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 Mitnik[¶]

First Draft: July 2004
This Draft: June 11, 2005

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. We also develop tests for the null hypotheses of a constant and a zero conditional average treatment effect. The latter is in practice more powerful than the commonly used test for a zero average treatment effect.

JEL Classification: C14, C21, C52
Keywords: *Average Treatment Effects*

[*] We are grateful for helpful comments by Richard Blundell, and by participants in seminars at the ASSA meetings in Philadelphia, University College London and the Malinvaud seminar at CREST. Financial support for this research was generously provided through NSF grant SES 0136789.

[†] Department of Economics, University of California at Berkeley, crump@econ.berkeley.edu.

[‡] Department of Economics, University of California at Los Angeles, hotz@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/.

[¶] Department of Economics, University of Miami, omitnik@exchange.sba.miami.edu.

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 Rotznitzky (1995). See Rosenbaum (2001), Heckman, Lalonde and Smith (1999), Wooldridge (2002), Blundell and Costa-Diaz (2002) and Imbens (2004) for surveys of this literature. 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.[1]

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) suggests 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 the matched data. All of these methods have some advantages as well as drawbacks. All of them change the estimand, at least in finite samples. They all do tend to reduce sensitivity of the final estimates to model specification. However, they rely on arbitrary choices regarding thresholds for discarding observations, and

[1] 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."

there are few formal results on their properties.

In this paper we propose a systematic approach to account for subpopulations with limited overlap in the covariates. This 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 possibly 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 leads under some conditions to choosing the subpopulation with the propensity score in an interval $[a, 1-a]$, with the optimal value of a 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 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 $(e(X) \cdot (1-e(X)))^{-1}$, and 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.

It is important to stress that these methods change the estimand. Instead of focusing on the traditional estimands, the population average treatment effect, the average effect for the subpopulation of the treated or another *a priori* defined subpopulation of interest, we focus on average effects for a statistically defined (weighted) subpopulation.[2] This change of focus is *not* motivated by an intrinsic interest in the subpopulation we ultimately estimate the average causal effect for. Rather, it acknowledges and addresses the difficulties in making inferences about the population of primary interest. Instead of reporting solely the potentially imprecise estimate for

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

the population average treatment effect we propose reporting both estimates for the population of interest and estimates for subpopulations where we can make more precise inferences. In settings where we cannot ascertain with much confidence the sign of the population average treatment effect such estimates may serve to demonstrate that there are subpopulations that benefit from or are harmed by the program, as well as the extent of this benefit or harm. It is also important to 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.

This change of estimand is uncommon in econometric analyses.[3] Typically in such analyses the estimand is defined *a priori*, followed by a presentation of estimates 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 or interest regions of the parameter space consistent with the model may be reported in a bounds analysis of the type developed by Manski (1990, 2003). Here our approach is different and to some extent complementary. Sacrificing some external validity by changing the sample from one that was potentially representative of the population of interest we potentially gain some internal validity by changing it to a sample where we can obtain more precise and credible estimates.[4] Our proposed stress on internal validity at the expense of external validity is similar to that in randomized experiments which are often carried out in populations unrepresentative of the population of interest.[5] More generally, the primacy of internal validity over external validity is advocated in many discussions of causality (see, for example, Shadish, Cook, and Campbell, 2002).

In interpreting our results it is also of interest to consider estimation of the average treatment effect under the assumption that it does not vary with the covariates.[6] This assumption is generally informative except in the case where the propensity score is constant. Under this assumption the model is a special case of Robinson's (1988) partial linear model. The efficient estimator for that case is identical to the efficient estimator in the heterogenous case for the weighted average treatment effect with the weights chosen to obtain the most precisely estimated average treatment effect.

We also develop a set of three new nonparametric tests. We first test the hypothesis that there is no variation in the conditional average treatment effect by covariates. Second, we test the hypothesis that the conditional average treatment effect is zero for all values of the covariates. Third, we test the hypothesis that the optimally weighted average treatment effect is equal to zero.

We illustrate these methods using three data sets. The first is 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

[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.

[4] 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.

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

[6] 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.

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 very different from the treated that it is essentially impossible to estimate the population average treatment effect. More relevant, given the fact that most of the researchers analyzing this data set have focused on the average effect for the treated, is that the variance for the optimal subsample is only 40% of that for the propensity score weighted sample (which estimates the effect on the treated).

The second data set, containing a sample of lottery players, was collected by Imbens, Rubin and Sacerdote (2001), They compare labor market outcomes for lottery winners and losers. Here the differences between the control and treatment group are much smaller, although they are still significantly different from zero at conventional levels. Here the optimal subpopulation approach suggests dropping 108 observations out of 496, and leads to an reduction in the variance of 60%.

The last example uses data from the Greater Avenue for INdependence (GAIN) experiments designed to evaluate labor market programs in California. We use data from the Los Angeles and Riverside locations to see if controls from one location can be used as a nonexperimental comparison group in the other location. Here the covariates are quite close. The optimal subpopulation approach suggests dropping only 407 observations out of 4035. The calculations suggest that even though the two subpopulations are close, this still leads to a decrease in the variance of 20%.

In all three cases the improvement in precision from focusing on the restricted sample is substantial. The additional improvement from moving from the optimal subpopulation to the optimally weighted sample is considerably smaller. The increased difficulty in interpretation of the weighted average treatment effect may not be worth this additional increase in precision.

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.

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

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

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

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

The key is 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 $X = x_0$ 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 $X = x_1$ 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

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

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 may wish to report also the estimator for the precisely estimable average treatment effect to convey the information 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.

In this paper 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 the set of covariates such that the propensity score is in the closed interval $[a, 1-a]$. The optimal value of the boundary point a 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, the generalization of (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, $\sigma_w^2(x) = \mathrm{Var}(Y(w)|X = x)$, 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 $c > 0$,

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

In addition 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*

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

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.[7] 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 estimators of the population average treatment effect τ_P but also as estimators of the conditional average treatment effect τ_C. As an estimator of τ_C the efficient estimator $\hat{\tau}_H$ has asymptotic variance

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

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:

Theorem 4.2 (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*

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

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 shows that the efficient estimator for $\tau_{P,g}$, as an estimator for the conditional average treatment effect version of this estimand,

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

has asymptotic variance

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

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

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

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

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

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 (4.8) 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

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

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.[8] 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.[9] 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 modelling decisions such as the inclusion of covariates or functional form assumptions.

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 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

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

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

$$V_C^{\text{eff}\prime}(\mathcal{A}) = \frac{1}{q(\mathcal{A})} \cdot \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} \middle| X \in \mathcal{A}\right]. \tag{6.9}$$

[8] 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.

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

We look for an optimal $\mathcal{A}$, denoted by $\mathcal{A}^*$, that minimizes the asymptotic variance (6.9) 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 $a_0 = 0$. Calculate

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

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

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

leading to

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

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

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

7 The Optimally Weighted Average Treatment Effect

Lemma 7.1 *Suppose Assumptions – hold, and that $\sigma_0^2(x) = \sigma_1^2(x) = \sigma^2$ and that $\tau(x) = \tau$ for all x. Then the Optimally Weighted Average Treatment Effect (OWATE) is τ_{g^*}, where*

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

8 Testing

In this section we discuss some statistical tests. Most of the statistical tests discussed in this literature focus on the appropriateness of the various estimators (Heckman and Hotz, 1989). Some attempt to provide assessments of the unconfoundedness assumptions (Imbens, 2004). We focus on three different hypotheses concerning the conditional average treatment effect $\tau(x)$. The first hypothesis we consider is that the conditional average treatment effect is zero for all values of the covariates. The second one is the hypothesis that $\tau(x)$ is constant as a function of x. The third test concerns the hypothesis that the optimally weighted conditional average treatment effect τ_{C,g^*} is equal to zero. The latter test is very simple. The previous results lead to a root-N consistent estimator that is asymptotically normal with zero asymptotic bias so that we can use a simple Wald test. The other tests are more complext, requiring comparisons of nonparametric estimators of regression functions over the entire support of the covariates.

There are some obvious relationships between the null hypotheses. H'_0 implies H'_0 and H''_0. The last two hypotheses, H'_0 and H''_0 combined imply H_0. There is no direct relationship between H'_0 and H''_0.

In order to develop these tests we need estimators for the two regression functions. We use the series estimator for the regression function $\mu_w(x)$ developed by Imbens, Newey and Ridder (2004). Let K denote the number of terms in the series. As the basis we use power series. Let $\lambda = (\lambda_1, ..., \lambda_d)$ be a multi-index of dimension d, that is, a d-dimensional vector of non-negative integers, with $|\lambda| = \sum_{k=1}^{d} \lambda_k$, and let $x^{\lambda} = x_1^{\lambda_1} \dots x_d^{\lambda_d}$. Consider a series $\{\lambda(r)\}_{r=1}^{\infty}$ containing all distinct vectors such that $|\lambda(r)|$ is nondecreasing. Let $p_r(x) = x^{\lambda(r)}$, where $P_r(x) = (p_1(x), ..., p_r(x))'$. Given the assumptions below the expectation $\Omega_K = \mathbb{E}[P_K(X)P_K(X)'|W = 1]$ is nonsingular for all K. Hence we can construct a sequence $R_K(x) = \Omega_K^{-1/2} P_K(x)$ with $\mathbb{E}[R_K(X)R_K(X)'|W = 1] = I_K$. Let $R_{kK}(x)$ be the kth element of the vector $R_K(x)$. It will be convenient to work with this sequence of basis function $R_K(x)$. The nonparametric series estimator of the regression function $\mu_w(x)$, given K terms in the series, is given by:

$$\hat{\mu}_w(x) = r^K(x)' \left(\sum_{W_i=w} R_K(X_i)R_K(X_i)' \right)^{-} \sum_{W_i=w} R_K(X_i)Y_i = R_K(x)'\hat{\gamma}_{w,K},$$

where

$$\hat{\gamma}_{w,K} = \left(\sum_{W_i=w} R_K(X_i)R_K(X_i)' \right)^{-} \sum_{W_i=w} R_K(X_i)Y_i.$$

Define the $N_w \times K$ matrix $R_{w,K}$ with rows equal to $R_K(X_i)$ for units with $W_i = w$, and Y_w to be the N_w vector with elements equal to Y_i for the same units, so that $\hat{\gamma}_{w,K} = (R_{w,K}'R_{w,K})^{-1}(R_{w,K}'Y_w)$. Note that we use A^- here to denote a generalized inverse of A.

Given the estimator $\hat{\mu}_{w,K}(x)$ we estimate the error variance σ_w^2 as

$$\hat{\sigma}_{w,K}^2 = \frac{1}{N_w} \sum_{i|W_i=w} (Y_i - \hat{\mu}_{w,K}(X_i))^2.$$

Let $\Omega_{w,K}$ be the limiting variance of $\sqrt{N}\hat{\gamma}_{w,K}$ as the sample size increases for fixed K. We estimate this variance as

$$\hat{\Omega}_{w,K} = \hat{\sigma}_{w,K}^2 \cdot (R_{w,K}'R_{w,K}/N)^{-1}.$$

We make the following assumptions.

Assumption 8.1 (Distribution of Covariates)
$X \in \mathbb{X} \subset \mathbb{R}^d$, where $\mathbb{X}$ is the Cartesian product of intervals $[x_{jL}, x_{jU}]$, $j = 1, \dots, d$, with $x_{jL} < x_{jU}$. The density of X is bounded away from zero on $\mathbb{X}$.

Assumption 8.2 (Propensity Score)
(i) The propensity score is bounded away from zero and one.
(ii) The propensity score is s times continuously differentiable.

Assumption 8.3 (Conditional Outcome Distributions)
(i) The two regression functions $\mu_w(x)$ are t times continuously differentiable.
(ii) the conditional variance of $Y_i(w)$ given $X_i = x$ is equal to σ_w^2.

Assumption 8.4 (Rates for Series Estimators)
$K = N^\nu$, with $< \nu <$.

We assume homoskedasticty, although this assumption is not essential and can be relaxed to allow the conditional variance to depend on$x$, as long as it is bounded from above and below.

8.1 Testing the Null Hypothesis of Zero Conditional Average Treatment Effects

Here we are interested in the null hypothesis that the conditional average treatment effect $\tau(x)$ is zero for all values of the covariates. Formally,

$$\mathrm{H}_0 : \ \forall\ x \in \mathbb{X},\ \tau(x) = 0.$$

To test this hypothesis we compare estimators for $\mu_1(x)$ and $\mu_0(x)$. Given our use of series estimators we can compare the estimated parameters $\hat{\gamma}_{0,K}$ and $\hat{\gamma}_{1,K}$. Specifically, we use as the test statistic for the test of the null hypothesis H_0 the normalized quadratic form

$$T = \Big((\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K})'(\hat{\Omega}_{1,K} + \hat{\Omega}_{0,K})^{-1}(\hat{\gamma}_{1,K} - \hat{\gamma}_{0,K}) - K \Big) / \sqrt{2K}.$$

Theorem 8.1 *Suppose Assumptions ??-?? hold. Then if $\tau(x) = 0$ for all $x \in \mathbb{X}$,*

$$T \xrightarrow{d} \mathcal{N}(0, 1).$$

Proof: See Appendix.

8.2 Testing the Null Hypothesis of a Constant Conditional Average Treatment Effect

$$\mathrm{H'}_0 : \ \exists\ \tau_0,\ \text{such that}\ \forall\ x \in \mathbb{X},\ \tau(x) = \tau_0.$$

Let $\hat{\tau}_{C,g^*}$ be an estimator for τ_{C,g^*}.

For testing the null hypothesis $\mathrm{H'}_0$ we use the test statistic

$$T' = \sum_{i=1}^{N} \left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) - \hat{\tau}_{C,g^*}\right)^2 \cdot g^*(X_i).$$

8.3 Testing the Null Hypothesis of a Zero Average Treatment Effect

$$\mathrm{H''}_0 : \ \tau_{C,g^*} = 0.$$

For the testing the third null hypothesis we use the fact that $\hat{\tau}_{g^*}$ has a limiting normal distribution.

9 Some Illustrations Based on Real Data

In this section we apply the methods developed in this paper to three data sets. The data sets differ by the amount of balance between controls and treated, to highlight the effectiveness and importance of ensuring balance in a range of settings. In each case we first calculate the optimal cutoff point e^* based on an estimate of the propensity score. We report the number of observations discarded by the proposed sample selection. We also report the estimated asymptotic variance for four cases. First, the efficiency bound for the average treatment effect using the full sample. Second, the efficiency bound for the selected sample. Third, the efficiency bound for the optimally weighted sample. Fourth, we report the efficiency bound for the average effect for the treated.

9.1 The Lalonde Data

The first data set we use is a data set originally put together by Lalonde (1986), and subsequently used by Dehejia and Wahba (1999) and Smith and Todd (2004). The sample we use here is the one used by Dehejia and Wahba. The treatment of interest is a job training program. The trainees are drawn from an experimental evaluation of this program. The control group is a sample drawn from the Panel Study of Income Dynamics (PSID). The control and treatment group are very unbalanced. Table 1 presents some summary statistics. The fourth and fifth column present the averages for each of the covariates separately for the control and treatment group. Consider for example the average earnings in the year prior to the program, earn '75. For the control group from the PSID this 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.

For this data set we estimate the propensity score using a logistic model with all nine covariates entering linearly. We then use the estimated propensity score to calculate the optimal cutoff point, a in the notation of Lemma ?. The optimal cutoff point is $a = 0.0660$. The number of observations that should be discarded according to this criterion 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.

The next table presents asymptotic standard errors for four estimands. First the standard error for the population average treatment effect. Second, the asymptotic standard error for the average treatment effect in the subpopulation with $a < e(x) < 1 - a$, for the optimal value of $a = 0.0660$. Third, the standard error for the optimally weighted average treatment effect τ_g^*. Fourth, the asymptotic standard error for the average treatment effect for the treated.

The second row in this table presents ratios of the asymptotic standard error to the asymptotic standard error for the population average treatment effect. There is a huge gain to moving from the population average treatment effect to any of the three other estimands. This follows from 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 areas is difficult, and even under the assumptions made it can only be done with great uncertainty. For this example this is well known in the literature. See for example Dehejia and Wahba (1999). More interesting 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 optimal area (2.58 versus 1.62), with the latter still substantially larger than the standard error for the optimally weighted average treatment effect (1.28).

Appendix

Before proving Theorem 8.1 we present a couple of preliminary results.

Lemma A.1 *Suppose Assumptions XX-XX hold. Then* (i)

$$\left\|\hat{\Omega}_{w,K} - \Omega_{w,K}\right\| = O_p\left(\zeta(K)K^{\frac{1}{2}}N^{-\frac{1}{2}}\right),$$

and (ii) *The eigenvalues of* $\Omega_{w,K}$ *are bounded and bounded away from zero and* (iii) *The eigenvalues of* $\hat{\Omega}_{w,K}$ *are bounded and bounded away from zero if* $O_p\left(\zeta(K)K^{\frac{1}{2}}N^{-\frac{1}{2}}\right) = o_p(1)$.

Proof: We will generalize the proof found in Imbens, Newey and Ridder (2004). For (i) we will show

$$\mathbb{E}\left[\left\|\hat{\Omega}_{w,K} - \Omega_{w,K}\right\|^2\right] \leq C \cdot \zeta(K)^2 K/N$$

so that the result follows by Markov's inequality.

$$\begin{aligned}
& \mathbb{E}\left[\left\|\hat{\Omega}_{w,K} - \Omega_{w,K}\right\|^2\right] \\
= \; & \mathbb{E}\left[\left\|\left(R'_{w,K}R_{w,K}/N_w\right) - \Omega_{w,K}\right\|^2\right] \\
= \; & \mathbb{E}\left[\mathrm{tr}\left(\left(R'_{w,K}R_{w,K}/N_w\right) - \Omega_{w,K}\right)'\left(\left(R'_{w,K}R_{w,K}/N_w\right) - \Omega_{w,K}\right)\right] \\
= \; & \mathbb{E}\left[\mathrm{tr}\left(R'_{w,K}R_{w,K}R'_{w,K}R_{w,K}/N_w^2 - \Omega_{w,K}\left(R'_{w,K}R_{w,K}/N_w\right) - \left(R'_{w,K}R_{w,K}/N_w\right)\Omega_{w,K} + \Omega^2_{w,K}\right)\right] \\
= \; & \mathrm{tr}\left(\mathbb{E}\left[R'_{w,K}R_{w,K}R'_{w,K}R_{w,K}/N_w^2\right] - \Omega_{w,K}\mathbb{E}\left[R'_{w,K}R_{w,K}/N_w\right] - \mathbb{E}\left[R'_{w,K}R_{w,K}/N_w\right]\Omega_{w,K} + \Omega^2_{w,K}\right) \\
= \; & \mathrm{tr}\left(\mathbb{E}\left[R'_{w,K}R_{w,K}R'_{w,K}R_{w,K}/N_w^2\right] - 2\cdot\Omega^2_{w,K} + \Omega^2_{w,K}\right) \\
= \; & \mathrm{tr}\left(\mathbb{E}\left[R'_{w,K}R_{w,K}R'_{w,K}R_{w,K}/N_w^2\right]\right) - \mathrm{tr}\left(\Omega^2_{w,K}\right)
\end{aligned}$$

The second term is

$$\mathrm{tr}(\Omega^2_{w,K}) = \sum_{k=1}^{K}\sum_{l=1}^{K}\left(\mathbb{E}\left[R_{kK}(X)R_{lK}(X)|W=w\right]\right)^2 \tag{A.1}$$

The first term is

$$\begin{aligned}
& \mathrm{tr}\left(\mathbb{E}\left[R'_{w,K}R_{w,K}R'_{w,K}R_{w,K}\right]/N_w^2\right) \\
= \; & \mathbb{E}\left[\sum_{k=1}^{K}\sum_{l=1}^{K}\left(\sum_{i|W_i=w}^{N} R_{kK}(X_i)R_{lK}(X_i)\right)^2\right]/N_w^2 \\
= \; & \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i|W_i=w}^{N}\sum_{j|W_j=w}^{N}\mathbb{E}\left[R_{kK}(X_i)R_{lK}(X_i)R_{lK}(X_j)R_{kK}(X_j)|W=w\right]/N_w^2
\end{aligned}$$

We can then partition this expression into terms with $i=j$,

$$\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i|W_i=w}^{N}\mathbb{E}\left[R_{kK}(X_i)^2R_{lK}(X_i)^2|W=w\right]/N_w^2 \tag{A.2}$$

and with terms $i \neq j$,

$$N_w(N_w-1)\sum_{k=1}^{K}\sum_{l=1}^{K}\left(\mathbb{E}\left[R_{kK}(X)R_{lK}(X)|W=w\right]\right)^2/N_w^2 \tag{A.3}$$

Combining equations (1), (2) and (3) yields,

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\Omega}_{w,K}-\Omega_{w,K}\right\|^2\right] &= \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i|W_i=w}^{N}\mathbb{E}\left[R_{kK}(X_i)^2R_{lK}(X_i)^2|W=w\right]/N_w^2 \\
&\quad +N_w(N_w-1)\sum_{k=1}^{K}\sum_{l=1}^{K}\left(\mathbb{E}\left[R_{kK}(X)R_{lK}(X)|W=w\right]\right)^2/N_w^2 \\
&\quad -\sum_{k=1}^{K}\sum_{l=1}^{K}\left(\mathbb{E}\left[R_{kK}(X)R_{lK}(X)|W=w\right]\right)^2 \\
&= \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i|W_i=w}^{N}\mathbb{E}\left[R_{kK}(X_i)^2R_{lK}(X_i)^2|W=w\right]/N_w^2 \\
&\quad -\sum_{k=1}^{K}\sum_{l=1}^{K}\left(\mathbb{E}\left[R_{kK}(X)R_{lK}(X)|W=w\right]\right)^2/N_w \\
&< \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i|W_i=w}^{N}\mathbb{E}\left[R_{kK}(X_i)^2R_{lK}(X_i)^2|W=w\right]/N_w^2 \\
&= \frac{1}{N_w^2}\sum_{i|W_i=w}^{N}\mathbb{E}\left[\sum_{k=1}^{K}R_{kK}(X_i)^2\sum_{l=1}^{K}R_{lK}(X_i)^2|W=w\right] \\
&\leq \frac{1}{N_w^2}\sum_{i|W_i=w}^{N}\zeta(K)^2\cdot\mathbb{E}\left[\sum_{l=1}^{K}R_{lK}(X_i)^2|W=w\right] \\
&= \frac{1}{N_w^2}\sum_{i|W_i=w}^{N}\zeta(K)^2\cdot\sum_{l=1}^{K}\mathbb{E}\left[R_{lK}(X_i)^2|W=w\right] \\
&= \frac{1}{N_w}\zeta(K)^2\cdot\mathrm{tr}\left(\Omega_{w,K}\right) \\
&\leq \frac{1}{N_w}\zeta(K)^2\cdot K\cdot\lambda_{max}\left(\Omega_{w,K}\right) \\
&\leq C\cdot K\zeta(K)^2/N
\end{aligned}$$

where the fifth line follows by

$$\zeta(K)=\sup_x\|R_K(x)\|=\sup_x\left(\sum_{k=1}^{K}R_{kK}^2(x)\right)^{\frac{1}{2}}$$

which then implies that

$$\sum_{k=1}^{K}R_{kK}^2(x)\leq\zeta(K)^2.$$

The eighth line follows since the maximum eigenvalue of $\Omega_{w,K}$ is $O(1)$ (see below).

For (ii), let us first show that for any two positive semi-definite matrices A and B, and conformable vectors c and d, if $A\geq B$ in a positive semi-definite sense, then for

$$\lambda_{min}(A)=\min_{c'c=1}c'Ac=c^{*\prime}Ac^*,\quad\lambda_{min}(B)=\min_{d'd=1}d'Ad=d^{*\prime}Ad^*,$$

we have that,

$$\lambda_{min}(A) = c^{*\prime}Ac^* \geq c^{*\prime}Bc^* \geq d^{*\prime}Bd^* = \lambda_{min}(B).$$

Now, let $f_w(x) = f_{X|W}(x|W = w)$ and recall that $\Omega_{w,K} = \mathbb{E}\left[R_K(X)R_K(X)'|W = w\right]$ where $\Omega_{1,K}$ is normalized to equal I_K. Next define

$$c(x) = f_0(x)/f_1(x)$$

and note that by Assumptions 8.1 and 8.2 we have that

$$0 < \underline{c} \leq c(x) \leq \bar{c}.$$

Thus we may define $c(x) \equiv \underline{c} + \tilde{c}(x)$ so that,

$$\begin{aligned}
\Omega_{0,K} &= \mathbb{E}\left[R_K(x)R_K(x)'|W = 0\right] \\
&= \int R_K(x)R_K(x)'f_0(x)dx \\
&= \int R_K(x)R_K(x)'c(x)f_1(x)dx \\
&= \int R_K(x)R_K(x)'\left(\underline{c} + \tilde{c}(x)\right)f_1(x)dx \\
&= \underline{c}\int R_K(x)R_K(x)'f_1(x)dx + \int R_K(x)R_K(x)'\tilde{c}(x)f_1(x)dx \\
&= \underline{c}\cdot\Omega_{1,K} + \int R_K(x)R_K(x)'\tilde{c}(x)f_1(x)dx \\
&= \underline{c}\cdot\Omega_{1,K} + \tilde{C}
\end{aligned}$$

Where $\tilde{C}$ is a positive semi-definite matrix. Thus,

$$\Omega_{0,K} \geq \underline{c}\cdot\Omega_{1,K} \Rightarrow \lambda_{min}\left(\Omega_{0,K}\right) \geq \underline{c}\cdot\lambda_{min}\left(\Omega_{1,K}\right) = \underline{c}$$

so that the minimum eigenvalue of $\Omega_{0,K}$ is bounded away from zero. The minimum eigenvalue of $\Omega_{1,K}$ is bounded away from zero by construction. Then note that for a positive definite matrix A, $1/\lambda_{min}(A) = \lambda_{max}(A^{-1})$, so that the eigenvalues of $\Omega_{w,K}$ are also bounded from above.

For (iii) consider the minimum eigenvalue of $\hat{\Omega}_{w,K}$:

$$\begin{aligned}
\lambda_{min}\left(\hat{\Omega}_{w,K}\right) &= \min_{c'c=1} c'\left(\hat{\Omega}_{w,K}\right)c \\
&= \min_{c'c=1}\left(c'\left(\Omega_{w,K}\right)c + c'\left(\hat{\Omega}_{w,K} - \Omega_{w,K}\right)c\right) \\
&\geq \min_{c'c=1} c'\left(\Omega_{w,K}\right)c + \min_{d'd=1} d'\left(\hat{\Omega}_{w,K} - \Omega_{w,K}\right)d \\
&= \lambda_{min}\left(\Omega_{w,K}\right) + \lambda_{min}\left(\hat{\Omega}_{w,K} - \Omega_{w,K}\right) \\
&\geq \lambda_{min}\left(\Omega_{w,K}\right) - \left\|\hat{\Omega}_{w,K} - \Omega_{w,K}\right\| \\
&= \lambda_{min}\left(\Omega_{w,K}\right) - O_p\left(\zeta(K)K^{\frac{1}{2}}N^{-\frac{1}{2}}\right)
\end{aligned}$$

Where the fifth line follows since for a symmetric matrix A

$$\|A\|^2 = \text{tr}\left(A^2\right) \geq \lambda_{min}\left(A^2\right) = \lambda_{min}(A)^2,$$

and since the norm is nonnegative

$$\|A\| \geq -\lambda_{min}(A).$$

The last line follows by part (i). □

Newey (1994) showed that $\zeta(K)$ is $O(K)$, so this lemma implies that if $K^3/N \to 0$ (as implied by Assumption XX), $\|\hat{\Omega}_{w,K} - \Omega_{w,K}\| = o_p(1)$.

Next, define the pseudo true value $\gamma^*_{w,K}$ as

$$\gamma^*_{w,K} \equiv \left(\mathbb{E}[R_K(X)R_K(X)'|W=w]\right)^{-1}\mathbb{E}[R_K(X)Y|W=w] = \Omega^{-1}_{w,K}\mathbb{E}[R_K(X)Y|W=w].$$

and

$$\tilde{\gamma}_{w,K} \equiv \gamma^*_{w,K} + \Omega^{-1}_{w,K}R'_{w,K}\varepsilon_w/N_w$$

where

$$\varepsilon_w \equiv Y_w - \mu_w(\mathbf{X}).$$

Then we can write $\sqrt{N_w}(\tilde{\gamma}_{w,K} - \gamma^*_{w,K})$ as

$$\Omega^{-1}_{w,K}\frac{1}{\sqrt{N_w}}R'_{w,K}\varepsilon_w = \frac{1}{\sqrt{N_w}}\sum_{i|W_i=w}^{N}\Omega^{-1}_{w,K}R_K(X_i)\varepsilon_i$$

with

$$\mathbb{E}[\Omega^{-1}_{w,K}R_K(X_i)\varepsilon_i] = \Omega^{-1}_{w,K}\mathbb{E}\left[R_K(X_i)\mathbb{E}\left[\varepsilon_i|X_i\right]\right] = 0$$

and

$$\mathbb{V}\left[\Omega^{-1}_{w,K}R_K(X_i)\varepsilon_i\right] = \sigma^2_w \cdot \Omega^{-1}_{w,K}$$

Therefore,

$$S_{w,K} \equiv \frac{1}{\sqrt{N_w}}\sum_{i|W_i=w}^{N}\left[\sigma^2_w \cdot \Omega_{w,K}\right]^{-\frac{1}{2}}R_K(X_i)\varepsilon_i \equiv \frac{1}{\sqrt{N_w}}\sum_{i|W_i=w}^{N}Z_i$$

is a normalized summation of N_w independent random vectors distributed with expectation $\mathbf{0}$ and variance-covariance matrix I_K.
Denote the distribution of $S_{w,K}$ by Q_{N_w} and define $\beta_3 \equiv \sum_{i=1}^{N}\mathbb{E}\left\|\frac{Z_i}{\sqrt{N}}\right\|^3$. Then, by Theorem 1.3, Götze (1991), provided that $K \geq 6$,

$$sup_{\mathcal{A}\in A_K}\left|Q_{N_w}(\mathcal{A}) - \Phi(\mathcal{A})\right| \leq C_K\beta_3 N^{-\frac{1}{2}}$$

where A_K is the class of all measurable convex sets in K-dimensional Euclidean space, C_K is $O(K)$, and Φ is a multivariate standard Gaussian distribution.

Lemma A.2 *Suppose that $K(N) = N^\nu$ where $\nu < \frac{2}{11}$. Then,*

$$\sup_{\mathcal{A}\in A_K}\left|Q_{N_w}(\mathcal{A}) - \Phi(\mathcal{A})\right| \to 0$$

Proof: First we will show that β_3 is $O(K^{\frac{9}{2}}N^{-\frac{1}{2}})$

$$\begin{aligned}
\beta_3 \equiv \sum_{i|W_i=w}^{N} \mathbb{E}\left\|\frac{Z_i}{\sqrt{N_w}}\right\|^3 &= N_w^{-\frac{3}{2}} \sum_{i|W_i=w}^{N} \mathbb{E}\left\|\left[\sigma_w^2 \cdot \Omega_{w,K}\right]^{-\frac{1}{2}} R_K(X_i)\varepsilon_i\right\|^3 \\
&= \left(N_w \cdot \sigma_w^2\right)^{-\frac{3}{2}} \sum_{i|W_i=w}^{N} \mathbb{E}\left\|\Omega_{w,K}^{-\frac{1}{2}} R_K(X_i)\varepsilon_i\right\|^3 \\
&\leq \left(N_w \cdot \sigma_w^2\right)^{-\frac{3}{2}} \sum_{i|W_i=w}^{N} \mathbb{E}\left[\|\Omega_{w,K}^{-\frac{1}{2}}\|^3 \, \|R_K(X_i)\varepsilon_i\|^3\right]
\end{aligned}$$

First, consider

$$\|\Omega_{w,K}^{-\frac{1}{2}}\|^3 = \left[\text{tr}(\Omega_{w,K}^{-1})\right]^{\frac{3}{2}} \leq \left[K \cdot \lambda_{max}(\Omega_{w,K}^{-1})\right]^{\frac{3}{2}} \leq C \cdot K^{\frac{3}{2}}$$

which is $O(K^{\frac{3}{2}})$ because $\lambda_{min}(\Omega_{w,K})$ is bounded away from zero by Lemma 0.1. Next, consider

$$\mathbb{E}\left\|R_K(X_i)\varepsilon_i\right\|^3 \leq \sup_x \left\|R_K(x)\right\|^3 \cdot \mathbb{E}\left|\varepsilon_i\right|^3 \leq C \cdot K^3$$

where the third moment of ε_i is bounded by Assumption XX and so the factor is $O(K^3)$. Since σ_w^2 is also bounded by Assumption XX, β_3 is $O(K^{\frac{9}{2}}N^{-\frac{1}{2}})$. Thus,

$$C_K \beta_3 N_w^{-\frac{1}{2}} = K \sum_{i|W_i=w}^{N} \mathbb{E}\left\|\frac{Z_i}{\sqrt{N_w}}\right\|^3 N_w^{-\frac{1}{2}} \leq C \cdot K \cdot K^{\frac{9}{2}} N_w^{-\frac{1}{2}} \cdot N_w^{-\frac{1}{2}} = C \cdot K^{\frac{11}{2}} N^{-1}$$

and the result follows. □

We may proceed further to detail conditions under which the quadratic form, $S_{w,K}'S_{w,K}$, properly normalized, converges to a univariate standard Gaussian distribution. The quadratic form $S_{w,K}'S_{w,K}$ can be written as

$$S_{w,K}'S_{w,K} = \sum_{j=1}^{K}\Big(\frac{1}{\sqrt{N_w}} \sum_{i|W_i=w}^{N} Z_{ij}\Big)^2$$

where Z_{ij} is the j^{th} element of the vector Z_i. Thus, $S_{w,K}'S_{w,K}$ is a sum of K uncorrelated, squared random variables with each random variable converging to a standard Gaussian distribution by the previous result. Intuitively, this sum should converge to a χ^2 random variable with K degrees of freedom.

Lemma A.3 *Under Assumptions XX-XX,*

$$\sup_c \left|\mathrm{P}(S_{w,K}'S_{w,K} \leq c) - \chi_K^2(c)\right| \to 0.$$

Proof: Define the set $A(c) \equiv \left\{S \in \mathbb{R}^K \,|\, S'S \leq c\right\}$. Note that $A(c)$ is a measurable convex set in $\mathbb{R}^K$. Also note that for $Z_K \sim \mathcal{N}(0, I_K)$, we have that $\chi_K^2(c) = \mathrm{P}[Z_K'Z_k \leq c]$. Then,

$$\begin{aligned}
\sup_c \left|\mathrm{P}[S_{w,K}'S_{w,K} \leq c] - \chi_K^2(c)\right| &= \sup_c \left|\mathrm{P}(S_{w,K}'S_{w,K} \leq c) - \mathrm{P}(Z_K'Z_K \leq c)\right| \\
&= \sup_c \left|\mathrm{P}(S_{w,K} \in A(c)) - \mathrm{P}(Z_K \in A(c))\right| \\
&\leq \sup_{\mathcal{A} \in A_K} \left|Q_{N_w}(\mathcal{A}) - \Phi(\mathcal{A})\right| \\
&\leq C_K \beta_3 N^{-\frac{1}{2}} \\
&= O(K^{\frac{11}{2}} N^{-1})
\end{aligned}$$

which is o(1) for $\nu < \frac{2}{11}$ by Lemma 0.2. □

The proper normalization of the quadratic form yields the studentized version, $(S'_{w,K}S_{w,K} - K)/\sqrt{2K}$. This converges to a standard Gaussian distribution by the following lemma.

Lemma A.4 *Under Assumptions XX-XX,*

$$\sup_c \left| \mathrm{P}\left(\frac{S'_{w,K}S_{w,K} - K}{\sqrt{2K}} \leq c \right) - \Phi(c) \right| \to 0.$$

Proof:

$$\begin{aligned} & \sup_c \left| \mathrm{P}\left(\frac{S'_{w,K}S_{w,K} - K}{\sqrt{2K}} \leq c \right) - \Phi(c) \right| \\ = & \sup_c \left| \mathrm{P}\left(S'_{w,K}S_{w,K} \leq K + c\sqrt{2K} \right) - \Phi(c) \right| \\ \leq & \sup_c \left| \mathrm{P}\left(S'_{w,K}S_{w,K} \leq K + c\sqrt{2K} \right) - \chi^2(K + c\sqrt{2K}) \right| + \sup_c \left| \chi^2(K + c\sqrt{2K}) - \Phi(c) \right| \end{aligned}$$

The first term goes to zero by Lemma 0.3. For the second term we may apply the Berry-Esséen Theorem which yields,

$$\sup_c \left| \mathrm{P}\left(\frac{Z'_K Z_K - K}{\sqrt{2K}} \leq c \right) - \Phi(c) \right| \leq C \cdot K^{-\frac{1}{2}}.$$

Thus for $\nu > 0$ the right-hand side converges to zero as well and the result is established. □

In order to proceed we need the following selected results from Imbens, Newey and Ridder (2004). These results establish convergence rates for the estimators of the regression function.

Lemma A.5 (Imbens, Newey and Ridder (2004)): *Suppose Assumptions XX-XX hold. Then,*

(i) there is a sequence $\gamma^0_{w,K}$ *such that*

$$\sup_x \left| \mu_w(x) - R_K(x)'\gamma^0_{w,K} \right| \equiv \sup_x \left| \mu_w(x) - \mu^0_{w,K} \right| = O(K^{-\frac{s}{d}})$$

(ii)

$$\sup_x \left| R_K(x)'\gamma^*_{w,K} - R_K(x)'\gamma^0_{w,K} \right| \equiv \sup_x \left| \mu^*_{w,K} - \mu^0_{w,K} \right| = O_p\left(\zeta(K)^2 K^{-\frac{s}{d}} \right)$$

(iii)

$$\left\| \gamma^*_{w,K} - \gamma^0_{w,k} \right\| = O(\zeta(K) K^{-\frac{s}{d}})$$

(iv)

$$\left\| \hat{\gamma}_{w,K} - \gamma^0_{w,k} \right\| = O_p(K^{\frac{1}{2}} N^{-\frac{1}{2}} + K^{-\frac{s}{d}})$$

The following lemma describes the limiting distribution of the infeasible test statistic.

Lemma A.6 *Under Assumptions XX-XX,*

$$\left(N_w \cdot \left(\hat{\gamma}_{w,K} - \gamma^*_{w,K} \right)' \left(\hat{\sigma}^2_{w,K} \cdot \hat{\Omega}^{-1}_{w,K} \right)^{-1} \left(\hat{\gamma}_{w,K} - \gamma^*_{w,K} \right) - K \right) / \sqrt{2K} \xrightarrow{d} \mathcal{N}(0,1)$$

Proof: We need only show that,

$$\left\|\left[\hat{\sigma}^2_{w,K}\cdot\hat{\Omega}^{-1}_{w,K}\right]^{-\frac{1}{2}}\sqrt{N_w}\left(\hat{\gamma}_{w,K}-\gamma^*_{w,K}\right)-S_{w,K}\right\|=o_p(1).$$

then the result follows by Lemmas 0.2, 0.3, and 0.4.

First, notice that we can rewrite $\hat{\gamma}_{w,K}$ as

$$\hat{\gamma}_{w,K}=\gamma^*_{w,K}+\hat{\Omega}^{-1}_{w,K}R'_{w,K}\varepsilon_{w,K}/N_w$$

where

$$\varepsilon_{w,K}\equiv Y_w-R_{w,K}\gamma^*_{w,K},$$

with ith row equal to

$$\varepsilon_{Ki}=Y_i-R_K(X_i)'\gamma^*_{w,K}.$$

Then,

$$\begin{aligned}
&\left\|\left[\hat{\sigma}^2_{w,K}\cdot\hat{\Omega}^{-1}_{w,K}\right]^{-\frac{1}{2}}\sqrt{N_w}\left(\hat{\gamma}_{w,K}-\gamma^*_{w,K}\right)-S_{w,K}\right\|\\
=\;&\left\|\left[\hat{\sigma}^2_{w,K}\cdot\hat{\Omega}^{-1}_{w,K}\right]^{-\frac{1}{2}}\sqrt{N_w}\cdot\hat{\Omega}^{-1}_{w,K}\cdot R'_{w,K}\varepsilon_{w,K}/N_w-\left[\sigma^2_w\cdot\Omega_{w,K}\right]^{-\frac{1}{2}}\sqrt{N_w}\cdot R'_{w,K}\varepsilon_w/N_w\right\|\\
=\;&\left\|\hat{\sigma}^{-1}_{w,K}\hat{\Omega}^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_{w,K}/\sqrt{N_w}-\sigma^{-1}_w\Omega^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_w/\sqrt{N_w}\right\|\\
=\;&\Big\|\hat{\sigma}^{-1}_{w,K}\hat{\Omega}^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_{w,K}/\sqrt{N_w}-\hat{\sigma}^{-1}_{w,K}\hat{\Omega}^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_w/\sqrt{N_w}\\
&\quad+\hat{\sigma}^{-1}_{w,K}\hat{\Omega}^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_w/\sqrt{N_w}-\sigma^{-1}_w\hat{\Omega}^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_w/\sqrt{N_w}\\
&\quad+\sigma^{-1}_w\hat{\Omega}^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_w/\sqrt{N_w}-\sigma^{-1}_w\Omega^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_w/\sqrt{N_w}\Big\|\\
\le\;&\left\|\hat{\sigma}^{-1}_{w,K}\hat{\Omega}^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_{w,K}/\sqrt{N_w}-\hat{\sigma}^{-1}_{w,K}\hat{\Omega}^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_w/\sqrt{N_w}\right\|\\
&+\left\|\hat{\sigma}^{-1}_{w,K}\hat{\Omega}^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_w/\sqrt{N_w}-\sigma^{-1}_w\hat{\Omega}^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_w/\sqrt{N_w}\right\|\\
&+\left\|\sigma^{-1}_w\hat{\Omega}^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_w/\sqrt{N_w}-\sigma^{-1}_w\Omega^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_w/\sqrt{N_w}\right\|\\
=\;&\left|\hat{\sigma}^{-1}_{w,K}\right|\left\|\hat{\Omega}^{-\frac{1}{2}}_{w,K}R'_{w,K}\left(\varepsilon_{w,K}-\varepsilon_w\right)/\sqrt{N_w}\right\| &\text{(A.4)}\\
&+\left|\hat{\sigma}^{-1}_{w,K}-\sigma^{-1}_w\right|\left\|\hat{\Omega}^{-\frac{1}{2}}_{w,K}\cdot R'_{w,K}\varepsilon_w/\sqrt{N_w}\right\| &\text{(A.5)}\\
&+\left|\sigma^{-1}_w\right|\left\|\left(\hat{\Omega}^{-\frac{1}{2}}_{w,K}-\Omega^{-\frac{1}{2}}_{w,K}\right)R'_{w,K}\varepsilon_w/\sqrt{N_w}\right\| &\text{(A.6)}
\end{aligned}$$

First, consider equation (4),

$$\begin{aligned}
&\left|\hat{\sigma}^{-1}_{w,K}\right|\left\|\hat{\Omega}^{-\frac{1}{2}}_{w,K}R'_{w,K}\left(\varepsilon_{w,K}-\varepsilon_w\right)/\sqrt{N_w}\right\|\\
=\;&\left(\sigma^{-1}_w+o_p\left(N^{-\frac{1}{2}}\right)\right)\cdot\left\|\hat{\Omega}^{-\frac{1}{2}}_{w,K}R'_{w,K}\left(\varepsilon_{w,K}-\varepsilon_w\right)/\sqrt{N_w}\right\|\\
=\;&\left(O\left(1\right)+o_p\left(N^{-\frac{1}{2}}\right)\right)\cdot\left\|\hat{\Omega}^{-\frac{1}{2}}_{w,K}R'_{w,K}\left(\varepsilon_{w,K}-\varepsilon_w\right)/\sqrt{N_w}\right\|
\end{aligned}$$

where

$$\begin{aligned}
& \mathbb{E}\left\|\hat{\Omega}_{w,K}^{-\frac{1}{2}} R_{w,K}' \left(\varepsilon_{w,K}-\varepsilon_w\right)/\sqrt{N_w}\right\|^2 \\
= \;& \mathbb{E}\left[\frac{1}{N_w}\operatorname{tr}\left(\left(\varepsilon_{w,K}-\varepsilon_w\right)' R_{w,K}\hat{\Omega}_{w,K}^{-1} R_{w,K}'\left(\varepsilon_{w,K}-\varepsilon_w\right)\right)\right] \\
= \;& \mathbb{E}\left[\left(\left(\varepsilon_{w,K}-\varepsilon_w\right)' R_{w,K}\left(R_{w,K}'R_{w,K}\right)^{-1} R_{w,K}'\left(\varepsilon_{w,K}-\varepsilon_w\right)\right)\right] \\
\leq \;& \mathbb{E}\left[\left(\varepsilon_{w,K}-\varepsilon_w\right)'\left(\varepsilon_{w,K}-\varepsilon_w\right)\right] \\
= \;& \mathbb{E}\left[\left(\mu_w(\mathbf{X})-R_{w,K}\gamma_{w,K}^*\right)'\left(\mu_w(\mathbf{X})-R_{w,K}\gamma_{w,K}^*\right)\right] \\
\leq \;& N_w\cdot\sup_x\left|\mu_w(x)-R_K(x)'\gamma_{w,K}^*\right|^2 \\
\leq \;& N_w\cdot\sup_x\left(\left|\mu_w(x)-R_K(x)'\gamma_{w,K}^0\right|+\left|R_K(x)'\gamma_{w,K}^0-R_K(x)'\gamma_{w,K}^*\right|\right)^2 \\
= \;& N_w\left(O\left(K^{-\frac{s}{d}}\right)+O\left(\zeta(K)^2K^{-\frac{s}{d}}\right)\right)^2 \\
= \;& O\left(N\right)\cdot\left(O\left(\zeta(K)^2K^{-\frac{s}{d}}\right)\right)^2
\end{aligned}$$

so that equation (4) is $O_p\left(N^{\frac{1}{2}}\zeta(K)^2K^{-\frac{s}{d}}\right)$ by Markov's inequality and consistency of the sample variance. The third line follows since $(I_{N_w}-R_{w,K}(R_{w,K}'R_{w,K})^{-1}R_{w,K}')$ is a projection matrix and so it is positive semi-definite. The seventh line follows from Lemma 0.5 (*i*) and (*ii*).

Now consider equation (5),

$$\left|\hat{\sigma}_{w,K}^{-1}-\sigma_w^{-1}\right|\left\|\hat{\Omega}_{w,K}^{-\frac{1}{2}}\cdot R_{w,K}'\varepsilon_w/\sqrt{N_w}\right\|$$

The first factor is $o_p(N^{-\frac{1}{2}})$ and

$$\begin{aligned}
& \mathbb{E}\left\|\hat{\Omega}_{w,K}^{-\frac{1}{2}}\cdot R_{w,K}'\varepsilon_w/\sqrt{N_w}\right\|^2 \\
= \;& \mathbb{E}\left[\frac{1}{N_w}\operatorname{tr}\left(\varepsilon_w' R_{w,K}\hat{\Omega}_{w,K}^{-1}R_{w,K}'\varepsilon_w\right)\right] \\
= \;& \mathbb{E}\left[\operatorname{tr}\left(\varepsilon_w' R_{w,K}\left(R_{w,K}'R_{w,K}\right)^{-1}R_{w,K}'\varepsilon_w\right)\right] \\
= \;& \mathbb{E}\left[\operatorname{tr}\left(R_{w,K}\left(R_{w,K}'R_{w,K}\right)^{-1}R_{w,K}'\varepsilon_w\varepsilon_w'\right)\right] \\
= \;& \operatorname{tr}\left(\mathbb{E}\left[R_{w,K}\left(R_{w,K}'R_{w,K}\right)^{-1}R_{w,K}'\mathbb{E}\left[\varepsilon_w\varepsilon_w'|\mathbf{X}\right]\right]\right) \\
= \;& \sigma_w^2\cdot\operatorname{tr}\left(\mathbb{E}\left[R_{w,K}\left(R_{w,K}'R_{w,K}\right)^{-1}R_{w,K}'\right]\right) \\
= \;& \sigma_w^2\cdot\mathbb{E}\left[\operatorname{tr}\left(R_{w,K}\left(R_{w,K}'R_{w,K}\right)^{-1}R_{w,K}'\right)\right] \\
= \;& \sigma_w^2\cdot\mathbb{E}\left[\operatorname{tr}\left(\left(R_{w,K}'R_{w,K}\right)^{-1}R_{w,K}'R_{w,K}\right)\right] \\
= \;& \sigma_w^2\cdot\operatorname{tr}\left(I_K\right) \\
= \;& \sigma_w^2\cdot K
\end{aligned}$$

so that the second factor is $O\left(K^{\frac{1}{2}}\right)$. Thus equation (5) is $o_p\left(K^{\frac{1}{2}}N^{-\frac{1}{2}}\right)$.

Finally, consider equation (6),

$$\begin{aligned}
& \left|\sigma_w^{-1}\right|\left\|\left(\hat{\Omega}_{w,K}^{-\frac{1}{2}}-\Omega_{w,K}^{-\frac{1}{2}}\right)R_{w,K}'\varepsilon_w/\sqrt{N_w}\right\| \\
\leq \;& C\cdot\left\|\hat{\Omega}_{w,K}^{-\frac{1}{2}}-\Omega_{w,K}^{-\frac{1}{2}}\right\|\left\|R_{w,K}'\varepsilon_w/\sqrt{N_w}\right\|
\end{aligned}$$

The first factor is $O_p\left(\zeta(K)K^{\frac{1}{2}}N^{-\frac{1}{2}}\right)$ by Lemma 0.1 and the continuous mapping theorem, and

$$\begin{aligned}
& \mathbb{E}\left\|R'_{w,K}\varepsilon_w/\sqrt{N_w}\right\|^2 \\
= \;& \mathbb{E}\left[\frac{1}{N_w}\mathrm{tr}\left(\varepsilon'_w R_{w,K}R'_{w,K}\varepsilon_w\right)\right] \\
= \;& \mathbb{E}\left[\frac{1}{N_w}\mathrm{tr}\left(R'_{w,K}\varepsilon_w\varepsilon'_w R_{w,K}\right)\right] \\
= \;& \mathrm{tr}\left(\frac{1}{N_w}\mathbb{E}\left[R'_{w,K}\mathbb{E}\left[\varepsilon_w\varepsilon'_w|\mathbf{X}\right]R_{w,K}\right]\right) \\
= \;& \sigma^2_w\cdot\mathrm{tr}\left(\mathbb{E}\left[R'_{w,K}R_{w,K}/N_w\right]\right) \\
= \;& \sigma^2_w\cdot\mathrm{tr}\left(\Omega_{w,K}\right) \\
\leq \;& \sigma^2_w\cdot K\cdot\lambda_{max}\left(\Omega_{w,K}\right) \\
\leq \;& C\cdot K
\end{aligned}$$

so that the second factor is $O\left(K^{\frac{1}{2}}\right)$ by Assumption XX, Lemma 0.1 (*ii*) and Markov's inequality. Thus, equation (3) is $O_p\left(\zeta(K)KN^{-\frac{1}{2}}\right)$.

Combining these results yields:

$$\begin{aligned}
& \left\|\left[\hat{\sigma}^2_{w,K}\cdot\hat{\Omega}^{-1}_{w,K}\right]^{-\frac{1}{2}}\sqrt{N_w}\left(\hat{\gamma}_{w,K}-\gamma^*_{w,K}\right)-S_{w,K}\right\| \\
= \;& O_p\left(N^{\frac{1}{2}}\zeta(K)^2K^{-\frac{s}{d}}\right)+o_p\left(K^{\frac{1}{2}}N^{-\frac{1}{2}}\right)+O_p\left(\zeta(K)KN^{-\frac{1}{2}}\right) \\
= \;& O_p\left(N^{\frac{1}{2}}K^{(2-\frac{s}{d})}\right)+o_p\left(K^{\frac{1}{2}}N^{-\frac{1}{2}}\right)+O_p\left(K^2N^{-\frac{1}{2}}\right)
\end{aligned}$$

All three terms are $o_p(1)$ by Assumption XX and for $\frac{s}{d}>\frac{4\nu+1}{2\nu}$. □

Proof of Theorem 8.1: From the previous lemma we have that [10]

$$T^*\equiv\left(N_w\cdot\left((\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})-(\gamma^*_{1,K}-\gamma^*_{0,K})\right)'\cdot\hat{V}^{-1}\cdot\left((\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})-(\gamma^*_{1,K}-\gamma^*_{0,K})\right)-K\right)/\sqrt{2K}$$

converges in distribution to a $\mathcal{N}(0,1)$ random variable, where $\hat{V}$ is defined as

$$\hat{V}\equiv(\hat{\sigma}^2_{0,K}\cdot\hat{\Omega}^{-1}_{0,K}+\hat{\sigma}^2_{1,K}\cdot\hat{\Omega}^{-1}_{1,K}).$$

To complete the proof we must show that under our assumptions $|T^*-T|=o_p(1)$, where T is defined as

$$T\equiv\left(N_w\cdot(\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})'\cdot\hat{V}^{-1}\cdot(\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})-K\right)/\sqrt{2K}.$$

Note that under the null hypothesis $\mu_1(x)=\mu_0(x)$ so we may choose the same approximating sequence $\gamma^0_{1,K}=\gamma^0_{0,K}$ for $\mu^0_{1,K}(x)=\mu^0_{0,K}(x)$. Then,

$$\begin{aligned}
\left\|\gamma^*_{1,K}-\gamma^*_{0,K}\right\| &= \left\|\gamma^*_{1,K}-\gamma^0_{1,K}+\gamma^0_{0,K}-\gamma^*_{0,K}\right\| \\
&\leq \left\|\gamma^*_{1,K}-\gamma^0_{1,K}\right\|+\left\|\gamma^0_{0,K}-\gamma^*_{0,K}\right\| \\
&= O(\zeta(K)K^{-\frac{s}{d}}) \qquad \text{(A.7)}
\end{aligned}$$

[10] For simplicity of notation we assume $N_1=N_0$

by Lemma 0.5 (*iii*), and

$$\begin{aligned}\|\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K}\| &= \left\|\hat{\gamma}_{1,K}-\gamma^{0}_{1,K}+\gamma^{0}_{0,K}-\hat{\gamma}_{0,K}\right\| \\ &\leq \left\|\hat{\gamma}_{1,K}-\gamma^{0}_{1,K}\right\|+\left\|\gamma^{0}_{0,K}-\hat{\gamma}_{0,K}\right\| \\ &= O_p(K^{\frac{1}{2}}N^{-\frac{1}{2}}+K^{-\frac{s}{d}}) \qquad \text{(A.8)}\end{aligned}$$

by Lemma 0.5 (*iv*). So then,

$$\begin{aligned}|T^{*}-T| &= \Big|\Big(N_w\cdot\left((\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})-(\gamma^{*}_{1,K}-\gamma^{*}_{0,K})\right)'\hat{V}^{-1}\left((\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})-(\gamma^{*}_{1,K}-\gamma^{*}_{0,K})\right)-K\Big)/\sqrt{2K} \\ &\quad -\Big(N_w\cdot(\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})'\hat{V}^{-1}(\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})-K\Big)/\sqrt{2K}\Big| \\ &= \frac{N_w}{\sqrt{2K}}\cdot\Big|\left((\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})-(\gamma^{*}_{1,K}-\gamma^{*}_{0,K})\right)'\hat{V}^{-1}\left((\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})-(\gamma^{*}_{1,K}-\gamma^{*}_{0,K})\right) \\ &\quad -(\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})'\hat{V}^{-1}(\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})\Big| \\ &= \frac{N_w}{\sqrt{2K}}\cdot\left|-2\cdot(\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})'\hat{V}^{-1}\left(\gamma^{*}_{1,K}-\gamma^{*}_{0,K}\right)+\left(\gamma^{*}_{1,K}-\gamma^{*}_{0,K}\right)'\hat{V}^{-1}\left(\gamma^{*}_{1,K}-\gamma^{*}_{0,K}\right)\right| \\ &\leq \frac{N_w}{\sqrt{2K}}\cdot\left(2\cdot\left|(\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})'\hat{V}^{-1}\left(\gamma^{*}_{1,K}-\gamma^{*}_{0,K}\right)\right|+\left|\left(\gamma^{*}_{1,K}-\gamma^{*}_{0,K}\right)'\hat{V}^{-1}\left(\gamma^{*}_{1,K}-\gamma^{*}_{0,K}\right)\right|\right)\end{aligned}$$

Consider the first term,

$$\begin{aligned}2\cdot\left|(\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})'\hat{V}^{-1}(\gamma^{*}_{1,K}-\gamma^{*}_{0,K})\right| &= 2\cdot\left|\operatorname{tr}\left((\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K})'\hat{V}^{-1}(\gamma^{*}_{1,K}-\gamma^{*}_{0,K})\right)\right| \\ &\leq 2\cdot\|\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K}\|\cdot\left\|\gamma^{*}_{1,K}-\gamma^{*}_{0,K}\right\|\cdot\lambda_{max}(\hat{V}^{-1}) \\ &\leq C\cdot\|\hat{\gamma}_{1,K}-\hat{\gamma}_{0,K}\|\cdot\left\|\gamma^{*}_{1,K}-\gamma^{*}_{0,K}\right\|+o_p(1) \\ &= \left(O_p(K^{\frac{1}{2}}N^{-\frac{1}{2}}+K^{-\frac{s}{d}})\cdot O(\zeta(K)K^{-\frac{s}{d}})\right)\end{aligned}$$

Where the third line follows from Lemma 0.1 (*iii*) and Assumption XX. The last line follows from equations (7) and (8).

Now, consider the second term,

$$\begin{aligned}\left|(\gamma^{*}_{1,K}-\gamma^{*}_{0,K})'\hat{V}^{-1}(\gamma^{*}_{1,K}-\gamma^{*}_{0,K})\right| &= \left|\operatorname{tr}\left((\gamma^{*}_{1,K}-\gamma^{*}_{0,K})'\hat{V}^{-1}(\gamma^{*}_{1,K}-\gamma^{*}_{0,K})\right)\right| \\ &\leq \left\|\gamma^{*}_{1,K}-\gamma^{*}_{0,K}\right\|^2\cdot\lambda_{max}(\hat{V}^{-1}) \\ &\leq C\cdot\left\|\gamma^{*}_{1,K}-\gamma^{*}_{0,K}\right\|^2+o_p(1) \\ &= O(\zeta(K)^2K^{-\frac{2s}{d}})\end{aligned}$$

Where the third line follows from Lemma 0.1 (*iii*) and Assumption XX. The last line follows from equation (7).

So then,

$$\begin{aligned}|T^{*}-T| &= \frac{N}{\sqrt{2K}}\cdot\left(O_p(K^{\frac{1}{2}}N^{-\frac{1}{2}}+K^{-\frac{s}{d}})\cdot O(\zeta(K)K^{-\frac{s}{d}})+O(\zeta(K)^2K^{-\frac{2s}{d}})\right) \\ &= O_p\left(N^{\frac{1}{2}}\zeta(K)K^{-\frac{s}{d}}\right)+O_p\left(N\zeta(K)K^{-(\frac{1}{2}+\frac{2s}{d})}\right)+O\left(N\zeta(K)^2K^{-(\frac{1}{2}+\frac{2s}{d})}\right)\end{aligned}$$

For $\frac{s}{d}>\frac{2\nu+1}{2\nu}$ all three terms are $o_p(1)$ and the result follows. □

Table 1: Covariate Balance for Lalonde Data

	mean	stand.	mean		Normalized Dif. in Treat. and Contr. Ave's				
		dev.	contr.	treat.	all	[t-stat]	$a < e(x)$	optimal	prop score
							$< 1 - a$	weights	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

References

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

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.

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.

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

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.

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.

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.

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.

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 4: Covariate Balance for Lottery Data

	mean	standard	mean	mean	Normalized Dif. in Treat. and Contr. Ave's				
		deviation	controls	treated	all	[tstat]	$a < e(x)$	optimal	prop score
							$< 1 - a$	weights	weighted
year won	6.23	1.18	6.38	6.06	-0.27	[-3.0]	-0.19	-0.18	-0.19
# tickets bought	3.33	2.86	2.19	4.57	0.83	[9.9]	0.42	0.42	0.86
education	13.73	2.20	14.43	12.97	-0.66	-7.8]	-0.47	-0.42	-0.46
work then	0.78	0.41	0.77	0.80	0.08	[0.9]	-0.03	-0.01	0.02
male	0.63	0.48	0.67	0.58	-0.19	[-2.1]	-0.12	-0.10	-0.13
age won	50.22	13.68	53.21	46.95	-0.46	[-5.2]	-0.26	-0.22	-0.38
earn -6	0.01	0.01	0.02	0.01	-0.27	[-3.0	-0.14	-0.15	-0.19
earn -5	0.01	0.01	0.02	0.01	-0.28	[-3.2	-0.17	-0.18	-0.21
earn -4	0.01	0.01	0.02	0.01	-0.30	[-3.6	-0.21	-0.20	-0.25
earn -3	0.01	0.01	0.02	0.01	-0.26	[-2.9	-0.20	-0.19	-0.21
earn -2	0.02	0.02	0.02	0.01	-0.27	[-3.0]	-0.21	-0.20	-0.20
earn -1	0.02	0.02	0.02	0.01	-0.22	[-2.5]	-0.19	-0.18	-0.17
work -6	0.69	0.46	0.69	0.70	0.03	[0.3]	0.07	0.02	0.05
work -5	0.71	0.45	0.68	0.74	0.14	[1.6]	0.10	0.09	0.12
work -4	0.71	0.45	0.69	0.73	0.09	[1.1]	0.02	0.05	0.10
work -3	0.70	0.46	0.68	0.73	0.13	[1.4]	0.03	0.05	0.11
work -2	0.71	0.46	0.68	0.74	0.15	[1.6]	0.06	0.06	0.15
work -1	0.71	0.45	0.69	0.74	0.10	[1.2]	0.03	0.01	0.17
log odds ratio	0.01	1.97	-1.12	1.25	1.20	[16.4]	0.72	0.67	1.03

Table 5: Subsample Sizes for Lottery Data: Propensity Score Threshold 0.0914

	$e(x) < a$	$a \leq e(x) \leq 1 - a$	$1 - a < e(x)$	all
controls	37	216	6	259
treated	4	172	61	237
all	41	388	67	496

Table 6: Asymptotic Standard Errors for Lottery Data

	ATE	OSATE	OWATE	ATT
Asymptotic Standard Error	1.6199	2.7586	1.0918	1.0055
Ratio to All	1.0000	1.7029	0.6740	0.6207

Table 7: Covariate Balance for Gain Data

	mean	standard	mean	mean	Normalized Dif. in Treat. and Contr. Ave's				
		deviation	controls	treated	all	[tstat]	$a < e(x)$	optimal	prop score
							$< 1 - a$	weights	weighted
earn q-1	268	974	214	423	0.21	[5.1]	0.21	0.17	0.24
earn q-2	297	1033	219	521	0.29	[6.8]	0.28	0.26	0.39
earn q-3	307	1049	221	554	0.32	[7.1]	0.30	0.27	0.46
earn q-4	292	1010	208	533	0.32	[7.3]	0.31	0.29	0.47
earn y-2	1166	3697	750	2363	0.44	[9.2]	0.42	0.39	0.72
earn y-3	595	2037	363	1262	0.44	[9.1]	0.42	0.39	0.75
unempl q-1	0.85	0.36	0.88	0.77	-0.30	[-7.4]	-0.28	-0.27	-0.31
unempl q-2	0.85	0.36	0.88	0.76	-0.32	[-8.0]	-0.30	-0.30	-0.38
unempl q-3	0.84	0.36	0.87	0.76	-0.32	[-8.0]	-0.30	-0.29	-0.39
unempl q-4	0.84	0.36	0.88	0.75	-0.35	[-8.7]	-0.33	-0.32	-0.43
unempl y-2	0.73	0.44	0.78	0.59	-0.42	[-10.8]	-0.38	-0.37	-0.47
unempl y-3	0.81	0.39	0.85	0.69	-0.40	[-9.9]	-0.37	-0.37	-0.50
education	8.62	5.01	8.18	9.87	0.34	[10.8]	0.21	0.21	0.18
age	37.28	8.68	38.48	33.82	-0.54	[-15.4]	-0.39	-0.42	-0.43
log odds ratio	-1.20	0.82	-1.35	-0.75	0.73	[20.2]	0.59	0.59	0.76

Table 8: Subsample Sizes for Gain Data: Propensity Score Threshold 0.0932

	$e(x) < a$	$a \leq e(x) \leq 1 - a$	$1 - a < e(x)$	all
controls	366	2629	0	2995
treated	39	999	2	1040
all	405	3628	2	4035

Table 9: Asymptotic Standard Errors for Gain Data

	ATE	ATT	OSATE	OWATE
Asymptotic Standard Error	0.1326	0.1286	0.1283	0.1211
Ratio to All	1.0000	0.9697	0.9676	0.9130