Propensity Score Analysis

Seminar in Criminology, Research and Analysis– Crim 7301
Week 5, 9/22/16
Andrew Wheeler

Class Overview

- Use Case for Propensity Score
- Weighting vs. Matching
- Balance Statistics

Use Case for Propensity Scores

- Treatment is *categorical* – you either get it or you don't. Treatment cannot be a continuous variable.
- You have a general idea of what predicts selection into treatment, but don't have an exact model.
- You have a large sample of potential control cases, but many may be very different than those in treated sample

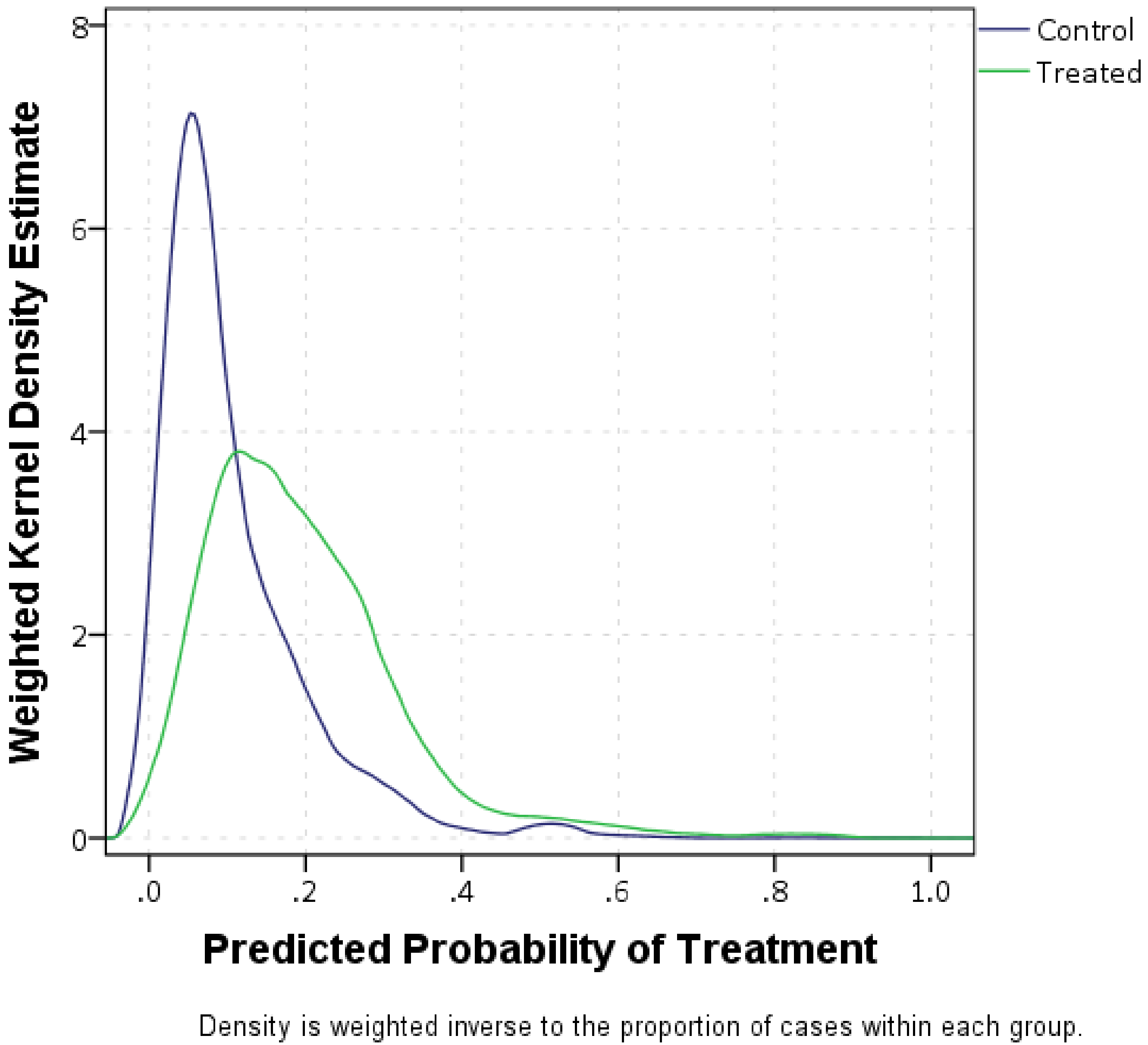
Calculating Propensity Scores

- Predict the probability of treatment (typically using logistic regression)

$$\text{Prob}(T) = f(\beta_0 + \beta_1 X_1 + \cdots \beta_k X_k)$$

- Then can use the predicted probability of treatment, $\hat{p}$, as regression weight:
  - Weight is $1/\hat{p}$ if treated
  - For non-treated it is $1/(1 - \hat{p})$
- Or can included propensity score on right hand side
- Or can match control cases based on $\hat{p}$

Calculating Propensity Scores



Weighting vs. Matching vs. OLS

- Benefits of matching:
  - Matching controls for non-linear effects *on the outcome* for control variables
  - Prevents making apples to oranges comparisons
  - Large weights can cause problems (see Freedman and Berk, 2008, *Weighting Regressions by Propensity Scores)*
  - can be comparable to OLS by using *doubly robust estimates*
  - can be simple, can conduct a t-test of mean differences for the matched control vs treatment sample
  - Can match even with partial missing data

- Drawbacks?
  - Ad-hoc
  - Can't use all the cases

Types of Matching

- Exact matching
- Nearest-Neighbor: Caliper sometimes ¼ standard deviation of the logit, sometimes some small probability difference (1~2%)
- Optimal Matching
- Stratification

Generally these do not make a large difference (as long as balanced obtained) – see work by Austin (2012) *A comparison of 12 algorithms for matching on the propensity score*

Can have multiple matches, diminishing returns after ~5.

Checking Balance

- Local tests for individual covariates:
  - t-tests of mean differences
  - Standardized bias: t = treatment sample, c = control sample, $\bar{X}$ = mean, S^2 = variance. (typically want 20% or less)

$$100 \cdot \frac{(\overline{X_t} - \overline{X_c})}{\sqrt{(S_t^2 + S_c^2)/2}}$$

  - Side by side histograms or other graphics (like a dot plot)

- Global tests
  - Refit the propensity score equation using the same covariates on the *matched* data and test that all coefficients equal zero, see http://blogs.worldbank.org/impactevaluations/tools-trade-joint-test-orthogonality-when-testing-balance

Checking Balance

Table 1: Covariate Balance on Crime Variables

Covariate	Set	Condition	N	Mean	Std. Dev.	T-Test	Std. Bias
Part 1 Violent	Full Samp.	Control	11514	0.9	2.0		
		Treated	345	2.5	3.8	15.0	55
	Matched Samp.	Control	328	1.5	2.4		
		Treated	328	2.0	2.6	2.6	20
Part 1 Non-Violent	Full Samp.	Control	11514	3.8	8.4		
		Treated	345	5.8	8.4	4.4	24
	Matched Samp.	Control	328	3.7	5.7		
		Treated	328	4.9	5.9	2.8	22

Other Tips

- Pro-tip with many variables – do data reduction! (More variables harder to balance them all)

- Software
  - R's MatchIt is the most flexible and tends to produce the best quality matches. Slower though and more computationally intensive
  - Stata has corrections for standard errors built in, but forces you to use weighted statistics (command *psmatch2* does not allow drawing without replacement)
  - SPSS *FUZZY* does not always do a good job drawing comparisons, will work for very large datasets though, and can customize drawing matches (see https://andrewpwheeler.wordpress.com/2015/05/20/fuzzy-matching-in-spss-using-a-custom-python-function/)

Homework & Next Weeks Class

Lab Assignment

- propensity score matching in R (using *MatchIt)*, or Stata (using *psmatch2*), or in SPSS (using *FUZZY*)

For Next Week – Difference in Difference Regression Designs

- readings, *Mostly Harmless*, *Chapter 5*
- Allison, P. (1990). Change scores as dependent variables in regression analysis, *Sociological Methodology* 20: 93-114.
- Maltz, M., Gordon, A., McDowell, D., and McClearly R. (1980). An artifact in pretest-posttest designs: How it can mistakenly make delinquency programs look effective. *Evaluation Review* 4(2): 225-240.