TECHNICAL WORKING PAPER SERIES

THE ROLE OF PROPENSITY SCORE IN ESTIMATING DOSE-RESPONSE FUNCTIONS

Guido W. Imbens

Technical Working Paper 237
http://www.nber.org/papers/T0237

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 1999

I am grateful to Alan Krueger for raising the question that this paper answers, and to Tom Belin and Geert Ridder for comments. This research was supported by grant SBR 9818644 from the NSF to the NBER. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

The Role of the Propensity Score in
Estimating Dose-Response Functions
Guido W. Imbens
NBER Technical Working Paper No. 237
April 1999

ABSTRACT

Estimation of average treatment effects in observational, or non-experimental, 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 and Rubin (1983) 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 the bias associated with differences in pre-treatment variables between treatment and control groups. The Rosenbaum-Rubin proposals deal exclusively with the case where treatment takes on only two values. In this paper an extension of this methodology is proposed that allows for estimation of average causal effects with multi-valued treatments while maintaining the advantages of the propensity score approach.

Guido W. Imbens
Department of Economics
8256 Bunche Hall
UCLA
405 Hilgard Ave.
Los Angeles, CA 90095
and NBER
imbens@econ.ucla.edu

1. Introduction

Estimation of average treatment effects in observational, or non-experimental, 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 covariance adjustment are often inadequate. Rosenbaum and Rubin (1983, RR from hereon), Rubin and Rosenbaum (1984) and Rubin and Thomas (1992) 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 the bias associated with differences in pre-treatment variables between treatment and control groups.

The RR proposals deal exclusively with the case where treatment takes on only two values. In many cases of interest, however, the treatment takes on more than two values. A physician may choose among three or more options to treat a patient. Alternatively, a drug may be applied in different doses, or a treatment may be applied over time periods of different length. In all cases the treatment takes on more than two values and the basic propensity score methodology does not apply. In this paper an extension of this methodology is proposed that allows for estimation of average causal effects with multi-valued treatments. In addition, a weaker concept of unconfoundedness is introduced, and closer links with the concept of missing at random from the missing data literature (Rubin, 1976; Little and Rubin, 1987) are established.

2. The Basic Setup

We are interested in the causal effect of some treatment on some outcome. The treatment, denoted by T, takes on values in a set $\mathcal{T}$. RR focus on the case where the treatment is binary, or $\mathcal{T} = \{0, 1\}$. Associated with each unit i and each value of the treatment t there is a potential outcome, $Y_i(t)$. We are interested in the average outcome, $E[Y(t)]$, for all values

of t, and particularly 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. We observe for each unit i, in a random sample of size N drawn from a large population, the treatment assigned and received, 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 formalized by using the concept of unconfoundedness (RR). Two versions of unconfoundedness are used in this paper. The first, labelled strong unconfoundedness, follows the definition in RR.

Definition 1 (Strong Unconfoundedness)
Assignment to treatment T is strongly unconfounded, given pre-treatment variables X, if

$$T \perp \left\{Y(t)\right\}_{t\in\mathcal{T}} \Big| X.$$

In addition a weaker version of unconfounded is used for which one additional piece of notation is required. Let $D_i(t)$ be the indicator, for unit i, of receiving treatment t:

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

Now weak unconfoundedness is defined:

Definition 2 (Weak Unconfoundedness)
Assignment to treatment T is weakly unconfounded, given pre-treatment variables X, if

$$D(t) \perp Y(t) \Big| X,$$

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

Weak unconfoundedness relaxes two aspects of strong unconfoundedness. Strong unconfoundedness 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. 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 binary treatment level indicator $D(t)$, rather than of the treatment level T. This second difference is immaterial in the binary treatment case.

The assumption of strong unconfoundedness has no testable implications. In addition it is difficult to think of applications where the weaker form would be plausible but the stronger form would not be. The importance of the difference between the two versions is more conceptual. The weak unconfoundedness concept is closely linked to the missing data interpretation of the problem of causal inference. For units with $D_i(t) = 0$ the variable $Y_i(t)$ is missing. Given the interest in estimating the population average of $Y_i(t)$, $E[Y(t)]$, the concern is with the representativeness of the average of $Y_i(t)$ in the subsample with $D_i(t) = 1$. In that interpretation there is no direct role for the treatment level actually assigned to units with $D_i(t) = 0$. All that matters is that they did not receive treatment level t. The main role for additional variables is in allowing the researcher to adjust for differences by defining subpopulations. However, because the other potential outcomes $Y_i(s)$, for $s \neq t$ are never observed for units with $D_i(t) = 1$, they can play no role in any adjustment procedures. The definition of weak unconfoundedness reflects this lack of relevance of the other potential outcomes. As a result, the definition of weak unconfoundedness is closely tied to the definition of Missing At Random (MAR, Rubin, 1976; Little and Rubin, 1987) in the missing data literature.

Weak unconfoundedness implies that within subpopulations defined by pre-treatment variables one can estimate average outcomes by conditioning on treatment status:

$$E[Y(t)|X = x] = E[Y(t)|D(t) = 1, X = x] \tag{1}$$

$$= E[Y(t)|T = t, X = x] = E[Y|T = t, X = x].$$

Average outcomes can be estimated by averaging these conditional means:

$$E[Y(t)] = E\Big[E[Y(t)|X]\Big].$$

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. This motivated the work by Rosenbaum and Rubin (1983, 1984) who proposed an alternative based on the propensity score that circumvents the need to condition on the entire set of pre-treatment variables.

3. The Propensity Score with Binary Treatments

In this section we assume that the treatment of interest takes on two values, or $\mathcal{T} = \{0, 1\}$ and review the relevant RR results, modified where applicable to rely on weak, rather than strong, unconfoundedness. RR define the propensity score as

Definition 3 (Propensity Score, Rosenbaum and Rubin (1983))
The propensity score is the conditional probability of receiving the treatment given the pre-treatment variables:

$$e(x) \equiv Pr(T = 1|X = x).$$

The first property of the propensity score is the balancing of the pre-treatment variables. This is a mechanical result that follows from the definition of the propensity score and does not require unconfoundedness.

Lemma 1 (Balancing of Pre-treatment Variables Given the Propensity Score with Binary Treatments, Rosenbaum and Rubin (1983))
Suppose the treatment is binary. Then:

$$T \perp X \Big| e(X).$$

Proof: See Appendix.

Combined with weak unconfoundedness, the balancing property leads to the key property of the propensity score:

Lemma 2 (Weak Unconfoundedness Given the Propensity Score with Binary Treatments, Rosenbaum and Rubin, 1983)
Suppose the treatment is binary, and suppose that assignment to treatment is weakly unconfounded. Then assignment to treatment is weakly unconfounded given the propensity score:

$$D(t) \perp Y(t) \,\Big|\, e(X),$$

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

Proof: See Appendix.

This result implies that instead of having to condition on the entire set of pre-treatment variables, it is sufficient to condition on a one-dimensional function of the pre-treatment variables, the propensity score. Formally,

Theorem 1 (Adjustment for Propensity Score given Weak Unconfoundedness)
Suppose assignment to a binary treatment is weakly unconfounded. Then
(i),

$$\mu(t,e) \equiv E[Y(t)|e(X)=e] = E[Y(t)|T=t, e(X)=e],$$

(ii),

$$E[Y(t)] = E[E[\mu(t,e(X))],$$

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

Proof: See Appendix.

An alternative method for exploiting the propensity score is through weighting by the inverse of the probability of receiving the treatment actually received, as in the Horvitz-Thompson (Horvitz and Thompson, 1952) estimator:

Theorem 2 (Weighting and the Propensity Score)

Suppose assignment to a binary treatment is weakly unconfounded. Then

$$E\left[\frac{Y \cdot T}{e(X)}\right] = E[Y(1)], \qquad \text{and,} \quad E\left[\frac{Y \cdot (1-T)}{1-e(X)}\right] = E[Y(0)].$$

Proof: See Appendix.

In many applications the researcher does not know the propensity score. In that case one has to estimate the propensity score before being able to use it for conditioning or weighting purposes in estimating average treatment effects. In practice, however, inference for average treatment effects is often less sensitive to specifications of the propensity score than to specification of the conditional expectation of the potential outcomes, and the two-step approach to inference through estimation of the propensity score has often been more effective than direct adjustment methods. Since the early work of Rosenbaum and Rubin (1983, 1984) these methods have generated widespread interest in various fields such as epidemiology (e.g., Harrel et al, 1990; Robins and Greenland, 1986), and social sciences (e.g., Dehejia and Wahba, 1998; Hahn, 1998; Heckman, Ichimura, Smith and Todd, 1998).

4. Multi-valued Treatments

In this section we allow the treatment of interest to take on integer values between 0 and K, or $\mathcal{T} = \{0, 1, \ldots, K\}$. We first modify the RR definition of the propensity score:

Definition 4 (Generalized Propensity Score)

The Generalized Propensity Score (GPS) is the conditional probability of receiving a particular level of the treatment given the pre-treatment variables:

$$r(t, x) \equiv Pr(T = t | X = x) = E[D(t)|X = x].$$

In terms of this notation the RR definition of the propensity score is $e(x) = r(1, x)$. The GPS will be used in different ways. First, it defines a single random variable as a transformation of

the two random variables T and X: $r(T, X)$. Second, it defines a family of random variables indexed by t as transformations of X alone: $r(t, X)$, for all $t \in \mathcal{T}$.

Similar to the standard propensity score, the GPS satisfies a balancing property by construction:

Lemma 3 (Balancing Given the Generalized Propensity Score)

$$D(t) \perp X \,\Big|\, r(t, X),$$

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

Proof: See Appendix.

First note that the conditioning argument changes with the level of the treatment. It is in general not true that $D(s) \perp X | r(t, X)$ for $s \neq t$.

Also note that in the binary case $T = D(1)$ and $r(1, X) = e(X)$, and hence we have conditional independence of T and X given the score. This is not true in general with multi-valued treatments. To guarantee conditional independence of a multi-valued treatment T and the pre-treatment variables X one would need to condition on the entire set of $K + 1$ scores $\{r(t, X)\}_{t \in \mathcal{T}}$. It is only in the binary treatment case that conditioning on the set $\{r(t, X)\}_{t \in \mathcal{T}}$, in that case equal to $\{r(0, X), r(1, X)\} = \{1 - r(1, X), r(1, X)\}$, is identical to conditioning only on a single score (the propensity score $e(X)$) because of the adding up of the assignment probabilities.

As in the binary treatment case, the balancing property does not require any form of unconfoundedness. It is a mechanical result implied by the definition of the score. It is however in combination with unconfoundedness that this balancing property is useful.

Lemma 4 (Weak Unconfoundedness Given the Generalized Propensity Score) *Suppose assignment to treatment* T *is weakly unconfounded given pre-treatment variables* X. *Then:*

$$D(t) \perp Y(t) \,\Big|\, r(t, X),$$

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

Proof: See Appendix.

Lemma 4 leads to the main result in the paper that one can estimate average outcomes by conditioning solely on the GPS.

Theorem 3 (Estimation of Average Potential Outcomes through Adjustment for the Generalized Propensity Score)
Suppose assignment to treatment is weakly unconfounded given pre-treatment variables X. *Then*
(i),

$$\beta(t,r) = E[Y(t)|r(t,X) = r] = E[Y|T = t, r(T,X) = r],$$

(ii),

$$E[Y(t)] = E[\beta(t, r(t,X))],$$

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

Proof: See Appendix.

Consider $\beta(t,r)$. It represents the conditional expectation of the outcome with two conditioning arguments. The first conditioning argument is the treatment level T. The second is the probability of receiving the treatment that was actually received, $r(T,X)$. To obtain the population average of $E[Y(t)]$ this conditional expectation is then averaged, evaluated at treatment level t and the probability of receiving treatment level t, $r(t,X)$. Note that the averaging is <u>not</u> after evaluating $\beta(t,r)$ at treatment level t and the probability of receiving the treatment actually received, $r(T,X)$.

As an alternative to the conditioning argument in Theorem 3, as in the binary case, one can use the score to weight the observations, using the following equality:

Theorem 4 (Weighting and the Generalized Propensity Score)
Suppose assignment to treatment is weakly unconfounded. Then

$$E\left[\frac{Y \cdot D(t)}{r(T,X)}\right] = E[Y(t)].$$

Proof: See Appendix.

5. Comparison with Binary Treatments

Here we discuss the key difference between the RR and the current approach. The RR 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 the difference in sample averages by treatment status is unbiased for the expected difference of $Y(1) - Y(0)$ within that subpopulation, that is, it is unbiased for the average causal effect. Another way of stating this is that the regression of the observed outcome on treatment level and propensity score has a causal interpretation. To get an estimate of the population average causal effect one then adds up the within-subpopulation estimates, weighted by population shares.

The GPS also partitions the population in subpopulations. Consider the subpopulation defined by $r(T,X) = r$. Within this subpopulation the average value of $Y(t)$ for units with treatment level t is an unbiased estimate of the average of $Y(t)$ for the subpopulation with $r(t,X) = r$. The reason is that this subpopulation with $T = t$ and $r(T,X) = r$ is the same as the subpopulation with $T = t$ and $r(t,X) = r$. However, the average of $Y(s)$ for units with $T = s$ in the same subpopulation with $r(T,X) = r$ is unbiased for the average of $Y(s)$ in a different subpopulation, namely that with $r(s,X) = r$. Hence no causal comparisons 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 $r(T,X)$ has no causal interpretation.

More formally, with

$$\beta(t,r) = E[Y|T=t, r(T,X)=r],$$

consider the difference $\beta(t,r) - \beta(s,r)$:

$$\beta(t,r) - \beta(s,r) = E\Big[Y(t)\Big|T=t, r(T,X)=r\Big] - E\Big[Y(s)\Big|T=s, r(T,X)=r\Big].$$

By weak unconfoundedness this is equal to

$$E\Big[Y(t)\Big|r(t,X)=r\Big] - E\Big[Y(s)\Big|r(s,X)=r\Big],$$

but there is no causal interpretation for this difference because the conditioning sets differ: $\{x|r(t,x)=r\} \neq \{x|r(s,x)=r\}$. To obtain a causal interpretation one needs to contract the conditioning set to the intersection of the two conditioning sets:

$$\begin{aligned} &E\Big[Y(t)\Big|T=t, r(t,X), r(s,X)\Big] - E\Big[Y(s)\Big|T=s, r(t,X), r(s,X)\Big] \\ &\quad = E\Big[Y(t)-Y(s)\Big|r(t,X), r(s,X)\Big]. \end{aligned}$$

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 expansion of the conditioning set can be avoided while still obtaining a causal interpretation of differences in outcomes by treatment status by virtue of the adding up of the two assignment probabilities. Consider the binary case with $K=1$ and the propensity score $e(x) = r(1,x)$. In their discussion of the propensity score methodology, RR demonstrate that conditional on the propensity score outcome differences by treatment status are unbiased for average treatment effects:

$$E\Big[Y(1)\Big|T=1, e(X)=e\Big] - E\Big[Y(0)\Big|T=0, e(X)=e\Big] = E\Big[Y(1)-Y(0)\Big|e(X)=e\Big].$$

To see the difference with the GPS, let us rewrite this in the GPS notation:

$$\beta(1,e) - \beta(0,1-e) = E\Big[Y(1)\Big|T=1, r(1,X)=e\Big] - E\Big[Y(0)\Big|T=0, r(0,X)=1-e\Big]$$

$$= E\Big[Y(1) - Y(0)\Big|r(1,X) = e\Big].$$

The reason this causal comparison requires no additional conditioning is because the conditioning sets are identical:

$$\Big\{x\Big|r(1,x) = 1 - e\Big\} = \Big\{x\Big|r(0,x) = e\Big\},$$

and hence

$$E\Big[Y(1) - Y(0)\Big|r(0,X), r(1,X)\Big] = E\Big[Y(1) - Y(0)\Big|1 - r(1,X), r(1,X)\Big]$$

$$= E\Big[Y(1) - Y(0)\Big|r(1,X)\Big].$$

In contrast, there is still no causal interpretation for the comparison conditional on the value of the GPS:

$$\beta(1,r) - \beta(0,r) = E\Big[Y(1)\Big|T = 1, r(1,X) = r\Big] - E\Big[Y(0)\Big|T = 0, r(0,X) = r\Big]$$

$$= E\Big[Y(1)\Big|r(1,X) = r\Big] - E\Big[Y(0)\Big|r(1,X) = 1 - r\Big],$$

because again the conditioning sets differ.

However, the lack of a causal interpretation of within the subpopulations does not invalidate the causal interpreation after averaging over the distribution of the score.

6. Implementation

Similar to the implementation of the binary treatment propensity score methodology, the implementation of the GPS method consists of three steps.

In the first step the score $r(t,x)$ is estimated. With a binary treatment the standard approach (e.g., Rubin and Rosenbaum, 1984; Rosenbaum, 1995) is to estimate the propensity score using a logistic regression. With a multi-valued treatment one may distinguish two cases of interest. First, consider the case where the values of the treatment are qualitatively distinct and without a logical ordering. For example, a physician may be choosing from

a set of distinct treatments, e.g., surgery, drug treatment, no treatment. In that case one may wish to use discrete response models. For example extensions of the logistic regression model such as multinomial logit or nested logit models may be appropriate. In the second case of interest the treatments correspond to ordered levels of a treatment. This may reflect the dose of a drug, or the time over which a treatment is applied. The ordering is likely to come with a corresponding degree of belief that the probability of receiving the treatment is a smooth function of the level of the treatment. In that case one may wish to impose smoothness of the score in t.

In the second step the conditional expectation $\beta(t,r) = E[Y|T = t, r(T,X) = r]$ of the outcome given treatment level t and the probability of receiving the treatment received $r(T,X)$ is estimated. Again there is a distinction between the case where the levels of the treatment are qualitatively distinct and the case where smoothness of the conditional expectation function in t is appropriate. In the first case estimation of the conditional expectations should be separate for different levels of the treatment. In the second case one may wish to impose smoothness, for example by constructing blocks based on values of both treatment and score, and within the blocks use covariance adjustment methods.

In the third step the average response at treatment level t, $\beta(t) = E[\beta(t, r(t,X))]$ is estimated as the average of the estimated conditional expectation, $\hat{\beta}(t, r(t,X))$, averaged over the appropriate distribution of the pre-treatment variables. Here the choice is in the appropriate distribution of the pre-treatment variables. The most obvious choice is the empirical distribution. In that case the estimand is the population average of the potential outcomes. However, this need not be the relevant distribution. It may be that the researcher is interested in the dose-response function for a particular subpopulation. In that case the averaging should be over the distribution of pre-treatment variables in that subpopulation. Another reason for choosing a different distribution of pre-treatment variables to average over concerns the resulting precision. If for some values of the pre-treatment variables, common in the population, particular levels of the treatment are rare, estimates of the population

average dose-response function are will be imprecise for those levels of the treatment. In that case it may be that by choosing a different distribution of pre-treatment variables a dose-response function can be precisely estimated over a larger range of treatment levels.

7. Overlap in the Distribution of Pre-treatment Variables

No procedure for adjusting for pre-treatment differences is likely to work well if there is insufficient overlap in the distribution of pre-treatment variables by treatment status. It is often difficult to assess whether there is sufficient overlap when there are many pre-treatment variables. It may be that there is considerable overlap in each pair of marginal covariate distributions without overlap in the joint distributions. RR suggest inspecting the distribution of the propensity score by treatment status as a descriptive tool for investigating the overlap. A sufficient condition for overlap in the joint pre-treatment variable distributions is that there is overlap in the two marginal propensity score distributions. The two distributions that are compared in the binary treatment case are $f(e(X)|T=1)$ and $f(e(X)|T=0)$. If the distributions show sufficient overlap, attempts to draw causal inferences are more likely to lead to satisfactory inferences.

A similar graphical diagnostic is avaiable for the multi-valued treatment case. Because the treatment can take on many values, the comparison requires more than two univariate distributions. Instead we compare for each value of t the univariate distribution of $r(t,X)$ conditional on $T=t$ with the same distribution conditional on $T\neq t$. As in the binary treatment case, the concern is with differences in the two distributions. If for a given value of t the distribution of $r(t,X)$ conditional on $T=t$ is similar to that conditional on $T\neq t$, then all adjustment methods are likely to perform well. If, however, there are substantial differences, the choice of adjustment method becomes more important and methods relying on functional form are likely to lose robustness. If one graphs the probability distribution of $r(t,X)$ given $T=t$ for all values of $t\in\mathcal{T}$ in a single graph, and similarly for the distribution of $r(t,X)$ conditional on $T\neq t$, the diagnostic requires the comparison of two

three dimensional graphs.

In applications it may be the case that there is sufficient overlap for pairs of values of the treatment for ranges of the pre-treatment variables, but not for others. Let us consider an example with three levels of the treatment, or $\mathcal{T} = \{0, 1, 2\}$, in some detail. Suppose there is a single pre-treatment variable X, uniformly distributed over the interval $[0, 3]$. If $X_i \in [0, 1)$, then with probability $(1-\varepsilon)/2$, for some ε close to zero, unit i receives treatment 0, and with probability $(1-\varepsilon)/2$ the unit receives treatment 1. If $X_i \in [1, 2)$, then with probability $(1-\varepsilon)/2$ unit i receives treatment 1, and with probability $(1-\varepsilon)/2$ the unit receives treatment 2, and finally $X_i \in [2, 3]$, then with probability $(1-\varepsilon)/2$ unit i receives treatment 0, and with probability $(1-\varepsilon)/2$ the unit receives treatment 2. Assume in addition that treatment assignment is weakly unconfounded. The score is

$$r(t,x) = \begin{cases} (1-\varepsilon)/2 & \text{if } t = 0,\ x \in [0,1) \cup [2,3], \\ \varepsilon & \text{if } t = 0,\ x \in [1,2), \\ (1-\varepsilon)/2 & \text{if } t = 1,\ x \in [0,2), \\ \varepsilon & \text{if } t = 1,\ x \in [1,3], \\ (1-\varepsilon)/2 & \text{if } t = 2,\ x \in [1,3], \\ \varepsilon & \text{if } t = 2,\ x \in [0,1), \\ 0 & \text{otherwise.} \end{cases}$$

Now consider estimation of the average potential outcome $Y(0)$. We cannot accurately estimate the population average because with $X \in [1, 2)$ there are few units assigned to treatment level 0. Similarly, the population average of $Y(1)$ cannot be precisely estimated because there are few units with $T = 1$ in the range $X \in [0, 1)$, and the population average of $Y(2)$ cannot be precisely estimated because there are few units with $T = 2$ and $X \in [1, 2)$. Hence we cannot make precise statements regarding the population effect of *any* of the treatments relative to any other. However, we can make precise statements regarding the relative effect of treatment level 1 versus treatment level 0 for the subpopulation with $X \in [0,)$. Similarly we can make precise statements regarding the relative effect of any other combination of treatments for different subpopulations.

8. Conclusion

In this paper an extension of the propensity score methodology developed by Rosenbaum and Rubin (1983) is proposed to deal with mult-valued treatments. As in the binary treatment case, this methodology allows the researcher to avoid estimating a conditional expectation of the outcome of interest as a function of all pre-treatment variables and instead requires only estimation of this conditional expectation as a function of one variable for each level of the treatment. In order to achieve this a weaker version of unconfoundedness is introduced that highlights links to the missing data literature.

Appendix

First I prove Lemmas 3 and 4, and Theorems 3 and 4. The earlier lemmas and theorems are then shown to be special cases of these results.

Proof of Lemma 3:

First,

$$Pr(D(t)=1|X,r(t,X))=E[D(t)|X,r(t,X)]=E[D(t)|X]=r(t,X),$$

because by definition $r(t,X)=E[D(t)|X]$. Second,

$$Pr(D(t)=1|r(t,X))=E[D(t)|r(t,X)]=E\Big[E[D(t)|X,r(t,X)]\Big|r(t,X)\Big]$$

$$E[r(t,X)|r(t,X)]=r(t,X).$$

Hence $Pr(D(t)=1|X,r(t,X))=Pr(D(t)=1|r(t,X))$ and conditionally on $r(t,X)$ the treatment indicator $D(t)$ and the pre-treatment variables X are independent. $\mathcal{QED}$.

Proof of Lemma 4:

First,

$$\begin{aligned}Pr(D(t)=1|Y(t),r(t,X))&=E[D(t)|Y(t),r(t,X)]\\&=E\Big[E[D(t)|Y(t),X,r(t,X)]\Big|Y(t),r(t,X)\Big]\\&=E[r(t,X)|Y(t),r(t,X)]=r(t,X).\end{aligned}$$

Second, as shown before in the proof for Lemma 3, $Pr(D(t)=1|r(t,X))=r(t,X)$. Hence $Pr(D(t)=1|Y(t),r(t,X))=Pr(D(t)=1|r(t,X))$, and conditionally on $r(t,X)$ the treatment indicator $D(t)$ and the potential outcome $Y(t)$ are independent. $\mathcal{QED}$.

Proof of Theorem 3:

Part (i). First,

$$E[Y|T=t,r(T,X)=r]=E[Y(t)|T=t,r(T,X)=r]$$

$$= E[Y(t)|T = t, r(t,X) = r] = E[Y(t)|D(t) = 1, r(t,X) = r],$$

which by unconfoundedness is equal to

$$E[Y(t)|r(t,X) = r].$$

Part (ii) follows directly by applying iterated expectations.

Proof of Theorem 4:

First rewrite the expectation as an iterated expectation with the inner expectation conditional on X:

$$E\left[\frac{Y \cdot D(t)}{r(T,X)}\right] = E\Big[E\Big[\frac{Y \cdot D(t)}{r(T,X)}\Big|X\Big]\Big].$$

Next, by conditioning on $D(t) = 1$ and multiplying by the probability of $D(t) = 1$ this is equal to

$$E\Big[E\Big[\frac{Y}{r(T,X)}\Big|D(t) = 1, X\Big] \cdot Pr(D(t) = 1|X)\Big].$$

Conditional on $D(t) = 1$, $Y = Y(t)$ and $r(T,X) = r(t,X)$, so this can be rewritten as:

$$E\Big[E\Big[\frac{Y(t)}{r(t,X)}\Big|D(t) = 1, X\Big] \cdot Pr(D(t) = 1|X)\Big].$$

Now the conditioning on $D(t)$ is irrelevant by the weak unconfoundedness assumption, so this is equal to:

$$E\Big[E\Big[\frac{Y(t)}{r(t,X)}\Big|X\Big] \cdot Pr(D(t) = 1|X)\Big].$$

Because $Pr(D(t) = 1|X) = r(t,X)$, this can be written as:

$$E\Big[E\Big[\frac{Y(t)}{r(t,X)}\Big|X\Big] \cdot r(t,X)\Big] = E\Big[E\Big[Y(t)\Big|X\Big]\Big] = E[Y(t)].$$

$\mathcal{QED}$.

Proof of Lemma 1:

With T binary, $D(1) = T$, and $e(X) = r(T,X)(1)$. Because the assumptions in Lemma 1 are the same as those in Lemma 3, it follows that

$$D(t) \perp X \,\Big|\, r(t,X),$$

for all $t \in \mathcal{T}$, and hence

$$D(1) \perp X \mid r(1, X).$$

With T binary, $D(1) = T$, and $e(X) = r(T, X)(1)$, so

$$T \perp X \mid e(X).$$

$\mathcal{QED}$.

Proof of Lemma 2:

The assumptions for Lemma 4 are satisfied. Hence

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

for all $t \in \mathcal{T}$, implying

$$D(1) \perp Y(1) \mid r(1, X),$$

and

$$D(0) \perp Y(0) \mid r(0, X).$$

With T binary the first is equivalent to

$$D(1) \perp Y(1) \mid e(X),$$

and the second to

$$D(0) \perp Y(0) \mid 1 - e(X),$$

with the latter expression equivalent to

$$D(0) \perp Y(0) \mid e(X).$$

Hence

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

for all t. $\mathcal{QED}$.

Proof of Theorem 1:

This follows directly from the proof for Theorem 3. $\mathcal{QED}$.

Proof of Theorem 2:

This follows directly from the proof for Theorem 4 $\mathcal{QED}$.

References

Dehejia, R., and S. Wahba, (1998), "Causal Effects in Non–experimental Studies: Re–evaluating the Evaluation of Training Programs," forthcoming, *Journal of the American Statistical Association*.

Hahn, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects" *Econometrica*, Vol. 66, 315-322.

Harrell, F., S. Marcus, P. Layde, S. Broste, E. Cook, D. Wagner, L. Muhlbaier, and S. Peck, (1990), "Statistical Methods in Support", *Journal of Clinical Epidimiology*, Vol. 43, S89-S98.

Heckman, J., H. Ichimura, J. Smith, and P. Todd, (1998), "Characterizing Selection Bias Using Experimental Data", *Econometrica*, Vol. 66, No. 5, p 1017–1098.

Horvitz, D., and D. Thompson, (1952), "A Generalization of Sampling Without Replacement from a Finite Population", *Journal of the American Statistical Association*, Vol. 47, 663-685.

Little, R., and D. Rubin, (1987), *Statistical Analysis with Missing Data*, New York, Wiley.

Robins, J., and S. Greenland, (1986), "The Role of Model Selection in Causal Inference from Nonexperimental Data", *American Journal of Epidemiology*, Vol. 123.

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

Rosenbaum, P., and D. Rubin, (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika*, 70, 1, 41–55.

Rosenbaum, P., and D. Rubin, (1984), "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, Vol 79, 516–524.

Rubin, D., (1976), "Inference and Missing Data," *Biometrika* 63, 581–592.

Rubin D., and N. Thomas, (1992), "Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Covariates", *Biometrika*, Vol. 79, 797-809.