Quasi-Experimental Strategies When Randomization Is Not Feasible: Propensity Score Matching

For presentation at the 9th Annual Child Welfare Waiver Demonstration Projects Meeting, Washington DC, June 9, 2005

Shenyang Guo
Judy Wildfire
School of Social Work
University of North Carolina at Chapel Hill
Acknowledgements

Support for this research is from two sources. The Children’s Bureau provided a grant for the development of innovative quantitative methods for child welfare research (Shenyang Guo, PI). The NC Dept of Health and Human Services is supporting an evaluation of NC’s Title IV-E Waiver Demonstration project (Judy Wildfire, PI). We thank Rick Barth, Mark Fraser, and Lynn Usher for comments and suggestions on this presentation.
Presentation Overview

- The challenges: situations under which randomized experiments are infeasible
- The counterfactual framework
- The fundamental assumption embedded in all evaluations
- The propensity score matching (PSM) approach
- An illustration
- Analytic plan for the evaluation of the NC Waiver Demonstration project
A PSM Resource Package

- “PSM References” compiled by Guo, Barth, and Gibbons (2003).


Limitations of Randomization


- The case of educational study: evaluation of the Catholic school effect vis-à-vis the public school effect on learning. Morgan (2001) found that the Catholic school effect is the strongest among those Catholic school students who are less likely to attend Catholic schools.
The Counterfactual Framework

- **Counterfactual**: what would have been the outcomes for persons being served had they not been served?

- The counterfactual framework was developed by Neyman (1923) and Rubin (1978).

- The key assumption of the counterfactual framework is that individuals selected into experimental groups have potential outcomes in both states: the one in which they are observed and the one in which they are not observed. A rigorous evaluation design should aim to accomplish a robust estimation of counterfactual.
The Fundamental Assumption

- Rosenbaum & Rubin (1983): \((Y_0, Y_1) \perp W \mid X\).

  Conditional on observed covariates, the service assignment should not correlate with the outcome under either service or control condition.

Violation of the Assumption Is Equivalent to Ignoring Threats to Internal Validity

- Internal validity relates to questions about ascertaining whether the intervention caused the changes in the target problem (Campbell & Stanley, 1963; Cook & Campbell 1979).

- Key among the nine threats outlined by Campbell & Stanley is selection.
Violation of the Assumption Also Is Equivalent to Violating An OLS Assumption

- An ordinary least square (OLS) regression model assumes no correlation of the error term with the explanatory variables. The assumption is often violated when researchers evaluate service effects from observational data.

- Consider the model: $Y_i = \alpha + \tau W_i + X_i' \beta + \varepsilon_i$

  The independent variable $w$ is usually correlated with the error term $\varepsilon$. The consequence is inconsistent and biased estimate about the treatment effect $\tau$. 
Development and Application of PSM: An Overview

- **Who?**
  - Heckman (1978, 1979)
  - Rosenbaum and Rubin (1983)

- **When is PSM appropriate?**
  - (1) Analyzing treatment effects from observational data;
  - (2) Evaluating effectiveness of intervention with a quasi-experimental design.

- **Software packages:** R; PSMATCH2 of Stata.
Run Logistic Regression:

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

1-to-1 or 1-to-n Match

- Nearest neighbor matching
- Caliper matching
- Mahalanobis
- Mahalanobis with propensity score added

Either

Multivariate analysis based on new sample
The fundamental challenge is that under welfare reform, counties and states have more freedom to choose their own policies.

It is a system reform initiative that combines individual-level, agency-level, and community strategies to change outcomes for children in child welfare. An evaluation must consider combined effect of all of these in determining the impact of the waiver.

Counties are allowed flexibility to define own “waiver” interventions that fit the needs and environment of the population that they serve. There is no one “silver bullet” being implemented across the state.
Evaluation Challenges Posed by Title IV-E Waiver Demonstration Projects in NC (2)

- It is not administratively feasible to assign individual families or children to distinct service conditions because the changes in approach to service reflect a fundamental shift in philosophy.

- Group randomization (Bloom and Raudenbush, 2004) is not feasible because participation reflects a voluntary commitment to reform.

- Counties participating in the Demonstration necessarily and desirably reflect a “selection bias”. 
An Illustration

A Monte Carlo Study
Simulating the Evaluation
of the Title IV-E Waiver
Demonstration Project in NC
Objective of the Monte Carlo Study

- To simulate the settings typically found in the Title IV-E Waiver Demonstration project. Through data simulations, we compare PSM to a “quasi-random” method. Because the population parameters are known in advance, we will show via the study the bias and efficiency of estimation associated with each method.
The Hypothetical Population

- Suppose in a hypothetical state, there are 20 counties participating in Waiver Demonstration, and another 20 counties forming a comparison group. The intervention is designed to reduce the number of placements among children with a substantiated report.

- Assume further: in each county, there are 1,000 children who meet the study criteria and eligible for the study. That is, the population $N = 1,000 \times 20 + 1,000 \times 20 = 40,000$.

- Assume further: the Waiver Counties have *lower* average income, and their children have *more severe* behavioral problems than the Comparison Counties (selection bias is built in).
Assume further: the intervention has a small effect (i.e., the hazard rate of placement since the first substantiated report for the Waiver group is 30% lower than that of the Comparison group).

Assume further: due to cost constraints, we cannot evaluate the outcomes for the entire population. Instead, we must create a sample comprising 2,000 Waiver children and 2,000 comparison children.
Various Types of Research and Sampling Designs

- **Pure random**: create a random sample of 4,000, and assign 2,000 to Waiver and 2,000 to Comparison purely randomly. This method is infeasible due to spillover effects.

- **Quasi-random**: draw a random sample of 2,000 from the 20,000 Waiver subjects, and a random sample of 2,000 from the 20,000 Comparison subjects.

- **PSM**: draw a random sample of 2,000 from the 20,000 Waiver subjects, and to find 2,000 best matches on propensity scores from the Comparison group.
Research Question for the Monte Carlo Study

Between the two approaches (i.e., quasi-random and PSM), which design is better?
### Descriptive Statistics of the Population

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Waiver (N=20000)</th>
<th>Comparison (N=20000)</th>
<th>P-value of t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>X1(Income)</td>
<td>96.19</td>
<td>14.61</td>
<td>100.44</td>
</tr>
<tr>
<td>X2(CBCL)</td>
<td>49.83</td>
<td>10.74</td>
<td>44.41</td>
</tr>
<tr>
<td>X3</td>
<td>.002</td>
<td>.99</td>
<td>.000</td>
</tr>
<tr>
<td>X4</td>
<td>69.13</td>
<td>2.85</td>
<td>69.15</td>
</tr>
<tr>
<td>X5</td>
<td>.77</td>
<td>.70</td>
<td>.77</td>
</tr>
<tr>
<td>X6</td>
<td>-.007</td>
<td>1.00</td>
<td>.003</td>
</tr>
</tbody>
</table>

### Correlations

<table>
<thead>
<tr>
<th>Waiver (Dummy)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>-.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>.25</td>
<td>-.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>.00</td>
<td>.15</td>
<td>.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td>-.00</td>
<td>-.00</td>
<td>-.00</td>
<td>-.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X5</td>
<td>.00</td>
<td>.01</td>
<td>-.01</td>
<td>.01</td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>X6</td>
<td>-.01</td>
<td>-.00</td>
<td>-.00</td>
<td>.01</td>
<td>-.01</td>
<td>.01</td>
</tr>
</tbody>
</table>
Estimated Survival Functions in Population: Small Effect Size of Waiver Intervention
Finding 1: The quasi-random approach violates the fundamental assumption, but PSM does not

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>X1(Income)</td>
<td>96.14</td>
<td>15.05</td>
<td>100.43</td>
</tr>
<tr>
<td>X2(CBCL)</td>
<td>49.91</td>
<td>10.51</td>
<td>44.42</td>
</tr>
<tr>
<td>X3</td>
<td>0.007</td>
<td>.99</td>
<td>-0.034</td>
</tr>
<tr>
<td>X4</td>
<td>69.09</td>
<td>2.83</td>
<td>69.11</td>
</tr>
<tr>
<td>X5</td>
<td>.75</td>
<td>.68</td>
<td>.79</td>
</tr>
<tr>
<td>X6</td>
<td>-.019</td>
<td>.99</td>
<td>.024</td>
</tr>
</tbody>
</table>

**Quasi-Random:** This pattern is shown by all 100 simulations.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>X1(Income)</td>
<td>96.49</td>
<td>14.71</td>
<td>96.48</td>
</tr>
<tr>
<td>X2(CBCL)</td>
<td>49.67</td>
<td>10.26</td>
<td>49.71</td>
</tr>
<tr>
<td>X3</td>
<td>0.026</td>
<td>1.011</td>
<td>0.038</td>
</tr>
<tr>
<td>X4</td>
<td>69.10</td>
<td>2.89</td>
<td>69.19</td>
</tr>
<tr>
<td>X5</td>
<td>.77</td>
<td>.71</td>
<td>.77</td>
</tr>
<tr>
<td>X6</td>
<td>-.011</td>
<td>1.006</td>
<td>.003</td>
</tr>
</tbody>
</table>

**PSM:**
This pattern is shown by all 100 simulations.

PSM used all 6 variables as matching variables.
Finding 2: Based on Cox regression analysis, PSM yields less error than the quasi-random approach (100 simulations using all predictors)

<table>
<thead>
<tr>
<th>Effect</th>
<th>MEAN PSM</th>
<th>Error PSM</th>
<th>SD PSM</th>
<th>MEAN RANDOM</th>
<th>Error RANDOM</th>
<th>SD RANDOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect_TRUE</td>
<td>-.37</td>
<td>-.369500</td>
<td>-.0005</td>
<td>-.371200</td>
<td>+.0012</td>
<td>.030424</td>
</tr>
</tbody>
</table>

Full Model via A PSM Procedure

Full Model via A Random Procedure
Finding 3: Even if a Cox model is not correctly specified (INCOME is dropped), PSM is more robust (100 Simulations)

<table>
<thead>
<tr>
<th>Effect</th>
<th>MEAN PSM</th>
<th>Error PSM</th>
<th>MEAN RANDOM</th>
<th>Error RANDOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE</td>
<td>-.37</td>
<td>-.00054</td>
<td>-.370560</td>
<td>+ .00056</td>
</tr>
<tr>
<td>PSM</td>
<td>-.369460</td>
<td>-.028394</td>
<td>.029753</td>
<td></td>
</tr>
<tr>
<td>RANDOM</td>
<td>-.370560</td>
<td>+ .00056</td>
<td>.029753</td>
<td></td>
</tr>
</tbody>
</table>
Limitations

- Propensity scores cannot adjust for unobserved covariates.
- PSM works better in larger samples.
- PSM can handle a covariate that is related to the service condition, but not to the outcome in the same way as a covariate with the same relation to the service condition, but strongly related to outcome (Rubin 1997).
Continuing Debate Among Researchers

- Much skepticism that PSM or any other nonexperimental approach can reliably simulate, much less, replace randomized experiments.


Context of the Debate

- Debate about PSM being better than a randomized experiment is moot with regard to evaluating Waiver Demonstrations in which a saturation approach makes individual or group randomization infeasible.

- Conclusion: PSM provides an analytic framework that can add an important degree of control in quasi-experimental evaluations.
NC Waiver Demonstration Project

- System reform effort that utilizes flexible funding to implement county-specific “demonstration” in 39 Waiver counties
- All Waiver counties agree to work to change these outcomes:
  - Reduce the likelihood of out-of-home placement while maintaining safety for children
  - Reduce the length of stay for children who must enter out-of-home placement
  - Reduce the likelihood of repeat abuse or neglect for all children
NC Waiver Demonstration Project

- County demonstrations include implementation of strategies targeted to individual clients, agency staff and community members that all combine to change the experiences of children involved with the child welfare system
Evaluation of the NC Waiver Demonstration Project

- Comparison group design with 39 Waiver counties and 30 comparison counties
- Outcomes for all children served in these counties are tracked using longitudinal data derived from administrative data files maintained by the state
- There is only limited services data available for all children in the state so it is not possible to adequately track the changes in service utilization at a client level with existing data files
Process Evaluation

- Selected 2 random samples of children in the Waiver counties
- Using PSM select 2 matched samples of children in the comparison counties using age, race and gender
- Implement a case record review for the samples to get supplemental information on services used at varying time points in the case
Why PSM?

- Waiver counties self-selected into the demonstration
- Comparison counties were selected so that the group as a whole matched Waiver counties on key county-level elements: rate of change in IV-E expenditures over the last several years, similar patterns of use of out-of-home placement, size of county
- There were age and race differences in the distribution of children served in the 2 groups
Why PSM?

- We could not collect services data on all children in the counties therefore had to rely on sampling strategy.
- PSM assured that the samples were similarly distributed on key individual characteristics.
Process Evaluation

- Use the sample data in conjunction with county aggregate data on service availability to describe service patterns for clients in Waiver and comparison counties.
- Repeat this process midway and at the end of demonstration period to assess changes due to the Waiver.
Outcome analysis

- Assess changes in outcomes for all children in Waiver counties versus comparison counties
- Supplement these analyses by examining outcomes for children in the samples
Current status

- Very early in evaluation
- Will be finishing up first round of case record abstractions in both Waiver and comparison counties this month
- Will begin looking at baseline services data from the samples and linking these data back to existing data sources to check validity and reliability of the data