Audits as Evidence: Experiments, Ensembles, and Enforcement

Pat Kline and Chris Walters
UC Berkeley and NBER

July 2019

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)
  - Audit/correspondence experiments (Bertrand and Mullainathan, 2004)
- Understanding variation in discrimination across employers is essential
  - For enforcing the law – e.g. targeting of EEOC investigations
  - For assessing effects on minority workers (Becker, 1957; Charles and Guryan, 2008)
- We develop tools for using correspondence experiments to detect illegal discrimination by individual employers

Agenda: Ensembles and Decisionmaking

- Correspondence studies send multiple applications to each job opening
- We view this as an *ensemble* of many small 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)
- Take the perspective of hypothetical auditor (e.g. the EEOC) who must make decisions about which employers to investigate
- Treat auditor's problem as an exercise in large scale testing (Efron, 2012)
- We develop methods and apply them to 3 experimental data sets

Setup and Notation

- Sample of J jobs, each receiving L_w white and L_b black applications (total $L = L_b + L_w$)
- $R_{j\ell} \in \{b, w\}$ indicates race of application ℓ to job j (randomly assigned)
- $Y_{j\ell} \in \{0, 1\}$ indicates a callback from job j to applicant ℓ
- (C_{jw}, C_{jb}) count callbacks for each race:

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

Potential Outcomes

- *Potential* callback to application ℓ to job j as a function of race r:

$$Y_{j\ell}(r) : \{b, w\} \to \{0, 1\}$$

- Observed callback outcome is $Y_{j\ell} = Y_{j\ell}(R_{j\ell})$
- Represent potential outcomes as job-specific function of race and other factors $U_{j\ell}$:

$$Y_{j\ell}(r) = Y_j(r, U_{j\ell})$$

- **Assumption 1**: *Stable job-specific callback rule:*

$$U_{j\ell} | R_{j1} ... R_{jL} \overset{iid}{\sim} Uniform(0, 1)$$

- Distribution of $U_{j\ell}$ does not depend on $\{R_{jk}\}_{k=1}^{L}$ by virtue of random assignment
- Key restriction is that the $U_{j\ell}$ are independent – rules out e.g. firms calling back first qualifed app and ignoring subsequent apps (test later)

Defining Discrimination

- Under Assumption 1, we have stable race-by-job callback probabilities in repeat experiments:

$$p_{jr} \equiv \int_0^1 Y_j\left(r,u\right)du,\ \ r \in \{b,w\}$$

- Define discrimination as $D_j \equiv 1\{p_{jb} \neq p_{jw}\}$
- Distinguish idiosyncratic/ex-post $(Y_{j\ell}(b) \neq Y_{j\ell}(w))$ vs. systematic/ex-ante $(p_{jb} \neq p_{jw})$ discrimination
- Systematic definition is relevant for prospective enforcement

Binomial Mixtures

- Under Assumption 1, callback counts $C_j = (C_{jw}, C_{jb})$ at employer j are generated by binomial trials:

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

$$\equiv f(c|p_{jw}, p_{jb})$$

- **Assumption 2**: *Random sampling*

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

- Observed callback probabilities are a mixture of binomials:

$$Pr(C_j = c) = \int f(c|p_w, p_b) dG(p_w, p_b) \equiv \bar{f}(c)$$

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

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

- One reason for interest in $G(\cdot, \cdot)$ is that it characterizes prevalence and severity of discrimination in the population
- Fraction of jobs that are *not* discriminating:

$$\pi^0 = \int_0^1 dG(p, p)$$

- Second reason: tool for deciding which jobs are discriminating
- By Bayes' rule, fraction discriminating among jobs with callback configuration C_j is:

$$\Pr(D_j = 1 | C_j) = \int_{p_w \neq p_b} f(C_j | p_w, p_b) dG(p_w, p_b) \times \frac{(1 - \pi^0)}{\bar{f}(C_j)}$$

Indirect Evidence

$$\Pr(D_j = 1 | C_j) = \int_{p_w \neq p_b} f(C_j | p_w, p_b) dG(p_w, p_b) \times \frac{(1 - \pi^0)}{\bar{f}(C_j)}$$

$$\equiv \mathcal{P}\left(\underbrace{C_j}_{\text{direct}}, \underbrace{G(\cdot,\cdot)}_{\text{indirect}}\right).$$

- "Posterior" blends direct evidence from an employer's own behavior with indirect evidence from the population from which it was drawn
- Key parameter: π^0 serves the role of "prior" probability of innocence
- How best to use indirect evidence in decisionmaking?

Auditor's Problem

- Consider an auditor (e.g. the EEOC) who knows $G(\cdot,\cdot)$ and must decide which employers to investigate
- Decision rule $\delta(c) : \{0...L_w\} \times \{0...L_b\} \rightarrow \{0,1\}$ maps callbacks to a binary inquiry decision
- Loss function depends on number of type I and type II errors:

$$\mathcal{L}_J(\delta) = \sum_{j=1}^{J} \left\{ \underbrace{\delta\left(C_j\right)\left(1 - D_j\right)}_{\text{Type I}} \kappa + \underbrace{\left[1 - \delta\left(C_j\right)\right] D_j}_{\text{Type II}} \gamma \right\}.$$

- The D_j are unknown, so the auditor minimizes expected loss (i.e. risk), $\mathcal{R}_J(G,\delta) = E\left[\mathcal{L}_J(\delta)\right]$
- Reasonable doubt: investigate when $\mathcal{P}(C_j, G) > \kappa/(\kappa+\gamma)$ ▸ details
- N.B.: Posterior threshold rule controls False Discovery Rate (FDR), while classical hypothesis test does not (Benjamini and Hochberg, 1995; Storey, 2003) ▸ details

Moments

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

- It turns out that some features of $G(\cdot,\cdot)$ are nonparametrically identified
- Observed callback frequencies are given by

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

$$= \left(\begin{array}{c} L_w \\ c_w \end{array}\right)\left(\begin{array}{c} L_b \\ c_b \end{array}\right) \sum_{x=0}^{L_w-c_w} \sum_{s=0}^{L_b-c_b} (-1)^{x+s} \left(\begin{array}{c} L_w - c_w \\ x \end{array}\right)\left(\begin{array}{c} L_b - c_b \\ s \end{array}\right)$$

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

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

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

- Implies identification of all moments $\mu(m,n)$ with $m \leq L_w$, $n \leq L_b$.
- Example: $Var(p_{jw} - p_{jb})$ identified as long as $\min\{L_w, L_b\} \geq 2$.

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): Gender discrimination in Mexico
- Estimation: GMM, and "shape-constrained" GMM requiring moments to be consistent with a coherent probability distribution ▸ 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

Table I: Descriptive statistics for resume correspondence studies

	Bertrand & Mullainathan (1)	Nunley et al. (2)	Arceo-Gomez & Campos-Vasquez (3)
Number of jobs	1,112	2,305	802
Applications per job	4	4	8
Treatment/control	Black/white	Black/white	Male/female
Design	Stratified 2x2	Sample 4 names w/out replacement	Stratified 4x4
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 (0.007)	-0.026 (0.007)	-0.030 (0.008)

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

Table III: 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)
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.012
	(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 III: 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)	(0.004)
$E[(p_b - E[p_b])^2]$	0.023	0.023
	(0.004)	(0.003)
$E[(p_w - E[p_w])(p_b - E[p_b])]$	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)
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.012	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:	0.00
	P-value:	1.000
Sample size	1,112	

Substantial Variation in Discrimination

Table VI.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
	(0.011)	(0.011)	(0.012)
Correlation with p_w	0.927	1.000	-0.717
	(0.055)	-	(0.089)

First Two Moments in Nunley et al. Data

Table IV: Moments of callback rate distribution, Nunley et al. 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 Extra Moments

Table IV: Moments of callback rate distribution, Nunley et al. data

Moment	(2,2) design (1)	(3,1) design (2)	(1,3) design (3)
$E[p_w]$	0.174	0.199	0.142
	(0.010)	(0.025)	(0.015)
$E[p_b]$	0.148	0.149	0.157
	(0.010)	(0.015)	(0.013)
$E[(p_w - E[p_w])^2]$	0.089	0.108	-
	(0.007)	(0.009)	
$E[(p_b - E[p_b])^2]$	0.085	-	0.083
	(0.007)		(0.008)
$E[(p_w - E[p_w])(p_b - E[p_b])]$	0.083	0.084	0.080
	(0.006)	(0.009)	(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	0.043	-
	(0.004)	(0.007)	
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.047	-	0.045
	(0.005)		(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 Does Not Reject

more tests

Table IV: Moments of callback rate distribution, Nunley et al. 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	0.199	0.142	0.027	0.177
	(0.010)	(0.025)	(0.015)		(0.007)
$E[p_b]$	0.148	0.149	0.157	0.854	0.153
	(0.010)	(0.015)	(0.013)		(0.007)
$E[(p_w - E[p_w])^2]$	0.089	0.108	-	0.097	0.095
	(0.007)	(0.009)			(0.004)
$E[(p_b - E[p_b])^2]$	0.085	-	0.083	0.857	0.084
	(0.007)		(0.008)		(0.004)
$E[(p_w - E[p_w])(p_b - E[p_b])]$	0.083	0.084	0.080	0.926	0.084
	(0.006)	(0.009)	(0.009)		(0.004)
$E[(p_w - E[p_w])^3]$	-	0.051	-		0.106
		(0.008)			(0.006)
$E[(p_b - E[p_b])^3]$	-	-	0.044		0.092
			(0.007)		(0.006)
$E[(p_w - E[p_w])^2(p_b - E[p_b])]$	0.044	0.043	-	0.875	0.040
	(0.004)	(0.007)			(0.002)
$E[(p_w - E[p_w])(p_b - E[p_b])^2]$	0.047	-	0.045	0.819	0.042
	(0.005)		(0.007)		(0.002)
$E[(p_w - E[p_w])^3(p_b - E[p_b])]$	-	0.034	-	-	0.035
		(0.005)			(0.002)
$E[(p_w - E[p_w])(p_b - E[p_b])^3]$	-	-	0.037	-	0.037
			(0.006)		(0.002)
$E[(p_w - E[p_w])^2(p_b - E[p_b])^2]$	0.036	-	-	-	0.038
	(0.004)				(0.002)
				J-statistic:	23.09
				P-value:	0.190
Sample size	1,146	544	550		2,240

Treatment Effects Are Variable and Skewed

Table VI.B: Treatment effect variation in Nunley et al. (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)	(0.009)
Correlation with p_w	0.944	1.000	-0.336
	(0.018)	-	(0.048)
Skewness	3.757	3.648	-4.450
	(0.074)	(0.087)	(0.405)

Thick Tail of Extreme Discriminators in AGCV

Table VI.C: Treatment effect variation in AGCV

	p_m (1)	p_f (2)	p_m - p_f (3)
Mean	0.114	0.140	-0.025
	(0.009)	(0.009)	(0.008)
Standard deviation	0.231	0.257	0.179
	(0.011)	(0.010)	(0.011)
Correlation with p_f	0.735	1.000	-0.483
	(0.035)	-	(0.051)
Skewness	4.067	3.748	-1.403
	(0.140)	(1.161)	(0.385)
Excess kurtosis	8.452	5.756	12.227
	(1.458)	(8.790)	(2.291)

Posteriors

Bounds on Priors and Posteriors

- Moments of $G(\cdot, \cdot)$ aren't enough to compute posterior $\mathcal{P}(C_j, G)$
- Conservative approach: use what we know about $G(\cdot, \cdot)$ to bound prior π^0 and posterior $\mathcal{P}(C_j, G)$
- Upper bound on prior share innocent:

$$\bar{\pi}^0 = \max_{G \in \mathscr{G}} \int_0^1 dG(p, p) \quad s.t. \quad \bar{f} = B\mu_G$$

- Following Tebaldi et al. (2019), search over space $\mathscr{G}$ of discretized bivarate CDFs
- Objective and constraints are linear in p.m.f associated with $G(\cdot, \cdot) \implies$ apply linear programming ▸ details
- Same approach can be used to bound other notions of discrimination, e.g. share not discriminating against blacks: $\int_{p_b \geq p_w} dG(p_b, p_w)$.

In BM, At Most 87% of Jobs Are Innocent

Table VII: Upper bounds on shares not discriminating, BM data

	Share not discriminating: $\Pr(p_w = p_b)$ (1)
	$\boxed{0.870}$
J-statistic:	29.26
P-value (bound = 1):	0.000

At Most 56% Making Two Total Calls Are Innocent

Table VII: Upper bounds on shares not discriminating, BM data

Callbacks	Share not discriminating: $\Pr(p_w = p_b)$ (1)
All	0.870
0	0.962
1	0.576
2	$\boxed{0.558}$
3	0.492
4	0.788
J-statistic:	29.26
P-value (bound = 1):	0.000

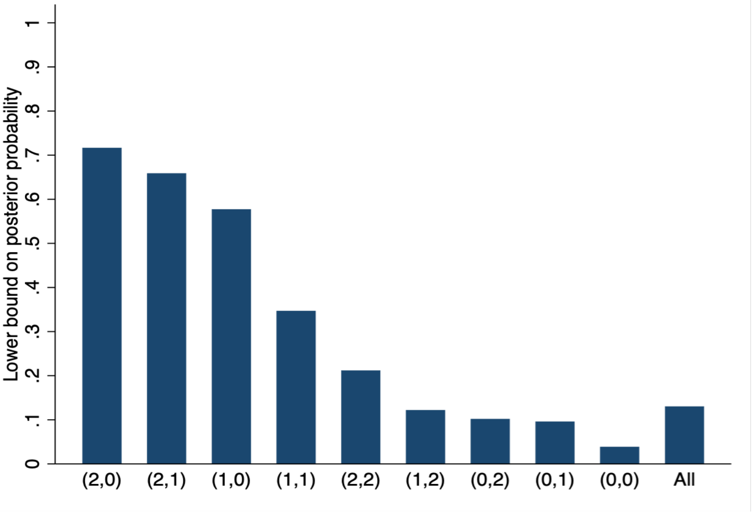
Cannot Reject Zero Discrimination Against Whites

Table VII: Upper bounds on shares not discriminating, BM data

Callbacks	Share not discriminating: $\Pr(p_w = p_b)$ (1)	Share not disc. against whites: $\Pr(p_w \geq p_b)$ (2)	Share not disc. against blacks: $\Pr(p_w \leq p_b)$ (3)
All	0.870	1.000	0.870
0	0.962	1.000	0.962
1	0.576	1.000	0.576
2	0.558	1.000	0.558
3	0.492	1.000	0.492
4	0.788	1.000	0.788
J-statistic:	29.26	0.00	29.26
P-value (bound = 1):	0.000	1.000	0.000

In BM, At Least 72% With $C_j = (2, 0)$ Discriminate

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



In Nunley et al., Cannot Reject $\Pr(p_{jw} \geq p_{jb}) = 1$

Table VIII: Upper bounds on shares not discriminating, Nunley et al. data

Design	Callbacks	Share not discriminating: $\Pr(p_w = p_b)$ (1)	Share not disc. against whites: $\Pr(p_w \geq p_b)$ (2)	Share not disc. against blacks: $\Pr(p_w \leq p_b)$ (3)
All	All	0.642	0.846	0.827
(2,2)	0	0.848	0.907	0.952
	1	0.328	0.815	0.567
	2	0.309	0.984	0.325
	3	0.179	0.933	0.264
	4	0.579	0.743	0.872
	J-statistic:	62.64	23.46	62.64
	P-value (bound = 1):	0.000	0.120	0.000

At Most 33% That Make Two Calls Have $p_{jw} \leq p_{jb}$

Table VIII: Upper bounds on shares not discriminating, Nunley et al. data

Design	Callbacks	Share not discriminating: $\Pr(p_w = p_b)$ (1)	Share not disc. against whites: $\Pr(p_w \geq p_b)$ (2)	Share not disc. against blacks: $\Pr(p_w \leq p_b)$ (3)
All	All	0.642	0.846	0.827
(2,2)	0	0.848	0.907	0.952
	1	0.328	0.815	0.567
	2	0.309	0.984	0.325
	3	0.179	0.933	0.264
	4	0.579	0.743	0.872
	J-statistic:	62.64	23.46	62.64
	P-value (bound = 1):	0.000	0.120	0.000

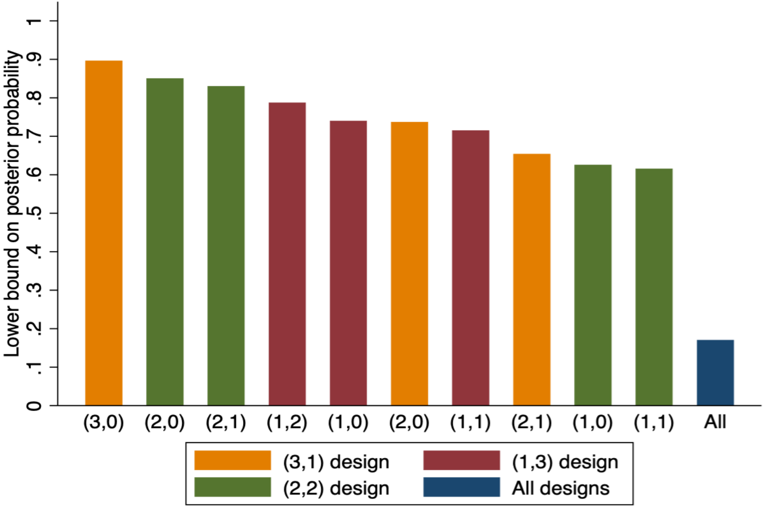
Informative Bounds In Other Designs and Callback Strata

Table VIII: Upper bounds on shares not discriminating, Nunley et al. data

Design	Callbacks	Share not discriminating: $\Pr(p_w = p_b)$ (2)	Share not disc. against whites: $\Pr(p_w \geq p_b)$ (3)	Share not disc. against blacks: $\Pr(p_w \leq p_b)$ (4)
All	All	0.642	0.846	0.827
(2,2)	0	0.848	0.907	0.952
	1	0.328	0.815	0.567
	2	0.309	0.984	0.325
	3	0.179	0.933	0.264
	4	0.579	0.743	0.872
(3,1)	0	0.853	0.898	0.964
	1	0.337	0.894	0.549
	2	0.332	0.998	0.336
	3	0.151	0.922	0.251
	4	0.566	0.767	0.837
(1,3)	0	0.839	0.916	0.936
	1	0.323	0.754	0.594
	2	0.326	0.958	0.369
	3	0.204	0.955	0.262
	4	0.581	0.723	0.893
	J-statistic:	62.64	23.46	62.64
	P-value (bound = 1):	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



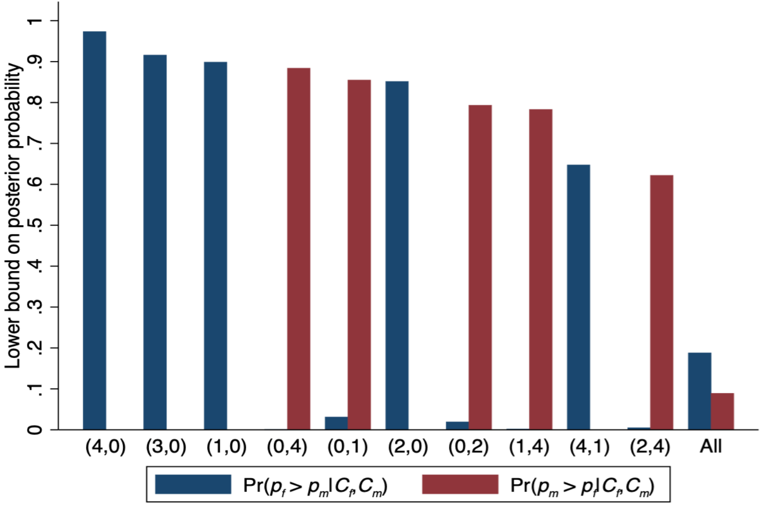
In AGCV, Discrimination in Both Directions

Table IX: Upper bounds on shares not discriminating, AGCV data

Callbacks	Share not discriminating: $\Pr(p_f = p_m)$ (1)	Share not disc. against women: $\Pr(p_f \geq p_m)$ (2)	Share not disc. against men: $\Pr(p_f \leq p_m)$ (3)
All	0.723	0.911	0.812
0	0.864	0.960	0.905
1	0.105	0.586	0.520
2	0.284	0.740	0.544
3	0.424	0.953	0.472
4	0.497	0.945	0.553
5	0.654	0.829	0.825
6	0.591	0.788	0.803
7	0.514	0.843	0.671
8	0.924	0.989	0.935
J-statistic:	369.66	33.88	359.95
P-value (bound = 1):	0.000	0.005	0.000

Lower Bounds on Posteriors Above 90%

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



Decisions

Decisions

- Consider auditor's decision problem under a particular parametric model for $G(\cdot, \cdot)$
- Detection/error tradeoff (DET) curve: Tradeoff between false accusation and successful detection for a fixed number of apps
- Build DET curves for three versions of Nunley et al. experiment:
  - Two black/two white, random covariates
  - Five black/five white, random covariates
  - Optimal 10-app combination of race/covariates

Parametric Model: Mixed Logit

- Logit model for callback to application ℓ at job j:

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

- $R_{j\ell}$ indicates race, $X_{j\ell}$ includes other randomly-assigned characteristics (GPA, experience, etc.)
- Normal/discrete type mixing distribution:

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

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

Discrimination is Rare But Intense

Table X: Mixed logit estimates, Nunley et al. data

			Types	
		Constant	No selection	Selection
		(1)	(2)	(3)
Distribution of logit(p_w):	α_0	-4.708	-4.931	-4.927
		(0.223)	(0.242)	(0.280)
	σ_α	4.745	4.988	4.983
		(0.223)	(0.249)	(0.294)
Discrimination intensity:	β_0	0.456	$\boxed{4.046}$	4.053
		(0.108)	$\boxed{(1.563)}$	(1.576)
Discrimination logit:	τ_0	-	-1.586	-1.556
			(0.416)	(1.098)
	τ_α	-	-	-0.005
				(0.180)
Fraction with $p_w \neq p_b$:		1.000	$\boxed{0.168}$	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

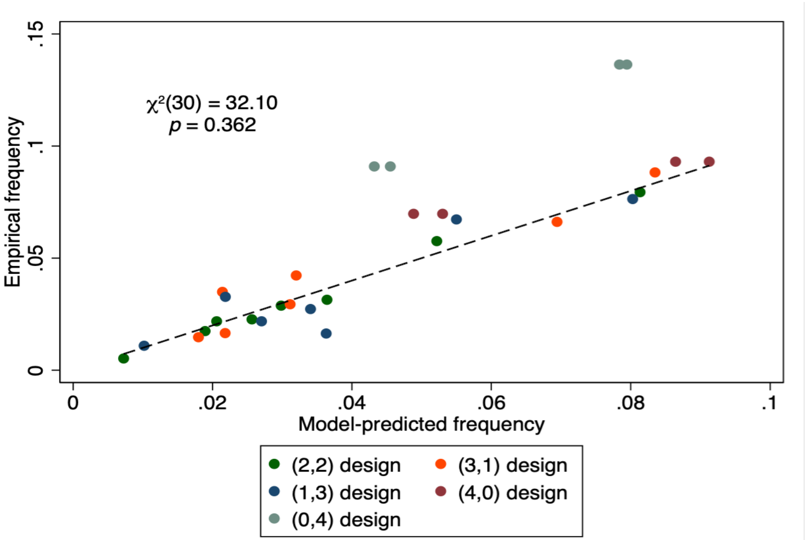
Discrimination is Not A "Luxury"

Table X: Mixed logit estimates, Nunley et al. data

			Types	
		Constant	No selection	Selection
		(1)	(2)	(3)
Distribution of logit(p_w):	α_0	-4.708	-4.931	-4.927
		(0.223)	(0.242)	(0.280)
	σ_α	4.745	4.988	4.983
		(0.223)	(0.249)	(0.294)
Discrimination intensity:	β_0	0.456	4.046	4.053
		(0.108)	(1.563)	(1.576)
Discrimination logit:	τ_0	-	-1.586	-1.556
			(0.416)	(1.098)
	τ_α	-	-	-0.005
				(0.180)
Fraction with $p_w \neq p_b$:		1.000	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

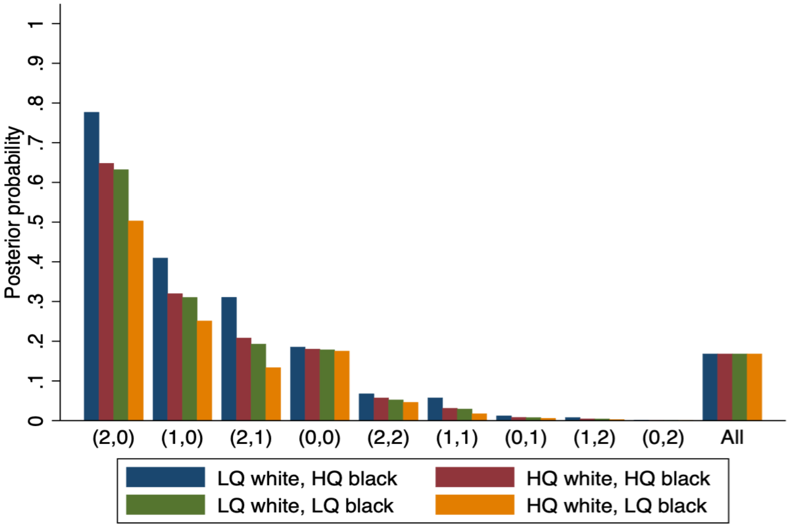
The Logit Model Fits Well

Figure IV: Mixed logit model fit



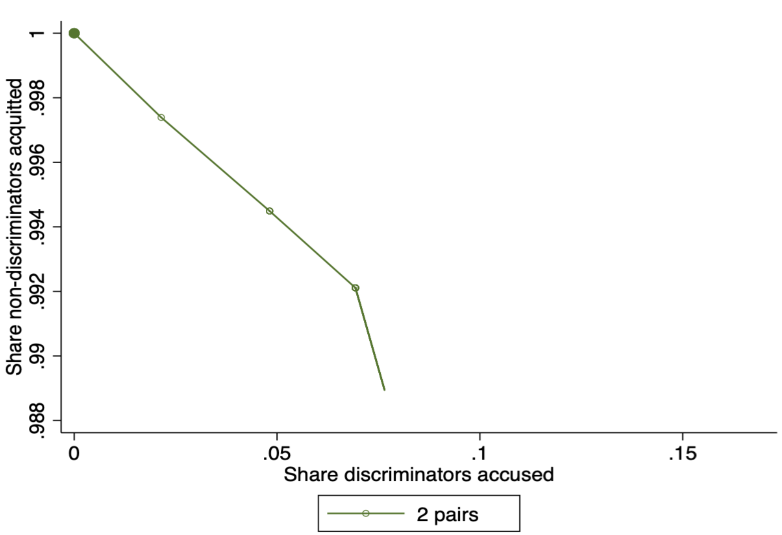
Covariates Generate Variation in Posteriors

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



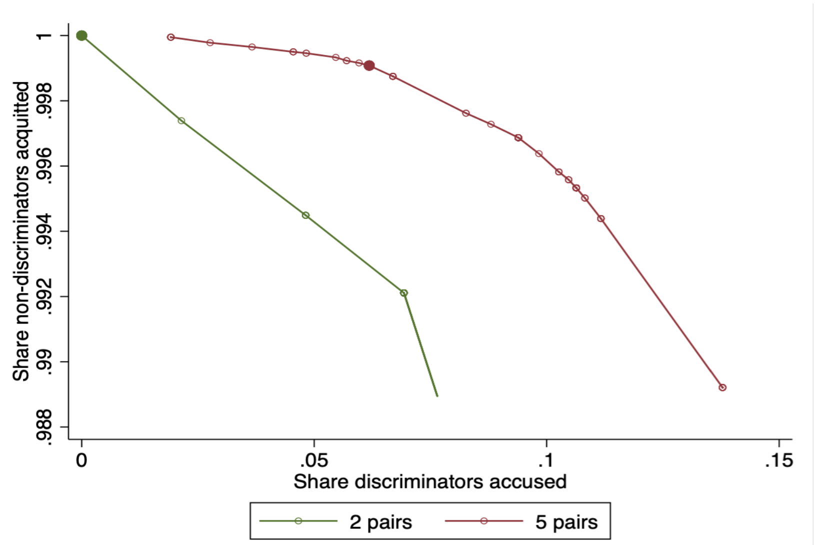
With 2 Pairs, 80% Threshold Yields Few Accusations

Figure VI: Detection/error tradeoffs, Nunley et al. data



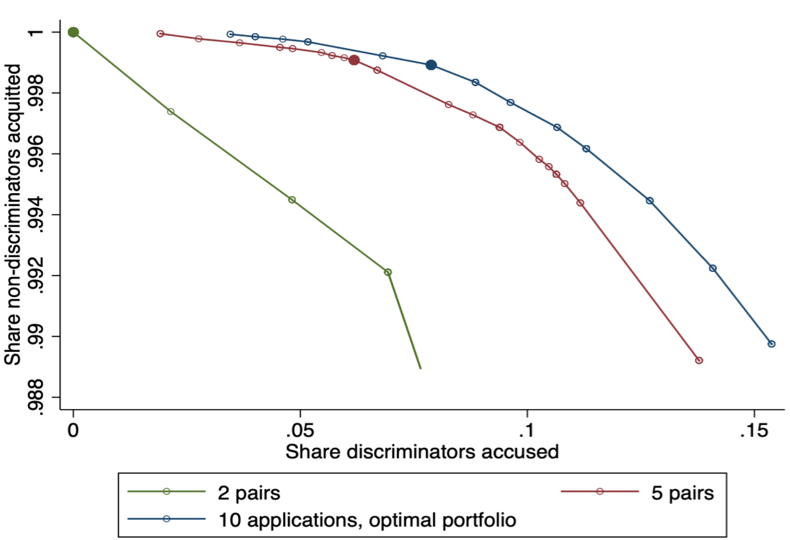
Sending 5 Pairs Boosts Detection Substantially

Figure VI: Detection/error tradeoffs, Nunley et al. data



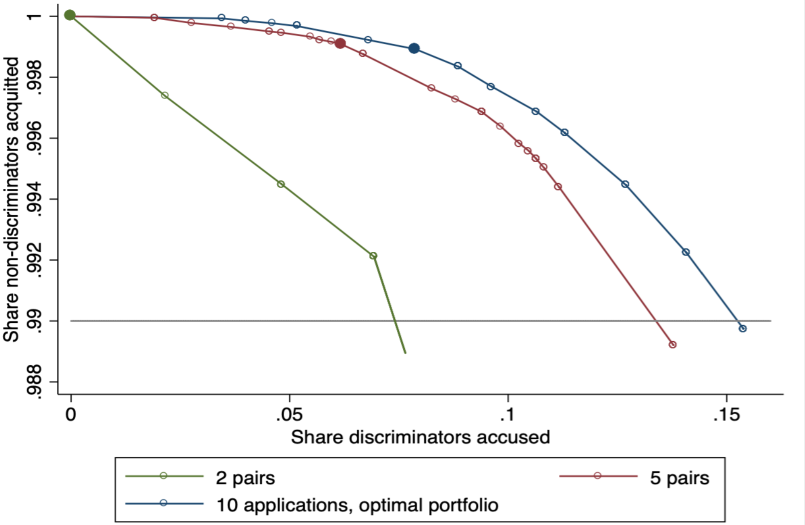
Optimizing Portfolio Yields Further Gains

Figure VI: Detection/error tradeoffs, Nunley et al. data



Fixing Size at 0.01 Yields More (Mostly False) Accusations

Figure VI: Detection/error tradeoffs, Nunley et al. data



Ambiguity

Auditing Under Ambiguity

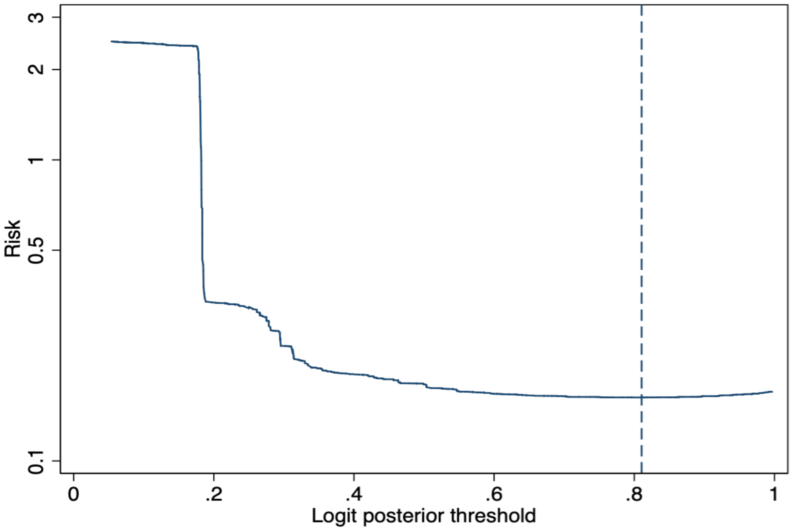
- How would decisions change if the auditor admits that $G(\cdot,\cdot)$ might not be logit?
- Important (extreme) benchmark for decisionmaking under ambiguity: minimax decision rule
- Minimax risk function and decision rule when auditor knows G lies in some identified set Θ:

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

- Think of δ^{mm} as an estimator of unobserved D_j's that "shrinks" towards a least favorable prior
- Contrast risk and decisions based upon mixed logit prior with minimax decisions details

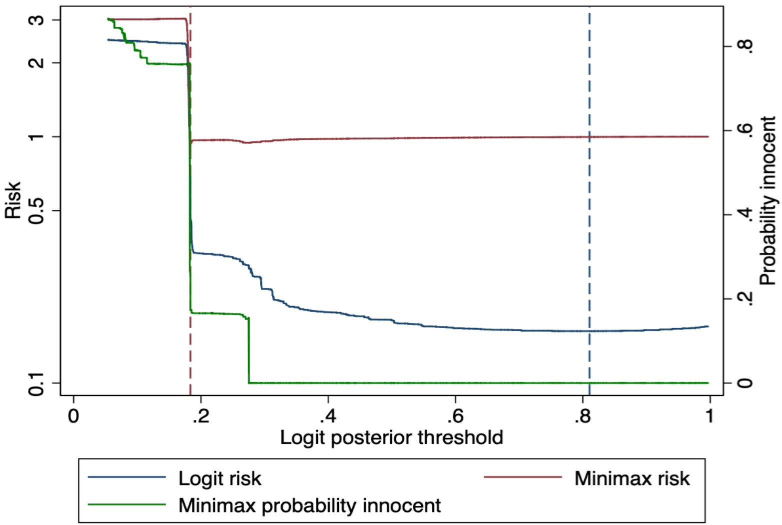
Logit Risk With $\kappa = 4, \ \gamma = 1$

Figure VII: Logit and minimax risk, Nunley et al. data



Minimax Decision Rule Is More Aggressive!

Figure VII: Logit and minimax risk, Nunley et al. data



Concluding Thoughts

- This paper develops and applies methods for detecting illegal discrimination by specific employers
- We find tremendous heterogeneity in discrimination – implies enforcement is a difficult inferential problem
- Nevertheless, favorable detection rates are achievable with relatively minor modifications to standard audit designs – suggests potential for real-time enforcement
- Methodological lessons:
  - Partial identification of response distribution does not preclude "borrowing strength" from the ensemble
  - Appropriate use of indirect evidence depends critically on investigator's loss function
- Question for future work: how do policy conclusions in other "empirical Bayes" evaluations of individual units (e.g. teachers, schools, hospitals, neighborhoods) vary with alternative notions of loss?

Bonus

Posterior Threshold Rule

- Risk $\mathcal{R}_J(G, \delta)$ can be rewritten

$$J \sum_{c_w=0}^{L_w} \sum_{c_b=0}^{L_b} \int \{\delta(c_w, c_b)(1 - \mathcal{P}(c_w, c_b, G))\kappa + [1 - \delta(c_w, c_b)]\mathcal{P}(c_w, c_b, G)\gamma\}$$

$$\times f(c_w, c_b | p_w, p_b) dG(p_w, p_b)$$

- Integrand is minimized by setting $\delta(c) = 0$ when $\mathcal{P}(c, G) \leq \frac{\kappa}{\kappa+\gamma}$ and $\delta(c) = 1$ otherwise

- Risk-minimizing decision rule is therefore

$$\delta(c) = 1\left\{\mathcal{P}(c, G) > \frac{\kappa}{\kappa+\gamma}\right\}.$$

back

$pFDR_J$ Control

- Let $N_J = \sum_{j=1}^{J} \delta(C_j)$ denote the total number of investigations
- Positive False Discovery Rate of Storey (2003) is defined:

$$pFDR_J = E\left[N_J^{-1} \sum_{j=1}^{J} \delta(C_j)(1 - D_j) | N_j \geq 1\right]$$

- Storey (2003) showed $pFDR_J = \Pr(D_j = 0 | \delta(C_j) = 1)$, so

$$pFDR_J = \Pr\left(D_J = 0 | \mathcal{P}(C_j, G) > \frac{\kappa}{\gamma+\kappa}\right)$$

$$\leq \Pr\left(D_j = 0 | \mathcal{P}(C_j, G) = \frac{\kappa}{\gamma+\kappa}\right) = \frac{\gamma}{\gamma+\kappa}.$$

- $\Pr(N_J \geq 1) \leq 1$, so posterior threshold rule also controls $FDR_J = pFDR_J \times Pr(N_J \geq 1)$.

back

Discretization of G

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

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

- $\{\pi_{kl}\}_{k=1,l=1}^{K,K}$ 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 π solves quadratic programming problem:

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

- M is a $dim(\mu) \times K^2$ matrix defined so that $M\pi = \mu$ for G_K
- Shape constrained moment estimates: $\hat{\mu} = M\hat{\pi}$
- 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:

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

- 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:

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

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

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

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

- 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

No Evidence That Callbacks Are Rival

Table II: Tests for dependence across trials

Nunley et al. data			AGCV data		
	Main effect	Leave-out mean		Main effect	Leave-out mean
Variable	(1)	(2)	Variable	(3)	(4)
Black	-0.028	-0.019	Married	0.001	0.002
	(0.010)	(0.027)		(0.008)	(0.033)
Female	0.010	0.009	Age	0.003	0.002
	(0.010)	(0.027)		(0.003)	(0.005)
High SES	-0.233	-0.674	Scholarship	-0.003	-0.060
	(0.174)	(0.522)		(0.010)	(0.050)
GPA	-0.043	-0.153	Predicted callback rate	-0.644	-0.136
	(0.066)	(0.198)		(0.504)	(0.888)
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	0.452		Joint p-value	0.589	
Sample size	9,220		Sample size	6,416	

back

Linear Programming

- Optimization problem for computing upper bound on share innocent:

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

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

$$\bar{f}\,(c_w, c_b) = \left(\begin{array}{c} L_w \\ c_w \end{array} \right) \left(\begin{array}{c} L_b \\ c_b \end{array} \right) \textstyle\sum_{k=1}^{K} \sum_{l=1}^{K} \pi_{kl}$$

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

- 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

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

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

- Maximal risk function for posterior cutoff q:

$$\mathcal{R}_J^m(q) = J \max_{\left\{\pi_{klk'l'}\right\}} \sum_{a \in \mathscr{A}_1} w_a$$

$$\times \left\{\Pr\left(\delta\left(C_j, a, q\right) = 1, D_j = 0\right)\kappa + \Pr\left(\delta\left(C_j, a, q\right) = 0, D_j = 1\right)\gamma\right\}$$

- $\mathscr{A}_1$ is list of possible quality configurations with corresponding probabilities w_a
- Constraints: $\pi_{klk'l'}$ 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 $\pi_{klk'l'}$ (see Appendix D)
- Set $K = 30$ when computing maximal risk in practice

back