OUTCOME TESTS

Ivan A. Canay
Northwestern University
ECON 481-3

1. Knowles, J., N. Persico, and P. Todd (2001). Racial Bias in Motor Vehicle Searches: Theory and Evidence, Journal of Political Economy, 109, 203–229.
2. Canay, I., M. Mogstad, and J. Mountjoy (2020). On the Use of Outcome Tests for Detecting Bias in Decision Making. Becker Friedman Institute Working Paper No. 2020–125

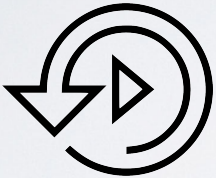
Past & Future

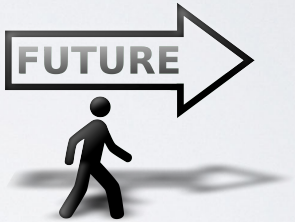
So Far

- Selection on Observables
- Roy Models and LATE
- MTEs
- Extrapolation

TODAY

- Outcome Tests
- Average vs Marginal behavior
- Roy Models in Outcome Tests
- Identification via MTEs





Outline

1. **A Model for Racial Bias in Motor Vehicle Searches**
2. Average-based Outcome Tests
3. A Roy Model for Racial Bias in Bail Decisions
4. Marginal-based Outcome Tests via MTEs

Motivation

Decisions of judges, lenders, journal editors, and other gatekeepers often lead to **disparities** in outcomes across affected groups.

Question

Whether, and to what extent, these group-level disparities are driven by relevant differences in underlying individual characteristics, or by biased decision makers.

Racial Bias in Vehicle Searches

Knowles, Persico, and Todd (2001, KPT)

- Police checking for illegal drugs more likely to search vehicles of black motorists than white motorists.
- KPT develops a model of police and motorist behavior to study bias.
- Idea is to separate statistical discrimination from prejudice/taste for bias/bias.
- KPT: prejudice (taste for discrimination) is a property of the officer's utility function, whereas statistical discrimination is a property of equilibrium.
- **Insight**: if an officer has the same cost of searching two subgroups, then the returns from searching will be equal across the subgroups.

Comments

- The idea of testing for discrimination by looking at differential outcomes is originally due to Becker (1957).
- Led to the so-called **Outcome Tests**.
- Different models and assumptions lead to different outcome tests.

Main Features

Main Features

- The model belongs to the literature on optimal auditing (Becker (68), Stigler (70))
- There are two sides of the market: **auditor** and **auditee** (a continuum of both)
- Both parties behave strategically.
- Police officers optimally decide whether to search a vehicle or not.
- Drivers decide whether to carry contraband or not.

Connection to Becker

- If police are prejudiced, the equilibrium returns to searching members of the group that is discriminated against will be below average.
- This idea, that tastes for discrimination lead to lower profits for the discriminators, originated with Becker (1957).
- **Becker (57)**: a firm that discriminates against minority employees uses labor inputs less efficiently, and so should have lower profits, than a non-discriminating firm.

The Model: Motorists

Motorists/Driver: they have type (r, v) where:

- Race: $R \in \{W, B\}$ is the race of the driver.
- Non-Race characteristics: $V \in \mathcal{V}$ denotes all **other** non-race characteristics.
- Decision: whether to carry contraband/drugs or not,
  - If they **do not** carry: payoff is zero whether or not the car is searched.
  - If they **carry**: payoff is $-\lambda_{\mathrm{s}}(r, v)$ if they are searched and $\lambda_{\mathrm{n}}(r, v)$ if not searched.
- Denote by $\gamma(r, v)$ the probability that a police officer searches a driver of type (r, v).
- Expected Payoff: to a motorist from carrying contraband is

$$\gamma(r, v)(-\lambda_{\mathrm{s}}(r, v)) + (1 - \gamma(r, v))\lambda_{\mathrm{n}}(r, v)$$

Note: V is not observed by the econometrician but may be observed by the police officer.

The Model: Officers

Police Officers: they don't have a type (cf. Anwar and Fang (06))

- Decision: whether to search a driver of type (r, v) or not.
- Goal: maximize the total number of convictions minus a cost of searching cars.
  - **Cost:** Marginal cost of searching a motorist of race r is denoted by $\tau(r) \in (0, 1)$.
  - **Benefit:** normalized to one (so that the cost is scaled as a fraction of the benefit).
- Let G denote the event that the motorist searched is guilty (i.e., with drugs in car)
- **Police Problem:** chooses the probability $\gamma(r, v)$ by solving

$$\max_{\{\gamma(r,v):r\in\{b,w\}\}} \sum_{r\in\{b,w\}} \int [P\{G|r,v\} - \tau(r)]\gamma(r,v)f(c,v)dv\,, \tag{1}$$

taking $P\{G|r, v\}$ as given.

Discrimination

Definition (Taste for Discrimination)

A police officer is racially prejudiced, or has taste for discrimination, if $\tau(w) \neq \tau(b)$.

Definition (Statistical Discrimination)

Assume $\tau(w) = \tau(b)$. Then an outcome exhibits statistical discrimination if $\gamma(b) \neq \gamma(w)$, where

$$\gamma(r) = \int \gamma(r, v) f(v|r) dv \, .$$

- Note that $\tau(\cdot)$ is preferences, while $\gamma(\cdot)$ is an equilibrium object.
- For Statistical discrimination we could use $\gamma(b, v) \neq \gamma(w, v)$ for some $v \in \mathcal{V}$ (later).

Outline

1. A Model for Racial Bias in Motor Vehicle Searches
2. **Average-based Outcome Tests**
3. A Roy Model for Racial Bias in Bail Decisions
4. Marginal-based Outcome Tests via MTEs

Equilibrium

- In equilibrium drivers randomize the decision to carry contraband and officers randomize the decision to search.

- For motorist to willing to randomize, the expected payoff to carry should be zero and so

$$\gamma^*(r,v) = \frac{\lambda_{\mathrm{n}}(r,v)}{\lambda_{\mathrm{n}}(r,v) + \lambda_{\mathrm{s}}(r,v)} \,.$$

  This ratio determines the police officer's search intensity.

- For a police officer to randomize, it must be the case that

$$P^*\{G|r,v\} = \tau(r) \quad \text{for all } r \in \{b,w\} \text{ and } v \in \mathcal{V} \,.$$

  Note: essentially imposes that $P^*\{G|r,v\}$ is flat in v. That is, in equilibrium motorists respond to officers and so they carry drugs with probability $\tau(r)$.

Implications

- Suppose that $\tau(w) = \tau(b) = \bar{\tau}$. Then, in equilibrium,

$$P^*\{G|b,v\} = \bar{\tau} = P^*\{G|w,v\}\,, \tag{2}$$

  but this **does not** imply that $\gamma^*(b,v) = \gamma^*(w,v)$. The equilibrium search intensity may be higher for blacks if,

$$\frac{\lambda_{\mathrm{n}}(w,v)}{\lambda_{\mathrm{n}}(w,v) + \lambda_{\mathrm{s}}(w,v)} < \frac{\lambda_{\mathrm{n}}(b,v)}{\lambda_{\mathrm{n}}(b,v) + \lambda_{\mathrm{s}}(b,v)}\,,$$

  e.g., the expected value of carrying drugs is higher for blacks for all (or some) v.

- This results forms the basis of the **outcome test**.

Outcome Test: intuition

- **Intuition:** if officers are profiling black drivers due to racial prejudice, they will search blacks even when the returns from searching them, i.e., the probabilities of successful searches against blacks, are smaller than those from searching whites.

- More precisely, if racial prejudice is the reason for racial profiling, then the success rate against the marginal black motorist (i.e., the last black motorist deemed suspicious enough to be searched) will be lower than the success rate against the marginal white motorist.

- **Difficulty:** researchers never observe search success rates against the **marginal** motorists (known as the inframarginality problem).

- **KPT model:** the success rate of the marginal motorist equals the success rate of the average motorist and this allows for a version of the outcome test based on averages: average-based outcome test.

Outcome Test: Result

- Equation (2) provides an outcome test for prejudice ($\tau(b) = \tau(w)$) that can be implemented without data on V and γ.

- It suffices to have data on the frequency of guilt by race conditional on being searched,

$$
\begin{aligned}
Q(w) &= \int P^*\{G|w,v\} \frac{\gamma^*(w,v) f(v|w)}{\int \gamma^*(w,v) f(v|w) dv} dv \\
&= \bar{\tau} \int \frac{\gamma^*(w,v) f(v|w)}{\int \gamma^*(w,v) f(v|w) dv} dv \quad \text{“key that } \tau \text{ does not depend on } v\text{”} \\
&= \bar{\tau} \\
&= \int P^*\{G|b,v\} \frac{\gamma^*(b,v) f(v|b)}{\int \gamma^*(b,v) f(v|b) dv} dv = Q(b)
\end{aligned}
$$

- **Average-based Outcome Test:** Conclude there is no bias if $Q(w) = Q(b)$.

Comments

- KPT find evidence of $Q(w) = Q(b)$. But also find evidence that blacks are searched more often than whites, $\gamma^*(b) > \gamma^*(w)$ (statistical discrimination).
- Motorists' response to the probability of being searched is key for this outcome test.
- If motorists did not react to the probability of being searched, testing for prejudice would require data on v.
- **Fundamental Assumption:** KPT implicitly assume that $\tau(r)$ does not depend on v - i.e., $\tau(r, v) \Rightarrow$ there is **no difference** between the marginal and the average search success rate.
- Anwar and Fang (2006) improve upon KPT, but still implicitly assume a condition like $\tau(r, v) = \tau(r)$.
- Brock, Cooley, Durlauf, Navarro (2012) show AOTs are not robust to letting $\tau(\cdot)$ depend on v provided $F(v|b) \neq F(v|w)$.

Outline

1. A Model for Racial Bias in Motor Vehicle Searches
2. Average-based Outcome Tests
3. **A Roy Model for Racial Bias in Bail Decisions**
4. Marginal-based Outcome Tests via MTEs

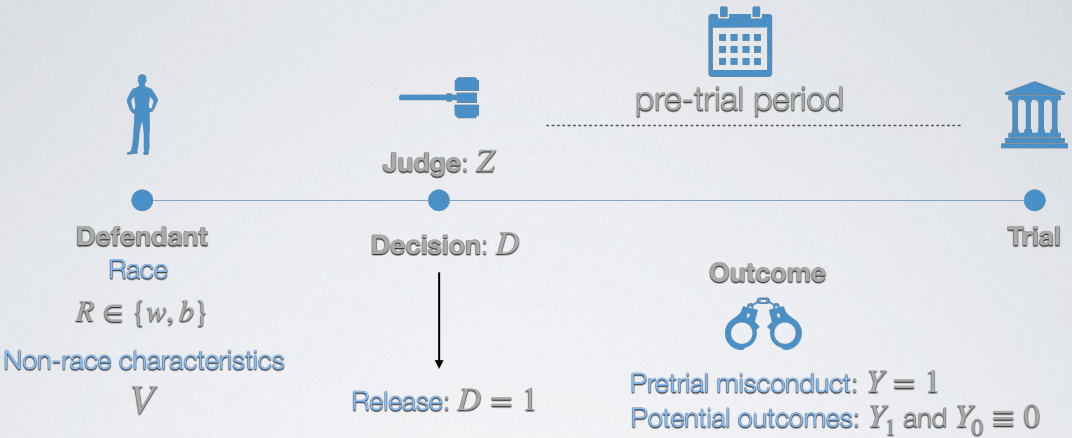
Overview

- Becker's (57,93) idea on how to test for racial bias:
  - **A.** Define when decisions are non-discriminatory or **unbiased**.
  - **B.** Derive **optimality conditions** for the agent's maximization problem.
  - **C.** Exploit implications on data and check if **consistent** with behavior.
- Inspired literature on discrimination to analyze group-level disparities
  - **A.** Early on, this led to comparisons of average-behavior across groups.
    but challenges arise when interpreting these results due to selection bias and inframarginality.
    see, Knowles et al. (2001); Anwar and Fang (2006); Persico (2009); among others.
  - **B.** More recently, random assignment of decisions makers opened new doors
    led to comparisons of marginal-behavior across groups (selection and inframarginality)
    but it also claimed greater levels of generality in the scope of the decision model
    see Arnold, et al. (**ADY**, 2018); Dobbie et. al. (2020); Arnold et al. (2021); among others.

Canay, Mogstad, and Mountjoy (20)

- Recent methods exploiting Marginal-based Outcome Tests (**MOTs**) share distinctive features.
  - One sided market ⇒ **E.g.,**: Bail Decisions in criminal justice system (ADY)
  - Decision model belongs to the class of **Generalized Roy models**: a cost/benefit analysis.
  - Institutional setting overcomes common challenges:
    (1) leverage random assignment of judges to avoid selection bias and
    (2) leverage near-continuous variation in judges to isolate marginal defendants.
  - **MOTs**: compare outcome probs. for marginal black vs. marginal white defendants.
- **CMM**: examines the **identifying power** of MOTs in Roy Models.
  - Formal results on **when and why** we can learn about racial bias via MOTs.
  - **Blueprint** for researchers who consider using MOTs to detect bias in these settings.
  - For concreteness: focuses on the pretrial release framework.
  - **Other settings**: results apply to other partial-equilibrium market settings.

Bail Setting



Marginal-based Outcome Test (MOTs)

Conclude there is racial bias when

$$P\{Y_1 = 1 | \text{marginal white}\} > P\{Y_1 = 1 | \text{marginal black}\} .$$

Summary of Results

- CMM study the suitability of the **Generalized Roy Model** and the **Extended Roy Model** to detect bias.
- CMM show MOTs are logical **invalid** in the Generalized Roy Model
  - Analyst may find bias when each judge is racially unbiased.
  - Analyst may conclude no bias when each judge is racially biased.
  - **Problem**: unbiased judges do not necessarily equalize outcome probs. of the marginal white and black defendants.
- CMM show MOTs are logical **valid** in the Extended Roy Model
  - **Restriction**: the only source of bias is racial.
- CMM study **identification** of MOTs
  - The Generalized Roy Model does not admit an MTE representation without further restrictions.
  - The Extended Roy Model automatically delivers a valid an MTE representation

Setting Up The Model

- **Variables**: Judges Z, Defendants (R, V), Decision D, Outcome Y.
- **Incarceration cost**: the cost of incarcerating a given defendant is $c(z)$.
- **Judge's information set**: Let $W(z)$ be the information set of judge z on defendants:
  1. R: Race
  2. V: Non-race characteristics that are not observed by the analyst
  3. X: Non-race characteristics observed by the analyst (ignored from here on).
  4. F: beliefs on $F(V|R)$ and $F(Y_1|R, V)$ - also captures prediction errors.
- **Taste for Discrimination**: Denote by $\beta(z, r, v) > 0$ the "taste" judge z has on a defendant (r, v).

Decision Problem

Judge z' minimizes expected cost, where "expected" here means the expectation conditional the **information set** $W(z')$:

$$\min_{d \in \{0,1\}} E[c(z')(1-d) + \beta(z', R, V)Y_d | W(z')] .$$

Generalized Roy Model

- The judge's optimization problem leads to the optimal decision D given by

$$D = I\Big\{\beta(z,R,V)E[Y_1|W(z)] \leq c(z)\Big\}\,.$$

- **Note:** $E[Y_1|W(z)]$ may not equal $E[Y_1|R,V]$ if judge z makes miscalculates risk.
- We can formalize this by letting $\lambda(z,r,v) > 0$ and assuming that

$$E[Y_1|W(z)] = \lambda(z,R,V)E[Y_1|R,V]\,.$$

When $\lambda(z,R,V) = 1$ judges do not makes prediction errors. We assume this for now.

Definition (Generalized Roy Model)

The decision rule in the **Generalized Roy Model** is given by

$$D = I\Big\{E[Y_1|R,V] \leq \tau(z,R,V)\Big\} \text{ where } \tau(z,R,V) \equiv \frac{c(z)}{\beta(z,R,V)}\,.$$

Generalized Roy Model: features

- **Generalized Roy Model**: judge z cost-benefit comparison is

$$D = I\Big\{E[Y_1|R,V] \leq \tau(Z,R,V)\Big\}$$

  - $E[Y_1|R,V]$ is the **expected cost** of release
  - $\tau(z,r,v)$ is the **expected benefit** of release (or expected cost of detain)

- **Main features of this model**:

1. Cost of release is tied to an observed outcome and does not vary across judges,

$$E[Y_1|R,V] \equiv P\{Y_1 = 1|R = r, V = v\}\ .$$

2. Benefit of release is unobserved to the analyst (object of interest).

3. Both R and V enter the cost function and the benefit functions.

4. Judge observes both R and V; analyst observes R but not V.

5. V is a scalar index of all possible non-race characteristics (signal of risk) - as in Anwar and Fang (2006).

Extended Roy Model

- A **common restriction** in the literature on average-based outcomes tests is

$$\beta(z,r,v) = \beta(z,r) \text{ for all } r \in \{w,b\} .$$

(Knowles, Persico, and Todd, 2001; Anwar and Fang, 2006; Persico, 2009; among others)

- **Implication**: No forms of bias other than racial.

- In our setting with MOTs, this restriction translates to

$$\tau(z,r,v) = \tau(z,r) \text{ for all } r \in \{w,b\}$$

and leads to the so-called **Extended Roy Model**

Definition (Extended Roy Model)

The decision rule in the **Extended Roy Model** is given by

$$D = I\Big\{E[Y_1|R,V] \leq \tau(z,R)\Big\} \text{ where } \tau(z,R) \equiv \frac{c(z)}{\beta(z,R)} . \tag{3}$$

Discussion

Taste for Discrimination

Models share same "legitimate" optimization problem and differ in their taste for discrimination parameter.

- **E** All bias in the decision making must manifest itself in only **one form**: as racial bias.
- **G** Bias can vary in magnitude with other characteristics (speech patterns, body weight, facial features).

Measurement/Prediction Error

if $\lambda \neq 1$ then $\tau = c/(\beta \times \lambda)$: a catch-all for residual factors beyond the probability of misconduct.

- **G** Any error $E[Y_1|W(z)] = \lambda(z,R,V)E[Y_1|R,V]$ can be **absorbed** by the function $\tau(z,R,V)$.
- **E** Any error $E[Y_1|W(z)] = \lambda(z,R,V)E[Y_1|R,V]$ must be **uncorrelated with** V (even if uncorrelated with race).

Distribution of V given R

Both model can handle $F_{V|w} = F_{V|b}$ and $F_{V|w} \neq F_{V|b}$ similarly. This feature separates AOTs from MOTs.

AOTs Have been shown to exhibit problems provided $F_{V|w} \neq F_{V|b}$ (Brock, et. al.).

MOTs CMM show their invalidity does not depend on whether $F_{V|w} \neq F_{V|b}$ or not.

Definition of Racial Bias

Definition (Bias in the Generalized Roy Model)

We say judge z is **racially unbiased** if

$$\beta(z, r, v) = \beta(z, v) \text{ for all } v \in \mathcal{V} .$$

We say judge z is **partially racially biased** against black defendants if

$$\beta(z, w, v) \leq \beta(z, b, v) \text{ for all } v \in \mathcal{V} \text{ with strict inequality for some interval } (\underline{v}, \bar{v}) \subseteq \mathcal{V} .$$

We say judge z is **completely racially biased** against black defendants if

$$\beta(z, w, v) < \beta(z, b, v) \text{ for all } v \in \mathcal{V} .$$

- **Note**: applies to defendants with the same non-race characteristics v.
- **Extended Roy Model**: no distinction between partially and completely biased since:

$$\beta(z, w) < \beta(z, b) \text{ in that case } .$$

Marginal Defendant

Marginal defendant is determined via the following **single crossing condition**.

Assumption (SC)

For all (z, r) *there exists a* **marginal defendant** $V^*_{z,r} \in \text{int}(\mathcal{V})$ *such that*

$$E[Y_1|r, V^*_{z,r}] = \tau(z, r, V^*_{z,r}) \,,$$

and $E[Y_1|r, V^*_{z,r}] < \tau(z, r, v)$ *for all* $v < V^*_{z,r}$ *and* $E[Y_1|r, V^*_{z,r}] > \tau(z, r, v)$ *for all* $v > V^*_{z,r}$ *.*

- Same SC for the **Extended Roy Model**: $E[Y_1|r, V^*_{z,r}] = \tau(z, r) \,,$

Marginal Defendant

Marginal defendant is determined via the following **single crossing condition**.

Assumption (SC)

For all (z, r) *there exists a* **marginal defendant** $V^*_{z,r} \in \text{int}(\mathcal{V})$ *such that*

$$E[Y_1|r, V^*_{z,r}] = \tau(z, r, V^*_{z,r}) \,,$$

and $E[Y_1|r, V^*_{z,r}] < \tau(z, r, v)$ *for all* $v < V^*_{z,r}$ *and* $E[Y_1|r, V^*_{z,r}] > \tau(z, r, v)$ *for all* $v > V^*_{z,r}$ *.*

- Same SC for the **Extended Roy Model**: $E[Y_1|r, V^*_{z,r}] = \tau(z, r)$,

- The **marginal-based outcome test** compares pretrial misconduct probabilities at the margin:

$$\begin{aligned} \text{evidence of racial bias}: \quad & E[Y_1|w, V^*_{z,w}] > E[Y_1|b, V^*_{z,b}] \\ \text{evidence of no bias}: \quad & E[Y_1|w, V^*_{z,w}] = E[Y_1|b, V^*_{z,b}] \,. \end{aligned}$$

Outline

1. A Model for Racial Bias in Motor Vehicle Searches
2. Average-based Outcome Tests
3. A Roy Model for Racial Bias in Bail Decisions
4. **Marginal-based Outcome Tests via MTEs**

Logically Valid Test

Definition

We say that the outcome test is logically valid if and only if

$$E[Y_1|w, V^*_{z,w}] = E[Y_1|b, V^*_{z,b}] \tag{4}$$

whenever judge z is racially unbiased, and

$$E[Y_1|w, V^*_{z,w}] > E[Y_1|b, V^*_{z,b}] \tag{5}$$

whenever judge z is (partially or completely) racially biased against black defendants.

- **Minimal requirement**: differences in outcome probabilities at the margin identify differences in the expected benefits (at least the sign).
- **Implication**: marginal white and black defendants should have the same probability of pretrial misconduct if and only if their common judge is racially unbiased.

Assumptions

Assumption (C)

The expected cost function $E[Y_1|r,v] : \{w,b\} \times \mathcal{V} \to \mathbf{R}$ is continuous and non-decreasing in v for $r \in \{w,b\}$.

Assumption (B)

The expected benefit function $\tau(z,r,\cdot) : \mathcal{V} \to \mathbf{R}$ is continuous and strictly decreasing in v for $r \in \{w,b\}$.

- **Note**: neither monotonicity nor continuity are required to prove our main results
- Main results hold under "local versions" of these conditions (in paper).

Main Result I

Theorem (MOTs in the GRM)

Let **SC**, **C**, *and* **B** *hold and consider the Generalized Roy Model and a judge* z'.

(I) *Suppose* z' *is racially unbiased, i.e.*

$$\tau(z', r, v) = \tau(z', v) \text{ for all } v \in \mathcal{V} .$$

The marginal white may exhibit a higher misconduct probability than the marginal black,

$$E[Y_1|w, V^*_{z',w}] > E[Y_1|b, V^*_{z',b}] .$$

(II) *Suppose* z' *is partially biased against black defendants, i.e.*

$$\tau(z, w, v) \geq \tau(z, b, v) \text{ for all } v \in \mathcal{V} \text{ with strict inequality for } v \in (\underline{v}, \bar{v}) .$$

The marginal black may exhibit identical misconduct probability than the marginal white,

$$E[Y_1|w, V^*_{z',w}] = E[Y_1|b, V^*_{z',b}] .$$

It may also be the case that $E[Y_1|w, V^*_{z',w}] < E[Y_1|b, V^*_{z',b}]$.

(III) *The results in (II) hold if* z' *is completely biased against black defendants.*

Comments

- **Problem**: judges equalize outcome probs. for the **same value of** v $\Rightarrow$ yet most often $V^*_{z,w} \neq V^*_{z,b}$.
- **Implications**: a test of H_0 : "no bias" based on estimating $E[Y_1|r, V^*_{z',r}]$ via MTEs:
  - May have **no size control** (over-rejects under the null)
  - May have **no power** or conclude bias in the opposite direction
  - Confidence intervals for the benefit differential are generally **invalid**.
- **Summary**: differences in $E[Y_1|r, V^*_{z',r}]$ are **uninformative** about differences in $\tau(z, r, v)$.
- The **GRM** has been **recently used** to implement MOTs in a variety of papers
  - See Arnold, et al. (2018); Dobbie et. al. (2020); Arnold et al. (2021); among others
  - CMM's theorem raises **concerns** to the level of generality claimed in those papers.

Related Results

- Results on the invalidity of MOTs in the Generalized Roy Model are **novel**.
- **Related**: Brock et al. (2012) on the scope of generality of **AOTs in two-sided markets**.
  e.g., those in Knowles et al. (2001); Anwar and Fang (2006); Persico (2009); among others.
- **Shared features**: allowing the taste for discrimination parameter $\beta(z, w)$ to depend on unobserved characteristics is problematic for the validity of the outcome tests.
- **Main Differences:**
  - Brock et al. (2012) discussed lack of robustness of original results based on possibly restrictive models.
  - Formal arguments about AOTs and MOTs are quite distinct.
  - Main driving force for invalidity of AOTs is the possibility that $F_{V|w} \neq F_{V|r}$.

    ...in particular, assuming $F_{V|w} = F_{V|r}$ renders AOTs generally valid.
  - CMM setting: MOTs are invalid in the Generalized Roy Model even if $F_{V|w} = F_{V|r}$.

Main Result II

Theorem (MOTs in the ERM)

Let **SC** *hold and consider the Extended Roy Model and a judge z'. Then, the MOTs are logically valid in the sense that*

$$\text{sign}\left(E[Y_1|w, V^*_{z',w}] - E[Y_1|b, V^*_{z',b}]\right) = \text{sign}\Big(\tau(z', w) - \tau(z', b)\Big) \,.$$

- The model allows the marginal defendant to exhibit different non-race characteristics, i.e.,

$$V^*_{z,w} \neq V^*_{z,b} \,,$$

even when judges are racially unbiased. This can be interpreted as statistical discrimination.

QUESTIONS?

Econometrics Behind MOTs

- **MOTs**: require misconduct probabilities of marginal white and black defendants.
- **Random Assignment**: $(Y_1, Y_0, R, V) \perp\!\!\!\perp Z$ - holds in the bail setting.
  $\Rightarrow$ Not enough to identify these quantities from the distribution of the observed data (Y, D, R, Z).
- **Generalized Roy Model:**
  - Without further restrictions **it is not possible** to recover misconduct probabilities of marginal defendants.
    $\Rightarrow$ Model restricts behavior of a given judge, but it does not impose restrictions on how judges differ in their behavior.
  - **Needed**: cross-judge restrictions.
  - **Important**: cross-judge restrictions do not alleviate the invalidity of MOTs in the Generalized Roy Model.
- **Contrasts**: the Extended Roy Model admits valid MTEs without further assumptions.

Monotonicity

Exogenous reassignment from one judge to another weakly increases the benefit function (and thus the likelihood of release) for every defendant of a given race:

$$\tau(z, r, v) \geq \tau(z', r, v) \text{ for all } v \in \mathcal{V} \qquad (\mathsf{M})$$

and any two judges z and z' and race $r \in \{w, b\}$. **Automatically holds** in the Extended Roy Model.

- Monotonicity says we could sort judges from most lenient to most severe.
- Monotonicity is key to obtain a valid **MTE representation** (next).
- **Claim**: Monotonicity is unrelated to the invalidity of MOTs in Theorem 1.

Marginal Probabilities as MTEs

- **MTE representation**: In order for the decision model to admit a valid MTE representation, one needs to argue that the model can take the form

$$D = I\Big\{U_r \leq p(z,r)\Big\} \quad \text{with} \quad U_r \sim U[0,1] \text{ and } p(z,r) = P\Big\{D = 1 | Z = z, R = r\Big\}.$$

- **Implication**: it allows the analyst to define the race-specific MTE functions as

$$MTE_r(u) \equiv E[Y_1 - Y_0 | U_r = u, R = r] = P\Big\{Y_1 = 1 \;\Big|\; U_r = u, R = r\Big\}.$$

- **Marginal defendants** are those with $U_r = p(z,r)$, so the misconduct rate of marginal defendants of race r facing judge z are identified by

$$MTE_r\Big(p(z,r)\Big).$$

Wrapping up: Extended Roy Model

Take Away: MOTs work in the context of an Extended Roy Model.

Extended Roy Model

$$D = I\Big\{E[Y_1|R,V] \leq \tau(z,R)\Big\} \text{ where } \tau(z,R) \equiv c(z)/\beta(z,R)\ .$$

1. (a) No forms of bias **other than racial**

   (b) A biased judge must be **equally biased** against all defendants of the same race.

   At odds with body of empirical evidence suggesting that prejudices can vary in magnitude with a wide range of characteristics, including skin tone, facial features, speech patterns, height, etc.

2. Judges' accuracy in predicting $E[Y_1|R,V]$ **cannot vary systematically** with V.

$$E[Y_1|W(z)] = \lambda(z,R,V)E[Y_1|R,V] \ \Rightarrow \tau(z,R,V) = \frac{c(z)}{\lambda(z,R,V)\beta(z,R)}\ .$$

That is: $\tau(\cdot)$ is a catch-all for residual factors beyond the probability of misconduct

At odds with work comparing decisions of judges with machine learning: Judges underestimate defendants with minor current charge/strong prior convictions.

THE END!