

Introduction to Propensity Score Matching: A New Device for Program Evaluation

Workshop Presented at the Annual Conference of the
Society for Social Work Research
New Orleans, January, 2004

Shenyang Guo, Ph.D.¹, Richard Barth, Ph.D.¹, and Claire Gibbons, MPH²
Schools of Social Work¹ and Public Health²
University of North Carolina at Chapel Hill

NSCAW data used to illustrate PSM were collected under funding by the Administration on Children, Youth, and Families of the U.S. Department of Health and Human Services. Findings do not represent the official position or policies of the U.S. DHHS. PSM analyses were partially funded by the Robert Wood Johnson Foundation and the Childrens Bureau's Child Welfare Research Fellowship. **Results are preliminary and not quotable.** Contact information: sguo@email.unc.edu

Outline

- Overview: Why Propensity Score Matching?
- Highlights of the key features of PSM
- Example: Does substance abuse treatment reduce the likelihood of child maltreatment re-report?

Why Propensity Score Matching?

- Theory of Counterfactuals
 - The fact is that some people receive treatment.
 - The counterfactual question is: “What would have happened to those who, in fact, did receive treatment, if they had not received treatment (or the converse)?”
 - Counterfactuals cannot be seen or heard—we can only create an estimate of them.
 - PSM is one “correction strategy” that corrects for the selection biases in making estimates.

Approximating Counterfactuals

- A range of flawed methods have long been available to us:
 - RCTs
 - Quasi-experimental designs
 - Matching on single characteristics that distinguish treatment and control groups (to try to make them more alike)

Limitations of Random Assignment

- Large RCTs take a long time and great cost to generate answers—analysis of existing data may more timely, yet acceptably accurate
- RCTs are not feasible when variables cannot be manipulated—e.g., some events in child welfare are driven by legal mandates
- Prior analysis of the need for withholding treatment should be done before RCTs are deemed necessary.

Limitations of Quasi-Experimental Designs

- Selection bias may be substantial
- Comparison groups used to make counterfactual claims may have warped counters and failing factials, leading to intolerably ambiguous findings

Limitations of Matching

- If the two groups do not have substantial overlap, then substantial error may be introduced:
 - E.g., if only the worst cases from the untreated “comparison” group are compared to only the best cases from the treatment group, the result may be regression toward the mean
 - makes the comparison group look better
 - Makes the treatment group look worse.

Propensity Score Matching

- Employs a predicted probability of group membership—e.g., treatment vs. control group-- based on observed predictors, usually obtained from logistic regression to create a counterfactual group
- Propensity scores may be used for matching or as covariates—alone or with other matching variables or covariates.

PSM Has Many Parents

- In 1983, **Rosenbaum and Rubin** published their seminal paper that first proposed this approach.
- From the 1970s, **Heckman** and his colleagues focused on the problem of selection biases, and traditional approaches to program evaluation, including randomized experiments, classical matching, and statistical controls. Heckman later developed “Difference-in-differences” method

PSM Has Skeptics, Too

Howard Bloom, MDRC

- Sees PSM as a somewhat improved version of simple matching, but with many of the same limitations
- Inclusion of propensity scores can help reduce large biases, but significant biases may remain
- Local comparison groups are best—PSM is no miracle maker (it cannot match unmeasured contextual variables)
- Short-term biases (2 years) are substantially less than medium term (3 to 5 year) biases—the value of comparison groups may deteriorate

Michael Sosin, University of Chicago

- Strong assumption that untreated cases were not treated at random
- Argues for using multiple methods and not relying on PSM

Limitations of Propensity Scores

- Large samples are required
- Group overlap must be substantial
- Hidden bias may remain because matching only controls for observed variables (to the extent that they are perfectly measured)

(Shadish, Cook, & Campbell, 2002)

Criteria for “Good” PSM

- Identify treatment and comparison groups with substantial overlap
- Match, as much as possible, on variables that are precisely measured and stable (to avoid extreme baseline scores that will regress toward the mean)
- Use a composite variable—e.g., a propensity score—which minimizes group differences across many scores

Risks of PSM

- They may undermine the argument for experimental designs—an argument that is hard enough to make, now
- They may be used to act “as if” a panel survey is an experimental design, overestimating the certainty of findings based on the PSM.

A Methodological Overview

- Reference list
- The crucial difference of PSM from conventional matching: match subjects on **one score** rather than **multiple variables**: “... the propensity score is a monotone function of the discriminant score” (Rosenbaum & Rubin, 1984).
- Continuum of complexity of matching algorithms
- Computational software
 - STATA – PSMATCH2
 - SAS SUGI 214-26 “GREEDY” Macro
 - S-Plus with FORTRAN Routine for difference-in-differences (Petra Todd)

General Procedure

Run Logistic Regression:

- Dependent variable: $Y=1$, if participate; $Y = 0$, otherwise.
- Choose appropriate conditioning (instrumental) variables.
- Obtain propensity score: predicted probability (p) or $\log[p/(1-p)]$.

Match Each Participant to One or More Nonparticipants on Propensity Score

- Nearest neighbor matching
- Caliper matching
- Mahalanobis metric matching in conjunction with PSM
- Stratification matching
- Difference-in-differences matching (kernel & local linear weights)

Multivariate analysis based on new sample

Nearest neighbor and caliper matching

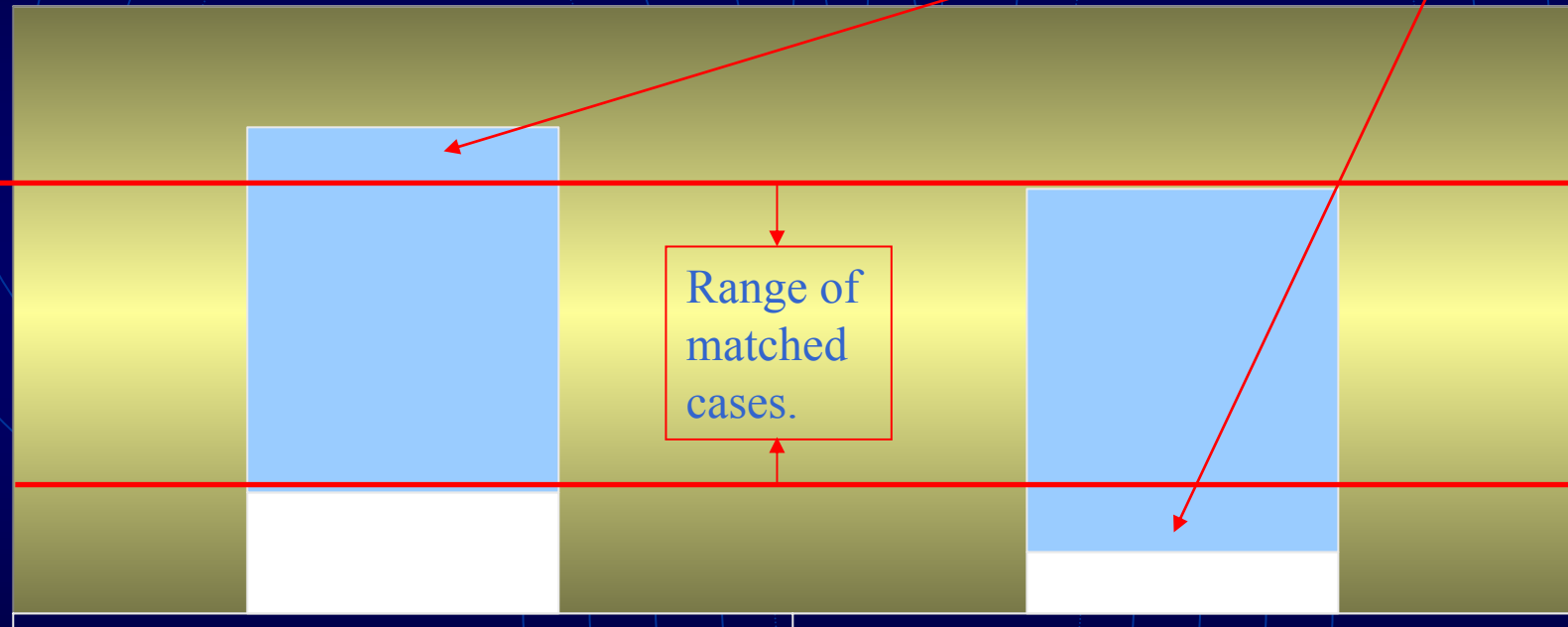
- **Nearest neighbor:** Randomly order the participants and nonparticipants, then select the first participant and find the nonparticipant with closest propensity score.
- **Caliper:** define a common-support region (e.g., .01 to .00001), and randomly select one nonparticipant that matches on the propensity score with the participant. SAS macro “GREEDY” does this.

Problem 1: Incomplete Matching or Inexact Matching?

- While trying to maximize exact matches (i.e., strictly “nearest” or narrow down the common-support region), cases may be excluded due to incomplete matching.
- While trying to maximize cases (i.e., widen the region), inexact matching may result.

Problem 2: Cases Are Excluded at Both Ends of the Propensity Score

Cases excluded



Participants

Nonparticipants

■ Predicted Probability

Mahalanobis Metric Matching: A Conventional Method

- Use this method to choose one nonparticipant from multiple matches.
- Procedure:
 - 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;

- The nonparticipant, j , with the minimum distance $d(i,j)$ is chosen as the match for participant i , and both are removed from the pool;
- Repeat the above process until matches are found for all participants.

Mahalanobis in Conjunction with PSM

Mahalanobis metric matching is a conventional method. However, the literature suggests two advanced methods that combine the Mahalanobis method with the propensity score matching: (1) Mahalanobis metric matching including the propensity score, and (2) Nearest available Mahalanobis metric matching within calipers defined by the propensity score.

Stratification

One of several methods developed for missing data imputation

1. Group sample into five categories based on propensity score (quintiles).
2. Within each quintile, there are r participants and n nonparticipants. Use “approximate Bayesian bootstrap” method to conduct matching or resampling.

Heckman's Difference-in-Differences Matching Estimator (1)

Fundamental difference

(i.e., counterfactual or program effect) one attempts to estimate. It holds *only* when each participant matches to one nonparticipant.

$$D_{t,t'}(X) = \underbrace{E(Y_{1t} - Y_{0t'} | X, D = 1)}_{\text{Participants' before-after difference}} - \underbrace{E(Y_{0t} - Y_{0t'} | X, D = 0)}_{\text{Nonparticipants' before-after difference}}$$

Participants' before-after difference:

Average differences in outcome Y for participants with characteristics X between pre-intervention (t') and post-intervention (t).

Nonparticipants' before-after difference:

Sample average outcome differences for nonparticipants with characteristics X between times t' and t.

Heckman's Difference-in-Differences Matching Estimator (2)

Difference-in-differences

Applies when each participant matches to *multiple nonparticipants*.

Weight
(see the following two slides)

$$\hat{\alpha}_{KDM} = \frac{1}{n_1} \sum_{i \in I_1 \cap S_p} \{ (Y_{1ti} - Y_{0ti}) - \sum_{j \in I_0 \cap S_p} W(i, j) (Y_{0tj} - Y_{0t'j}) \}$$

Total number of participants

Participant i in the set of common-support.

Multiple nonparticipants who are in the set of common-support (matched to i).

Difference

.....in.....

Differences

Heckman's Difference-in-Differences Matching Estimator (3)

Weights $W(i,j)$ (distance between i and j) can be determined by using one of two methods:

1. Kernel matching:

$$W(i, j) = \frac{G\left(\frac{P_j - P_i}{\alpha_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{\alpha_n}\right)}$$

where $G(\cdot)$ is a kernel function and α_n is a bandwidth parameter.

Heckman's Difference-in-Differences Matching Estimator (4)

2. Local linear weighting function:

$$W(i, j) = \frac{G_{ij} \sum_{k \in I_0} G_{ik} (P_k - P_i)^2 - [G_{ij} (P_j - P_i)] \left[\sum_{k \in I_0} G_{ik} (P_k - P_i) \right]}{\sum_{j \in I_0} G_{ij} \sum_{k \in I_0} G_{ij} (P_k - P_i)^2 - \left(\sum_{k \in I_0} G_{ik} (P_k - P_i) \right)^2}$$

Heckman's Difference-in-Differences Matching Estimator (5)

A Summary of Procedures to Implement Heckman's Approach:

1. Obtain propensity score;
2. For each participant, identify all nonparticipants who match on the propensity score (i.e., determine common-support set);
3. Calculate before-after difference for each participant;
4. Calculate before-after differences for multiple nonparticipants using kernel weights or local linear weights;
5. Evaluate difference-in-differences.

Heckman's Contributions to PSM (In Our Opinion)

- Unlike traditional matching, his estimator requires the use of longitudinal data, that is, outcomes before and after intervention;
- His estimator employs recent advances in matching (kernel and local linear weights);
- By doing this, the estimator is more robust: it eliminates temporarily-invariant sources of bias that may arise, when program participants and nonparticipants are geographically mismatched or from differences in survey questionnaire.

Illustrating Example

The Likely Impact of Substance Abuse Treatment on Re-abuse Reporting Among CWS Involved Families

Collaboration:

- NSCAW data
- Robert Wood Johnson support, under SAPRP, for analysis
- Children's Bureau Faculty Fellowship Award to Shenyang Guo supports development of workshop and website materials

<http://sswnt5.sowo.unc.edu/VRC/Lectures/index.htm>

Research Question, Data, and Challenge

- **Research Question:** Does substance abuse treatment reduce the likelihood of re-reports over an 18 month follow-up period?
- **Data:** National Survey of Child and Adolescent Well-being (NSCAW).
 - National probability sample of CWS cases, limited to “in-home” cases where the primary caregiver is female (90% of all CWS cases)
- **Challenge:** Selection bias: how can we address the concern that cases that did not get substance abuse treatment (SAT) did not need it?

Define AOD Treatment[^]

- Caregiver report:
 - Currently receiving any type of treatment for AOD problem
 - Admitted to hospital for AOD problem
 - Stayed overnight in program for AOD problem
 - Went to ER for AOD problem
 - Visited clinic/doctor for AOD problem
- CWW report:
 - Received service for AOD problem once referred

[^]During first 12 months following this CWS investigation

Sample

| | | |
|--|-----------|---------|
| Female, in-home caregivers | 3669 | |
| | Unmatched | Matched |
| With non-missing AOD services variable | 2758 | 574 |
| Final regression predicting re-reports | 2529 | 520 |

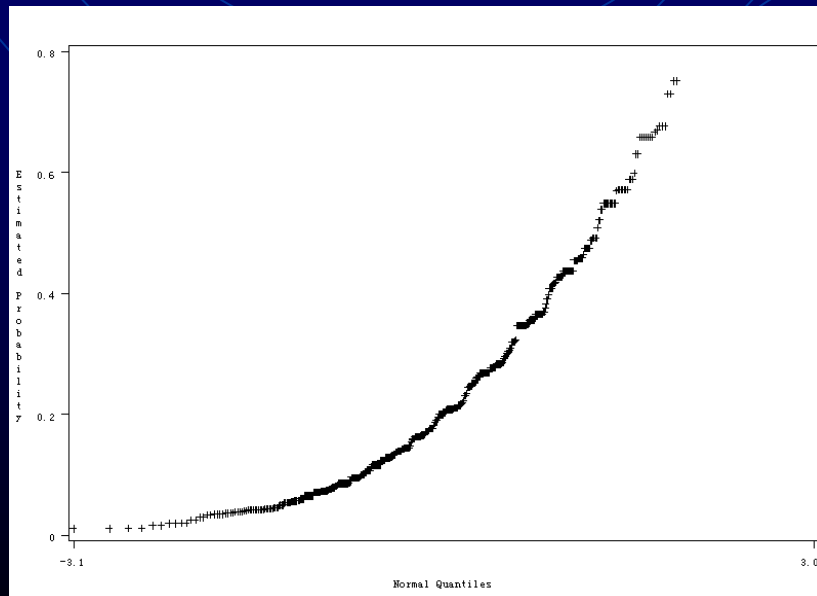
Identify Variables With Likely Linkage to Substance Abuse Treatment Use

- Marital status
- Education
- Poverty
- Employment
- Closed/open
- Child race/ethnicity
- Child age
- Caregiver age
- Trouble paying for basic necessities
- Caregiver mental health
- Caregiver arrest
- Prior AOD treatment
- Maltreatment type

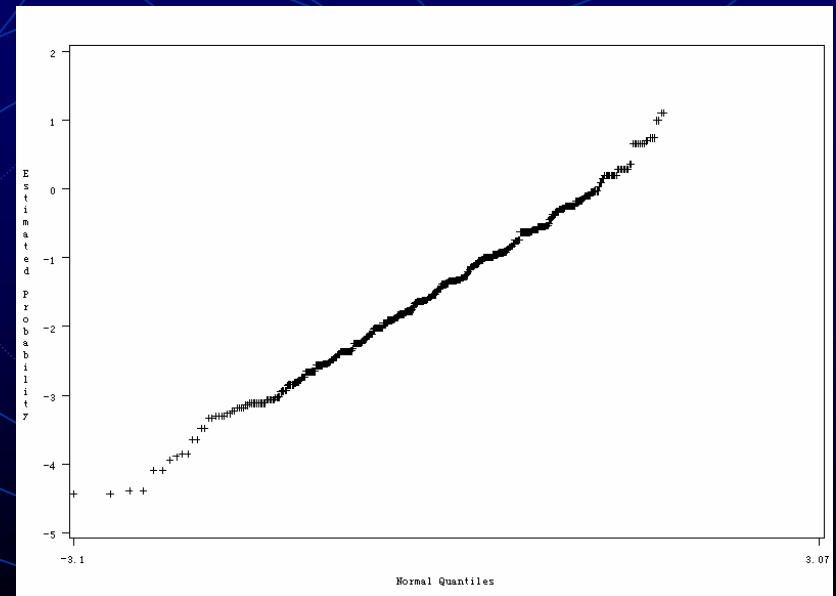
Generate & Transform Propensity Scores

1. Logistic regression to generate predicted probabilities
2. Q-Q Plots Testing Normality of Predicted Probability and Predicted Logit

Predicted Probability



Predicted logit -- $\log [p/(1-p)]$



Matching of Predicted Probabilities

- We employed the **caliper matching** method to match the sample of AOD service users (n=298) to the sample of AOD service nonusers (n=2,460) based on the predicted logit.
- Software: SAS macro “GREEDY”.

Sample Differences Before and After Matching

- Before matching (n=2,758: 2,460 AOD service users and 298 nonusers), **all** 13 variables except marital status and caregiver age are statistically significant.
- After matching (n=520: 260 AOD service users and 260 nonusers), only **two** variables (education & poverty) remain significant.
- This indicates that users and nonusers in the new sample share almost exactly the same characteristics, and selection bias has been mitigated in the new sample.

Second Stage Analysis: What Variables Predict Likelihood of Re-report?

- Substance abuse service receipt (Y/N)
- Child age
- Caregiver age
- Prior child welfare service receipt
- Caregiver mental health problem
- Number of children
- Trouble paying for basic necessities
- Active domestic violence
- Open/closed
- Receipt of welfare (e.g., TANF)
- Child has major special needs or behavior problems

Significant Predictors of Re-reports, Unmatched Sample (n=2,758)

• Unweighted

- AOD services (.67)**
- Prior CWS (.67)**
- CG MH problem (.78)**
- Trouble paying for basic necessities (.53)**
- Welfare receipt (.76)*
- Child has major special needs or behavior problems (.75)**

• Weighted

- AOD services (3.24)*
- Child has major special needs or behavior problems (1.99)*

Significant Predictors of Re-reports, with PSM (n=520)

- Unweighted
 - Prior CWS (.60)*
- Weighted
 - Prior CWS (3.06)**
 - Welfare receipt (3.6)*

Finding:

Once selection bias is mitigated through matching, substance abuse services in the first 12 months are not significantly associated with the likelihood of re-reports over 18 months in the weighted and unweighted data.

*p<.05, **p<.01

Interpretation Issues

- **Weighted or Unweighted Data**
 - Weights are no longer correct after resampling
 - Unweighted data does not reflect the population
- **Meaning of other coefficients in model**
 - PSM would need to be conducted to resample to test each other intervention
- **Generalizability to the entire population**
 - Excluded cases are different than PSM cases—some of these cases might benefit from the intervention

Potential Areas of Application:

- Use national sample as a benchmark, greatly reduce cost of evaluation of new intervention program
- Better model causal effect heterogeneity
- Missing data imputation

Possible Applications of PSM to Social Work Evaluation

- When designing a new intervention, one may only create a treatment group (i.e., no randomized control group), carefully select a national sample (i.e., NSCAW, AHEAD, PSID), and then use PSM to match the treatment sample to the national sample to assess impact of intervention.
- Using any existing survey data (e.g., within NSCAW), one may use PSM to better evaluate the heterogeneity of causal effects, for example, the impact of parental use of substance-abuse services on children's well-being or outcomes.

A Paradigm Shift in Program Evaluation Implications of PSM

- Problems and biases in the case of social experiment
- Selection bias: self-selection, bureaucratic selection: whether or not can randomization control for these biases?
- A criticism to all conventional methods:
 - ❑ No randomized control;
 - ❑ No simple matching;
 - ❑ No simple statistical control;
- A paradigm shift in evaluation of counterfactuals!!

Thank You Very Much

Questions?

