Part B: Selection on Observables

B2: Propensity Score Methods

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2024

Outline

1. Propensity score methods
2. Applications

Propensity score results (Rosenbaum-Rubin 1983)

- Consider binary D. Recall $p(X) \equiv P(D = 1 \mid X)$
- **Proposition 1**: $D \perp\!\!\!\perp X \mid p(X)$
  - i.e. propensity score balances X between treated and control groups
  - Proof: $P(D = 1 \mid X, p(X)) = P(D = 1 \mid X) = p(X)$
- **Proposition 2:** $D \perp\!\!\!\perp (Y_0, Y_1) \mid X \implies D \perp\!\!\!\perp (Y_0, Y_1) \mid p(X)$
  - i.e., under CIA, controlling for scalar $p(X)$ is enough
    - ⋆ A stronger version of the OVB idea
  - Proof: $P\left(D = 1 \mid p(X), Y_0, Y_1\right) = \mathbb{E}\left[\mathbb{E}\left[D \mid p(X), X, Y_0, Y_1\right] \mid p(X), Y_0, Y_1\right] = p(X)$, doesn't depend on (Y_0, Y_1)

P-score methods: Steps

1. Obtain $p(X)$
   - Known in stratified RCTs
   - Parametric estimation, e.g. logit of D on X
   - Non-parametric regression of D on X
2. Assess overlap
   - Compare p-score distributions in treated & control groups
3. Verify balance
   - Within bins of $\hat{p}(X)$ compare X among treated and controls
   - If balance fails (with sufficiently many bins), make the p-score model richer
4. Adjust for p-score differences between treated and control
   - Regression, matching, blocking, reweighting

P-score adjustment methods: Regression

- With constant effects, enough to control linearly

$$Y_i = \beta D_i + \gamma p(X_i) + \text{error}$$

  - *Exercise*: Why?
  - *Exercise*: if $p(X)$ is estimated from a linear probability model, both ways are numerically the same as linearly controlling for X
    - Regression and pscore methods are related

- With heterogeneous effects, this yields the variance-weighted average of effects (*Exercise*: Why?)

P-score adjustment methods: Blocking/Matching

- **Matching:** For each treated obs., find the untreated one with the closest $p(X_i)$
  - Discard untreated observations with p-score outside the range for the treated
- **Blocking** (stratifying)**:**
  - Split data into bins of $p(X_i)$
  - Estimate difference-in-means within bins
  - Average across bins weighting by # obs. (ATE) or # treated obs. (ATT)

P-score adjustment methods: Reweighting (IPW)

- In the bin with $p(X_i) = \pi$ we have fraction π of observed $Y_i(1)$ (for treated) and fraction $1 - \pi$ of comparable but missing $Y_i(1)$ (for controls)
- Horvitz-Thompson (1952) "inverse probability weighting" (IPW): reweighting by $1/\pi$ makes the sample of $Y_i(1)$ representative

$$\mathbb{E}\left[\frac{YD}{p(X)}\right] = \mathbb{E}\left[\frac{Y_1 D}{p(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{Y_1 D}{p(X)} \mid X\right]\right] = \mathbb{E}\left[Y_1\right]$$

- Similarly for Y_0: $\mathbb{E}\left[\frac{Y(1-D)}{1-p(X)}\right] = \mathbb{E}\left[Y_0\right]$. Thus, under CIA+overlap:

$$ATE = \mathbb{E}\left[\left(\frac{D}{p(X)} - \frac{1-D}{1-p(X)}\right) Y\right] = \mathbb{E}\left[\frac{D - p(X)}{p(X)(1-p(X))} \cdot Y\right]$$

- *Exercise:* derive the reweighting expression for ATT

P-score adjustment methods: Reweighting (2)

- Plug-in (Horvitz-Thompson) estimator: $\widehat{ATE}_{HT} = \frac{1}{N}\sum_i \left(\frac{D_i}{\hat{p}(X_i)} - \frac{1-D_i}{1-\hat{p}(X_i)}\right) Y_i$
- Issue: weights on treated ($\frac{1}{N}\frac{D_i}{\hat{p}(X_i)}$) and control ($\frac{1}{N}\frac{1-D_i}{1-\hat{p}(X_i)}$) do not exactly sum to 1
  - E.g. adding a constant to Y_i changes the estimator
- Normalizing the weights improves performance: the Hajek estimator

$$\widehat{ATE}_{Hajek} = \frac{\sum_i \frac{D_i Y_i}{\hat{p}(X_i)}}{\sum_i \frac{D_i}{\hat{p}(X_i)}} - \frac{\sum_i \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}}{\sum_i \frac{1-D_i}{1-\hat{p}(X_i)}}$$

- Convenient implementation: regression of Y_i on D_i with weights $\frac{D_i}{\hat{p}(X_i)} + \frac{1-D_i}{1-\hat{p}(X_i)}$ (and no covariates)

P-score methods: Warnings

- P-score methods remove selection bias but are generally inefficient
  - E.g. in an RCT controlling for X improves efficiency, while $p(X) = \mathit{const}$
  - Although reweighting by a non-parametrically estimated p-score is efficient with fixed $\dim(X)$ (Hirano, Imbens, Ridder 2003)
    - Estimation is necessary even if true $p(X)$ is known (as in an RCT)
  - King and Nielsen (2019) similarly prefer matching on X to pscore matching
- Finite-sample properties may be poor because of dividing by $\hat{p}(X_i)$ and $1 - \hat{p}(X_i)$
  - "Balancing" approaches estimate the inverse p-score directly (cf. Ben-Michael, Feller, Hirschberg, Zubizarreta 2021)

Inference

These estimators first estimate the p-score and then use it $\Longrightarrow$ inference is difficult

- For p-score reweighting, see Hirano-Imbens-Ridder (2003)
- For p-score matching, see Abadie-Imbens (2016)
- Bootstrap may work (except matching) but this have not been proved
- Or just ignore the error from the p-score estimation
    - This tends to *overestimate* the SE

NSW application

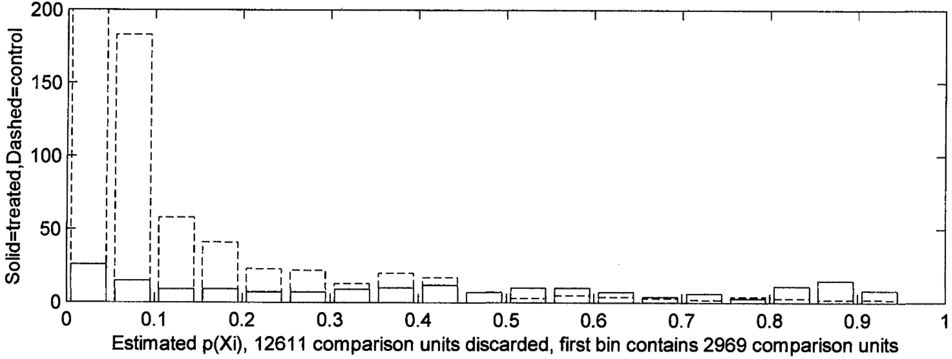
- Dehejia and Wahba (1999, 2002) show p-score methods, especially p-score matching, perform much better than Lalonde's regression controls
- They search among logit specifications until they achieve balance

Table 4. Sample Means of Characteristics for Matched Control Samples

Matched samples	*No. of observations*	*Age*	*Education*	*Black*	*Hispanic*	*No degree*	*Married*	*RE74 (U.S. $)*	*RE75 (U.S. $)*
NSW	185	25.81	10.35	.84	.06	.71	.19	2,096	1,532
MPSID-1	56	26.39	10.62	.86	.02	.55	.15	1,794	1,126
		[2.56]	[.63]	[.13]	[.06]	[.13]	[.12]	[1,406]	[1,146]
MPSID-2	49	25.32	11.10	.89	.02	.57	.19	1,599	2,225
		[2.63]	[.83]	[.14]	[.08]	[.16]	[.16]	[1,905]	[1,228]
MPSID-3	30	26.86	10.96	.91	.01	.52	.25	1,386	1,863
		[2.97]	[.84]	[.13]	[.08]	[.16]	[.16]	[1,680]	[1,494]
MCPS-1	119	26.91	10.52	.86	.04	.64	.19	2,110	1,396
		[1.25]	[.32]	[.06]	[.04]	[.07]	[.06]	[841]	[563]
MCPS-2	87	26.21	10.21	.85	.04	.68	.20	1,758	1,204
		[1.43]	[.37]	[.08]	[.05]	[.09]	.08	[896]	[661]
MCPS-3	63	25.94	10.69	.87	.06	.53	.13	2,709	1,587
		[1.68]	[.48]	[.09]	[.06]	[.10]	[.09]	[1,285]	[760]

- Note small samples: most controls are discarded

NSW application: Overlap



(Dehejia and Wahba 1999, Figure 2)

- Limited overlap with the CPS control group (not shown: even worse for PSID)

NSW application: Results

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID and CPS

	NSW earnings less comparison group earnings		NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score					
			Quadratic in score[b]	Stratifying on the score			Matching on the score	
	(1) Unadjusted	(2) Adjusted[a]	(3)	(4) Unadjusted	(5) Adjusted	(6) Observations[c]	(7) Unadjusted	(8) Adjusted[d]
NSW	1,794	1,672						
	(633)	(638)						
PSID-1[e]	−15,205	731	294	1,608	1,494	1,255	1,691	1,473
	(1,154)	(886)	(1,389)	(1,571)	(1,581)		(2,209)	(809)
PSID-2[f]	−3,647	683	496	2,220	2,235	389	1,455	1,480
	(959)	(1,028)	(1,193)	(1,768)	(1,793)		(2,303)	(808)
PSID-3[f]	1,069	825	647	2,321	1,870	247	2,120	1,549
	(899)	(1,104)	(1,383)	(1,994)	(2,002)		(2,335)	(826)
CPS-1[g]	−8,498	972	1,117	1,713	1,774	4,117	1,582	1,616
	(712)	(550)	(747)	(1,115)	(1,152)		(1,069)	(751)
CPS-2[g]	−3,822	790	505	1,543	1,622	1,493	1,788	1,563
	(670)	(658)	(847)	(1,461)	(1,346)		(1,205)	(753)
CPS-3[g]	−635	1,326	556	1,252	2,219	514	587	662
	(657)	(798)	(951)	(1,617)	(2,082)		(1,496)	(776)

Caveats

- IW (Lecture 2) report that p-score methods do not eliminate a significant "effect" of NSW on 1975 (pre-NSW) earnings
  - *Exercise:* What are the two possible explanations?
- Smith and Todd (2005a) show other pscore specifications don't perform as well
  - Dehejia (2005) replies that having 2 years of pre-treatment earnings is important, and it's visible in the covariate balance test
- Smith and Todd (2005a,b) point out other sensitivities of Dehejia and Wahba
  - Suggest diff-in-diff + matching works better

Failure of CIA

- Arceneaux, Gerber, Green (2006): a similar analysis of a "Get Out the Vote" program
  - Randomized phone calls but not everyone answers
  - IV methods identify the ATT — could we get it using selection on observables instead?
  - Covariates are similar to Lalonde; bigger sample allows exact matching
  - Yet, the estimates do not match the truth $\Longrightarrow$ CIA must not hold