Caliendo & Kopeinig (2005)

Causality reading group, October 14th 2019
Stéphanie van der Pas
Some Practical Guidance for the Implementation of Propensity Score Matching

Marco Caliendo
Sabine Kopeinig

May 2005
Figure 1: PSM - Implementation Steps

- **Step 0:** Decide between PSM and CVM
- **Step 1:** Propensity Score Estimation (sec. 3.1)
- **Step 2:** Choose Matching Algorithm (sec. 3.2)
- **Step 3:** Check Overlap/Common Support (sec. 3.3)
- **Step 4:** Matching Quality/Effect Estimation (sec. 3.4-3.7)
- **Step 5:** Sensitivity Analysis (sec. 4)

CVM: Covariate Matching, PSM: Propensity Score Matching
Roy-Rubin potential outcome model

had he not received the treatment. The standard framework in evaluation analysis to formalise this problem is the potential outcome approach or Roy-Rubin-model (Roy (1951), Rubin (1974)). The main pillars of this model are individuals, treatment

Roy (1951): town with two occupations:

- rabbit hunting;
- trout fishing.

Core ideas:
- people will choose the occupation at which they have a comparative advantage;

- distribution of earnings within an occupation is affected by the self-selection of persons into an occupation.


Basic concepts (Rubin (2004))

**Unit**
The person, place, or thing upon which a treatment will operate, at a particular time.

**Treatment**
An intervention, the effects of which (on some particular measurement on the units) the investigator wishes to assess relative to no intervention (i.e., the “control”).

**Potential outcomes**
The values of a unit’s measurement of interest after (a) application of the treatment and (b) non-application of the treatment (i.e., under control).

**Causal effect**
For each unit, the comparison of the potential outcome under treatment and the potential outcome under control.

**The fundamental problem of causal inference**
We can observe at most one of the potential outcomes for each unit.
Causal effect

Rubin (2004)
For each unit, the comparison of the potential outcome under treatment and the potential outcome under control.

Elements of causal inference
The following example motivates the namings “cause” and “effect”:

Example 3.2 (Cause-effect interventions) Suppose that the distribution \( P_{C,E} \) is entailed by an SCM \( \mathcal{C} \)

\[
C := N_C \\
E := 4 \cdot C + N_E,
\]
with \( N_C, N_E \overset{iid}{\sim} \mathcal{N}(0, 1) \), and graph \( C \rightarrow E \). Then,

\[
P^e_E = \mathcal{N}(0, 17) \neq \mathcal{N}(8, 1) = P^e_{E; do(C=2)} = P^e_{E|C=2} \\
\neq \mathcal{N}(12, 1) = P^e_{E; do(C=3)} = P^e_{E|C=3}.
\]

Intervening on \( C \) changes the distribution of \( E \). But on the other hand,

\[
P^e_{C; do(E=2)} = \mathcal{N}(0, 1) = P^e_C = P^e_{C; do(E=3)} = (\neq P^e_{C|E=2}).
\]
Causal effect

Pearl (2009), p. 46-47

Definition 2.3.6 (Inferred Causation)

*Given* \( \hat{P} \), a variable \( C \) has a causal influence on variable \( E \) if and only if there exists a directed path from \( C \) to \( E \) in every minimal latent structure consistent with \( \hat{P} \).

We view this definition as normative because it is based on one of the least disputed norms of scientific investigation: Occam’s razor in its semantical casting. However, as with any scientific inquiry, we make no claims that this definition is guaranteed to always identify stable physical mechanisms in Nature. It identifies the mechanisms we can plausibly infer from nonexperimental data; moreover, it guarantees that any alternative mechanism will be less trustworthy than the one inferred because the alternative would require more contrived, hindsight adjustment of parameters (i.e., functions) to fit the data.

Pearl (2009), p.103 (emphasis and formatting mine)

“ [...] one cannot overemphasize the importance of the **conceptual clarity that structural equations offer** vis-à-vis the potential-outcome model. [...] The thought of having to express, defend, and manage formidable counterfactual relationships of this type may explain why **the enterprise of causal inference is currently viewed with such awe and despair among rank-and-file epidemiologists and statisticians** - and why most economists and social scientists continue to use structural equations instead of the potential-outcome alternatives [...]”.
Neyman-Rubin potential outcome model

Binary treatment indicator:

\[ D_i = \begin{cases} 
1 & \text{individual } i \text{ is treated} \\
0 & \text{otherwise} 
\end{cases} \]

(Potential) outcomes for individual \( i \):

\( (Y_i(0), Y_i(1)) \)

Treatment effect for an individual \( i \):

\( \tau_i = Y_i(1) - Y_i(0) \)
Estimands

Average treatment effect on the treated:

$$\tau_{ATT} = E[\tau \mid D = 1] = E[Y(1) \mid D = 1] - E[Y(0) \mid D = 1].$$

“Using the mean outcome of untreated individuals \( E[Y(0) \mid D = 0] \) is in non-experimental studies usually not a good idea.”

Average treatment effect:

$$\tau_{ATE} = E[Y(1) - Y(0)].$$
Regression and matching estimates of the effects of elite college attendance on educational and career achievement

Jennie E. Brand *, Charles N. Halaby

Department of Sociology, University of Wisconsin-Madison, Madison, WI 53706, USA
Available online 19 August 2005

Abstract

This paper adopts a potential outcome approach to identify and estimate the average treatment effect of attending an elite college on educational and career achievement. A central purpose is to compare the estimates yielded by regression and matching methods of adjusting for the endogeneity of elite college attendance. The analysis follows a high school graduation and college entry cohort across four decades of labor force participation, and estimates elite college effects on educational attainment, occupational socioeconomic status at early-, mid-, and late-career, and wages at mid- and late-career. The findings suggest that attending an elite college yields an advantage with respect to educational achievement and occupational status; results for wages are mixed. One prominent pattern is that the returns to attending an elite college for those who did attend are small by comparison to those that would have been achieved by otherwise equivalent students who attended non-elite institutions.
Matching

Covariate

Treatment

Outcome
Matching
Propensity score matching

The exogenous measures constitute a fairly comprehensive representation of variables that have figured prominently in sociological and economic studies of college attendance and college selectivity. The outcome variables tap major dimensions of educational, occupational, and economic achievement that have appeared in the stratification literature over the past three decades.

Table 1
Descriptions of pre-college exogenous and career outcome variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ability and academics</strong></td>
<td></td>
</tr>
<tr>
<td>Class rank</td>
<td>High school grades percentile rank 0-99</td>
</tr>
<tr>
<td>Mental ability</td>
<td>Henmon-Nelson 11th grade IQ scores</td>
</tr>
<tr>
<td>College track</td>
<td>Dummy variable (met UW-Madison requirements = 1)</td>
</tr>
<tr>
<td>Math</td>
<td>Semesters of high school math</td>
</tr>
<tr>
<td>Private/public high school</td>
<td>Dummy variable (private = 1)</td>
</tr>
<tr>
<td><strong>Family background</strong></td>
<td></td>
</tr>
<tr>
<td>Parent’s income</td>
<td>Truncated at $99,800; started log</td>
</tr>
<tr>
<td>Father’s (head’s) SEI</td>
<td>1957 Occupation; Duncan 1970 SEI 0-96;</td>
</tr>
<tr>
<td>Mother’s education</td>
<td>Years of schooling completed</td>
</tr>
<tr>
<td>Catholic</td>
<td>Dummy variable (Catholic = 1)</td>
</tr>
<tr>
<td>Jewish</td>
<td>Dummy variable (Jewish = 1)</td>
</tr>
<tr>
<td>Intact family</td>
<td>Dummy variable (living with both parents = 1)</td>
</tr>
<tr>
<td>Number of siblings</td>
<td>Number of siblings</td>
</tr>
<tr>
<td>Rural/urban residence</td>
<td>Dummy variable (rural = 1)</td>
</tr>
<tr>
<td><strong>Career outcomes</strong></td>
<td></td>
</tr>
<tr>
<td>Graduated college</td>
<td>Dummy variable (bachelor’s = 1)</td>
</tr>
<tr>
<td>Advanced degree</td>
<td>Dummy variable (masters or PhD = 1)</td>
</tr>
<tr>
<td>Occupational status, first job</td>
<td>Duncan 1970 SEI 0-96</td>
</tr>
<tr>
<td>Occupational status 1974</td>
<td>Duncan 1970 SEI 0-96</td>
</tr>
<tr>
<td>Wage 1974</td>
<td>Hourly wage (started log)</td>
</tr>
<tr>
<td>Occupational status 1992</td>
<td>Duncan 1970 SEI 0-96</td>
</tr>
<tr>
<td>Wage 1992</td>
<td>Hourly wage (started log)</td>
</tr>
</tbody>
</table>
The central role of the propensity score in observational studies for causal effects

By Paul R. Rosenbaum

Departments of Statistics and Human Oncology, University of Wisconsin, Madison, Wisconsin, U.S.A.

And Donald B. Rubin

University of Chicago, Chicago, Illinois, U.S.A.

Summary

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Both large and small sample theory show that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates. Applications include: (i) matched sampling on the univariate propensity score, which is a generalization of discriminant matching, (ii) multivariate adjustment by subclassification on the propensity score where the same subclasses are used to estimate treatment effects for all outcome variables and in all subpopulations, and (iii) visual representation of multivariate covariance adjustment by a two-dimensional plot.
Assumptions - “The assignment mechanism is critical”

**Stable Unit-Treatment-Value Assumption (SUTVA)**
There is only one form of the treatment and one form of the control, and there is no interference among units.

Rubin (2004):

<table>
<thead>
<tr>
<th>You take:</th>
<th>Asp</th>
<th>Not</th>
<th>Asp</th>
<th>Not</th>
</tr>
</thead>
<tbody>
<tr>
<td>I take:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Potential Outcomes and Values in Example

<table>
<thead>
<tr>
<th>Unit</th>
<th>Y₁(Asp, Asp)</th>
<th>Y₁(Not, Not)</th>
<th>Y₁(Asp, Not)</th>
<th>Y₁(Not, Asp)</th>
<th>Y₂(Asp, Asp)</th>
<th>Y₂(Not, Not)</th>
<th>Y₂(Asp, Not)</th>
<th>Y₂(Not, Asp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 = you</td>
<td>Y₁(Asp, Asp) = 0</td>
<td>Y₁(Not, Not) = 100</td>
<td>Y₁(Asp, Not) = 50</td>
<td>Y₁(Not, Asp) = 75</td>
<td>Y₂(Asp, Asp) = 0</td>
<td>Y₂(Not, Not) = 100</td>
<td>Y₂(Asp, Not) = 100</td>
<td>Y₂(Not, Asp) = 0</td>
</tr>
<tr>
<td>2 = me</td>
<td>Y₂(Asp, Asp) = 0</td>
<td>Y₂(Not, Not) = 100</td>
<td>Y₂(Asp, Not) = 100</td>
<td>Y₂(Not, Asp) = 0</td>
<td>Y₁(Asp, Asp) = 0</td>
<td>Y₁(Not, Not) = 0</td>
<td>Y₁(Asp, Not) = 0</td>
<td>Y₁(Not, Asp) = 0</td>
</tr>
</tbody>
</table>

“Perhaps when I have headaches, I complain a great deal to you, inducing whatever headpain you have to increase.”
Assumptions - strong ignorability

Common support \[ 0 < P(D = 1 | X) < 1 \]

Unconfoundedness \[ Y(0), Y(1) \perp D | X \]
Rubin (2004): “The perfect doctor”

<table>
<thead>
<tr>
<th>Y(0)</th>
<th>Y(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>W</th>
<th>Y(0)</th>
<th>Y(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>?</td>
<td>14</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>?</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>?</td>
<td>9</td>
</tr>
</tbody>
</table>

Observed \[ \bar{y}_1 - \bar{y}_0 = 5.6 \neq -2. \]
Rosenbaum and Rubin (1983)

A balancing score \( b(x) \) is a function of the observed covariates such that the conditional distribution of \( x \) given \( b(x) \) is the same for the treated \((z = 1)\) and control units \((z = 0)\):

\[
x \perp z \mid b(x)
\]

The propensity score is the conditional probability of assignment to treatment one, given the covariates.

\[
e(x) = \mathbb{P}(z = 1 \mid x)
\]

**Theorem 1**
The propensity score is a balancing score.
Rosenbaum and Rubin (1983)

**Theorem 3**
If treatment assignment is strongly ignorable given $x$, then it is strongly ignorable given any balancing score $b(x)$.

**Theorem 4**
If treatment assignment is strongly ignorable and $b(x)$ is a balancing score, then

$$
\mathbb{E}[r_1 \mid b(x), z = 1] - \mathbb{E}[r_0 \mid b(x), z = 0] = \mathbb{E}[r_1 - r_0 \mid b(x)]
$$
Should we do propensity score matching at all?

Zhao (2004):
- The propensity score is perceived to be “less data-hungry”, but individuals matched on the propensity score may have quite different treatment outcomes;

- PSM is a good approach when covariates and treatment indicator are highly correlated, but does not perform well in small sample sizes;

- Mahalanobis matching is relatively robust under different settings;

- Incorporating outcome information into the matching metric might be a promising approach.
Why Propensity Scores Should Not Be Used for Matching

Gary King\textsuperscript{1} and Richard Nielsen\textsuperscript{2}

\textsuperscript{1}Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA. Email: king@harvard.edu, URL: http://GaryKing.org
\textsuperscript{2}Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. Email: rnielsen@mit.edu, URL: http://www.mit.edu/~rnielsen

Abstract
We show that propensity score matching (PSM), an enormously popular method of preprocessing data for causal inference, often accomplishes the opposite of its intended goal—thus increasing imbalance, inefficiency, model dependence, and bias. The weakness of PSM comes from its attempts to approximate a completely randomized experiment, rather than, as with other matching methods, a more efficient fully blocked randomized experiment. PSM is thus uniquely blind to the often large portion of imbalance that can be eliminated by approximating full blocking with other matching methods. Moreover, in data balanced enough to approximate complete randomization, either to begin with or after pruning some observations, PSM approximates random matching which, we show, increases imbalance even relative to the original data. Although these results suggest researchers replace PSM with one of the other available matching methods, propensity scores have other productive uses.

\textit{Keywords:} matching, propensity score matching, coarsened exact matching, Mahalanobis distance matching, model dependence

We trace the PSM paradox to the particular way propensity scores interact with matching. Thus, our results do not necessarily implicate the many other productive uses of propensity scores, such as regression adjustment (Vansteelandt and Daniel 2014), inverse weighting (Robins, Hernan, and Brumback 2000), stratification (Rosenbaum and Rubin 1984), and some uses of the propensity score within other methods (e.g., Diamond and Sekhon 2012; Imai and Ratkovic 2014). Moreover, the mathematical theorems in the literature used to justify propensity scores in general, such as in Rosenbaum and Rubin (1983), are of course correct and useful elsewhere, but we show they are not relevant to the practice of matching.
Proof of Rosenbaum and Rubin (1983) is “although mathematically correct, [...] either of little use or misleading when applied to real data.”

The theorem
- “encourages researchers to settle for the lower standards of approximating only complete randomization” (rather than a blocked experiment);
- and would be “somewhat more useful if it were reversed”.

King and Nielsen (2019)