11.3.5 Understanding Propensity Scores

The method of propensity score (Rosenbaum and Rubin 1983), or propensity score matching (PSM) is the most developed and popular strategy for causal analysis in observational studies. It is not emphasized in this book, because it is an estimation method, designed to deal with the variability of finite samples, but does not add much to our understanding of the asymptotic, large sample limits, which is the main focus of the book. However, due to the prominence of propensity score methods in causal analysis, and recent controversies surrounding its usage, we devote this section to explain where it falls in the grand scheme of graphical models, admissibility, identifiability, bias reduction, and the statistical vs. causal dichotomy.

The method of propensity score is based on a simple, yet ingenious idea, of purely statistical character. Assuming a binary action (or treatment) X, and an arbitrary set S of measured covariates, the propensity score $L(s)$ is the probability that action $X = 1$ will be chosen by a participant with characteristics $S = s$, or

$$L(s) = P(X = 1 | S = s) \tag{11.9}$$

What Rosenbaum and Rubin showed is that, viewing L(s) as a function of S, hence, as a random variable, X and S are independent given $L(s)$, that is, $X \perp\!\!\!\perp S | L(s)$. In words, all units that map into the same value of $L(s)$ are comparable, or "balanced," in the sense that, within each stratum of L, treated and untreated units have the same distribution of characteristics S.[6]

To see the significance of this result, let us assume, for simplicity that $L(s)$ can be estimated separately and approximated by discrete strata $L = \{l_1, l_2, \dots, l_k\}$. The conditional independence $X \perp\!\!\!\perp S | L(s)$, together with the functional mapping $S \to L$ renders S and L c-equivalent in the sense defined in Section 11.3.3, Eq. (11.8), namely, for any Y,

$$\sum_s P(y | s, x) P(s) = \sum_l P(y | l, x) P(l). \tag{11.10}$$

This follows immediately by writing:[7]

$$
\begin{aligned}
\sum_l P(y | l, x) P(l) &= \sum_s \sum_l P(y | l, s, x) P(l) P(s | l, x) \\
&= \sum_s \sum_l P(y | s, x) P(l) P(s | l) \\
&= \sum_s P(y | s, x) P(s)
\end{aligned}
$$

Thus far we have not mentioned any causal relationship, nor the fact that Y is an outcome variable and that, eventually, our task would be to estimate the causal effect of X on Y. The c-equivalence of S and L merely implies that, if for any reason one wishes to estimate the "adjustment estimand" $\sum_s P(y | s, x) P(s)$, with S and Y two arbitrary sets of variables then, instead of summing over a high dimensional set S, one might as well sum over a one dimensional vector $L(s)$. The asymptotic estimate, in the limit of very large sample, would be the same in either method.

---

[6]This independence emanates from the special nature of the function $L(s)$, and is not represented in the graph, i.e., if we depict L as a child of S, L would not in general d-separate S from X.

[7]This also follows from the fact that condition C_2 is satisfied by the substitution $S_1 = S$ and $S_2 = L(s)$.

This *c*-equivalence further implies, and this is where causal inference first comes into the picture, that if one chooses to approximate the causal effect $P(y|do(x))$ by the adjustment estimand $E_s P(y|s, x)$ then, asymptotically, the same approximation can be achieved using the estimand $E_l P(y|l, x)$, where the adjustment is performed over the strata of L. The latter has the advantage that, for finite samples, each of the strata is less likely to be empty and each is likely to contain both treated and untreated units.

The method of propensity score can thus be seen as an efficient estimator of the adjustment estimand, formed by an arbitrary set of covariates S; it makes no statement regarding the appropriateness of S nor does it promise to correct for any confounding bias, or to refrain from creating new bias where none exists.

In the special case where S is admissible, that is,

$$P(y|do(x)) = E_s P(y|s, x) \tag{11.11}$$

L would be admissible as well, and we would then have an unbiased estimand of the causal effect,[8]

$$P(y|do(x)) = E_l P(y|l, x)$$

accompanied with an efficient method of estimating the right hand side. Conversely, if S in inadmissible, L would be inadmissible as well, and all we can guarantee is that the bias produced by the former would be faithfully and efficiently reproduced by the latter.

The controversy surrounding propensity score

Thus far, our presentation of propensity score leaves no room for misunderstanding, and readers of this book would find it hard to understand how a controversy could emerge from an innocent estimation method which merely offers an efficient way of estimating a statistical quantity that sometimes does, and sometimes does not coincide with the causal quantity of interest, depending on the choice of S.

But a controversy has developed recently, most likely due to the increased popularity of the method and the strong endorsement it received from prominent statisticians (Rubin 2007), social scientists (Morgan and Winship 2007; Berk and de Leeuw 1999), health scientists (Austin 2008), and economists (Heckman 1992). The popularity of the method has in fact grown to the point where some federal agencies now expect program evaluators to use this approach as a substitute for experimental designs (Peikes et al. 2008). This move reflects a general tendency among investigators to play down the cautionary note concerning the required admissibility of S, and to interpret the mathematical proof of Rosenbaum and Rubin as a guarantee that, in each strata of L, matching treated and untreated subjects somehow eliminates confounding from the data and contributes therefore to overall bias reduction. This tendency was further reinforced by empirical studies (Heckman et al. 1998; Dehejia and Wahba 1999) in which agreement was found between propensity score analysis and randomized trials, and in which the agreement was attributed to the ability of the

---

[8]Rosenbaum and Rubin (1983) proved the *c*-equivalence of S and L only for admissible S, which is unfortunate; it gives readers the impression that the propensity score matching somehow contributes to bias reduction.

former to "balance" treatment and control groups on important characteristics. Rubin has encouraged such interpretations by stating: "This application uses propensity score methods to create subgroups of treated units and control units...as if they had been randomized. The collection of these subgroups then "approximate" a randomized block experiment with respect to the observed covariates" (Rubin 2007).

Subsequent empirical studies, however, have taken a more critical view of propensity score, noting with disappointment that a substantial bias is sometimes measured when careful comparisons are made to results of clinical studies (Smith and Todd 2005; Luellen et al. 2005; Peikes et al. 2008).

But why would anyone play down the cautionary note of Rosenbaum and Rubin when doing so would violate the golden rule of causal analysis: No causal claim can be established by purely statistical method, be it propensity scores, regression, stratification, or any other distribution-based design. The answer, I believe, rests with the language that Rosenbaum and Rubin used to formulate the condition of admissibility, i.e., Eq. (1.11). The condition was articulated in the restricted language of potential-outcome, stating that the set S must render X "strongly ignorable," i.e., $\{Y_1, Y_0\} \perp\!\!\!\perp X | S$. As stated several times in this book, the opacity of "ignorability" is the Achilles heel of the potential-outcome approach – no mortal can apply this condition to judge whether it holds in even simple problems, with all causal relationships correctly specified, let alone partially specified problems that involve dozens of variables.[9]

The difficulty that most investigators experience in comprehending what "ignorability" means, and what judgment it summons them to exercise, has tempted them to assume that it is automatically satisfied, or at least is likely to be satisfied if one includes in the analysis as many covariates as possible. The prevailing attitude is that adding more covariates can cause no harm (Rosenbaum 2002, p. 76) or the need to reason about the causal relationships among those covariates, the treatment, the outcome and, most importantly, the confounders left unmeasured.

This attitude stands contrary to what students of graphical models have learned, and what this book has attempted to teach. The admissibility of S can only be established by appealing to the causal knowledge available to the investigator, and that knowledge, as we know from graph theory and the back-door criterion, makes bias reduction a non-monotonic operation, i.e., eliminating bias (or imbalance) due to one confounder may awaken and unleash bias due to dormant, unmeasured confounders. Examples are abound (e.g., Figure 6.3) where adding a variable to the analysis, not only is not needed but would introduce irreparable bias.

Another factor inflaming the controversy has been the general belief that the bias-reducing potential of propensity score methods can be tested experimentally by running case studies and comparing effect estimates obtained by propensity scores to those obtained by controlled randomized experiments (Shadish and Cook 2009).[10] This belief is

---

[9] Advocates of the potential outcome tradition are invited to inspect Figure 11.8 (b) (or any model, or story, or toy-example of their choice) and judge whether any subset of C renders X "strongly ignorable." This could easily be determined, of course, by the back-door criterion, but, unfortunately, graphs are still feared and misunderstood by some of the chief advocates of the potential-outcome camp (e.g., Rubin 2004).

[10] Such beliefs are encouraged by careless statements such as: "For dramatic evidence that such an analysis can

unjustified because the bias-reducing potential of propensity scores depends critically on the specific choice of S or, more accurately, on the cause-effect relationships among variables inside and outside S. Measuring significant bias in one problem instance (say, an educational program in Oklahoma) does not preclude finding zero bias in another (say, crime control in Arkansas), even under identical statistical distributions $P(x, s, y)$.

With these considerations in mind, one is justified in asking a social science type question: What is it about propensity scores that has inhibited a more general understanding of their promise and limitations?

Richard Berk in *Regression Analysis: A Constructive Critique* (Berk 2004) recalls similar phenomena in social science, where immaculate ideas were misinterpreted by the scientific community: "I recall a conversation with Don Campbell in which he openly wished that he had never written Campbell and Stanley (1966). The intent of the justly famous book, *Experimental and Quasi-Experimental Designs for Research*, was to contrast randomized experiments to quasi-experimental approximations and to strongly discourage the latter. Yet the apparent impact of the book was to legitimize a host of quasi-experimental designs for a wide variety of applied social science. After I got to know Dudley Duncan late in his career, he said that he often thought that his influential book on path analysis, *Introduction to Structural Equation Models* was a big mistake. Researchers had come away from the book believing that fundamental policy questions about social inequality could be quickly and easily answered with path analysis."

I believe that a similar cultural phenomenon has evolved around propensity scores.

It is not that Rosenbaum and Rubin were careless in stating the condition for success. Formally, they were very clear in warning practitioners that propensity scores work only under "strong ignorability" condition. However, what they failed to realize is that it is not enough to warn people against dangers they cannot recognize; to protect them from perilous adventures, we must also give them eye-glasses to spot the threats, and a meaningful language to reason about them. By failing to equip readers with tools (e.g., graphs) for recognizing how "strong ignorability" can be violated or achieved, they have encouraged a generation of researchers (including federal agencies) to assume that ignorability either holds in most cases, or can be made to hold by clever designs.

---

reach the same conclusion as an exactly parallel randomized experiment, see Shadish and Clark (2006, unpublished)" (Rubin 2007).