Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination

Patrick Kline and Chris Walters
UC Berkeley and NBER

September 2020

Labor market discrimination

Title VII of the Civil Rights of 1964 prohibits employment discrimination on the basis of race, sex, and other protected characteristics

- Empirical literature focuses on measuring market-level averages of discrimination (Altonji and Blank, 1999; Guryan and Charles, 2013)
  - Observational studies of "unexplained" gaps (Oaxaca, 1978)
  - Correspondence experiments (Bertrand and Mullainathan, 2004)
- Variation in discrimination across employers influences
  - Effects on minority workers (Becker, 1957; Charles and Guryan, 2008)
  - Difficulty of enforcing the law – e.g., targeting of EEOC investigations / charge priority system
- Today: tools for using correspondence experiments to quantify heterogeneity and detect discrimination by individual jobs

Correspondence studies as ensembles

Correspondence studies send multiple applications to each job opening

- We view such studies as *ensembles* of small micro-experiments
- Use the ensemble in service of two goals
  - Learn about the distribution of discrimination across employers
  - Interpret the evidence against particular employers – "indirect evidence" (Efron, 2010)
- Methodological contribution: extend non-parametric Empirical Bayes (EB) methods to settings where each experiment too small for normality to ensue
  - Shape constrained GMM for estimating heterogeneity moments
  - Robust posteriors for detection / decision-making

Preview of findings

Apply methods to three high-quality correspondence experiments

- Key findings
  - Tremendous heterogeneity: a few jobs discriminate intensely, most discriminate little
  - Discrimination against both genders
  - Imbalances in callback rates can provide robust evidence of discrimination by particular jobs
- Policy implications
  - 10 applications sufficient to reliably detect non-trivial share of discriminating jobs
  - Parametric EB decision rule yields performance close to minimax

Preliminaries

Setup and Notation

- Sample of J jobs, each receiving L_w white and L_b black applications (total $L = L_w + L_b$)
- $R_{j\ell} \in \{w, b\}$ indicates assigned race of application ℓ to job j
- *Potential* callbacks from job j to application ℓ as fn of race:

$$(Y_{j\ell}(w), Y_{j\ell}(b)) \in \{0,1\}^2$$

- Observed callback outcome is $Y_{j\ell} = Y_{j\ell}(R_{j\ell})$
- (C_{jw}, C_{jb}) count callbacks for each race:

$$C_{jw} = \sum_{\ell=1}^{L} 1\{R_{j\ell} = w\} Y_{j\ell}, \; C_{jb} = \sum_{\ell=1}^{L} 1\{R_{j\ell} = b\} Y_{j\ell} \ .$$

Bernoulli Trials

Assumption 1. *Bernoulli trials*:

$$Y_{j\ell}(r)|R_{j1}...R_{jL} \overset{iid}{\sim} Bernoulli(p_{jr}), \quad r \in \{w, b\}$$

- Potential outcomes are independent of $\{R_{jk}\}_{k=1}^{L}$ by virtue of random assignment
- Key restriction is that callbacks are *independent* trials
  - Rules out serial dependence ("runs") in callbacks
  - Rules out *interference* between apps – e.g., firms calling back first qualifed app and ignoring subsequent apps
- Surprisingly good approximation for $L \leq 8$.

Defining Discrimination

- Under Assumption 1, each job is characterized by a stable pair of race-by-job callback probabilities (p_{jw}, p_{jb})
- Define discrimination as $D_j = 1\{p_{jw} \neq p_{jb}\}$
- Distinguish idiosyncratic/ex-post $(Y_{j\ell}(w) \neq Y_{j\ell}(b))$ vs. systematic/ex-ante $(p_{jw} \neq p_{jb})$ discrimination
- Systematic definition is relevant for prospective enforcement: EEOC mission is to "**prevent** and remedy unlawful employment discrimination"

Binomial Mixtures

Probability of callback config $(C_{jw} = c_w, C_{jb} = c_b)$ at job j is:

$$f(c_w, c_b | p_{jw}, p_{jb}) = \binom{L_w}{c_w} p_{jw}^{c_w} (1 - p_{jw})^{L_w - c_w} \times \binom{L_b}{c_b} p_{jb}^{c_b} (1 - p_{jb})^{L_b - c_b}$$

Assumption 2. *Random sampling*:

$$(p_{jw}, p_{jb}) \overset{iid}{\sim} G(.,.)$$

- Unconditional callback probabilities are mixtures of binomials:

$$\Pr(C_{jw} = c_w, C_{jb} = c_b) = \int f(c_w, c_b | p_w, p_b) dG(p_w, p_b) \equiv \bar{f}(c_w, c_b)$$

- "Mixing distribution" $G(\cdot, \cdot)$ governs heterogeneity in callback rates across employers

Importance of $G(\cdot,\cdot)$

$G(\cdot,\cdot)$ characterizes prevalence and severity of discrimination

- Prevalence of discrimination:

$$\bar{\pi} = \Pr\left(D_j = 1\right) = \int_{p_w \neq p_b} dG(p_w, p_b)$$

- Severity reflected in moments

$$\int \left(p_w - p_b\right)^k dG\left(p_w, p_b\right)$$

Indirect Evidence

By Bayes' rule, prevalence of discrimination among jobs with callback configuration $(C_{jw} = c_w, C_{jb} = c_b)$ is:

$$\begin{aligned}
\pi\left(c_w, c_b\right) &= \Pr(D_j = 1 | C_{jw} = c_w, C_{jb} = c_b) \\
&= \frac{\int_{p_w \neq p_b} f(c_w, c_b | p_w, p_b) dG(p_w, p_b)}{\bar{f}(c_w, c_b)} \\
&= \mathcal{P}\left(\underbrace{c_w, c_b}_{\text{direct}}, \underbrace{G\left(\cdot, \cdot\right)}_{\text{indirect}}\right)
\end{aligned}$$

- "Posterior" $\mathcal{P}$ blends direct evidence on a job's own behavior with *indirect* evidence on the population from which it was drawn
- If "prior" $\bar{\pi} \in \{0, 1\}$, no need for direct evidence

Empirical Bayes

EB approach forms empirical posteriors

$$\hat{\pi}(c_w, c_b) = \mathcal{P}\left(c_w, c_b, \hat{G}(\cdot, \cdot)\right)$$

- Closely related to mult. testing literature on False Discovery Rates (Benjamini and Hochberg, 1995).
  - Here, $1 - \pi(c_w, c_b)$ corresponds to the pFDR of Storey (2002)
  - $\hat{\pi}(c_w, c_b)$ enables computation of "q-value" of detection rule
- Illustrate more complex uses of $\hat{G}$ when prevalence and intensity both important

Identification

Moments of $G(\cdot,\cdot)$

With $L \leq 20$, inappropriate to treat counts as truth plus normal noise (Brown, 2008).

- Obstructs identification of G but some moments identified
- Marginal callback probabilities are related to moments of G by

$$\bar{f}(c_w, c_b) = \mathbb{E}\left[\binom{L_w}{c_w} p_{jw}^{c_w}\left(1-p_{jw}\right)^{L_w-c_w} \times \binom{L_b}{c_b} p_{jb}^{c_b}\left(1-p_{jb}\right)^{L_b-c_b}\right]$$

$$= \binom{L_w}{c_w}\binom{L_b}{c_b} \sum_{x=0}^{L_w-c_w} \sum_{s=0}^{L_b-c_b} (-1)^{x+s} \binom{L_w-c_w}{x}\binom{L_b-c_b}{s}$$

$$\times \mathbb{E}\left[p_{jw}^{c_w+x} p_{jb}^{c_b+s}\right].$$

- Collect into system relating callback probs $\bar{f}$'s to moments $\mu(m,n) = \mathbb{E}[p_{jw}^m p_{jb}^n]$:

$$\bar{f} = B\mu \implies \mu = B^{-1}\bar{f}$$

Identification

Lemma 1. (Identification of Moments): *Under Assumptions 1 and 2, all moments $\mu(m,n)$ for $0 \leq m \leq L_w$ and $0 \leq n \leq L_b$ are identified.*

- Example: Variance of discrimination is

$$\mathbb{V}\left[p_{jb} - p_{jw}\right] = [\mu(0,2) - \mu(0,1)^2] + [\mu(2,0) - \mu(1,0)^2] - 2[\mu(1,1) - \mu(0,1)\mu(1,0)]$$

- Lemma 1 implies this variance is identified with two or more applications per race
- **Overdispersion** intuition: success probabilities must be heterogeneous if callback frequencies are more variable than would be predicted by Bernoulli uncertainty

Posteriors and prevalence

What features of G are needed to form posterior $\mathcal{P}\left(c_w, c_b, G\left(\cdot,\cdot\right)\right)$?

- Define $\bar{\pi}_t = \Pr\left(D_j = 1 | C_{wj} + C_{bj} = t\right)$ as prevalence in callback stratum $t \in \{0, ..., L\}$
- Exploiting binomial structure, can write posterior $\mathcal{P}$ as ▸ details

$$
1 - \underbrace{\underbrace{\left[1 - \bar{\pi}_{c_w + c_b}\right]}_{\text{prior that } D=0} \underbrace{\frac{\binom{L_w}{c_w}\binom{L_b}{c_b}}{\binom{L}{c_w + c_b}}}_{\text{likelihood if } D=0} \underbrace{\frac{\sum_{x=0}^{L_w} \bar{f}(x, c_w + c_b - x)}{\bar{f}(c_w, c_b)}}_{1/\text{marginal}}}_{\text{pFDR}}
$$

- Callback probs $\bar{f}$ identified $\Rightarrow \mathcal{P}$ known up to stratum specific prevalences $\{\bar{\pi}_t\}_{t=0}^{L}$

Robust Bayes approach: use identified moments μ to *bound* posterior $\mathcal{P}$

Bounds on prevalence

Sharp lower bound on prevalence of discrimination given callback probs $\bar{f}$:

$$\bar{\pi} \geq \min_{G \in \mathscr{G}} \int_{p_w \neq p_b} dG(p_w, p_b) \quad s.t. \quad \bar{f} = B\mu_G$$

- Search over space $\mathscr{G}$ of discretized bivarate CDFs (Noubiap et al., 2001)
- Objective and constraints are linear in p.m.f associated with $G(\cdot, \cdot)$ $\implies$ apply linear programming routine ▸ details
- Tighter bound than in FDR literature (Efron et al, 2001; Storey, 2002)

Same approach can be used to bound prevalence of *directional* notions of discrimination

- Share discriminating against blacks $\int_{p_b < p_w} dG(p_b, p_w)$
- Share "reverse" discriminating against whites $\int_{p_b > p_w} dG(p_b, p_w)$

Lower bounds on $\{\bar{\pi}_t\}_{t=0}^{L} \mapsto$ lower bounds on $\mathcal{P}$

Correspondence Experiments

Data

Apply methods to data from three resume correspondence studies:

- Bertrand and Mullainathan (2004): Racial discrimination in Boston/Chicago
- Nunley et al. (2015): Racial discrimination among recent college graduates in the US
- Arceo-Gomez and Campos-Vasquez (2014, "AGCV"): Gender discrimination in Mexico

Table I: Descriptive statistics for resume correspondance studies

	Bertrand & Mullainathan (1)	Nunley et al. (2)	Arceo-Gomez & Campos-Vasquez (3)
Number of jobs	1,112	2,305	799
Applications per job	4	4	8
Treatment/control	Black/white	Black/white	Male/female
Callback rates: Total	0.079	0.167	0.123
Treatment	0.063	0.154	0.108
Control	0.094	0.180	0.138
Difference	-0.031	-0.026	-0.033
	(0.007)	(0.007)	(0.008)

Are Callbacks Independent Trials?

Testing Assumption 1

Our key *iid* trials assumption has testable implications

- Test 1: Exploit information on order of resumes in AGCV
  - In strata defined by total callbacks, all possible sequences should be equally likely
  - With dependence would generally expect "runs" of consecutive successes/failures
  - Compare Pearson χ^2 and exact multinomial goodness of fit p-values (Cressie and Read, 1989) details
- Test 2: Look for interference using observed characteristics
  - Random assignment of resume characteristics $\implies$ some resumes face stronger competition
  - Ask whether callbacks are affected by characteristics of other applications to the same job
  - In Nunley et al. data, racial mix of resumes varies randomly – yields overidentification of some moments

No Evidence of Dependence in AGCV

Tests for dependence, AGCV data

Callbacks	Observations (1)	χ^2 statistic (2)	d.f. (3)	P-value (4)	Exact p-value (5)
Panel A. Four-application sequences					
1	142	1.4	3	0.708	0.794
2	99	10.0	5	0.075	0.155
3	64	3.2	3	0.367	0.513
Panel B. Eight-application sequences					
1	56	7.8	7	0.347	0.504
2	37	23.6	27	0.651	0.697
3	36	58.4	55	0.352	0.397
4	39	75.2	69	0.286	0.457
5	16	40.7	55	0.924	1.000
6	20	28.6	27	0.379	0.469
7	6	8.4	7	0.300	0.539

Panel C. Joint tests

Independence in all callback strata: χ^2 (247) = 242.7, p = 0.565

No order effects: χ^2 (7) = 5.3, p = 0.622

Regression of callback on order: coef. = -0.0021, s.e. = 0.0015, p = 0.147

Regression of callback on frac. females sent earlier: coef. = -0.003, s.e. = 0.013, p = 0.788

No Evidence of Dependence in AGCV

Tests for dependence, AGCV data

Callbacks	Observations (1)	χ^2 statistic (2)	d.f. (3)	*P*-value (4)	Exact *p*-value (5)
Panel A. Four-application sequences					
1	142	1.4	3	0.708	0.794
2	99	10.0	5	0.075	0.155
3	64	3.2	3	0.367	0.513
Panel B. Eight-application sequences					
1	56	7.8	7	0.347	0.504
2	37	23.6	27	0.651	0.697
3	36	58.4	55	0.352	0.397
4	39	75.2	69	0.286	0.457
5	16	40.7	55	0.924	1.000
6	20	28.6	27	0.379	0.469
7	6	8.4	7	0.300	0.539

Panel C. Joint tests

Independence in all callback strata: $\chi^2(247) = 242.7$, $p = 0.565$

No order effects: $\chi^2(7) = 5.3$, $p = 0.622$

Regression of callback on order: coef. = -0.0021, s.e. = 0.0015, $p = 0.147$

Regression of callback on frac. females sent earlier: coef. = -0.003, s.e. = 0.013, $p = 0.788$

No Evidence That Callbacks Are Rival in Nunley et al

Tests for dependence, NPRS data

Variable	Main effect (1)	Leave-out mean (2)
Black	-0.028	-0.019
	(0.010)	(0.027)
Female	0.010	0.009
	(0.010)	(0.027)
High SES	-0.233	-0.674
	(0.174)	(0.522)
GPA	-0.043	-0.153
	(0.066)	(0.198)
Business major	0.008	0.010
	(0.008)	(0.021)
Employment gap	0.011	0.034
	(0.009)	(0.023)
Current unemp.: 3+	0.013	0.005
	(0.012)	(0.032)
6+	-0.008	-0.038
	(0.012)	(0.029)
12+	0.001	0.021
	(0.012)	(0.032)
Past unemp.: 3+	0.029	0.065
	(0.012)	(0.031)
6+	-0.011	-0.016
	(0.012)	(0.033)
12+	-0.004	0.019
	(0.012)	(0.031)
Predicted callback rate	0.476	-0.041
	(0.248)	(0.626)
Joint *p* -value	$\boxed{0.452}$	
Sample size	9,220	

Moment Estimates

Moment Estimation

- Estimate moments by GMM, and “shape-constrained” GMM requiring moments to be consistent with a coherent probability distribution
- Shape-constrained estimator finds set of discrete $G(\cdot,\cdot)$’s that come closest to matching observed callback frequencies details
- Standard errors based on “numerical bootstrap” of Hong and Li (2017) details
- Test model restrictions using bootstrap method of Chernozhukov, Newey, and Santos (2015) details

First Two Moments of $G(\cdot,\cdot)$ Are Identified in BM

Table A.I: Moments of callback rate distribution, BM data

Moment	Estimate
$E[p_w]$	0.094
	(0.006)
$E[p_b]$	0.063
	(0.006)
$E[(p_w - E[p_w])^2]$	0.040
	(0.005)
$E[(p_b - E[p_b])^2]$	0.023
	(0.004)
$E[(p_w - E[p_w])(p_b - E[p_b])]$	0.028
	(0.004)
$E[(p_w - E[p_w])^2(p_b - E[p_b])]$	0.015
	(0.003)
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.023
	(0.003)
$E[(p_w - E[p_w])^2(p_b - E[p_b])^2]$	0.010
	(0.003)
Sample size	1,112

Shape Constraints Do Not Bind

Table A.I: Moments of callback rate distribution, BM data

Moment	No constraints (1)	Shape constraints (2)
$E[p_w]$	0.094	0.094
	(0.006)	(0.007)
$E[p_b]$	0.063	0.063
	(0.006)	(0.006)
$E[(p_w - E[p_w])^2]$	0.040	0.040
	(0.005)	(0.005)
$E[(p_b - E[p_b])^2]$	0.023	0.023
	(0.004)	(0.004)
$E[(p_w - E[p_w])(p_b - E[p_b])]$	0.028	0.028
	(0.004)	(0.003)
$E[(p_w - E[p_w])^2(p_b - E[p_b])]$	0.015	0.014
	(0.003)	(0.002)
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.023	0.012
	(0.003)	(0.002)
$E[(p_w - E[p_w])^2(p_b - E[p_b])^2]$	0.010	0.010
	(0.003)	(0.002)
	J-statistic:	0.0
	P-value:	1.00
Sample size	1,112	

Substantial Variation in Discrimination

Table III.A: Treatment effect variation in BM (2004)

	p_b (1)	p_w (2)	$p_b - p_w$ (3)
Mean	0.063	0.094	-0.031
	(0.006)	(0.007)	(0.006)
Standard deviation	0.152	0.199	0.082
	(0.012)	(0.012)	(0.016)
Correlation with p_w or p_f	0.927	1.00	-0.717
	(0.051)	-	(0.119)

First Two Moments in Nunley et al. Data

Table A.II: Moments of callback rate distribution, NPRS data

Moment	(2,2) design
$E[p_w]$	0.174 (0.010)
$E[p_b]$	0.148 (0.010)
$E[(p_w - E[p_w])^2]$	0.089 (0.007)
$E[(p_b - E[p_b])^2]$	0.085 (0.007)
$E[(p_w - E[p_w])(p_b - E[p_b])]$	0.083 (0.006)
$E[(p_w - E[p_w])^2(p_b - E[p_b])]$	0.044 (0.004)
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.047 (0.005)
$E[(p_w - E[p_w])^2(p_b - E[p_b])^2]$	0.036 (0.004)
Sample size	1,146

Extra Designs Identify Additional Moments

Table A.II: Moments of callback rate distribution, NPRS data

Moment	(2,2) design (1)	(3,1) design (2)	(1,3) design (3)
$E[p_w]$	0.174 (0.010)	0.199 (0.025)	0.142 (0.015)
$E[p_b]$	0.148 (0.010)	0.149 (0.015)	0.157 (0.013)
$E[(p_w - E[p_w])^2]$	0.089 (0.007)	0.108 (0.009)	-
$E[(p_b - E[p_b])^2]$	0.085 (0.007)	-	0.083 (0.008)
$E[(p_w - E[p_w])(p_b - E[p_b])]$	0.083 (0.006)	0.084 (0.009)	0.080 (0.009)
$E[(p_w - E[p_w])^3]$	-	0.051 (0.008)	-
$E[(p_b - E[p_b])^3]$	-	-	0.044 (0.007)
$E[(p_w - E[p_w])^2(p_b - E[p_b])]$	0.044 (0.004)	0.043 (0.007)	-
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.047 (0.005)	-	0.045 (0.007)
$E[(p_w - E[p_w])^3(p_b - E[p_b])]$	-	0.034 (0.005)	-
$E[(p_w - E[p_w])(p_b - E[p_b])^3]$	-	-	0.037 (0.006)
$E[(p_w - E[p_w])^2(p_b - E[p_b])^2]$	0.036 (0.004)	-	-
Sample size	1,146	544	550

Joint Test of All Restrictions Fails to Reject

Table A.II: Moments of callback rate distribution, NPRS data

Moment	(2,2) design (1)	(3,1) design (2)	(1,3) design (3)	P-value (4)	Combined estimates (5)
	Design-specific estimates				
$E[p_w]$	0.174 (0.010)	0.199 (0.025)	0.142 (0.015)	0.027	0.177 (0.007)
$E[p_b]$	0.148 (0.010)	0.149 (0.015)	0.157 (0.013)	0.854	0.153 (0.007)
$E[(p_w - E[p_w])^2]$	0.089 (0.007)	0.108 (0.009)	-	0.097	0.095 (0.005)
$E[(p_b - E[p_b])^2]$	0.085 (0.007)	-	0.083 (0.008)	0.857	0.084 (0.005)
$E[(p_w - E[p_w])(p_b - E[p_b])]$	0.083 (0.006)	0.084 (0.009)	0.080 (0.009)	0.926	0.084 (0.004)
$E[(p_w - E[p_w])^3]$	-	0.051 (0.008)	-		0.106 (0.007)
$E[(p_b - E[p_b])^3]$	-	-	0.044 (0.007)		0.092 (0.006)
$E[(p_w - E[p_w])^2(p_b - E[p_b])]$	0.044 (0.004)	0.043 (0.007)	-	0.875	0.040 (0.002)
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.047 (0.005)	-	0.045 (0.007)	0.819	0.042 (0.002)
$E[(p_w - E[p_w])^3(p_b - E[p_b])]$	-	0.034 (0.005)	-	-	0.035 (0.002)
$E[(p_w - E[p_w])(p_b - E[p_b])^3]$	-	-	0.037 (0.006)	-	0.037 (0.002)
$E[(p_w - E[p_w])^2(p_b - E[p_b])^2]$	0.036 (0.004)	-	-	-	0.038 (0.002)
				J-statistic:	$\boxed{23.1}$
				P-value:	$\boxed{0.190}$
Sample size	1,146	544	550		2,240

Treatment Effects Are Variable and Skewed

Table III.B: Treatment effect variation in NPRS (2015)

	p_b (1)	p_w (2)	$p_b - p_w$ (3)
Mean	0.153	0.177	-0.023
	(0.007)	(0.007)	(0.005)
Standard deviation	0.290	0.308	0.102
	(0.008)	(0.007)	(0.012)
Correlation with p_w or p_f	0.944	1.00	-0.336
	(0.017)	-	(0.066)
Skewness	3.76	3.65	-4.45
	(0.08)	(0.08)	(0.82)

Thick Tail of Extreme Discriminators in AGCV

Table III.C: Treatment effect variation in AGCV (2014)

	p_m (1)	p_f (2)	$p_m - p_f$ (3)
Mean	0.109	0.137	-0.028
	(0.009)	(0.010)	(0.008)
Standard deviation	0.229	0.257	0.178
	(0.012)	(0.011)	(0.014)
Correlation with p_w or p_f	0.738	1.00	-0.498
	(0.039)	-	(0.058)
Skewness	4.04	3.74	-1.64
	(0.13)	(0.10)	(0.56)
Excess kurtosis	8.59	5.91	13.6
	(1.13)	(0.71)	(3.5)

Prevalence and Posteriors

In BM, At Least 13% of Jobs Discriminate

Lower bounds on discrimination probabilities, BM data

	Share discriminating: $\Pr(p_w \neq p_b)$
	(1)
	$\boxed{0.130}$
J-statistic:	29.26
P-value (bound = 0):	0.000

At Least 44% Making Two Total Calls Discriminate

Lower bounds on discrimination probabilities, BM data

Callbacks	Share discriminating: $\Pr(p_w \neq p_b)$ (1)
All	0.130
0	0.038
1	0.424
2	$\boxed{0.442}$
3	0.508
4	0.212
J-statistic:	29.26
P-value (bound = 0):	0.000

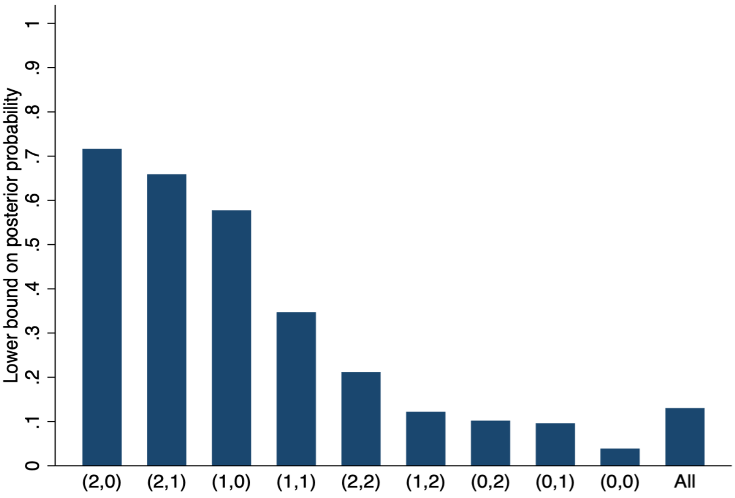
Cannot Reject Absence of Discrimination Against Whites

Lower bounds on discrimination probabilities, BM data

Callbacks	Share discriminating: $\Pr(p_w \neq p_b)$ (1)	Share disc. against whites: $\Pr(p_w < p_b)$ (2)	Share disc. against blacks: $\Pr(p_b < p_w)$ (3)
All	0.130	0.000	0.130
0	0.038	0.000	0.038
1	0.424	0.000	0.424
2	0.442	0.000	0.442
3	0.508	0.000	0.508
4	0.212	0.000	0.212
J-statistic:	29.26	0.00	29.26
P-value (bound = 0):	0.000	1.000	0.000

At Least 72% With $(C_{jw}, C_{jb}) = (2, 0)$ Discriminate

Figure I: Lower bounds on posterior probabilities of discrimination, BM data



In Nunley et al., Cannot Reject $\Pr(p_{jw} < p_{jb}) = 0$

Lower bounds on discrimination probabilities, Nunley et al. data

Callbacks	Share discriminating: $\Pr(p_w \neq p_b)$ (1)	Share disc. against whites: $\Pr(p_w < p_b)$ (2)	Share disc. against blacks: $\Pr(p_b < p_w)$ (3)
All	0.358	0.154	0.173
0	0.152	0.093	0.048
1	0.672	0.185	0.433
2	0.691	0.016	0.675
3	0.821	0.067	0.736
4	0.421	0.257	0.128
J-statistic:	62.64	23.46	62.64
P-value (bound = 0):	0.000	0.120	0.000

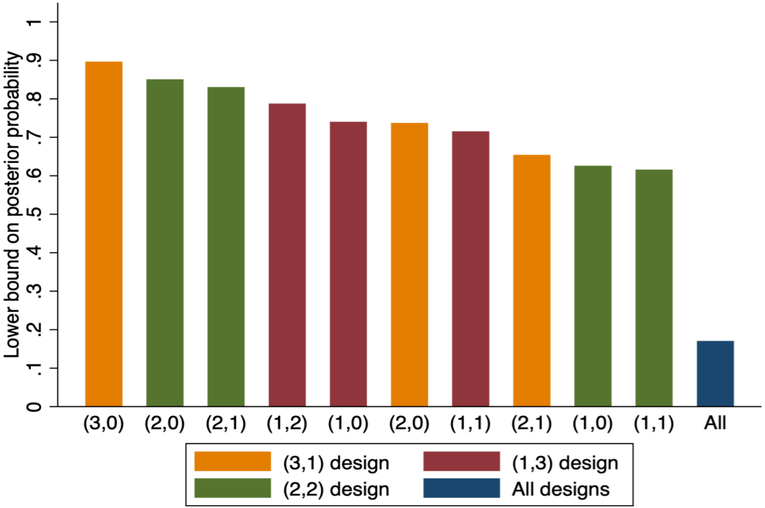
At Least 68% That Make Two Calls Have $p_{jb} < p_{jw}$

Lower bounds on discrimination probabilities, Nunley et al. data

Callbacks	Share discriminating: $\Pr(p_w \neq p_b)$ (1)	Share disc. against whites: $\Pr(p_w < p_b)$ (2)	Share disc. against blacks: $\Pr(p_b < p_w)$ (3)
All	0.358	0.154	0.173
0	0.152	0.093	0.048
1	0.672	0.185	0.433
2	0.691	0.016	$\boxed{0.675}$
3	0.821	0.067	0.736
4	0.421	0.257	0.128
J-statistic:	62.64	23.46	62.64
P-value (bound = 0):	0.000	0.120	0.000

Lower Bounds on Posteriors Above 85%

Figure II: Lower bounds on posterior probabilities of discrimination, Nunley et al. data



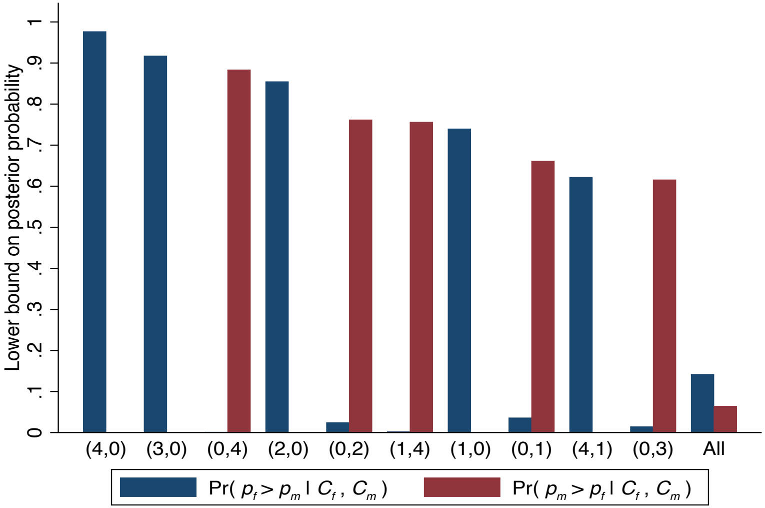
In AGCV, Discrimination Against Both Men and Women

Lower bounds on discrimination probabilities, AGCV data

Callbacks	Share discriminating: $\Pr(p_f \neq p_m)$ (1)	Share disc. women: $\Pr(p_f < p_m)$ (2)	Share disc. against men: $\Pr(p_m < p_f)$ (3)
All	0.207	0.064	0.142
0	0.065	0.023	0.042
1	0.721	0.307	0.414
2	0.708	0.226	0.481
3	0.584	0.050	0.533
4	0.518	0.053	0.465
5	0.320	0.153	0.167
6	0.372	0.176	0.197
7	0.453	0.122	0.331
8	0.069	0.008	0.062
J-statistic:	427.8	27.1	421.0
P-value:	0.000	0.018	0.000

Lower Bounds on Posteriors Above 90%

Figure III: Lower bounds on posterior probabilities of discrimination, AGCV data



Detection Error Tradeoffs

Experimental Design and Detection Error Tradeoffs

Results so far establish that some callback patterns produce high posterior probabilities of discrimination even with few applications per job

- But few jobs produce these patterns. Can correspondence experiments serve as a useful tool for detecting discrimination when prevalence is low?
- Consider alternative hypothetical experiments based on models fit to the Nunley et al. (2015) data
- Take the perspective of hypothetical regulator who knows $G(\cdot, \cdot)$ and must decide which jobs to investigate based upon callbacks
  - Investigations are costly, want to detect most extreme discriminators
  - Start with a parametric model for $G(\cdot, \cdot)$ then ask how regulator's decisions are affected by second-guessing parametric assumptions
  - Detection/error tradeoff (DET) curves: tradeoff between true negatives and true positives for a fixed number of apps

Mixed Logit

Logit model for callback to application ℓ at job j:

$$\Pr\left(Y_{j\ell} = 1 | \alpha_j, \beta_j, R_{j\ell}, X_{j\ell}\right) = \Lambda\left(\alpha_j - \beta_j 1\{R_{j\ell} = b\} + X'_{j\ell}\psi\right).$$

- $\Lambda(x) \equiv \exp(x)/(1+\exp(x))$ is the logistic CDF
- $R_{j\ell}$ indicates race, $X_{j\ell}$ includes other randomly-assigned characteristics (GPA, experience, etc.)
- Two-type mixing:

$$\alpha_j \sim N\left(\alpha_0, \sigma_\alpha^2\right),$$

$$\beta_j = \begin{cases} \beta_0, & \text{with prob. } \Lambda(\tau_0 + \tau_\alpha \alpha_j), \\ 0, & \text{with prob. } 1 - \Lambda(\tau_0 + \tau_\alpha \alpha_j). \end{cases}$$

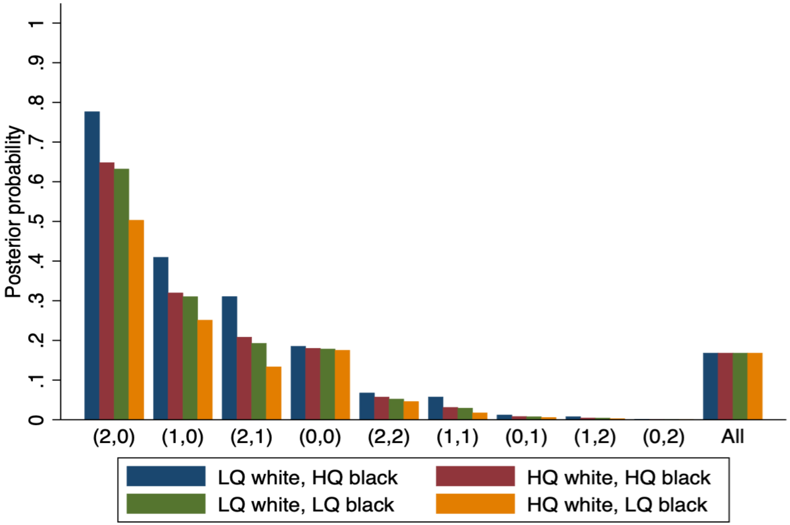
Discrimination is Rare But Intense

Table V: Mixed logit parameter estimates, NPRS data

			Types	
		Constant	No selection	Selection
		(1)	(2)	(3)
Distribution of logit(p_w):	α_0	-4.71	-4.93	-4.93
		(0.22)	(0.24)	(0.28)
	σ_α	4.74	4.99	4.98
		(0.22)	(0.25)	(0.29)
Discrimination intensity:	β_0	0.456	4.05	4.05
		(0.108)	(1.56)	(1.58)
Discrimination logit:	τ_0	-	-1.59	-1.56
			(0.42)	(1.10)
	τ_α	-	-	-0.005
				(0.180)
Fraction with $p_w \neq p_b$:		1.00	0.168	0.170
Log-likelihood		-2,792.1	-2,788.2	-2,788.2
Parameters		15	16	17
Sample size		2,305	2,305	2,305

Covariates Generate Variation in Posteriors

Figure IV: Mixed logit estimates of posterior discrimination probabilities, Nunley et al. data



Regulator's Problem

Consider a regulator who knows G and must choose whether to investigate, $\delta_j \in \{0,1\}$, based upon callbacks (C_{jw}, C_{jb})

- Regulator seeks to minimize loss function:

$$\mathcal{L}_j(\delta_j) = \delta_j \times \left(\kappa - \Lambda\left(\Lambda^{-1}(p_{jw}) - \Lambda^{-1}(p_{jb})\right)\right)$$

- Intuition: regulator wants to investigate employers with large logit coefficients on race

Optimal decision rule $\delta(C_{jb}, C_{jw})$ minimizes expected loss (risk)

$$\mathcal{R}(G,\delta) \equiv \mathbb{E}[\mathcal{L}_j(\delta(C_{jw}, C_{jb}))]$$

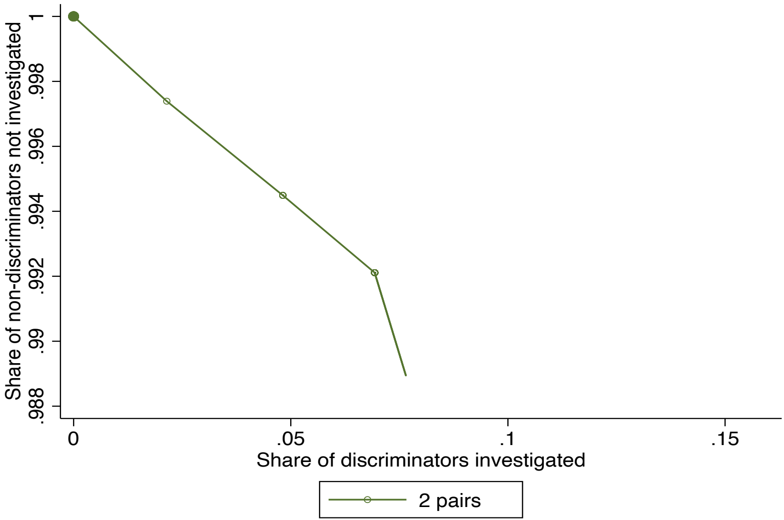
- In the two-type mixed logit, this results in a posterior cutoff rule:

$$\delta(C_{jw}, C_{jb}) = 1\left\{\mathcal{P}(C_{jw}, C_{jb}, G(.,.)) > \frac{\kappa - 1/2}{\Lambda(\beta_0) - 1/2}\right\}$$

- Focus on example where κ is such that posterior cutoff is 80%.

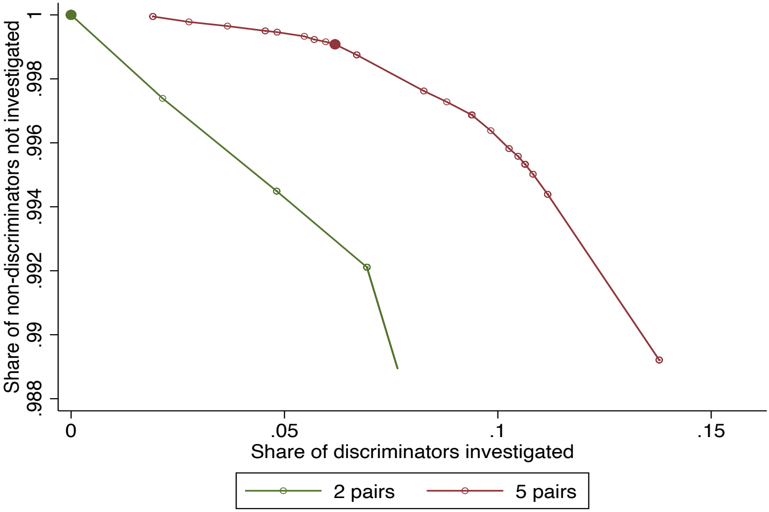
With 2 Pairs, 80% Threshold Yields Few Investigations

Figure V: Detection/error tradeoffs, NPRS data



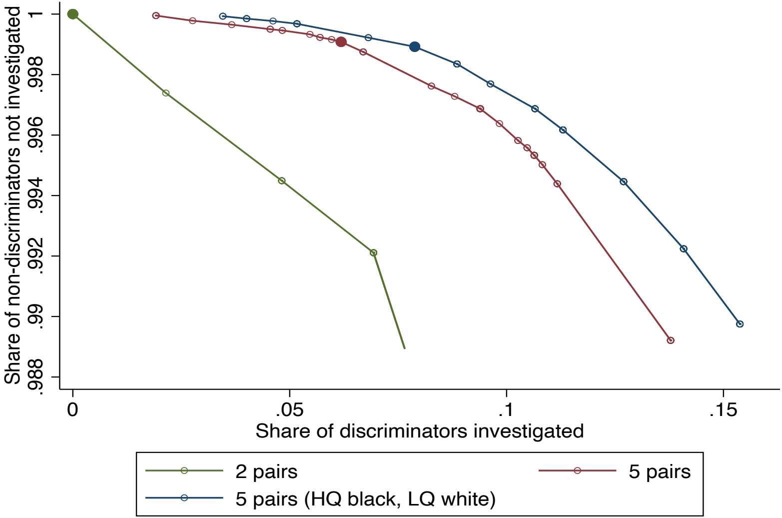
Sending 5 Pairs Boosts Detection Substantially

Figure V: Detection/error tradeoffs, NPRS data



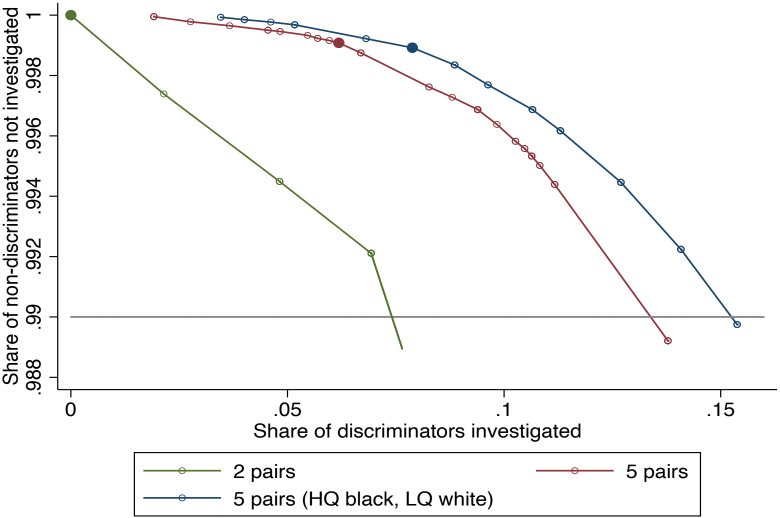
Leveraging Covariates Yields Further Gains

Figure V: Detection/error tradeoffs, NPRS data



Fixing Size at 0.01 Yields More (Mostly False) Accusations

Figure V: Detection/error tradeoffs, NPRS data



Accommodating Ambiguity

Beyond Logit: Policy When Partially Identified

- How would decisions change if the regulator fears that $G(\cdot,\cdot)$ is not logit?
- Important (extreme) benchmark for decisionmaking under ambiguity: minimax decision rule
- Max risk function and minimax decision rule when auditor knows G lies in some identified set Θ:

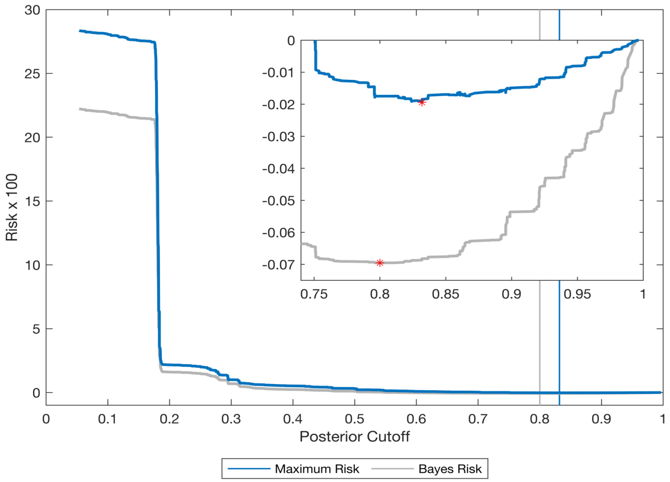
$$\mathcal{R}_m(\Theta,\delta) \equiv \sup_{G\in\Theta} \mathcal{R}(G,\delta), \;\; \delta^{mm} \equiv \arg\inf_{\delta} \mathcal{R}_m(\Theta,\delta)$$

- Minimax regulator chooses δ^{mm} to minimize risk, assuming nature will select the least favorable distribution in Θ in response to any decision rule ("Γ-minimax")
- Manage space of decision rules by considering a restricted set defined by logit posterior cutoffs
- Contrast risk and decisions based upon mixed logit prior and minimax

details

Minimax Regulator Chooses Slightly Higher Threshold

Figure VI: Bayes and minimax risk, NPRS data



Concluding Thoughts

- Tremendous heterogeneity in discrimination $\Rightarrow$ enforcing equal opportunity is a difficult inferential problem
  - Results today suggest favorable detection rates achievable with minor modifications to standard audit designs

- Ongoing work
  - Jobs vs firms: is bad behavior clustered in particular companies? How to construct reliable rankings?
  - Optimal experimental design: dynamic auditing to detect effects at lower cost

- Methods applicable to other settings where behavioral responses of individual units are of interest. Examples:
  - Workplace safety audits (Levine et al., 2012)
  - Choice experiments (Halevy et al., 2018)
  - Evaluating schools / teachers (Chetty et al., 2014; Angrist et al, 2017)

Bonus

Dynamic Auditing (Avivi, Kline, Rose, Walters, in progress)

Letting H_n denote the job history information available as of app $\#n$, we can write the value function:

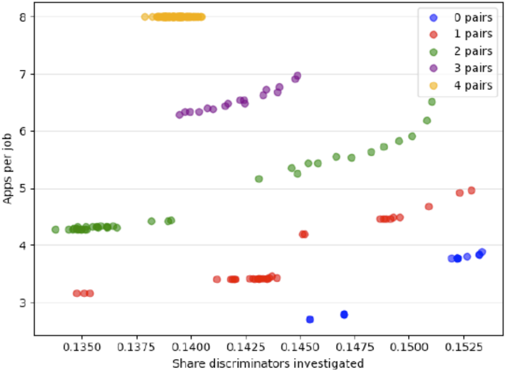
$$V(H_n) = \begin{cases} \max\left\{ \underbrace{\max_{r\in\{w,b\},x\in\{hi,lo\}} v_{rx}(H_n)}_{\text{send optimal app}}, \underbrace{v_I(H_n)}_{\text{investigate}}, \underbrace{0}_{\text{give up}} \right\} & \text{if } n<K \\ \max\left\{ \underbrace{v_I(H_n)}_{\text{investigate}}, \underbrace{0}_{\text{give up}} \right\} & \text{if } n=K \end{cases}$$

where r is race, x is quality, $K=8$ is max $\#$ of apps to a job and:

- $v_{rx}(H_n) = -c + \mathbb{E}_n[V(H_{n+1})]$
- $v_I(H_n) = \underbrace{\int \mathbb{E}_n[p_{jw}(x) - p_{jb}(x)]\, dF(x)}_{\text{investigation yield}} - \underbrace{\kappa}_{\text{cost}}$

Dynamic auditor (0 pairs) requires <1/2 as many apps to detect discriminators as static auditor (4 pairs)

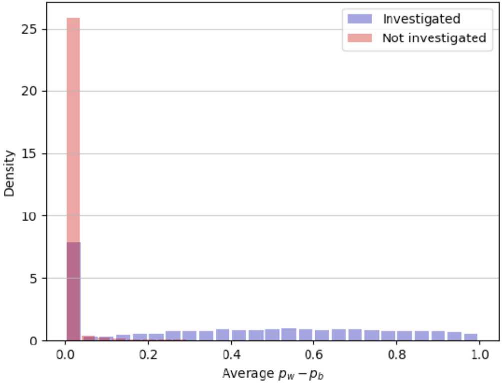
Figure 2: Share discriminators investigated and number of applications per job, when 5.5-6% of jobs are investigated



Note: This figure presents the share of discriminators investigated and the average number of applications per job when 5.5-6% of jobs are investigated by different number of initial pairs. The estimates are coming from simulating the auditor decision assuming the data was generated from the censored logit model in column (1) of Table (1). The curves are generated by varying κ between 0.01 to 0.09 and c between 0.00001 to 0.004.

Dynamic auditing detects intense discriminators

Figure 6: The conditional distribution of average $p_w - p_b$ given investigation status, starting with 0 pairs, $\kappa = 0.13$, $c = 0.0001$



Note: This figures presents the conditional distribution of average $p_w - p_b$ given investigation status for the dynamic auditor with payoff parameters $\kappa = 0.13$ and $c = 0.0001$. The densities were generated by simulating the auditor decision assuming the data was generated from the censored logit model in column (1) of Table (1).

Discretization of G

- We approximate $G(p_w, p_b)$ with the discrete distribution:

$$G_K(p_w, p_b) = \sum_{k=1}^{K} \sum_{l=1}^{K} \eta_{kl} 1\{p_w \leq \varrho(k, l), p_b \leq \varrho(l, k)\}$$

- $\{\eta_{kl}\}_{k=1,l=1}^{K,K}$ are probability masses
- $\{\varrho(k, l), \varrho(l, k)\}_{k=1,l=1}^{K,K}$ are a set of mass point coordinates generated by

$$\varrho(x, y) = \underbrace{\frac{\min\{x, y\} - 1}{K}}_{\text{diagonal}} + \underbrace{\frac{\max\{0, x - y\}^2}{K(1 + K - y)}}_{\text{off-diagonal}}.$$

- Gives a two-dimensional grid with K^2 elements, equally spaced along the diagonal and quadratically spaced off the diagonal according to distance from diagonal

Shape Constrained GMM

- Let $\tilde{f}$ denote vector of empirical callback frequencies
- Shape constrained GMM estimator of η solves quadratic programming problem:

$$\hat{\eta} = \arg\inf_{\eta} \ (\tilde{f} - BM\eta)' W (\tilde{f} - BM\eta) \ s.t. \ \eta \geq 0, \ \mathbf{1}'\eta = 1.$$

- M is a $dim(\mu) \times K^2$ matrix defined so that $M\eta = \mu$ for G_K
- Yields shape constrained moment estimates: $\hat{\mu} = M\hat{\eta}$
- W is weighting matrix – use two-step optimal weighting
- Set $K = 150$ for GMM estimation

back

Hong and Li (2017) Standard Errors

- Bootstrap μ^* solves QP problem replacing $\tilde{f}$ with $(\tilde{f} + J^{-1/4} f^*)$, where elements of f^* given by:

$$\sqrt{J}\left[\frac{\sum_j \omega_j^* 1\{C_{jw}=c_w, C_{jb}=c_b\}}{\sum_j \omega_j^*} - \frac{\sum_j 1\{C_{jw}=c_w, C_{jb}=c_b\}}{J}\right].$$

- Weights ω_j^* drawn iid from exponential distribution with mean 0 and variance 1
- Standard errors for $\phi(\hat{\mu})$ computed as standard deviation of $J^{-1/4}[\phi(\mu^*) - \phi(\hat{\mu})]$ across bootstrap replications

back

Chernozhukov et al. (2015) Goodness of Fit Test

- "J-test" goodness of fit statistic:

$$T_n = \inf_{\eta} \, (\tilde{f} - BM\eta)'\hat{\Sigma}^{-1}(\tilde{f} - BM\eta) \; s.t. \; \eta \geq 0, \; \mathbf{1}'\eta = 1$$

- Letting F^* denote (centered) bootstrap analogue of $\tilde{f}$ and W^* a weighting matrix, bootstrap test statistic is

$$T_n^* = \inf_{\pi, h} \, (F^* - BM\eta)'W^*(F^* - BM\eta)$$

$$s.t. \; (\tilde{f} - BM\eta)'W(\tilde{f} - BM\eta) = T_n, \; \eta \geq 0, \; \mathbf{1}'\eta = 1, \; h \geq -\eta, \; 1'h = 0.$$

- As in the full sample, conduct two-step GMM estimation in bootstrap replications
- Calculate p-value as fraction of bootstrap samples with $T_n^* > T_n$
- Solve via Second Order Cone Programming

back

Testing for Dependence Across Trials

- Consider set of J_k jobs making k total calls
- Under the null of *iid* trials, all sequences yielding k successes are equally likely
  - With $L = 4$ and $k = 2$, six possible sequences: (1,1,0,0), (1,0,1,0), (1,0,0,1), (0,1,1,0), (0,1,0,1), (0,0,1,1)
- Test statistic:

$$\hat{T}_k = \sum_{s=1}^{q_k^{-1}} \frac{(\hat{q}_{s,k} - q_k)^2}{q_k(1 - q_k)/J_k}$$

- $\hat{q}_{s,k}$ is empirical frequency of sequence s among those with k calls, $q_k = \begin{pmatrix} L \\ k \end{pmatrix}^{-1}$ is expected frequency under the null
- Under the null $\hat{T}_k$ is χ^2 distributed with $\begin{pmatrix} L \\ k \end{pmatrix} - 1$ degrees of freedom

back

Importance of $\bar{\pi}_t$

- Define the t-conditional quantities where $t = c_w + c_b$ is total callbacks

$$(p_{wj}, p_{bj}) \,|\, C_{wj} + C_{bj} = t \sim G_t(\cdot, \cdot)$$

$$\bar{f}_t(c_w) = \frac{\bar{f}(c_w, t - c_w)}{\sum_{x=0}^{L_w} \bar{f}(x, t - x)}$$

$$f_t(c_w | p_w, p_b) = \frac{f(c_w, t - c_w | p_w, p_b)}{\sum_{x=0}^{L_w} f(x, t - x | p_w, p_b)}$$

- Note by standard sufficiency arguments that

$$f_t(c_w | p, p) = \frac{\binom{L_w}{c_w}\binom{L_b}{t - c_w}}{\binom{L}{t}} = B(t, c_w)$$

Importance of $\bar{\pi}_t$

- Now rewrite the posterior $\mathcal{P}\left(c_w, c_b, G\left(\cdot,\cdot\right)\right)$ as $\mathcal{P}_t\left(c_w, G\left(\cdot,\cdot\right)\right)$

$$
\begin{aligned}
\mathcal{P}_t\left(c_w, G\left(\cdot,\cdot\right)\right) &= \frac{\displaystyle\int_{p_w \neq p_b} f_t(c_w|p_w, p_b)dG_t(p_w, p_b)}{\bar{f}_t(c_w)} \\
&= 1 - \frac{\displaystyle\int_p f_t(c_w|p, p)dG_t(p, p)}{\bar{f}_t(c_w)} \\
&= 1 - B\left(t, c_w\right)\frac{\int_p dG_t(p, p)}{\bar{f}_t(c_w)} \\
&= 1 - B\left(t, c_w\right)\frac{1 - \bar{\pi}_t}{\bar{f}_t(c_w)}
\end{aligned}
$$

- Note $\bar{f}_t(c_w)$ is identified from experimental frequencies, so only unknown here is $\bar{\pi}_t$!

back

Linear Programming

- Optimization problem for computing lower bound on share discriminating:

$$\max_{\{\eta_{kl}\}} \sum_{l=1}^{K}\sum_{k=1}^{K}\eta_{kl}1\{\varrho(k,l)=\varrho(l,k)\} \; s.t. \sum_{k=1}^{K}\sum_{l=1}^{K}\eta_{kl}=1, \quad \eta_{kl}\geq 0$$

- Additional moment constraints for all (c_w, c_b):

$$\bar{f}\left(c_w, c_b\right) = \binom{L_w}{c_w}\binom{L_b}{c_b} \textstyle\sum_{k=1}^{K}\sum_{l=1}^{K}\eta_{kl}$$

$$\times\varrho\left(k,l\right)^{c_w}\left(1-\varrho\left(k,l\right)\right)^{L_w-c_w}\varrho\left(l,k\right)^{c_b}\left(1-\varrho\left(l,k\right)\right)^{L_b-c_b}.$$

- Set $K = 900$ for computing bounds

back

Computing Maximum Risk

- Letting H and L refer to high and low quality covariate values, we approximate $G(p_w^H, p_w^L, p_b^H, p_b^L)$ with

$$G_K(p_w^H, p_w^L, p_b^H, p_b^L) = \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{k'=1}^{K}\sum_{l'=1}^{K} \eta_{klk'l'}$$

$$\times 1\left\{p_w^H \le \varrho(k,l), p_w^L \le \varrho(k',l'), p_b^H \le \varrho(l,k), p_b^L \le \varrho(l',k')\right\}.$$

- Maximal risk function for posterior cutoff q:

$$\mathcal{R}_J^m(q) =$$

$$J \max_{\{\eta_{klk'l'}\}} \sum_{l \in \mathscr{A}_1} w_l \mathbb{E}\left[\delta(C_j, l, q)\left\{\kappa - \Lambda\left(\sum_{x \in \{H,L\}} \frac{\Lambda^{-1}(p_{wj}^x) - \Lambda^{-1}(p_{bj}^x)}{2}\right)\right\} | L_j = l\right]$$

- $\mathscr{A}_1$ is list of possible quality configurations with corresponding probabilities w_a
- Constraints: $\eta_{klk'l'}$ positive and sum to 1, along with matching a list of logit-smoothed callback frequencies
- Joint probabilities $\Pr\left(\delta\left(C_j, a, q\right) = 1, D_j = d\right)$ linear in $\eta_{klk'l'}$ (see Appendix D)
- Set $K = 30$ when computing maximal risk in practice

back