Propensity Score Analysis and Strategies for Its Application to Services Training Evaluation

Shenyang Guo, Ph.D.
School of Social Work
University of North Carolina at Chapel Hill
June 14, 2011

For the 14th Annual National Human Services Training Evaluation Symposium
Cornell University
Overview

1. Types of evaluation problems
2. Why and when propensity score analysis is needed
3. Conceptual frameworks and assumptions
4. Overview of corrective methods
5. Greedy matching
6. Strategies to apply PSA to human services training evaluation
1. Types of evaluation problems
Three Different Types of Evaluation Problems

1. Policy evaluation
2. Program evaluation
3. Training evaluation
Heckman (2005, pp. 7-9) distinguished three broad classes of policy evaluation questions:

1. evaluating the impact of previous interventions on outcomes (internal validity);
2. forecasting the impacts of interventions implemented in one environment in other environments (external validity); and
3. forecasting the impacts of interventions never historically experienced for other environments (using history to forecast the consequences of new policies).
Program 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?*
Program Evaluation: 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, testing effects, and more.
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).


Randomized clinical trial is a gold standard.
Selection Bias in Evaluation of Quasi-Experimental Programs

Total Sample

Individual decision to participate

Administrator’s decision to select

Control group

Drop out

Continue

Treatment group

Drop out

Continue

Individual Decision not to participate in experiment

Administrator’s decision not to select

Source: Maddala, 1983, p. 266
Training Evaluation

Cindy Parry & Robin Leak:

- Training evaluation concerns the extent to which knowledge, attitudes and skills learned in a training context are possibly applied to the job.
- Evaluators aims to evaluate the effectiveness of transfer of learning (TOF), that is, how learned behaviors are generalized to the job context AND maintained over a period of time.

In my view, if we view training is an intervention, then training evaluation is analogous to program evaluation where randomization is infeasible. Under this context, PSA is applicable.
2. Why and when propensity score analysis is needed
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 (\( \tau \)) 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 (\( \tau \)) 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 may 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
3. 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 \(E(Y_0\mid W=0)\) really represents \(E(Y_0\mid W=1)\)?

In a series of papers, Heckman and colleagues criticized this assumption.

Consider \(E(Y_1\mid W=1) - E(Y_0\mid W=0)\). Add and subtract \(E(Y_0\mid W=1)\), we have

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

The standard estimator provides unbiased estimation if and only if \(E(Y_0\mid W=1) = E(Y_0\mid W=0)\).

In many empirical projects, \(E(Y_0\mid W=1) \neq E(Y_0\mid W=0)\).
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

<table>
<thead>
<tr>
<th>Sources of randomness</th>
<th>Statistical Causal Model</th>
<th>Econometric Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implicit</td>
<td>Explicit</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Models of conditional counterfactuals</th>
<th>Statistical Causal Model</th>
<th>Econometric Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implicit</td>
<td>Explicit</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mechanism of intervention for determining counterfactuals</th>
<th>Statistical Causal Model</th>
<th>Econometric Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothetical randomization</td>
<td>Many mechanisms of hypothetical interventions including randomization; mechanism is explicitly modeled</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment of interdependence</th>
<th>Statistical Causal Model</th>
<th>Econometric Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recursive</td>
<td>Recursive or Simultaneous systems</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Social/market interactions</th>
<th>Statistical Causal Model</th>
<th>Econometric Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ignored</td>
<td>Modeled in general Equilibrium frameworks</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Projections to different populations?</th>
<th>Statistical Causal Model</th>
<th>Econometric Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does not project</td>
<td>Projects</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parametric?</th>
<th>Statistical Causal Model</th>
<th>Econometric Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonparametric</td>
<td>Becoming nonparametric</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Range of questions answered</th>
<th>Statistical Causal Model</th>
<th>Econometric Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>One focused treatment effect</td>
<td>In principle, answers many possible questions</td>
<td></td>
</tr>
</tbody>
</table>

Source: Heckman (2005, p.87)
4. Overview of Corrective Methods
Four Models Described by Guo & Fraser (2010)

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)

Four Models Described by Guo & Fraser (2010)

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: 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 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 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)

<table>
<thead>
<tr>
<th>Procedure Name &amp; Useful References</th>
<th>Chapter &amp; Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chapter 4</strong></td>
<td></td>
</tr>
<tr>
<td>Heckman (StataCorp, 2003)</td>
<td>sampleSelection</td>
</tr>
<tr>
<td>treatreg (StataCorp, 2003)</td>
<td></td>
</tr>
<tr>
<td><strong>Chapter 5</strong></td>
<td></td>
</tr>
<tr>
<td>Rosenbaum &amp; Rubin's (1983) propensity score matching</td>
<td></td>
</tr>
<tr>
<td>psmatch2 (Leuven &amp; Sianesi, 2003)</td>
<td>cem (Deheja &amp; Wahba, 1999; Iacus, King, &amp; Porro, 2008)</td>
</tr>
<tr>
<td></td>
<td>Matching (Sekehon, 2007)</td>
</tr>
<tr>
<td></td>
<td>Matchit (Ho, Imai, King, &amp; Stuart, 2004)</td>
</tr>
<tr>
<td></td>
<td>PSAgraphics (Helmreich &amp; Pruzek, 2008)</td>
</tr>
<tr>
<td></td>
<td>WhatIf (King &amp; Zeng, 2006; King &amp; Zeng, 2007)</td>
</tr>
<tr>
<td></td>
<td>USPS (Obenchain, 2007)</td>
</tr>
<tr>
<td></td>
<td>gbm (McCaffrey, Ricgeway, &amp; Morral, 2004)</td>
</tr>
<tr>
<td>Generalized boosted regression</td>
<td>boost (Schonlau, 2007)</td>
</tr>
<tr>
<td>Optimal maching (Rosenbaum, 2002a)</td>
<td>optmatch (Hansen, 2007)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Procedure Name &amp; Useful References</th>
<th>Chapter &amp; Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chapter 6</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Chapter 7</strong></td>
<td></td>
</tr>
<tr>
<td>Kernel-based machining</td>
<td></td>
</tr>
<tr>
<td>psmatch2 (Leuven &amp; Sianesi, 2003)</td>
<td></td>
</tr>
<tr>
<td><strong>Chapter 8</strong></td>
<td></td>
</tr>
<tr>
<td>Rosenbaum's (2002a) sensitivity analysis</td>
<td></td>
</tr>
<tr>
<td>rbounds (Gangl, 2007)</td>
<td></td>
</tr>
<tr>
<td>rbounds (Keele, 2008)</td>
<td></td>
</tr>
</tbody>
</table>
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/
5. Greedy propensity score matching (Rosenbaum & Rubin, 1983)
Greedy matching

- Nearest neighbor: \( C(P_i) = \min_j |P_i - P_j|, \ j \in I_0 \)
  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 \( \varepsilon \) 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
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 Mahalanobis metric matching within calipers defined by the propensity score (need your own programming).
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

\texttt{psmatch2} for Greedy Matching

\begin{verbatim}
// Nearest neighbor within caliper (.25*SD=.401)
psmatch2 aodserv, pscore(logit1) caliper(0.401) ///
noreplacement descending
\end{verbatim}

Program Name

Name of treatment variable

Specification of caliper size: = .25*SD

Name of the propensity score variable saved from logistic regression
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.
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.
Example of Greedy Matching (7)

Additional examples of child welfare research employing greedy matching:

6. Strategies to apply PSA to human services training evaluation
Strategy I: Yoke Comparison Group Using National Survey Data (1)

- RCT or experimental approach
- Nonexperimental approach
- Yoke training group with a comparison group from a national representative survey
- A small to large match

A randomized clinical controlled trial

![Diagram showing treated and control groups with arrows indicating the comparison group and survey or panel study.](diagram.png)
Strategy I: Yoke Comparison Group Using National Survey Data (2)

An example:

- Glisson, Dukes, & Green (2006) employs randomized blocks to evaluate the effects of a one-year training (i.e., the Availability, Responsiveness, and Continuity [ARC] organizational intervention).

- If one replicates the ARC study but only has training data for the intervention group, one can employs the National Survey of Child and Adolescent Well-Being (NSCAW) as a big pool to form a comparison group. That is, no additional effort is required to collect data from a control group.
Strategy I: Yoke Comparison Group Using National Survey Data (3)

- Key requirement: both the training and national survey datasets should contain common outcome variables and key matching variables.

- The ARC study aims to implement an organizational intervention to reduce casework turnover, and to create positive organizational climates and culture in a child welfare and juvenile justice system.

- NSCAW is a national representative panel data designed to address a range of questions about the outcomes of children who are involved in child welfare system across the country. It now has two cohorts of study children from 92 counties in 36 states.
Strategy I: Yoke Comparison Group Using National Survey Data (4)

- NSCAW contain many outcome variables developed by Glisson to measure organizational climates and culture. To evaluate the ARC training effects, evaluators should collect these outcome variables from their own training group. This is often not a problem.

- To make a valid evaluation, researchers should make sure that both the training data and NSCAW data contain key variables to control for selection bias (e.g., urban/rural, caseload, type of maltreatments, etc.).

- Evaluators then use PSA to create a comparison group from NSCAW, and evaluate the outcome differences between the training and comparison groups.

- The net difference on outcomes after PSA may be a better measure of transfer of learning.
Strategy II: Use Data from any Nonequivalent Comparison Group

- The same strategy can be applied to a comparison between the training group and any nonequivalent comparison group that is readily available to the evaluator.
- These nonequivalent comparison groups may be agencies in similar counties or states that do not receive a targeted training.
- Again, the key requirement is that both the training and comparison groups should contain key variables on outcomes as well as matching variables.
- PSA is a strategy to make the two groups as comparable as possible (i.e., free of selection bias on the observed measures) so that researchers can discern the effectiveness of transfer of learning.
References (1)


References (2)


Thank you!

Questions and Discussion