strength exceeds the officer's credal threshold, i.e., if $p > t$. If a search is conducted then drugs are found with probability p. Therefore, both the search threshold and the parameters of the signal distribution are treated as latent variables to be inferred from the data. By applying the model to the data, we can make inferences about each of the latent variables.

If we apply this model to the Connecticut data we find that the search threshold for whites is [ADD] and the search threshold for blacks is [ADD]. In their original paper, the authors show that blacks face a lower search threshold in North Carolina (7% for blacks compared to 15% for whites).

The threshold test is not measuring false positive parity. Indeed, this was never the point of Becker's hit rate test. The point of that test was to use information about outcomes

to make inferences about thresholds. But as we saw we cannot do this except in the sort of idealized circumstances articulated in KPT. With Simoiu et al. (2017)'s approach we estimate thresholds directly.

Some problems remain, however. While this approach avoids the problem of infra-marginality – a problem that was created by Becker's attempt to go from hit rates to thresholds – it leaves itself vulnerable to its own version of omitted variable bias. The authors give an example of their own:

> If officers have a lower threshold for searching drivers when they suspect possession of cocaine rather than marijuana, and black and Hispanic drivers are disproportionately likely to be suspected of carrying cocaine, then the threshold test could mistakenly infer discrimination where there is none. Unfortunately, the suspected offense motivating a search is not recorded in our data, and so we cannot directly test for such an effect (pg. 1211).

This sort of thing would be true for any other characteristic associated with culpability that is not recorded in the data. So by solving one interpretive problem we leave ourselves open to another. Of course, if we assume the KPT rationality assumptions then we could equally defend the threshold test - if there exist endogenous indicators of cocaine possession, and drivers are aware such indicators affect an officer's search threshold, then drivers would respond accordingly and any such indicator would lose its probative value in equilibrium.

8 Concluding remarks

While all three tests produced similar results the approaches are strikingly different. Simoiu et al. (2017)'s test explicitly assumes that officers *do not* observe race in order to compare thresholds across race. Their defense of this is that it is unlawful for officers to condition on race. I find this to be a questionable rationale for the modeling choice. After all, we are in the business of searching for discrimination so it seems odd to assume that one kind of it is not happening in order to evaluate the extent to which the other is. Becker's approach attempts to draw conclusions about thresholds on the basis of information about average outcomes. Meanwhile, the marginal effects analysis uses the relative contribution of a driver's race to the probability of getting searched in order to draw inferences about the officer's threshold for conducting a search. All three approaches face some variant of omitted variable bias, and all three can be defended under idealized conditions. Perhaps, then, this is the most we can hope for. The contribution of this essay has been to articulate a normative standard for unjustified disparate impact – the Lockean threshold test, due to its interpretation in terms of epistemic risk – and to construct a Bayesian hierarchical model for evaluating the relative role of race (the Bayesian marginal effects model).

References

Anderson, E. (1999). What is the point of equality? *Ethics 109*, 287–337.

Anwar, F. and H. Fang (2006). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review 96*, 127–151.

Arrow, K. J. (1973). The theory of discrimination. In O. Ashenfelter and A. Rees (Eds.), *Discrimination in Labor Markets*. Princeton: Princeton University Press.

Ayres, I. (2001). *Pervasive Prejudice? Unconventional Evidence of Racial and Gender Discrimination*. Chicago: University of Chicago Press.

Ayres, I. (2005). Three tests for measuring unjustified disparate impacts in organ transplantation: The problem of 'included variable' bias. *Perspectives in Biology and Medicine 48*(1), S68–S87.

Ayres, I. and F. E. Vars (2000). Determinants of citations to articles in law reviews. *Journal of Legal Studies 29*(1), 427–450.

Babic, B. (2018). A Theory of Epistemic Risk. *Draft*.

Becker, G. S. (1957). *The Economics of Discrimination*. Chicago: University of Chicago Press.

Becker, G. S. (1993). Nobel Lecture: The Economic Way of Looking at Behavior. *Journal of Political Economy 101*(1), 385–409.

Borooah, V. K. (2001). Racial bias in police stops and searches: An economic analysis. *European Journal of Political Economy 17*(1), 17–37.

Borsboom, D., W.-J. Romeijn, and J. M. Wicherts (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods 13*(2), 75–98.

Carpenter, B., A. Gelman, L. M. Hoffman, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Stan (2016). A probabilistic programming language. *Journal of Statistical Software*.

DeGroot, M. H. and M. J. Schervish (2012). *Probability and Statistics* (4th ed.). New York: Wiley.

Duane, S., K. A. D. P. B. J. and D. Roweth (1987). Hybrid monte carlo. *Physics Letters B 195*, 216–222.

Dworkin, R. (1986). *Law's Empire*. Cambridge: Harvard University Press.

Dworkin, R. (2000). *Sovereign Virtue*. Cambridge: Harvard University Press.

Fuller, L. (1964). *The Morality of Law*. New Haven: Yale University Press.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis* (3rd ed.). New York: CRC Press (Taylor & Francis).

Gelman, A. and C. R. Shalizi (2013). Philosophy and the pratice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology 66*, 8–38.

Knowles, J., N. Persico, and P. Todd (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy 109*, 203–229.

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in Inductive Logic and Probability*, Volume 2, pp. 263–293. Berkeley: University of California Press.

Lindley, D. V. (1985). *Making Decisions*. New York: Wiley.

Mackay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Neal, R. M. (1995). An improved acceptance procedure for the hybrid monte carlo algorithm. *Journal of Computational Physics 111*, 194–203.

Pierson, E., S. Corbett-Davies, and S. Goel (2017). Fast threshold tests for detecting discrimination. *Working Paper*.

Pierson, E., C. Simoiu, J. Overgoor, S. Corbett-Davies, V. Ramachandran, C. Phillips, and S. Goel (2017). A large-scale analysis of racial disparities in police stops across the united states. *Working Paper*.

Quinn, W. (1989). Actions, intentions, and consequences: The doctrine of double effect. *Philosophy and Public Affairs 18*(4), 334–351.

Raz, J. (1995). *Ethics in the Public Domain: Essays in the Morality of Law and Politics*. Oxford: Oxford University Press.

Robert, C. P. (2007). *The Bayesian Choice: From Decision Theoretic Foundations to Computational Implementation*. Springer.

Simoiu, C., S. Corbett-Davies, and S. Goel (2017). The Problem of Infra-Marginality in Outcome Tests for Discrimination. *The Annals of Applied Statistics 11*(3), 1193–1216.

Thacher, D. (2002). From racial profiling to racial equality: Rethinking equity in police stops and searches. *Working Paper*.