Difference-in-Difference Estimators of Prejudice: An Examination of the Existing Test and An Alternative

Hanming Fang
Duke University

Nicola Persico
New York University

April 2009
(Preliminary)

Introduction

- Racial disparities, and also related, gender disparities, are ubiquitous.
- Racial disparities have been documented in employment, in medical care, in lending, in motor vehicle stops and searches, and in all phases of law enforcement such as jury selection, prosecution and sentencing.
- A key question is whether the documented racial and gender disparities in outcomes are due to racial bias, a term that we use to refer to the presence of a psychic utility differential accrued by "the treators" (i.e. the employers, the doctors, the lenders, the police officers, the judges etc.) in making their decisions on "the treated" (the job applicant, the patient, etc.) when the treated have different races.

Introduction

- But how to test whether racial or gender bias are partly responsible for the racial and gender disparities?
- The attempt to infer bias from statistical data is hindered by many data issues, most notably the problem of "missing variables:" among the legitimate sources of disparate impact are productive characteristics of the treated, which are potentially correlated with race, but may not be observed by the researcher.
- There is a growing list of empirical papers which adopt different identification strategies in an (implicit or explicit) effort to circumvent the missing variable, and other, problems.
- Usually, the issue of identification is dealt with at a rather intuitive level in these papers.

Introduction

- In particular, there have been quite some papers that use DD based on the whether or not race or gender is observable to the treators to test for the presence of racial or gender bias.
- To the best of our knowledge, the question of what is exactly estimated from the DD analysis has never been formally addressed.
- It is simply assumed, by some intuitive reasoning at most, that the DD estimator is informative of the prejudice.
- A plausible intuition is the notion that *behavior in the color-blind environment is by definition unbiased*, and any disparities that arise when race becomes observable are due to bias.

Introduction

- However, the treator's behavior in the color-blind environment is not necessarily unbiased if there are other observables that are correlated with race.
- That is, racial or gender prejudice still indirectly impacts the behavior of the treators even when race or gender of the treated is not directly observable. Thus the treators' behavior under the color-blindness setup does *not* represent the statistical benchmark of the behavior in the absence of racial prejudice.
- Given this observation, it is then not difficult to understand why DD estimator does not necessarily inform us about the racial prejudice at all.
- *The ideal statistical benchmark should be the treators' behavior when they do not have racial prejudice, which is very different from their behavior when only the race of the treated is concealed.*

Agenda of the Paper

1. Analyze what is the exact necessary and sufficient conditions under which DD based estimator on the variation in the observability of race informs us about prejudice.
   - The condition turns out to be that race (or gender) does not provide any useful information to the treater to predict the legitimate part of her payoff given other signals about the treated that are observable to the treater.
2. To the extent that such assumptions may be strong, we propose an alternative outcome-based test that work more generally.
   - We show that if there is no racial prejudice, not observing race always lowers the treator's legitimate payoff (i.e. search success rates, or loan repayment rates etc.) relative to that when race is observable; however, the legitimate payoff of the treator may be higher when race is unobservable when the treaters are racially prejudiced.
   - This test does not always have power, but it is nonetheless informative in some situations.
3. Explore the general DD estimators based on other variations.

A General Treatment Game: Players

- There are two types of actors:
  - the potential discriminators – whom we refer to as "*the treators*";
  - those potentially discriminated – whom we refer to as "*the potentially treated.*"
- There is a mass 1 of treators, indexed by s distributed with density $f(s)$.
- The potentially treated consists of a mass M of agents, divided into groups indexed by g, where each g represents a specific combination of all their innate characteristics, observable or not.
- We write $g \in G$, where G could be a finite, numerable or continuous set.
- To fix ideas we will assume that G is continuous. Each group g has mass $\mu(g)$, so that $\int_G d\mu(g) = M$.

A General Treatment Game: Information Structure

- In our discussion, it is important to be specific about what the treator observes when she makes the treatment decision, and which elements she does not observe.
- We will model information as a partition over the set G.
- Let $\mathcal{T}_s$ (stands for "type") denotes a partition over the set G with generic element T_s.
  - T_s represents a group of the potentially treated all of whom are identical in the eyes of treator s, though they may be heterogeneous in ways not observed by the treator.
- Treator s's actions (treatment) need to be measurable with respect to $\mathcal{T}_s$.
- The treator knows μ.

Race (or Gender)

- An observable characteristics for which discrimination is hypothesized is somehow salient, say race or gender. For concreteness, we will say it is race.
- There are only two races. $r = a, w$.
- We assume that the researcher knows the race of the treated;
  - Formally, this assumption is equivalent to the assumption that $\mathcal{D}$ is at least as fine as the two-element partition $\mathcal{R} = \{A, W\}$ of G, where A denotes the set containing all g's who have race a, and W denotes the complementary set.
- Treators do not necessarily all know the race of each population member.
- We will study "difference in difference" tests for racial bias, based on some variation in the treator's problem.

Strategies and Payoffs

- A strategy for treator s is a vector of probabilities $p_s(g)$, one for each group g.
- Since treator s cannot distinguish two groups g and g' if they both belong to the same T_s, a strategy needs to be measurable with respect to $\mathcal{T}_s$.

Definition

A strategy for treator s is a function $p_s(g) \in [0,1]$ such that

$$p_s(g) \in [0,1] \text{ is measurable with respect to } \mathcal{T}_s \tag{1}$$

$$\int_G p_s(g)\, d\mu(g) \leq C_s. \tag{2}$$

Strategies and Payoffs for the Treator

- A strategy for the treated is a function $a(g) \in [0,1]$ which specifies the action for each member of group g.
- The expected payoff of the *treator* if she follows strategy $p(\cdot)$ is given by

$$U_s(p(g), a(g)) = \int_G \left[\overbrace{u(p(g), a(g))}^{\text{Legitimate}} - \overbrace{p(g)\, t_s(g)}^{\text{Illegitimate}} \right] d\mu(g). \quad (3)$$

- Increasing $t_s(g)$ makes treator s less inclined to treat group g.

Strategies and Payoffs for the Treated

- For each g we will denote by

$$\mathbf{p}(g) = \int_0^1 p_s(g) f(s)\, ds$$

  the aggregate action by the treators against g.
- For each g, the *treated's payoff* is

$$v(a, \mathbf{p}(g), g).$$

- Note that v only depends on the aggregate treatment $\mathbf{p}(g)$, and not on the treatment of any individual treator.

Equilibrium

- The equilibrium of the game generates the data we get to observe.

Definition

An equilibrium of the treatment game given $[\mathcal{T}_s]_s$ is a set of strategies $a^*(g), [p_s^*(g)]_s$ such that
(a) for each g, $a^*(g)$ maximizes $v(a, \mathbf{p}^*(g), g)$;
(b) for each s, $p_s^*(g)$ maximizes $U_s(p(g), a^*(g))$; and, finally,
(c) for each g, $\mathbf{p}^*(g) = \int_0^1 p_s^*(g) f(s)\, ds$.

Assumptions on $t_s(g)$ and Definitions

- We now impose some restrictions on $t_s(g)$ that is typically assumed, implicitly or explicitly, in the discrimination literature.

Definition

Treator s is **Unbiased**

$$t_s(g) = t_s \text{ for all } g$$

Treator s is **Biased against** A

$$t_s(g) = t_s(A) > t_s(W) = t_s(g') \text{ for all } g \in A, g' \in W$$

- For now consider the case that $t_s(A), t_s(W)$ do not vary over the treators, i.e. $t_s(A) = t(A)$ and $t_s(W) = t(W)$ for all s.
- We will consider the case where $t_s(A)$ and $t_s(W)$ vary by s later.

Do the Treated Respond to the Behavior of the Treated?

- Our framework accommodates the case in which the treated responds to the treators' strategy, but for now I will focus on the non-response case, which could be accommodated by assuming:

Assumption

$v(a, \mathbf{p}(g), g) = -\infty$ *except for* $a = \alpha(g)$.

- The way this assumption is formulated implies that the treated's optimal action $a^*(g)$ necessarily equals $\alpha(g)$ independent of $\mathbf{p}(g)$.
- For today, I will focus on this case ignoring the treated's problem.
- This may be reasonable assumption in some settings: e.g. *patients vs. doctors.*
- This assumptions imply that the treated population under scrutiny does not vary as we change other aspects of the environment.
- This may not be a good assumption in some applications: *highway stops during day vs. during night.*

The Decision Problem of the Treator

- For a given information structure $\mathcal{T}_s$ of a treator s, he solves:

$$\max_{p(\cdot)} \int_G \left[u\left(p\left(g\right), \alpha\left(g\right)\right) - p\left(g\right) t_s\left(g\right)\right] d\mu\left(g\right)$$

$$p\left(g\right) \text{ measurable with respect to } \mathcal{T}_s$$

$$\int_G p_s\left(g\right) d\mu\left(g\right) \leq C_s.$$

- There are several dimensions in which variations could potentially exploited to learn about $t_s\left(g\right)$:
  - Variations in $\mathcal{T}_s$;
  - Variations in t_s across s;
  - Variations in C_s.

The Action-Based DD Test Based on Variations in $\mathcal{T}_s$

- Denote by $\mathcal{T}_s^O$ an information set in which race is observable to the treator, and denote by $\mathcal{T}_s^U$ the corresponding information set where race is not observable but is otherwise identical to $\mathcal{T}_s^O$.
- $P_U^*(R)$ denotes the amount of treatment devoted to race R in the setup where the race is Unobservable to the treator, and $P_O^*(R)$ is its counterpart when race is Observable.
- The DD test:

$$DD = \overbrace{\underbrace{[P_U^*(W) - P_U^*(A)]}_{\text{1st difference}} - [P_O^*(W) - P_O^*(A)]}^{\text{2nd difference}} < 0$$
$$\Leftrightarrow \{\text{ treators are biased against } A\}.$$

When Does DD Work?

- Let

$$w(p,T) = \frac{1}{\mu(T)} \int_T u(p, \alpha(g))\, d\mu(g),$$

represents the expected payoff from treating group T.

Assumption

Consider any two groups $T_s, T_s' \in \mathcal{T}_s^O$ which differ only because of race. We assume

$$w(\cdot, T_s) = w(\cdot, T_s').$$

Proposition

Assume the above Assumption holds. Suppose the treator is biased against A. Then the DD test is valid.

- Intuition.
- Remark: Only the sign, not the magnitude of the DD estimate is informative.

Counter Example 1

- An unbiased officer who maximizes the probability of finding contraband can stop and search 100 people.

Color of Car \ Driver Race	Black	White
Dark Colored Car	0.5 (50)	0.4 (50)
Light Colored Car	0 (70)	0.6 (70)

- We can verify that:

$$\begin{aligned} P_O^*(B) &= 30/120 = 25\% \\ P_O^*(W) &= 70/120 \approx 58.33\%. \\ P_U^*(B) &= P_U^*(W) = 50/120 \approx 41.67\%. \end{aligned}$$

- Thus the DD estimator is

$$DD = (.4167 - .4167) - (.25 - .5833) = 33.33\%.$$

- The DD estimator would conclude that there is racial prejudice *against whites.*

Counter Example 2

- Suppose that the officer is prejudiced against black drivers: $t(B) = 0.21$ and $t(W) = 0.4$ and no capacity constraint.

Color of Car \ Driver Race	Black	White
Dark Colored Car	$0.20\ (80)$	$0.45\ (20)$
Light Colored Car	$0.40\ (20)$	$0.50\ (80)$

- When the officer can observe race, he will search all white drivers; he will only search black drivers in light colored cars.
- When race is not observable, however, he will use a threshold

$$\begin{aligned} E\left[t(r)\,|\text{Dark Colored Car}\right] &= \sum_{r\in\{W,B\}} t(r) \times \Pr\left(r|\text{Dark Colored Car}\right) \\ &= 0.248; \end{aligned}$$

while the guilty rate among drivers of dark colored cars is given by

$$\frac{.2 \times 80 + .45 \times 20}{100} = 0.25.$$

- Hence,

$$\begin{aligned} P_O^*(B) &= 20/100 = 20\% \\ P_O^*(W) &= 100\%; \\ P_U^*(B) &= P_U^*(W) = 100\%. \end{aligned}$$

- Thus the DD estimator is

$$DD = (1.00 - 1.00) - (.20 - 1.00) = 80\%.$$

- Thus DD estimator would have led us to conclude that there is racial prejudice *against whites*, even though the officer is prejudiced against blacks.

Outcomes-Based Test Based on Variations in $\mathcal{T}_s$

Proposition

As race becomes observable, if a treator is unbiased then the legitimate portion of his payoff increases.

- The proposition suggests a very general test that can be carried out provided that we can construct $u\left(p^*\left(g\right), \alpha\left(g\right)\right)$ with the available data on outcomes.
- If we see that the legitimate portion of the treator's payoff is higher when race is not observable, we can reject no prejudice.
- The test has low power, but could be informative.

Potential Applications of the Above Test

- `Prosper.com` peer-to-peer lending
- Finding not allowing the race of the borrowers to be revealed lead to increases in the profitability of the loans would provide evidence of prejudice by the lenders.

DD Test for Relative Prejudice Using Variations in t_s

- Now suppose that there are two types of treators, 1 and 2.
- Suppose that we are interested in finding whether treator 1 is **less prejudiced against** A than treator 2, i.e.,

$$t_1(A) - t_1(W) < t_2(A) - t_2(W).$$

- The DD test for the above relative prejudice is

$$DD = \overbrace{\underbrace{[P_1^*(W) - P_1^*(A)]}_{\text{1st difference}} - [P_2^*(W) - P_2^*(A)]}^{\text{2nd difference}} < 0$$

$$\Leftrightarrow \{\text{ treator 1 is less biased against } A \text{ than treator 2}\}$$

- This interpretation is not warranted when treators also differ in capacity $C_1 \neq C_2$.

Counter Example 3

- There are a total of 150 patients who might benefit from further testing.
  - Of those, 50 whites and 50 blacks definitely need further testing;
  - The remaining 50 patients are whites who might potentially benefit from additional testing, but seem less urgent.
- Suppose that two doctors are both unbiased.
- An unbiased doctor (treator 1) can refer $C_1 = 100$ patients for further testing. Under this setup, the doctor will refer 50 whites and 50 black patients.
- Doctor 2 has a bigger budget $C_2 = 150$, the doctor will refer 100 whites and 50 blacks. We have

$$\begin{aligned} P_1^*(W) - P_1^*(A) &= 50 - 50 = 0 \\ P_2^*(W) - P_2^*(A) &= 100 - 50 = 50 \end{aligned}$$

hence $DD = -50 < 0$ and the DD test would conclude that doctor 2 is more biased against blacks.

Rank Order Test

- A rank order test (instead of DD) similar in spirit to Anwar and Fang (2006) could work.

Proposition

Consider two treators with treator-specific C_i *and* $t_i(\cdot)$*. Suppose* $P_1^*(W) > P_2^*(W)$ *and* $P_1^*(A) < P_2^*(A)$*. Then it cannot be that treator 1 is less biased against* A *than treator 2.*

Rank Order Test vs. DD Test

- The rank order test: treator 1 is more biased against A than treator 2 if

$$P_1^*(W) - P_2^*(W) > 0 \quad \text{and} \quad P_1^*(A) - P_2^*(A) < 0.$$

- These two inequality imply (but are not equivalent!) to

$$\begin{aligned} P_1^*(W) - P_2^*(W) &> P_1^*(A) - P_2^*(A), \\ &\Leftrightarrow [P_1^*(W) - P_1^*(A)] - [P_2^*(W) - P_2^*(A)] > 0. \end{aligned}$$

  which is the DD test statistic.

Corollary

The DD test is different from the rank order test and it is biased towards over rejecting the null of no discrimination.

An Example

- In Price and Wolfers (2008), the relative bias of white v. black referees is inferred from the number of fouls called against white/black players.
- Using our notation, treator 1 corresponds to a majority-white refereeing crew, treator 2 to a majority-black refereeing crew.
- $P_i^* (R)$ represents the average number of fouls per game assessed by crew i on a player of race R.

An Example

- Price and Wolfers report $P_1^*(A) = 4.330$, $P_2^*(A) = 4.329$, $P_1^*(W) = 4.954$, and $P_2^*(W) = 5.023$.
- The test offered in our proposition reads:

$$P_1^*(A) > P_2^*(A) \text{ and } P_1^*(W) < P_2^*(W).$$

  This test could not reject with any degree of confidence the hypothesis of no relative bias ($P_1^*(A) < P_2^*(A)$).
- The (possibly improper) DD test statistic reads

$$\begin{aligned} P_1^*(A) - P_1^*(W) &> P_2^*(A) - P_2^*(W) \\ 4.330 - 4.954 &> 4.329 - 5.023 \end{aligned}$$

  which provides stronger support for the bias hypothesis.

Things to Do

- Other variations
- The case where the treated respond to the anticipated treator behavior;
- The case where the treated and the treators sort.