Biometrika (2000), **87**, 3, *pp.* 706–710

Printed in Great Britain

The role of the propensity score in estimating dose-response functions

By GUIDO W. IMBENS

Department of Economics, University of California at Los Angeles, Los Angeles, California 90095, U.S.A.

imbens@econ.ucla.edu

Summary

Estimation of average treatment effects in observational studies often requires adjustment for differences in pre-treatment variables. If the number of pre-treatment variables is large, standard covariance adjustment methods are often inadequate. Rosenbaum & Rubin (1983) propose an alternative method for adjusting for pre-treatment variables for the binary treatment case based on the so-called propensity score. Here an extension of the propensity score methodology is proposed that allows for estimation of average casual effects with multi-valued treatments.

Some key words: Causal inference; Dose-response function; Multivalued treatment; Observational study; Propensity score; Unconfoundedness.

1. Introduction

Estimation of average treatment effects in observational studies often requires adjustment for differences in pre-treatment variables. If the number of pre-treatment variables is large and their distribution varies substantially with treatment status, standard adjustment methods such as covariance adjustment are often inadequate. Rosenbaum & Rubin (1983, 1984) propose an alternative method for adjusting for pre-treatment variables based on the propensity score, the conditional probability of receiving the treatment given pre-treatment variables. They demonstrate that adjusting solely for the propensity score removes all bias associated with differences in the pre-treatment variables. The Rosenbaum–Rubin proposals deal exclusively with binary-valued treatments. In many cases of interest, however, treatments take on more than two values. Here an extension of the propensity score methodology is proposed that allows for estimation of average causal effects with multi-valued treatments. The key insight is that for estimation of average causal effects it is not necessary to divide the population into subpopulations where causal comparisons are valid, as the propensity score does; it is sufficient to divide the population into subpopulations where average potential outcomes can be estimated.

2. The model

We are interested in the average causal effect of some treatment on some outcome. The treatment, denoted by T, takes on values in a set $\mathscr{T}$. Associated with each unit i and each value of the treatment t there is a potential outcome, denoted by $Y_i(t)$. We are interested in average outcomes, $E\{Y(t)\}$, for all values of t, and in particular in differences of the form $E\{Y(t)-Y(s)\}$, the average causal effect of exposing all units to treatment t rather than treatment s. The average here is taken over the population of interest, which may be the population the sample is drawn from, or some subpopulation thereof. More generally we can look at average differences of functions of $Y(t)$ for different values of t, such as the distribution function of $Y(t)$ at a point. We observe, for each unit i

in a random sample of size N drawn from a large population, the treatment T_i, the outcome associated with that treatment level $Y_i \equiv Y_i(T_i)$, and a vector of pre-treatment variables X_i.

The key assumption, maintained throughout the paper, is that adjusting for pre-treatment differences solves the problem of drawing causal inferences. This is formalised by using the concept of unconfoundedness. Let $D_i(t)$ be the indicator of receiving treatment t:

$$D_i(t) = \begin{cases} 1 & \text{if } T_i = t, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 1 (*Weak unconfoundedness*). *Assignment to treatment T is weakly unconfounded, given pre-treatment variables X, if*

$$D(t) \perp Y(t) \mid X,$$

for all $t \in \mathscr{T}$.

Rosenbaum & Rubin (1983) make the stronger assumption that

$$T \perp \{Y(t)\}_{t \in \mathscr{T}} \mid X,$$

which requires the treatment T to be independent of the entire set of potential outcomes. Instead, weak unconfoundedness requires only pairwise independence of the treatment with each of the potential outcomes, like the assumption used in Robins (1995). In addition weak unconfoundedness only requires the independence of the potential outcome $Y(t)$ and the treatment to be 'local' at the treatment level of interest, that is independence of the indicator $D(t)$, rather than of the treatment level T. The definition of weak unconfoundedness is similar to that of 'missing at random' (Rubin, 1976; Little & Rubin, 1987, p. 14) in the missing data literature.

Although in substantive terms the weak unconfoundedness assumption is not very different from the assumption used by Rosenbaum & Rubin (1983), it is important that one does not need the stronger assumption to validate estimation of the expected value of $Y(t)$ by adjusting for X:

$$E\{Y(t) \mid X\} = E\{Y(t) \mid D(t) = 1, X\} = E\{Y \mid D(t) = 1, X\} = E\{Y \mid T = t, X\}. \qquad (1)$$

Average outcomes can then be estimated by averaging these conditional means:

$$E\{Y(t)\} = E[E\{Y(t) \mid X\}].$$

In practice it can be difficult to estimate $E\{Y(t)\}$ in this manner when the dimension of X is large, because the first step requires estimation of the expectation of $Y(t)$ given the treatment level and all pre-treatment variables, and this motivated Rosenbaum & Rubin (1983) to develop the propensity score methodology.

3. The propensity score with binary treatments

In the binary treatment context with $\mathscr{T} = \{0, 1\}$, Rosenbaum & Rubin (1983) define the propensity score as the conditional probability of receiving the treatment given the pre-treatment variables:

$$e(x) \equiv \text{pr}(T = 1 \mid X = x).$$

If assignment to treatment is weakly unconfounded given the pre-treatment variables, then assignment to treatment is weakly unconfounded given the propensity score:

$$D(t) \perp Y(t) \mid e(X),$$

for all $r \in \mathscr{T}$. This result implies that, instead of having to adjust for all pre-treatment variables, it is sufficient to adjust for the propensity score $e(X)$.

An alternative method for exploiting the propensity score is through weighting by the inverse of the probability of receiving the treatment actually received (Rosenbaum, 1987), similar to the

Horvitz–Thompson estimator (Horvitz & Thompson, 1952). By weak unconfoundedness, we have

$$E\left\{\frac{YT}{e(X)}\right\} = E\{Y(1)\}, \quad E\left\{\frac{Y(1-T)}{1-e(X)}\right\} = E\{Y(0)\},$$

which can be used to estimate the average causal effect $E\{Y(1) - Y(0)\}$.

4. The propensity score with multi-valued treatments

Here we allow the treatment of interest to take on integer values between 0 and K, so that $\mathscr{T} = \{0, 1, \ldots, K\}$. First, we modify the Rosenbaum–Rubin definition of the propensity score.

Definition 2 (*Generalised propensity score*). *The generalised propensity score is the conditional probability of receiving a particular level of the treatment given the pre-treatment variables*:

$$r(t, x) \equiv \mathrm{pr}(T = t \mid X = x) = E\{D(t) \mid X = x\}.$$

Suppose assignment to treatment T is weakly unconfounded given pre-treatment variables X. Then, by the same argument as in the binary treatment case, assignment is weakly unconfounded given the generalised propensity score: $D(t) \perp Y(t) \mid r(t, X)$, for all $t \in \mathscr{T}$. This is the point where using the weak form of the unconfoundedness assumption is important. There is in general no scalar function of the covariates such that the level of the treatment T is independent of the set of potential outcomes $\{Y(t)\}_{t \in \mathscr{T}}$. Such a scalar function may exist if additional structure is imposed on the assignment mechanism; see for example Joffe & Rosenbaum (1999).

Since weak unconfoundedness given all pretreatment variables implies weak unconfoundedness given the generalised propensity score, one can estimate average outcomes by conditioning solely on the generalised propensity score.

Theorem 1. *Suppose assignment to treatment is weakly unconfounded given pre-treatment variables X. Then, for all $t \in \mathscr{T}$,*

(i) $\beta(t, r) \equiv E\{Y(t) \mid r(t, X) = r\} = E\{Y \mid T = t, r(T, X) = r\}$,

(ii) $E\{Y(t)\} = E\{\beta(t, r(t, X))\}$.

As with the implementation of the binary treatment propensity score methodology, the implementation of the generalised propensity score method consists of three steps. In the first step the score $r(t, x)$ is estimated. With a binary treatment the standard approach (Rosenbaum & Rubin, 1984; Rosenbaum, 1995, p. 79) is to estimate the propensity score using a logistic regression. With a multi-valued treatment one may distinguish two cases of interest. First, if the values of the treatment are qualitatively distinct and without a logical ordering, such as surgery, drug treatment and no treatment, one may wish to use discrete response models such as the multinomial or nested logit. Secondly, if the treatments correspond to ordered levels of a treatment, such as the dose of a drug or the time over which a treatment is applied, one may wish to impose smoothness of the score in t.

In the second step the conditional expectation $\beta(t, r) = E\{Y \mid T = t, r(T, X) = r\}$ is estimated. Again the implementation may be different in the case where the levels of the treatment are qualitatively distinct from the case where smoothness of the conditional expectation function in t is appropriate.

In the third step the average response at treatment level t is estimated as the average of the estimated conditional expectation, $\hat{\beta}(t, r(t, X))$, averaged over the distribution of the pre-treatment variables. Note that to get the average $E\{Y(t)\}$ the second argument in the conditional expectation $\beta(t, r)$ is evaluated at $r(t, X_i)$, not at $r(T_i, X_i)$.

As an alternative to the above implementation one can use the inverse of the generalised propensity score to weight the observations, using the following equality:

$$E\left\{\frac{YD(t)}{r(T, X)}\right\} = E\{Y(t)\}.$$

It appears difficult to exploit smoothness of the outcome in the level of the treatment in this weighting approach. Similarly, matching approaches where units are grouped in a way to allow causal comparisons within matches appear less well suited to the multi-valued treatment case.

5. Comparison with binary treatments

The Rosenbaum–Rubin propensity score partitions the population into subpopulations where valid causal comparisons can be made. Within the subpopulation with propensity score equal to $e(X)=e$, the average value of $Y(1)$ for treated units is unbiased for the subpopulation average value of $Y(1)$, and similarly for the average value of $Y(0)$ for control units. Hence in this subpopulation the difference in sample averages by treatment status is unbiased for the average causal effect. In other words, the regression of the observed outcome on treatment level and propensity score has a causal interpretation.

The generalised propensity score also partitions the population into subpopulations. Within the subpopulation with $r(T, X)=r$, the average value of $Y(t)$ for units with treatment level t is an unbiased estimator of the average of $Y(t)$ for the subpopulation with $r(t, X)=r$. However, in the same subpopulation the average of $Y(s)$ for units with $T=s$ is unbiased for the average of $Y(s)$ in a different subpopulation, namely that with $r(s, X)=r$. Hence no causal comparison can be drawn within the subpopulation defined by $r(T, X)=r$, and the regression of observed outcome Y on treatment level T and the score $r(T, X)$ does not have a causal interpretation. Formally,

$$\beta(t, r)-\beta(s, r)=E\{Y(t) \mid T=t, r(T, X)=r\}-E\{Y(s) \mid T=s, r(T, X)=r\}.$$

By weak unconfoundedness this is equal to

$$E\{Y(t) \mid r(t, X)=r\}-E\{Y(s) \mid r(s, X)=r\},$$

which has no causal interpretation because the conditioning sets differ. To obtain a causal interpretation one needs to condition on the intersection of the two conditioning sets:

$$E\{Y(t) \mid T=t, r(t, X), r(s, X)\}-E\{Y(s) \mid T=s, r(t, X), r(s, X)\}=E\{Y(t)-Y(s) \mid r(t, X), r(s, X)\},$$

a point also made in Lechner (2000). However, in general such causal interpretations require conditioning on an additional variable. This is exactly what the propensity score approach attempts to avoid.

In the binary treatment case the additional conditioning can be avoided by virtue of the fact that the two assignment probabilities add up to unity. Rosenbaum & Rubin (1983) demonstrate that, conditional on the propensity score, outcome differences by treatment status are unbiased for average treatment effects:

$$E\{Y(1) \mid T=1, e(X)=e\}-E\{Y(0) \mid T=0, e(X)=e\}=E\{Y(1)-Y(0) \mid e(X)=e\}.$$

To see the difference with the generalised propensity score, let us rewrite this in the new notation, with $r(1, x)=e(x)$ and $r(0, x)=1-e(x)$:

$$\begin{aligned}\beta(1, e)-\beta(0, 1-e) &= E\{Y(1) \mid r(1, X)=e\}-E\{Y(0) \mid r(0, X)=1-e\} \\ &= E\{Y(1)-Y(0) \mid r(1, X)=e\}.\end{aligned}$$

The reason that this causal comparison requires no additional conditioning is because the conditioning sets are identical: $\{x \mid r(1, x)=1-e\}=\{x \mid r(0, x)=e\}$. Thus

$$\begin{aligned}E\{Y(1)-Y(0) \mid r(0, X), r(1, X)\} &= E\{Y(1)-Y(0) \mid 1-r(1, X), r(1, X)\} \\ &= E\{Y(1)-Y(0) \mid r(1, X)\}.\end{aligned}$$

In contrast, there is no causal interpretation conditional on the value of the generalised propensity score:

$$\beta(1, r)-\beta(0, r)=E\{Y(1) \mid r(1, X)=r\}-E\{Y(0) \mid r(1, X)=1-r\},$$

because again the conditioning sets differ.

However, the lack of a causal interpretation within the subpopulations does not invalidate the causal interpretation after averaging over the distribution of the score. The key insight is that, by averaging these expectations over different arguments, $\beta(t, r)$ over $r(t, X)$ to get $E\{Y(t)\}$, and $\beta(s, r)$ over $r(s, X)$ to get $E\{Y(s)\}$, we achieve the causal interpretation.

Acknowledgement

I am grateful to Alan Krueger for raising the question that this paper answers, and to Tom Belin, Gary Chamberlain and Geert Ridder for comments. This research was supported by a grant from the National Science Foundation to the National Bureau of Economic Research.

References

Horvitz, D. & Thompson, D. (1952). A generalization of sampling without replacement from a finite population. *J. Am. Statist. Assoc.* **47**, 663–85.

Joffe, M. & Rosenbaum, P. (1999). Invited commentary: propensity scores. *Am. J. Epidem.* **150**, 1–7.

Lechner, M. (2000). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluations of Active Labour Market Policies in Europe*, Ed. M. Lechner and F. Pfeiffer. To appear. Heidelberg: Physica.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Robins, J. (1995). Discussion of ‘Causal diagrams in empirical research’ by J. Pearl. *Biometrika* **82**, 695–8.

Rosenbaum, P. (1987). Model-based direct adjustment. *J. Am. Statist. Assoc.* **82**, 387–94.

Rosenbaum, P. (1995). *Observational Studies.* New York: Springer Verlag.

Rosenbaum, P. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rosenbaum, P. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Statist. Assoc.* **79**, 516–24.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–92.

[*Received January* 1999. *Revised February* 2000]