Thank you for joining us today!

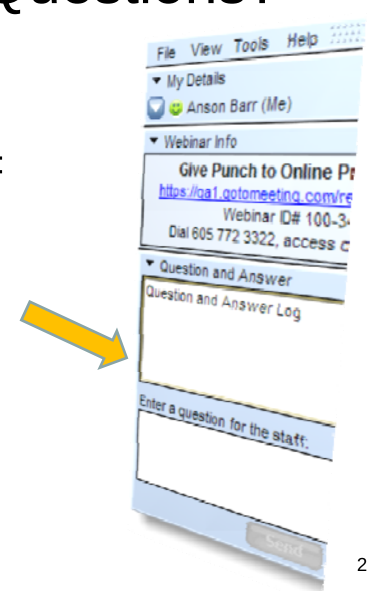
- If you haven't dialed into the audio (telephone) portion, please do so now:
  1-866-516-5393
  Access Code: 33403311
- If you are experiencing technical problems with the GoToWebinar program (visual portion), contact the help desk:
  1-800-263-6317
  Reference Webinar ID: 336202560
- Today's presentation and handouts are available for download at:
  http://www.cffutures.com/webinars

The webinar will begin shortly. Thank you for your patience.

How Do I Ask Questions?

For your convenience, there are two ways to ask questions during this webinar presentation:

1. Type and send your questions through the Question and Answer log located on the bottom half on your panel/dashboard.
2. There will also be time at the end of the webinar for you to ask questions via the conference line.



Propensity Score Matching
Strategies for Evaluating Substance Abuse Services for Child Welfare Client, Session I of II

I) Introductions

Ken DeCerchio

II) Overview: Propensity Score Matching

Shenyang Guo

III) Conceptual Frameworks and Assumptions

Shenyang Guo

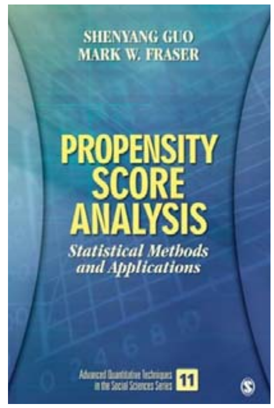
IV) Overview: Corrective Methods

Shenyang Guo

V) Greedy Propensity Score Matching

Shenyang Guo

VI) Discussion/Questions



Part I – Overview of Propensity Score Matching

1. Why and when propensity score analysis is needed
2. Conceptual frameworks and assumptions
3. Overview of corrective methods
4. Greedy propensity score matching

1. Why and when propensity score analysis is needed

Purpose of Evaluation

The field of program evaluation is distinguished principally by cause-effect studies that aim to answer a key question:

To what extent can the net difference observed in outcomes between treated and nontreated groups be attributed to an intervention, given that all other things are held constant?

Note. The term “intervention research” refers to the design and evaluation of programs.

Internal Validity and Threats

- Internal validity – the validity of inferences about whether the relationship between two variables is causal (Shadish, Cook, & Campbell, 2002).
- In program evaluation and observational studies in general, researchers are concerned about threats to internal validity. These threats are factors affecting outcomes other than intervention or the focal stimuli. There are nine types of threats.*
- Selection bias is the most problematic one!

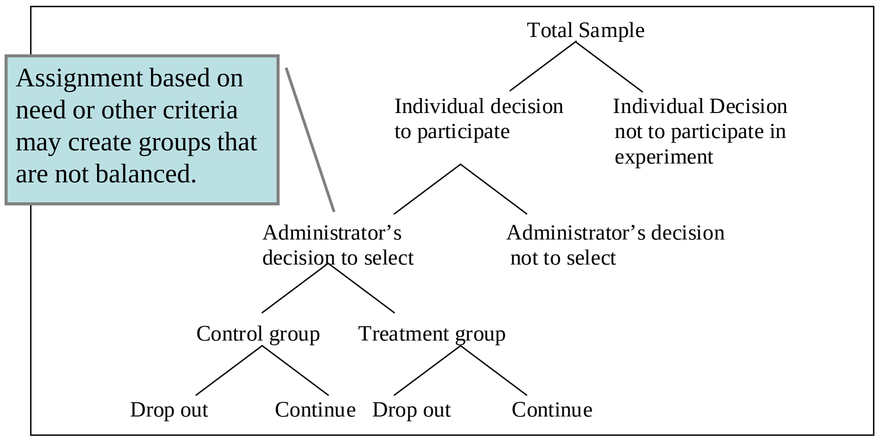
*These include differential attrition, maturation, regression to the mean, instrumentation, and testing effects.

Fisher's Randomized Experiment

- A theory of observational studies must have a clear view of the role of randomization, so it can have an equally clear view of the consequences of its absence (Rosenbaum, 2002).
- Fisher's book, *The Design of Experiments* (1935/1971), introduced the principles of randomization, demonstrating them with the example of testing a British lady's tea tasting ability.

Sources of Selection

Example of Selection Bias:
Decision Tree for Evaluation of Social Experiments



Source: Maddala, 1983, p. 266

Why and when propensity score analysis is needed? (1)

Need 1: Remove Selection Bias

The randomized clinical trial is the "gold standard" in outcome evaluation. However, in social and health research, RCTs are not always practical, ethical, or even desirable. Under such conditions, evaluators often use quasi-experimental designs, which – in most instances – are vulnerable to selection. Propensity score models help to remove selection bias.

Example: In an evaluation of the effect of Catholic versus public school on learning, Morgan (2001) found that the Catholic school effect is strongest among Catholic school students who are less likely to attend Catholic schools.

Why and when propensity score analysis is needed? (2)

Need 2: Analyze causal effects in observational studies

- Observational data - those that are not generated by mechanisms of randomized experiments, such as surveys, administrative records, and census data.
- To analyze such data, an ordinary least square (OLS) regression model using a dichotomous indicator of treatment does not work, because in such model the error term is correlated with explanatory variables. The violation of OLS assumption will cause an inflated and asymptotically biased estimate of treatment effect.

The Problem of Contemporaneous Correlation in Regression Analysis

Consider a routine regression equation for the outcome, Y_i:

$$Y_i = \alpha + \tau W_i + \beta X_i + e_i$$

where W_i is a dichotomous variable indicating intervention, and X_i is the vector of covariates for case i.

In this approach, we wish to estimate the effect (τ) of treatment (W) on Y_i by controlling for observed confounding variables (X_i).

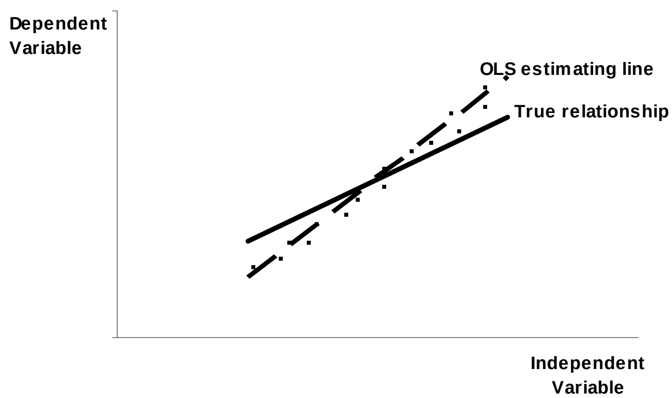
When randomization is compromised or not used, the correlation between W and e may not be equal to zero. As a result, the ordinary least square estimator of the effect of intervention (τ) may be biased and inconsistent. W is not exogenous.

How Big Is This Problem?

Very big! The majority of nonrandomized studies that have used statistical controls to balance treatment and nontreatment groups <u>may</u> have produced erroneous findings.

Note. The amount of error in findings will be related to the degree to which the error term is NOT independent of explanatory measures, including the treatment indicator. This problem applies to any statistical model in which the independence of the error term is assumed.

Consequence of Contemporaneous Correlation: Inflated (Steeper) Slope and Asymptotical Bias



Source: Kennedy (2003), p.158

2. Conceptual frameworks and assumptions

The Neyman-Rubin Counterfactual Framework (1)

- **Counterfactual**: what would have happened to the treated subjects, had they not received treatment?
- One of the seminal developments in the conceptualization of program evaluation is the Neyman (1923) – Rubin (1978) **counterfactual framework**. The key assumption of this framework is that individuals selected into treatment and nontreatment groups have potential outcomes in both states: the one in which they are observed and the one in which they are not observed. This framework is expressed as:

$$Y_i = W_i Y_{1i} + (1 - W_i) Y_{0i}$$

- The key message conveyed in this equation is that to infer a causal relationship between W_i (the cause) and Y_i (the outcome) the analyst cannot directly link Y_{1i} to W_i under the condition $W_i = 1$; instead, the analyst must check the outcome of Y_{0i} under the condition of $W_i = 0$, and compare Y_{0i} with Y_{1i}.

The Neyman-Rubin Counterfactual Framework (2)

- There is a crucial problem in the above formulation: Y_{0i} is not observed. Holland (1986, p. 947) called this issue the "fundamental problem of causal inference."
- The Neyman-Rubin counterfactual framework holds that a researcher can estimate the counterfactual by examining the average outcome of the treatment participants (i.e., $E(Y_1|W=1)$]) and the average outcome of the nontreatment participants (i.e., $E(Y_0|W=0)$)in the population. Because both outcomes are observable, we can then define the treatment effect as a mean difference:

$$\tau = E(Y_1|W=1) - E(Y_0|W=0)$$

- Under this framework, the evaluation of $E(Y_1|W=1) - E(Y_0|W=0)$ can be thought as an effort that uses $E(Y_0|W=0)$ to estimate the counterfactual $E(Y_0|W=1)$. The central interest of the evaluation is not in $E(Y_0|W=0)$, but in $E(Y_0|W=1)$.

The Neyman-Rubin Counterfactual Framework (3)

- With sample data, evaluators can estimate the average treatment effect as:

$$\hat{\tau} = E(\hat{y}_1 \mid w = 1) - E(\hat{y}_0 \mid w = 0)$$

- The real debate about the classical experimental approach centers on the question: whether $E(Y_0|W=0)$ really represents $E(Y_0|W=1)$?
- In a series of papers, Heckman and colleagues criticized this assumption.
- Consider $E(Y_1|W=1) – E(Y_0|W=0)$. Add and subtract $E(Y_0|W=1)$, we have

  $\{E(Y_1|W=1) – E(Y_0|W=1)\} + \{E(Y_0|W=1) - E(Y_0|W=0)\}$

  The standard estimator provides unbiased estimation if and only if $\mathbf{E(Y_0|W=1) = E(Y_0|W=0)}$.

  In many empirical projects, $\mathbf{E(Y_0|W=1) \neq E(Y_0|W=0)}$.

The Neyman-Rubin Counterfactual Framework (4)

Heckman & Smith (1995) - Four Important Questions:

- What are the effects of factors such as subsidies, advertising, local labor markets, family income, race, and sex on program application decision?
- What are the effects of bureaucratic performance standards, local labor markets and individual characteristics on administrative decisions to accept applicants and place them in specific programs?
- What are the effects of family background, subsidies and local market conditions on decisions to drop out from a program and on the length of time taken to complete a program?
- What are the costs of various alternative treatments?

The Fundamental Assumption: Strongly Ignorable Treatment Assignment

- Rosenbaum & Rubin (1983)

$$(Y_0, Y_1) \perp W \mid X.$$

- Different versions: “unconfoundedness” and “ignorable treatment assignment” (Rosenbaum & Robin, 1983), “selection on observables” (Barnow, Cain, & Goldberger, 1980), “conditional independence” (Lechner 1999), and “exogeneity” (Imbens, 2004)

The SUTVA assumption (1)

- To evaluate program effects, statisticians also make the *Stable Unit Treatment Value Assumption*, or SUTVA (Rubin, 1986), which says that the potential outcomes for any unit do not vary with the treatments assigned to any other units, and there are no different versions of the treatment.
- Imbens (on his Web page) uses an aspirin example to interpret this assumption, that is, the first part of the assumption says that taking aspirin has no effect on your headache, and the second part of the assumption rules out differences on outcome due to different aspirin tablets.

The SUTVA assumption (2)

- According to Rubin, SUTVA is violated when there exists interference between units or there exist unrepresented versions of treatments.
- The SUTVA assumption imposes *exclusion* restrictions on outcome differences. Because of this reason, economists underscore the importance of analyzing *average treatment effects for the subpopulation of treated units*, which is frequently more important than the effect on the population as a whole. This is especially a concern when evaluating the importance of a narrowly targeted program, e.g., a labor-market intervention.
- What statisticians and econometricians called "evaluating average treatment effects for the treated" is similar to the *efficacy subset analysis* found in the literature of intervention research.

Two traditions (1)

There are two traditions in modeling causal effects when random assignment is not possible or is compromised: the econometric versus the statistical approach

- ➢ The econometric approach emphasizes the structure of selection, and, therefore, underscores a direct modeling of selection bias
- ➢ The statistical approach assumes that selection is random conditional on covariates.

- ➢ Both approaches emphasize a direct control of observed covariates by using conditional probability of receiving treatment
- ➢ The two approaches are based on different assumptions for their correction models and differ on the level of restrictiveness of assumptions

Two traditions (2)

Heckman's econometric model of causality (2005) and the contrast of his model to the statistical model

	Statistical Causal Model	**Econometric Models**
Sources of randomness	Implicit	Explicit
Models of conditional counterfactuals	Implicit	Explicit
Mechanism of intervention for determining counterfactuals	Hypothetical randomization	Many mechanisms of hypothetical interventions including randomization; mechanism is explicitly modeled
Treatment of interdependence	Recursive	Recursive or Simultaneous systems
Social/market interactions	Ignored	Modeled in general Equilibrium frameworks
Projections to different populations?	Does not project	Projects
Parametric?	Nonparametric	Becoming nonparametric
Range of questions answered	One focused treatment effect	In principle, answers many possible questions

Source: Heckman (2005, p.87)

3. Overview of Corrective Methods

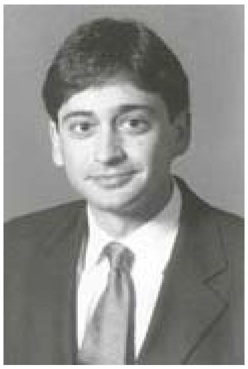
Four Models Described by Guo & Fraser (2010)

1. Heckman’s sample selection model (Heckman, 1976, 1978, 1979) and its revised version estimating treatment effects (Maddala, 1983)

Four Models Described by Guo & Fraser (2010)

2. Propensity score matching (Rosenbaum & Rubin, 1983), optimal matching (Rosenbaum, 2002), propensity score weighting, modeling treatment dosage, and related models

Four Models Described by Guo & Fraser (2010)



3. Matching estimators (Abadie & Imbens, 2002, 2006)

Four Models Described by Guo & Fraser (2010)

4. Propensity score analysis with nonparametric regression (Heckman, Ichimura, & Todd, 1997, 1998)

General Procedure for Propensity Score Matching Summarized by Guo & Fraser (2010)

Step 1: Logistic regression
- Dependent variable: log odds of receiving treatment
- Search an appropriate set of conditioning variables (boosted regression, etc.)
- Estimated propensity scores: predicted probability (p) or log[(1-p)/p].

Step 2:Analysis using propensity scores
- Analysis of weighted mean differences using kernel or local linear regression (difference-in-differences model of Heckman et al.)

Step 2: Analysis using propensity scores:
- Multivariate analysis using propensity scores as weights

Step 2: Matching
- Greedy match (nearest neighbor with or without calipers)
- Mahalanobis with or without propensity scores
- Optimal match (pair matching, matching with a variable number of controls, full matching)

Step 3: Post-matching analysis
- Multivariate analysis based on matched sample

Step 3: Post-matching analysis
- Stratification (subclassification) based on matched sample

Computing Software Packages for Running the Four Models (Stata & R)

	Procedure Name & Useful References	
Chapter & Methods	**Stata**	**R**
Chapter 4		
Heckman (1978, 1979) sample seletion model & Madalla (1983) treatment effect model	*heckman* (StataCorp, 2003)	*sampleSelection* (Toomet Henningsen, 2008)
	treatreg (StataCorp, 2003)	
Chapter 5		
Rosenbaum & Rubin's (1983) propensity score matching	*psmatch2* (Leuven & Sianesi, 2003)	*cem* (Deheja & Wahba, 1999; Iacus, King, & Porro, 2008)
		Matching (Sekehon, 2007)
		MatchIt (Ho, Imai, King, & Stuart, 2004)
		PSAgraphics (Helmreich & Pruzek, 2008)
		WhatIf (King & Zeng, 2006; King & Zeng, 2007)
		USPS (Obenchain, 2007)
Generalized boosted regression	*boost* (Schonlau, 2007)	*gbm* (McCaffrey, Ricgeway, & Morral, 2004)
Optimal maching (Rosenbaum, 2002a)		*optmatch* (Hansen, 2007)

	Procedure Name & Useful References	
Chapter & Methods	**Stata**	**R**
Post-matching covariance imbalance check (Haviland, Nagin, & Rosenbaum, 2007)	*imbalance* (Guo, 2008a)	
Hodges-Lehmann aligned rank test after optimal matching (Haviland, Nagin, & Rosenbaum, 2007; Lehmann, 2006)	*hl* (Guo, 2008b)	
Chapter 6		
Abadie & Imbens (2002, 2006) matching estimagors	*nnmatch* (Abadie,Drukker, Herr, & Imbens, 2004)	*Matching* (Sekehon, 2007)
Chapter 7		
Kernel-based maching (Heckman, Ichimura, & Todd, 1997, 1998)	*psmatch2* (Leuven & Sianesi, 2003)	
Chapter 8		
Rosenbaum's (2002a) sensitivity analysis	*rbounds* (Gangl, 2007)	*rbounds* (Keele, 2008)

Other Corrective Models

- Regression discontinuity designs
- Instrumental variables approaches (Guo & Fraser [2010] reviews this method)
- Interrupted time series designs
- Bayesian approaches to inference for average treatment effects

The Companion Website of Guo & Fraser (2010)

- All data and syntax files of the examples used in the book are available in the following website:

  http://ssw.unc.edu/psa/

4.Greedy propensity score matching (Rosenbaum & Rubin, 1983)

Rosenbaum and Rubin PSM (1)

Greedy matching

- Nearest neighbor: $C(P_i) = \min_j |P_i - P_j|, \quad j \in I_0$

  The nonparticipant with the value of P_j that is closest to P_i is selected as the match and A_i is a singleton set.
- Caliper: A variation of nearest neighbor: A match for person i is selected only if $|P_i - P_j| < \varepsilon, \; j \in I_0$

  where ε is a pre-specified tolerance. Recommended caliper size: $.25\sigma_p$
- 1-to-1 Nearest neighbor within caliper (The is a common practice)
- 1-to-n Nearest neighbor within caliper

Rosenbaum and Rubin PSM (2)

Mahalanobis metric matching:

- Mahalanobis without p-score: Randomly ordering subjects, calculate the distance between the first participant and all nonparticipants. The distance, d(i,j) can be defined by the Mahalanobis distance:

$$d(i, j) = (u - v)^T C^{-1} (u - v)$$

  where u and v are values of the matching variables for participant i and nonparticipant j, and C is the sample covariance matrix of the matching variables from the full set of nonparticipants.
- Mahalanobis metric matching with p-score added (to u and v).
- Nearest available Mahalandobis metric matching within calipers defined by the propensity score (need your own programming).

Rosenbaum and Rubin PSM (3)

Multivariate analysis at Step-3

One may perform routine multivariate analysis. These analyses may include:

- multiple regression
- generalized linear model
- survival analysis
- structural equation modeling with multiple-group comparison, and
- hierarchical linear modeling (HLM)

As usual, we use a dichotomous variable indicating treatment versus control in these models.

Sample Syntax Running Stata ***psmatch2*** for Greedy Matching

Specification of caliper size: = ***.25*SD***

```
// Nearest neighbor within caliper (.25*SD=.401)
psmatch2 aodserv, pscore(logit1) caliper(0.401) ///
    noreplacement descending
```

Program Name

Name of treatment variable

Name of the propensity score variable saved from logistic regression

Example of Greedy Matching (1)

Research Questions

The association between parental substance abuse and child welfare system involvement is well-known but little understood. This study aims to address the following questions: Whether or not these children are living in a safe environment? Does substance abuse treatment for caregivers affect the risk of child maltreatment re-report?

Example of Greedy Matching (2)

Data and Study Sample

- A secondary analysis of the National Survey of Child and Adolescent Well-Being (NSCAW) data.
- It employed NSCAW of two waves: baseline information between October 1999 and December 2000, and the 18-months follow-up. The sample for this study was limited to 2,758 children who lived at home (e.g., were not in foster care) and whose primary caregivers were female.

Example of Greedy Matching (3)

Measures

- The choice of explanatory variables (i.e., matching variables) in the logistic regression model is crucial. We chose these variables based on a review of substance abuse literature to determine what characteristics were associated with treatment receipt.
- We found that these characteristics fall into four categories: *demographic characteristics; risk factors; prior receipt of substance abuse treatment; and need for substance abuse services.*

Example of Greedy Matching (4)

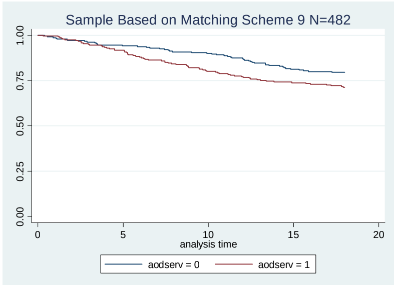
Analytic Plan*:*

- ***"3 x 2 x 2 design" = 12 Matching Schemes: Three*** logistic regression models (i.e., each specified a different set of matching variables); ***Two*** matching algorithms (i.e., nearest neighbor within caliper and Mahalanobis), and ***Two*** matching specifications (i.e., for nearest neighbor we used two different specifications on caliper size, and for Mahalanobis we used one with and one without propensity score as a covariate to calculate the Mahalanobis metric distances).
- Outcome analysis: survival model using Kaplan-Meier estimator evaluating difference in the survivor curve.

Example of Greedy Matching (5)

Findings:

- Children of substance abuser service users appear to live in an environment that elevates risk of maltreatment and warrants continued protective supervision.
- The analysis based on the original sample without controlling for heterogeneity of service receipt masked the fact that substance abuse treatment may be a marker for greater risk.



Example of Greedy Matching (6)

For more information about this example, see

- Guo & Fraser, 2010, pp.175-186.
- Guo, S., Barth, R.P., & Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review 28:* 357-383
- Barth, R.P., Gibbons, C., & Guo, S. (2006). Substance abuse treatment and the recurrence of maltreatment among caregivers with children living at home: A propensity score analysis. *Journal of Substance Abuse Treatment 30:* 93-104.

Example of Greedy Matching (7)

Additional examples of child welfare research employing greedy matching:

- Barth, R.P., Lee, C.K., Wildfire, J., & Guo, S. (2006). A comparison of the governmental costs of long-term foster care and adoption. *Social Service Review 80(1):* 127-158.
- Barth, R.P., Guo, S., McCrae, J. (2008). Propensity score matching strategies for evaluating the success of child and family service programs. *Research on Social Work Practice 18,* 212-222.
- Weigensberg, E.C.; Barth, R.P., Guo, S. (2009). Family group decision making: A propensity score analysis to evaluate child and family services at baseline and after 36-months. *Children and Youth Services Review, 31*, 383-390.

Contact Information

Ken DeCerchio, MSW, CAP
Program Director
Center for Children & Family Futures
Phone: (714) 505-3525
Cell: (850) 459-3329
E-mail: kdecerchio@cffutures.org
Website: www.ncsacw.samhsa.gov

Shenyang Guo, PhD
Professor of Social Work
University of North Carolina
Phone: (919) 843-2455
Email: sguo@email.unc.edu

46

QUESTIONS AND DISCUSSION

47

THANK YOU!

PLEASE TAKE A BRIEF MOMENT TO COMPLETE OUR EVALUATION.

YOU WILL BE RE-DIRECTED TO THE EVALUATION AFTER EXITING THIS WEBINAR.

48