CFDR Summer Methods Seminar: Introduction to Propensity Score Analysis

Matthew VanEseltine
Assistant Professor
Bowling Green State University

June 26, 2013

Outline

1. **Introduction**
2. Matching and Counterfactuals
3. The Propensity Score Matching Process
   a. Estimate Propensity Scores
   b. Match Cases
   c. Estimate Treatment Effects
   d. Demonstrate Robustness
4. Elaborations and Complications
5. End

A Practical Introduction

This workshop is a practical introduction to propensity score analysis (PSA), a relatively new approach to estimating treatment effects with nonexperimental data.

Whereas regression models attempt to balance data by including controls, PSA involves matching cases based on their predicted likelihood to experience values of the independent variable of interest.

The simplest forms of PSA use discrete treatments (e.g., imprisoned or not; became married or remained unmarried) and is best suited for studies of longitudinal data with few moderating or mediating variables.

The workshop will cover various types of matching, strengths and weaknesses of the approach, and tests of robustness, with examples in Stata.

It’s Here

Uses and Mentions of “Propensity Score,” 2003–2013

2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
AJS	SF	AJS	ASR	AJS	AJS	ASR	AJS	AJS	AJS	AJS
AJS		Crim	Crim	Crim	ASR	SF	AJS	ASR	AJS	AJS
Crim		JMF	Crim	Crim	SF	SF	ASR	ASR	ASR	AJS
		JMF		JMF	SF	SF	ASR	ASR	ASR	ASR
		JMF			Crim	Crim	ASR	ASR	SF	SF
		JMF			Crim	Crim	ASR	ASR	SF	Crim
		JMF			JMF	Crim	ASR	ASR	SF	JMF
					JMF	JMF	ASR	ASR	SF	JMF
							ASR	SF	SF	JMF
							SF	SF	SF	JMF
							SF	Crim	Crim	
							SF	Crim	Crim	
							SF	Crim	Crim	
							SF	JMF	Crim	
							Crim		JMF	
							Crim		JMF	
							Crim		JMF	
							Crim		JMF	
							JMF		JMF	
							JMF			
							JMF			
							JMF			
							JMF			
							JMF			

Matching and Counterfactuals

1. Introduction
2. **Matching and Counterfactuals**
3. The Propensity Score Matching Process
   a. Estimate Propensity Scores
   b. Match Cases
   c. Estimate Treatment Effects
   d. Demonstrate Robustness
4. Elaborations and Complications
5. End

Matching and Counterfactuals

Counterfactual reasoning comes at the same kinds of questions we often address with regression. We are often interested estimating treatment effects:

- What was the effect of some life experience, choice or circumstances on some outcome?

Ideally, the causal effect on each case: treated and untreated.

- The perfect study would look at both of your outcomes given the treatment.
- This is Rubin’s “fundamental problem of causal inference.”

	Y^1	Y^0
Treatment (W = 1)	observed	*counterfactual*
Control (W = 0)	*counterfactual*	observed

- We have a missing data problem.
- So we are going to use our best available data to estimate the missing data.

Matching and Counterfactuals

Take a cue from randomized experiments.

- There, random treatment creates identical groups (well – unimportantly different).

What if we could figure out everything that ‘matters’ and just match people together?

Statistically construct sufficiently identical groups

- Exact matching has been around a long time – but handles very few covariates.
- Propensity score analysis instead extracts the relevant information from those covariates (likelihood to receive treatment) to make its matches.

Before I get into the process, a concluding introductory thought:

- Propensity score analysis is not magical (and see Shadish 2013).

Estimate Propensity Scores

1. Introduction
2. Matching and Counterfactuals
3. The Propensity Score Matching Process
   a. **Estimate Propensity Scores**
      - **Propensity Score**
      - **Model Covariates**
      - **Common Support**

   b. Match Cases

   c. Estimate Treatment Effects

   d. Demonstrate Robustness
4. Elaborations and Complications
5. End

Propensity Scores

A propensity score is the probability of being assigned to a treatment.

- Randomized experiments build this into the design. A coin flip: $P(W_i = 1) = 0.5$
- With observational data, we are instead going to try to estimate it.
- An estimated likelihood, given a vector of observed covariates: $P(W_i = 1|X_i)$
- Matching cases on propensity score will approximately **balance** treated and untreated.

An unbiased estimate of the treatment effects **if** we can satisfy requirements, primarily:

1. Strongly Ignorable Treatment Assumption:
   Treatment is independent of outcomes conditional on observed covariates.
2. Stable Unit Treatment Value Assumption:
   One version of the treatment; one case's treatment does not affect another's outcome.

Most commonly, a binomial regression model.

- Regress the treatment on the covariates: propensity score is the predicted probability.
- Alternatives are on the way, particularly generalized boosted modeling (*boost* in Stata).

Model Covariates

Which covariates to include? Remember the goal: predicting selection into treatment. As with regression, a lot comes down to this, and quality matters more than quantity.

There is a debate about quantity:

- "Including a variable that is related to treatment, but not outcome, does not improve balance and reduces the number of matched pairs available for analysis" (Austin et al. 2007). "Familiar econometric rules apply" about the tradeoffs of adding nonsignificant covariates (Ho et al. 2007).

  … as in regression, include covariates jointly related to treatment and outcome.
- "Unless a variable can be excluded because there is a consensus that it is unrelated to the outcome or is not a proper covariate, it is advisable to include it in the propensity score model even if it is not statistically significant." (Rubin 1997)

  … when in doubt, toss it in.

Pay attention to time order! Covariates → Treatment → Outcome.

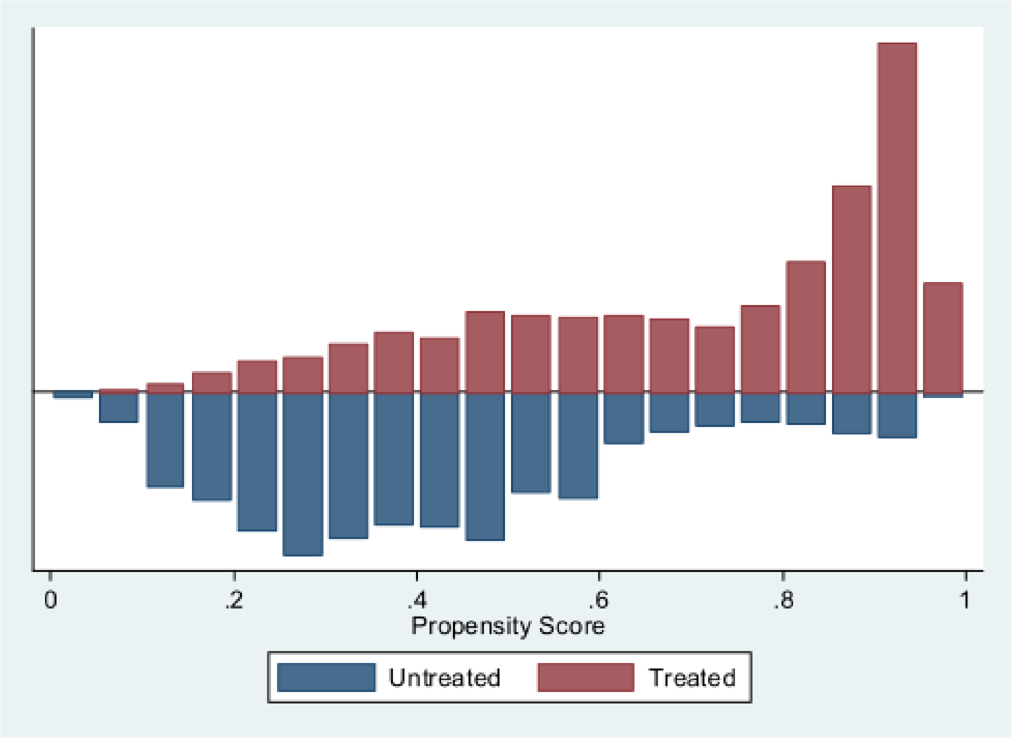
Common Support

Limit inferences to a range with information on both treated and untreated cases.

- Throw out all treated with higher propensity scores than the highest untreated.
- Throw out all untreated with lower propensity scores than the lowest treated.

And now is a good time to look at the distribution of your propensity score.

This is *psgraph* in Stata:



Match Cases

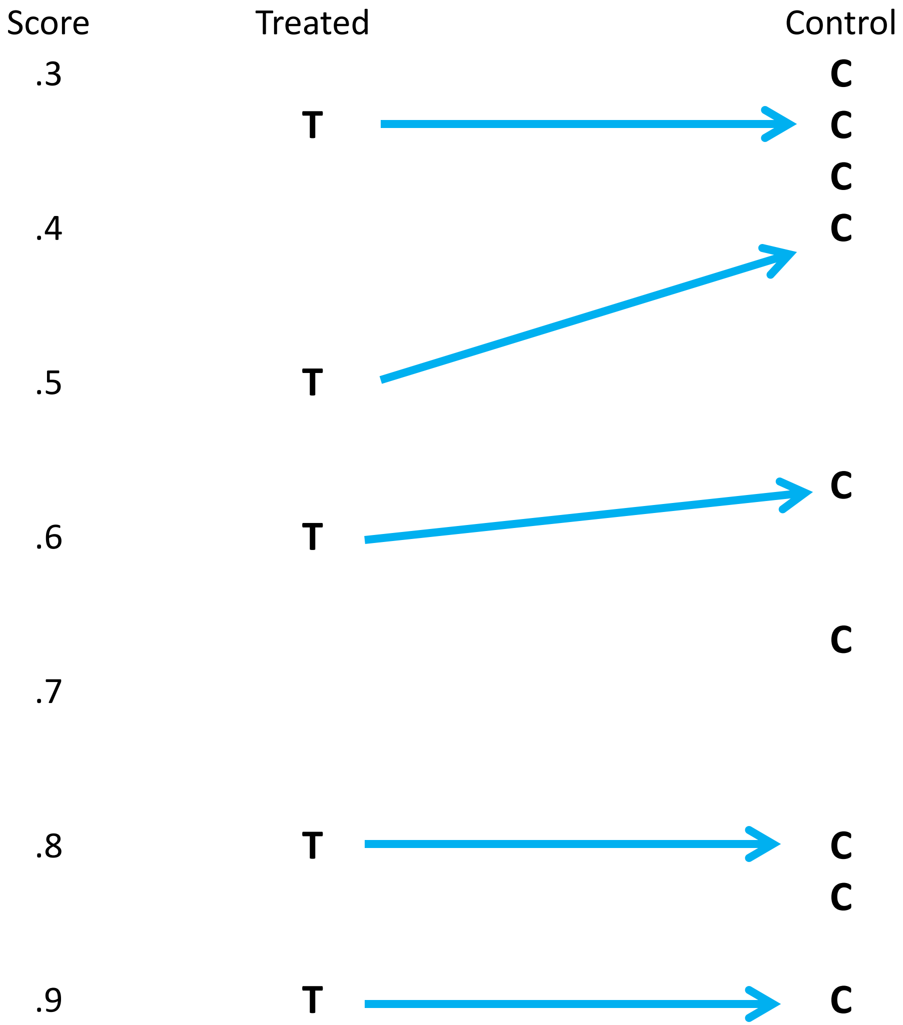
1. Introduction
2. Matching and Counterfactuals
3. The Propensity Score Matching Process
    a. Estimate Propensity Scores
    b. Match Cases
    - **Matching Algorithms**
    - **Check for Balance**

    c. Estimate Treatment Effects
    d. Demonstrate Robustness
4. Elaborations and Complications
5. End

Matching Algorithms

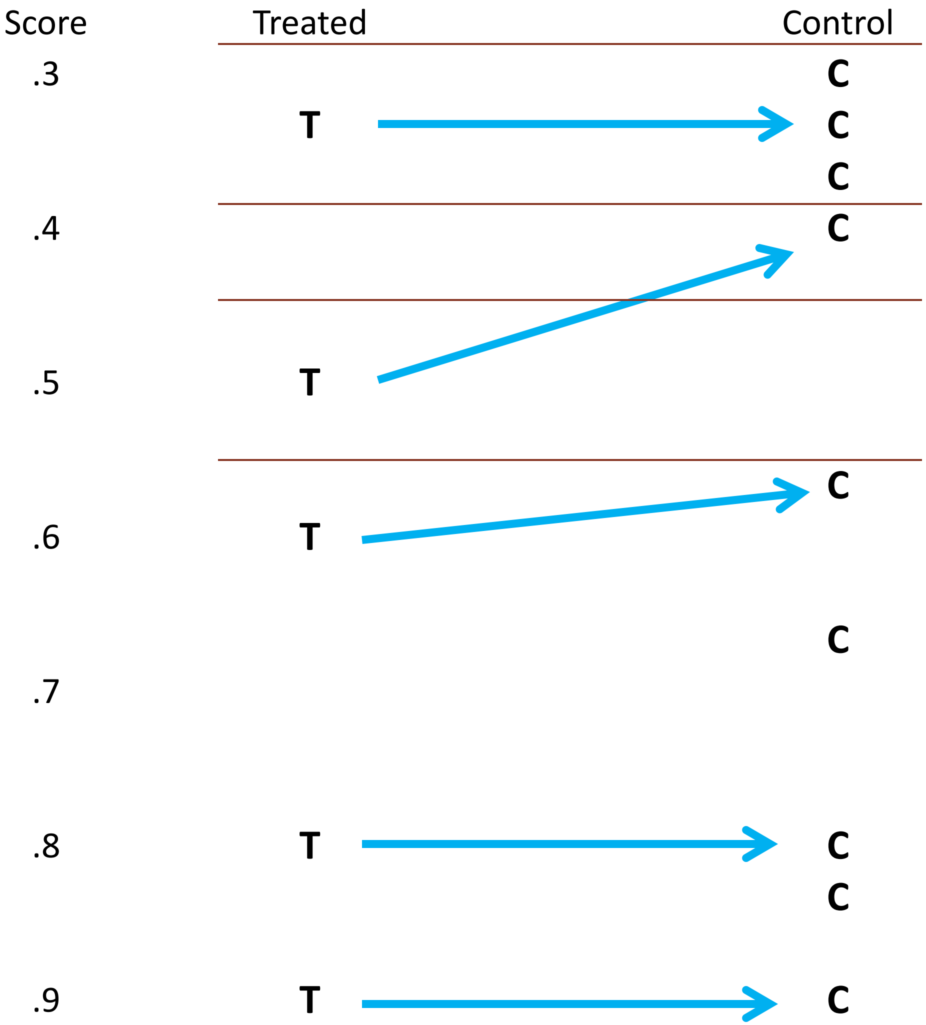
Score	Treated	Control
.3		**C**
	T	**C**
		C
.4		**C**
.5	**T**	
		C
.6	**T**	
		C
.7		
.8	**T**	**C**
		C
.9	**T**	**C**

Nearest-Neighbor



Score	Treated	Control
.3		C
	T	C
		C
.4		C
.5	T	
		C
.6	T	
		C
.7		
.8	T	C
		C
.9	T	C

NN within Caliper



Score Treated Control

.3 C

T C

C

.4 C

.5 T

C

.6 T

C

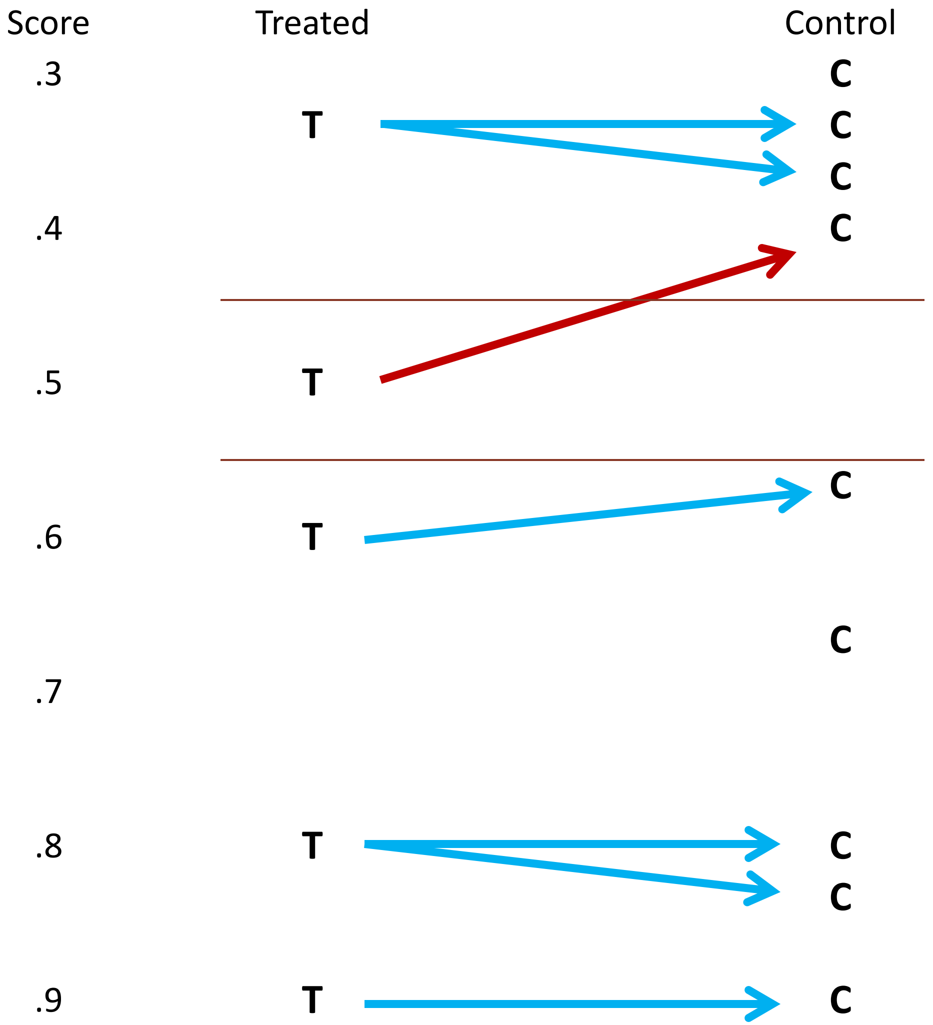
.7

.8 T C

C

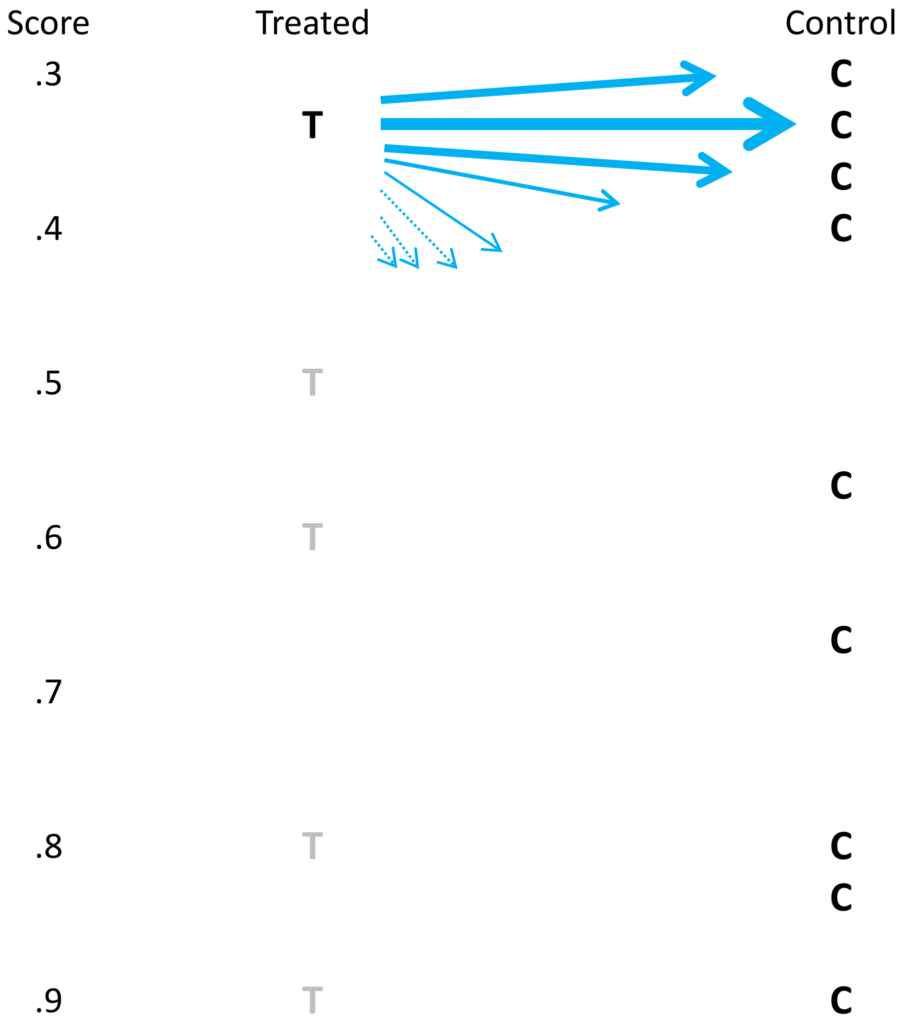
.9 T C

2-to-1 within Caliper



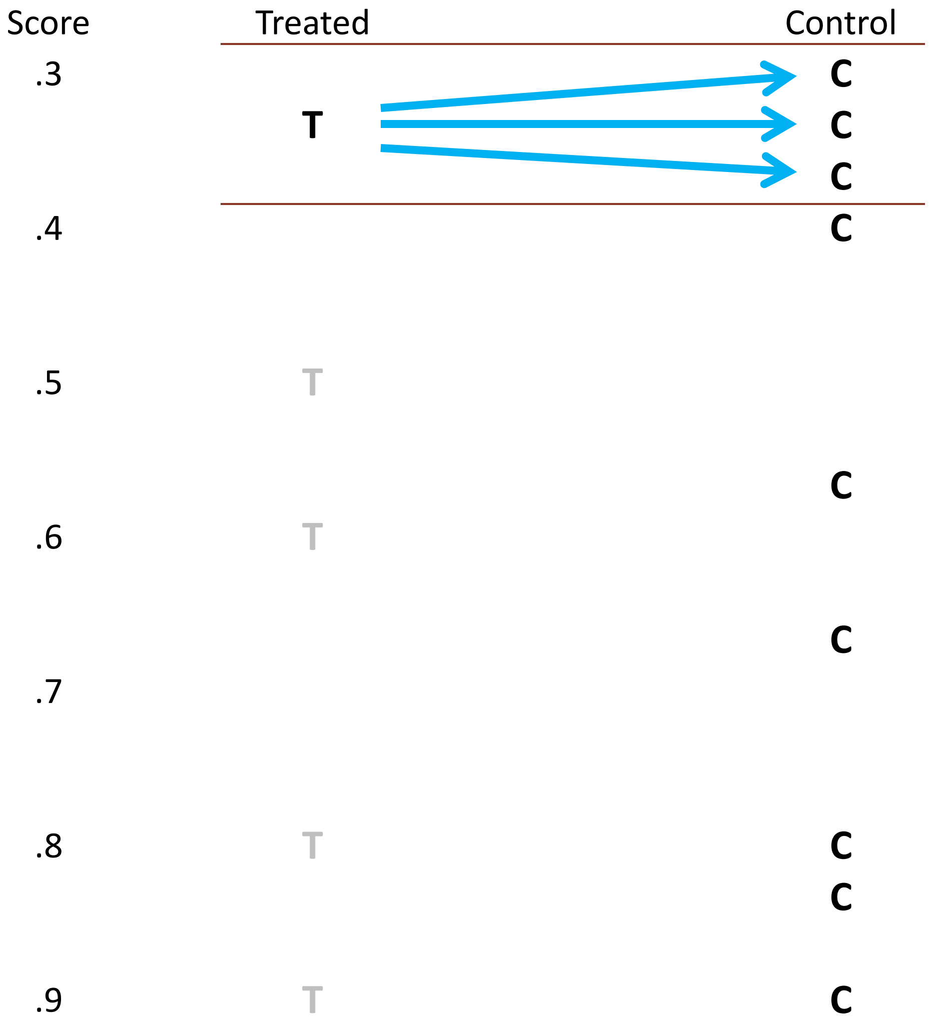
Score	Treated	Control
.3		C
	T	C
		C
.4		C
.5	T	
		C
.6	T	
		C
.7		
.8	T	C
		C
.9	T	C

Kernel (Gaussian)



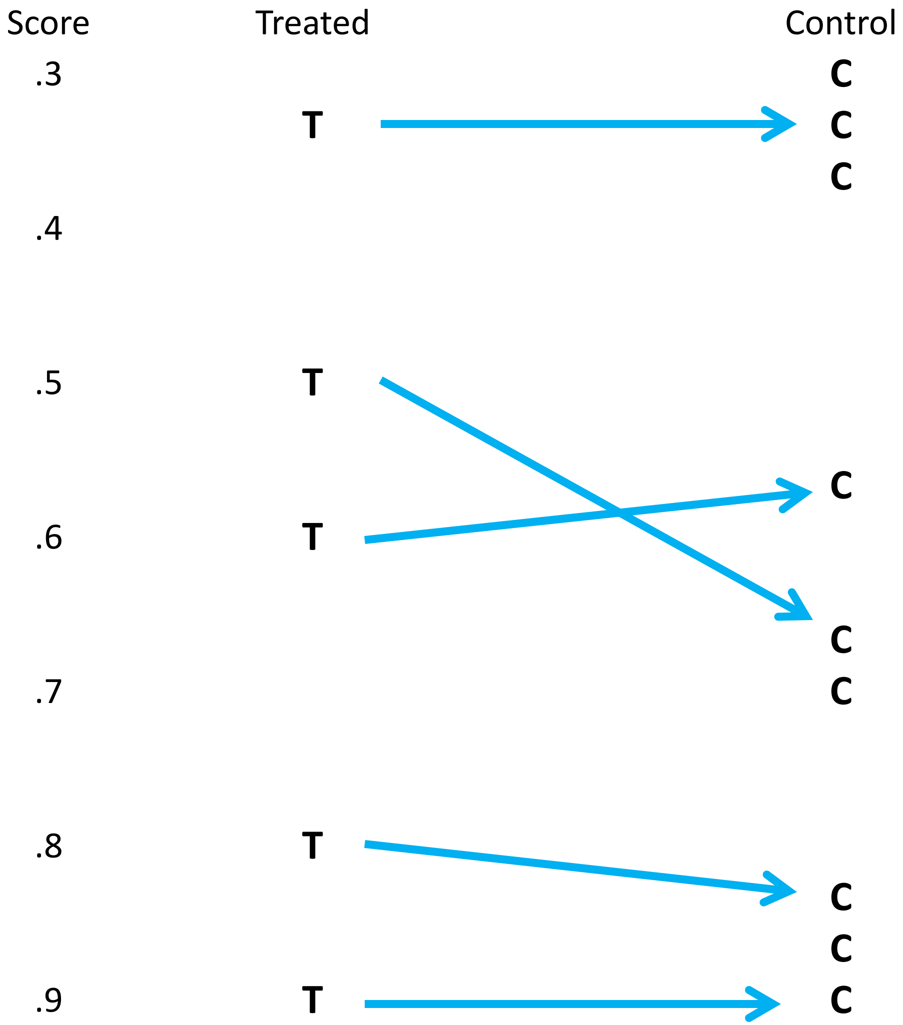
Score	Treated	Control
.3		C
	T	C
		C
.4		C
.5	T	
		C
.6	T	
		C
.7		
.8	T	C
		C
.9	T	C

Kernel (Uniform)



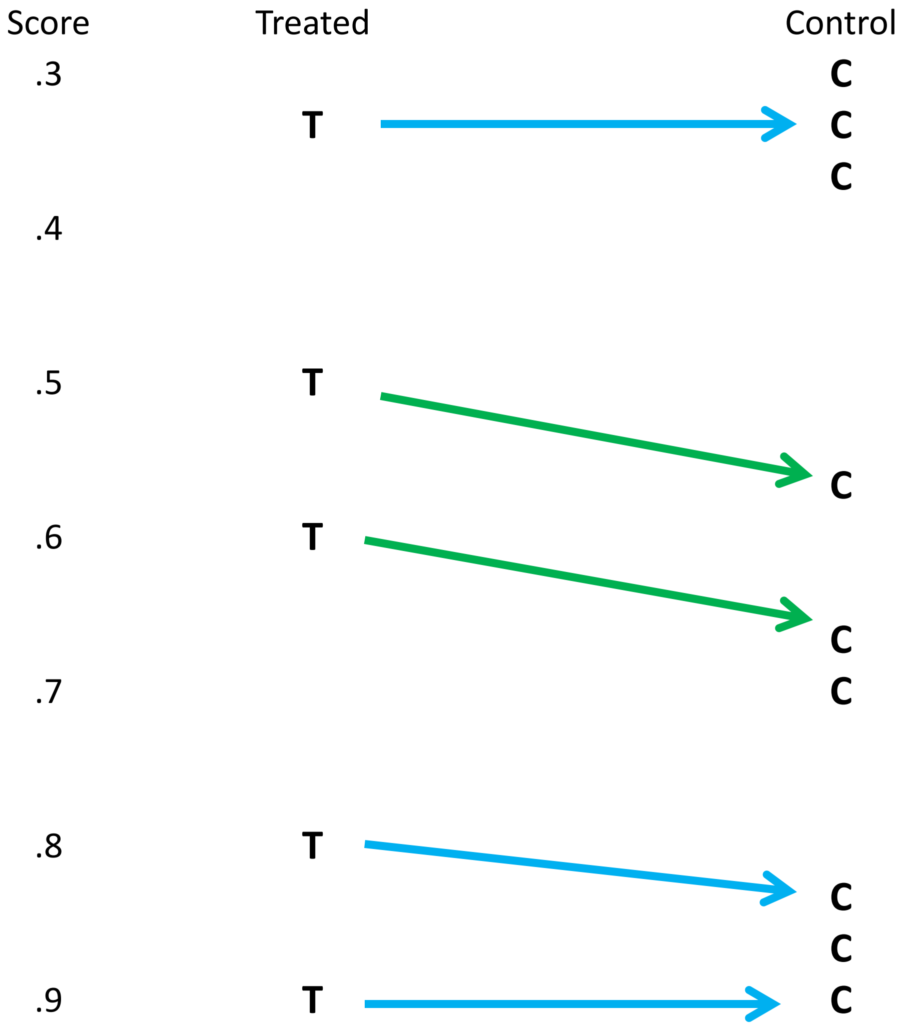
Score	Treated	Control
.3		C
	T	C
		C
.4		C
.5	T	
		C
.6	T	
		C
.7		
.8	T	C
		C
.9	T	C

Greedy vs. Optimal



Score	Treated	Control
.3		C
	T	C
		C
.4		
.5	T	
		C
.6	T	
		C
.7		C
.8	T	
		C
		C
.9	T	C

Greedy vs. Optimal



Matching Algorithms

A range of matching protocols… so how do you decide?

Tradeoffs between variance and bias, completeness and accuracy.

- With many untreated, many-to-one improves efficiency.
- Kernel is preferred when treated and control groups have "quite different" distributions (Apel and Sweeten 2010).

Nearest-neighbor, without replacement, with caliper is a good default.

- Recommended width: 0.2 of the SD of the logit of the propensity score (Austin 2011)

For us, availability probably matters. See references for software links.

- In Stata, *pscore* supports nearest-neighbor, kernel, and radius matching.
  - And *psmatch2* adds Mahalanobis to that list.
- Optimal matching is available in R, but not Stata (yet).

And as always, consult your field's literature for standard expectations.

Check for Balance

A critical evaluation of the Strongly Ignorable Treatment Assumption:

> Conditional on covariates *X*, the assignment to treatment *W* will be independent of the potential outcomes *Y* if observable covariates are held constant.

This also known as unconfoundedness, selection on observables, conditional independence, exogeneity… the same concern underlies controls in OLS.

Did the propensity score successfully balance the data on **observed** covariates? Compare differences between treated and untreated for each covariate before and after matching.

- Can use standardized bias (Rosenbaum and Rubin 1985b): under 10 as a rule of thumb
  - Available for Stata in *pstest*.
- Can use a general guideline: less than 25% of the SD of X (Ho et al. 2007)
- No hard rules. You'll see t-tests being used, though they're debatable here.
- In Stata, *pscore* can test for balance by strata: don't make it angry.

Estimate Treatment Effects

1. Introduction
2. Matching and Counterfactuals
3. The Propensity Score Matching Process
   a. Estimate Propensity Scores
   b. Match Cases
   c. **Estimate Treatment Effects**
      - **Average Treatment Effect on the Treated (ATT)**
      - **Alternatives to Matching**
   d. Demonstrate Robustness
4. Elaborations and Complications
5. End

Treatment Effect Estimate

Really estimating **average treatment effect on the treated** (ATT).

- Methodologically, we have been matching controls to our treated cases.
- Conceptually, the treated could have gone untreated (more than the other way around).

The most commonly used packages in Stata are *pscore* and *psmatch2*.

Becker, Sascha O. and Andrea Ichino. 2002. “Estimation of Average Treatment Effects Based on Propensity Scores.” *The Stata Journal* 2(4): 358–377.

Alternatives to Matching

Stratification on the propensity score.

- Bin the sample into quintiles (or finer) by propensity score.
- Five subclasses are expected to remove 90% of bias from modeled covariates.
- Favored not for the overall estimate as much as the substantive value.

Stratification

Score	Treated	Control	Strata
.3			
	T	C	
			1
.4		C	
		C	
	T		
			2
.5		C	
	T	C	
.6	T		
			3
		C	
	T		
.7		C	
			4
	T	C	
.8	T		
	T	C	
	T		5
.9	T	C	

Alternatives to Matching

Stratification on the propensity score.

- Bin the sample into quintiles (or finer) by propensity score.
- Five subclasses are expected to remove 90% of bias from modeled covariates.
- Favored not for the overall estimate as much as the substantive value.

Covariate adjustment on the propensity score.

- Regress the outcome on the treatment, controlling for the propensity score.
- Restrict to on-support cases; can also trim the sample; can add controls.
- Will usually give you results extremely similar to ordinary regression approach.

Demonstrate Robustness

1. Introduction
2. Matching and Counterfactuals
3. The Propensity Score Matching Process
    a. Estimate Propensity Scores
    b. Match Cases
    c. Estimate Treatment Effects
    d. **Demonstrate Robustness**
        - **Rosenbaum Bounds**
        - **Modifications and Variations**
4. Elaborations and Complications
5. End

Rosenbaum Bounds

Rosenbaum (2002): A sensitivity analysis for observational studies should ask “what the unmeasured covariate would have to be like to alter the conclusions of the study.”

- Really showing us “what happens when we violate ignorable treatment assumption”?
- The standard formal sensitivity analysis for propensity score matching in sociology.

Gamma (Γ) is a hypothetical odds ratio of an increase in odds of the outcome for treated cases due to unobserved covariate(s).

- 1.0 (no bias) up by intervals to 2.0.
- “How bad does it have to be?”
- “When do I lose confidence in my effect?”
- Better to show confidence intervals.

Available for Stata as *rbounds*.

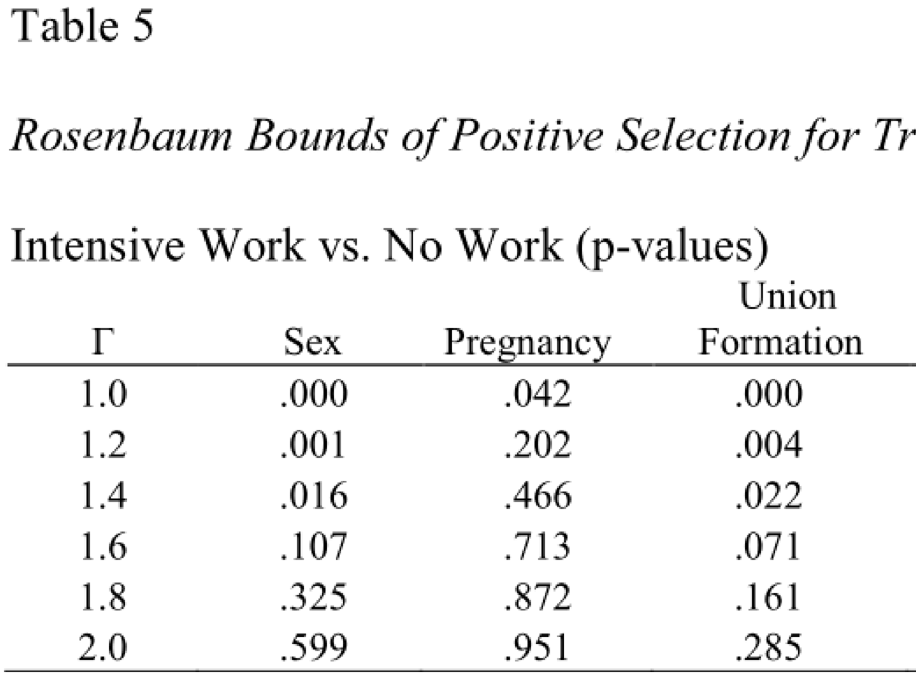


Table 5

Rosenbaum Bounds of Positive Selection for Tr

Intensive Work vs. No Work (p-values)

Γ	Sex	Pregnancy	Union Formation
1.0	.000	.042	.000
1.2	.001	.202	.004
1.4	.016	.466	.022
1.6	.107	.713	.071
1.8	.325	.872	.161
2.0	.599	.951	.285

Modifications and Variations

Some (like in a regression approach) depends on your research question.
Different formulations of outcomes? Definitions of treatment? Subgroups?

Especially important covariate(s)?

- Might separate analyses into subgroups
- Might include covariate as an exact match

Shenyang Guo recommends a “3 x 2 x 2” design, for twelve total treatment estimates

- 3 alternative propensity score models
- 2 different matching algorithms
- 2 variations in matching specifications
- “Which choice should I make?” “Both!”

Elaborations and Complications

1. Introduction
2. Matching and Counterfactuals
3. The Propensity Score Matching Process
    a. Estimate Propensity Scores
    b. Match Cases
    c. Estimate Treatment Effects
    d. Demonstrate Robustness
4. **Elaborations and Complications**
5. End

What About…

Sample size?

- At least 1,000–1,500 is recommended (Shadish 2013).

Missing data?

- Some disagreement. Listwise is most common. Multiple imputation might be okay.

Clustered data?

- Still an open issue. Multilevel matching is on the way. (Within-group or within-person.)

Weighting?

- You can use population weights on the final estimates if it makes sense.
- Sampling weights are generally out.

Non-binary treatment variable?

- Multinomial, ordinal, continuous… on the way, and examples in the literature.

End

1. Introduction
2. Matching and Counterfactuals
3. The Propensity Score Matching Process
   a. Estimate Propensity Scores
   b. Match Cases
   c. Estimate Treatment Effects
   d. Demonstrate Robustness
4. Elaborations and Complications
5. **End**

Points for Review

1. Why is propensity score matching appropriate for this research question and data?
2. Was the treatment adequately conceptualized and measured?
3. Did the propensity score model include appropriate covariates?
4. Was the support condition addressed and enforced?
5. Were all covariates shown to be successfully balanced?
6. Did the authors justify their choices of matching procedures and their parameters?
7. Did the authors demonstrate the robustness of their results?

For longer lists, see Apel and Sweeten (2010:559–560), Guo and Fraser (2010:321–326).

Key Features

The reasons *I* find propensity score matching especially advantageous or compelling:

- Counterfactual framework.
- Common support and “apples to apples” comparisons.
- Nonparametric estimates of treatment effects.
- Rosenbaum bounds and sensitivity testing.
- Generally promotes good habits.

Morgan, Winship, and Harding (2007): "Matching represents an intuitive method for addressing causal questions, primarily because it pushes the analyst to confront the process of causal exposure as well as the limitations of available data."

Acknowledgements

Special thanks are due to certain Penn State faculty. My training in this methodology was founded formally by D. Wayne Osgood, and more informally by Jeremy Staff and Michael Massoglia (now at Wisconsin).

This work was supported by the Center for Family and Demographic Research, Bowling Green State University, which has core funding from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R24HD050959-09).

My Stata example uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website. No direct support was received from grant P01-HD31921 for this analysis.

Examples in Handout

Apel, Robert, Arjan A.J. Blokland, Paul Nieuwbeerta, and Marieke van Schellen. 2009. “The Impact of Imprisonment on Marriage and Divorce: A Risk Set Matching Approach.” *Journal of Quantitative Criminology* 26:269–300. **Sensitivity, null treatment comparison.**

Frisco, Michelle L., Chandra Muller, and Kenneth Frank. 2007. “Parents’ Union Dissolution and Adolescents’ School Performance: Comparing Methodological Approaches.” *Journal of Marriage and Family* 69:721–741. **Nearest neighbor, kernel, OLS comparison, rare treatment.**

Yanovitzky, Itzhak, Elaine Zanutto, and Robert Hornik. 2005, “Estimating causal effects of public health education campaigns using propensity score methodology.” *Evaluation and Program Planning* 28:209–220. **Pre- and post-adjustment, stratification by quintile, dosage.**

Meier, Ann M. 2007. “Adolescent First Sex and Subsequent Mental Health.” *American Journal of Sociology* 112:1811–1847. **Full adjustment table, score distribution, many subgroups.**

King, Ryan D., Michael Massoglia, and Ross Macmillan. 2007. “The Context of Marriage and Crime: Gender, the Propensity to Marry, and Offending in Early Adulthood.” *Criminology* 45:33–65. **OLS comparison, gender subgroups, stratification within gender.**

References

Introduction and Overview

Apel, Robert J. and Gary Sweeten. 2010. "Propensity Score Matching in Criminology and Criminal Justice" in *Handbook of Quantitative Criminology,* edited by Alex R. Piquero and David Weisburd. New York, NY: Springer.

Steiner, Peter M. and David Cook. 2013. "Matching and Propensity Scores" in *The Oxford Handbook of Quantitative Methods*, edited by Todd L. Little. New York, NY: Oxford University Press.

Full-Scale Coverage

Guo, Shenyang and Mark W. Fraser. 2010. *Propensity Score Analysis: Statistical Methods and Applications*. Thousand Oaks, CA: Sage.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference*. New York, NY: Cambridge University Press.

References

Elizabeth Stuart maintains a page with information on propensity score procedures available to researchers. This will be a good place to look if you go beyond Stata:
http://www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html

Stata:

Becker, Sascha O. and Andrea Ichino. 2002. “Estimation of Average Treatment Effects Based on Propensity Scores.” *Stata Journal* 2(4): 358-377. [***pscore***]

Leuven, Edwin and Barbara Sianesi. 2003. “PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.” [***psmatch2***]

R:

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. *MatchIt*.
http://gking.harvard.edu/matchit

References

Some – Not All – Core Technical Work

Rosenbaum, Paul. R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55.

Rosenbaum, Paul. R. and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79: 516–524.

Rosenbaum, Paul. R. and Donald B. Rubin. 1985a. "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." *The American Statistician* 39: 33–38.

Rosenbaum, Paul. R. and Donald B. Rubin. 1985b. "The Bias Due to Incomplete Matching." *Biometrics* 41: 103–116.

Rosenbaum, Paul R. 2002. *Observational Studies*, 2nd Ed. New York, NY: Springer.

Rubin, Donald B. 1997. "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Annals of Internal Medicine* 127: 757–763.

References

Materials from Other Presentations

Chen, Vivien W., and Krissy Zeiser. 2007. “Implementing Propensity Score Matching Causal Analysis with Stata” at the Population Research Institute, Penn State University, February 27, 2008. Slides: http://help.pop.psu.edu/help-by-statistical-method/propensity-matching/Intro%20to%20P-score_Sp08.pdf

Guo, Shenyang. 2010. “Overview of Propensity Score Matching” at Children and Family Futures, September 1, 2010. Slides: http://cffutures.com/files/webinar-handouts/Presentation%20Propensity%20Scoring%20Session%20I.pdf

Massoglia, Michael. 2012. “Propensity Score Models” at Indiana University, November 30, 2012. Slides: http://www.indiana.edu/~wim/docs/11_30_2012_massoglia_Propensity%20Score%20Models-IU.pptx

Stuart, Elizabeth. 2011. ”The Why, When, and How of Propensity Score Methods for Estimating Causal Effects” at Society for Prevention Research, May 31, 2011. Slides: http://www.preventionresearch.org/wp-content/uploads/2011/07/SPR-Propensity-pc-workshop-slides.pdf