Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand*

Richard K. Crump† V. Joseph Hotz‡ Guido W. Imbens§ Oscar A. Mitnik¶

First Draft: July 2004
This Draft: May 2006

Abstract

Estimation of average treatment effects under unconfoundedness or selection on observables is often hampered by lack of overlap in the covariate distributions. This lack of overlap can lead to imprecise estimates and can make commonly used estimators sensitive to the choice of specification. In such cases researchers have often used informal methods for trimming the sample or focused on subpopulations of interest. In this paper we develop formal methods for addressing such lack of overlap in which we sacrifice some external validity in exchange for improved internal validity. We characterize optimal subsamples where the average treatment effect can be estimated most precisely, as well optimally weighted average treatment effects. We show the problem of lack of overlap has important connections to the presence of treatment effect heterogeneity: under the assumption of constant conditional average treatment effects the treatment effect can be estimated much more precisely. The efficient estimator for the treatment effect under the assumption of a constant conditional average treatment effect is shown to be identical to the efficient estimator for the optimally weighted average treatment effect.

JEL Classification: C14, C21, C52

Keywords: *Average Treatment Effects, Causality, Unconfoundedness, Overlap, Treatment Effect Heterogeneity*

*We are grateful for helpful comments by Richard Blundell and Michael Lechner, and by participants in seminars at the ASSA meetings in Philadelphia, University College London, UCLA, UC-Berkeley, MIT, Harvard, the Malinvaud seminar at CREST, and the IAB Empirical Evaluation of Labour Market Programmes conference.

†Department of Economics, University of California at Berkeley, crump@econ.berkeley.edu, http://socrates.berkeley.edu/∼ crump/.

‡Department of Economics, University of California at Los Angeles, hotz@econ.ucla.edu, http://www.econ.ucla.edu/hotz/.

§Department of Agricultural and Resource Economics, and Department of Economics, University of California at Berkeley, 661 Evans Hall, Berkeley, CA 94720-3880, imbens@econ.berkeley.edu, http://elsa.berkeley.edu/users/imbens/.

¶Dept of Economics, University of Miami, omitnik@miami.edu, http://moya.bus.miami.edu/∼ omitnik/.

1 Introduction

There is a large literature on estimating average treatment effects under assumptions of unconfoundedness or ignorability following the seminal work by Rubin (1973, 1978) and Rosenbaum and Rubin (1983a). Researchers have developed estimators based on regression methods (e.g., Hahn, 1998, Heckman, Ichimura and Todd, 1998), matching (e.g., Rosenbaum, 1989, Abadie and Imbens, 2004), and methods based on the propensity score (e.g., Rosenbaum and Rubin, 1983a, Hirano, Imbens and Ridder, 2003). Related methods for missing data problems are discussed in Robins, Rotnitzky and Zhao (1995) and Robins and Rotnitzky (1995).[1] In practice an important concern in implementing all these methods is that one needs sufficient overlap between covariate distributions in the two subpopulations. Even if there exist areas with sufficient overlap, there may be other parts of the covariate space with few units of one of the treatment levels. Such areas of limited overlap can lead to estimators for average treatment effects with poor finite sample properties. In particular, such estimators can have substantial bias, large variances, as well as considerable sensitivity to the exact specification of the regression functions or propensity score. Heckman, Ichimura and Todd (1997) and Dehejia and Wahba (1999) point out the empirical relevance of this overlap issue.[2]

One strand of the literature has focused on assessing the robustness of existing estimators to a variety of potential problems including lack of overlap. See, for example, Rosenbaum and Rubin (1983b), Imbens (2003), and Ichino, Mealli, and Nannicini (2005). A second strand of the literature focuses on developing new estimators that are more robust and precise. With this goal in mind researchers have proposed discarding or downweighting observations with covariates in areas with limited overlap. A number of specific methods have been proposed for implementing this. In simplest setting, with a discrete covariate, Rubin (1977)and Lee (2005b) suggest simply discarding all units with covariate values with either no treated or no control units. Rubin and Cochran (1973) suggest caliper matching where potential matches are dropped if the within-match difference in propensity scores exceeds some threshold level. Dehejia and Wahba (1999) focus on the average treatment effect for the treated and suggest discarding all controls with estimated propensity scores below the smallest value of the propensity score among the treated. Heckman, Ichimura, Smith and Todd (1997) and Heckman, Ichimura and Todd (1998) drop units from the analysis if the estimated density of the covariate distribution conditional on treatment status is below some threshold. Ho, Imai, King and Stuart (2004) propose preprocessing the data by matching units and carrying out parametric inferences using only the matched data.

All of these methods for dealing with limited overlap in the covariates of the two treatment subpopulations have some advantages as well as some drawbacks. For our purposes, we note that all of these methods involve changing the *estimand*, at least in finite samples. While the

[1] See Rosenbaum (2001), Heckman, Lalonde and Smith (1999), Wooldridge (2002), Blundell and Costa-Diaz (2002), Imbens (2004) and Lee (2005a) for surveys of this literature.

[2] Dehejia and Wahba (1999) write: "... our methods succeed for a transparent reason: They only use the subset of the comparison group that is comparable to the treatment group, and discard the complement." Heckman, Ichimura and Todd (1997) write "A major finding of this paper is that comparing the incomparable – i.e., violating the common support condition for the matching variables – is a major sources of evaluation bias as conventionally measured."

resulting estimators do tend to reduce sensitivity of the final estimates to model specification, they rely on arbitrary choices regarding thresholds for discarding observations, i.e., on exactly how the estimand is changed. Furthermore, there are few formal results on their properties.

In this paper, we propose a systematic approach to dealing with subpopulations for which there is limited overlap in the covariates. Our approach has asymptotic optimality properties under some conditions and is straightforward to implement. We consider two specific methods. First, we focus on average treatment effects within a selected subpopulation defined in terms of covariate values. Conditioning on a subpopulation reduces the effective sample size, thus increasing the variance of the estimated average treatment effect. However, if the subpopulation is chosen appropriately, it may be possible to estimate the average treatment within this subpopulation more precisely than the average effect for the entire population despite the smaller sample size. It turns out that in general this tradeoff is well defined and, under some conditions, leads to choosing the observations for the subpopulation that has propensity scores in an interval $[\alpha, 1-\alpha]$, where the optimal cutoff value of α solely determined by the distribution of the propensity score. We refer to this as the Optimal Subpopulation Average Treatment Effect (OSATE).

Second, we consider weighted average treatment effects with the weights depending only on the covariates. The first approach of choosing a subpopulation can be viewed as a special case in this framework where the weight function is restricted to be an indicator function. Without imposing this restriction we characterize the weight function that leads to the most precisely estimated average treatment effect. Note that this class of estimands with weights depending on the covariates includes the average treatment effect for the treated where the weight function is proportional to the propensity score. Under the same conditions as before, the optimal weight function will again be a function of the propensity score alone, proportional to the product of the propensity score and one minus the propensity score. We refer to this as the Optimally Weighted Average Treatment Effect (OWATE).

The switch to average treatment effect for an optimally selected subpopulation or to a optimally weighted average treatment effect has a second benefit beyond the increase in precision. The subpopulations for treated and control group in this selected or weighted population tend to be more balanced in the distribution of the covariates. This is a consequence of the fact that, under homoskedasticity, the variance of the conditional average treatment effect is proportional to $(e(X)\cdot(1-e(X)))^{-1}$. Thus, lowering the weight on high-variance observations increases the weight on observations with propensity scores close to 1/2. The increased balance in the selected or weighted sample reduces the sensitivity of any estimators to changes in the specification. In the extreme case, where the selected sample is completely balanced in covariates in the two treatment arms, one can simply use the average difference in outcomes between treated and control units.

As is the case with some of the methods for estimating treatment effects noted above, the methods we propose change the estimand relative to the one (or ones) of original focus. This is somewhat uncommon in econometric analyses.[3] Typically, the estimand is defined, *a priori*, as

[3]One exception is the local average treatment effect introduced by Imbens and Angrist (1994), which is the average effect of the treatment for the subpopulation of compliers.

is the case with the population average treatment effect, the average effect for the subpopulation of the treated or another *a priori* defined subpopulation of interest. In these cases, estimates are produced that turn out to be more or less precise, depending on the actual data. In cases where even large data sets would not permit point identification of the estimand, regions of the parameter space consistent with the model may be reported in a bounds analysis of the type developed by Manski (1990, 2003).

In this paper, we focus on average effects for a statistically defined (weighted) subpopulation.[4] This change of focus is *not* motivated, *per se*, by an intrinsic interest in the subpopulation for which we ultimately estimate the average causal effect. Rather, it acknowledges and addresses the difficulties in making inferences about the population of primary interest. This approach has several potential justifications. First, it focuses on achieving precise estimates. By changing the sample from one that was potentially representative of the population of interest, we can gain greater internal validity, although, in doing so, we may sacrifice some of the external validity of the resulting estimates.[5] Our proposed approach of placing greater stress on internal versus external validity is similar to that found in the design of randomized experiments which are carried out on populations unrepresentative of the population of interest in order to improve the precision of the inferences to be drawn.[6] More generally, the primacy of internal validity over external validity is advocated in many discussions of causal inference (see, for example, Shadish, Cook, and Campbell, 2002).

Second, our approach is well-suited to situations where the primary interest is to determine whether a treatment may harm or benefit *some* group in a broader population. For example, one may be interested whether there is any evidence that a particular drug could harm or have side effects for some group of patients in a well-defined population. In this context, obtaining greater precision in the estimation of a treatment effect, even if it is not for the entire population, is warranted. We note that the subpopulation for which these estimands are valid are defined in terms of the observed covariate values so that one can determine, for each individual, whether they are in the relevant subpopulation or not.

Third, our approach can provide useful, albeit auxiliary, information when making inferences about the treatment effects for fixed populations. Thus, instead of only reporting the potentially imprecise estimate for the population average treatment effect, one can also report the estimates for the subpopulations where we can make more precise inferences.

In interpreting our results, it also is of interest to consider the estimation of the average treatment effect under the assumption that it does not vary with the covariates.[7] This assumption can be quite informative except in the case where the propensity score is constant. Under the assumption that the treatment effect does not vary with covariates, the model is a special case of the partial linear model studied by Robinson (1988), Stock (1989) and Robins, Mark

[4] This is also true for the method proposed by Heckman, Ichimura and Todd, (1998).

[5] A separate issue is that in practice in many cases even the original sample is not representative of the population of interest. For example, we are often interested in policies that would extend small pilot versions of job training programs to different locations and times.

[6] Even in those settings this can be controversial and lead to misleading conclusions.

[7] The possible presence of heterogeneity of the treatment effect is an important consideration in much of this literature. See for applications Dehejia and Wahba (1999), Lechner (2002) and others.

and Newey (1992).[8] As we discuss below, the efficient estimator for that case turns out to be identical to the efficient estimator in the heterogeneous case for the weighted average treatment effect with the weights chosen to obtain the most precisely estimated average treatment effect. In Crump, Hotz, Imbens and Mitnik (2005), we exploit this fact to develop non-parametric tests of treatment effect heterogeneity with respect to covariates that characterize subpopulations of a population of interest.

Finally, it is important to note that our calculations are not tied to a specific estimator. The results formally refer to differences in the efficiency bound for different subpopulations. As a consequence, they are relevant for all efficient estimators, including the ones proposed by Hahn (1998), Hirano, Imbens and Ridder (2003), Imbens, Newey and Ridder (2004), Robins, Rotnitzky and Zhao (1995). Although not directly applicable to estimators that do not reach the efficiency bound, such as the nearest neighbor matching estimators in Abadie and Imbens (2002) and the local linear estimators in Heckman, Ichimura and Todd (1998), the close relation between those estimators and the efficient ones suggests that with matching the same issues are relevant.

We illustrate these methods, using data from the non-experimental part of a data set on labor market programs previously used by Lalonde (1986), Dehejia and Wahba (1999), Smith and Todd (2005) and others. In this data set, the overlap issue is a well known problem, with the control and treatment group far apart on some of the most important covariates including lagged values for the outcome of interest, yearly earnings. Here the optimal subpopulation method suggests dropping 2363 out of 2675 observations (leaving only 312 observations, or 12% of the original sample) in order to minimize the variance. Calculations suggest that this lowers the variance by a factor 1/160000, reflecting the fact that most of the controls are so different from the treated that it is essentially impossible to estimate the population average treatment effect. More relevant is that the variance for the optimal subsample is only 40% of that for the propensity score weighted sample, which estimates the average effect on the treated. Such potential gains in precision of an estimated treatment effect have gone largely unnoticed because most of the researchers analyzing this data set have focused almost exclusively on the average treatment effect for the treated,

The remainder of the paper is organized as follows. In section 2, we present a simple example in which there is a single and scalar covariate is used in the estimation of the average treatment effect. This example allows us to illustrate how the precision of the estimates varies with changes in the estimand. Section 3 characterizes the general set up we use throughout the paper, section 4 reviews existing results for the efficient estimation of treatment effects and section 5 reviews the previous approaches to dealing with limited overlap when estimating treatment effects. In section 6, we define and characterize the properties of the OSATE and OWATE estimators. In section 7, we present the application to the Lalonde data.

[8] Stock (1989) also focuses on estimating the effect of a policy intervention, but he formulates the problem differently. In his approach the intervention does not change the relationship between the covariates and the outcome. Instead it changes the distribution of the covariates in a known manner. The estimand is then the difference between the average value of the regression function given the induced distribution of the covariates and the average value of the regression function given the current distribution of the covariates.

2 A Simple Example

To set the stage for the issues to be discussed in this paper, consider an example with a scalar covariate X taking on two values, 0 and 1. Let N_x be the sample size for the subsample with $X = x$, and let $N = N_0 + N_1$ be the total sample size. Also let $p = N_1/N$ be the population share of $X = 1$ units. Let the average treatment effect conditional on the covariate be equal to τ_x. The population average treatment effect is then $\tau = p\cdot\tau_1+(1-p)\cdot\tau_0$. Let N_{xw} be the number of observations with covariate $X_i = x$ and treatment indicator $W_i = w$. Also, let $e_x = N_{x1}/N_x$ be the propensity score for $x = 0, 1$. Finally, let $\bar{y}_{xw} = \sum_{i=1}^{N} Y_i \cdot 1\{X_i = x, W_i = w\}/N_{xw}$ be the average within each of the four subpopulations. Assume that the variance of $Y(w)$ given $X_i = x$ is σ^2 for all x.

The natural estimator for the treatment effects for each of the two subpopulations are

$$\hat{\tau}_0 = \bar{y}_{01} - \bar{y}_{00}, \qquad \text{and} \qquad \hat{\tau}_1 = \bar{y}_{11} - \bar{y}_{10},$$

with variances

$$V(\hat{\tau}_0) = \sigma^2 \cdot \left(\frac{1}{N_{00}} + \frac{1}{N_{01}}\right) = \frac{\sigma^2}{N\cdot(1-p)} \cdot \frac{1}{e_0\cdot(1-e_0)},$$

and

$$V(\hat{\tau}_1) = \sigma^2 \cdot \left(\frac{1}{N_{10}} + \frac{1}{N_{11}}\right) = \frac{\sigma^2}{N\cdot p} \cdot \frac{1}{e_1\cdot(1-e_1)}.$$

The estimator for the population average treatment effect is

$$\hat{\tau} = p\cdot\hat{\tau}_1 + (1-p)\cdot\hat{\tau}_0.$$

Because the two estimates $\hat{\tau}_0$ and $\hat{\tau}_1$ are independent, the variance of the population average treatment effect is

$$\begin{aligned} V(\hat{\tau}) &= p^2\cdot V(\hat{\tau}_1) + (1-p)^2\cdot V(\hat{\tau}_0) \\ &= \frac{\sigma^2}{N}\cdot\left(\frac{p}{e_1\cdot(1-e_1)} + \frac{1-p}{e_0\cdot(1-e_0)}\right) = \frac{\sigma^2}{N}\cdot\mathbb{E}\left[\frac{1}{e_X\cdot(1-e_X)}\right]. \end{aligned}$$

The first point of the paper concerns the comparison of $V(\hat{\tau})$, $V(\hat{\tau}_0)$, and $V(\hat{\tau}_1)$). Define $V_{\min} = \min(V(\hat{\tau}), V(\hat{\tau}_0), V(\hat{\tau}_1)$. Then

$$V_{\min} = \begin{cases} V(\hat{\tau}_0) & \text{if} \quad & & (e_1(1-e_1))/(e_0(1-e_0)) & \leq (1-p)/(2-p), \\ V(\hat{\tau}) & \text{if} \quad (1-p)/(2-p) \leq & & (e_1(1-e_1))/(e_0(1-e_0)) & \leq (1+p)/p, \\ V(\hat{\tau}_1) & \text{if} \quad & (1+p)/p \leq & (e_1(1-e_1))/(e_0(1-e_0)). & \end{cases} \tag{2.1}$$

Which estimator has the smallest variance depends on the relative sizes of the two subsamples, p, and the ratio of the product of the propensity score and one minus the propensity score, $e_1(1-e_1)/(e_0(1-e_0))$. If the propensity score for units with $X = 0$ is close to zero or one, we

cannot estimate the average treatment effect for this subpopulation precisely. In that case, the ratio $e_1(1-e_1)/(e_0(1-e_0))$ will be high and we may be able to estimate the average treatment effect for the $X = x_1$ subpopulation more accurately than for the population as a whole, even though we may lose a substantial number of observations by discarding units with $X_i = 0$. Similarly, if the propensity score for the $X = 1$ subpopulation is close to zero or one, the ratio $e_1(1-e_1)/(e_0(1-e_0))$ is close to zero, and we may be able to estimate the average treatment effect for the $X = x_0$ subpopulation more accurately than for the population as a whole. If the ratio is close to one, we can estimate the average effect for the population as a whole more accurately than for either of the two subpopulations.

The second advantage of focusing on subpopulation average treatment effects is in this case obvious. Within the two subpopulations we can estimate the within-subpopulation average treatment effect without bias by simply differencing average treatment and control outcomes. Thus our results are not sensitive to the choice of estimator, whereas in the population as a whole there is potentially substantial bias from simply differencing average outcomes.

The second point is that one need not limit the choice to the three average treatment effects discussed so far. More generally, one may wish to focus on a weighted average treatment effect

$$\tau_\lambda = \lambda \cdot \tau_1 + (1-\lambda) \cdot \tau_0,$$

for fixed λ, which can be estimated as

$$\hat{\tau}_\lambda = \lambda \cdot \hat{\tau}_1 + (1-\lambda) \cdot \hat{\tau}_0,$$

The variance for this weighted average treatment effect is

$$\begin{aligned} V(\hat{\tau}_\lambda) &= \lambda^2 \cdot V(\hat{\tau}_1) + (1-\lambda)^2 \cdot V(\hat{\tau}_0) \\ &= \lambda^2 \cdot \frac{\sigma^2}{N \cdot p} \cdot \frac{1}{e_1 \cdot (1-e_1)} + (1-\lambda)^2 \cdot \frac{\sigma^2}{N \cdot (1-p)} \cdot \frac{1}{e_0 \cdot (1-e_0)}. \end{aligned}$$

The variance is minimized at

$$\lambda^* = \frac{1/V(\hat{\tau}_1)}{1/V(\hat{\tau}_1) + 1/V(\hat{\tau}_0)} = \frac{p \cdot e_1 \cdot (1-e_1)}{(1-p) \cdot e_0 \cdot (1-e_0) + p \cdot e_1 \cdot (1-e_1)}. \tag{2.2}$$

with the minimum value for the variance equal to

$$V(\tau_{\lambda^*}) = \frac{\sigma^2}{N} \cdot \frac{1}{((1-p) \cdot e_0 \cdot (1-e_0) + p \cdot e_1 \cdot (1-e_1))} = \frac{\sigma^2}{N} \cdot \frac{1}{\mathbb{E}[e_X \cdot (1-e_X)]}.$$

The ratio of the variance for the population average to the variance for the optimally weighted average treatment effect is

$$\begin{aligned} V(\tau_P)/V(\tau_{\lambda^*}) &= \mathbb{E}\left[\frac{1}{e_X \cdot (1-e_X)}\right] \Big/ \frac{1}{\mathbb{E}[e_X \cdot (1-e_X)]} \qquad (2.3) \\ &= \mathbb{E}\left[\frac{1}{V(e_X)}\right] \Big/ \frac{1}{\mathbb{E}[V(e_X)]}. \end{aligned}$$

By Jensen's inequality this is greater than one if $V(e_X) > 0$, that is, if the propensity score varies across the population.

In summary, suppose in this case one is interested in the population average treatment effect τ. One may find that the efficient estimator is imprecise. This is consistent with two different states of the world. In one state the average effect for both of the subpopulations are also imprecisely estimated, and in effect one cannot say much about the effect of the treatment at all. In the other state of the world it is still possible to learn something about the effect of the treatment because one of the subpopulation average treatment effects can be estimated precisely. In that case – which corresponds to the propensity score for one of the two subpopulations being close to zero or one – one also may wish to report the estimator for the precisely estimable average treatment effect to convey the information that the data contain about the effect of the treatment. It is important to stress that the message of the paper is not that one should report $\hat{\tau}_m$ or $\hat{\tau}_f$ instead of $\hat{\tau}$. Rather, in cases where $\hat{\tau}_m$ or $\hat{\tau}_f$ are precisely estimable and $\hat{\tau}$ is not, one should report both.

Below, we generalize this analysis to the case with a vector of potentially continuously distributed covariates. We study the existence and characterization of a partition of the covariates space into two subsets. For one of the subpopulations, the average treatment effect is at least as accurately estimable as that for any other subset of the covariate space. This leads to a generalization of (2.1). Under some assumptions, this problem has a well-defined solution and these subpopulations have a very simple characterization, namely a set of values of the covariates for which the propensity score is in the closed interval $[\alpha, 1-\alpha]$. The optimal value of the boundary point α is determined by the distribution of the propensity score and its calculation is straightforward. In addition, we characterize the optimally weighted average treatment effect and its variance, which generalizes (2.2) and (2.3).

3 Set Up

The basic framework is standard in this literature (e.g., Rosenbaum and Rubin, 1983; Hahn, 1998; Heckman, Ichimura and Todd, 1998; Hirano, Imbens and Ridder, 2003). We have a random sample of size N from a large population. For each unit i in the sample, let W_i indicate whether the treatment of interest was received, with $W_i = 1$ if unit i receives the treatment of interest, and $W_i = 0$ 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. We observe W_i and Y_i, where

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

In addition, we observe a vector of pre-treatment variables, or covariates, denoted by X_i. Define the two conditional means, $\mu_w(x) = \mathbb{E}[Y(w)|X = x]$, the two conditional variances, $\sigma_w^2(x) = \text{Var}(Y(w)|X = x)$, the conditional average treatment effect $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x] = \mu_1(x) - \mu_0(x)$, and the propensity score, the probability of selection into the $e(x) = \Pr(W = 1|X = x) = \mathbb{E}[W|X = x]$.

Initially we focus on two average treatment effects. The first is the (super-)population average treatment effect

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

We also consider the conditional average treatment effect:

$$\tau_C = \frac{1}{N}\sum_{i=1}^{N} \tau(X_i),$$

where we condition on the observed set of covariates. The reason for focusing on the second one is twofold. First, it is analogous to the common conditioning on covariates in regression analysis. Second, it can be estimated more precisely if there is indeed variation in the treatment effect by covariates.

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:

Assumption 3.1 (Unconfoundedness)

$$W \perp (Y(0), Y(1)) \;\Big|\; X. \tag{3.4}$$

In addition we assume there is overlap in the covariate distributions:

Assumption 3.2 (Overlap)
For some $c > 0$,

$$c \leq e(x) \leq 1 - c.$$

For estimation, we often need smoothness conditions on the two regression functions $\mu_w(x)$ and the propensity score $e(x)$.

4 Efficiency Bounds

Next, we review some results for efficient estimation of treatment effects. First we discuss efficient estimators previously developed by Hahn (1998) and Hirano, Imbens and Ridder (2003) for treatment effects allowing for heterogeneity in the treatment effects. Second, we present some results for efficient estimation of treatment effects under a variety of assumptions that restrict the heterogeneity of the treatment effects. This setting is closely related to the partial linear model developed by Robinson (1988).

Hahn (1998) calculates the efficiency bound for τ_P.

Theorem 4.1 (HAHN, 1998) *Suppose Assumptions 3.1 and 3.2 hold. Then the semiparametric efficiency bounds for τ is*

$$V_P^{\text{eff}} = \mathbb{E}\left[(\tau(X)-\tau)^2 + \frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right]. \tag{4.5}$$

Proof: See Hahn (1998).

Robins, Rotznitzky and Zhao (1995) present a similar result in a missing data setting.

Hahn (1998) also proposes an estimator that achieves the efficiency bound.[9] Hahn's estimator is asymptotically linear,

$$\hat{\tau}_H = \frac{1}{N}\sum_{i=1}^{N} \psi(Y_i, W_i, X_i) + o_p\left(N^{-1/2}\right),$$

where

$$\psi(y,w,x) = w\cdot\frac{y-\mu_1(x)}{e(x)} - (1-w)\cdot\frac{y-\mu_0(x)}{1-e(x)} + \mu_1(x) - \mu_0(x) - \tau.$$

One implication of this representation is that we can view Hahn's estimator – as well as the other efficient estimators – not only as an estimator of the population average treatment effect, τ_P, but also as an estimator of the conditional average treatment effect τ_C. As an estimator of τ_C, the efficient estimator $\hat{\tau}_H$ has asymptotic variance

$$\mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right]. \tag{4.6}$$

Furthermore, these estimators are efficient for τ_C:

Theorem 4.2 (EFFICIENCY BOUND FOR τ_C) *Suppose Assumptions 3.1 and 3.2 hold. Then the semiparametric efficiency bounds for τ_C is*

$$V_C^{\text{eff}} = \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right]. \tag{4.7}$$

Proof: See Appendix.

Next we consider a larger set of estimands. Instead of looking at the average treatment effect within a subpopulation we consider weighted average treatment effects of the form

$$\tau_{P,g} = \mathbb{E}[\tau(X)\cdot g(X)]/\mathbb{E}[g(X)],$$

for nonnegative functions $g(\cdot)$. For estimands of this type the efficiency bound is given in the following theorem:

[9] Other efficient estimators have been proposed by Hirano, Imbens and Ridder (2003) and Imbens, Newey and Ridder (2004).

Theorem 4.3 (Hirano, Imbens and Ridder, 2003) *Suppose Assumptions 3.1 and 3.2 hold, and suppose that $g(\cdot)$ is known. Then the semiparametric efficiency bounds for τ_g is*

$$V_{P,g}^{\text{eff}} = \frac{1}{\mathbb{E}[g(X)]^2} \cdot \mathbb{E}\left[g(X)^2 \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} + (\tau(X)-\tau_g)^2\right)\right]$$

Proof: See Hirano, Imbens and Ridder (2003).

Again there is an asymptotically linear estimator that achieves this efficiency bound. The same argument as above therefore establishes that the efficient estimator for $\tau_{P,g}$, as an estimator for the conditional average treatment effect version of this estimand,

$$\tau_{C,g} = \sum_{i=1}^{N} \tau(X_i) \cdot g(X_i) \Big/ \sum_{i=1}^{N} g(X_i),$$

has asymptotic variance

$$V_{C,g}^{\text{eff}} = \frac{1}{\mathbb{E}[g(X)]^2} \cdot \mathbb{E}\left[\frac{g(X)^2}{e(X)}\sigma_1^2(X) + \frac{g(X)^2}{1-e(X)}\sigma_0^2(X)\right]. \tag{4.8}$$

Next we consider the case where the weights depend on the propensity score: $g(x) = h(e(x))$. This will be useful later when some estimands of interest have this form. If the propensity score is known this is a special case of the previous result, but if the propensity score is unknown the efficiency bound changes.

Theorem 4.4 (Weighted Average Treatment Effects with Weights Depending on the Propensity Score) *Suppose Assumptions 3.1 and 3.2 hold, and suppose that the weights are a function of the propensity score: $g(x) = h(e(x))$ with $h(\cdot)$ known and $e(x)$ unknown. Then the semiparametric efficiency bounds for τ_g is*

$$V_{P,g}^{\text{eff}} = \frac{1}{\mathbb{E}[g(X)]^2} \cdot \mathbb{E}\left[g(X)^2 \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} + (\tau(X)-\tau_g)^2\right)\right]$$
$$+\frac{1}{\mathbb{E}[g(X)]^2} \cdot \mathbb{E}\left[e(X)(1-e(X)) \cdot [h'(e(X))]^2(\tau(X)-\tau_g)^2\right],$$

where $h'(a)$ is the first derivative of $h(a)$.

Proof: See Appendix.

A special case of this arises when $h(a) = a$ so that the weights are proportional to the propensity score. In that case, the estimand is the average effect for the treated. The efficiency bound for this case under the unknown propensity score was previously derived by Hahn (1998, Theorem 1). It is equal to

$$V_{P,t}^{\text{eff}} = \frac{1}{\mathbb{E}[e(X)]^2} \cdot \mathbb{E}\left[e(X)^2 \cdot \left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right) + e(X) \cdot (\tau(X)-\tau_t)^2\right].$$

Finally, we consider the case where we know that the average treatment effect does not vary by covariates.

Assumption 4.1 (Constant Conditional Average Treatment Effect)
For all x, $\mu_1(x) - \mu_0(x) = \tau$.

This assumption is slightly weaker than assuming a constant treatment effect. Under this assumption the efficiency bound is a generalization of the bound given in Robins, Mark and Newey (1992) to the heteroskedastic case:

Theorem 4.5 (Robins, Mark and Newey, 1992) *Suppose Assumptions 3.1, 3.2, and 4.1 hold. Then the semiparametric efficiency bounds for τ is*

$$V_{\text{cons}}^{\text{eff}} = \left(\mathbb{E}\left[\left(\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} \right)^{-1} \right] \right)^{-1}. \tag{4.9}$$

Proof: See Robins, Mark and Newey (1992).

It is interesting to compare the efficiency bound for τ under the constant average treatment effect assumption given in (4.9) with the efficiency bound for the average conditional treatment effect τ_C given in (4.6). By Jensen's inequality the former is smaller, unless $\sigma_1^2(x)/e(x) + \sigma_0^2(x)/(1-e(x))$ is constant. Under homoskedasticity the ratio of the variances V_C^{eff} and $V_{\text{cons}}^{\text{eff}}$ reduces to

$$\mathbb{E}\left[\frac{1}{V(W|X)} \right] \Big/ \frac{1}{\mathbb{E}[V(W|X)]},$$

the same expression we obtained in the binary covariate case. This ratio is greater than one unless the propensity score is constant. If the propensity score takes on values close to zero or one this ratio can be large. The implication is that knowledge of the treatment effect being constant as a function of the covariates can be very valuable.

5 Previous Approaches to Dealing with Limited Overlap

In empirical application, there is often concern about the overlap assumption (e.g., Dehejia and Wahba, 1999; Heckman, Ichimura, and Todd, 1997). To ensure that there is sufficient overlap researchers have sometimes trimmed their sample by excluding observations with propensity scores close to zero or one. Cochran and Rubin (1977) suggest caliper matching where units whose match quality is too low according to the distance in terms of the propensity score are left unmatched.

Dehejia and Wahba (1999) focus on the average effect for the treated, They suggest dropping all control units with an estimated propensity score lower than the smallest value, or larger than the largest value, for the estimated propensity score among the treated units. Formally, they first estimate the propensity score. Let the estimated propensity score for unit i be $\hat{e}(X_i)$. Then let $\bar{e}_1$ be the minimum of the $\hat{e}(X_i)$ among treated units and let $\overline{e}_1$ be the maximum of the $\hat{e}(X_i)$ among treated units. DW then drop all control units such that $\hat{e}(X_i) < \bar{e}_1$ or $\hat{e}(X_i) > \overline{e}_1$.

Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998) also focus on the average effect for the treated. They propose discarding units with covariate values at which the estimated density is below some threshold. The precise method is as

follows.[10] First they estimate the propensity score $\hat{e}(x)$. Next, they estimate the density of the estimated propensity score in both treatment arms. Let $\hat{f}_w(e)$ denote the estimated density of the estimated propensity score. The specific estimator they use is a kernel estimator

$$\hat{f}_w(e) = \frac{1}{N_w \cdot h} \sum_{i|W_i=w} K\left(\frac{\hat{e}(X_i) - e}{h}\right),$$

with bandwidth h.[11] First HIT discard observations with $\hat{f}_0(\hat{e}(X_i))$ or $\hat{f}_1(\hat{e}(X_i))$ exactly equal to zero leaving J observations. Observations with the estimated denstiy equal to zero may exist when the kernel has finite support. Smith and Todd, for example, use a quadratic kernel with $K(u) = (u^2 - 1)^2$ for $|u| \leq 1$ and zero elsewhere. Next, they fix a quantile q (Smith and Todd use $q = 0.02$). Among the J observations with positive densities they rank the $2J$ values of $\hat{f}_0(\hat{e}(X_i))$ and $\hat{f}_1(\hat{e}(X_i))$. They then drop units i with $\hat{f}_0(\hat{e}(X_i))$ or $\hat{f}_1(\hat{e}(X_i))$ less than or equal to c_q, where c_q is the largest real number such that

$$\frac{1}{2J} \sum_{i=1}^{J} \left(1\{\hat{f}_0(\hat{e}(X_i)) < c_q\} + 1\{\hat{f}_1(\hat{e}(X_i)) < c_q\}\right) \leq q.$$

Ho, Imai, King and Stuart (2004) propose combining any specific parametric procedure that the researcher may wish to employ with a nonparametric first stage in which the units are matched to the closest unit of the opposite treatment. This typically leads to a data set that is much more balanced in terms of covariate distributions between treated and control. It therefore thus reduces sensitivity of the parametric model to specific modeling decisions such as the inclusion of covariates or functional form assumptions.

King et al (2005): convex hull.

All these methods tend to make the estimators more robust to specification decisions. However, few formal results are available on the properties of these procedures.

6 Alternative Estimands

6.1 The Optimal Subpopulation Average Treatment Effect

First we consider trimming the sample by excluding units with covariates outside of a set $\mathcal{A}$, where $\mathcal{A} \subset \mathbb{X}$, with $\mathbb{X} \subset \mathbb{R}^k$ the covariate space. For a given set $\mathcal{A}$ we define a corresponding average treatment effect $\tau_C(\mathcal{A})$:

$$\tau_C(\mathcal{A}) = \int_{\mathcal{A}} \tau(x) f(x) dx.$$

The efficiency bound for this parameter is

$$V_C^{\text{eff}}(\mathcal{A}) = \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} \middle| X \in \mathcal{A}\right].$$

[10] See Heckman, Ichimura and Todd (1997) and Smith and Todd (2005) for details, and Ham, Li and Reagan (2005) for an application of this method.

[11] In their application Smith and Todd (2005) use Silverman's rule of thumb to choose the bandwidth.

Because the relative size of the subpopulation in $\mathcal{A}$ is $q(\mathcal{A}) = \Pr(X \in \mathcal{A})$, the efficiency bound normalized by the original sample size is

$$V_C^{\text{eff}\prime}(\mathcal{A}) = \frac{1}{q(\mathcal{A})} \cdot \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} \middle| X \in \mathcal{A} \right]. \tag{6.10}$$

We look for an optimal $\mathcal{A}$, denoted by $\mathcal{A}^*$, that minimizes the asymptotic variance (6.10) among all subsets $\mathcal{A}$.

There are two competing effects. First, by excluding units with covariate values outside the set $\mathcal{A}$ one reduces the effective sample size from N to $N \cdot q(\mathcal{A})$. This will increase the asymptotic variance, normalized by the original sample size, by a factor $1/q(\mathcal{A})$. Second, by discarding units with high values for $\sigma_1^2(X)/e(X) + \sigma_0^2(X)/(1-e(X))$ (that is, units with covariate values such that it is difficult to estimate the average treatment effect) one can lower the conditional expectation $\mathbb{E}[\sigma_1^2(X)/e(X) + \sigma_0^2(X)/(1-e(X))|X \in \mathcal{A}]$. Optimally choosing $\mathcal{A}$ involves balancing these two effects. The following theorem gives the formal result for the optimal $\mathcal{A}^*$ that minimizes the asymptotic variance.

Theorem 6.1 (OSATE)
Let $\underline{f} \leq f(x) \leq \overline{f}$, and $\sigma^2(x) \leq \overline{\sigma^2}$ for $w = 0, 1$ and all $x \in \mathbb{X}$. We consider sets $\mathcal{A} \subset \mathbb{X}$ that are elements of the sigma algebra of Borel subsets of $\mathbb{R}^k$. Then the Optimal Subpopulation Average Treatment Effect (OSATE) is $\tau_C(\mathcal{A}^*)$, where, if

$$\sup_{x \in \mathbb{X}} \frac{\sigma_1^2(x) \cdot (1-e(x)) + \sigma_0^2(x) \cdot e(x)}{e(x) \cdot (1-e(x))} \leq 2 \cdot \mathbb{E}\left[\frac{\sigma_1^2(X) \cdot (1-e(X)) + \sigma_0^2(X) \cdot e(X)}{e(X) \cdot (1-e(X))} \right],$$

then $\mathcal{A}^* = \mathbb{X}$ and otherwise,

$$\mathcal{A}^* = \left\{ x \in \mathbb{X} \middle| \frac{\sigma_1^2(x) \cdot (1-e(x)) + \sigma_0^2(x) \cdot e(x)}{e(x) \cdot (1-e(x))} \leq a \right\},$$

where a is a positive solution to

$$a = 2 \cdot \mathbb{E}\left[\frac{\sigma_1^2(X) \cdot (1-e(X)) + \sigma_0^2(X) \cdot e(X)}{e(X) \cdot (1-e(X))} \middle| \frac{\sigma_1^2(X) \cdot (1-e(X)) + \sigma_0^2(X) \cdot e(X)}{e(X) \cdot (1-e(X))} < a \right].$$

Proof: See Appendix.

The result in this theorem simplifies under homoskedasticity.

Corollary 6.1 Optimal Overlap Under Homoskedasticity *Suppose that $\sigma_w^2(x) = \sigma^2$ for all $w \in \{0, 1\}$ and $x \in \mathbb{X}$. If*

$$\sup_{x \in \mathbb{X}} \frac{1}{e(x) \cdot (1-e(x))} \leq 2 \cdot \mathbb{E}\left[\frac{1}{e(X) \cdot (1-e(X))} \right],$$

then $\mathcal{A}^* = \mathbb{X}$. Otherwise,

$$\mathcal{A}^* = \left\{ x \in \mathbb{X} \middle| \frac{1}{e(x) \cdot (1-e(x))} \leq a \right\},$$

where a is a solution to

$$a = 2 \cdot \mathbb{E}\left[\frac{1}{e(X) \cdot (1-e(X))} \middle| \frac{1}{e(X) \cdot (1-e(X))} < a \right].$$

We can find the smallest value of a that satisfies the first order conditions – and which therefore must correspond to a local minimum for $g(a)$ – by iteratively solving equation (**??**). Start with $\alpha_0 = 0$. Calculate

$$\gamma_k = \gamma(\alpha_k) = \mathbb{E}[(e \cdot (1-e))^{-1} | \alpha_k \leq e \leq 1 - \alpha_k].$$

Note that $\gamma_k > 4$. Then solve α_k by solving for the solution in $(0, 1/2)$ of

$$\frac{1}{\alpha_{k+1} \cdot (1 - \alpha_{k+1})} = 2 \cdot \gamma_k,$$

leading to

$$\alpha_{k+1} = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{1}{2 \cdot \gamma_k}}.$$

In an application we would typically not know the propensity score. In that case, we would carry out the calculations with the conditional expectation $\mathbb{E}[(e \cdot (1-e))^{-1} | \alpha \leq e \leq 1 - \alpha]$ replaced by

$$\sum_{i=1}^{N} \frac{1}{e(X_i) \cdot (1 - e(X_i))} \cdot 1\{\alpha \leq e(X_i) \leq 1 - \alpha\} \Big/ \sum_{i=1}^{N} 1\{\alpha \leq e(X_i) \leq 1 - \alpha\}.$$

Estimating the optimal value of α is not particularly difficult. However, it is also unlikely that the variance is very sensitive to the exact cutoff point. It may therefore be sufficient to approximate the optimal α. If the marginal distribution of the propensity score is uniform on the unit interval we can numerically calculate the exact value for the optimal α. This turns out to be 0.1018, suggesting that using the interval $[0.1018, 0.8982]$ would give a asymptotic variance close to optimal.

6.2 The Optimally Weighted Average Treatment Effect

In this section, we consider weighted average treatment effects of the form

$$\tau_g = \int_x g(x) \cdot \tau(x) dF(x) \Big/ \int_x g(x) dF(x).$$

The following theorem gives the most precisely estimable weighte average treatment effect.

Theorem 6.2 (OWATE)
Suppose Assumptions – hold. Then the Optimal Weighted Average Treatment Effect (OWATE) is τ_{g^*}, where

$$g^*(x) = \left(\frac{\sigma_1^2(x)}{e(x)} + \frac{\sigma_0^2(x)}{1 - e(x)} \right)^{-1},$$

Proof: See Appendix.

Corollary 6.2 *Suppose Assumptions – hold, and that $\sigma_0^2(x) = \sigma_1^2(x) = \sigma^2$ for all x. Then the Optimally Weighted Average Treatment Effect (OWATE) is τ_{g^*}, where*

$$g^*(x) = e(x) \cdot (1 - e(x)).$$

7 Some Illustrations Based on Real Data

In this section, we apply the methods developed in this paper to data from a set of manpower training programs. We first calculate the optimal cutoff point α based on an estimate of the propensity score. We report the number of observations discarded by the proposed sample selection. We also report the asymptotic variance for five alternative estimands relative to that for the average treatment effect for the full sample. These five estimands include the average effect for the controls, the average effect for the treated, the OSATE, the OWATE, and the average effect for those with a propensity score between 0.1 and 0.9. The latter is assess the sensitivity to the choice of cutoff values for the propensity score.

7.1 The Lalonde Data

The data set we use was originally analyzed by Lalonde (1986) and subsequently by Dehejia and Wahba (1999) and Smith and Todd (2004). In particular, the particular samples we yse are the ones used in Dehejia and Wahba. The treatment of interest is a job training program. The data for trainees are drawn from an experimental evaluation of this program. Rather than using the randomly assigned control group from this evaluation, we analyze data for a (non-randomly assigned) control group taken from the Panel Study of Income Dynamics (PSID). These control and treatment groups are very unbalanced.

Table 1 presents some summary statistics for these data. The fourth and fifth column present the averages for each of the covariates separately for the comparison and treatment groups. Consider, for example, the average earnings of sample members in the year prior to the program, earn '75. For the control group from the PSID this, the mean of this variable is 19.06, in thousands of dollars. For the treatment group, it is only 1.53. Given that the standard deviation is 13.88, this is a very large difference of 1.26 standard deviations, suggesting that simple covariance adjustments are unlikely to lead to credible inferences.

Using these two samples, we estimate the propensity score of program participation using a logistic model with all nine covariates entering linearly. We then use the estimated propensity score to calculate the optimal cutoff point, α in the homoskedasticity case. The optimal cutoff point is $\alpha = 0.0660$. Based on this cutoff point, the number of observations that should be discarded according is substantial. Out of the original 2675 observations (2490 controls and 185 treated), only 312 are left (183 controls and 129 treated). In Table 3, we present the number of observations in the various categories.

Table 2 presents the asymptotic standard errors for the four estimands. The first is the standard error for the population average treatment effect (ATE). The second is the asymptotic standard error for the average treatment effect for the treated (ATT). The third is the asymptotic standard error for the average treatment effect in the subpopulation with $\alpha < e(x) < 1-\alpha$, for the optimal value of $\alpha = 0.0660$ (OSATE), while the fourth is the standard error for the optimally weighted average treatment effect (OWATE).

The second row in Table 2 shows the ratios of the asymptotic standard error to the asymptotic standard error for the population average treatment effect. There is a huge gain in precision in moving from the population average treatment effect to any of the three other estimands.

This gain is due to the huge differences between the treated and control covariate distributions. As a result of these differences, there are large areas in the covariate space where there are essentially no treated units. Hence, estimating the average treatment effects in those regions of the covariate space is difficult, and can only be done with great imprecision, even under the assumptions made. This finding has been noted by others (see Dehejia and Wahba, 19999) for this particular data set. What the previous investigations based on this data set have not noticed, however, is the fact that there is still a large difference in asymptotic standard errors between the three other estimands. The asymptotic standard error for the average effect for the treated is much larger than for the OSATE estimator (2.58 versus 1.62) which, in turn is substantially larger than the standard error for the OWATE estimator (1.62 versuss 1.28). Overall, this particular example provides a nice illustration of the potential benefits, at least with respect to precision, of optimally choosing one's estimand when estimating treatment effects.

8 Conclusion

Appendix

Proof of Theorem 4.2: Suppose we have an estimator $\tilde{\tau}$ for τ_C that asymptotically linear with influence function $\phi(Y, W, X)$, so that

$$\tilde{\tau} - \frac{1}{N}\sum_{i=1}^{N}\tau(X_i) = \frac{1}{N}\sum_{i=1}^{N}\phi(Y_i, W_i, X_i) + o_p\left(N^{-1/2}\right).$$

The Hahn estimator $\hat{\tau}_H$ satisfies

$$\hat{\tau} - \frac{1}{N}\sum_{i=1}^{N}\tau(X_i) = \frac{1}{N}\sum_{i=1}^{N}\psi(Y_i, W_i, X_i) + o_p\left(N^{-1/2}\right),$$

where

$$\psi(y, w, w) = \phi(y, w, x) - \tau(x) = w \cdot \frac{y - \mu_1(x)}{e(x)} - (1 - w) \cdot \frac{y - \mu_0(x)}{1 - e(x)}.$$

Note that $\mathbb{E}[\psi(Y, W, X)|X] = 0$ so that $\mathbb{E}[\psi(Y, W, X) \cdot (\tau(X) - \tau_P)] = 0$. For $\tilde{\tau}$ to be unbiased asymptotically it must be that $\mathbb{E}[\phi(Y, W, X)|X = x] = 0$ for all x, again implying that $\mathbb{E}[\phi(Y, W, X) \cdot (\tau(X) - \tau_P)] = 0$.
For $\tilde{\tau}$ to be more efficient than $\hat{\tau}$ as an estimator for τ_C it must be that $\mathbb{E}[\phi(Y, W, X)^2] < \mathbb{E}[\psi(Y, W, X)^2]$. Because $\hat{\tau}_H$ is efficient for τ_P, it must be that $\mathbb{E}[(\phi(Y, W, X) + \tau(X) - \tau_P)^2] \geq \mathbb{E}[(\psi(Y, W, X) + \tau(X) - \tau_P)^2]$. Because $\mathbb{E}[\phi(Y, W, X) \cdot (\tau(X) - \tau_P)] = 0$ and $\mathbb{E}[\psi(Y, W, X) \cdot (\tau(X) - \tau_P)] = 0$ this cannot both be true. □

Proof of Theorem 4.4: The derivation of the efficiency bound follows that of Hahn (1998). The density of $(Y(0), Y(1), T, X)$ with respect to some σ-finite measure is

$$\begin{aligned} q(y(0), y(1), t, x) &= f(y(0), y(1)|t, x)f(t|x)f(x) \\ &= f(y(0), y(1)|x)f(t|x)f(x) \\ &= f(y(0), y(1)|x)p(x)^t(1 - p(x))^{1-t}f(x) \end{aligned}$$

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

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

where $f_t(y|x) = f_{Y(T)|X}(y(t)|x) = \int f(y(1 - t), y|x)dy(1 - t)$. Consider a regular parametric submodel indexed by θ, with density

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

which is equal to the true density $q(y, t, x)$ for $\theta = \theta_0$. The score is given by

$$\frac{d}{d\theta}\ln q(y, t, x|\theta) = s(y, t, x|\theta) = t \cdot s_1(y|x, \theta) + (1 - t) \cdot s_0(y|x, \theta) + s_x(x, \theta) + \frac{t - p(x)}{p(x)(1 - p(x))} \cdot p'(x, \theta)$$

where

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

The tangent space of the model is the set of functions

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

where $a(x)$ is any square-integrable measurable function of x and s_1, s_0, and s_x satisfy

$$\int s_1(y,x) f_1(y|x) dy = \mathbb{E}[s_1(Y(1), X)|X = x] = 0, \quad \forall x,$$
$$\int s_0(y,x) f_0(y|x) dy = \mathbb{E}[s_0(Y(0), X)|X = x] = 0, \quad \forall x,$$
$$\int s_x(x) f(x) dx = \mathbb{E}[s_x(X)] = 0.$$

The parameter of interest is

$$\tau_h = \frac{\iint h(p(x)) y f_1(y|x) f(x) dy dx - \iint h(p(x)) y f_0(y|x) f(x) dy dx}{\int h(p(x)) f(x) dx}.$$

Thus, for the parametric submodel indexed by θ, the parameter of interest is

$$\tau_h(\theta) = \frac{\iint h(p(x,\theta)) y f_1(y|x,\theta) f(x,\theta) dy dx - \iint h(p(x,\theta)) y f_0(y|x,\theta) f(x,\theta) dy dx}{\int h(p(x,\theta)) f(x,\theta) dx}.$$

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

$$\frac{\partial \tau_h(\theta_0)}{\partial \theta} = \mathbb{E}[F_{\tau_h}(Y,T,X) \cdot s(Y,T,X|\theta_0)].$$

First, we will calculate $\frac{\partial \tau_h(\theta_0)}{\partial \theta}$. Let $\mu_h = \int h(p(x)) f(x) dx$. Then,

$$\begin{aligned}
\frac{\partial \tau_h(\theta_0)}{\partial \theta} &= \frac{1}{\mu_h} \Bigg[\iint h(p(x,\theta_0)) y \left[s_1(y|x,\theta_0) f_1(y|x,\theta_0) - s_0(y|x,\theta_0) f_0(y|x,\theta_0) \right] f(x,\theta_0) dy dx \\
&\quad + \int h(p(x,\theta_0)) \left[\tau(x) - \tau_h \right] s_x(x|\theta_0) f(x,\theta_0) dx \\
&\quad + \int h'(p(x,\theta_0)) p'(x,\theta_0) \left[\tau(x) - \tau_h \right] f(x,\theta_0) dx \Bigg]
\end{aligned}$$

where $h'(p(x)) = \frac{d}{dp(x)} h(p(x))$. The following choice for F_{τ_h} satisfies the condition:

$$\begin{aligned}
F_{\tau_h}(Y,T,X) &= \frac{T \cdot h(p(X))}{\mu_h \cdot p(X)} (Y - \mathbb{E}[Y(1)|X]) - \frac{(1-T) \cdot h(p(X))}{\mu_h \cdot (1 - p(X))} (Y - \mathbb{E}[Y(0)|X]) \\
&\quad + \frac{h(p(X))}{\mu_h} (\tau(X) - \tau_h) + \frac{(T - p(X)) \cdot h'(p(X))}{\mu_h} (\tau(X) - \tau_h).
\end{aligned}$$

To see this consider the following product,

$$
\begin{aligned}
& F_{\tau_h}(Y,T,X)\cdot s(Y,T,X|\theta_0) \\
&= \frac{T^2\cdot h(p(X))}{\mu_h\cdot p(X)} s_1(Y,X)(Y-\mathbb{E}[Y(1)|X]) && \text{(A.1)}\\
&+ 0 && \text{(A.2)}\\
&+ \frac{T\cdot h(p(X))}{\mu_h\cdot p(X)} s_x(X)(Y-\mathbb{E}[Y(1)|X]) && \text{(A.3)}\\
&+ \frac{T(T-p(X))\cdot h(p(X))}{\mu_h\cdot p(X)^2(1-p(X))} p'(X)(Y-\mathbb{E}[Y(1)|X]) && \text{(A.4)}\\
&- 0 && \text{(A.5)}\\
&- \frac{(1-T)^2\cdot h(p(X))}{\mu_h\cdot (1-p(X))} s_0(Y,X)(Y-\mathbb{E}[Y(0)|X]) && \text{(A.6)}\\
&- \frac{(1-T)\cdot h(p(X))}{\mu_h\cdot (1-p(X))} s_x(X)(Y-\mathbb{E}[Y(0)|X]) && \text{(A.7)}\\
&- \frac{(1-T)\cdot (T-p(X))\cdot h(p(X))}{\mu_h\cdot p(X)(1-p(X))^2} p'(X)(Y-\mathbb{E}[Y(0)|X]) && \text{(A.8)}\\
&+ \frac{T\cdot h(p(X))}{\mu_h} s_1(Y,X)(\tau(X)-\tau_h) && \text{(A.9)}\\
&+ \frac{(1-T)\cdot h(p(X))}{\mu_h} s_0(Y,X)(\tau(X)-\tau_h) && \text{(A.10)}\\
&+ \frac{h(p(X))}{\mu_h} s_x(X)(\tau(X)-\tau_h) && \text{(A.11)}\\
&+ \frac{(T-p(X))\cdot h(p(X))}{\mu_h\cdot p(X)(1-p(X))} p'(X)(\tau(X)-\tau_h) && \text{(A.12)}\\
&+ \frac{T(T-p(X))\cdot h'(p(x))}{\mu_h} s_1(Y,X)(\tau(X)-\tau_h) && \text{(A.13)}\\
&+ \frac{(1-T)(T-p(X))\cdot h'(p(x))}{\mu_h} s_0(Y,X)(\tau(X)-\tau_h) && \text{(A.14)}\\
&+ \frac{(T-p(X))\cdot h'(p(x))}{\mu_h} s_x(X)(\tau(X)-\tau_h) && \text{(A.15)}\\
&+ \frac{(T-p(X))^2\cdot h'(p(x))}{\mu_h\cdot p(X)(1-p(X))} p'(X)(\tau(X)-\tau_h). && \text{(A.16)}
\end{aligned}
$$

Consider each expectation in turn. Equation (1) yields,

$$\begin{aligned}
&\mathbb{E}\left[\frac{T^2 \cdot h(p(X))}{\mu_h \cdot p(X)} s_1(Y,X)(Y - \mathbb{E}[Y(1)|X])\right] \\
&\quad = \mathbb{E}\left[\frac{T \cdot h(p(X))}{\mu_h \cdot p(X)} s_1(Y(1),X)Y(1)\right] - \mathbb{E}\left[\frac{T \cdot h(p(X))}{\mu_h \cdot p(X)} s_1(Y(1),X)\mathbb{E}[Y(1)|X]\right] \\
&\quad = \mathbb{E}\left[\frac{h(p(X))}{\mu_h \cdot p(X)} \mathbb{E}[T \cdot s_1(Y(1),X)Y(1)|X]\right] \\
&\qquad -\mathbb{E}\left[\frac{h(p(X))}{\mu_h \cdot p(X)} \mathbb{E}[T \cdot s_1(Y(1),X)|X] \cdot \mathbb{E}[Y(1)|X]\right] \\
&\quad = \mathbb{E}\left[\frac{h(p(X))}{\mu_h} \mathbb{E}[s_1(Y(1),X)Y(1)|X]\right] - \mathbb{E}\left[\frac{h(p(X))}{\mu_h} \mathbb{E}[s_1(Y(1),X)|X] \cdot \mathbb{E}[Y(1)|X]\right] \\
&\quad = \mathbb{E}\left[\frac{h(p(X))}{\mu_h} \mathbb{E}[s_1(Y(1),X)Y(1)|X]\right] \\
&\quad = \frac{1}{\mu_h} \iint h(p(x)) y s_1(y,x) f_1(y|x) f(x) dy dx.
\end{aligned}$$

Equation (3) yields,

$$\begin{aligned}
&\mathbb{E}\left[\frac{T \cdot h(p(X))}{\mu_h \cdot p(X)} s_x(X)(Y - \mathbb{E}[Y(1)|X])\right] \\
&\quad = \mathbb{E}\left[\frac{T \cdot h(p(X))}{\mu_h \cdot p(X)} s_x(X)Y(1)\right] - \mathbb{E}\left[\frac{T \cdot h(p(X))}{\mu_h \cdot p(X)} s_x(X)\mathbb{E}[Y(1)|X]\right] \\
&\quad = \mathbb{E}\left[\frac{h(p(X))}{\mu_h \cdot p(X)} s_x(X)\mathbb{E}[TY(1)|X]\right] - \mathbb{E}\left[\frac{h(p(X))}{\mu_h} s_x(X)\mathbb{E}[Y(1)|X]\right] \\
&\quad = \mathbb{E}\left[\frac{h(p(X))}{\mu_h} s_x(X)\mathbb{E}[Y(1)|X]\right] - \mathbb{E}\left[\frac{h(p(X))}{\mu_h} s_x(X)\mathbb{E}[Y(1)|X]\right] \\
&\quad = 0.
\end{aligned}$$

Equation (4) yields,

$$\begin{aligned}
&\mathbb{E}\left[\frac{T(T - p(X)) \cdot h(p(X))}{\mu_h \cdot p(X)^2(1 - p(X))} p'(X)(Y - \mathbb{E}[Y(1)|X])\right] \\
&\quad = \mathbb{E}\left[\frac{T(T - p(X)) \cdot h(p(X))}{\mu_h \cdot p(X)^2(1 - p(X))} p'(X)(Y(1) - \mathbb{E}[Y(1)|X])\right] \\
&\quad = \mathbb{E}\left[\frac{h(p(X))}{\mu_h \cdot p(X)^2(1 - p(X))} p'(X)\mathbb{E}[T(T - p(X))Y(1)|X]\right] \\
&\qquad -\mathbb{E}\left[\frac{h(p(X))}{\mu_h \cdot p(X)^2(1 - p(X))} p'(X)\mathbb{E}[T(T - p(X))|X] \cdot \mathbb{E}[Y(1)|X]\right] \\
&\quad = \mathbb{E}\left[\frac{h(p(X))}{\mu_h \cdot p(X)} p'(X)\mathbb{E}[Y(1)|X]\right] - \mathbb{E}\left[\frac{h(p(X))}{\mu_h \cdot p(X)} p'(X)\mathbb{E}[Y(1)|X]\right] \\
&\quad = 0.
\end{aligned}$$

Equation (6) yields,

$$
\begin{aligned}
&-\mathbb{E}\left[\frac{(1-T)^2\cdot h(p(X))}{\mu_h\cdot(1-p(X))}s_0(Y,X)(Y-\mathbb{E}[Y(0)|X])\right]\\
&\quad=-\mathbb{E}\left[\frac{(1-T)\cdot h(p(X))}{\mu_h\cdot(1-p(X))}s_0(Y(0),X)Y(0)\right]\\
&\qquad+\mathbb{E}\left[\frac{(1-T)\cdot h(p(X))}{\mu_h\cdot(1-p(X))}s_0(Y(0),X)\mathbb{E}[Y(0)|X]\right]\\
&\quad=-\mathbb{E}\left[\frac{h(p(X))}{\mu_h\cdot(1-p(X))}\mathbb{E}[(1-T)s_0(Y(0),X)Y(0)|X]\right]\\
&\qquad+\mathbb{E}\left[\frac{h(p(X))}{\mu_h\cdot(1-p(X))}\mathbb{E}[(1-T)s_0(Y(0),X)|X]\mathbb{E}[Y(0)|X]\right]\\
&\quad=-\mathbb{E}\left[\frac{h(p(X))}{\mu_h}\mathbb{E}[s_0(Y(0),X)Y(0)|X]\right]+\mathbb{E}\left[\frac{h(p(X))}{\mu_h}\mathbb{E}[s_0(Y(0),X)|X]\mathbb{E}[Y(0)|X]\right]\\
&\quad=-\mathbb{E}\left[\frac{h(p(X))}{\mu_h}\mathbb{E}[s_0(Y(0),X)Y(0)|X]\right]\\
&\quad=-\frac{1}{\mu_h}\iint h(p(x))ys_0(y,x)f_0(y|x)f(x)dydx.
\end{aligned}
$$

Equation (7) yields,

$$
\begin{aligned}
&-\mathbb{E}\left[\frac{(1-T)\cdot h(p(X))}{\mu_h\cdot(1-p(X))}s_x(X)(Y-\mathbb{E}[Y(0)|X])\right]\\
&\quad=-\mathbb{E}\left[\frac{(1-T)\cdot h(p(X))}{\mu_h\cdot(1-p(X))}s_x(X)Y(0)\right]+\mathbb{E}\left[\frac{(1-T)\cdot h(p(X))}{\mu_h\cdot(1-p(X))}s_x(X)\mathbb{E}[Y(0)|X]\right]\\
&\quad=-\mathbb{E}\left[\frac{h(p(X))}{\mu_h\cdot(1-p(X))}s_x(X)\mathbb{E}[(1-T)Y(0)|X]\right]+\mathbb{E}\left[\frac{h(p(X))}{\mu_h}s_x(X)\mathbb{E}[Y(0)|X]\right]\\
&\quad=-\mathbb{E}\left[\frac{h(p(X))}{\mu_h}s_x(X)\mathbb{E}[Y(0)|X]\right]+\mathbb{E}\left[\frac{h(p(X))}{\mu_h}s_x(X)\mathbb{E}[Y(0)|X]\right]\\
&\quad=0
\end{aligned}
$$

Equation (8) yields,

$$
\begin{aligned}
&-\mathbb{E}\left[\frac{(1-T)\cdot(T-p(X))\cdot h(p(X))}{\mu_h\cdot p(X)(1-p(X))^2}p'(X)(Y-\mathbb{E}[Y(0)|X])\right]\\
&\quad=-\mathbb{E}\left[\frac{(1-T)\cdot(T-p(X))\cdot h(p(X))}{\mu_h\cdot p(X)(1-p(X))^2}p'(X)(Y(0)-\mathbb{E}[Y(0)|X])\right]\\
&\quad=-\mathbb{E}\left[\frac{h(p(X))}{\mu_h\cdot p(X)(1-p(X))^2}p'(X)\mathbb{E}[(1-T)(T-p(X))Y(0)|X]\right]\\
&\qquad+\mathbb{E}\left[\frac{h(p(X))}{\mu_h\cdot p(X)(1-p(X))^2}p'(X)\mathbb{E}[(1-T)(T-p(X))|X]\mathbb{E}[Y(0)|X]\right]\\
&\quad=-\mathbb{E}\left[\frac{h(p(X))}{\mu_h\cdot(1-p(X))}p'(X)\mathbb{E}[Y(0)|X]\right]+\mathbb{E}\left[\frac{h(p(X))}{\mu_h\cdot(1-p(X)))}p'(X)\mathbb{E}[Y(0)|X]\right]\\
&\quad=0.
\end{aligned}
$$

Equation (9) yields,

$$\begin{aligned}
&\mathbb{E}\left[\frac{T \cdot h(p(X))}{\mu_h} s_1(Y, X)(\tau(X) - \tau_h)\right] \\
&\quad = \mathbb{E}\left[\frac{T \cdot h(p(X))}{\mu_h} s_1(Y(1), X)(\tau(X) - \tau_h)\right] \\
&\quad = \mathbb{E}\left[\frac{h(p(X))}{\mu_h} \mathbb{E}[T s_1(Y(1), X)|X](\tau(X) - \tau_h)\right] \\
&\quad = \mathbb{E}\left[\frac{p(X) \cdot h(p(X))}{\mu_h} \mathbb{E}[s_1(Y(1), X)|X](\tau(X) - \tau_h)\right] \\
&\quad = 0.
\end{aligned}$$

Equation (10) yields,

$$\begin{aligned}
&\mathbb{E}\left[\frac{(1-T) \cdot h(p(X))}{\mu_h} s_0(Y, X)(\tau(X) - \tau_h)\right] \\
&\quad = \mathbb{E}\left[\frac{(1-T) \cdot h(p(X))}{\mu_h} s_0(Y(0), X)(\tau(X) - \tau_h)\right] \\
&\quad = \mathbb{E}\left[\frac{h(p(X))}{\mu_h} \mathbb{E}[(1-T) s_0(Y(0), X)|X](\tau(X) - \tau_h)\right] \\
&\quad = \mathbb{E}\left[\frac{(1-p(X)) \cdot h(p(X))}{\mu_h} \mathbb{E}[s_0(Y(0), X)|X](\tau(X) - \tau_h)\right] \\
&\quad = 0.
\end{aligned}$$

Equation (11) yields,

$$\begin{aligned}
&\mathbb{E}\left[\frac{h(p(X))}{\mu_h} s_x(X)(\tau(X) - \tau_h)\right] \\
&\quad = \frac{1}{\mu_h} \int h(p(x))(\tau(x) - \tau_h) s_x(x) f(x) dx
\end{aligned}$$

Equation (12) yields,

$$\begin{aligned}
&\mathbb{E}\left[\frac{(T - p(X)) \cdot h(p(X))}{\mu_h \cdot p(X)(1-p(X))} p'(X)(\tau(X) - \tau_h)\right] \\
&\quad = \mathbb{E}\left[\frac{\mathbb{E}[(T - p(X))|X] \cdot h(p(X))}{\mu_h \cdot p(X)(1-p(X))} p'(X)(\tau(X) - \tau_h)\right] \\
&\quad = 0.
\end{aligned}$$

Equation (13) yields,

$$\begin{aligned}
&\mathbb{E}\left[\frac{T(T - p(X)) \cdot h'(p(X))}{\mu_h} s_1(Y, X)(\tau(X) - \tau_h)\right] \\
&\quad = \mathbb{E}\left[\frac{T(T - p(X)) \cdot h'(p(X))}{\mu_h} s_1(Y(1), X)(\tau(X) - \tau_h)\right] \\
&\quad = \mathbb{E}\left[\frac{h'(p(X))}{\mu_h} \mathbb{E}[T(T - p(X)) s_1(Y(1), X)|X](\tau(X) - \tau_h)\right] \\
&\quad = \mathbb{E}\left[\frac{p(X)(1-p(X)) \cdot h'(p(X))}{\mu_h} \mathbb{E}[s_1(Y(1), X)|X](\tau(X) - \tau_h)\right] \\
&\quad = 0.
\end{aligned}$$

Equation (14) yields,

$$\begin{aligned}
&\mathbb{E}\left[\frac{(1-T)(T-p(X))\cdot h'(p(X))}{\mu_h}s_0(Y,X)(\tau(X)-\tau_h)\right] \\
&\quad=\mathbb{E}\left[\frac{(1-T)(T-p(X))\cdot h'(p(X))}{\mu_h}s_0(Y(0),X)(\tau(X)-\tau_h)\right] \\
&\quad=\mathbb{E}\left[\frac{h'(p(X))}{\mu_h}\mathbb{E}[(1-T)(T-p(X))s_0(Y(0),X)|X](\tau(X)-\tau_h)\right] \\
&\quad=\mathbb{E}\left[\frac{p(X)(1-p(X))\cdot h'(p(X))}{\mu_h}\mathbb{E}[s_0(Y(0),X)|X](\tau(X)-\tau_h)\right] \\
&\quad=0.
\end{aligned}$$

Equation (15) yields,

$$\begin{aligned}
&\mathbb{E}\left[\frac{(T-p(X))\cdot h'(p(X))}{\mu_h}s_x(X)(\tau(X)-\tau_h)\right] \\
&\quad=\mathbb{E}\left[\frac{\mathbb{E}[(T-p(X))|X]\cdot h'(p(X))}{\mu_h}s_x(X)(\tau(X)-\tau_h)\right] \\
&\quad=0.
\end{aligned}$$

Equation (16) yields,

$$\begin{aligned}
&\mathbb{E}\left[\frac{(T-p(X))^2\cdot h'(p(X))}{\mu_h\cdot p(X)(1-p(X))}p'(X)(\tau(X)-\tau_h)\right] \\
&\quad=\mathbb{E}\left[\frac{(T^2+p(X)^2-2\cdot Tp(X))\cdot h'(p(X))}{\mu_h\cdot p(X)(1-p(X))}p'(X)(\tau(X)-\tau_h)\right] \\
&\quad=\mathbb{E}\left[\frac{\mathbb{E}[T+p(X)^2-2\cdot Tp(X)|X]\cdot h'(p(X))}{\mu_h\cdot p(X)(1-p(X))}p'(X)(\tau(X)-\tau_h)\right] \\
&\quad=\mathbb{E}\left[\frac{p(X)(1-p(X)\cdot h'(p(X))}{\mu_h\cdot p(X)(1-p(X))}p'(X)(\tau(X)-\tau_h)\right] \\
&\quad=\mathbb{E}\left[\frac{h'(p(X))}{\mu_h}p'(X)(\tau(X)-\tau_h)\right] \\
&\quad=\frac{1}{\mu_h}\int h'(p(x))p'(x)[\tau(x)-\tau_h]f(x)dx
\end{aligned}$$

Since $F_{\tau_h}\in\mathcal{T}$, the variance bound is

$$\begin{aligned}
\mathbb{E}[F_{\tau_h}(Y,T,X)^2] &= \mathbb{E}\left[\frac{[h(p(X))]^2}{(\mu_h)^2\cdot p(X)}\cdot\mathbb{V}[Y(1)|X]\right]+\mathbb{E}\left[\frac{[h(p(X))]^2}{(\mu_h)^2\cdot(1-p(X))}\cdot\mathbb{V}[Y(0)|X]\right] \\
&\quad+\mathbb{E}\left[\frac{[h(p(X))+(T-p(X))\cdot h'(p(X))]^2}{(\mu_h)^2}(\tau(X)-\tau_h)^2\right] \\
&= \mathbb{E}\left[\frac{[h(p(X))]^2}{(\mu_h)^2\cdot p(X)}\cdot\mathbb{V}[Y(1)|X]\right]+\mathbb{E}\left[\frac{[h(p(X))]^2}{(\mu_h)^2\cdot(1-p(X))}\cdot\mathbb{V}[Y(0)|X]\right] \\
&\quad+\mathbb{E}\left[\frac{[h(p(X))]^2+p(X)(1-p(X))\cdot[h'(p(X))]^2}{(\mu_h)^2}(\tau(X)-\tau_h)^2\right]
\end{aligned}$$

For the special case of $h(p(x))=p(x)$ the semiparametric efficiency bound is,

$$\mathbb{E}\left[\frac{p(X)}{(\mu_h)^2}\cdot\mathbb{V}[Y(1)|X]\right]+\mathbb{E}\left[\frac{p(X))^2}{(\mu_h)^2\cdot(1-p(X))}\cdot\mathbb{V}[Y(0)|X]\right]+\mathbb{E}\left[\frac{p(X)}{(\mu_h)^2}(\tau(X)-\tau_h)^2\right].$$

For the special case of $h(p(x)) = p(x)(1-p(x))$ the semiparametric efficiency bound is,

$$\mathbb{E}\left[\frac{p(X)(1-p(X))^2}{(\mu_h)^2}\cdot\mathbb{V}[Y(1)|X]\right]+\mathbb{E}\left[\frac{p(X)^2(1-p(X))}{(\mu_h)^2}\cdot\mathbb{V}[Y(0)|X]\right]$$
$$+\mathbb{E}\left[\frac{p(X)^2(1-p(X))^2+p(X)(1-p(X))\cdot(1-2\cdot p(X))^2}{(\mu_h)^2}(\tau(X)-\tau_h)^2\right]$$

which simplifies to

$$\mathbb{E}\left[\frac{p(X)(1-p(X))^2}{(\mu_h)^2}\cdot\mathbb{V}[Y(1)|X]\right]+\mathbb{E}\left[\frac{p(X)^2(1-p(X))}{(\mu_h)^2}\cdot\mathbb{V}[Y(0)|X]\right]$$
$$+\mathbb{E}\left[\frac{p(X)(1-p(X))(3p(X)^2-3p(X)+1)}{(\mu_h)^2}(\tau(X)-\tau_h)^2\right].$$

References

Abadie, A., and G. Imbens, (2002), "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," NBER technical working paper # 283.

Angrist, J., (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66(2): 249-288.

Blundell, R. and M. Costa-Dias (2002), "Alternative Approaches to Evaluation in Empirical Microeconomics," Institute for Fiscal Studies, Cemmap working paper cwp10/02.

Bolthausen, E., and F. Götze (1993), "The Rate of Convergence for Multivariate Sampling Statistics," *The Annals of Statistics*, V. 21: 1692-1710.

Crump, R., V. J. Hotz, G. Imbens and O. Mitnik, (2005), "Nonparametric Tests for Treatment Effect Heterogeneity', Unpublished manuscript, Departments of Economics, Miami, UC-Berkeley and UCLA.

Dehejia, R., and S. Wahba, (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, 94: 1053-1062.

Frölich, M. (2002), "What is the Value of knowing the propensity score for estimating average treatment effects", Department of Economics, University of St. Gallen.

Götze, F., (1991), "On the Rate of Convergence in te Multivariate CLT," *The Annals of Probability*, Vol 19(2), 724-739.

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

Hampel, F., E. Ronchetti, P. Rousseeuw and W. Stahel (2005), *Robust Statistics: The Approach Base on Influence Fuctions,* Wiley Series in Probability and Statistics.

Härdle, W., and E. Mammen, (1993), "Comparing Nonparametric Versus Parametric Regression Fits," *The Annals of Statistics*, Vol 21(4), 1926-1947.

Heckman, J., and V. J. Hotz, (1989), "Alternative Methods for Evaluating the Impact of Training Programs," (with discussion), *Journal of the American Statistical Association.*, 84(804): 862-874.

Heckman, J., H. Ichimura, and P. Todd, (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies* 64(4): 605-654.

Heckman, J., H. Ichimura, and P. Todd, (1998), "Matching as an Econometric Evaluation 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(5): 1017-1098.

Heckman, J., R. LaLonde, and J. Smith, (1999), "The economics and econometrics of active labor market programs," in O. Ashenfelter and D. Card (eds.), *Hanbook of Labor Economics*, Vol. 3A, North-Holland, Amsterdam, 1865-2097.

Hirano, K., and G. Imbens (2001), "Estimation of Causal Effects Using Propensity Score Weighting: An Application of Data on Right Hear Catherization," *Health Services and Outcomes Research Methodology*, 2: 259-278.

Hirano, K., G. Imbens, and G. Ridder, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4): 1161-1189.

Ho, D., K. Imai, G. King, and E. Stuart, (2004), "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," mimeo, Department of Government, Harvard University.

Horowitz, J., and V. Spokoiny, (2001), "An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model Against a Nonparametric Alternative," *Econometrica*, 69(3): 599-631.

Huber, P. J. (2004), *Robust Statists*, Wiley Series in Probability and Statistics.

Ichino, A., F. Mealli, and T. Nannicini, (2005), "Sensitivity of Matching Estimators to Unconfoundedness. An Application to the Effect of Temporary Work on Future Employment," mimeo, European University Institute.

Imbens, G. (2003), "Sensivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, Papers and Proceedings.

Imbens, G., (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review, *Review of Economics and Statistics*, 86(1): 1-29.

Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 61(2): 467-476.

Imbens, G., W. Newey and G. Ridder, (2003), "Mean-squared-error Calculations for Average Treatment Effects," unpublished manuscript, Department of Economics, UC Berkeley.

Lalonde, R.J., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76: 604-620.

Lechner, M, (2002), "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies," *Review Economics and Statistics*, 84(2): 205-220.

Lechner, M, (2002), "Some Practical Issues in the Evaluation of Heterogenous Labour Market Programmes by Matching Methods," *Journal of the Royal Statistical Society*, Series A, 165: 659–82.

Lee, M.-J., (2005a), *Micro-Econometrics for Policy, Program, and Treatment Effects* Oxford University Press, Oxford.

Lee, M.-J., (2005b), "Treatment Effect and Sensitivity Analysis for Self-selected Treatment and Selectively Observed Response," unpublished manuscript, School of Economics and Social Sciences, Singapore Management University.

Manski, C., (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80: 319-323.

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

Neumeyer, N., and H. Dette, (2003), "Nonparametric Comparison of Regression Curves: An Empirical Process Approach," *The Annals of Statistics*, Vol 31, 880-920.

Pinkse, J., and P. Robinson, (1995), "Pooling Nonparametric Estimates of Regression Functions with a Similar Shape," in *Statistical Methods of Econometrics and Quantitative Economics: A Volume in Honour of C.R. Rao*, G.S. Maddala, P.C.B. Phillips and T.N. Srinivisan, eds., 172-197.

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

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

Robins, J.M., S. Mark, and W. Newey, (1992), "Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders," *Biometrics*, 48(2): 479-495.

Robinson, P., (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 67: 645-662.

Rosenbaum, P., (1989), "Optimal Matching in Observational Studies", *Journal of the American Statistical Association*, 84: 1024-1032.

Rosenbaum, P., (2001), *Observational Studies*, second edition, Springer Verlag, New York.

Rosenbaum, P., and D. Rubin, (1983a), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70: 41-55.

Rosenbaum, P., and D. Rubin, (1983b), "Assessing the Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45: 212-218.

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. B., (1978), "Bayesian inference for causal effects: The Role of Randomization", *A*nnals of Statistics, 6: 34-58.

Shadish, W., T. Cook, and D. Campbell, *Experimental and Quasi-Experimental Designs*, Houghton Mifflin, Boston, MA.

Smith, J., and P. Todd, (2005), "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of Econometrics*, 125: 305-353.

Stock, J., (1989), "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84(406): 567-575.

Wooldridge, J., (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Zhao, Z., (2004), "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics and an Application", *Review of Economics and Statistics*, 86(1): 91-107.

Table 1: Covariate Balance for Lalonde Data

	mean	stand. dev.	mean contr.	mean treat.	Normalized Dif. in Treat. and Contr. Ave's all	[t-stat]	$a < e(x) < 1-a$	optimal weights	prop score weighted
age	34.23	10.50	34.85	25.82	-0.86	[-16.0]	-0.18	-0.25	-0.35
educ	11.99	3.05	12.12	10.35	-0.58	[-11.1]	-0.04	-0.08	-0.12
black	0.29	0.45	0.25	0.84	1.30	[21.0]	0.20	0.27	0.37
hispanic	0.03	0.18	0.03	0.06	0.15	[1.5]	0.07	-0.01	-0.08
married	0.82	0.38	0.87	0.19	-1.76	[-22.8]	-0.81	-0.79	-0.70
unempl '74	0.13	0.34	0.09	0.71	1.85	[18.3]	0.78	0.78	1.19
uenmpl '75	0.13	0.34	0.10	0.60	1.46	[13.7]	0.51	0.47	0.90
earn '74	18.23	13.72	19.43	2.10	-1.26	[-38.6]	-0.20	-0.23	-0.26
earn '75	17.85	13.88	19.06	1.53	-1.26	[-48.6]	-0.14	-0.18	-0.18
log odds ratio	-7.87	4.91	-8.53	1.08	1.96	[53.6]	0.42	0.48	0.57

Table 2: Asymptotic Standard Errors for Lalonde Data

	ATE	ATT	OSATE	OWATE
Asymptotic Standard Error	636.58	2.58	1.62	1.29
Ratio to All	1.0000	0.0040	0.0025	0.0020

Table 3: Subsample Sizes for Lalonde Data: Propensity Score Threshold 0.0660

	$e(x) < a$	$a \leq e(x) \leq 1-a$	$1-a < e(x)$	all
controls	2302	183	5	2490
treated	9	129	47	185
all	2311	312	52	2675