TECHNICAL WORKING PAPER SERIES

EFFICIENT ESTIMATION OF AVERAGE TREATMENT EFFECTS USING THE ESTIMATED PROPENSITY SCORE

Keisuke Hirano
Guido W. Imbens
Geert Ridder

Technical Working Paper 251
http://www.nber.org/papers/T0251

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2000

We thank Gary Chamberlain, Jinyong Hahn, Donald Rubin, and seminar participants at the University of Chicago, the University of California at Davis, and Harvard-MIT for comments. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

Efficient Estimation of Average Treatment Effects Using the
Estimated Propensity Score
Keisuke Hirano, Guido W. Imbens, Geert Ridder
NBER Technical Working Paper No. 251
March 2000

ABSTRACT

We are interested in estimating the average effect of a binary treatment on a scalar outcome. If assignment to the treatment is independent of the potential outcomes given pretreatment variables, biases associated with simple treatment-control average comparisons can be removed by adjusting for differences in the pre-treatment variables. Rosenbaum and Rubin (1983, 1984) show that adjusting solely for differences between treated and control units in a scalar function of the pre-treatment, the propensity score, also removes the entire bias associated with differences in pre-treatment variables. Thus it is possible to obtain unbiased estimates of the treatment effect without conditioning on a possibly high-dimensional vector of pre-treatment variables. Although adjusting for the propensity score removes all the bias, this can come at the expense of efficiency. We show that weighting with the inverse of a nonparametric estimate of the propensity score, rather than the true propensity score, leads to efficient estimates of the various average treatment effects. This result holds whether the pre-treatment variables have discrete or continuous distributions. We provide intuition for this result in a number of ways. First we show that with discrete covariates, exact adjustment for the estimated propensity score is identical to adjustment for the pre-treatment variables. Second, we show that weighting by the inverse of the estimated propensity score can be interpreted as an empirical likelihood estimator that efficiently incorporates the information about the propensity score. Finally, we make a connection to results to other results on efficient estimation through weighting in the context of variable probability sampling.

Keisuke Hirano
Department of Economics
UCLA
Los Angeles, CA 90095
khirano@econ.ucla.edu

Guido W. Imbens
Department of Economics
UCLA
Los Angeles, CA 90095
and NBER
imbens@econ.ucla.edu

Geert Ridder
Department of Economics
Johns Hopkins University
Baltimore, MD 21218
gridder@jhu.edu

1. Introduction

Estimating the average effect of a binary treatment on a scalar outcome is a basic goal of many empirical studies in economics. If assignment to the treatment is *unconfounded*, that is, independent of the potential outcomes conditional on pre-treatment variables, the average treatment effect for the subpopulation with a given value of the pre-treatment variables can be estimated by simply taking the difference between the treatment and control averages in that subpopulation. The population average treatment effect can then be estimated by weighting these subpopulation estimates by the distribution of the pre-treatment variables. If there are many pre-treatment variables, this strategy may not be desirable or even feasible. An appealing alternative approach is based on the *propensity score*, the conditional probability of receiving treatment given pre-treatment variables. Rosenbaum and Rubin (1983, 1984) show that adjusting solely for differences in the propensity score between treated and control units removes all bias associated with differences in the pre-treatment variables. Recent applications of these methods in economics include Heckman, Ichimura and Todd (1997), Dehejia and Wahba (1999), Hotz, Imbens and Mortimer (1999), and Lechner (1999).

Although adjusting for the propensity score removes all bias, it may do so at the expense of efficiency. Hahn (1998) and Heckman, Ichimura and Todd (1998) show that adjusting only for the known propensity score can result in efficiency losses compared to adjusting for all pre-treatment variables. However, Rosenbaum (1987) and Rubin and Thomas (1997) demonstrate that using parametric estimates of the propensity score, rather than the true propensity score, can avoid some of these efficiency losses. Rotnitzky and Robins (1995) make a similar point in the context of regression models in the presence of missing data where the missing data are *missing at random* (Rubin, 1976; Little and Rubin, 1987). They show that weighting by the inverse of a parametric estimate of the selection probability is more efficient than weighting by the inverse of the true selection probability.

In this paper we propose estimators based on the estimated propensity score that are fully efficient for estimation of population average treatment effects. Our estimators weight observations by the inverse of nonparametric estimates of the propensity score, rather than the true propensity score. We use results from Newey (1994) to calculate the variances of these semiparametric estimators, and show that they achieve the semiparametric efficiency bounds obtained in Hahn (1998). We provide intuition for this result in a number of different ways. First, we show that with discrete covariates, the estimator based on weighting by the inverse of the estimated propensity score is identical to an efficient estimator that directly controls for all pre-treatment variables (e.g., Hahn, 1998). Second, we show in the case where the propensity score is known, the proposed estimator can be interpreted as an empirical likelihood estimator (e.g., Imbens, Spady and Johnson, 1998) that efficiently incorporates the information about the propensity score. Finally, we make a connection to results involving efficient estimation with estimated rather than population weights in the context of stratified sampling (e.g., Lancaster, 1990; Wooldridge, 1999).

In the next section we lay out the problem and discuss earlier work. In Section 3 we provide some intuition for our efficiency results by examining a simplified version of the problem. In Section 4 we give the formal conditions under which weighting by the estimated propensity score results in an efficient estimator, in four separate cases. The first case is the missing data case studied by Robins and Rotnitzky (1995) and Rotnitzky and Robins (1995), with the missing data assumed to be missing at random. In the second case we focus on efficient estimation of the population average treatment effect, one of the cases studied by Hahn (1998). In the third case we focus on a weighted average treatment effect with a known weight function. Finally we look at the case where the weight function is proportional to the propensity score, and thus the parameter of interest is the average treatment effect for the treated (Rubin, 1977; Heckman and Robb, 1985). Recent work on estimation of this parameter includes Heckman, Ichimura and Todd (1997, 1998) and Hahn (1998).

2. The Basic Setup and Previous Results

2.1 The Model

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 receives the treatment of interest, and $T_i = 0$ if unit i receives the control treatment. Using the potential outcome notation, let $Y_i(0)$ denote the outcome for unit i under control and $Y_i(1)$ the outcome under treatment.[1] We observe T_i and Y_i, where

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

In addition, we observe a vector of pre-treatment variables, or covariates, denoted by X_i. We shall focus on the population average treatment effect:

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

We shall also discuss estimation of weighted average treatment effects

$$\tau_g = \frac{\int E[Y(1) - Y(0)|X = x] g(x) dF(x)}{\int g(x) dF(x)},$$

and the average effect for the treated:

$$\tau_{\text{treated}} = E[Y(1) - Y(0)|T = 1].$$

The central problem of evaluation research is that for unit i we observe $Y_i(0)$ or $Y_i(1)$, but never both. Without further restrictions, the treatment effects are not consistently estimable. To solve the identification problem, we maintain throughout the paper 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. Formally:

[1] Implicit in this notation is the stability assumption or SUTVA (Rubin, 1978) that units are not affected by receipt of treatment by others, and that there is only one version of the treatment.

Assumption (Unconfoundedness)

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

Define the average effect conditional on pre-treatment variables:

$$\tau(x) \equiv E[Y(1) - Y(0)|X = x]$$

Note that $\tau(x)$ is estimable under the unconfoundedness assumption, because

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

The population average treatment effect can be obtained by averaging the $\tau(x)$ over the distribution of X:

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

In practice, the strategy of forming cells and comparing units with exactly the same value of X may fail if X takes on too many distinct values. To avoid having to match units by the values of all pre-treatment variables, Rosenbaum and Rubin (1983, 1984) developed an approach based on the *propensity score*, the probability of selection into the treatment group:

$$e(x) = Pr(T = 1|X = x). \tag{2}$$

We assume that this probability is bounded away from zero and one. Their key insight was that if treatment and potential outcomes are independent conditional on all pre-treatment variables, they are also independent conditional on the conditional probability of receiving treatment given pre-treatment variables, that is, conditional on the propensity score. Formally, unconfoundedness implies

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

(See Rosenbaum and Rubin (1983) for the proof of this result and further discussion.) Thus, to obtain unbiased estimates of the average treatment effect, it is only necessary to match on a scalar variable. An alternative approach, analogous to the Horvitz-Thompson (1952) estimator, is to reweight the observations by the inverse of their selection probabilities and take the weighted average as an estimate of the treatment effect. See also Rosenbaum (1987). This weighting estimator can be written as

$$\hat{\tau}_{ht} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{t_i \cdot y_i}{e(x_i)} - \frac{(1-t_i) \cdot y_i}{1 - e(x_i)} \right].$$

Although adjusting for differences in the propensity score, either through weighting or through regression adjustment, removes all bias associated with differences in the pre-treatment variables, it may do so at a price. Compared to estimators that adjust for differences in all pre-treatment variables there may be a loss of efficiency, as pointed out by Hahn (1998) and Heckman, Ichimura and Todd (1998).[2]

2.2 Previous Results

The model set out above, and related models, have been examined by many researchers. Hahn (1998) studies the same model as the current paper, calculates the efficiency bound, and proposes an efficient estimator. His estimator imputes the missing potential outcomes given covariates, and requires nonparametric estimation of the two conditional expectations $\mu_t(x) = E[Y|T = t, X = x]$ for $t = 0, 1$. Hahn also shows that the estimator for the population average treatment effect based conditioning on the true propensity score does not in general reach the efficiency bound, and that in fact knowledge of the propensity score does not affect the semiparametric efficiency bound. In addition Hahn considers inference for the average treatment effect on the treated and concludes that for that estimand knowledge of the propensity score is indeed informative. He also derives efficient estimators for that

[2] A separate issue is whether standard asymptotic theory provides adequate approximations to the sampling distributions of estimators based on initial nonparametric estimates of conditional means, when the dimension of the conditioning variable is high. See for example Robins and Ritov (1995) and Angrist and Hahn (1999). We do not address this issue here.

case.

Heckman, Ichimura and Todd (1998) focus on the average treatment effect for the subpopulation of the treated. They consider estimators based on nonparametric kernel regressions of the outcome on treatment status and either covariates or the propensity score. They conclude that in general there is no clear ranking of their estimators; under some conditions the estimator based on adjustment for all covariates is superior to the estimator based on adjustment for the propensity score, and under other conditions the second estimator is to be preferred. Lack of knowledge of the propensity score does not alter this conclusion.

Rosenbaum (1987) and Rubin and Thomas (1997) investigate the differences between using the estimated and the true propensity score when the propensity score belongs to a parametric family. They conclude that there can be efficiency gains from using the estimated propensity score. Rosenbaum (1987) interprets this as a bias conditional on an ancillary statistic that is removed by using the estimated propensity score. Lancaster (1990) makes a similar point in the context of choice-based sampling.

Robins and Rotnitzky (1995) and Robins, Rotnitzky and Zhao (1995) and Rotnitzky and Robins (1995) study inference for parameters in a regression model with missing data, using essentially the *missing at random* (MAR, Rubin, 1976; Little and Rubin, 1987) assumption. They calculate the efficiency bound and note that conditioning on the true selection probability, or weighting by its inverse, does not lead to an efficient estimator. Rotnitzky and Robins (1995) show that when the selection probability model has a parametric form, weighting by the inverse of the estimated selection probability is more efficient than weighting by the inverse of the true selection probability, and suggest it may be possible to achieve efficiency by allowing the dimension of the model for the selection probability to grow with the sample size. For this case Robins and Rotnitzky (1995) propose an estimator that achieves the efficiency bound. Their estimator, like Hahn's estimator in a different context, requires an initial estimate of the conditional expectation of the outcome given pre-treatment variables,

$\mu_1(x) = E[Y|T = 1, X = x]$. Unlike Hahn, who uses this conditional expectation to impute the missing outcomes, Robins and Rotnitzky use it to formulate a parametric model for the selection probabilities and estimate the parameters by a weighted regression with the weights equal to the inverse of the estimated selection probabilities.

3. A Simple Example with Binary Covariates

To develop some intuition for the formal results that will be presented in Section 4, we consider the simpler problem of estimating the population average of a variable Y, $\beta_0 = E[Y]$, given a random sample of size N of the triple $(T_i, X_i, T_i \cdot Y_i)$. In other words, T_i and X_i are observed for all units in the sample, but Y_i is only observed if $T_i = 1$. We provide a heuristic argument for efficiency of estimated weights, deferring a formal result to Section 4.

The analog to the unconfoundedness assumption here is the assumption that the Y_i are missing at random, or

$$T_i \;\perp\; Y_i \;\Big|\; X_i.$$

The role of the propensity score is played here by the selection probability:

$$p(x) = E[T|X = x] = Pr(T = 1|X). \tag{4}$$

First, we restrict our attention in this section to the case with a single binary covariate. Let N_{tx} denote the number of observations with $t_i = t$ and $x_i = x$, for $t, x \in \{0, 1\}$, and let $N_{\cdot x} = N_{0x} + N_{1x}$ be the number of observations with $x_i = x$. Furthermore, suppose the true selection probability is constant, equal to $p_0(x) = 1/2$ for all $x \in \{0, 1\}$.[3] The normalized variance bound for β_0 is

$$V_{bound} = 2 \cdot E\left[V(Y|X)\right] + V\left(E[Y|X]\right),$$

which can be calculated from results in Robins and Rotnitzky (1995) or Hahn (1998).

[3] Thus the missing data are *missing completely at random* (MCAR, Rubin, 1976; Little and Rubin, 1987).

We shall consider four estimators. First, consider estimating the population average by the sample average for complete observations:

$$\hat{\beta}_{comp} = \sum_{i=1}^{N} y_i \cdot t_i \Big/ \sum_{j=1}^{N} t_j. \tag{5}$$

Simple calculations show that under the MCAR assumptions, and with 50% of the observations missing on average, this estimator has normalized variance

$$V_{comp} = 2 \cdot E[V(Y|X)] + 2 \cdot V(E[Y|X]) = 2V(Y),$$

strictly larger than the variance bound V_{bound}.

The second, "true weights" estimator weights the observed outcomes by the inverse of the true selection probability:

$$\hat{\beta}_{tw} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i \cdot t_i}{p(x_i)} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i \cdot t_i}{1/2}. \tag{6}$$

Its large sample normalized variance is

$$V_{tw} = 2 \cdot E[V(Y|X)] + V(E[Y|X]) + E\left[E[Y|X]^2\right] = 2V(Y) + E(Y^2),$$

even larger than the variance for $\hat{\beta}_{comp}$.

The third estimator weights the observed outcomes by the inverse of a nonparametric estimate of the selection probability. This estimator is the main focus of the paper, and it will be discussed in Section 4 in more general settings. In the current context, the estimated selection probability is simply the proportion of observed outcomes for a given value of the covariate. For units with $x_i = 0$, the proportion of observed outcomes is $N_{10}/N_{\cdot 0}$, and for units with $x_i = 1$, the proportion of observed outcomes is $N_{11}/N_{\cdot 1}$. Thus the estimated selection probability is

$$\hat{p}(x) = \begin{cases} N_{10}/N_{\cdot 0} & \text{if } x = 0, \\ N_{11}/N_{\cdot 1} & \text{if } x = 1. \end{cases}$$

Then the proposed "estimated weights" estimator is:

$$\hat{\beta}_{ew} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i \cdot t_i}{\hat{p}(x_i)}. \tag{7}$$

The normalized variance of this estimator is equal to the variance bound:

$$V_{ew} = 2 \cdot E[V(Y|X)] + V(E[Y|X]).$$

So in this simple case, not only does the weighting estimator with nonparametrically estimated weights have a lower variance than the estimator using the "true" weights, but it is fully efficient in the sense of achieving the variance bound. In the remainder of this section we shall provide some intuition for this result that suggests why this efficiency property may carry over to case with the continuous and vector-valued covariates, as well as with general dependence of the selection probability or propensity score on the covariates.

To help understand why the estimated weights approach is efficient, it is useful to consider a fourth estimator. Let

$$\hat{\mu}(x) = \sum_{i|x_i=x} y_i \cdot t_i \Big/ \sum_{i|x_i=x} t_i,$$

for $x \in \{0,1\}$ be the non-parametric estimator for the conditional regression function

$$\mu(x) = E[Y|T=1, X=x].$$

Now consider the following, "regression-on-covariates" estimator:

$$\hat{\beta}_{rc} = \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}(x_i).$$

Substituting for the estimator $\hat{\mu}(x_i)$ it can be shown that this estimator is numerically identical to an estimator based on averaging the observed and imputed outcomes:

$$\hat{\beta}_{rc} = \frac{1}{N} \sum_{i=1}^{N} t_i \cdot y_i + (1 - t_i) \cdot \hat{\mu}(x_i).$$

This estimator averages over all the observations, using observed values when they are available, and imputing estimated values when they are not. Hahn (1998) proposed using imputation estimators similar to this to estimate treatment effects, and showed that they are efficient in the cases he studies.[4]

In the current setting, with a single binary covariate, we can rewrite the regression-on-covariates estimator as

$$\begin{aligned}\hat{\beta}_{rc} &= \frac{1}{N}\Big[(N_{00}+N_{10})\cdot\hat{\mu}(0)+(N_{01}+N_{11})\cdot\hat{\mu}(1)\Big] \\ &= \frac{1}{N}\left[\ (N_{00}+N_{10})\cdot\frac{\sum_{i:x_i=0} t_i\cdot y_i}{N_{10}}+(N_{01}+N_{11})\cdot\frac{\sum_{i:x_i=1} t_i\cdot y_i}{N_{11}}\ \right] \\ &= \frac{1}{N}\sum_i \frac{t_i\cdot y_i}{w_i}\end{aligned}$$

where

$$w_i = \begin{cases} N_{10}/N_{\cdot 0} & \text{if } x_i = 0 \\ N_{11}/N_{\cdot 1} & \text{if } x_i = 1, \end{cases}$$

[4] An alternative to the Hahn estimator and the estimator proposed in the current paper is an estimator proposed by Robins and Rotnitzky (1995). First one estimates the conditional expectation of the outcome given pre-treatment variables, $\hat{\mu}(x)$. In the second step a logistic regression model involving a single unknown parameter δ is estimated by maximum likelihood:

$$Pr(T=1|X=x) = \frac{\exp(\delta\cdot(\hat{\mu}(x)-\hat{\beta}))}{1+\exp(\delta\cdot(\hat{\mu}(x)-\hat{\beta}))},$$

for a preliminary estimate of the parameter of interest $\hat{\beta}$. (At $\delta = 0$ this model reduces to the true selection probability, equal to 1/2 in this case.) The inverse of the weight is then constructed as

$$w_i = \frac{\exp(\hat{\delta}\cdot(\hat{\mu}(x_i)-\hat{\beta}))}{1+\exp(\hat{\delta}\cdot(\hat{\mu}(x_i)-\hat{\beta}))},$$

and finally, the population mean is estimated by

$$\hat{\beta}_{rr} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i\cdot t_i}{w_i}.$$

Although numerically different from the estimator with nonparametric weights in this single binary regressor case, the Robins-Rotnitzky estimator also reaches the variance bound.

These weights are identical to $\hat{p}(x_i)$, and thus the two estimators $\hat{\beta}_{rc}$ and $\hat{\beta}_{ew}$ are identical. This numerical equivalence between the nonparametric regression estimator and the estimator with the nonparametric weights can be shown to hold for any sample with discrete pre-treatment variables. This implies that the nonparametric weights estimator is fully efficient in the discrete pre-treatment variable case, and since the formulation of the bound does not rely on discreteness, one might expect using the same reasoning as Chamberlain (1987) in the context of GMM estimation, that the estimator is also efficient when the pre-treatment variables are continuous.

A second interpretation of the estimator that is directly suggestive of its efficiency properties is based on a generalized method of moments (GMM) representation (Hansen, 1982). Under the assumption that the selection probability is $p(x) = 1/2$, we can estimate β_0 using the single moment restriction $E[\psi_1(Y, X, T, \beta_0)] = 0$, with

$$\psi_1(y, t, x, \beta) = y \cdot t/p(x) - \beta = \frac{y \cdot t}{1/2} - \beta.$$

The GMM estimator based on this single moment restriction, given knowledge of the selection probability, is the "true-weights" estimator $\hat{\beta}_{tw}$. However, thise estimator is not necessarily efficient, because it ignores the additional information that is available in the form of knowledge of the selection probability:

$$E[T|X = x] = p(x) = 1/2.$$

We can write this additional information in moment restriction form as

$$E\left[T - 1/2|X\right] = 0.$$

With a binary covariate this conditional moment restriction corresponds to two marginal moment restrictions, $E[\psi_2(Y, T, X, \beta_0)] = 0$, with:

$$\psi_2(y, t, x, \beta) = \begin{pmatrix} x \cdot (t - 1/2) \\ (1 - x) \cdot (t - 1/2) \end{pmatrix}.$$

Estimating β_0 in a generalized method of moments framework using the moments $\psi_1(\cdot)$ and $\psi_2(\cdot)$ leads to a fully efficient estimator.[5] We can implement the GMM estimator in different ways. The standard GMM approach of Hansen (1982) estimates an optimal weight matrix and then minimizes a quadratic form involving the average moments. Here it is of particular interest to consider an alternative, the empirical likelihood estimator (e.g., Qin and Lawless, 1994; Imbens, 1997; Imbens, Spady and Johnson, 1998). The empirical likelihood estimator is based on maximization, both over a nuisance parameter $\pi = (\pi_1, \ldots, \pi_N)$ and over the parameter of interest β, of the logarithm of the empirical likelihood function

$$L(\pi) = \sum_{i=1}^{N} \ln \pi_i, \tag{8}$$

subject to three sets of restrictions:
(i), the adding-up restriction $\sum_{i=1}^{N} \pi_i = 1$;
(ii), the restriction for the identifying moment $\psi_1(\cdot)$,

$$\sum_{i=1}^{N} \pi_i \cdot \left(\frac{y_i \cdot t_i}{1/2} - \beta \right) = 0; \tag{9}$$

and

(iii), the two restrictions from knowledge of the selection probability, the additional moments $\psi_2(\cdot)$:

$$\sum_{i=1}^{N} \pi_i \cdot x_i \cdot (t_i - 1/2) = 0, \tag{10}$$

and

$$\sum_{i=1}^{N} \pi_i \cdot (1 - x_i) \cdot (t_i - 1/2) = 0.$$

[5] Although $\psi_2(\cdot)$ does not depend on the parameter of interest, $\psi_2(\cdot)$ is generally correlated with $\psi_1(\cdot)$. Thus there can be efficiency gains from using both sets of moment conditions. See, e.g., Hellerstein and Imbens (1999), and Qian and Schmidt (1999).

¿From the restriction (11), we can concentrate out $\hat{\beta}_{el}$ by noting that a solution $(\hat{\pi}, \hat{\beta}_{el})$ will satisfy

$$\hat{\beta}_{el} = \sum_{i=1}^{N} \hat{\pi}_i \cdot y_i \cdot (2t_i) \Big/ \sum_{i=1}^{N} \hat{\pi}_i = \sum_{i=1}^{N} 2\hat{\pi}_i \cdot y_i \cdot t_i.$$

Solving for $\hat{\pi}_i$ by maximizing (8) subject to the adding-up restriction and (12), we find

$$\hat{\pi}_i = \left(1 + \frac{N_{11}/N_{\cdot 1} - 1/2}{1/4} \cdot x_i \cdot (t_i - 1/2) + \frac{N_{10}/N_{\cdot 0} - 1/2}{1/4} \cdot (1 - x_i) \cdot (t_i - 1/2)\right)^{-1}.$$

Substituting this into the solution for β gives

$$\hat{\beta}_{el} = \sum_{i=1}^{N} 2\hat{\pi}_i \cdot y_i \cdot t_i = \hat{\beta}_{ew}.$$

This interpretation suggests moving from the "true-weights" estimator to the "estimated-weights" estimator increases efficiency in the same way that adding moment restrictions in a generalized method of moments framework improves efficiency. A similar finding appears in Crepon, Kramarz, and Trognon (1998). They consider GMM estimation where some of the parameters can be designated as nuisance parameters. They show that GMM estimation using a reduced set of moment conditions, in which nuisance parameters are replaced by solutions to the sample analogs of the remaining moment conditions, is asymptotically equivalent to using the full set of moment conditions. Their results also imply that using the true values of the nuisance parameters may lead to efficiency losses in some contexts.

A third interpretation of the efficiency gain from weighting by the inverse of the estimated rather than the true propensity score builds on connections to the literature on weighting in stratified sampling. Translated to our setting, the results by Lancaster (1990) suggest studying the distribution of the various estimators conditional on the ancillary statistics $\sum t_i$, $\sum x_i$ and $\sum t_i \cdot x_i$. Conditional on those three statistics the true-weights estimator is biased, while the estimated-weights estimator remains unbiased. Rosenbaum (1987) discusses this issue specifically in the context of estimated versus true propensity scores. Wooldridge (1999) finds similar results in general variable probability and stratified sampling settings, in which observations are first drawn randomly from the population, and then retained or discarded with some probability that depends on its stratum. Wooldridge shows that weighted versions of standard M-estimators, where the weighting is by the inverse of the sampling probabilities, will lead to appropriate estimators. In addition, he shows that efficiency gains are possible by using estimated rather than population versions of the weights.

4. Efficient Estimation Using Estimated Weights

In this section we present the main results of the paper. We discuss four separate cases. First, we extend the example of the previous section to allow for continuous covariates and a missing data mechanism that depends on the covariates. Second, we consider the problem of estimating the population average treatment effect under the unconfoundedness assumption. Third, we consider estimation of weighted average treatment effects, a generalization of the population average treatment effect case. Finally, we consider estimation of the effect of the treatment on the treated, which in our setup will follow directly from the general weighted average treatment effect problem. This will shed additional light on Hahn's (1998) result that for this parameter knowledge of the propensity score affects the efficiency bound.

4.1 Estimating Population Averages with Outcomes Missing at Random

The first case we consider is a general version of the example in Section 3. We are interested in estimating a population mean, when the variable of interest is missing for some units and the missing data mechanism satisfies the MAR assumption. For each unit, in a random sample of size N from the population of interest, there is a triple (Y, T, X), with T binary. We observe $(T, X, T \cdot Y)$. The first assumption is

Assumption 1 (Missing At Random)

$$T \perp Y \Bigm| X.$$

Let $p_0(x)$ be the selection probability, that is, the probability of observing Y given $X = x$:

$$p_0(x) = E[T|X = x] = Pr(T = 1|X = x).$$

We use the framework of Newey (1994) for deriving the variance of the semiparametric estimator for β_0 based on an initial nonparametric estimator for $p_0(x)$. We can characterize β_0 through the moment equation:

$$E\left[\psi(Y, T, X, \beta_0, p_0(X))\right] = 0,$$

where

$$\psi(y, t, x, \beta, p(x)) = \frac{y \cdot t}{p(x)} - \beta.$$

We are interested in estimators for β_0 based on nonparametric estimators for the selection probability $p_0(\cdot)$. We estimate $p_0(\cdot)$ with a series estimator. For $K = 1, 2, \ldots,$ let

$$r^K(x) = (r_{1K}(x), r_{2K}(x) \ \ldots, r_{KK}(x))'$$

be a $K-$vector of functions. Let

$$r^K = \Big((r^K(x_1), \ldots, r^K(x_N)\Big)',$$

denote the matrix obtained by evaluating $r^K(\cdot)$ at the observed values of X, and let

$$\underline{t} = (t_1, \ldots, t_n)',$$

be the vector of observed values of T. Then

$$\hat{\pi} = (r^{K\prime} r^K)^{-} r^{K\prime} \underline{t},$$

where A^- is a generalized inverse of A, is the vector of least squares estimates in a regression of $\underline{t}$ on r^K, and

$$\hat{p}(x) = r^K(x)\hat{\pi}.$$

More specifically we consider power series. Let $\lambda = (\lambda_1, \dots, \lambda_r)'$ be an r-dimensional vector of nonnegative integers (multi-indices), with norm $|\lambda| = \sum_{j=1}^r \lambda_j$. Let $x^\lambda = \prod_{j=1}^r x_j^{\lambda_j}$. Let $(\lambda(k))_{k=1}^\infty$ be a sequence that includes all distinct multi-indices and satisfies $|\lambda(k)| \leq |\lambda(k+1)|$. For such a sequence $\lambda(k)$ we consider the series

$$r_{kK}(x) = x^{\lambda(k)}.$$

Given the estimate $\hat{p}(x)$ for the selection probability $p_0(x)$, we estimate the population mean $\beta_0 = E(Y)$ by setting the average moment evaluated at the estimated selection probability equal to zero as a function of β:

$$\sum_{i=1}^N \psi(y_i, t_i, x_i, \beta, \hat{p}(x_i)) = 0.$$

Given the form of the moment condition, this leads to the estimator

$$\hat{\beta} = \frac{1}{N} \sum_{i=1}^N \frac{y_i \cdot t_i}{\hat{p}(x_i)}.$$

In addition to the missing at random assumption, Assumption 1, the following assumptions are used to derive the properties of the estimator. First, we restrict the distribution of X and Y:

Assumption 2 (Distribution of X)
(i) the support $\mathcal{X}$ of the r-dimensional covariate X is a Cartesian product of compact intervals, $\mathcal{X} = \prod_{j=1}^r [x_{lj}, x_{uj}]$,

(ii), the density of X is bounded from 0 on $\mathcal{X}$.

Assumption 3 (Distribution of Y)
(i) $E(Y^2) < \infty$,
(ii), $\mu(x) = E(Y|X = x)$ is continuously differentiable for all $x \in \mathcal{X}$.

The next assumption requires sufficient smoothness of the selection probability.

Assumption 4 (Selection Probability)
The true selection probability $p_0(x)$ satisfies the following conditions: For all $x \in \mathcal{X}$
(i) $p_0(x)$ is continuously differentiable of order $s > 3 \cdot r$ with r the dimension of X ,
(ii), $p_0(x) \geq \underline{p} > 0$.

Finally, we restrict the rate at which additional terms are added to the series approximation to $p_0(x)$, depending on the dimension of X and the number of derivatives of $p_0(x)$.

Assumption 5 (Series Estimator)
The series estimator of $p_0(x)$ is a power series estimator with $K = N^\nu$ for some $1/(2 \cdot \alpha) < \nu < 1/6$ with $\alpha = s/r$.

Under these conditions we can state the first result.

Theorem 1 *Suppose Assumptions 1-5 hold. Then:*
(i)

$$\hat{\beta} \xrightarrow{p} \beta_0,$$

(ii),

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, V(E[Y|X]) + E[V(Y|X)/p_0(X)]),$$

and (iii), $\hat{\beta}$ reaches the semiparametric efficiency bound.

Proof: see Appendix.

Remark: In Section 3, this result was shown for binary X. Theorem 1 establishes the result for continuous X. If X has both continuous and discrete components, this can be easily dealt with, at additional notational expense.

4.2 Estimating Average Treatment Effects

In this section we focus on efficient estimation of the average treatment effect. We postulate for each unit the existence of a pair of potential outcomes $(Y(0), Y(1))$ and are interested in the average treatment effect, $\tau_0 = E[Y(1) - Y(0)]$.

We modify Assumption 1 to require conditional independence of the pair of potential outcomes and treatment assignment:

Assumption 1' (Unconfounded Treatment Assignment)

$$T \perp (Y(0), Y(1)) \Big| X.$$

Assumption 3 is modified to reflect the presence of two potential outcomes:

Assumption 3' (Distribution of $Y(0), Y(1)$)
(i) $E(Y(0)^2) < \infty$ *and* $E(Y(1)^2) < \infty$*,*
(ii), $E(Y(0)|X = x)$ *and* $E(Y(1)|X = x)$ *are continuously differentiable for all* $x \in \mathcal{X}$*.*

Finally, Assumption 4 is modified to require the propensity score to be bounded away from both zero and one:

Assumption 4' (Propensity Score)
The true propensity score $e_0(x) \equiv Pr(T = 1|X = x)$ *satisfies the following conditions: For all* $x \in \mathcal{X}$
(i) $e_0(x)$ *is continuously differentiable of order* $s > 3 \cdot r$ *with* r *the dimension of* X*,*
(ii), $0 < \underline{e} \leq e_0(x) \leq \bar{e} < 1$.

We estimate τ_0 by first estimating the propensity score the same way the selection probability was estimated before through series estimation. Then $\hat{\tau}$ is the solution to

$$\sum_{i=1}^{N} \psi(y_i, t_i, x_i, \tau, \hat{e}(x_i)) = 0,$$

where

$$\psi(y, t, x, \tau, e(x)) = \frac{y \cdot t}{e(x)} - \frac{y \cdot (1-t)}{1-e(x)} - \tau, \tag{11}$$

so that

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i \cdot t_i}{\hat{e}(x_i)} - \frac{y_i \cdot (1-t_i)}{1-\hat{e}(x_i)}.$$

The formal result is:

Theorem 2 *Suppose Assumptions 1', 2, 3', 4', and 5 hold. Then:*
(i)

$$\hat{\tau} \xrightarrow{p} \tau_0,$$

(ii),

$$\sqrt{N}(\hat{\tau} - \tau_0) \xrightarrow{d} \mathcal{N}(0, V),$$

with

$$V = V(E[Y(1) - Y(0)|X]) + E[V(Y(1)|X)/e_0(X)] + E[V(Y(0)|X)/(1 - e_0(X))]),$$

and (iii), $\hat{\tau}$ reaches the semiparametric efficiency bound.

Proof: see Appendix.

4.3 Estimating the Average Treatment Effect for Subpopulations

We generalize the previous result to τ_g, the weighted average treatment effect for a known weighting function $g(x)$. One motivation for considering this estimand is that by choosing $g(x)$ appropriately, one can define treatment effects for subpopulations defined by X. In addition, by choosing $g(x)$ appropriately, one can recover the average effect of the treatment on the treated, as will be discussed below.

A semiparametric efficiency bound for τ_g has not been previously calculated in the literature. The next result shows that our estimator is efficient.

Theorem 3 *The semiparametric efficiency bound for estimation of τ_g is*

$$E\left[\frac{g(X)^2}{(\mu_g)^2 e_0(X)} V(Y(1)|X)\right] + E\left[\frac{g(X)^2}{(\mu_g)^2(1-e_0(X))} V(Y(0)|X)\right]$$
$$+E\left[\frac{g(X)^2}{(\mu_g)^2}\left(E(Y(1)|X) - E(Y(0)|X) - \tau_g\right)^2\right],$$

where

$$\mu_g = E(g(X)).$$

Proof: See Appendix.

To estimate τ_g, we use the following moment function:

$$\psi(y,t,x,\tau_g,e(x)) = g(x)\cdot\left(\frac{y\cdot t}{e(x)} - \frac{y\cdot(1-t)}{1-e(x)} - \tau_g\right). \tag{12}$$

This leads to the estimator

$$\hat{\tau}_g = \sum_i g(x_i)\left[\frac{y_i\cdot t_i}{\hat{e}(x_i)} - \frac{y_i\cdot(1-t_i)}{1-\hat{e}(x_i)}\right] / \sum_i g(x_i).$$

Similar reasoning to the previous results gives the following results, which establishes that this estimator is efficient:

Theorem 4 *Suppose Assumptions 1', 2, 3', 4', and 5 hold, that g(x) is bounded from above and that $\mu_g \equiv E(g(X)) > 0$. Then*
(i),

$$\hat{\tau}_g \xrightarrow{p} \tau_g,$$

(ii),

$$\sqrt{N}(\hat{\tau}_g - \tau_g) \xrightarrow{d} \mathcal{N}(0, V),$$

with

$$\begin{aligned} V &= E\left[\frac{g(X)^2}{\mu_g^2 e_0(X)} V(Y(1)|X)\right] + E\left[\frac{g(X)^2}{\mu_g^2(1-e_0(X))} V(Y(0)|X)\right] \\ &\quad + E\left[\frac{g(X)^2}{\mu_g^2}\left(E(Y(1)|X) - E(Y(0)|X) - \tau_g\right)^2\right]. \end{aligned}$$

Proof: See Appendix.

Remark: We could weaken Assumption 4'(ii), the assumption that the propensity score is bounded away from 0 and 1, by the assumption that $g(x)/e_0(x)$ and $g(x)/(1-e_0(x))$ are bounded from above. Thus, if there is insufficient overlap in the distributions of the treated and untreated units, one may wish to choose $g(\cdot)$ to restrict attention to a subpopulation for which there is sufficiently large probability of observing both treated and untreated units.

4.4 Estimating the Average Treatment Effect for the Treated

The average treatment effect for the treated (Rubin, 1977; Heckman and Robb, 1985) is a special case of the weighted average treatment effect, corresponding to the weighting function $g(x) = e_0(x) = Pr(T = 1|X = x)$. Thus we can use the moment equation:

$$\psi(y, t, x, \tau_{treated}, e(x)) = e_0(x) \cdot \left(\frac{y \cdot t}{e(x)} - \frac{y \cdot (1-t)}{1-e(x)} - \tau_{treated}\right). \tag{13}$$

Notice that we assume that $e_0(x)$ is a known function. However, the inverse weights will still be estimated nonparametrically in our aproach. The next result, which follows from Theorem 4, shows that this estimator achieves the efficiency bound calculated by Hahn (1998) for estimation of the effect of treatment on the treated, assuming that the propensity score is known.

Corollary 5 *Suppose that Assumptions 1', 2, 3', 4', and 5 hold. Then*
(i),

$$\hat{\tau}_{treated} \xrightarrow{p} \tau_{treated},$$

(ii),

$$\sqrt{N}(\hat{\tau}_{treated} - \tau_{treated}) \xrightarrow{d} \mathcal{N}(0, V),$$

with

$$\begin{aligned} V &= E\left[\frac{e_0(X)}{\mu_e^2}V(Y(1)|X)\right] + E\left[\frac{e_0(X)^2}{\mu_e^2(1-e_0(X))}V(Y(0)|X)\right] \\ &\quad + E\left[\frac{e_0(X)^2}{\mu_e^2}\left(E(Y(1)|X) - E(Y(0)|X) - \tau_{treated}\right)^2\right] \end{aligned}$$

where

$$\mu_e = E(e_0(X)).$$

and (iii), $\hat{\tau}_{treated}$ *achieves the semiparametric efficiency bound.*

Proof: See Appendix.

Thus, using the estimated propensity score again leads to a fully efficient estimator.

If the propensity score is not known, then Hahn (1998) shows that this affects the efficiency bound for the effect of treatment on the treated. Our previous estimator cannot be used since it makes use of $e_0(x)$. However, we can modify the moment function to be

$$\psi(y, t, x, \tau_{treated}, e(x)) = e(x) \cdot \left(\frac{y \cdot t}{e(x)} - \frac{y \cdot (1-t)}{1-e(x)} - \tau_{treated}\right).$$

Thus our modifed estimator will use the estimated propensity score in place of $e_0(x)$ in the weighting of observations. Call this estimator $\hat{\tau}_{te}$. The next theorem shows that this estimator is efficient if the propensity is not known.

Theorem 6 *Suppose that Assumptions 1', 2, 3', 4', and 5 hold. Then*
(i),

$$\hat{\tau}_{te} \xrightarrow{p} \tau_{treated},$$

(ii),

$$\sqrt{N}(\hat{\tau}_{te} - \tau_{treated}) \xrightarrow{d} \mathcal{N}(0, V),$$

with

$$\begin{aligned} V &= E\left[\frac{e_0(X)}{\mu_e^2} V(Y(1)|X)\right] + E\left[\frac{e_0(X)^2}{\mu_e^2(1-e_0(X))} V(Y(0)|X)\right] \\ &\quad + E\left[\frac{e_0(X)}{\mu_e^2}\left(E(Y(1)|X) - E(Y(0)|X) - \tau_{treated}\right)^2\right] \end{aligned}$$

where

$$\mu_e = E(e_0(X)).$$

and (iii), $\hat{\tau}_{te}$ achieves the semiparametric efficiency bound for estimation of $\tau_{treated}$ when the propensity score is not known.

Proof: See Appendix.

5. Conclusion

In this paper we have studied efficient estimation of various average treatment effects under an unconfounded treatment assignment assumption. Although weighting observations by the inverse of the true propensity score does not lead to efficient estimators, weighting each observation by the inverse of a nonparametric estimate of the propensity score does lead to efficient estimators. We provide intuition for this result through connections to estimators based on adjustment for covariates, empirical likelihood estimators, and estimators from the literature on variable probability sampling.

Unlike previously proposed estimators the estimators proposed in this paper do not require nonparametric estimation of the regression function of the outcome on the covariates. They do, however, require nonparametric estimation of the propensity probability. The results underline the important role played by the propensity score in estimation of average causal effects.

Appendix

Proof of Theorem 1:

Throughout we use the sup norm for functions $||g(\cdot)|| = \sup_{x\in\mathcal{X}} |g(x)|$ with $|.|$ the usual Euclidean norm. We indicate that $\psi(y,t,x,\beta,p(\cdot))$ is a functional, i.e. a real valued function of the function p, by denoting the argument as $p(\cdot)$ instead of $p(x)$. The sup norm of ψ is $||\psi(y,t,x,\beta,p(\cdot))|| = \sup_{\tilde{x}\in\mathcal{X}} |\psi(y,t,x,\beta,p(\tilde{x}))|$. To simplify the notation, we use C to denote a generic constant in a bound.

The result follows from Theorem 6.1 in Newey (1994). We first check the following conditions, corresponding to Assumptions 5.4-5.6 and 6.1-6.6 in Newey (1994).

Condition 1 *There are $\varepsilon > 0$, and $b(y,t,x), \tilde{b}(y,t,x) > 0$, with $E(b(Y,T,X)), E(\tilde{b}(Y,T,X)) < \infty$, and a compact subset $\mathcal{B}$ of $\Re$ with $\beta_0 \in \mathcal{B}$ such that for all $\beta \in \mathcal{B}$*
(i) $\psi(y,t,x,\beta,p_0(\cdot))$ is continuous in β with probability one,
(ii), $||\psi(y,t,x,\beta,p_0(\cdot))|| < b(y,t,x)$,
(iii), $||\psi(y,t,x,\beta,p(\cdot)) - \psi(y,t,x,\beta,p_0(\cdot))|| < \tilde{b}(y,t,x) \cdot ||p(\cdot) - p_0(\cdot)||^{\varepsilon}$.

Because β_0 is finite by Assumption 3(i), we can choose for $\mathcal{B}$ any bounded and closed set that contains β_0. The first part of Condition 1 is trivially satisfied, because the moment function is linear in β. For the second part, note that

$$||\psi(y,t,x,\beta,p_0(\cdot)|| \leq \left|\left|\frac{y\cdot t}{p_0(\cdot)}\right|\right| + |\beta|$$

$$\leq |y|/\underline{p} + \sup_{\beta\in\mathcal{B}} |\beta|,$$

where $\underline{p} = \inf_{x\in\mathcal{X}} p(x) > 0$ by Assumption 4(ii). Take $b(y,t,x) = |y|/\underline{p} + \sup_{\beta\in\mathcal{B}} |\beta|$ which has finite expectation by Assumption 3(i) and the fact that $\mathcal{B}$ is bounded. For the third part, note that

$$||\psi(y,t,x,\beta,p(\cdot)) - \psi(y,t,x,\beta,p_0(\cdot))|| = \left|\left|\frac{y \cdot t}{p(\cdot)p_0(\cdot)} \cdot \Big(p_0(\cdot) - p(\cdot)\Big)\right|\right|$$

$$\leq \frac{|y|}{\underline{p}^2} \cdot ||p_0(\cdot) - p(\cdot)||.$$

Take $\tilde{b}(y,t,x) = |y|/\underline{p}^2$, which is positive and has a finite expectation by Assumptions 3(i) and 4(ii), and take $\varepsilon = 1$.

Condition 2 $E[\psi(Y,T,X,\beta,p_0(\cdot))] = 0$ *has a unique solution in* $\mathcal{B}$.

We have $E(YT|X) = E(Y|T = 1, X)\Pr(T = 1|X) = E(Y|X)\Pr(T = 1|X)$ where the last equality follows from Assumption 1. Because $p_0(x) = \Pr(T = 1|X = x)$ is bounded away from zero on $\mathcal{X}$ by Assumption 4(ii),

$$E\left(\frac{YT}{p_0(X)}\right) = E\left[\frac{E(YT|X)}{p_0(X)}\right] = E(Y) = \beta_0.$$

Comment: Conditions 1 and 2 imply that $\hat{\beta}$ is weakly consistent if the nonparametric estimator $\hat{p}$ converges in probability to p_0 in the sup norm.

Condition 3 *(i)* β_0 *is an interior point of the compact set* $\mathcal{B}$*,*
(ii), There is a neighborhood $\mathcal{N}$ *of* β_0 *and* $\varepsilon > 0$ *such that for all* $\beta \in \mathcal{N}$ *and* $||p(\cdot) - p_0(\cdot)|| < \varepsilon$, $\psi(y,t,x,\beta,p(\cdot))$ *is differentiable with respect to* β *and the expected value of this derivative is nonzero at* $\beta = \beta_0, p(\cdot) = p_0(\cdot)$*,*
(iii), Condition 1 is satisfied for the derivative of the moment function with respect to β*,*
(iv), $E(||\psi(y,t,x,\beta_0,p_0(\cdot))||^2) < \infty$.

Part (i) is satisfied by an appropriate choice of $\mathcal{B}$. Because the derivative is equal to -1, part (ii) is trivially satisfied for all $\beta \in \mathcal{B}$ and $\varepsilon > 0$. Part (iii) is trivially satisfied for the same reason. For part (iv) note that

$$E(||\psi(Y,T,X,\beta_0,p_0(\cdot))||^2) \leq E\left[\left(\frac{|YT|}{\underline{p}} + \beta_0\right)^2\right] \leq \frac{E(Y^2)}{\underline{p}^2} + \beta_0^2 + 2\frac{E(|Y|)|\beta_0|}{\underline{p}}$$

and this is finite by assumption 3.

Condition 4 *$E((T - p_0(X))^2|X = x)$ is bounded.*

Since T is binary and $p_0(x)$ is the conditional expectation of T given $X = x$, $E((T - p_0(X))^2|X = x)$ is the conditional variance of T, which equals $p_0(x)(1-p_0(x)) \leq 1$ for all x.

Condition 5 *For each K there is a nonsingular $K \times K$ matrix A_K such that for $R^K(x) = A_K r^K(x)$:*
(i) the smallest eigenvalue of $E[R^K(X)R^K(X)']$ is bounded away from zero uniformly in K,
(ii), $R^K(x)$ is a subvector of $R^{K+1}(x)$ for all K,
(iii), for each K there is a nonzero K vector γ such that $\gamma' R^K(x)$ is a nonzero constant for $x \in \mathcal{X}$.

We use the fact that the series is a power series. Together with the Assumption 2 this implies the conditions for Lemma A.15 in Newey (1995) are satisfied. This lemma implies for each K there is a nonsingular matrix A_K such that (i), for $R^K(x) = A_K r^K(x)$, the smallest eigenvalue of $E[R^K(X)R^K(X)']$ bounded away from zero uniformly in K, and (ii) $R^K(x)$ is a subvector of $R^{K+1}(x)$ for all K. Hence for this modified series Conditions 5(i) and (ii) are satisfied. Condition 5(iii) is also satisfied since $r_{1K}(x) \equiv 1$ for all K, and thus for any K vector $\tilde{\gamma}$ with a first component that is not equal to 0, $\tilde{\gamma}' r^K(x) = \tilde{\gamma}_1 \neq 0$ for all $x \in \mathcal{X}$. Because A_K is nonsingular (iii) holds if we set $\gamma = (A_K^{-1})'\tilde{\gamma}$.

Comment: Note that the linear transformation from the series $r^K(x)$ to $R^K(x) = A_K r^K(x)$ does not affect the estimate for β, so we can consider estimation based on the transformed series $R^K(\cdot)$, as we shall do in the sequel.

Comment: We do not consider the case where K is estimated (but K does depend on the number of observations N). This simplifies the rate conditions ($\underline{K} = \overline{K}$ in Condition 6.2 of Newey (1994)).

Condition 6 *There are constants C and $\alpha = s/r$ with s as in Assumption 4 and r the dimension of x such that for all K there is a K vector π_K such that*

$$||p_0(\cdot) - R^K(\cdot)'\pi_K|| \leq CK^{-\alpha}.$$

Condition 6 implies that Assumption 6.3 of Newey (1994) holds with $d = 0$. This condition holds by Lemma A.12 of Newey (1995) (again set $d = 0$).

Condition 7 *There is a function $D(y, t, x, p(\cdot); \beta, \tilde{p}(\cdot))$, linear in $p(\cdot)$, and $b(y, t, x)$ such that, if $||\tilde{p}(\cdot) - p_0(\cdot)||$ and $|\beta - \beta_0|$ are sufficiently small, then*

$$\Big|\Big|\psi\Big(y, t, x, \beta, p(\cdot)\Big) - \psi\Big(y, t, x, \beta, \tilde{p}(\cdot)\Big) - D\Big(y, t, x, p(\cdot) - \tilde{p}(\cdot); \beta, \tilde{p}(\cdot)\Big)\Big|\Big|$$

$$\leq b(y, t, x) \cdot ||p(\cdot) - \tilde{p}(\cdot)||^2,$$

with $E[b(Y, T, X)]$ finite.

For Condition 7 choose

$$D\Big((y, t, x, p(\cdot); \beta, \tilde{p}(\cdot)\Big) = -\frac{y \cdot t}{\tilde{p}(x)^2} \cdot p(x).$$

Note that $D\left(y,t,x,p(\cdot)-\tilde{p}(\cdot);\beta,\tilde{p}(\cdot)\right)$ is the linear term in the Taylor series expansion of $\psi\left(y,t,x,\beta,p(\cdot)\right)$ around $\tilde{p}$. Then

$$\Big|\Big|\psi\Big(y,t,x,\beta,p(\cdot)\Big)-\psi\Big(y,t,x,\beta,\tilde{p}(\cdot)\Big)-D\Big(y,t,x,p(\cdot)-\tilde{p}(\cdot);\beta,\tilde{p}(\cdot)\Big)\Big|\Big|$$

$$=\left|\left|\frac{y\cdot t}{p(\cdot)}-\frac{y\cdot t}{\tilde{p}(\cdot)}+\frac{y\cdot t}{\tilde{p}(\cdot)^2}\cdot(p(\cdot)-\tilde{p}(\cdot))\right|\right|$$

$$=|y\cdot t|\cdot\left|\left|\frac{1}{\tilde{p}(\cdot)^2\cdot p(\cdot)}\cdot(p(\cdot)-\tilde{p}(\cdot))^2\right|\right|$$

$$\leq\frac{|y|}{\underline{p}^3}\cdot\Big|\Big|p(\cdot)-\tilde{p}(\cdot)\Big|\Big|^2.$$

because $||g(.)^2||\leq||g(.)||^2$. By Assumption 3 $b(Y,T,X)=|Y|/\underline{p}^3$ has a finite expected value, and thus Condition 7 is satisfied. This is part (i) of Assumption 6.4 of Newey (1994).

Condition 8 *For the $\alpha=s/r$ as in Condition 6,*

$$\zeta_0(K)\cdot\Big((K/N)^{1/2}+K^{-\alpha}\Big)\longrightarrow 0,$$

and

$$\sqrt{N}\cdot\zeta_0(K)^2\cdot\Big(K/N+K^{-2\alpha}\Big)\longrightarrow 0,$$

where $\zeta_0(K)=||R^K(\cdot)||$.

This is the second part of Newey's Assumption 6.4. By Lemma A.15 in Newey (1995) $\zeta_0(K)\leq CK$. Hence, to satisfy Condition 8, we first show that

$$(K^3/N)^{1/2}\longrightarrow 0$$

and

$$(K^6/N)^{1/2} \longrightarrow 0$$

The second limit implies the first, and the second limit holds if K increases at a rate less than 1/6 as in Assumption 5. In addition,

$$K^{1-\alpha} \longrightarrow 0$$

and

$$N^{1/2}K^{2-2\alpha} \longrightarrow 0$$

The first limit holds because Assumption 4 implies $\alpha > 3$. The second holds if K increases at a rate greater than $1/(4\alpha - 4)$. Assumption 5 implies that the rate exceeds $1/2\alpha$, which, as long as $\alpha > 2$ (which holds by Assumption 4), implies that the rate exceeds $1/(4\alpha - 4)$.

Condition 9 *There is a $b(y,t,x)$, with $E[b(Y,T,X)^2]$ finite, such that for the function $D(y,t,x,p(\cdot);\beta,\tilde{p}(\cdot))$ in Condition 7,*

$$||D\left(y,t,x,p(\cdot);\beta_0,p_0(\cdot)\right)|| \leq b(y,t,x) \cdot ||p(\cdot)||.$$

Although this is a stronger statement than Assumption 6.5 in Newey (1994) where $d > 0$, it follows from our assumptions. Set $b(y,t,x) = |y|/(2\underline{p})^2|$ so that $E[b(Y,T,X)^2] < \infty$ by Assumption 3. Then

$$\left|\left|D\left(y,t,x,p(\cdot);\beta_0,p_0(\cdot)\right)\right|\right| = \left|\left|-\frac{y \cdot t}{p_0(\cdot)^2} \cdot p(\cdot)\right|\right|$$

$$\leq b(y,t,x) \cdot ||p(\cdot)||.$$

Condition 10 *For the $\alpha = s/r$ as in Condition 6*

$$\left(\sum_{k=1}^{K} ||R_{kK}(\cdot)||^2\right)^{1/2} \cdot \left((K/N)^{1/2} + K^{-\alpha}\right) \longrightarrow 0.$$

This is a weaker statement than the second part of Assumption 6.5 in Newey (1994) where $d > 0$. However, Conditions 9 and 10 are sufficient for (A.4) on p. 1376 of Newey (1994). Using Lemma A.15 in Newey (1995) we have for $k = 1, \ldots, K$

$$||R_{kK}(\cdot)|| = \zeta_0(K) \leq CK.$$

Hence,

$$\left(\sum_{k=1}^{K} ||R_{kK}(\cdot)||^2\right)^{1/2} \leq CK^{3/2},$$

and thus

$$\left(\sum_{k=1}^{K} ||R_{kK}(\cdot)||^2\right)^{1/2} \cdot (K/N)^{1/2} \leq C\left(\frac{K^4}{N}\right)^{1/2} \longrightarrow 0,$$

because K increases at a rate slower than 1/4, by Assumption 5. Also,

$$\left(\sum_{k=1}^{K} ||R_{kK}(\cdot)||^2\right)^{1/2} \cdot K^{-\alpha} \leq CK^{3/2-\alpha} \longrightarrow 0,$$

for $\alpha > 3/2$, which holds by Assumption 4. Conditions 9 and 10 imply Assumption 6.5 of Newey (1994).

Condition 11 *There is a function $\delta(x)$ with $E(\delta(X)^2) < \infty$ such that for the function $D(y, t, x, p(\cdot); \beta, \tilde{p}(\cdot))$ in Condition 7,*

$$E\left[D\left(Y, T, X, p(X); \beta_0, p_0(X)\right)\right] = E[\delta(X) \cdot p(X)],$$

for all $p(\cdot)$.

Let $\delta(X)$ be the conditional expectation of $D(\cdot)$ given X:

$$\delta(X) = E\left[D\left(Y, T, X, p(X), \beta_0, p_0(X)\right)\Big|X\right] = E\left[-\frac{Y \cdot T}{p_0(X)^2} \cdot p(X)\Big|X\right]$$

$$= -\frac{E[Y|X]}{p_0(X)} \cdot p(X).$$

Then by the law of iterated expectations

$$E\left[D\left(Y, T, X, p(X); \beta_0, p_0(X)\right)\right] = E\Big[E\left[D\left(Y, T, X, p(X); \beta_0, p_0(X)\right)\Big|X\right]\Big]$$

$$= E[\delta(X) \cdot p(X)].$$

Moreover

$$E(\delta(X)^2) \leq \frac{E(E(Y|X)^2)}{\underline{p}^2} \leq \frac{E(Y^2)}{\underline{p}^2}$$

which is bounded by Assumption 3(i).

Condition 12 *For $K = 1, 2, \ldots,$ there are $K-$vectors π_K and ξ_K such that:*
(i), $N \cdot E\left[|\delta(X) - \xi_K' R^K(X)|^2\right] \cdot E\left[|p_0(X) - \pi_K' R^K(X)|^2\right] \longrightarrow 0,$
(ii), $K\zeta_0(K)^4/N \longrightarrow 0,$
(iii), $\zeta_0(K)^2 \cdot E\left[|p_0(X) - \pi_K' R^K(X)|^2\right] \longrightarrow 0,$
(iv) $E\left[|\delta(X) - \xi_K' R^K(X)|^2\right] \longrightarrow 0.$

By Lemma A.12 in Newey (1995) (see also Lorentz (1986), chapter 8, Theorem 8) there is a $K-$vector π_K such that for $\alpha = s/r$ as in Condition 6

$$||p_0(\cdot) - \pi_K' R^K(\cdot)|| \leq CK^{-\alpha},$$

Because the norm is the supremum over the support of X

$$E\left[||p_0(X) - \pi_K' R^K(X)||^2\right] \leq ||p_0(\cdot) - \pi_K' R^K(\cdot)||^2 \leq CK^{-2\alpha}.$$

Another application of Lemma A.12 gives that there is a $K-$vector ξ_K such that

$$E\left[||\delta(X) - \xi_K' R^K(X)|\right] \le CK^{-\tilde{\alpha}}$$

with $\tilde{\alpha} > 0$ because $\delta(\cdot)$ is continuously differentiable by Assumption 3 ($E(Y|X = x)$ is continuously differentiable by this assumption) and hence $\tilde{\alpha} \ge 1/r > 0$. Thus,

$$N \cdot E\left[||\delta(X) - \xi_K' R^K(X)||^2\right] \cdot E\left[||p(X) - \pi_K' R^K(X)||^2\right] \le CNK^{-2\alpha-2\tilde{\alpha}}$$

$$\le CNK^{-2\alpha} \longrightarrow 0.$$

This holds if K increases at a rate greater than $1/(2\alpha)$ and this is guaranteed by Assumption 5. This establishes (i). For part (ii) we invoke Lemma A.15 of Newey (1995). If K increases at a rate less than 1/5 the limit holds and this is true by Assumption 5. For part (iii), we combine the bound used in the verification of part (i), and the bound on $\zeta_0(K)$, to show that the limit is 0 if $\alpha > 1$ which holds by Assumption 4. The limit in part (iv) is 0 if $\tilde{\alpha} > 0$ and this already has been established.

Given that Conditions 1-12 are satisfied, it follows that Theorem 6.1 in Newey (1994) applies. Hence $\hat{\beta}$ is consistent for β_0, and

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, V),$$

for

$$V = \text{Var}\Big(\psi(Y, T, X, \beta_0, p_0(X)) + \delta(X) \cdot (T - p_0(X)\Big)$$

$$= \text{Var}\left(\frac{Y \cdot T}{p_0(X)} - \beta_0\right) + \text{Var}\left(-\frac{\mu(X)}{p_0(X)} \cdot (T - p_0(X)\right)$$

$$+2 \cdot \text{Cov}\left(\frac{Y \cdot T}{p_0(X)} - \beta_0, -\frac{\mu(X)}{p_0(X)} \cdot (T - p_0(X))\right).$$

Define $\sigma^2(X) = V(Y|X)$. Let

$$V_1 = E\left[\left(\frac{Y \cdot T}{p_0(X)} - \beta_0\right)^2\right]$$

$$= E\left[\left(\frac{Y \cdot T}{p_0(X)} - \beta_0\right) \cdot \frac{Y \cdot T}{p_0(X)}\right]$$

$$= E\left[E\left[\left(\frac{Y \cdot T}{p_0(X)} - \beta_0\right) \cdot \frac{Y \cdot T}{p_0(X)}\middle| T = 1, X\right] \cdot Pr(T = 1|X)\right]$$

$$= E\left[E\left[\frac{Y^2 \cdot T}{p_0(X)^2} - \beta_0 \cdot \frac{Y \cdot T}{p_0(X)}\middle| T = 1, X\right] \cdot Pr(T = 1|X)\right]$$

$$= E\left[\left(\frac{\mu(X)^2 + \sigma^2(X)}{p_0(X)^2} - \beta_0 \cdot \frac{\mu(X)}{p_0(X)}\right) \cdot p_0(X)\right]$$

$$= E\left[\frac{\mu(X)^2}{p_0(X)}\right] + E\left[\frac{\sigma^2(X)}{p_0(X)}\right] - \beta_0 \cdot E[\mu(X)]$$

$$= E\left[\frac{\mu(X)^2}{p_0(X)}\right] + E\left[\frac{\sigma^2(X)}{p_0(X)}\right] - \beta_0^2.$$

Also let

$$V_2 = E\left[\left(-\frac{\mu(X)}{p_0(X)} \cdot (T - p_0(X))\right)^2\right]$$

$$= E\left[\left(\frac{\mu(X)}{p_0(X)} \cdot (T - p_0(X))\right) \cdot \frac{\mu(X)}{p_0(X)} \cdot T\right]$$

$$= E\left[E\left[\left(\frac{\mu(X)}{p_0(X)} \cdot (T - p_0(X))\right) \cdot \frac{\mu(X)}{p_0(X)} \cdot T\middle| T = 1, X\right] \cdot Pr(T = 1|X)\right]$$

$$= E\left[\frac{\mu(X)^2}{p_0(X)^2} \cdot (1 - p_0(X)) \cdot p_0(X)\right]$$

$$= E\left[\frac{\mu(X)^2}{p_0(X)}\right] - E[\mu(X)^2].$$

Finally, let

$$V_{12} = -E\left[\left(\frac{Y\cdot T}{p_0(X)} - \beta_0\right)\cdot\left(\frac{\mu(X)}{p_0(X)}\cdot(T - p_0(X))\right)\right]$$

$$= -E\left[\frac{Y\cdot T}{p_0(X)}\cdot\left(\frac{\mu(X)}{p_0(X)}\cdot(T - p_0(X))\right)\right]$$

$$= -E\left[E\left[\frac{Y\cdot T}{p_0(X)}\cdot\left(\frac{\mu(X)}{p_0(X)}\cdot(T - p_0(X))\right) \mid T = 1, X\right]\cdot Pr(T = 1|X)\right]$$

$$= -E\left[\frac{\mu(X)}{p_0(X)}\cdot\left(\frac{\mu(X)}{p_0(X)}\cdot(1 - p_0(X))\right)\cdot p_0(X)\right]$$

$$= -E\left[\frac{\mu(X)^2}{p_0(X)}\cdot(1 - p_0(X))\right] = -E\left[\frac{\mu(X)^2}{p_0(X)}\right] + E[\mu(X)^2].$$

Then

$$V = V_1 + V_2 + 2\cdot V_{12}$$

$$= E\left[\frac{\mu(X)^2}{p_0(X)}\right] + E\left[\frac{\sigma^2(X)}{p_0(X)}\right] - \beta_0^2 + E\left[\frac{\mu(X)^2}{p_0(X)}\right] - E[\mu(X)^2]$$

$$-2\cdot E\left[\frac{\mu(X)^2}{p_0(X)}\right] + 2\cdot E[\mu(X)^2]$$

$$= E[\mu(X)^2] - \beta_0^2 + E\left[\frac{\sigma^2(X)}{p_0(X)}\right]$$

$$= V(E[Y|X]) + E[V(Y|X)/p_0(X)].$$

Since this variance is also the semiparametric efficiency bound (see Robins and Rotnitzky (1995) and Hahn (1998)), the estimator is efficient.

□

Proof of Theorem 2:

The proof follows the same argument as the previous proof, so we sketch the modifications required. The key step involves choosing

$$D(y,t,x,e(\cdot),\tau,\tilde{e}(\cdot)) = -\left(\frac{y\cdot t}{\tilde{e}(x)^2} + \frac{y\cdot(1-t)}{(1-\tilde{e}(x))^2}\right)\cdot e(x),$$

and

$$\delta(x) = -\frac{\mu_1(x)}{e_0(x)} - \frac{\mu_0(x)}{1-e_0(x)}.$$

The normalized variance of the estimator is then:

$$V = V_1 + V_2 + 2\cdot V_{12},$$

where

$$V_1 = E\left[\psi(Y,T,X,\tau_0,e_0(\cdot))^2\right] = E\left[\left(\frac{Y\cdot T}{e_0(X)} - \frac{Y\cdot(1-T)}{e_0(X)} - \tau_0\right)^2\right]$$

$$= E\left[\frac{V(Y(1)|X)}{e_0(X)}\right] + E\left[\frac{V(Y(0)|X)}{1-e_0(X)}\right] - \tau_0^2 + E\left[\frac{E[Y(1)|X]^2}{e_0(X)}\right] + E\left[\frac{E[Y(0)|X]^2}{1-e_0(X)}\right],$$

$$V_2 = E\left[(\delta(X)\cdot(T-e_0(X))^2\right] = E\left[\left(\left(-\frac{\mu_1(x)}{e_0(x)} - \frac{\mu_0(x)}{1-e_0(x)}\right)\cdot(T-e_0(X))\right)^2\right]$$

$$= E\left[\frac{(1-e_0(X))E[Y(1)|X]^2}{e_0(X)}\right] + E\left[\frac{e_0(X)E[Y(0)|X]^2}{1-e_0(X)}\right] + 2E\left[E[Y(0)|X]E[Y(1)|X]\right],$$

and

$$V_{12} = E\left[\psi(Y,T,X,\tau_0,e_0(\cdot))\cdot(\delta(X)\cdot(T-e_0(X))\right]$$

$$= E\left[\left(\frac{Y\cdot T}{e_0(X)} - \frac{Y\cdot(1-T)}{e_0(X)} - \tau_0\right)\cdot\left(\left(-\frac{\mu_1(x)}{e_0(x)} - \frac{\mu_0(x)}{1-e_0(x)}\right)\cdot(T-e_0(X))\right)\right]$$

$$= -E\left[\frac{E[Y(1)|X]^2 \cdot (1 - e_0(X))}{e_0(X)}\right] - E\left[\frac{E[Y(0)|X]^2 \cdot e_0(X)}{1 - e_0(X)}\right]$$

$$-2 \cdot E\left[E[Y(1)|X] \cdot E[Y(0)|X]\right].$$

This adds up to

$$V = V(E[Y(1) - Y(0)|X]) + E[V(Y(1)|X)/e_0(X)] + E[V(Y(0)|X)/(1 - e_0(X))]),$$

which is equal to the semiparametric efficiency bound for estimation of τ.

□

Proof of Theorem 3: The derivation of the efficiency bound follows the proof in Hahn (1998). The density of $(Y(0), Y(1), T, X)$ with respect to some $\sigma-$finite measure is

$$q(y(0), y(1), t, x) = f(y(0), y(1)|x)e(x)^t(1 - e(x))^{1-t}f(x).$$

The density of the observed data (y, t, x), using the unconfoundedness assumption, is

$$q(y, t, x) = [f_1(y|x)e(x)]^t\,[f_0(y|x)(1 - e(x))]^{1-t}\,f(x),$$

where $f_1(\cdot|x) = \int f(y(0), \cdot|x)dy(0)$, and $f_0(\cdot|x) = \int f(\cdot, y(1)|x)dy(1)$. Consider a regular parametric submodel indexed by θ, with density

$$q(y, t, x|\theta) = [f_1(y|x, \theta)e(x)]^t\,[f_0(y|x, \theta)(1 - e(x))]^{1-t}\,f(x, \theta),$$

which equals $q(y, t, x)$ for $\theta = \theta_0$. Note that θ does not enter into the term $e(x)$, because we are assuming that the propensity score is known. The score is given by

$$\frac{d}{d\theta}\log q(y, t, x|\theta) = s(y, t, x|\theta) = t \cdot s_1(y|x, \theta) + (1 - t) \cdot s_0(y|x, \theta) + s_x(x, \theta),$$

where

$$\begin{aligned} s_1(y|x,\theta) &= \frac{d}{d\theta}\log f_1(y|x,\theta), \\ s_0(y|x,\theta) &= \frac{d}{d\theta}\log f_0(y|x,\theta), \\ s_x(x,\theta) &= \frac{d}{d\theta}\log f(x,\theta). \end{aligned}$$

The tangent space of the model is the set of functions

$$\mathcal{S} = \{t \cdot s_1(y,x) + (1-t)\cdot s_0(y,x) + s_x(x)\}$$

for s_1, s_0, and s_x satisfying

$$\begin{aligned} &\int s_1(y,x)f_1(y|x)dy = 0, \forall x \\ &\int s_0(y,x)f_0(y|x)dy = 0, \forall x \\ &\int s_x(x)f(x)dx = 0. \end{aligned}$$

We are interested in estimating

$$\tau_g \equiv \frac{\int\int g(x)yf_1(y|x)f(x)dydx - \int\int g(x)yf_0(y|x)f(x)dydx}{\int g(x)f(x)dx}$$

So for the parametric submodel indexed by θ,

$$\tau_g(\theta) \equiv \frac{\int\int g(x)yf_1(y|x,\theta)f(x,\theta)dydx - \int\int g(x)yf_0(y|x,\theta)f(x,\theta)dydx}{\int g(x)f(x,\theta)dx}$$

We need to find a function $F_\tau(y,t,x)$ such that for all regular parametric submodels,

$$\frac{\partial \tau_g(\theta_0)}{\partial\theta} = E\left[F_\tau(Y,T,X)s(Y,T,X|\theta_0)\right]$$

First we calculate $\frac{\partial\tau_g(\theta)}{\partial\theta}$. Let $\mu_g \equiv \int g(x)f(x)dx$. Then

$$\frac{\partial \tau_g(\theta_0)}{\partial\theta} =$$

$$\frac{1}{\mu_g}\left[\int\int g(x)ys_1(y|x,\theta_0)f_1(y|x,\theta_0)f(x,\theta_0)dydx - \int\int g(x)ys_0(y|x,\theta_0)f_0(y|x,\theta_0)f(x,\theta_0)dydx\right]$$

$$+\frac{1}{\mu_g}\left[\int g(x)\left\{E[Y(1)-Y(0)|X=x]-\tau_g\right\}s_x(x,\theta_0)f(x,\theta_0)dx\right].$$

The following choice for F_τ satisfies the condition:

$$F_\tau(Y,T,X) = \frac{T\cdot g(X)}{\mu_g\cdot e(X)}\left(Y - E[Y(1)|X]\right) - \frac{(1-T)\cdot g(X)}{\mu_g\cdot(1-e(X))}\left(Y-E[Y(0)|X]\right)$$

$$+\frac{g(X)}{\mu_g}\left(E[Y(1)-Y(0)|X]-\tau_g\right).$$

Hence τ_g is pathwise differentiable. By Theorem 2, in section 3.3 of Bickel, Klaassen, Ritov, and Wellner (1993), the variance bound is the expected square of the projection of $F_\tau(Y,T,X)$ on $\mathcal{S}$. Since $F_\tau \in \mathcal{S}$, the variance bound is

$$E[F_\tau(Y,T,X)^2] = E\left[\frac{g(X)^2}{(\mu_g)^2 e_0(X)}V(Y(1)|X)\right] + E\left[\frac{g(X)^2}{(\mu_g)^2(1-e_0(X))}V(Y(0)|X)\right]$$
$$+E\left[\frac{g(X)^2}{(\mu_g)^2}\left(E(Y(1)|X)-E(Y(0)|X)-\tau_g\right)^2\right]$$

□

Proof of Theorem 4:

We choose

$$D(y,t,x,e(\cdot),\tau_g,\tilde{e}(\cdot)) = -g(x)\left[\frac{y\cdot t}{\tilde{e}(x)^2}+\frac{y\cdot(1-t)}{(1-\tilde{e}(x))^2}\right]\cdot e(x),$$

and

$$\delta(x) = -g(x)\left[\frac{\mu_1(x)}{e_0(x)}+\frac{\mu_0(x)}{1-e_0(x)}\right]$$

The normalized variance of the estimator is

$$V = M^{-1}\Omega M^{-1'}$$

where

$$M = E\left[\frac{\partial\psi(Y,T,X,\tau_g,e_0)}{\partial\tau_g}\right] = -E(g(X)) = -\mu_g,$$

$$\Omega = Var\left[\psi(Y,T,X,\tau_g,e_0) + \delta(X)(T - e_0(X))\right].$$

We can write

$$\Omega = \Omega_1 + \Omega_2 + 2\Omega_{12},$$

where

$$\Omega_1 = E\left[\psi(Y,T,X,\tau_g,e_0)^2\right],$$

$$\Omega_2 = E\left[(\delta(X)(T - e_0(X)))^2\right],$$

and

$$\Omega_{12} = E\left[\psi(Y,T,X,\tau_g,e_0)\delta(X)(T - e_0(X))\right].$$

Straightforward calculations show that

$$\begin{aligned}\Omega_1 \quad = \quad & E\left[\frac{g(X)^2}{e_0(X)}V[Y(1)|X]\right] + E\left[\frac{g(X)^2}{1-e_0(X)}V[Y(0)|X]\right] \\ & +\tau_g^2 E\left[g(X)^2\right] - 2\cdot\tau_g E\left[g(X)^2\left(\mu_1(X) - \mu_0(X)\right)\right] \\ & +E\left[\frac{g(X)^2}{e_0(X)}\mu_1(X)^2\right] + E\left[\frac{g(X)^2}{1-e_0(X)}\mu_0(X)^2\right].\end{aligned}$$

$$\Omega_2 = E\left[\frac{g(X)^2(1-e_0(X))\mu_1(X)^2}{e_0(X)}\right] + E\left[\frac{g(X)^2 e_0(X)\mu_0(X)^2}{1-e_0(X)}\right] + 2\cdot E\left[g(X)^2\mu_0(X)\mu_1(X)\right]$$

$$\Omega_{12} = -E\left[\frac{g(X)^2(1-e_0(X))\mu_1(X)^2}{e_0(X)}\right] - E\left[\frac{g(X)^2 e_0(X)\mu_0(X)^2}{1-e_0(X)}\right] - 2\cdot E\left[g(X)^2\mu_0(X)\mu_1(X)\right].$$

Combining these results gives

$$
\begin{aligned}
V &= E\left[\frac{g(X)^2}{\mu_g^2 e_0(X)}V(Y(1)|X)\right] + E\left[\frac{g(X)^2}{\mu_g^2(1-e_0(X))}V(Y(0)|X)\right] \\
&\quad + E\left[\frac{g(X)^2}{\mu_g^2}\left(E(Y(1)|X) - E(Y(0)|X) - \tau_g\right)^2\right]
\end{aligned}
$$

□

Proof of Corollary 5: Apply Theorem 4 with $g(x) = e_0(x)$, and compare to the variance bound calculated in Hahn (1998).

□

Proof of Theorem 6:

We choose

$$D(y,t,x,e(\cdot),\tau,\tilde{e}(\cdot)) = \left[-\frac{y\cdot(1-t)}{(1-\tilde{e}(x))^2} - \tau\right]\cdot e(x),$$

and

$$\delta(x) = \left[-\frac{\mu_0(x)}{1-e_0(x)} - \tau\right]$$

The normalized variance of the estimator is

$$V = M^{-1}\Omega M^{-1'}$$

where

$$M = E\left[\frac{\partial\psi(Y,T,X,\tau,e_0)}{\partial\tau}\right] = -E(e(X)) = -\mu_e,$$

$$\Omega = Var\left[\psi(Y,T,X,\tau,e_0) + \delta(X)(T - e_0(X))\right].$$

We can write

$$\Omega = \Omega_1 + \Omega_2 + 2\Omega_{12},$$

where

$$\Omega_1 = E\left[\psi(Y,T,X,\tau,e_0)^2\right],$$

$$\Omega_2 = E\left[(\delta(X)(T-e_0(X)))^2\right],$$

and

$$\Omega_{12} = E\left[\psi(Y,T,X,\tau,e_0)\delta(X)(T-e_0(X))\right].$$

Straightforward calculations show that

$$\begin{aligned}\Omega_1 &= E\left[e_0(X)V[Y(1)|X]\right] + E\left[\frac{e_0(X)^2}{1-e_0(X)}V[Y(0)|X]\right] \\ &\quad +\tau^2 E\left[e_0(X)^2\right] - 2\cdot\tau E\left[e_0(X)^2\left(\mu_1(X)-\mu_0(X)\right)\right] \\ &\quad +E\left[e_0(X)^2\mu_1(X)^2\right] + E\left[\frac{e_0(X)^2}{1-e_0(X)}\mu_0(X)^2\right].\end{aligned}$$

$$\Omega_2 = E\left[\frac{e_0(X)}{1-e_0(X)}\mu_0(X)^2\right] + 2\cdot\tau E\left[e_0(X)\mu_0(X)\right] + \tau^2 E\left[(1-e_0(X))e_0(X)\right].$$

$$\Omega_{12} = -E\left[e_0(X)\mu_0(X)\mu_1(X)\right] - \tau E\left[e_0(X)(1-e_0(X))\mu_1(X)\right]$$

$$-E\left[\frac{e_0(X)^2}{1-e_0(X)}\mu_0(X)^2\right] - \tau E\left[e_0(X)^2\mu_0(X)\right].$$

Combining these results gives the variance, which we compare to the efficiency bound in Hahn (1994).

□

References

Angrist, J. D., and J. Hahn, (1999) "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects," NBER Technical Working Paper 241.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A., (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Chamberlain, G., (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions", *Journal of Econometrics* 34, 305–334.

Crepon, B., Kramarz, F., and Trognon, A., (1998), "Parameters of Interest, Nuisance Parameters and Orthogonality Conditions: an Application to Autoregressive Error Component Models," *Journal of Econometrics* 82, 135-156.

Dehejia, R., and S. Wahba, (1999) "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Assocation* 94, 1053-1062.

Hahn, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (2), 315-331.

Hansen, L., (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica* 50, 1029-1054.

Heckman, J., H. Ichimura, and P. Todd, (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies* 64, 605-654.

Heckman, J., H. Ichimura, and P. Todd, (1998), "Matching As An Econometric Evaluations Estimator," *Review of Economic Studies* 65, 261-294.

Heckman, J., and Robb., R., (1985), "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market*

Data, New York: Cambridge University Press.

Hellerstein, J., and G. Imbens, (1999), "Imposing Moment Restrictions from Auxiliary Data by Weighting," *Review of Economics and Statistics.*

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

Hotz, V. J., G. Imbens, and J. Mortimer, (1999), "Predicting the Efficacy of Future Training Programs using Past Experiences," NBER Technical Working Paper T0238.

Imbens, G., (1997), "One-step Estimators in Overidentified Generalized Method of Moments Estimator," *Review of Economic Studies.*

Imbens, G., R. Spady, and P. Johnson, (1998), "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica* 66 (2), 333-357.

Lancaster, T., (1990), "A Paradox in Choice-based Sampling", mimeo, Department of Economics, Brown University.

Lechner, M., (1999), "Earnings and Employment Effects of Continuous Off-the-job Training in East Germany after Unification," *Journal of Business and Economic Statistics* 17, 74-90.

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

Lorentz, G., (1986), *Approximation of Functions*, New York, Chelsea Publishing Company.

Newey, W., (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62, 1349-1382.

Newey, W., (1995), "Convergence Rates for Series Estimators," in *Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C. R. Rao*, Maddala, Phillips, and Srinivasan (eds.), Cambridge, Basil Blackwell.

Newey, W., and D. McFadden, (1994), "Large Sample Estimation," in *Handbook of*

Econometrics, Vol. 4, Engle and McFadden (eds.), North Holland.

Qian, H., and P. Schmidt, (1999), "Improved Instrumental Variables and Generalized Method of Moments Estimators", *Journal of Econometrics* 91, 145-169.

Qin, and J. Lawless, (1994), "Generalized Estimating Equations," *Annals of Statistics* 22, 300-325.

Robins, J., (1998), "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.

Robins, J., and Y. Ritov, (1997), "Towards a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models," *Statistics in Medicine* 16, 285-319.

Robins, J., and A. Rotnitzky, (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90 (429), 122-129.

Robins, J., A. Rotnitzky, and L. Zhao, (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90 (429), 106-121.

Rosenbaum, P., (1987), "Model-Based Direct Adjustment", *Journal of the American Statistical Association* 82, 387-394.

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

Rosenbaum, P., and D. Rubin, (1983), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society*, Series B, 45.

Rosenbaum, P., and D. Rubin, (1985), "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*

79, 516–524.

Rotnitzky, A., and J. Robins, (1995), "Semiparametric Regression Estimation in the Presence of Dependent Censoring," *Biometrika* 82 (4), 805-820.

Rubin, D., (1976), "Inference and Missing Data," *Biometrika* 63, 581–92.

Rubin, D., (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2(1), 1-26.

Rubin, D., and N. Thomas, (1992), "Affinely Invariant Matching Methods with Ellipsoidal Distributions," *Annals of Statistics* 20 (2) 1079-1093.

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

Rubin, D., and N. Thomas, (1996), "Matching Using Estimated Propensity Scores: Relating Theory to Practice," *Biometrics* 52, 249-264.

Wooldridge, J., (1999), "Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples", *Econometrica* 67, No. 6, 1385-1406.