Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score

Keisuke Hirano, Guido W. Imbens, and Geert Ridder

USC Center for Law, Economics & Organization Research Paper No. C02-13

USC
THE LAW SCHOOL

CENTER FOR LAW, ECONOMICS AND ORGANIZATION RESEARCH PAPER SERIES

Sponsored by the John M. Olin Foundation

University of Southern California Law School
Los Angeles, CA 90089-0071

Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score[1]

Keisuke Hirano – University of Miami[2]
Guido W. Imbens – University of California at Berkeley[3] and NBER
Geert Ridder – University of Southern California[4]

First Draft, February 2000
This Draft, July, 2002

[1]We thank Gary Chamberlain, Jinyong Hahn, James Robins, Donald Rubin, Jeffrey Wooldridge, four anonymous referees, and seminar participants at the University of Chicago, UC Davis, the University of Michigan, Michigan State University, UC Irvine, the University of Miami, and Harvard-MIT for comments. Financial support for this research was generously provided through NSF grants SBR-9818644 and SES 0136789 (Imbens) and ?? (Hirano).

[2]Department of Economics, University of Miami, P. O. Box 248126, Coral Gables FL 33124-6550, khirano@miami.edu.

[3]University of California, Department of Economics, and Department of Agricultural and Resource Economics, 661 Evans Hall, #3880, Berkeley, CA 94720-3880, imbens@econ.berkeley.edu.

[4]Department of Economics, University of Southern California, Los Angeles, CA, gridder@usc.edu.

Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score

Hirano, Imbens and Ridder

Abstract

We are interested in estimating the average effect of a binary treatment on a scalar outcome. If assignment to the treatment is unconfounded, that is, independent of the potential outcomes given covariates, biases associated with simple treatment-control average comparisons can be removed by adjusting for differences in the covariates. Rosenbaum and Rubin (1983a) show that adjusting solely for differences between treated and control units in a scalar function of the covariates, the propensity score, also removes all biases associated with differences in covariates. Although adjusting for the propensity score removes all the bias, this can come at the expense of efficiency, as shown by Hahn (1998), Heckman, Ichimura, Todd (1998), and Rotnitzky and Robins (1995). We show that weighting by the inverse of a nonparametric estimate of the propensity score, rather than the true propensity score, leads to efficient estimates of the average treatment effect. We provide intuition for this result by showing that this estimator can be interpreted as an empirical likelihood estimator that efficiently incorporates the information about the propensity score.

1. Introduction

Estimating the average effect of a binary treatment or policy on a scalar outcome is a basic goal of many empirical studies in economics. If assignment to the treatment is unconfounded (i.e., independent of potential outcomes conditional on covariates or pre-treatment variables, an assumption also known as selection on observables), the average treatment effect can be estimated by averaging within-subpopulation differences of treatment and control averages. If there are many covariates, this strategy may not be desirable or even feasible. An alternative approach is based on the propensity score, the conditional probability of receiving treatment given covariates. Rosenbaum and Rubin (1983a, 1985) show that, under the assumption of unconfoundedness, adjusting solely for differences in the propensity score between treated and control units removes all bias associated with differences in the covariates. Recent applications of propensity score methods in economics include Dehejia and Wahba (1999), Heckman, Ichimura and Todd (1997), and Lechner (1999).

Although adjusting for differences in the propensity score removes all bias, it need not be as efficient as adjusting for differences in all covariates, as shown by Hahn (1998), Heckman, Ichimura and Todd (1998), and Robins, Mark and Newey (1992). However, Rosenbaum (1987), Rubin and Thomas (1997), and Robins, Rotnitzky and Zhao (1995) show that using parametric estimates of the propensity score, rather than the true propensity score, can avoid some of these efficiency losses.

In this paper we propose estimators based on adjusting for nonparametric estimates of the propensity score that are fully efficient for estimation of average treatment effects. Our estimators weight observations by the inverse of nonparametric estimates of the propensity score, rather than the true propensity score. Extending results from Newey (1994) to derive the large sample properties of these semiparametric estimators, we show that they achieve the semiparametric efficiency bound. We also show that in the case the propensity score is known the proposed estimators can be interpreted as empirical likelihood estimators (e.g., Qin and Lawless, 1994; Imbens, Spady and Johnson, 1998) that efficiently incorporate the information about the propensity score.

If the propensity score is known, as, for example, in randomized experiments, these estimators can be used to improve efficiency over simply differencing treatment and control averages. In that case an attractive choice for the non-parametric series estimator for the propensity score uses the true propensity score as the leading term in the series. The estimators can also be used in the case where the propensity score is unknown, as an alternative to the previously proposed efficient estimators that require nonparametric estimation of functions in addition to the propensity score.

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. Section 5 concludes.

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, for $i = 1, \ldots, N$, let T_i indicate whether the treatment of interest was received, with $T_i = 1$ if unit i receives the active treatment, and $T_i = 0$ if unit i receives the control treatment. Using the potential outcome notation popularized by Rubin (1974), 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 $Y_i \equiv T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)$. In addition, we observe a vector of covariates denoted by X_i.[2] Initially we focus on the population average treatment effect:

$$\tau \equiv \mathbb{E}[Y(1) - Y(0)]. \tag{1}$$

We shall also discuss estimation of weighted average treatment effects

$$\tau_{wate} \equiv \frac{\int \mathbb{E}[Y(1) - Y(0)|X = x]g(x)dF(x)}{\int g(x)dF(x)}, \tag{2}$$

[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.

[2] These variables are assumed not to be affected by the treatment.

where $g(\cdot)$ is a known function of the covariates.[3] In the special case where the weight function $g(x)$ is equal to the propensity score $p(x) = \mathbb{P}(T = 1|X = x)$, this leads under the unconfoundedness assumption to the average effect for the treated:

$$\tau_{\text{treated}} \equiv \mathbb{E}[Y(1) - Y(0)|T = 1]. \tag{4}$$

The central problem of evaluation research is that for unit i we observe either $Y_i(0)$ or $Y_i(1)$, but never both. To solve the identification problem, we maintain throughout the paper the unconfoundedness assumption (Rubin, 1978; Rosenbaum and Rubin, 1983a), also known as the selection–on–observables assumption (Barnow, Cain, and Goldberger, 1980), which asserts that conditional on the observed covariates, the treatment indicator is independent of the potential outcomes. Formally:

Assumption 1 (Unconfounded Treatment Assignment)

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

As Heckman, Ichimura and Todd (1998) point out, for identification of the average treatment effect τ this assumption can be weakened to mean independence ($\mathbb{E}[Y(t)|T, X] = \mathbb{E}[Y(t)|X]$ for $t = 0, 1$). If one is interested in the average effect for the treated, the assumption can be further weakened to only require that $\mathbb{E}[Y(0)|T, X] = \mathbb{E}[Y(0)|X]$. In this paper we focus on the full independence assumption, to be consistent with much of the literature.

Under unconfoundedness we can estimate the average treatment effect conditional on covariates, $\tau(x) \equiv \mathbb{E}[Y(1) - Y(0)|X = x]$, because

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

[3] An alternative estimand which we do not consider here is the direct weighted average treatment effect of the form

$$\tau_{dwate} = \frac{\int \mathbb{E}[Y(1) - Y(0)|X = x]g(x)dx}{\int g(x)dx}, \tag{3}$$

where the weighting is only over the known function $g(x)$. Note that in general $F(x)$ is unknown so that knowledge of $g(x)$ does not imply knowledge of $g(x)dF(x)$ and the other way around; estimation strategies for the two estimands in (2) and (3) are in general different. Estimands of the latter type can be fit into the framework of Robins and Ritov (1997).

$$= \mathbb{E}[Y|T=1, X=x] - \mathbb{E}[Y|T=0, X=x].$$

In turn, the population average treatment effect can be obtained by averaging the $\tau(x)$ over the distribution of X: $\tau = \mathbb{E}[\tau(X)]$. 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.[4] To avoid the need to match units on the values of all covariates, Rosenbaum and Rubin (1983a, 1985) developed an approach based on the propensity score, the probability of selection into the treatment group:

$$p(x) \equiv \mathbb{P}(T=1|X=x) = \mathbb{E}[T|X=x], \tag{5}$$

which is assumed to be bounded away from zero and one. Their key insight was that if treatment and potential outcomes are independent conditional on all covariates, they are also independent conditional on the conditional probability of receiving treatment given covariates, that is, conditional on the propensity score. Formally, as shown by Rosenbaum and Rubin (1983a), unconfoundedness implies

$$T \perp (Y(0), Y(1)) \,\Big|\, p(X), \tag{6}$$

implying that adjustment for the propensity score suffices for removing all biases associated with differences in the covariates.

2.2 Previous Results

The model set out above, and related models, have been examined by many researchers. In an important paper Hahn (1998), studying the same model as in the current paper, calculates the semiparametric efficiency bounds, and proposes efficient estimators, for τ and $\tau_{treated}$. Hahn's estimator for τ, which is efficient irrespective of whether the propensity score is known, nonparametrically estimates the two conditional expectations $\mathbb{E}[YT|X=x]$ and $\mathbb{E}[Y(1-T)|X=x]$ as well as the propensity score $p(x)$, and then imputes the missing

[4] 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, especially when the dimension of the conditioning variable is high. For discussions of these issues, see Robins and Ritov (1997) and Angrist and Hahn (1999) and references therein.

potential outcomes as $\hat{Y}_i(1) = \hat{\mathbb{E}}[YT|X_i]/\hat{p}(X_i)$ and $\hat{Y}_i(0) = \hat{\mathbb{E}}[Y(1-T)|X_i]/(1-\hat{p}(X_i))$. Hahn shows that the estimator for the population average treatment effect conditioning only on the true propensity score rather than on the full set of covariates does not in general reach the efficiency bound. In addition Hahn concludes that for estimating $\tau_{treated}$ knowledge of the propensity score is informative and derives efficient estimators both with and without such knowledge. A difference between Hahn's estimators and our proposed estimators is that Hahn requires nonparametric estimation of the propensity score as well as the two conditional means $\mathbb{E}[YT|X = x]$ and $\mathbb{E}[Y(1-T)|X = x]$, whereas our proposed estimator only requires nonparametric estimation of the propensity score.

Heckman, Ichimura and Todd (1997, 1998) and Heckman, Ichimura, Smith and Todd (1998) focus on the average treatment effect for the treated $\tau_{treated}$. They consider estimators based on local linear 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. Our results show that by making the specification of the propensity score sufficiently flexible, this approach leads to a fully efficient estimator.

Robins, Mark and Newey (1992), Robins and Rotnitzky (1995), Robins, Rotnitzky and Zhao (1995), and Rotnitzky and Robins (1995) study the related problem of inference for parameters in regression models where some data are Missing At Random (MAR, Rubin, 1976; Little and Rubin, 1987). Rotnitzky and Robins (1995) show, in parametric settings, that weighting using the estimated rather than true selection probability can improve efficiency, 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 missing data case Robins and Rotnitzky (1995) also propose an efficient estimator which relies on an initial consistent, but not necessarily efficient, estimator of the full population parameters. The estimator proposed in this paper (and also the Hahn estimator) is efficient, but does not require an initial consistent estimator.

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, $\beta_0 = \mathbb{E}[Y]$, 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 formal results to Section 4.

The analogue to the unconfoundedness assumption here is the assumption that the Y_i are Missing At Random (MAR, Rubin, 1976), or

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

The role of the propensity score is played here by the selection probability: $p(x) = \mathbb{E}[T|X = x] = \mathbb{P}(T = 1|X)$. First, we restrict our attention in this section to the case with a single binary covariate.[5] Let N_{tx} denote the number of observations with $T_i = t$ and $X_i = x$, for $t, x \in \{0, 1\}$. Furthermore, suppose the true selection probability is constant, $p(x) = 1/2$ for all $x \in \{0, 1\}$.[6] The normalized variance bound for β_0 is

$$V_{bound} = 2 \cdot \mathbb{E}\left[\mathbb{V}(Y|X)\right] + \mathbb{V}\left(\mathbb{E}[Y|X]\right), \tag{7}$$

the variance of the maximum likelihood estimator.

[5] An efficient estimator is easily obtained by averaging the within-subsample difference of treatment/control averages. It can also be found by specializing the more general estimators in Robins and Rotnitzky (1995) and Hahn (1998) to this simple case. The discussion here is solely intended to convey intuition for the formal results that will be presented in Section 4.

[6] Thus the missing data are Missing Completely At Random (MCAR, Rubin, 1976; Little and Rubin, 1987).

The "true weights" estimator weights the complete observations by the inverse of the true selection probability:

$$\hat{\beta}_{tw} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i \cdot T_i}{p(X_i)} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i \cdot T_i}{1/2}. \tag{8}$$

Its large sample normalized variance is

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

strictly larger than the variance bound (7).

The second estimator weights the complete observations 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 setting, 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 $N_{10}/(N_{00} + N_{10})$, and for units with $X_i = 1$, the proportion of observed outcomes is $N_{11}/(N_{01} + N_{11})$. Thus the estimated selection probability is

$$\hat{p}(x) = \begin{cases} N_{10}/(N_{00} + N_{10}) & \text{if } x = 0, \\ N_{11}/(N_{01} + N_{11}) & \text{if } x = 1. \end{cases}$$

Then the proposed "estimated weights" estimator is:

$$\hat{\beta}_{ew} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i \cdot T_i}{\hat{p}(X_i)}. \tag{9}$$

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

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

Thus, 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. This will suggest 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.

An alternative interpretation of the estimated-weights estimator 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 $\mathbb{E}[\psi_1(Y, X, T, \beta_0)] = 0$, with

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

The GMM estimator based on the single moment restriction $\psi_1(\cdot)$, given knowledge of the selection probability, is the true-weights estimator $\hat{\beta}_{tw}$ in (8). However, this estimator is not necessarily efficient, because it ignores the additional information that is available in the form of knowledge of the selection probability. This additional information can be written in moment condition form as $\mathbb{E}[T - p(X)|X] = \mathbb{E}[T - 1/2|X] = 0$. With a binary covariate this conditional moment restriction corresponds to two marginal moment restrictions, $\mathbb{E}[\psi_2(Y, T, X, \beta_0)] = 0$, 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.[7] Here it is of particular interest to consider the empirical likelihood estimator (e.g., Qin and Lawless, 1994; Imbens, 1997; Kitamura and Stutzer, 1997; Imbens, Spady and Johnson, 1998), which 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:

$$L(\pi) = \sum_{i=1}^{N} \ln \pi_i, \tag{10}$$

[7]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 as in seemingly unrelated regressions. See, e.g., Hellerstein and Imbens (1999) and Qian and Schmidt (1999).

subject to the adding-up restriction $\sum_i \pi_i = 1$ and the moment conditions $\sum_i \pi_i \psi(y_i, t_i, x_i, \beta) = 0$. Solving for $\hat{\pi}_i$ and $\hat{\beta}_{el}$ by maximizing (10) subject to the restrictions, leads, after some manipulation, to:

$$\hat{\pi}_i = \left(1 + \frac{\frac{N_{11}}{N_{01}+N11} - 1/2}{1/4} \cdot x_i \cdot (t_i - 1/2) + \frac{\frac{N_{10}}{N_{00}+N_{10}} - 1/2}{1/4} \cdot (1 - x_i) \cdot (t_i - 1/2)\right)^{-1},$$

which in turn implies

$$\hat{\beta}_{el} = \sum_{i=1}^{N} 2 \cdot \hat{\pi}_i \cdot Y_i \cdot T_i = \hat{\beta}_{ew},$$

equal to the estimated weights estimator.

The above discussion generalizes directly to the case with general discrete covariates. With continuous covariates knowledge of the propensity score implies a conditional moment restriction corresponding to an infinite number of unconditional moment restrictions (e.g., Chamberlain, 1987). Using a series estimator for the propensity score captures the information content of such a conditional moment restriction by a sequence of unconditional moment restrictions.

The empirical likelihood interpretation suggests that 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) who find that 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, whereas using the true values of the nuisance parameters may lead to efficiency losses. These results are also linked to the literature on weighting in stratified sampling. Translated to our treatment effect 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. In a general discussion of weighted M-estimators Wooldridge (1999, 2002) shows that weighting by the inverse of estimated rather than population probabilities can lead to efficiency gains.

4. Efficient Estimation Using Estimated Weights

In this section we present the main results of the paper. We discuss three distinct cases. First, we consider the problem of estimating the population average treatment effect under the unconfoundedness assumption. This includes as a special case the extension of the binary-covariate MAR example of the previous section to continuous covariates. Second, we consider estimation of weighted average treatment effects. Finally, we consider estimation of the effect of the treatment on the treated, which in the known propensity score case will follow directly from the solution to the general weighted average treatment effect case. This discussion will shed additional light on Hahn's (1998) interesting result that for this parameter knowledge of the propensity score affects the efficiency bound, as well as on the findings in Heckman, Ichimura and Todd (1998) that in the case of the average treatment effect for the treated neither using the true nor using the estimated propensity score dominates the other.

4.1 Estimating Population Average Treatment Effects

In this section we use the set up from Section 2 with for each unit a pair of potential outcomes $(Y(0), Y(1))$ and focus on efficient estimation of the population average treatment effect, $\tau^* = \mathbb{E}[Y(1) - Y(0)]$.[8] As before, $p(x) = \mathbb{P}(T = 1|X = x)$ is the propensity score, the probability of receiving the active treatment. We maintain the unconfoundedness assumption. Define $\mu_w(x) \equiv \mathbb{E}[Y(w)|X = x]$ and $\sigma_w^2(x) = \mathbb{V}(Y(w)|X = x)$ to be the conditional mean and variance of $Y(w)$ respectively. Under unconfoundedness we have $\mu_w(x) = \mathbb{E}[Y|W = w, X = x]$ and $\sigma_w^2(x) = \mathbb{V}(Y|W = w, X = x)$. We can characterize τ^* through the moment equation:

$$\mathbb{E}\left[\psi(Y, T, X, \tau^*, p^*(X))\right] = 0,$$

[8] Whenever necessary to avoid confusion we will use a superscript $*$ to denote true (population) values, so that τ^* denotes the population average treatment effect and $p^*(x)$ the true (population) propensity score.

where

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

Given an estimator $\hat{p}(x)$ for the propensity score, we estimate τ^* by setting the average moment evaluated at the estimated selection probability equal to zero as a function of τ: $\frac{1}{N}\sum_{i=1}^{N}\psi(Y_i,T_i,X_i,\hat{\tau},\hat{p}(X_i)) = 0$, leading to the estimator

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i\cdot T_i}{\hat{p}(X_i)} - \frac{Y_i\cdot(1-T_i)}{1-\hat{p}(X_i)}\right). \tag{12}$$

Because $p^*(x)$ is a conditional expectation this semiparametric estimation problem directly fits into the framework of Newey (1994), and his results using least squares estimators for $p^*(x)$ based on series apply (see the working paper version, Hirano, Imbens and Ridder, 2000). However, because $p^*(x)$ is a probability such an approach has the unattractive feature that it approximates a probability by a linear function. We therefore estimate $p^*(x)$ in a sieve approach (e.g., Geman and Hwang, 1982) by the Series Logit Estimator (SLE). 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. Although the theory is derived for general sequences of approximating functions, the most common class of functions are power series. Let $\lambda = (\lambda_1, \ldots, \lambda_r)'$ be an r-dimensional vector of nonnegative integers (multi-indices), with norm $|\lambda| = \sum_{j=1}^{r}\lambda_j$, let $(\lambda(k))_{k=1}^{\infty}$ be a sequence that includes all distinct multi-indices and satisfies $|\lambda(k)| \leq |\lambda(k+1)|$, and let $x^\lambda = \prod_{j=1}^{r} x_j^{\lambda_j}$. For a sequence $\lambda(k)$ we consider the series $r_{kK}(x) = x^{\lambda(k)}$. If we denote the logistic cdf by $L(a) = \exp(a)/(1+\exp(a))$, the SLE for $p^*(x)$ is defined by $\hat{p}(x) = L(R^K(x)'\hat{\pi}_K^*)$ with

$$\hat{\pi}_K = \arg\max_{\pi}\sum_{i=1}^{N}\left(T_i\cdot\ln L(R^K(X_i)'\pi) + (1-T_i)\cdot\ln(1-L(R^K(X_i)'\pi))\right).$$

In Appendix A we discuss the relevant asymptotic theory.

In addition to the unconfoundedness assumption the following assumptions are used to derive the properties of the estimator. First, we restrict the distribution of X, $Y(0)$ and $Y(1)$:

Assumption 2 (Distribution of X)

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

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

Assumption 3 (Distribution of $Y(0), Y(1)$)

(i), $\mathbb{E}[Y(0)^2] < \infty$ and $\mathbb{E}[Y(1)^2] < \infty$,

(ii), $\mu_0(x)$ and $\mu_1(x)$ are continuously differentiable for all $x \in \mathbb{X}$.

The next assumption requires sufficient smoothness of the propensity score.

Assumption 4 (Selection Probability)

The propensity score $p^*(x)$ satisfies the following conditions: For all $x \in \mathbb{X}$

(i), $p^*(x)$ is continuously differentiable of order $s \geq 4 \cdot r$ where r is the dimension of X ,

(ii), $p^*(x)$ is bounded away from zero and one: $0 < \underline{p} \leq p^*(x) \leq \overline{p} < 1$.

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

Assumption 5 (Series Estimator)

The series logit estimator of $p^*(x)$ uses a power series with $K = N^\nu$ for some $1/(4(s/r-1)) < \nu < \frac{1}{9}$.

The restriction on the derivatives (Assumption $4(i)$) guarantees the existence of a ν that satisfies the conditions in Assumption 5. Under these conditions we can state the first result.

Theorem 1 *Suppose Assumptions 1-5 hold. Then:*

(i), $\hat{\tau} \xrightarrow{p} \tau^*$,

(ii), $\sqrt{N}(\hat{\tau} - \tau^*) \xrightarrow{d} \mathcal{N}(0, V)$, where

$$V = \mathbb{E}\left[\left(\left(\frac{YT}{p^*(X)} - \frac{Y(1-T)}{1-p^*(X)} - \tau^*\right) - \left(\frac{\mu_1(X)}{p^*(x)} + \frac{\mu_0(X)}{1-p^*(X)}\right)(T - p^*(X))\right)^2\right]$$

$$= \mathbb{E}\left[(\tau(X) - \tau)^2 + \frac{\sigma_1^2(X)}{p^*(X)} + \frac{\sigma_0^2(X)}{1-p^*(X)}\right],$$

and

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

Proof: see Appendix B.

Remark 1: This result also covers the extension of the binary-covariate MAR example in Section 3 to the continuous covariate case. Just set $Y = 0$ if $T = 0$ and set $Y(0)$ identically equal to 0.

Remark 2: Theorem 1 establishes the result for continuous X. If X has both continuous and discrete components, this can be dealt with in a conceptually straighforward manner by using the continuous covariate estimator within samples homogenous in the discrete covariates, at the expense of additional notation.

Derivations presented in Appendix B show that the estimator in Theorem 1 can be represented as asymptotically linear:

$$\hat{\tau} = \tau^* + \frac{1}{N}\sum_{i=1}^{N}\Big(\psi(Y_i, T_i, X_i, \tau^*, p^*(X_i)) + \alpha(T_i, X_i)\Big) + o_p(1/\sqrt{N}),$$

where $\psi(\cdot)$ is defined in (11) and

$$\alpha(t, x) = -\left(\frac{\mu_1(x)}{p^*(x)} + \frac{\mu_0(x)}{1 - p^*(x)}\right)(t - p^*(x)). \tag{13}$$

The known-weights estimator, (12) with $\hat{p}(x)$ replaced by $p^*(x)$, is asymptotically linear with score function $\psi(\cdot)$. The function $\alpha(t, x)$ represents the effect on the score function of estimating $p^*(x)$. Its first factor, $-(\mu_1(x)/p^*(x) + \mu_0(x)/(1 - p^*(x)))$, is the conditional expectation of the derivative of the moment condition $\psi(y, t, x, \tau^*, p^*(x))$ with respect to $p^*(x)$. Hence, the score linearizes the estimator with respect to τ (trivial since the estimator is already linear in τ) and $p(\cdot)$.

The asymptotically linear representation of $\hat{\tau}$ implies that its asymptotic variance equals

$$\mathbb{E}\left[\Big(\psi(Y, T, X, \tau^*, p^*(X)) + \alpha(T, X)\Big)\right]^2, \tag{14}$$

shown in the appendix to be equal to the variance expression in Theorem 1. We estimate this variance by replacing the unknown quantities τ, $p^*(\cdot)$ and $\alpha(\cdot)$ by estimates and replacing the expectation by a sample average:

$$\hat{V} = \frac{1}{N}\sum_{i=1}^{N} \left(\psi(Y_i, T_i, X_i, \hat{\tau}, \hat{p}(X_i)) + \hat{\alpha}(T_i, X_i)\right)^2. \tag{15}$$

The estimation of $\alpha(t,x)$ requires some additional explanation. The second factor, $t - p^*(x)$ is estimated as $t - \hat{p}(x)$. The first factor, $-(\mu_1(x)/p^*(x) + \mu_0(x)/(1-p^*(x)))$, can be written as the conditional expectation of $-(YT/p^*(X)^2 + Y(1-T)/(1-p^*(X))^2)$ given X. We therefore estimate the first factor in $\alpha(t,x)$ by nonparametric regression of $-(YT/\hat{p}(X)^2 + Y(1-T)/(1-\hat{p}(X))^2)$ on X, using the same series approach as we used for estimating $p^*(x)$. Thus

$$-\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_iT_i}{\hat{p}(X_i)^2} + \frac{Y_i(1-T_i)}{(1-\hat{p}(X_i))^2}\right)R^K(X_i)\right)'\left(\frac{1}{N}\sum_{i=1}^{N} R^K(X_i)R^K(X_i)'\right)^{-1} R^K(x),$$

with $R^K(x)$ the same series of approximating functions as before, is used as an estimator for $-(\mu_1(x)/p^*(x) + \mu_0(x)/(1-p^*(x)))$, and the function $\alpha(t,x)$ is estimated by $\hat{\alpha}(t,x)$:

$$\hat{\alpha}(t,x) = -\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_iT_i}{\hat{p}(X_i)^2} + \frac{Y_i(1-T_i)}{(1-\hat{p}(X_i))^2}\right)R^K(X_i)\right)' \tag{16}$$

$$\left(\frac{1}{N}\sum_{i=1}^{N} R^K(X_i)R^K(X_i)'\right)^{-1} R^K(x)(t - \hat{p}(x)).$$

The following theorem describes the formal result.

Theorem 2 *Suppose Assumptions 1-5 hold. Then $\hat{V}$ is consistent for V.*

Proof: see Appendix B.

In practive bootstrapping methods may be valuable alternatives to the above variance estimator.

4.2 Estimating the Weighted Average Treatment Effect

In this section we generalize the previous result to τ_{wate}, the weighted average treatment effect for a known weight function $g(x)$. One motivation for considering this estimand is that by choosing $g(x)$ appropriately, we can obtain treatment effects for subpopulations defined by X. In addition, by choosing $g(x)$ equal to the propensity score $p^*(x)$, we can recover the average effect of the treatment on the treated, as will be discussed below.

To estimate τ_{wate}, we use the following moment function:

$$\psi(y, t, x, \tau_{wate}, p(x)) = g(x) \cdot \left(\frac{y \cdot t}{p(x)} - \frac{y \cdot (1-t)}{1-p(x)} - \tau_{wate} \right), \tag{17}$$

leading to the estimator

$$\hat{\tau}_{wate} = \sum_i g(X_i) \left[\frac{Y_i \cdot T_i}{\hat{p}(X_i)} - \frac{Y_i \cdot (1-T_i)}{1-\hat{p}(X_i)} \right] \Big/ \sum_i g(X_i).$$

This estimator is asymptotically linear with the score function

$$\hat{\tau}_{wate} = \frac{1}{\mathbb{E}[g(X)]} \frac{1}{N} \sum_{i=1}^{N} \Big(\psi(Y_i, T_i, X_i, \tau_{wate}, p^*(x)) + \alpha(T_i, X_i) \Big) + o_p(1/\sqrt{N}),$$

where now

$$\alpha(t, x) = -g(x) \cdot \left(\frac{\mu_1(x)}{p^*(x)} + \frac{\mu_0(x)}{1-p^*(x)} \right) (t - p^*(x)).$$

The asymptotic variance can be estimated as

$$\hat{V} = \frac{1}{\left(\sum_i g(X_i)/N\right)^2} \frac{1}{N} \sum_{i=1}^{N} \left(\psi(Y_i, T_i, X_i, \hat{\tau}_{wate}, \hat{p}(X_i)) + \hat{\alpha}(T_i, X_i) \right)^2,$$

with an estimator for $\alpha(t, x)$ analogous to that for the average treatment effect:

$$\hat{\alpha}(t, x) = -g(x) \frac{1}{N} \sum_{i=1}^{N} \left(\left(\frac{Y_i T_i}{\hat{p}_K(X_i)^2} + \frac{Y_i (1-T_i)}{(1-\hat{p}_K(X_i))^2} \right) R^K(X_i) R^K(X_i) \right)'$$

$$\left(\frac{1}{N} \sum_{i=1}^{N} R^K(X_i) R^K(X_i)' \right)^{-1} R^K(x)(t - \hat{p}_K(x)).$$

Similar reasoning to the previous results gives the following results:

Theorem 3 *Suppose Assumptions 1-5 hold, that g(x) is bounded from above and that* $\mathbb{E}[g(X)] > 0$. *Then*
(i), $\hat{\tau}_{wate} \xrightarrow{p} \tau_{wate}$,
(ii), $\sqrt{N}(\hat{\tau}_{wate} - \tau_{wate}) \xrightarrow{d} \mathcal{N}(0, V)$, *with*

$$V = \frac{1}{\mathbb{E}[g(X)]^2}\mathbb{E}\left[g(X)^2\left(\tau(X) - \tau_{wate}\right)^2 + \frac{g(X)^2}{p^*(X)}\sigma_1^2(X) + \frac{g(X)^2}{1-p^*(X)}\sigma_0^2(X)\right]$$

and (iii), $\hat{V}$ *is consistent for* V.

The proof for this theorem follows the same line of argument as that for Theorems 1 and 2 and is omitted.

Remark: We could weaken Assumption 4(ii), the assumption that the propensity score is bounded away from 0 and 1, by the assumption that $g(x)/p^*(x)$ and $g(x)/(1-p^*(x))$ 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.

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

Theorem 4 *The semiparametric efficiency bound for estimation of* τ_{wate} *is*

$$V = \frac{1}{\mathbb{E}[g(X)]^2}\mathbb{E}\left[g(X)^2\left(\tau(X) - \tau_{wate}\right)^2 + \frac{g(X)^2}{p^*(X)}\sigma_1^2(X) + \frac{g(X)^2}{1-p^*(X)}\sigma_0^2(X)\right].$$

Proof: See Appendix B.

4.3 Estimating the Average Treatment Effect for the Treated

Under unconfoundedness the average treatment effect for the treated (Rubin, 1977; Heckman and Robb, 1985, Heckman, Ichimura and Todd, 1997, 1998) is a special case of the weighted average treatment effect, corresponding to the weighting function $g(x) = p^*(x)$. To see this first note that under unconfoundedness

$$\tau_{treated} = \mathbb{E}[Y(1) - Y(0)|T=1] = \mathbb{E}\left[\mathbb{E}[Y(1) - Y(0)|X, T=1]\Big|T=1\right]$$

$$= \mathbb{E}\left[\mathbb{E}[Y(1) - Y(0)|X]\Big|T = 1\right] \mathbb{E}[\tau(X)|T = 1].$$

Second, the latter is equal to

$$\mathbb{E}[\tau(X)|T = 1] = \int \tau(x) dF(x|T = 1) = \int \tau(x) p^*(x) dF(x) \Big/ \int p^*(x) dF(x),$$

which is τ_{wate} with $g(x)$ equal to $p^*(x)$. Hence we can use the moment equation (17) with $p^*(x)$ substituted for $g(x)$:

$$\psi(y, t, x, \tau_{treated}, p(x)) = p^*(x) \cdot \left(\frac{y \cdot t}{p(x)} - \frac{y \cdot (1-t)}{1 - p(x)} - \tau_{treated}\right). \tag{18}$$

The estimator is the solution to

$$0 = \sum_{i=1}^{N} p^*(X_i) \cdot \left(\frac{Y_i \cdot T_i}{\hat{p}(X_i)} - \frac{Y_i \cdot (1 - T_i)}{1 - \hat{p}(X_i)} - \tau_{treated}\right), \tag{19}$$

with the same nonparametric series estimator $\hat{p}(x)$ as before.

The next result, which follows directly 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 1 *Suppose that Assumptions 1-5 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*

$$V = \frac{1}{\mathbb{E}[p^*(X)]^2} \mathbb{E}\left[p^*(X)^2 \left(\tau(X) - \tau_{treated}\right)^2 + p^*(X)\sigma_1^2(X) + \frac{p^*(X)^2}{1 - p^*(X)} \sigma_0^2(X)\right],$$

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

The proof for this corollary is omitted as the result directly follows from Theorem 4.

Note that in the moment function (18) the propensity score appears in two places, first as $p^*(x)$ multiplying the remainder of the moment function where it replaced the general weight function $g(x)$ in (17), and second as $p(x)$ in the denominator of the two terms. We only use the estimated propensity score in the second part in the efficient estimator in (19). The

result of the theorem above implies that this is more efficient than using the true propensity score everywhere and solving

$$0 = \sum_{i=1}^{N} p^*(X_i) \cdot \left(\frac{Y_i \cdot T_i}{p^*(X_i)} - \frac{Y_i \cdot (1 - T_i)}{1 - p^*(X_i)} - \tau_{treated} \right), \tag{20}$$

or using the estimated propensity score everywhere, which amounts to solving

$$0 = \sum_{i=1}^{N} \hat{p}(X_i) \cdot \left(\frac{Y_i \cdot T_i}{\hat{p}(X_i)} - \frac{Y_i \cdot (1 - T_i)}{1 - \hat{p}(X_i)} - \tau_{treated} \right). \tag{21}$$

A direct implication of this result is that the simple sample average of the outcomes for the treated $\sum_i Y_i T_i / \sum_i T_i$ is less efficient for the population average $\mathbb{E}[Y(1)|T = 1]$ than the weighted average $\sum_i Y_i T_i (p^*(X_i)/\hat{p}(X_i)) / \sum_i p^*(X_i)$ where the weights are the ratio of the true and estimated propensity score. Another implication is that the estimators characterized by (20) and (21) cannot in general be ranked in terms of efficiency as there are effects of opposite signs (e.g., Heckman, Ichimura and Todd, 1997).

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 $\hat{\tau}_{treated}$ cannot be used since it makes use of $p^*(x)$. However, we can use the estimated propensity score in place of $p^*(x)$ in the weighting of observations as in (21). Call this estimator $\hat{\tau}_{te}$. The next theorem shows that this estimator is efficient if the propensity is not known.

Theorem 5 *Suppose that Assumptions 1-5 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*

$$V = \frac{1}{\mathbb{E}[p^*(X)]^2} \mathbb{E}\left[p^*(X) \left(\tau(X) - \tau_{treated}\right)^2 + p^*(X)\sigma_1^2(X) + \frac{p^*(X)^2}{1 - p^*(X)} \sigma_0^2(X) \right],$$

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

The proof goes along the same lines as that for Theorems 1 and 2 and is omitted.

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 empirical likelihood estimators, and estimators from the literature on variable probability sampling.

The estimators proposed in this paper require fewer functions to be estimated nonparametrically than other efficient estimators previously proposed in the literature. Which estimators have more attractive finite sample properties, and which have more attractive computational properties, remain open questions. The results underline the important role played by the propensity score in estimation of average causal effects.

Appendix A: Logistic Series Estimator

In this appendix we derive the relevant properties of the logistic series estimator, which can be interpreted as a sieve estimator (e.g., Geman and Hwang, 1982). Let $r^K(x) = (r_{1K}(x), \ldots, r_{KK}(x))'$ be a K-vector of functions. The triangular array of functions $r^K(x), K = 1, 2, \ldots$ is the basis for the approximation of the propensity score. In particular, we approximate a function $f : \mathbb{R}^r \to \mathbb{R}$ by $\gamma_K' r^K(x)$. Because $\gamma_K' r^K(x) = \gamma_K' A_K^{-1} A_K r^K(x)$ we can also use $R^K(x) = A_K r^K(x)$ as the basis of approximation. By choosing A_K appropriately we obtain a system of orthogonal (with respect to some weight function) polynomials. The properties of the series logit estimator and the proof of Theorem 1 are mostly for a general system of approximating functions. We indicate where the properties of the approximating class of functions are used.

In particular, we consider approximation by power series. One possible choice for a

triangular sequence of powers of x is

$$r^1(x) = 1, \quad r^2(x) = \begin{bmatrix} 1 \\ x_1 \end{bmatrix}, \quad r^{r+1}(x) = \begin{bmatrix} 1 \\ x_1 \\ . \\ . \\ . \\ x_r \end{bmatrix}, \quad r^{r+1}(x) = \begin{bmatrix} 1 \\ x_1 \\ . \\ . \\ . \\ x_r \\ x_1^2 \end{bmatrix} \tag{22}$$

Linear combinations of the elements of the vectors $r^K(x)$ are the approximating power series. A power series in which $x_1, \ldots, x_r$ are included up to power n has $(n+1)^r$ terms. Hence, if we use the sequence in (22) and set $K = (n+1)^r$, then $r^K(x)'\gamma_K$ has powers in all variables up to n. If the function f is s times continuously differentiable and $K = (n+1)^r$, then by Theorem 8, p. 90, in Lorentz (1986) there is a K-vector γ_K such that for $R^K(x) = A_K r^K(x)$,

$$\sup_{x \in \mathbb{X}} \left| f(x) - R^K(x)'\gamma_K \right| < Cn^{-s} = CK^{-\frac{s}{r}} \tag{23}$$

Here, and in the sequel, C denotes a generic positive constant[9], and $\mathbb{X}$ is the support of X. If we multiply the constant C in (23) by $2^{\frac{r}{s}}$, then this inequality holds for all K.

The series logit estimator of the population propensity score $p^*(x)$ is $\hat{p}_K(x) = L(R^K(x)'\hat{\pi}_K)$ with $L(z) = \exp(z)/(1+\exp(z))$ the logistic cdf, and

$$\hat{\pi}_K = \arg\max_{\pi} \sum_{i=1}^{N} \left(T_i \ln L(R^K(X_i)'\pi) + (1 - T_i) \ln(1 - L(R^K(X_i)'\pi)) \right). \tag{24}$$

For $N \to \infty$ and K fixed we have $\hat{\pi}_K \xrightarrow{p} \pi_K^*$, with π_K^* the pseudo true value:

$$\pi_K^* = \arg\max_{\pi} \mathbb{E}_X \left[p_0(X) \ln L(R^K(X)'\pi) + (1 - p_0(X)) \ln(1 - L(R^K(X)'\pi)) \right]. \tag{25}$$

We define the pseudo true propensity score: $p_K^*(x) = L(R^K(x)'\pi_K^*)$.

In the proofs for the theorems we need some properties of this logit estimator. For those properties it is convenient to distinguish between the deterministic difference between the true propensity score and the psuedo true propensity score and the stochastic difference

[9] If two constants are needed, we will use the generic notation C_1, C_2.

between the estimated propensity score and the pseudo true value. In the remainder of this appendix we therefore derive (i) a uniform bound on the difference between $p^*(x)$ and $p_K^*(x)$ and (ii) a bound on the sampling variance in the form of the stochastic order of $\|\hat{\pi}_K - \pi_K^*\|$.

The support of X is assumed to be a compact subset of $\mathbb{R}^r$. Moreover, we assume that $p^*(x)$ is s times continuously differentiable on $\mathbb{X}$, and that $p^*(x)$ is bounded from away 0 and 1 on $\mathbb{X}$. To ensure that the approximation of p^* is between 0 and 1 we do not approximate p^*, but rather the log odds ratio which is also s times continuously differentiable and bounded on $\mathbb{X}$. Hence by (23) there is a π_K such that

$$\sup_{x\in\mathbb{X}} \left| \ln\left(\frac{p^*(x)}{1-p^*(x)}\right) - R^K(x)'\pi_K \right| < CK^{-\frac{s}{r}} \tag{26}$$

From (26), for all $x \in \mathbb{X}$

$$L(R^K(x)'\pi_K - CK^{-\frac{s}{r}}) - L(R^K(x)'\pi_K) < p^*(x) - L(R^K(x)'\pi_K) < \tag{27}$$

$$< L(R^K(x)'\pi_K + CK^{-\frac{s}{r}}) - L(R^K(x)'\pi_K)$$

By the mean value theorem applied to the lower and upper bound and by $L'(R^K(x)\tilde{\pi}) = L(R^K(x)'\tilde{\pi})(1 - L(r^K(x)'\tilde{\pi}))$ being bounded at intermediate values $\tilde{\pi}$,[10] we find that the lower and upper bound are bounded by $-\frac{1}{4}CK^{-\frac{s}{r}}$ and $\frac{1}{4}CK^{-\frac{s}{r}}$, respectively. Hence, there is a π_K such that

$$\sup_{x\in\mathbb{X}} \left| p^*(x) - L(R^K(x)'\pi_K) \right| < CK^{-\frac{s}{r}} \tag{28}$$

We now obtain a uniform bound on the error in the approximation of p^* by p_K. We will use the matrix norm $||A|| = \sqrt{\mathrm{tr}(A'A)}$. Note that this is the usual Euclidean norm if A is a column vector[11]. If A is a scalar, we denote the norm by $|A|$. By (25) π_K^* is a solution to

[10] Below we show that we can make this assumption without loss of generality, if we impose a restriction on the sequences $K(N)$.

[11] It is useful to list some properties of this norm. Let A and B be $K \times K$ matrices and c be a K vector. Then $\|AB\|^2 = \sum_i \sum_j \left(\sum_k a_{ik}b_{kj}\right)^2$. Applying the vector Cauchy-Schwartz inequality to the inner sum we find $\|AB\| \leq \|A\|\|B\|$. By the maximum inequality for quadratic forms $\|Ac\| \leq \sqrt{\lambda_{max}(A'A)}\|c\|$ which gives a sharp upper bound (the upper bound $\|A\|\|c\|$ is not sharp in general). We also frequently use the Cauchy-Schwartz inequality for expectations that implies that for nonnegative random variables X, Y, $\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$.

the first order condition

$$\mathbb{E}_X\left[(p^*(X) - L(R^K(X)'\pi_K^*))R^K(X)\right] = 0$$

Hence the average approximation error is orthogonal to all components of R^K. If, as we assume, $R^K(x)$ has a constant component for all K, we have

$$\mathbb{E}_X\left[(p^*(X) - L(R^K(X)'\pi_K^*))\right] = 0 \tag{29}$$

However, we need a uniform bound on the absolute approximation error. By the mean value theorem there is a $\tilde{\pi}_K$ such that[12]

$$L(R^K(x)'\pi_K) = L(R^K(x)'\pi_K^*) + L'(R^K(x)'\tilde{\pi}_K)R^K(x)'(\pi_K - \pi_K^*) \tag{30}$$

By adding and subtracting $L(R^K(X)'\pi_K)$ in (29) we obtain

$$\left|\mathbb{E}_X\left[L(R^K(X)'\pi_K) - L(R^K(X)'\pi_K^*)\right]\right| \leq \tag{31}$$

$$\leq \mathbb{E}_X\left[\left|p^*(X) - L(R^K(X)'\pi_K)\right|\right] \leq CK^{-\frac{s}{r}}$$

Substitution of (30) in the left-hand side of this inequality gives, because $L'(R^K(x)'\tilde{\pi}_K)$ is bounded from 0 on $\mathbb{X}$ and $\|\mathbb{E}_X(R^K(X))\| > 0$ if $R^K(x)$ has a constant component

$$\left|\mathbb{E}_X\left[L(R^K(X)'\pi_K) - L(R^K(X)'\pi_K^*)\right]\right| \tag{32}$$

$$= \left|\mathbb{E}_X\left[L'(R^K(X)'\tilde{\pi}_K)R^K(X)'(\pi_K - \pi_K^*)\right]\right| \geq C\|\pi_K - \pi_K^*\|$$

Combining (31) and (32) we obtain

$$\|\pi_K - \pi_K^*\| \leq CK^{-\frac{s}{r}} \tag{33}$$

Hence by (30) and the Cauchy-Schwartz inequality

$$\sup_{x\in\mathbb{X}}\left|p^*(x) - L(R^K(x)'\pi_K^*)\right| \leq \tag{34}$$

[12] $\tilde{\pi}_K$ is a generic notation for an intermediate value in an application of the mean value theorem.

$$\sup_{x\in\mathbb{X}}\left|p^*(x)-L(R^K(x)'\pi_K)\right|+\sup_{x\in\mathbb{X}}\left|L(R^K(x)'\pi_K^*)-L(R^K(x)'\pi_K)\right|\leq$$

$$\leq C_1K^{-\frac{s}{r}}+C_2\sup_{x\in\mathbb{X}}\|R^K(x)\|\|\pi_K-\pi_K^*\|\leq C\sup_{x\in\mathbb{X}}\|R^K(x)\|K^{-\frac{s}{r}}$$

For orthogonal polynomials Newey (1994ab, 1997) gives the bound

$$\sup_{x\in\mathbb{X}}\|R^K(x)\|=O(K) \tag{35}$$

In general, the bound depends on the array of approximating functions that is used. In the sequel we use the notation

$$\zeta(K)=\sup_{x\in\mathbb{X}}\|R^K(x)\| \tag{36}$$

Combining (34) and (36), we have

$$\sup_{x\in\mathbb{X}}\left|p^*(x)-L(R^K(x)'\pi_K^*)\right|=O(K^{-s/r}\zeta(K)),$$

as a uniform rate on the difference between the propensity score and the pseudo true propensity score.[13]

For the second part of this appendix we derive the stochastic order of $\|\hat{\pi}_K-\pi_K^*\|$ if K increases without bounds with N. Let $K(N)$ be a sequence of values of K with $\lim_{N\to\infty}K(N)=\infty$. We obtain a bound on the variance of $\|\hat{\pi}_{K(N)}-\pi_{K(N)}\|$ if N is large. $\hat{\pi}_{K(N)}$ satisfies the first order conditions

$$\sum_{i=1}^{N}R^{K(N)}(X_i)(T_i-L(R^{K(N)}(X_i)'\hat{\pi}_{K(N)}))=0 \tag{37}$$

By the mean value theorem we obtain

$$\tilde{\Sigma}_{K(N)}(\hat{\pi}_{K(N)}-\pi_{K(N)})=V_{K(N)} \tag{38}$$

[13]Note that although by (26) there is a π_K such that

$$\sup_{x\in\mathbb{X}}\left|p^*(x)-L(R^K(x)'\pi_K)\right|=O(K^{-s/r}),$$

the order for the approximation at the pseudo true value π_K^* is not as good at $O(K^{-s/r}\zeta(K))$.

with

$$\tilde{\Sigma}_K = \frac{1}{N}\sum_{i=1}^{N} R^K(X_i)R^K(X_i)'L'(R^K(X_i)'\tilde{\pi}_K) \tag{39}$$

$$V_K = \frac{1}{N}\sum_{i=1}^{N} R^K(X_i)(T_i - L(R^K(X_i)'\pi_K^*)) \tag{40}$$

If the smallest eigenvalue of the matrix $S_K = \mathbb{E}_X\left[R^K(X)R^K(X)'\right]$ is bounded from 0 for all K[14], the same is true for the smallest eigenvalue of

$$\Sigma_K = \mathbb{E}_X\left[R^K(X)R^K(X)'L'(R^K(X)'\pi_K^*)\right] \tag{41}$$

for all K. Moreover, if $L'(R^K(x)'\tilde{\pi}_K)$ is bounded from 0 on $\mathbb{X}$, then the smallest eigenvalue of $\tilde{\Sigma}_{K(N)}$ is bounded from below by a positive constant(independent of $K(N)$) times the smallest eigenvalue of

$$\hat{S}_{K(N)} = \frac{1}{N}\sum_{i=1}^{N} R^{K(N)}(X_i)R^{K(N)}(X_i)' \tag{42}$$

and this eigenvalue is bounded from 0 with probability 1, if $\|\hat{S}_{K(N)} - S_K\| \xrightarrow{p} 0$ (Newey (1995), Lemma A.4) which is the case if $\frac{\zeta(K(N))^4}{N} \to 0$ (Newey (1995), Lemma A.10) with $\zeta(K)$ defined in (36). This condition on $K(N)$, which we refer to as the large sample identification condition, ensures that the logit model is identified if $N \to \infty$, even if the number of terms in the series increases without bounds.

Under the assumptions mentioned in the previous paragraph

$$\hat{\pi}_{K(N)} - \pi_{K(N)} = \tilde{\Sigma}_{K(N)}^{-1} V_{K(N)} + o_P(1) \tag{43}$$

By the properties of the trace norm

$$\mathbb{E}\left[\left\|\tilde{\Sigma}_{K(N)}^{-1}\sqrt{N}V_{K(N)}\right\|^2\right] \leq \mathbb{E}\left[\frac{1}{\lambda_{min}\left(\tilde{\Sigma}_{K(N)}\right)}\left\|\sqrt{N}V_{K(N)}\right\|^2\right] \tag{44}$$

[14]If $R^K(x)$ are orthogonal polynomials, a sufficient condition for this is that the density of X is bounded from 0 on $\mathbb{X}$.

Because $\lambda_{min}\left(\tilde{\Sigma}_{K(N)}\right)$ is bounded from 0 with probability 1 if $N \to \infty$, the right hand side is bounded by (if N is large)

$$C\mathbb{E}\left[\|\sqrt{N}V_{K(N)}\|^2\right] = C\text{tr}(\Sigma_{K(N)}) \leq C\zeta(K(N)). \tag{45}$$

Thus

$$\mathbb{E}(\|\tilde{\Sigma}_{K(N)}^{-1}V_{K(N)}\|^2) \leq C\frac{\zeta(K(N))}{N}$$

and we conclude that

$$\|\hat{\pi}_{K(N)} - \pi_{K(N)}\| = O\left(\sqrt{\frac{\zeta(K(N))}{N}}\right) + o(1). \tag{46}$$

The derivations above assume in a number of places that $L'(R^{K(N)}(x)'\tilde{\pi}_{K(N)})$ is bounded from 0 on $\mathbb{X}$ with probability 1 and for some range of sequences $K(N)$. We shall show that we can make this assumption without loss of generality, if the sequence $K(N)$ satisfies the large sample identification condition $\zeta(K(N))^4/N \to 0$ without assuming that $\sup_{\pi,x\in\mathbb{X}} \pi L'(R^K(x)'\pi)$ is bounded away from zero. In particular, let

$$\eta_K = \inf_{x\in\mathbb{X}} L(R^K(x)'\pi_K^*)(1 - L(R^K(x)'\pi_K^*)) \tag{47}$$

Because p^* is bounded from 0, η_K is strictly positive if K is greater than or equal to say K_0 if p^* is sufficiently often differentiable (see (34)). We can remedy this by substituting

$$\max\left\{\eta_{K_0}/2, L(R^{K(N)}(x)'\tilde{\pi}_{K(N)})(1 - L(R^{K(N)}(x)'\tilde{\pi}_{K(N)}))\right\}, \tag{48}$$

for $L'(R^{K(N)}(x)'\tilde{\pi}_{K(N)})$, which is bounded away from zero. Now consider (use the Cauchy-Schwartz inequality)

$$\sup_{x\in\mathbb{X}} |R^{K(N)}(x)'(\hat{\pi}_{K(N)} - \pi_{K(N)})| \leq \sup_{x\in\mathbb{X}} \|R^{K(N)}(x)\| \cdot \|\hat{\pi}_{K(N)} - \pi_{K(N)})\| = \tag{49}$$

$$= \zeta(K(N))O_P\left(\frac{\sqrt{\zeta(K(N))}}{\sqrt{N}}\right)$$

Hence if the large sample identification condition on $K(N)$ holds, then the left hand side converges to 0 in probability. Now by the mean value theorem

$$\sup_{x\in\mathbb{X}} |L(R^{K(N)}(x)'\hat{\pi}_{K(N)}) - L(R^{K(N)}(x)'\pi_{K(N)})| \leq \tag{50}$$

$$\leq \frac{1}{4} \sup_{x\in\mathbb{X}} |R^{K(N)}(x)'(\hat{\pi}_{K(N)} - \pi_{K(N)})| = o_P(1)$$

and we conclude

$$\lim_{N\to\infty} \Pr\left(\inf_{x\in\mathbb{X}} L(R^{K(N)}(x)'\hat{\pi}_{K(N)} \right) > \varepsilon) = 1 \tag{51}$$

An analogous result shows that for these sequences $K(N)$ the estimated propensity score is bounded from 1 with probability 1 and we conclude that $L(R^{K(N)}(x)'\tilde{\pi}_{K(N)})(1-L(R^{K(N)}(x)'\tilde{\pi}_{K(N)}))$ is bounded from 0 with probability 1, as required. To be precise, the error that we make is $o_P(1)$ uniformly over sequences $K(N)$ that satisfy the large sample identification condition. Hence, the lower bound in the definition (48) is not needed. The result that the estimated propensity score is bounded from 0 on $\mathbb{X}$ with probability 1 and for sequences $K(N)$ that satisfy the large sample identification condition is used below.

In the sequel we use the properties (34) and (46) of the series logit estimator of the propensity score. In (46) we can write K for $K(N)$ with the understanding that this stochastic order holds for all sequences $K(N)$ with $\zeta(K(N))^4/N \to 0$, so that that the right-hand side of (46) is effectively $o_P(1)$.

Appendix B: Proofs of Theorems

Proof of Theorem 1:

To ease the notational burden we present the proof for the special case with $Y(0) = 0$ with probability one. This can be interpreted as the special case where one is interested in estimating the average outcome $\beta = \mathbb{E}[Y(1)]$, where $Y(1)$ is missing at random conditional on the covariates X. Thus it is the direct extension of the binary-covariate example in Section 3. Since the average treatment effect case simply amounts to estimating two averages where in both cases the variables are missing at random, the argument for the general case is exactly

analogous, only involving substantially longer equations. In the proof we therefore follow the missing at random set up with interest in $\beta = \mathbb{E}[Y(1)]$, the missing at random assumption $Y(1) \perp T|X$, and a random sample of $(T_i, X_i, Y_i)_{i=1}^N$, where $Y_i = Y_i(1) \cdot T_i$.

The estimated weight estimator $\hat{\beta}_{ew}$ is

$$\hat{\beta}_{ew} = \frac{1}{N} \sum_{i=1}^{N} \frac{T_i \cdot Y_i}{\hat{p}_K(X_i)} \tag{52}$$

with $\hat{p}_K(X_i) = L(R^K(X_i)'\hat{\pi}_K)$. The key part of the proof is to show that

$$\left| \sqrt{N}(\hat{\beta}_{ew} - \beta_0) - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \left(\frac{T_i \cdot Y_i}{p^*(X_i)} - \beta_0 \right) - \frac{\mu_1(X_i)}{p^*(X_i)} (T_i - p^*(X_i)) \right\} \right| = o_P(1) \tag{53}$$

This implies that $\hat{\beta}_{ew}$ is asymptotically linear, i.e. behaves asymptotically as a sample average, with score function $\psi(Y, T, X, \beta_0, p^*(\cdot)) + \alpha(T, X)$, where

$$\psi(y, t, x, \beta, p(\cdot)) = \frac{t \cdot y}{p(x)} - \beta, \qquad \text{and} \quad \alpha(t, x) = -\frac{\mu_1(x)}{p^*(x)} \cdot (t - p^*(x))$$

The first term of the score function, $\psi(\cdot)$, is equal to the score that would obtain if we substitute the population probability p^* for the estimator $\hat{p}_K$ in (52). The second term, $\alpha(\cdot)$, gives the contribution of estimating p^* to the asymptotic distribution of $\hat{\beta}_{ew}$. This contribution is linear in $T - p^*(X)$. Hence, the score linearizes the estimator with respect to β (trivial since the estimator is already linear in β) and $p(\cdot)$. The asymptotic variance of $\hat{\beta}_{ew}$ is equal to the variance of $\psi(Y, T, x, \beta_0, p^*(X)) + \alpha(T, X)$ (note that its mean is 0). The three components of this variance are

$$\mathbb{E}[\psi(Y, T, X, \beta_0, p^*(\cdot))^2] = \mathbb{E}\left[\frac{\mu_1(X)^2}{p^*(X)} \right] + \mathbb{E}\left[\frac{\sigma_1^2(X)}{p^*(X)} \right] - \beta_0^2,$$

$$\mathbb{E}[\alpha(T, X)^2] = \mathbb{E}\left[\frac{\mu_1(X)^2}{p^*(X)} \right] - \mathbb{E}[\mu_1(X)^2],$$

$$\mathbb{E}[\psi(Y, T, X, \beta_0, p^*(\cdot)) \cdot \alpha(T, X)] = -\mathbb{E}\left[\frac{\mu_1(X)^2}{p^*(X)} \right] + \mathbb{E}[\mu_1(X)^2],$$

so that

$$\mathbb{E}[\psi(Y,T,X,\beta_0,p^*(\cdot))+\alpha(T,X)^2]=\mathbb{E}[\mu_1(X)^2]-\beta_0^2+\mathbb{E}\left[\frac{\sigma_1^2(X)}{p^*(X)}\right]$$

$$=\mathbb{V}(\mathbb{E}[Y(1)|X])+\mathbb{E}[\mathbb{V}(Y(1)|X)/p^*(X)],$$

which is the variance in Theorem 1, specialized to the case with $\mu_0(x)=\sigma_0^2(x)=0$.

In the proof of (53) we rewrite the difference by adding and subtracting a number of terms, so that we can bound the differences. We give the asymptotic order of all differences, which makes it easier to understand the role of the assumptions. We have

$$\sqrt{N}(\hat{\beta}_{ew}-\beta_0)=\frac{1}{\sqrt{N}}\sum_{i=1}^N\left(\frac{T_iY_i}{\hat{p}_K(X_i)}-\frac{T_iY_i}{p^*(X_i)}+\frac{T_iY_i}{p^*(X_i)^2}(\hat{p}_K(X_i)-p^*(X_i))\right) \tag{54}$$

$$+\frac{1}{\sqrt{N}}\sum_{i=1}^N\left(-\frac{T_iY_i}{p^*(X_i)^2}(\hat{p}_K(X_i)-p^*(X_i))+\int_{\mathbb{X}}\frac{\mu_1(x)}{p^*(x)}(\hat{p}_K(x)-p^*(x))\mathrm{d}F_0(x)\right) \tag{55}$$

$$-\sqrt{N}\int_{\mathbb{X}}\frac{\mu_1(x)}{p^*(x)}(\hat{p}_K(x)-p^*(x))\mathrm{d}F_0(x)-\frac{1}{\sqrt{N}}\sum_{i=1}^N\tilde{\delta}_K(X_i)\frac{T_i-p_K(X_i)}{\sqrt{p_K(X_i)(1-p_K(X_i))}} \tag{56}$$

$$+\frac{1}{\sqrt{N}}\sum_{i=1}^N(\tilde{\delta}_K(X_i)-\delta_K(X_i))\frac{T_i-p_K(X_i)}{\sqrt{p_K(X_i)(1-p_K(X_i))}} \tag{57}$$

$$+\frac{1}{\sqrt{N}}\sum_{i=1}^N\left(\delta_K(X_i)\frac{T_i-p_K(X_i)}{\sqrt{p_K(X_i)(1-p_K(X_i))}}-\delta_0(X_i)\frac{T_i-p^*(X_i)}{\sqrt{p^*(X_i)(1-p^*(X_i))}}\right) \tag{58}$$

$$+\frac{1}{\sqrt{N}}\sum_{i=1}^N\left\{\left(\frac{T_i\cdot Y_i}{p^*(X_i)}-\beta_0\right)+\delta_0(X_i)\frac{T_i-p^*(X_i)}{\sqrt{p^*(X_i)(1-p^*(X_i))}}\right\} \tag{59}$$

In this expression F_0 is the population cdf of X and

$$\tilde{\delta}_K(x)=-\int_{\mathbb{X}}\frac{\mu_1(z)}{p^*(z)}L'(R^K(z)'\tilde{\pi}_K)R^K(z)'\mathrm{d}F_0(z)\tilde{\Sigma}_K^{-1}\sqrt{L'(R^K(x)'\pi_K^*)}R^K(x) \tag{60}$$

$$\delta_K(x) = -\int_{\mathbb{X}} \frac{\mu_1(z)}{p^*(z)} L'(R^K(z)'\pi_K^*) R^K(z)' \mathrm{d}F_0(z) \Sigma_K^{-1} \sqrt{L'(R^K(x)'\pi_K^*)} R^K(x) \tag{61}$$

$$\delta_0(x) = -\frac{\mu_1(x)}{p^*(x)} \sqrt{p^*(X_i)(1-p^*(X_i))} \tag{62}$$

Note that (59) is equal to the linearized expression for $\sqrt{N}(\hat{\beta}_{ew} - \beta_0)$. To show that the estimator is indeed asymptotically linear, we must derive bounds on the terms (54)-(58). If a bound depends on both K and N, we derive the bound for sequences $K(N)$ that go to ∞ with N. Because during the derivation some restrictions on these sequences are imposed, the resulting bounds are not uniform in K. We have seen this type of argument in the derivation of the order of $\|\hat{\pi}_{K(N)} - \pi_{K(N)}\|$ where we imposed the large sample identification condition $\zeta(K(N))^4/N \to 0$.

Below we present the bounds on the terms (54)-(58). Details for the derivations for these bounds are available from the authors. The bound for (54) is

$$\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{T_i Y_i}{\hat{p}_K(X_i)} - \frac{T_i Y_i}{p^*(X_i)} + \frac{T_i Y_i}{p^*(X_i)^2} (\hat{p}_K(X_i) - p^*(X_i)) \right) \right|$$

$$= O_P\left(\frac{\zeta(K(N))^3}{\sqrt{N}} \right) + O_P\left(\sqrt{N} \zeta(K(N))^2 K(N)^{-2\frac{s}{r}} \right) + O_P\left(\zeta(K(N))^{5/2} K(N)^{-\frac{s}{r}} \right)$$

The bound for (55) is

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(-\frac{T_i Y_i}{p^*(X_i)^2} (\hat{p}_K(X_i) - p^*(X_i)) + \int_{\mathbb{X}} \frac{\mu_1(x)}{p^*(x)} (\hat{p}_K(x) - p^*(x)) \mathrm{d}F_0(x) \right)$$

$$= O_P(\zeta(K(N)) K(N)^{-\frac{s}{r}}) + O_P\left(\frac{\zeta(K(N))^2}{\sqrt{N}} \right)$$

The bound for (56) is

$$\left| \sqrt{N} \int_{\mathbb{X}} \frac{\mu_1(x)}{p^*(x)} (p_{K(N)}(x) - p^*(x)) \mathrm{d}F_0(x) \right| < C\sqrt{N} \zeta(K(N)) K(N)^{-\frac{s}{r}}$$

$$= O(\sqrt{N} \zeta(K(N)) K(N)^{-\frac{s}{r}})$$

The bound for (57) is

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\hat{\delta}_{K(N)}(X_i)-\delta_{K(N)}(X_i))\frac{T_i-p_{K(N)}(X_i)}{\sqrt{p_{K(N)}(X_i)(1-p_{K(N)}(X_i))}}\right| = O_P\left(\frac{\zeta(K(N))^{9/2}}{N^{1/2}}\right)$$

The bound for (58) is

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\delta_{K(N)}(X_i)\frac{T_i-p_{K(N)}(X_i)}{\sqrt{p_{K(N)}(X_i)(1-p_{K(N)}(X_i))}}-\delta_0(X_i)\frac{T_i-p^*(X_i)}{\sqrt{p^*(X_i)(1-p^*(X_i))}}\right)\right|$$

$$= O_P\left(\max\left(K(N)^{-\frac{1}{2}\frac{t}{r}},\zeta(K(N))K(N)^{-\frac{s}{r}}\right)\right)$$

From these five expressions we obtain

$$\left|\sqrt{N}(\hat{\beta}_{ew}-\beta_0)-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\left(\frac{T_i.Y_i}{p^*(X_i)}-\beta_0\right)-\frac{\mu_1(X_i)}{p^*(X_i)}(T_i-p^*(X_i))\right\}\right| \tag{63}$$

$$= O_P\left(\frac{\zeta(K(N))^3}{\sqrt{N}}\right)+O_P\left(\sqrt{N}\zeta(K(N))^2K(N)^{-2\frac{s}{r}}\right)+O_P\left(\zeta(K(N))^{5/2}K(N)^{-\frac{s}{r}}\right)$$

$$+O_P(\zeta(K(N))K(N)^{-\frac{s}{r}})+O_P\left(\frac{\zeta(K(N))^2}{\sqrt{N}}\right)+O(\sqrt{N}\zeta(K(N))K(N)^{-\frac{s}{r}})+O_P\left(\frac{\zeta(K(N))^{9/2}}{\sqrt{N}}\right)$$

$$+O_P\left(\max\left(K(N)^{-\frac{1}{2}\frac{t}{r}},\zeta(K(N))K(N)^{-\frac{s}{r}}\right)\right)$$

$$= O_P\left(\sqrt{N}\zeta(K(N))^2K(N)^{-2\frac{s}{r}}\right)+O_P\left(\zeta(K(N))^{5/2}K(N)^{-\frac{s}{r}}\right)+O_P\left(\frac{\zeta(K(N))^{9/2}}{\sqrt{N}}\right)$$

Note that the second term of the final expression is a bias term, the third a variance term, and the first a combination of a variance and bias term.

As noted $\zeta(K)$ depends on the sequence of approximating functions. For power series we have $\zeta(K)=O(K)$. If we consider sequences $K(N)=N^c$ we can find the range of c for which (63) is $o_P(1)$. Substitution in the right-hand side of (63) gives that the first term on

the right hand side requires that $c > \frac{1}{4(s/r-1)}$, the second that $s/r > 5/2$ and the third that $c < 1/9$. These inequalities can be simultaneously satisfied if $s/r \geq 4$. □

Proof of Theorem 2:

Define

$$\Psi_K = -\frac{1}{N}\sum_{i=1}^{N} \frac{Y_i T_i}{p^*(X_i)^2} R^K(X_i) \tag{64}$$

$$\hat{\Psi}_K = -\frac{1}{N}\sum_{i=1}^{N} \frac{Y_i T_i}{\hat{p}_K(X_i)^2} R^K(X_i) \tag{65}$$

$$\hat{\Sigma}_K = \frac{1}{N}\sum_{i=1}^{N} R^K(X_i) R^K(X_i)' \tag{66}$$

Then $\Psi_K' \hat{\Sigma}_K^{-1} R^K(x)$ is the predicted value in a least squares series regression of $-\frac{Y_i T_i}{p^*(X_i)^2}$ on $R^K(X_i)$[15]. This predicted value estimates $-\frac{\mathrm{E}(Y|x)}{p^*(x)}$ which is the conditional expectation (given $X = x$) of the derivative of the moment condition with respect to p^*. The usual bound for series estimators applies

$$\sup_{x\in\mathbb{X}} \left| \Psi_K' \hat{\Sigma}_K^{-1} R^K(x) + \frac{\mu_1(x)}{p^*(x)} \right| \leq C_1 \zeta(K(N)) O_P\left(\sqrt{\frac{\zeta(K(N))}{N}}\right) + C_2 K(N)^{-\frac{s'}{r}} \tag{67}$$

with s' the number of continuous derivatives of $\mu_1(x)$. Also

$$\left|\left| \hat{\Psi}_K - \Psi_K \right|\right| = \left|\left| \frac{1}{N}\sum_{i=1}^{N} \frac{(\hat{p}_K(X_i) - p^*(X_i))(p^*(X_i) + \hat{p}_K(X_i))}{\hat{p}_K(X_i)^2 p^*(X_i)^2} Y_i T_i R^K(X_i) \right|\right| \tag{68}$$

$$\leq \frac{1}{N}\sum_{i=1}^{N} \left| \frac{p^*(X_i) + \hat{p}_K(X_i)}{\hat{p}_K(X_i)^2 p^*(X_i)^2} \right| |\hat{p}_K(X_i) - p^*(X_i)|\, |Y_i| T_i \left|\left| R^K(X_i) \right|\right|$$

As in the proof of Theorem 1 we have that $\hat{p}_K(x)$ is bounded from 0 and 1 on $\mathbb{X}$ if $N \to \infty$ and hence we have the following bound for (68)

$$C \sup_{x\in\mathbb{X}} |\hat{p}_K(x) - p^*(x)| \sup_{x\in\mathbb{X}} \left|\left| R^K(x) \right|\right| \frac{1}{N}\sum_{i=1}^{N} |Y_i| + o_P(1) \tag{69}$$

[15]The number of terms in this series estimator need not be equal to that in the series estimator of the propensity score. The notation can be changed to reflect this.

$$= C_1\zeta(K(N))^2 O_P\left(\sqrt{\frac{\zeta(K(N))}{N}}\right) + C_2\zeta(K(N))^2 K(N)^{-\frac{s}{r}}$$

We use the bounds (67) and (69) to obtain a bound on

$$\hat{\alpha}_K(t,x) - \alpha(t,x) = \left(\hat{\Psi}_K - \Psi_K\right)' \hat{\Sigma}_K^{-1} R^K(x)(t - \hat{p}_K(x)) \tag{70}$$

$$+ \left[\Psi'_K \hat{\Sigma}_K^{-1} R^K(x) + \frac{\mu_1(x)}{p^*(x)}\right](t - \hat{p}_K(x)) + \frac{\mu_1(x)}{p^*(x)}(\hat{p}_K(x) - p^*(x))$$

Under the asymptotic identification condition

$$\sup_{x\in\mathbb{X}} |\hat{\alpha}_K(t,x) - \alpha(t,x)| \leq C_1 \left|\left|\hat{\Psi}_K - \Psi_K\right|\right| \sup_{x\in\mathbb{X}} \left|\left|R^K(x)\right|\right| \tag{71}$$

$$+C_2 \sup_{x\in\mathbb{X}} \left|\Psi'_K \hat{\Sigma}_K^{-1} R^K(x) + \frac{\mu_1(x)}{p^*(x)}\right| + C_3 \sup_{x\in\mathbb{X}} \mu_1(x) \sup_{x\in\mathbb{X}} |\hat{p}_K(x) - p^*(x)|$$

Because $\mathbb{X}$ is compact and $\mu_1(x)$ is continuous, $\sup_{x\in\mathbb{X}} \mu_1(x) < \infty$. Substitution of the bounds (67) and (69), collecting terms of the same order and omitting terms of lower order gives the bound

$$\sup_{x\in\mathbb{X}} |\hat{\alpha}_K(t,x) - \alpha(t,x)| \tag{72}$$

$$\leq C_1\zeta(K(N))^3 O_P\left(\sqrt{\frac{\zeta(K(N))}{N}}\right) + C_2\zeta(K(N))^3 K(N)^{-\frac{s}{r}} + C_3 K(N)^{-\frac{s'}{r}}$$

It can be shown that the difference between (15) and (14) is bounded by (72) (details of these calculations are available from the authors). Under the rates specified in Theorem 1 this bound is $o_p(1)$. Hence (15) is a consistent estimator for (14).

□

Proof of Theorem 4: 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

$$\tau_{wate} \equiv \frac{\int\int g(x) y f_1(y|x) f(x) dy dx - \int\int g(x) y f_0(y|x) f(x) dy dx}{\int g(x) f(x) dx}$$

So for the parametric submodel indexed by θ,

$$\tau_{wate}(\theta) \equiv \frac{\int\int g(x) y f_1(y|x,\theta) f(x,\theta) dy dx - \int\int g(x) y f_0(y|x,\theta) f(x,\theta) dy dx}{\int g(x) f(x,\theta) dx}$$

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

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

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

$$\frac{\partial \tau_{wate}(\theta_0)}{\partial \theta} =$$

$$\frac{1}{\mu_g}\left[\int\int g(x) y s_1(y|x,\theta_0) f_1(y|x,\theta_0) f(x,\theta_0) dy dx - \int\int g(x) y s_0(y|x,\theta_0) f_0(y|x,\theta_0) f(x,\theta_0) dy dx\right]$$

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

The following choice for F_τ satisfies the condition:

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

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

Hence τ_{wate} 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

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

□

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.

Barnow, B., G. Cain and A. Goldberger (1980), "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies*, vol. 5, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage.

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.

Geman, S., and C. Hwang, (1982), "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *Annals of Statistics*, Vol. 10, No. 2, 401-414.

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., H. Ichimura, J. Smith, and P. Todd, (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66, 1017-1098.

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* 81, 1-14.

Hirano, K., G. Imbens, and G. Ridder, (2000), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," NBER Technical Working Paper 251.

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 0238.

Imbens, G., (1997), "One-step Estimators in Overidentified Generalized Method of Moments Estimator," *Review of Economic Studies* 64, 359-383.

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

Kitamura, Y., and M. Stutzer, (1997), "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, Vol. 65, 861-874.

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. and D. 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., S. Mark, and W. Newey, (1992) "Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders," *Biometrics*, 48, 479-495.

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 Statis-*

tical 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. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

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.

Wooldridge, J, (2002), "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification," Institute for Fiscal Studies, cemmap working paper cwp11/02.