Statistics II

Week 4: **Matching**

Content for Today

1. Assignment reminders
2. Any questions about the last assignment?
3. Review of core concepts from lecture
4. Matching in R
5. Any questions about this week’s assignment?

Assignment Reminders

Assignment Reminders

- Please take the peer reviews seriously. Remember, they may be double-blind for you, but not for the instructors ;-)
- Do not grade based on how someone codes. There are tons of ways to accomplish the same goal, and as long as the answer is correct, they deserve full points.
- If someone gives you a poor-quality review, mark the accuracy portion with a 1 or 2. This will flag it for the professors to review.
- Please keep knitting and submitting the html files.
- Always check how the html looks - determine if it is something you’re proud to submit.
- There may be times when your html file is too large for `ghclass` to handle (which can be partially avoided by not printing whole data sets, etc.) If it happens that you get something to review that does not have an html file, please just knit it locally on your machine.

Questions about the assignments?

Lecture Review

Randomization

Remember: Causal effects are identified if treatment is randomly assigned.

- **Unconditional randomization**: Randomly assign a fixed number of subjects from the study population to the treatment.
    - This can be inefficient if the outcome varies vastly for reasons unrelated to the treatment. For example, environmental factors may vary widely in treatment and control groups and have more of an effect than the treatment itself, making treatment effects challenging to identify.
- **Conditional randomization:** Choose relevant covariates pre-treatment (if they are knowable), build/stratify groups within particular combinations of these covariates, and then randomly assign treatment within these groups.
- **Paired randomization:** Two subjects per combination of covariate value, one of which is randomly assigned to treatment.
    - This is a good option when we have many variables and/or small sample sizes.

Exact Matching

1. Use theoretical and empirical knowledge to identify relevant confounder(s) X

2. Starting from treated subjects, select a match from the control group with exactly the same value(s) on X

3. Drop subjects off “common support” (unmatched subjects)

4. Estimate causal effect as the average difference in Y across pairs of matched subjects:

$$E(Y \mid D = 1, X) - E(Y \mid D = 0, X)$$

Non-exact Matching

Exact matching can lead to a very small final sample size, especially if there are lots of covariates to consider.

Instead, we can match based on **propensity scores:** a measure of the probability of a unit of being in the treatment group. Propensity scores are based on the idea that we want to match observations on treated units, and therefore want to find control observations that *look* as if they are treated.

> In other words, based on the characteristics of this unit, what is the likelihood that it has been treated?

This is usually modeled with logit/probit regression.

Balance Tests

Difference-in-means test (t-test) of our pretreatment variables in *X* for the treatment and control groups

> Reminder: Propensity score matching is done using the probability of a unit of being in the treatment group based on a set of variables *X*. Our balance tests are performed on these set of *X*.

Capitalize on the propensity score tautology: If treatment probability is identically distributed in treatment and control group, all *X* will be balanced.

- If *X* are balanced, then estimated propensity scores are ok
- If *X* are not balanced, re-specify treatment probability model

A balance table shows difference between unmatched and matched groups on various *X* variables.

Logistic Regression

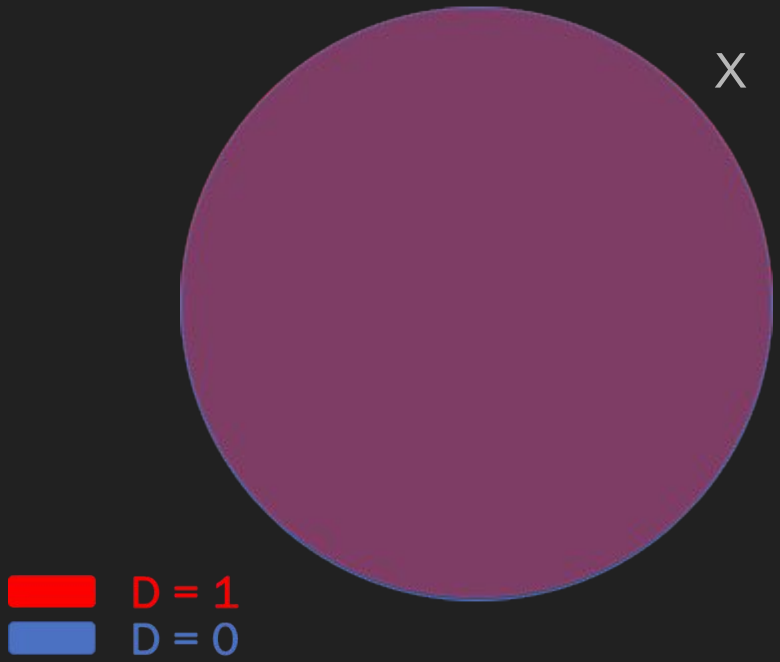
Similar to linear regression, except we’re working with a binary categorical outcome variable.

Instead of fitting a line to the data, logistic regression fits an S-shaped curve (sigmoid) that goes from 0 to 1; It tells you the probability of outcome based on the covariate(s) - this is our propensity score!

Coefficients are presented in terms of log(odds), so be careful with interpretation: it is not as straightforward as with linear regression.

Common Support

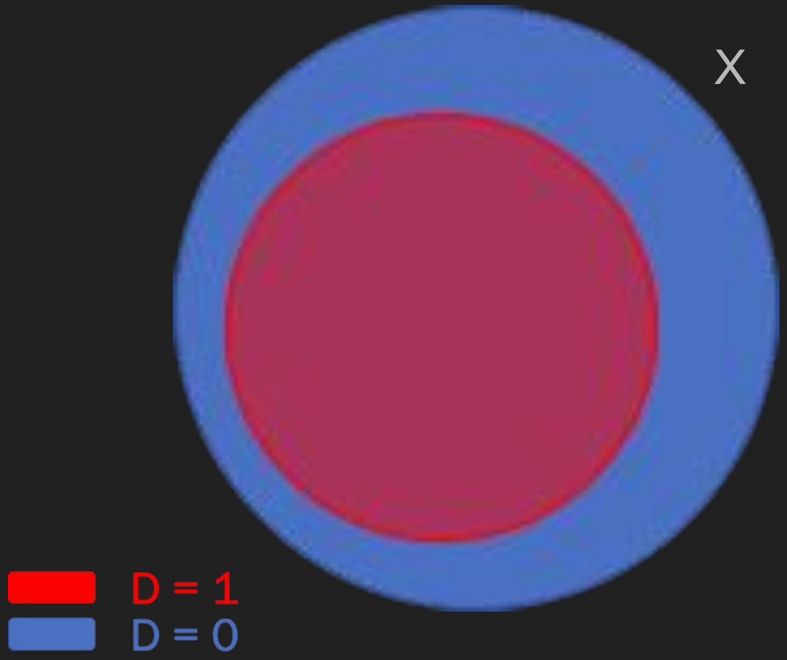
ATE



Name	Y	D	X
Jake	10	1	3
Gina	8	1	2
Terry	6	1	1
Rosa	8	0	3
Charles	6	0	2
Ray	4	0	1

In this case, we have full common support. Meaning that the distributions of X under both treatment and control are equal. We can match and gather the ATE.

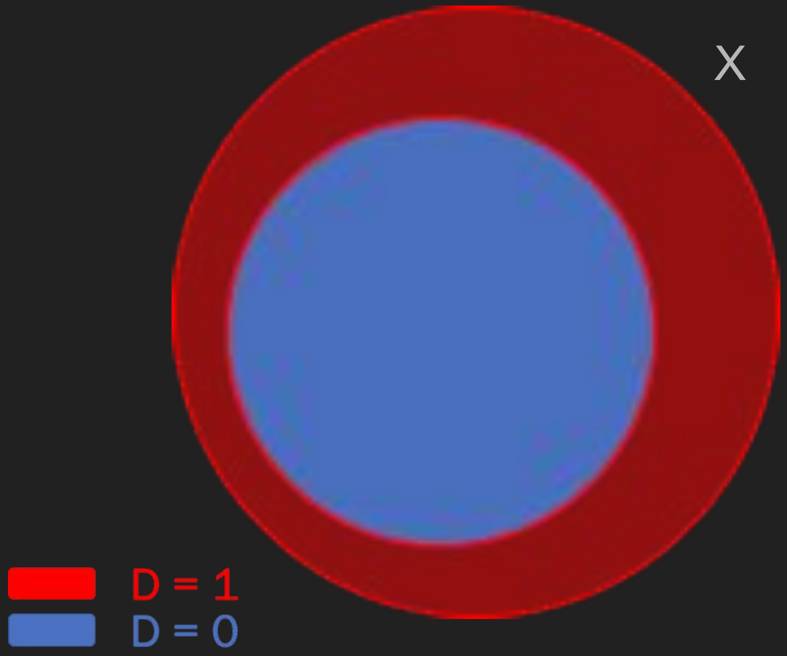
ATT



Name	Y	D	X
Jake	10	1	3
Gina	12	1	3
Terry	8	1	2
Rosa	6	0	3
Charles	3	0	2
Ray	1	0	1

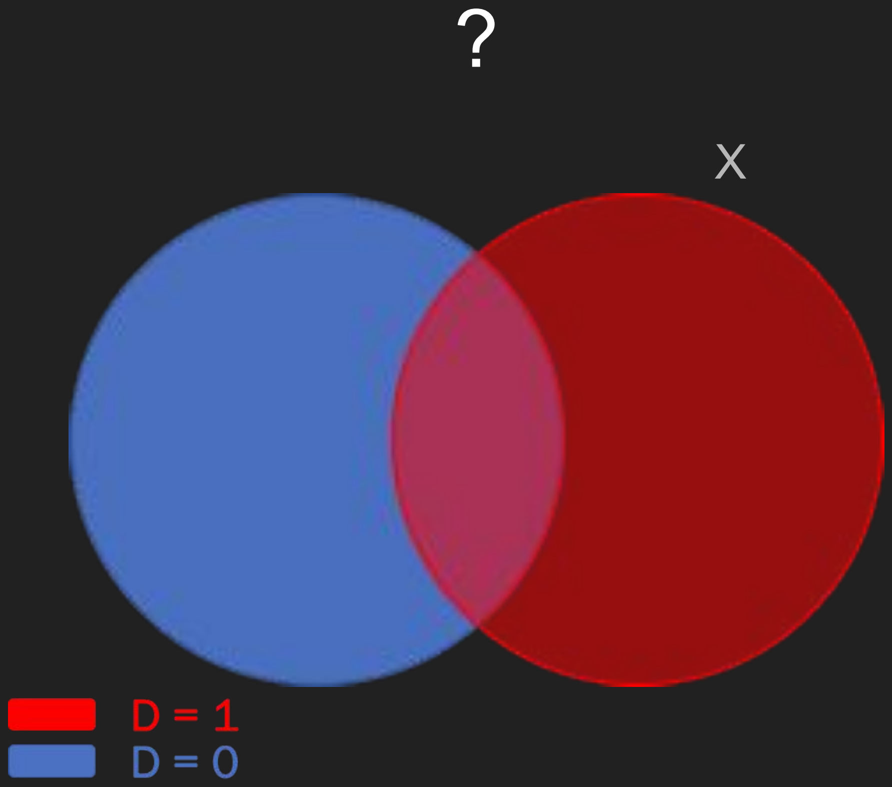
In this case, all our treated units have common support, but not our controls. We can match and gather the Average Treatment Effect for the Treated (ATT).

ATC



Name	Y	D	X
Jake	15	1	3
Gina	10	1	2
Terry	5	1	1
Rosa	10	0	3
Charles	6	0	2
Ray	4	0	2

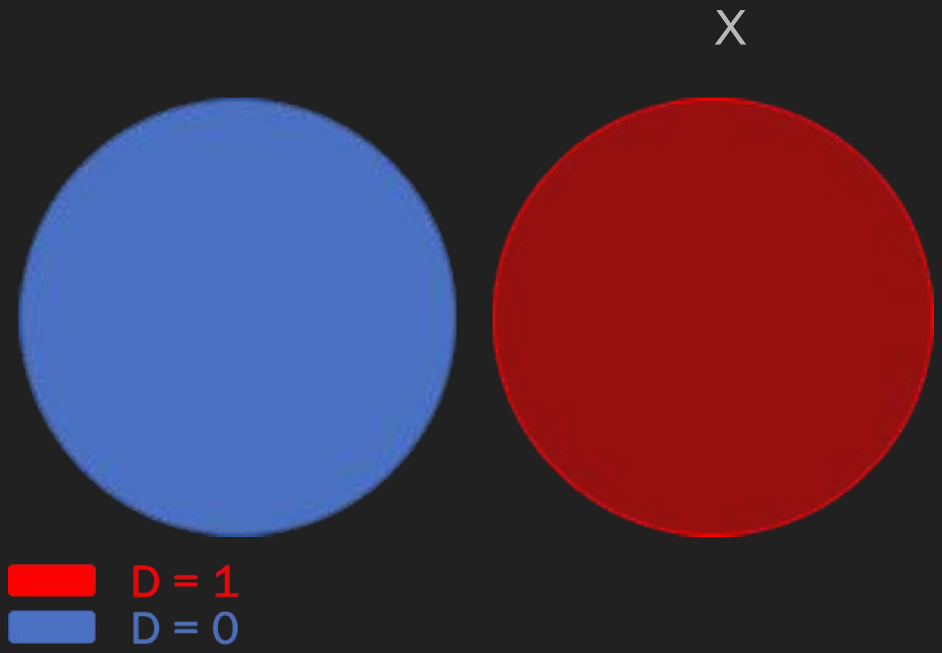
In this case, all our control units have common support, but not our treated. We can match and gather the **Average Treatment Effect for the Controls (ATC)**.



Name	Y	D	X
Jake	15	1	6
Gina	10	1	4
Terry	5	1	1
Rosa	10	0	1
Charles	6	0	2
Ray	4	0	3

In this case, only some of our control and treated units have common support. The results gathered would render a local estimate.

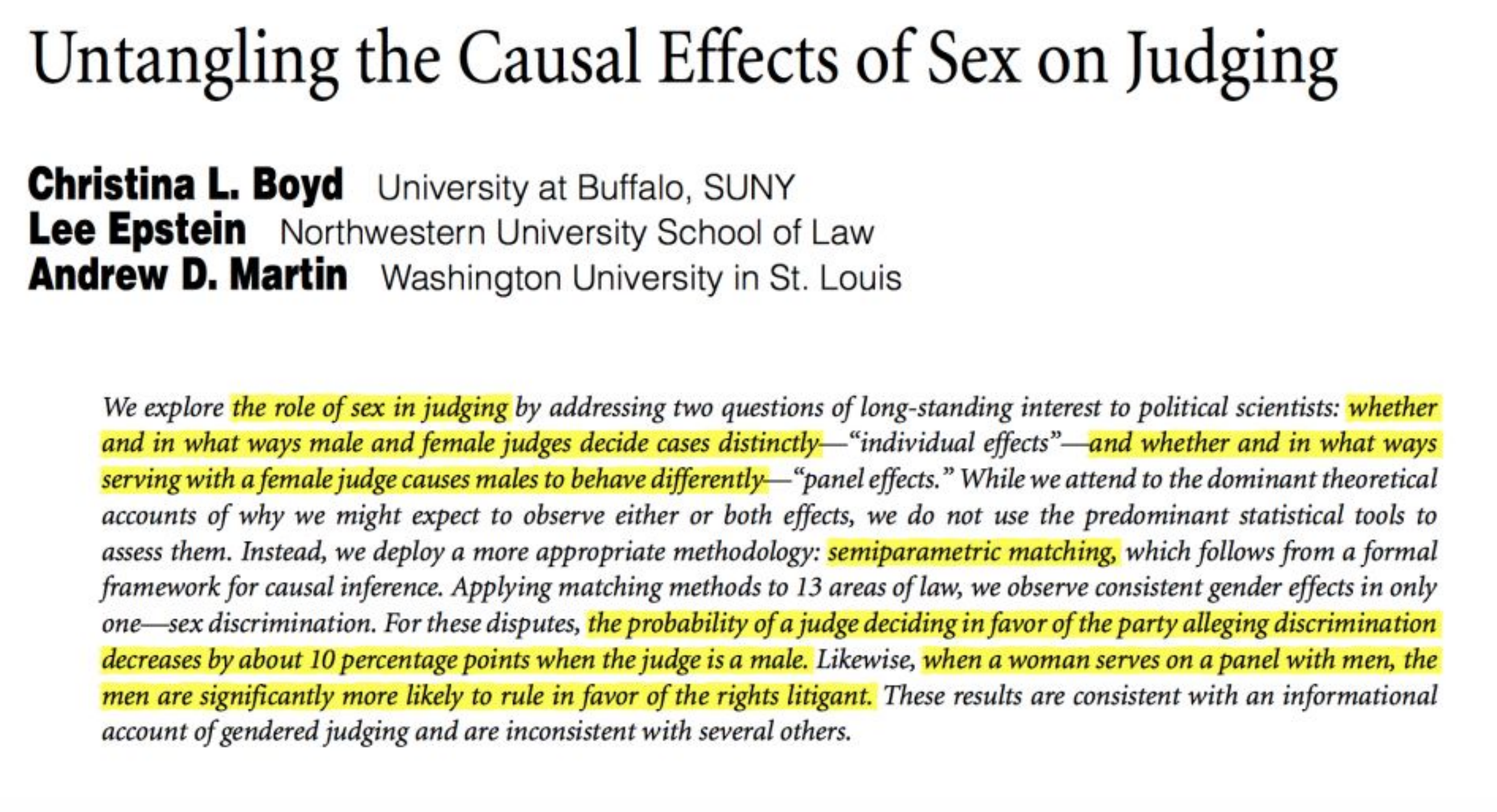
No common support



Name	Y	D	X
Jake	15	1	6
Gina	10	1	5
Terry	5	1	4
Rosa	10	0	1
Charles	6	0	2
Ray	4	0	3

In this case, none of our control and treated units have common support. We could not proceed with matching since our units are non-comparable in their levels of X.

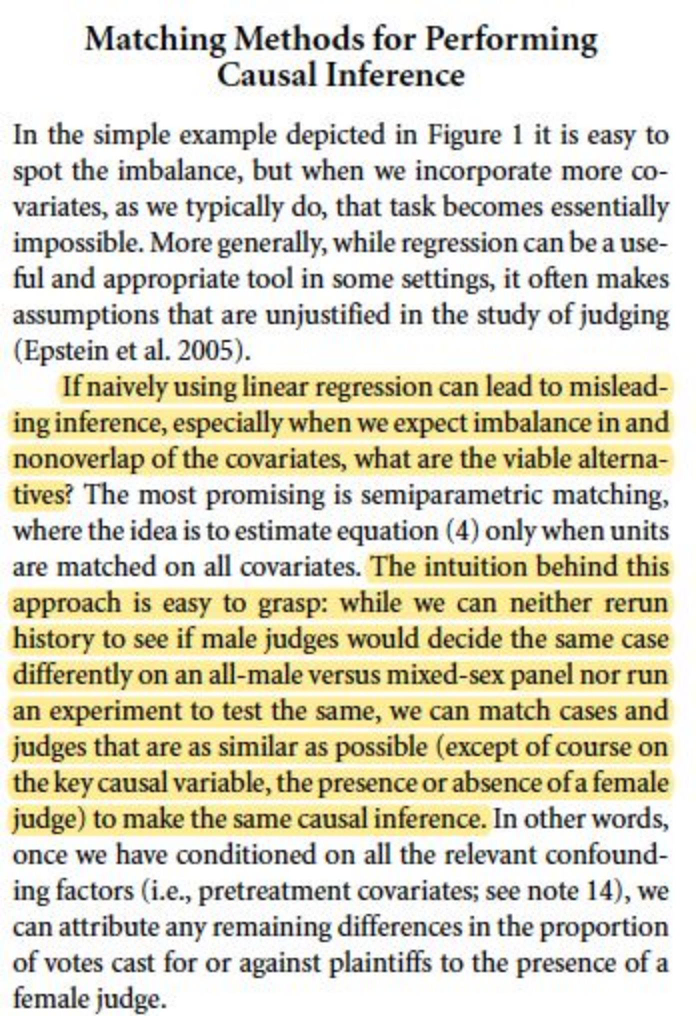
Application Paper



Untangling the Causal Effects of Sex on Judging

Christina L. Boyd University at Buffalo, SUNY
Lee Epstein Northwestern University School of Law
Andrew D. Martin Washington University in St. Louis

We explore the role of sex in judging by addressing two questions of long-standing interest to political scientists: whether and in what ways male and female judges decide cases distinctly—"individual effects"—and whether and in what ways serving with a female judge causes males to behave differently—"panel effects." While we attend to the dominant theoretical accounts of why we might expect to observe either or both effects, we do not use the predominant statistical tools to assess them. Instead, we deploy a more appropriate methodology: semiparametric matching, which follows from a formal framework for causal inference. Applying matching methods to 13 areas of law, we observe consistent gender effects in only one—sex discrimination. For these disputes, the probability of a judge deciding in favor of the party alleging discrimination decreases by about 10 percentage points when the judge is a male. Likewise, when a woman serves on a panel with men, the men are significantly more likely to rule in favor of the rights litigant. These results are consistent with an informational account of gendered judging and are inconsistent with several others.

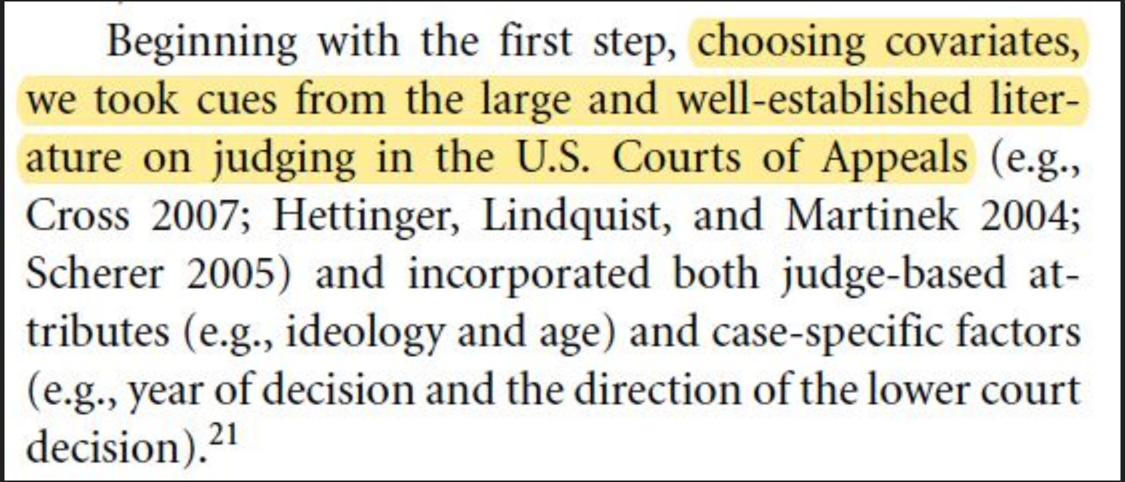


Matching Methods for Performing Causal Inference

In the simple example depicted in Figure 1 it is easy to spot the imbalance, but when we incorporate more covariates, as we typically do, that task becomes essentially impossible. More generally, while regression can be a useful and appropriate tool in some settings, it often makes assumptions that are unjustified in the study of judging (Epstein et al. 2005).

If naively using linear regression can lead to misleading inference, especially when we expect imbalance in and nonoverlap of the covariates, what are the viable alternatives? The most promising is semiparametric matching, where the idea is to estimate equation (4) only when units are matched on all covariates. The intuition behind this approach is easy to grasp: while we can neither rerun history to see if male judges would decide the same case differently on an all-male versus mixed-sex panel nor run an experiment to test the same, we can match cases and judges that are as similar as possible (except of course on the key causal variable, the presence or absence of a female judge) to make the same causal inference. In other words, once we have conditioned on all the relevant confounding factors (i.e., pretreatment covariates; see note 14), we can attribute any remaining differences in the proportion of votes cast for or against plaintiffs to the presence of a female judge.

Intuition behind matching



Beginning with the first step, choosing covariates, we took cues from the large and well-established literature on judging in the U.S. Courts of Appeals (e.g., Cross 2007; Hettinger, Lindquist, and Martinek 2004; Scherer 2005) and incorporated both judge-based attributes (e.g., ideology and age) and case-specific factors (e.g., year of decision and the direction of the lower court decision).[21]

What to feed the model?

How did they approach it?

“For this task, we used **“nearest-neighbor”** matching **with replacement**; that is, for each “mixed-sex” observation (or female judge, for the individual analysis), the “all-male” observation (or male judge) that has the closest propensity score is selected. We implemented this approach by matching observations from the control group (e.g., male judges on all-male panels) multiple times (“with replacement”).

What do these plots tell us?

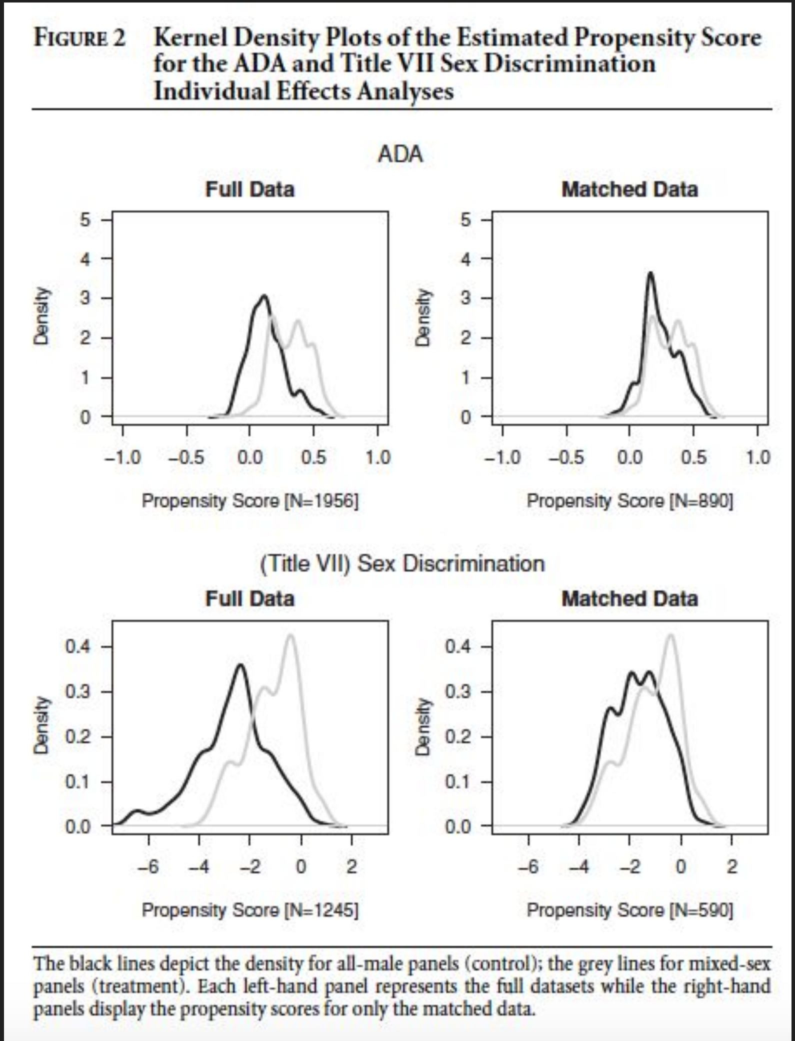


FIGURE 2 **Kernel Density Plots of the Estimated Propensity Score for the ADA and Title VII Sex Discrimination Individual Effects Analyses**

The black lines depict the density for all-male panels (control); the grey lines for mixed-sex panels (treatment). Each left-hand panel represents the full datasets while the right-hand panels display the propensity scores for only the matched data.

Propensity score distribution before and after the match.

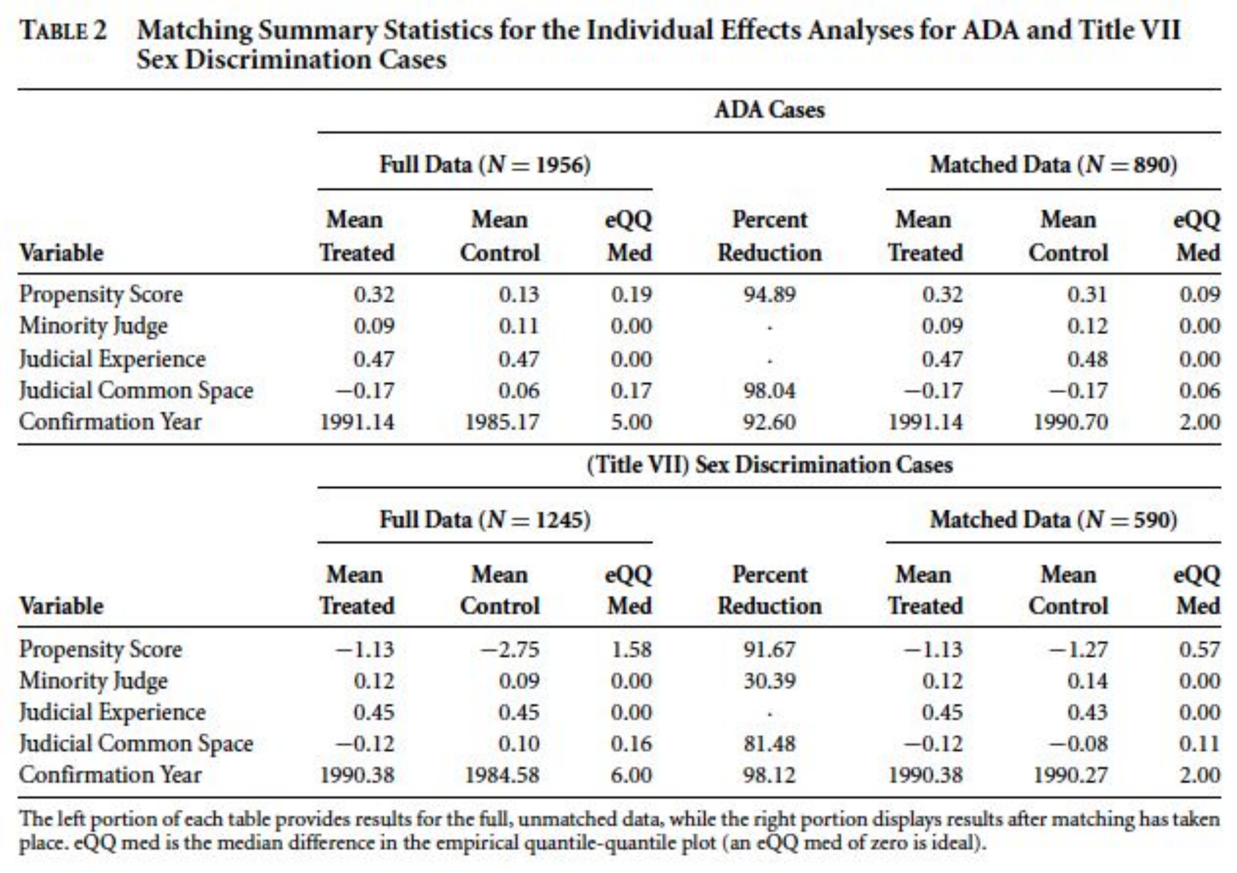


TABLE 2 Matching Summary Statistics for the Individual Effects Analyses for ADA and Title VII Sex Discrimination Cases

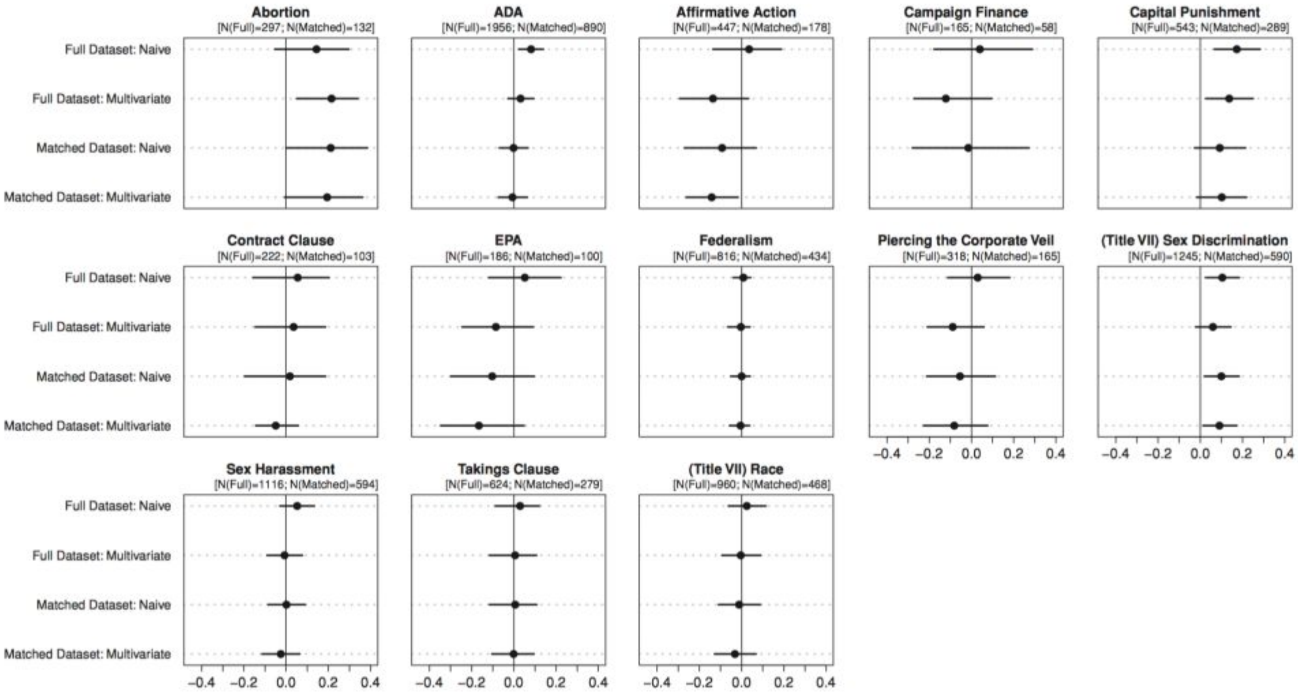
	ADA Cases						
	Full Data ($N = 1956$)				Matched Data ($N = 890$)		
Variable	Mean Treated	Mean Control	eQQ Med	Percent Reduction	Mean Treated	Mean Control	eQQ Med
Propensity Score	0.32	0.13	0.19	94.89	0.32	0.31	0.09
Minority Judge	0.09	0.11	0.00	.	0.09	0.12	0.00
Judicial Experience	0.47	0.47	0.00	.	0.47	0.48	0.00
Judicial Common Space	−0.17	0.06	0.17	98.04	−0.17	−0.17	0.06
Confirmation Year	1991.14	1985.17	5.00	92.60	1991.14	1990.70	2.00

	(Title VII) Sex Discrimination Cases						
	Full Data ($N = 1245$)				Matched Data ($N = 590$)		
Variable	Mean Treated	Mean Control	eQQ Med	Percent Reduction	Mean Treated	Mean Control	eQQ Med
Propensity Score	−1.13	−2.75	1.58	91.67	−1.13	−1.27	0.57
Minority Judge	0.12	0.09	0.00	30.39	0.12	0.14	0.00
Judicial Experience	0.45	0.45	0.00	.	0.45	0.43	0.00
Judicial Common Space	−0.12	0.10	0.16	81.48	−0.12	−0.08	0.11
Confirmation Year	1990.38	1984.58	6.00	98.12	1990.38	1990.27	2.00

The left portion of each table provides results for the full, unmatched data, while the right portion displays results after matching has taken place. eQQ med is the median difference in the empirical quantile-quantile plot (an eQQ med of zero is ideal).

Balance statistics before and after the match.

FIGURE 4 Dotplots of Average Treatment Effects (ATEs) for Individual Effects Across 13 Issue Areas



The lines represent 95% confidence intervals for the average treatment effect. For every issue area, the first two models are logistic regression models fit to each full, unbalanced dataset. The naive model includes only the judge's sex as a covariate. The other model includes the judge's sex and a number of controls, including ideology. The next two models show the ATE after nearest-neighbor matching with replacement on the estimated propensity score. The first is for a difference of proportions analysis. The second is for a logistic regression model with the judge's sex and a number of controls including ideology.

Dotplot of effect coefficients

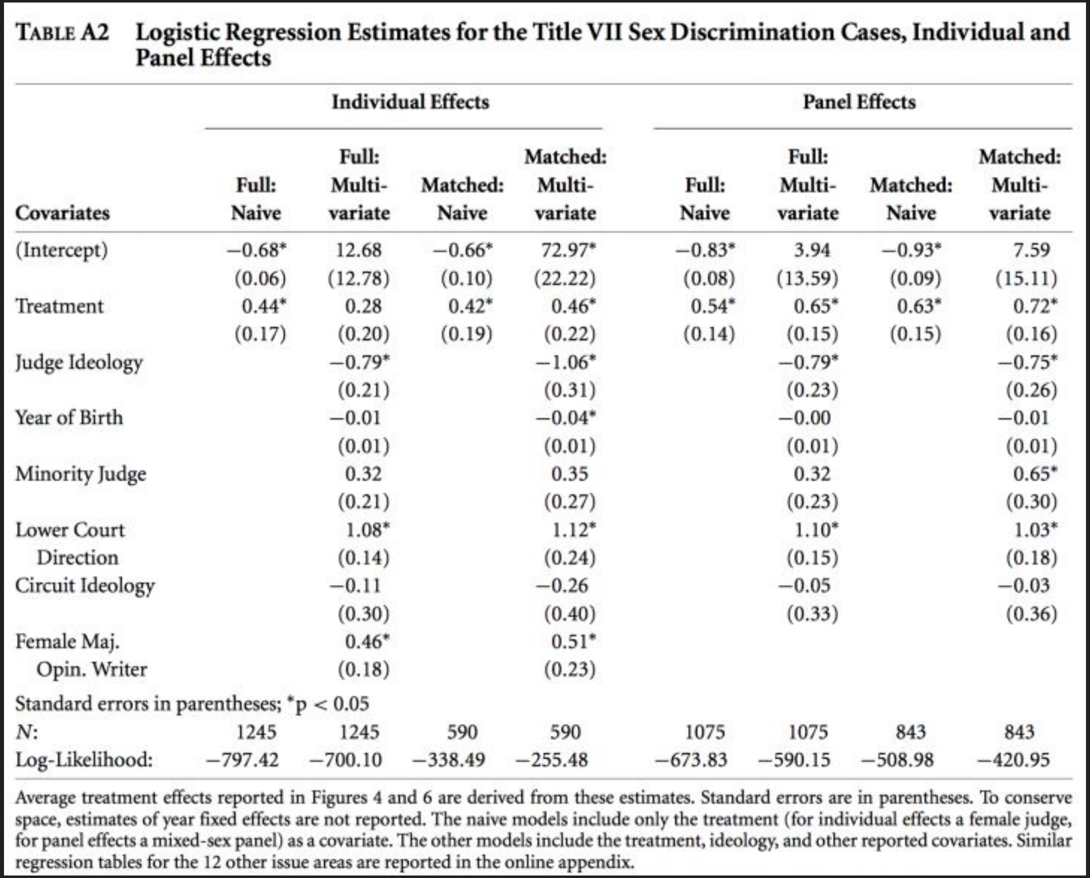


Table A2 Logistic Regression Estimates for the Title VII Sex Discrimination Cases, Individual and Panel Effects

	Individual Effects				Panel Effects			
Covariates	Full: Naive	Full: Multivariate	Matched: Naive	Matched: Multivariate	Full: Naive	Full: Multivariate	Matched: Naive	Matched: Multivariate
(Intercept)	−0.68*	12.68	−0.66*	72.97*	−0.83*	3.94	−0.93*	7.59
	(0.06)	(12.78)	(0.10)	(22.22)	(0.08)	(13.59)	(0.09)	(15.11)
Treatment	0.44*	0.28	0.42*	0.46*	0.54*	0.65*	0.63*	0.72*
	(0.17)	(0.20)	(0.19)	(0.22)	(0.14)	(0.15)	(0.15)	(0.16)
Judge Ideology		−0.79*		−1.06*		−0.79*		−0.75*
		(0.21)		(0.31)		(0.23)		(0.26)
Year of Birth		−0.01		−0.04*		−0.00		−0.01
		(0.01)		(0.01)		(0.01)		(0.01)
Minority Judge		0.32		0.35		0.32		0.65*
		(0.21)		(0.27)		(0.23)		(0.30)
Lower Court Direction		1.08*		1.12*		1.10*		1.03*
		(0.14)		(0.24)		(0.15)		(0.18)
Circuit Ideology		−0.11		−0.26		−0.05		−0.03
		(0.30)		(0.40)		(0.33)		(0.36)
Female Maj. Opin. Writer		0.46*		0.51*				
		(0.18)		(0.23)				
Standard errors in parentheses; *p < 0.05								
N:	1245	1245	590	590	1075	1075	843	843
Log-Likelihood:	−797.42	−700.10	−338.49	−255.48	−673.83	−590.15	−508.98	−420.95

Average treatment effects reported in Figures 4 and 6 are derived from these estimates. Standard errors are in parentheses. To conserve space, estimates of year fixed effects are not reported. The naive models include only the treatment (for individual effects a female judge, for panel effects a mixed-sex panel) as a covariate. The other models include the treatment, ideology, and other reported covariates. Similar regression tables for the 12 other issue areas are reported in the online appendix.

Regression output of effects

Questions?