Health Services & Outcomes Research Methodology 2:259–278, 2001

Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization

KEISUKE HIRANO
Department of Economics, University of Miami, PO Box 248126, Coral Gables, FL 33124-6550
E-mail: khirano@miami.edu

GUIDO W. IMBENS
Department of Economics, University of California, 549 Evans Hall, #3880, Berkeley, CA 94720-3880
E-mail: imbens@econ.berkeley.edu

Received February 2, 2001; revised August 22, 2001; accepted January 3, 2002

Abstract. We consider methods for estimating causal effects of treatments when treatment assignment is unconfounded with outcomes conditional on a possibly large set of covariates. Robins and Rotnitzky (1995) suggested combining regression adjustment with weighting based on the propensity score (Rosenbaum and Rubin, 1983). We adopt this approach, allowing for a flexible specification of both the propensity score and the regression function. We apply these methods to data on the effects of right heart catheterization (RHC) studied in Connors et al (1996), and we find that our estimator gives stable estimates over a wide range of values for the two parameters governing the selection of variables.

Keywords: casual inference, propensity score, treatment effects, right heart catheterization, variable selection

1. Introduction

A central goal of health outcomes research is to estimate the causal effect of a treatment on an outcome of interest. If assignment to treatment is based on randomization, such inferences are often straightforward. In many circumstances, however, random assignment is infeasible, either for ethical or practical reasons. Even if it is feasible, the randomization may be compromised by noncompliance and other missing data problems. Without randomization, or if the randomization is compromised by missing data problems, simple comparisons of treated and untreated outcomes will not generally yield valid estimates of causal effects. However, in some observational studies, it may be reasonable to assume that treatment assignment is unconfounded with potential outcomes conditional on a sufficiently rich set of covariates or pretreatment variables.

Given unconfoundedness, various methods have been proposed for estimating causal effects. Some rely on estimating the conditional regression function of the outcomes given covariates (e.g., Robins, Rotnitzky and Zhao, 1995; Robins and Rotnitzky, 1995; Hahn, 1998; Heckman, Ichimura and Todd, 1997, 1998). Others use the propensity score (Rosenbaum and Rubin, 1983, 1985) in matching procedures or regression adjustment.

Hirano, Imbens and Ridder (2000) propose a Horvitz-Thompson type estimator based on weighting by the inverse of the assignment probabilities, with the assignment probabilities estimated nonparametrically. Abadie and Imbens (2001) suggest a matching procedure that combines pairwise matching without replacement and covariance adjustment.

Many of the existing estimators for average causal effects under unconfoundedness require the researcher to make a large number of choices concerning which variables to include in the specification of the propensity score and/or the specification of the conditional mean of the outcome. In this paper we propose a specific class of estimators of average causal effects that requires relatively few decisions to be made operational. Our estimators use a flexible estimate of the propensity score to construct weights, and uses these weights in a weighted regression of the outcome on treatment and covariates. It is based on earlier work (Hirano, Imbens and Ridder, 2000) that shows that if the propensity score is estimated in a sufficiently flexible manner, a weighting-based estimator can achieve the semiparametric efficiency bound for estimation of average causal effects calculated by Robins, Rotnitzky and Zhao (1995) and Hahn (1998). Robins and Rotnitzky (1995) have suggested combining such weighting with regression adjustment, and demonstrated consistency under the assumption that a parametric model applies to either the propensity score or the regression function, but not necessarily to both. We use a simple criterion to select which of a potentially large set of covariates enter into the construction of the weights and in the regression adjustment. In particular, we propose to base these decisions on the strength of the marginal correlation between the treatment and each of the covariates separately, and on the conditional correlation of the outcome and each of the covariates given the treatment. Each estimator in the class we propose is characterized by two cutoff values; the first governs the restrictiveness of the specification of the two regression functions, and the second governs the restrictiveness the specification of the propensity score. This general class of estimators includes a number of standard ones, such as the simple difference in average treatment and control outcomes, and the estimator that adjusts for all covariates through regression, as well as estimators that rely purely on weighting to remove bias.

We apply these estimators to data on the effect of Right Heart Catherization (RHC), previously analyzed by Connors et al. (1996). We find that for intermediate values of the variable selection parameters, our estimator gives more stable estimates than for values that rely solely on regression adjustment or solely on propensity score weighting. We conclude that in practice one may wish to combine regression adjustment and weighting rather than rely solely on one of these methods to remove bias.

2. Efficient Estimation of Average Causal Effects under Unconfounded Treatment Assignment

We begin by reviewing some recent work on estimation of treatment effects, and propose an estimator that combines weighting based on the estimated propensity score, with regression adjustment.

2.1. Basic Setup and Weighting Estimators

Suppose we have a random sample of size N from a large population. For each unit i in the sample, let T_i indicate whether the treatment of interest was received, with $T_i = 1$ if unit i received the treatment of interest, and $T_i = 0$ if unit i received the control treatment. Using the potential outcomes notation (e.g., Rubin, 1974), let $Y_i(0)$ denote the outcome for unit i under control and $Y_i(1)$ the outcome under treatment. For unit i the treatment effect is $Y_i(1) - Y_i(0)$. We are interested in the average effect of the treatment in the population,

$$\tau = E[Y_i(1) - Y_i(0)].$$

However we will also discuss methods of estimating the average effect on the treated,

$$\tau_t = E[Y_i(1) - Y_i(0)|T_i = 1],$$

which is of interest if one wishes to evaluate the effect of the treatment on the subpopulation that is likely to take up the treatment. The difficulty in estimating either of these average treatment effects is that we only observe $Y_i(0)$ or $Y_i(1)$, but never both. Formally, we observe T_i and Y_i, where

$$Y_i = Y_i(T_i) = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0).$$

In addition, we observe a K-dimensional vector of pre-treatment variables, or covariates, denoted by X_i.

Throughout the analysis we make the unconfoundedness assumption (Rubin, 1978; Rosenbaum and Rubin, 1983), which asserts that conditional on the pre-treatment variables, the treatment indicator is independent of the potential outcomes:

$$T \perp (Y(0), Y(1))|X. \tag{1}$$

(Here and at other points below, we suppress the i subscript for notational conveniance.)

In other words, within subpopulations defined by values of the covariates, we have random assignment. In addition we assume that for all values of the covariates the probability of receiving either treatment is strictly positive. Formally, defining the propensity score as

$$e(x) = Pr(T = 1|X = x),$$

we asssume that

$$0 < e(x) < 1,$$

for all x.

Both these assumptions may be controversial in applications. The first assumption requires that all variables that affect both outcome and the likelihood of receiving the

treatment are observed. Although this is not testable, it clearly is a very strong assumption, and one that need not generally be applicable. We view it as a useful starting point for two reasons. One is that in some studies, like the Connors et al. (1996) study of right heart catherization, researchers have carefully investigated which variables are most likely to confound any comparison between treated and control units and made attempts to observe all such variables. Even if these attempts are not completely successful, the assumption that all relevant variables are observed may be a reasonable approximation, especially if much information about pre-treatment outcomes is available. Second, any alternative assumption that does not rely on unconfoundedness while allowing for consistent estimation of the average treatment effects must make alternative untestable assumptions. Whereas the unconfoundedness assumption implies that the best matches are units that differ only in their treatment status, but otherwise are identical, alternative assumptions implicitly match units that differ in the pre-treatment characteristics. Often such assumptions are even more difficult to justify. The unconfoundedness assumption therefore may be a natural starting point after comparing average outcomes for treated and control units to adjust for observable pre-treatment differences.

The second assumption, that the propensity score is bounded away from zero and one, is in principle testable. If there are values of the covariates for which the probability of receiving the treatment is zero or one, we cannot compare treated and control units at such values. In that case we have to limit comparisons to sets of values where there is sufficient overlap in the propensity score among treated and controls. For further discussion see Rubin (1977) and Heckman, Ichimura and Todd (1997).

2.2. Regression Adjustment

The unconfoundedness assumption (1) validates the comparison of treated and control units with the same value of the covariates. The treatment effect for the subpopulation with $X = x$ can be written as:

$$\begin{aligned}\tau(x) &= E[Y(1) - Y(0)|X = x] \\ &= E[Y(1)|T = 1, X = x] - E[Y(0)|T = 0, X = x] \\ &= E[Y|T = 1, X = x] - E[Y|T = 0, X = x],\end{aligned}$$

where both terms on the right-hand side can be estimated from a random sample of (Y, T, X). The average treatment effect τ can then be estimated using the equality

$$\tau = E[\tau(X)].$$

One way to implement this approach is to approximate the two conditional means by linear functions (e.g., Rubin, 1977):

$$E[Y|T = t, X = x] = \beta_{t0} + \beta_{t1}'x.$$

One can then estimate the parameters of these two regression functions by least squares methods applied separately to the subsamples of treated and control units, and estimate the average treatment effect as

$$\hat{\tau} = \hat{\beta}_{10} - \hat{\beta}_{00} + (\hat{\beta}_{11} - \hat{\beta}_{01})'\bar{x}, \tag{2}$$

where $\bar{x}$ is the sample average of the covariates. An alternative way of writing this estimator is as the least squares estimate of τ in the expanded regression on the entire sample

$$Y_i = \alpha_0 + \tau \cdot T_i + \alpha_1' X_i + \alpha_2'(X_i - \bar{x}) \cdot T_i + \varepsilon_i.$$

This representation will be useful later when we combine the regression adjustment with weighting. Note that the linearity is not really restrictive, as we can include functions of the original covariates in the vector x.

2.3. Weighting using the Propensity Score

When the dimension of X is large, it may be difficult to include all covariates in the regression, and thus to estimate accurately the two regression functions $\mu_t(x) = E[Y|T = t, X = x]$. To address this problem, Rosenbaum and Rubin (1983) developed the propensity score methodology. Their key insight was that given the unconfoundedness assumption (1), treatment assignment and the potential outcomes are independent conditional on a scalar function of the covariates, the conditional probability of assignment:

$$T \perp (Y(0), Y(1)) | e(X). \tag{3}$$

Thus, adjusting for the propensity score removes the bias associated with differences in the observed covariates in the treated and control groups. One way to implement this approach is to reweight treated and control observations to make them representative of the population of interest (as in Horvitz-Thompson (1952) estimators for stratified sampling). First consider the expectation

$$E\left[\frac{Y \cdot T}{e(X)}\right].$$

Conditional on $X = x$ the expectation of this expression is

$$E\left[\frac{Y \cdot T}{e(X)}\middle| X = x\right] = E[Y(1)|X = x].$$

Hence, the marginal expectation of $YT/e(X)$ is equal to $E[Y(1)]$. More generally,

$$E\left[\frac{Y \cdot T}{e(X)} - \frac{Y \cdot (1-T)}{1-e(X)}\right] = E[Y(1) - Y(0)] = \tau.$$

This reasoning suggests the simple weighting estimator:

$$\hat{\tau} = \sum_{i=1}^{N} \frac{t_i \cdot y_i}{\hat{e}(x_i)} \Big/ \sum_{i=1}^{N} \frac{t_i}{\hat{e}(x_i)} - \sum_{i=1}^{N} \frac{(1-t_i)y_i}{1-\hat{e}(x_i)} \Big/ \sum_{i=1}^{N} \frac{1-t_i}{1-\hat{e}(x_i)}, \tag{4}$$

where $\hat{e}(x)$ is an estimate of the propensity score. Note that in this estimator we normalize the weights so they add up to one in each treatment group. Although the above argument shows that they do add up in expectation to one, they do not do so exactly without the normalization. Hirano, Imbens, and Ridder (2000) show that if the propensity score is estimated nonparametrically using a series estimator, then the resulting estimate is asymptotically efficient. This is of interest because Robins and Rotnitzky (1995), Rubin and Thomas (1996), and Hahn (1998) show that adjusting for the *true* propensity score in general leads to inefficient estimates.

2.4. Combining Weighting with Regression Adjustment

The class of estimators we consider combine weighting and regression adjustment. As Robins and Rotnizky (1995) point out, as long as only one of the models, either that for the conditional mean of $Y(0)$ and $Y(1)$ given covariates, or that for the treatment indicator given covariates, is correctly specified, the resulting estimator will be consistent. Our estimators allow for increasingly flexible models in both dimensions, and may therefore be relatively robust compared to estimators that rely on very parsimonious specifications of one of the two components. Specifically, we consider estimators based on weighted least squares estimation of the regression function

$$Y_i = \alpha_0 + \tau \cdot T_i + \alpha_1' Z_i + \alpha_2'(Z_i - \bar{Z}) \cdot T_i + \varepsilon_i,$$

where the Z_i are a subset of the covariates X_i, with sample average $\bar{Z}$. Ideally the weights we would like to use are

$$\omega(t,x) = \frac{t}{e(x)} + \frac{1-t}{1-e(x)},$$

but with the propensity score unknown we replace the true propensity score by the estimated score, leading to the weights

$$\hat{\omega}(t,x) = \frac{t}{\hat{e}(x)} + \frac{1-t}{1-\hat{e}(x)}.$$

Our estimates of the propensity score are based on a logistic regression model:

$$Pr(T_i = 1|V_i = v) = \frac{\exp(v'\gamma)}{1 + \exp(v'\gamma)},$$

where V_i is a subset of the vector of covariates X_i.

Note that this class of estimators includes both the regression estimator given in (2) (by not including any covariates in V in the logistic regression), and the weighting estimator given in (4) (by not including any covariates in Z in the linear regression).

Given that we restrict the class of estimators considered to those where both the regression functions and the propensity score are linear in a subset of the covariates, the only remaining decisions concern the choice of the subsets. The linearity is not necessarily restrictive, as we can include higher order terms and interactions of the original covariates in the vector of covariates. However, we assume that any such higher order terms are already included in Z and V, and do not consider adding to either Z or V any functions of the covariates not already included in X. Hence the problem is an example of the classic subset selection problem in regression (e.g., Miller, 1990). Here we consider a very simple pair of rules, characterized by only a single degree of freedom.

First, consider the set of variables to be included in the propensity score. With K equal to the dimension of X, we estimate K logistic regressions. The kth logistic regression specifies

$$Pr(T_i = 1|X_{ik} = x_k) = \frac{\exp(\gamma_{k0} + \gamma_{k1} \cdot x_k)}{1 + \exp(\gamma_{k0} + \gamma_{k1} \cdot x_k)}.$$

After estimating this logistic regression by maximum likelihood we compute the t-statistic for the test of the null hypothesis that the slope coefficient γ_{k1} is equal to zero. If the t-statistic is larger in absolute value than t_{prop}, this variable will be included in V, the vector of covariates used in the final specification of the propensity score. After estimating all K logistic regressions we end up with the subset of covariates whose marginal correlation with the treatment indicator is relatively high. We orthogonalize the set of selected covariates, and use these to estimate the propensity score.

Similarly, we estimate K linear regressions of the type

$$Y_i = \beta_{k0} + \beta_{k1} \cdot T_i + \beta_{k2} \cdot X_{ik} + \varepsilon_i. \tag{5}$$

Again we calculate the t-statistic for the test of the null hypothesis that the slope coefficient β_{k2} is equal to zero in each of these regressions, and now select for Z all the covariates with a t-statistic larger in absolute value than t_{reg}. Thus, we include in the final regression all covariates which have substantial correlation with the outcome conditional on the treatment. As in the propensity score component, we orthogonalize the selected covariate matrix to improve numerical stability.

The rules within this class are easy to implement, but they are not necessarily fully optimal. For example, they do not take account of correlations between the different

covariates. More complex rules, however, may require estimation of the basic models for all subsets of regressors. In examples like ours, with over seventy covariates, this could be prohibitively expensive. In addition, an advantage of our class of estimators is that it includes as special cases a number of simple, commonly used estimators, so that we can compare a number of standard approaches within one framework.

The two remaining choices are the cutoff values for the t-statistics (t_{prop}, t_{reg}). We consider all pairs with t_{prop} and t_{reg} in the set $\{0, 1, 2, 4, 8, 16, \infty\}$. Some of the pairs included in this set are of particular interest. Choosing $(t_{prop}, t_{reg}) = (\infty, \infty)$ amounts to estimating the average treatment effect by the difference in treatment-control averages. Choosing $(t_{prop}, t_{reg}) = (\infty, 0)$ amounts to estimating τ by linear regression with all covariates and no weighting. At the other extreme, $(t_{prop}, t_{reg}) = (0, \infty)$ amounts to estimating the propensity score with all covariates and using this for the weighting, but without any additional covariance adjustment. More generally we consider pairs of values that allow some variables to enter in the propensity score and some to enter in the regression, or both, depending on their correlation with the treatment and their conditional correlation with the outcome.

Finally, we consider estimation of standard errors for the estimators taking into account estimation error in the propensity score and the regression adjustment. In the Appendix, we provide expressions for the asymptotic variance of the general estimator, using standard results on M-estimators. These standard errors are conditional on the correct specification of both the regression model and the propensity score. However, they can also be justified by allowing the parametrization to become more flexible as the sample size gets larger, using, for example, the results on series estimation in Newey (1994).

2.5. The Average Treatment Effect for the Treated

The discussion so far has focused on estimation of the population average treatment effect. If we wish to estimate the average treatment effect for the treated subpopulation, two modifications are required. Instead of subtracting the population average of the covariates included in the regression in the interaction term with the treatment indicator, one should subtract the average for the treated:

$$Y_i = \alpha_0 + \tau \cdot T_i + \alpha_1' Z_i + \alpha_2'(Z_i - \bar{Z}_1) \cdot T_i + \varepsilon_i,$$

where $\bar{Z}_1$ is the sample average of Z for the subsample of the treated units. The second modification is in the weights. Instead of being equal to $1/\hat{e}(z)$ for the treated and $1/(1 - \hat{e}(z))$, for the controls, the weights are now unity for the treated units and $\hat{e}(x)/(1 - \hat{e}(x))$ for the control units:

$$\omega(t, z) = t + (1 - t) \cdot \frac{\hat{e}(z)}{1 - \hat{e}(z)}.$$

3. Re-analysis of SUPPORT data on Right Heart Catheterization

In an influential study, Connors et al. (1996) used a propensity score matching approach (Rosenbaum and Rubin, 1983) to study the effectiveness of Right Heart Catheterization (RHC) in an observational setting, using data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). RHC is a diagnostic procedure used for critically ill patients. The SUPPORT study collected data on hospitalized adult patients at 5 medical centers in the U.S. Based on information from a panel of experts a rich set of variables relating to the decision to perform the RHC, as well as detailed outcome data, were collected. Table 1 lists the covariates we use in our analysis. Further information about the study can be found in Connors et al. (1996) and Murphy and Cluff (1990).

Connors et al. found that after adjusting for ignorable treatment assignment conditional on a range of covariates, RHC appeared to lead to lower survival than not performing RHC. This conclusion contradicted popular perception among practitioners that RHC was beneficial. Their primary analysis matched treated and untreated patients on the basis of the propensity score, with each unit matched at most once. This approach is called "case matching" or "pairwise matching" (e.g., Rosenbaum, 1985). While simple and conceptually appealing, it is generally less efficient than the approaches we suggest here, and construction of confidence intervals and hypothesis tests that take into account the estimation of the propensity score and the matching procedure can be difficult.

We have data on 5735 individuals, 2184 treated and 3551 controls. For each individual we observe treatment status, equal to 1 if RHC was applied within 24 hours of admission, and 0 otherwise, outcome (an indicator for survival at 30 days), and 72 covariates. Table 2 gives means of the covariates by treatment status. For each covariate, we calculate the t-statistic for the difference in means between untreated and treated observations. It is clear that the two treatment groups differ signficantly on many of the covariates.

We estimate the propensity score, using the logistic model outlined above. Figure 1 shows the distribution of estimated propensity scores, where all the covariates have been used in the specification of the propensity score. Separate histograms are generated for the control and treated groups. Our main concern here is to make sure that there is sufficient overlap between the two groups. We see that while the two groups obviously differ, in both groups the support of the estimated propensity scores is nearly the entire unit interval. Figure 2 shows the distribution of estimated propensity scores, where the variable selection rule has been applied to select the regressors in the logistic regression. The variable selection rule applied here selects all covariates that have a t-statistic at least equal to two. This rule led us to include 56 of the 72 covariates. Qualitatively, the estimates of the propensity score based on this selection of covariates leads to a figure similar to Figure 1.

The last two columns of Table 2 give the means of the control and treated groups, after weighting based on the estimated propensity score with variable selection (leading to the inclusion of 56 out of the 72 covariates). The weighting brings most of the means much closer together, although a few variables become slightly less balanced after the weighting. For most of the covariates, however, weighting by the propensity score appears to balance the control and treatment groups extremely well.

Table 1. SUPPORT Covariates

age	Age (years)
sex	Female
raceblack	Black
raceother	Other
edu	Education (years)
income1	Income $11–$25k
income2	Income $25–$50k
income3	Income > $50k
ins_care	Medicare
ins_pcare	Private & Medicare
ins_caid	Medicaid
ins_no	No Insurance
ins_carecaid	Medicare & Medicaid
cat1_copd	COPD
cat1_mosfsep	MOSF w/Sepsis
cat1_mosfmal	MOSF w/Malignancy
cat1_chf	CHF
cat1_coma	Coma
cat1_cirr	Cirrhosis
cat1_lung	Lung Cancer
cat1_colon	Colon Cancer
cat2_mosfsep	MOSF w/Sepsis
cat2_coma	Coma
cat2_mosfmal	MOSF w/Malignancy
cat2_lung	Lung Cancer
cat2_cirr	Cirrhosis
cat2_colon	Colon Cancer
resp	Respiratory diagnosis
card	Cardiovascular diagnosis
neuro	Neurological diagnosis
gastr	Gastrointestinal diagnosis
renal	Renal diagnosis
meta	Metabolic diagnosis
hema	Hematological diagnosis
seps	Sepsis diagnosis
trauma	Trauma diagnosis
ortho	Orthopedic diagnosis
das2d3pc	DASI — Duke Activity Status Index
dnr1	Do Not Resuscitate status on day 1
ca_yes	Cancer — localized
ca_meta	Cancer — metastatic
surv2md1	Estimate of prob. of surviving 2 months
aps1	APACHE score
scoma1	Glasgow coma score
wtkilo1	Weight
temp1	Temperature
meanbp1	Mean Blood Pressure
resp1	Respiratory Rate
hrt1	Heart Rate
pafi1	PaO2/FI02 ratio

Table 1. *continued*

paco21	PaCO2
ph1	PH
wblc1	WBC
hema1	Hematocrit
sod1	Sodium
pot1	Potassium
crea1	Creatinine
bili1	Bilirubin
alb1	Albumin
cardiohx	Cardiovascular symptoms
chfhx	Congestive Heart Failure
dementhx	Dementia, stroke or cerebral infarct, Parkinson's disease
psychhx	Psychiatric history, active psychosis or severe depression
chrpulhx	Chronic pulmonary disease, severe pulmonary disease
renalhx	Chronic renal disease, chronic hemodialysis or peritoneal dialysis
liverhx	Cirrhosis, hepatic failure
gibledhx	Upper GI bleeding
malighx	Solid tumor, metastatic disease, chronic leukemia/myeloma, acute leukemia, lymphoma
immunhx	Immunosuppression, organ transplant, HIV, Diabetes Mellitus, Connective Tissue Disease
transhx	transfer (> 24 hours) from another hospital
amihx	Definite myocardial infarction
wt0	weight = 0

We also calculate two matching estimators suggested by Abadie and Imbens (2001). In standard matching estimators (e.g., Rosenbaum, 1985), including the estimators used by Connors et al. (1996), each treated unit is matched to a single control, with each control being used at most once. If no adequate control unit can be found for a particular treated unit, it is discarded. This method implies that the numerical answers can actually depend on the order in which the treated units are matched. In contrast, the estimator proposed by Abadie and Imbens (2001) matches each treated unit to the closest control, and then each control to the closest treated unit, in each case with replacement. This implies that pairs are not necessarily independent, and the standard errors have to account for this. We report both a simple matching estimate based on this algorithm and a bias-adjusted estimate where, given the matched pairs, regression analysis is used to eliminate remaining bias. The simple matching estimate is -0.081 (standard error 0.017) and the bias-adjusted matching estimate is -0.063 (standard error 0.016).

Next, we turn to the estimates of the average causal effects under different choices for t_{prop} and t_{reg} as reported in Table 3. Standard errors are given in parentheses below the point estimates. Without any adjustment, either through the propensity score weighting or through regression, the estimated effect of the treatment is -0.074, based on the difference in average treatment and control outcomes. The last column presents the estimates that rely only on propensity score adjustment. There is a fairly wide range of estimates, as low as -0.074 and as high as -0.014. The bottom row presents estimates based on unweighted linear regression. Here the estimates range from -0.074 to -0.048. For comparison,

Table 2. Characteristics of untreated and treated groups

Variable	Untreated	Treated	t-stat	Untreated (weighted)	Treated (weighted)	t-stat (weighted)
age	61.76	60.74	−2.28	61.25	61.15	−0.19
sex	0.46	0.41	−3.42	0.44	0.43	−0.85
raceblack	0.16	0.15	−1.14	0.15	0.16	1.09
raceother	0.06	0.06	0.76	0.05	0.05	0.16
edu	11.56	11.85	3.35	11.68	11.71	0.39
income1	0.20	0.20	0.56	0.20	0.19	−1.19
income2	0.14	0.17	3.88	0.14	0.16	1.05
income3	0.07	0.08	2.19	0.07	0.07	0.12
ins_care	0.26	0.23	−2.79	0.25	0.23	−1.06
ins_pcare	0.21	0.22	1.26	0.22	0.21	−0.45
ins_caid	0.12	0.08	−4.77	0.11	0.11	0.01
ins_no	0.05	0.06	1.54	0.05	0.05	1.03
ins_carecaid	0.07	0.05	−2.19	0.06	0.06	0.30
cat1_copd	0.11	0.02	−13.57	0.07	0.06	−1.10
cat1_mosfsep	0.14	0.32	14.79	0.21	0.22	0.35
cat1_mosfmal	0.06	0.07	0.64	0.07	0.06	−1.09
cat1_chf	0.06	0.09	3.43	0.08	0.08	0.20
cat1_coma	0.09	0.04	−7.96	0.07	0.07	−0.02
cat1_cirr	0.04	0.02	−5.56	0.03	0.03	−0.10
cat1_lung	0.00	0.00	−3.77	0.00	0.00	−1.55
cat1_colon	0.00	0.00	−1.48	0.00	0.00	−0.65
cat2_mosfsep	0.11	0.19	7.81	0.15	0.15	0.34
cat2_coma	0.01	0.00	−3.40	0.01	0.01	−0.46
cat2_mosfmal	0.04	0.02	−4.34	0.03	0.04	0.57
cat2_lung	0.00	0.00	−2.28	0.00	0.00	−1.38
cat2_cirr	0.00	0.00	−1.22	0.00	0.00	−1.48
cat2_colon	0.00	0.00	0.32	0.00	0.00	0.73
resp	0.41	0.28	−10.01	0.37	0.36	−0.49
card	0.28	0.42	10.72	0.35	0.35	0.04
neuro	0.16	0.05	−13.74	0.11	0.10	−0.71
gastr	0.14	0.19	4.39	0.17	0.16	−0.20
renal	0.04	0.06	4.16	0.05	0.05	0.13
meta	0.04	0.04	−1.04	0.04	0.04	−0.43
hema	0.06	0.05	−2.30	0.06	0.06	0.55
seps	0.14	0.23	8.41	0.17	0.18	0.74
trauma	0.00	0.01	3.61	0.00	0.00	0.52
ortho	0.00	0.00	0.95	0.00	0.00	0.95
das2d3pc	20.37	20.70	2.32	20.37	20.64	1.42
dnr1	0.14	0.07	−8.67	0.11	0.10	−0.88
ca_yes	0.17	0.15	−2.66	0.16	0.17	0.61
ca_meta	0.07	0.05	−2.60	0.06	0.06	−0.13
surv2md1	0.60	0.56	−7.27	0.58	0.58	−0.01
aps1	50.93	60.73	18.27	55.67	56.15	0.51
scoma1	22.25	18.97	−4.09	20.87	20.65	−0.21
wtkilo1	65.04	72.36	9.47	68.69	69.10	0.31
temp1	37.63	37.59	−0.78	37.59	37.55	−0.57
meanbp1	84.86	68.19	−16.99	77.10	77.38	0.15
resp1	28.97	26.65	−6.07	28.09	28.64	0.99

Table 2. *continued*

hrt1	112.87	118.92	5.39	113.79	115.99	1.22
pafi1	240.62	192.43	−16.12	219.35	219.88	0.11
paco21	39.95	36.79	−9.43	38.90	38.50	−0.72
ph1	7.39	7.38	−4.37	7.38	7.38	0.77
wblc1	15.26	16.26	3.03	15.82	16.73	1.14
hema1	32.69	30.50	−10.10	31.59	31.42	−0.56
sod1	137.03	136.33	−3.39	136.78	136.72	−0.26
pot1	4.07	4.04	−0.99	4.15	3.97	−4.10
crea1	1.92	2.47	9.89	2.13	2.22	1.03
bili1	1.99	2.70	5.20	2.38	2.34	−0.23
alb1	3.16	2.97	−8.15	3.08	3.15	0.69
cardiohx	0.15	0.20	4.20	0.19	0.18	−0.50
chfhx	0.16	0.19	2.53	0.18	0.18	−0.08
dementhx	0.11	0.06	−6.17	0.09	0.08	−1.30
psychhx	0.08	0.04	−5.43	0.06	0.06	−0.38
chrpulhx	0.21	0.14	−7.21	0.18	0.17	−0.76
renalhx	0.04	0.04	1.15	0.05	0.04	−0.97
liverhx	0.07	0.06	−1.81	0.06	0.06	0.17
gibledhx	0.03	0.02	−2.65	0.03	0.02	−0.41
malighx	0.24	0.20	−3.75	0.22	0.23	0.45
immunhx	0.25	0.29	2.94	0.27	0.27	−0.32
transhx	0.09	0.14	6.10	0.11	0.11	0.71
amihx	0.02	0.04	2.67	0.03	0.03	0.07
wt0	0.10	0.06	−4.48	0.08	0.07	−0.59

consider the third row and column. In the third row, where the propensity score includes all covariates with a t-statistic of at least 2, for varying cutoff points for the regression adjustment, the range of estimates is $(-0.062, -0.053)$. Similarly in the third column, where all estimates are based on regression adjustment with covariates included with a t-statistic of 2, for varying specifications of the propensity score, the range is $(-0.068, -0.061)$. It is clear that by using a flexible specification of the propensity score the sensitivity to the specification of the regression function is dramatically reduced, and vice versa. We also note that these estimates agree closely with the bias-adjusted matching estimate of -0.063.

The standard errors for the estimates in Table 3 do not vary much with the specification. Most are between 0.012 and 0.016, in many cases slightly lower than for the matching estimates.

We can also examine how many of the covariates are included in the propensity score for the various cutoff points. With the cutoff point for inclusion in the propensity score equal to $t_{prop} = 0$ all 72 covariates are included. With $t_{prop} = 1$ we still include 66 of the covariates. With $t_{prop} = 2$ this goes down to 56. With $t_{prop} = 4$ only 32 of the covariates are included, and with $t_{prop} = 8$ and $t_{prop} = 16$ this goes down further to 15 and 1 respectively. Similarly, we consider the number of variables included in the regression function according to the various criteria. For $t_{reg} = 0$ again all 72 covariates are included. With $t_{reg} = 1$ 58 of the covariates are included, with $t_{reg} = 2$ we include 47 covariates, with

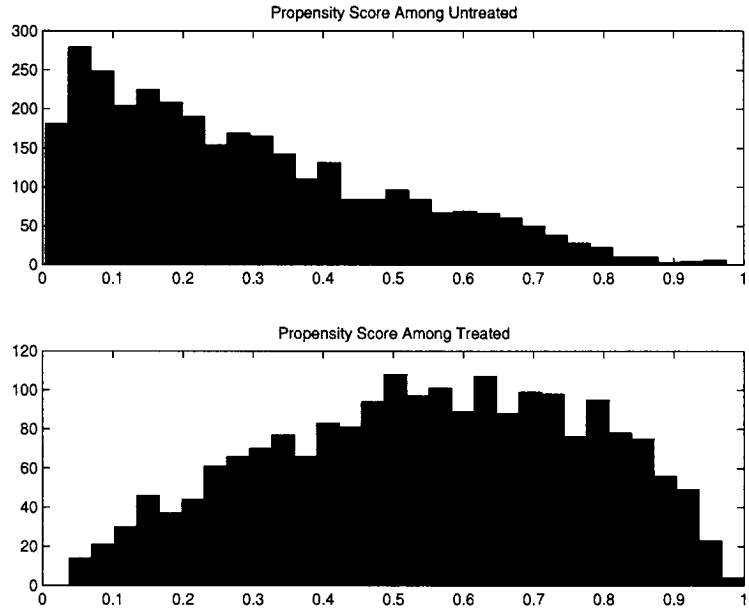

Figure 1. Propensity scores estimated using all covariates.

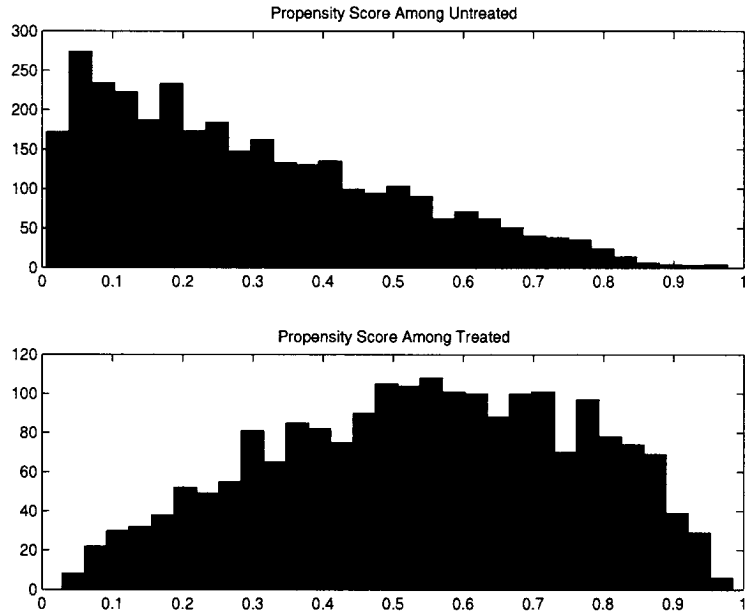

Figure 2. Propensity scores estimated using selected covariates.

Table 3. Estimates of effect of RHC (standard errors in parentheses)

		t_{reg}							
		0	1	2	4	8	16	∞	# of Cov.
t_{prop}	0	−0.062	−0.062	−0.063	−0.062	−0.061	−0.061	−0.060	72
		(0.015)	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)	(0.018)	
	1	−0.060	−0.060	−0.061	−0.059	−0.057	−0.055	−0.054	66
		(0.015)	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)	(0.018)	
	2	−0.060	−0.061	−0.062	−0.059	−0.057	−0.055	−0.053	56
		(0.015)	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)	(0.018)	
	4	−0.061	−0.063	−0.063	−0.060	−0.054	−0.054	−0.053	32
		(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.017)	
	8	−0.063	−0.064	−0.067	−0.066	−0.058	−0.059	−0.031	15
		(0.014)	(0.015)	(0.015)	(0.015)	(0.015)	(0.014)	(0.016)	
	16	−0.065	−0.067	−0.068	−0.065	−0.053	−0.048	−0.014	1
		(0.014)	(0.014)	(0.014)	(0.013)	(0.013)	(0.012)	(0.013)	
	∞	−0.065	−0.067	−0.068	−0.066	−0.054	−0.048	−0.074	0
		(0.014)	(0.014)	(0.014)	(0.013)	(0.012)	(0.012)	(0.013)	
	# of Cov.	72	58	47	29	10	4	0	

$t_{reg} = 4$ this goes down to 29, with $t_{reg} = 8$ it is 10, and with $t_{reg} = 16$ only 4 of the covariates are included.

Figure 3 presents a scatterplot of the 72 pairs of absolute values of the t-statistics used in the variable selection rules ($|t_{reg}|$, $|t_{prop}|$). To illustrate in another way the effect of our decision rule, we also calculated for each of the covariates the estimated effect of its inclusion on the bias of the average treatment effect, ignoring all the other covariates. This bias is calculated as the product of two regression coefficients. The first is the coefficient on the covariate in the regression of the outcome on the covariate and the treatment indicator, $\hat{\beta}_{k2}$ in equation 5. The second is the estimated coefficient on the treatment indicator in the regression

$$X_{ki} = \delta_{k0} + \delta_{k1} \cdot T_i + v_i.$$

The univariate bias associated with covariate k, defined as the difference between the estimated effect of the treatment in a regression where we include covariate k, $\hat{\beta}_{k1}$, and the estimated effect in a regression where we do not include covariate k, $\bar{Y}_1 - \bar{Y}_0$, is $\hat{\beta}_{k2} \cdot \hat{\delta}_{k1}$. For the covariates for which this bias is larger than 0.001 in absolute value the symbol "∗" is used in Figure 3, and for the others the symbol "o" is used. One can see that our t-statistic criterion is picking up those covariates which potentially have a large effect on the bias.

4. Conclusion

Estimation of causal effects under the unconfoundedness assumption can be challenging when the number of covariates is large and their functional relationship to the treatment

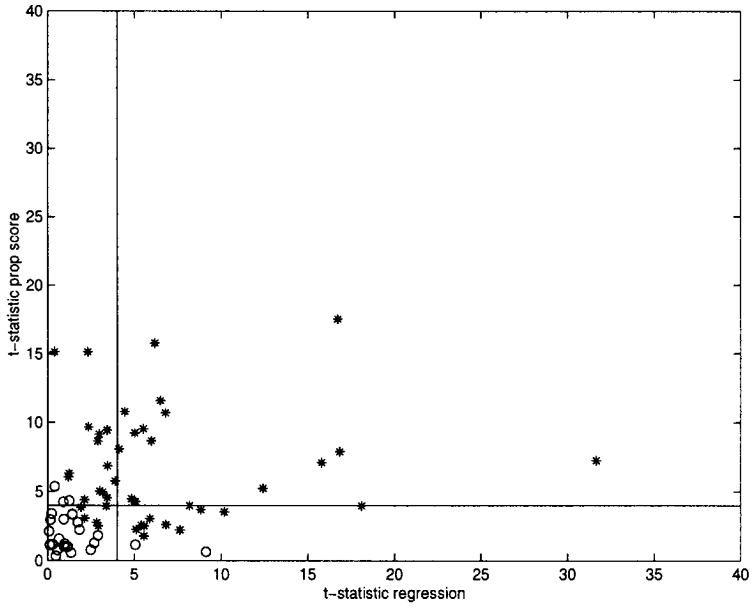

Figure 3. T-statistics for variable selection rules.

and outcome are not known precisely. By flexibly estimating both the propensity score and the conditional mean of the outcome given the treatment and the covariates, one can potentially guard against misspecification in a relatively general way. Here we propose a simple rule for deciding on the specification of the propensity score and the regression function. This rule only requires the specification of two readily interpretable cutoff values for variable selection, and is therefore relatively easy to implement and interpret. However, more work needs to be done to understand its properties, and also to investigate alternative approaches to variable selection in similar problems.

In our application to the SUPPORT study data, the estimator remained stable over a range of values for the two cutoffs, as long as some variables are included in both propensity score and regression adjustment. If no covariates are used in the regression adjustment, or no covariates are included in the propensity score, the estimates are more sensitive to the inclusion in the other component. It would be interesting to see if this robustness given both propensity score weighting and regression adjustment, compared to the sensitivity if only one method is used, extends to other applications.

Calculation of Standard Errors

Let $\alpha = (\alpha_0, \alpha_1', \alpha_2')'$, $\theta = (\gamma', \alpha')'$, and $z = (y, t, x)$. Recall that v are the variables in the logistic regression, and let w denote the regressors in the weighted regression.

Define the moment functions

$$\psi_1(z, \theta) = v\left(t - \frac{\exp(\gamma' v)}{1 + \exp(\gamma' v)}\right),$$

and

$$\begin{aligned}\psi_2(z, \theta) &= \omega(t, x) w(y - w'\alpha) \\ &= \left[\frac{t}{e} + \frac{(1-t)}{(1-e)}\right] w(y - w'\alpha) \\ &= [1 + \exp(\gamma' v)]\left(\frac{t}{\exp(\gamma' v)} + (1-t)\right) w(y - w'\alpha).\end{aligned}$$

The estimator can be defined as the solution to the sample moment equation

$$\frac{1}{n}\sum_{i=1}^{n} \psi(z_i, \theta) = 0$$

where

$$\psi(z, \theta) = \begin{pmatrix} \psi_1(z, \theta) \\ \psi_2(z, \theta) \end{pmatrix}.$$

By standard results on M-estimators, under θ

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Delta \Phi \Delta'),$$

where

$$\Phi = E[\psi(z, \theta)\psi(z, \theta)'],$$

$$\Delta = D^{-1},$$

$$D = E\left[\frac{\partial \psi}{\partial \theta'}\right].$$

To estimate the asymptotic variance use:

$$\hat{\Phi} = \frac{1}{n}\sum_i \psi(z_i, \hat{\theta})\psi(z_i, \hat{\theta})'.$$

$$\hat{D} = \frac{1}{n}\sum_i \frac{\partial \psi(z_i, \hat{\theta})}{\partial \theta'},$$

where the derivative of ψ can be calculated as

$$\frac{\partial \psi(z, \theta)}{\partial \theta'} = \begin{pmatrix} \frac{\partial \psi_1(z, \theta)}{\partial \gamma'} & \frac{\partial \psi_1(z, \theta)}{\partial \alpha'} \\ \frac{\partial \psi_2(z, \theta)}{\partial \gamma'} & \frac{\partial \psi_2(z, \theta)}{\partial \alpha'} \end{pmatrix}.$$

where

$$\frac{\partial \psi_1(z, \theta)}{\partial \gamma'} = \frac{\exp(\gamma' v)}{(1 + \exp(\gamma' v))^2} vv',$$

$$\frac{\partial \psi_1(z, \theta)}{\partial \alpha'} = 0,$$

$$\frac{\partial \psi_2(z, \theta)}{\partial \gamma'} = \left[\frac{-t}{\exp(\gamma' v)} + (1 - t)\exp(\gamma' v)\right](y - w'\delta)wv',$$

and

$$\frac{\partial \psi_2(z, \theta)}{\partial \alpha'} = -[1 + \exp(\gamma' v)]\left(\frac{t}{\exp(\gamma' v)} + (1 - t)\right)ww'.$$

For estimating the effect of the treatment on the treated we need to redefine w_i appropriately, and modify ψ_2 to be

$$\begin{aligned} \psi_2(z, \theta) &= \omega_1(t, x)w(y - w'\alpha) \\ &= \left[t + \frac{(1 - t)e}{1 - e}\right]w(y - w'\alpha) \\ &= [t + (1 - t)\exp(\gamma' v)]w(y - w'\alpha). \end{aligned}$$

Then we have

$$\frac{\partial\psi(z,\theta)}{\partial\theta'} = \begin{pmatrix} (\exp(\gamma'v)/(1+\exp(\gamma'v))^2)vv' & 0 \\ (1-t)\exp(\gamma'v)(y-w'\alpha)w\cdot v' & -[t+(1-t)\exp(\gamma'v)]ww' \end{pmatrix}.$$

Acknowledgments

We are grateful to the SUPPORT study for making their data available, to Enrico Moretti for help working with the data, and to Donald Rubin, Rajeev Dehejia, and three anonymous referees for comments. Sejin Min, Marcos Rangel and Yue Xu provided excellent research assistance. Financial support for this research was generously provided through NSF grants SES-9985257 (Hirano) and SBR-9818644 and SES-0136789 (Imbens).

References

A. Abadie and G. Imbens, "Simple and bias–corrected matching estimators for average treatment effects," unpublished manuscript, Department of Economics, UC Berkeley, 2001.

A. F. Connors, et al., "The effectiveness of right heart catheterization in the initial care of critically ill patients," *Jornal of the American Medical Association*, 276, pp. 889–897, 2001.

J. Hahn, "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66, pp. 315–331, 1998.

J. Heckman, H. Ichimura and P. Todd, "Matching as an econometric evaluation estimator: evidence from evaluating a job training program," *Review of Economic Studies*, 64, pp. 605–654, 1997.

J. Heckman, H. Ichimura and P. Todd, "Matching as an econometric evaluation estimator," *Review of Economic Studies*, 65, pp. 261–294, 1998.

K. Hirano, G. Imbens and G. Ridder, "Efficient estimation of average treatment effects using the estimated propensity score," NBER Technical Working Paper 251, 2000.

D. Horvitz and D. Thompson, "A generalization of sampling without replacement from a finite population," *Journal of the American Statistical Association*, 47, pp. 663–685, 1952.

A. Miller, *Subset Selection in Regression*, Chapman and Hall, London, 1990.

D. J. Murphy and L. E. Cluff, "SUPPORT: Study to understand prognoses and preferences for outcomes and risks of treatments—study design," *Journal of Clinical Epidemiology*, 43, pp. 1S–123S, 1990.

W. Newey, "The asymptotic variance of semiparametric estimators," *Econometrica*, 62, pp. 1349–1382, 1994.

J. Robins, "Marginal structural models versus structural nested models as tools for causal inference," to appear: AAAI Technical Report Series, Spring 1998 Symposium on Prospects for a Common Sense Theory of Causation, Stanford, CA, 1998.

J. Robins and Y. Ritov, "Towards a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models," *Statistics in Medicine*, 16, pp. 285–319, 1997.

J. Robins and A. Rotnitzky, "Semiparametric efficiency in multivariate regression models with missing data," *Journal of the American Statistical Association*, 90(429), pp. 122–129, 1995.

J. Robins, A. Rotnitzky and L. Zhao, "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *Journal of the American Statistical Association*, 90(429), pp. 106–121, 1995.

P. Rosenbaum, *Observational Studies*, Springer Verlag, 1985.

P. Rosenbaum, "Model-based direct adjustment," *Journal of the American Statistical Association*, 82, pp. 387–394, 1987.

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

P. Rosenbaum and D. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, 79, pp. 516–524, 1985.

A. Rotnitzky and J. Robins, "Semiparametric regression estimation in the presence of dependent censoring," *Biometrika*, 82(4), pp. 805–820, 1995.

D. Rubin and N. Thomas, "Affinely invariant matching methods with ellipsoidal distributions," *Annals of Statistics* 20(2), pp. 1079–1093, 1992.

D. Rubin and N. Thomas, "Characterizing the effect of matching using linear propensity score methods with normal distributions," *Biometrika*, 79, pp. 797–809, 1992.

D. Rubin and N. Thomas, "Matching using estimated propensity scores: Relating theory to practice," *Biometrics*, 52, pp. 249–264, 1996.