DEPARTMENT OF ECONOMICS
WORKING PAPERS

economics.ceu.hu

Inverse Propensity Score Weighted Estimation of Local Average Treatment Effects and a Test of the Unconfoundedness Assumption

by

Stephen G. Donald[1], Yu-Chin Hsu[2] and Robert P. Lieli[3]

2012/9

[1] Department of Economics, University of Texas, Austin.
[2] Department of Economics, University of Missouri, Columbia
[3] Department of Economics, Central European University, Budapest and the National Bank of Hungary, lielir@ceu.hu

Abstract

We propose inverse probability weighted estimators for the the local average treatment effect (LATE) and the local average treatment effect for the treated (LATT) under instrumental variable assumptions with covariates. We show that these estimators are asymptotically normal and efficient, and provide a higher order asymptotic mean squared error expansion for the LATE estimator. When the (binary) instrument satisfies a condition called one-sided non-compliance, we propose a Hausman-type test of whether treatment assignment is unconfounded conditional on some observables. The test is based on the fact that under one-sided non-compliance LATT coincides with the average treatment effect for the treated. We evaluate the effect of JTPA training programs on the earnings of participants to illustrate our methods. The unconfoundedness test suggests that treatment assignment among males is based partly on unobservables. In contrast, the hypothesis of random treatment assignment cannot be rejected among females.

Keywords: local average treatment effect, inverse probability weighted estimator, nonparametric estimation, unconfoundedness

Acknowledgements

We thank Jason Abrevaya, Yu-Wei Hsieh, Chung-Ming Kuan, Blaise Melly, Chris Taber and Ed Vytlacil for useful comments. All remaining errors are our responsibility

1 Introduction

Nonparametric estimation of average treatment effects from observational data is typically undertaken under one of two types of identifying conditions. The unconfoundedness assumption, in its weaker form, postulates that treatment assignment is mean-independent of potential outcomes conditional on a vector of observed covariates. The requirement that given these observables no other unobserved factor acts as a confounder for the mean effect is still a strong one and carries with it considerable identifying power. In particular, the average treatment effect (ATE) and the average treatment effect for the treated (ATT) are nonparametrically identified under this assumption. On the other hand, if unobservable confounders exist then instrumental variables—related to the outcome only through changing the likelihood of treatment—are typically utilized to learn about treatment effects. Uncoupled with additional structural assumptions, the availability of an instrumental variable (IV) is however not sufficient to identify ATE or ATT. In general, the IV will only identify the local average treatment effect (LATE; Imbens and Angrist 1994) and the local average treatment effect for the treated (LATT; Frölich and Lechner 2006; Hong and Nekipelov 2008). If one specializes to binary instruments, as we do in this paper, then the LATE and LATT parameters correspond to the average treatment effect over specific subgroups of the population. These subgroups are however dependent on the choice of the instrument and are generally unobservable. Partly for these reasons a number of authors have called into question the usefulness of LATE for program evaluation (Heckman 1997; Heckman and Urzúa 2009; Deaton 2009). In most such settings ATE and ATT are more natural and practically relevant parameters of interest—provided that they can be credibly identified and accurately estimated.[1]

When using instrumental variables, empirical researchers are often called upon to tell a "story" to justify their validity. As pointed out by Abadie (2003) and Frölich (2007), it is often easier to argue that the relevant IV conditions hold if conditioning on a vector of observable covariates is allowed. In particular, Frölich (2007) shows that in this scenario LATE is still nonparametrically identified[2] and proposes efficient estimators, based on nonparametric imputation and matching, for this quantity. Given the possible need to condition on a vector of observables to justify the IV

[1] In fairness, some of the criticism in Deaton (2009) goes beyond LATE, and also applies to ATE/ATT as a parameter of interest. See Imbens (2009) for a response to Deaton (2009).

[2] Of course, LATE will be nonparametrically identified in the subpopulations defined by the possible values of the covariates. The point is that it is also unconditionally identified.

assumptions, it is natural to ask whether treatment assignment itself might be unconfounded conditional on the same (or maybe a larger or smaller) vector of covariates. In this paper we propose a formal test of this hypothesis that relies on the availability of a specific kind of binary instrument for which LATT=ATT (so that the latter parameter is also identified). Establishing unconfoundedness under these conditions still offers at least two benefits: (*i*) it enables the estimation of an additional parameter of interest (namely, ATE) and (*ii*) it generally allows more efficient estimation of ATT than IV methods (we will argue this point in more detail later). To our knowledge this is the first test in the literature aimed at this task.

More specifically, the contributions of this work are threefold. Firstly, given a (conditionally) valid binary instrument, we propose alternative nonparametric IV estimators of LATE and LATT. These estimators rely on weighting by the estimated propensity score and are computed as the ratio of two estimators that are of the form proposed by Hirano et al. (2003), henceforth HIR. While Frölich (2007) conjectures in passing that such an estimator of LATE should be efficient, he does not provide a proof. We fill this gap in the literature and formally establish the first order asymptotic equivalence of our LATE estimator and Frölich's imputation/matching-based estimators (these are given by the ratio of estimators proposed by Hahn 1998). We also demonstrate that our LATT estimator is asymptotically efficient, i.e. first-order equivalent to that of Hong and Nekipelov (2008).

Secondly, we go beyond first order asymptotics and undertake a more careful comparison of LATE estimators based on imputation vs. inverse probability weighting through higher order mean square expansions. These expansions build on and extend previous work by Kalyanaraman (2009), Ichimura and Linton (2002) and Heckman et al. (1998), and can be employed to guide bandwidth selection in finite samples. We show that neither estimator dominates in terms of mean squared error, and offer some insight as to when one might outperform the other in finite samples. A criterion for choosing between the two estimators in practice is also provided. The theory is illustrated by a number of Monte Carlo exercises.

Thirdly, and most importantly, we propose a Hausman-type test for the unconfoundedness assumption. On the one hand, if a binary instrument satisfying "one-sided non-compliance" (Frölich and Melly 2008a) is available, then the LATT parameter associated with that instrument coincides with ATT, and is consistently estimable using the estimator we proposed. (Whether one-sided non-compliance holds is verifiable from the data.) On the other hand, if treatment assignment is

unconfounded given a vector of covariates, ATT can also be consistently estimated using the HIR estimator. If the unconfondedness assumption does not hold, then the HIR estimator will generally converge to a different limit. Thus, the unconfoundedness assumption can be tested by comparing our estimator of LATT with HIR's estimator of ATT. Of course, if the validity of the instrument itself is questionable, then the test should be more carefully interpreted as a joint test of the IV conditions and the unconfoundedness assumption. We investigate the finite sample size and power properties of our test using Monte Carlo simulations and apply it to data collected on training programs administered under the Job Training Partnership Act (JTPA).[3] Of interest is the effect of program participation on subsequent earnings, and eligibility serves as a binary instrument. For men we can strongly reject the hypothesis that the participation decision and potential earnings are unconfounded conditional on an observed vector of covariates. In contrast, the hypothesis of random treatment assignment cannot be rejected for women. Therefore, one can consistently estimate the average treatment effect for women but not for men.

The rest of the paper is organized as follows. In Section 2 we present a framework for defining and identifying causal effects nonparametrically. In Section 3 we propose nonparametric estimators of LATE and LATT, describe their asymptotic properties (first and second order), and compare our LATE estimator to Frölich's. The test for the unconfoundedness assumption is presented in Section 4, and the implications of unconfoundedness are discussed in more detail. Section 5 contains Monte Carlo simulations designed to illustrate the theoretical results in the paper, and the empirical application is given in Section 6. Section 7 summarizes and concludes. Proofs are collected in a technical appendix.

2 The basic framework and identification results

The following IV framework, augmented by covariates, is now standard in the treatment effect literature; see, e.g., Abadie (2003) or Frölich (2007) for a more detailed exposition. For each population unit (individual) one can observe the value of a binary instrument $Z \in \{0,1\}$ and a vector of covariates $X \in \mathbb{R}^k$. For $Z = z$, the random variable $D(z) \in \{0,1\}$ specifies individuals' potential treatment status with $D(z) = 1$ corresponding to treatment and $D(z) = 0$ to no treatment. The actually observed treatment status is then given by $D \equiv D(Z) = D(1)Z + D(0)(1 - Z)$.

[3]The same data set is analyzed by Abadie et al. (2002) and Frölich and Melly (2008b) among others.

Similarly, the random variable $Y(z,d)$ denotes the potential outcomes in the population that would obtain if one were to set $Z = z$ and $D = d$ exogenously. The following assumptions, taken from Abadie (2003) and Frölich (2007) with some modifications, describe the relationships between the variables defined above and justify Z being referred to as an instrument:

ASSUMPTION 1 Let $V = \big(Y(0,0), Y(0,1), Y(1,0), Y(1,1), D(1), D(0)\big)'$. There exists a subset X_1 of X such that

(i) (Moments): $E(V'V \mid X_1) < \infty$.

(ii) (Instrument assignment): $E(V \mid Z, X_1) = E(V \mid X_1)$ and $E(VV' \mid Z, X_1) = E(V \mid X_1)$.

(iii) (Exclusion of the instrument): $P[Y(1,d) = Y(0,d) \mid X_1] = 1$ for $d \in \{0,1\}$.

(iv) (First stage): $P[D(1) = 1 \mid X_1] > P[D(0) = 1 \mid X_1]$ and $0 < P(Z = 1 \mid X_1) < 1$.

(v) (Monotonicity): $P[D(1) \geq D(0) \mid X_1] = 1$.

Assumption $1(i)$ ensures the existence of the moments we will work with. Part (ii) states that, conditional on X_1, the instrument is exogenous with respect to the first and second moments of the potential outcome and treatment status variables. This is satisfied, for example, if the value of the instrument is completely randomly assigned. Part (iii) precludes the instrument from having a direct effect on potential outcomes. Part (iv) postulates that the instrument is (positively) related to the probability of being treated and implies that the distributions $X_1|Z = 0$ and $X_1|Z = 1$ have common support. Finally, the monotonicity of $D(z)$ in z, required in part (v), allows for three different types of population units with nonzero mass: compliers $[D(0) = 0, D(1) = 1]$, always takers $[D(0) = 1, D(1) = 1]$ and never takers $[D(0) = 0, D(1) = 0]$ (cf. Imbens and Angrist 1994). Of these, compliers are actually required to have positive mass—part (iv) rules out $P[D(1) = D(0)] = 1$. In light of these assumptions it is customary to think of Z as a variable that indicates whether an exogenous incentive to obtain treatment is present or as a variable signaling "intention to treat". We will denote the conditional probability $P(Z = 1 \mid X_1)$ by $q(X_1)$ and refer to it as the propensity score.

Given the exclusion restriction in part (iii), one can simplify the definition of the potential outcome variables as $Y(d) \equiv Y(1,d) = Y(0,d)$, $d = 0, 1$. The actually observed outcomes are then given by $Y \equiv Y(D) = Y(1)D + Y(0)(1 - D)$. The LATE ($\equiv \tau$) and LATT ($\equiv \tau_t$) parameters associated with the instrument Z are defined as

$$\begin{aligned} \tau &\equiv E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0] \\ \tau_t &\equiv E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0, D = 1]. \end{aligned}$$

LATE, originally due to Imbens and Angrist (1994), is the average treatment effect in the complier subpopulation. The LATT parameter was considered, for example, by Frölich and Lechner (2006) and Hong and Nekipelov (2008). LATT is the average treatment effect among those compliers who actually receive the treatment. Of course, in the subpopulation of compliers the condition $D = 1$ is equivalent to $Z = 1$, i.e. LATT can also be written as $E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0, Z = 1]$. In particular, if Z is an instrument that satisfies Assumption 1 unconditionally (say Z is assigned completely at random), then LATT coincides with LATE. While LATT may well be an interesting parameter in its own right, our interest in it is motivated mainly by the fact that it can serve as a bridge between the IV assumptions and unconfoundedness (this connection will be developed shortly).

Under Assumption 1 one can also interpret LATE/LATT as the ATE/ATT of Z on Y divided by the ATE/ATT of Z on D. More formally, define $W(z) \equiv D(z)Y(1) + (1 - D(z))Y(0)$ and $W \equiv W(Z) = ZW(1) + (1 - Z)W(0)$. It is easy to see that $W = DY(1) + (1 - D)Y(0) = Y$ and, as we show in Appendix A,

$$\tau = E[W(1) - W(0)] \Big/ E[D(1) - D(0)] \tag{1}$$

$$\tau_t = E[W(1) - W(0) \mid Z = 1] \Big/ E[D(1) - D(0) \mid Z = 1]. \tag{2}$$

The quantities on the rhs of (1) and (2) are nonparametrically identified from the joint distribution of the observables (Y, D, Z, X_1). In particular, we present the following identification result:

Theorem 1 *Under Assumption 1,*

$$E[W(1) - W(0)] = E\left[\frac{ZY}{q(X_1)} - \frac{(1-Z)Y}{1-q(X_1)}\right] \equiv \Delta \tag{3}$$

$$E[D(1) - D(0)] = E\left[\frac{ZD}{q(X_1)} - \frac{(1-Z)D}{1-q(X_1)}\right] \equiv \Gamma \tag{4}$$

$$E[W(1) - W(0)|Z = 1] = E\left[q(X_1)\left(\frac{ZY}{q(X_1)} - \frac{(1-Z)Y}{1-q(X_1)}\right)\right] \Big/ E[q(X_1)] \equiv \Delta_t \tag{5}$$

$$E[D(1) - D(0)|Z = 1] = E\left[q(X_1)\left(\frac{ZD}{q(X_1)} - \frac{(1-Z)D}{1-q(X_1)}\right)\right] \Big/ E[q(X_1)] \equiv \Gamma_t. \tag{6}$$

That is, $\tau = \Delta/\Gamma$ *and* $\tau_t = \Delta_t/\Gamma_t$.

These results are not entirely new in the literature—they are implied by, for example, Theorem 3.1 in Abadie (2003). The equality $\tau = \Delta/\Gamma$ is also stated by Frölich (2007). While the result $\tau_t = \Delta_t/\Gamma_t$ is strictly speaking new, it is not very surprising. For easy reference, a proof for Theorem 1 is provided in Appendix A.

We now turn to the discussion of unconfoundedness, also termed "ignorable treatment assignment" by Rubin (1978). We say that treatment assignment is unconfounded conditional on a subset X_2 of the vector X if

ASSUMPTION 2 (Unconfoundedness): $Y(1)$ and $Y(0)$ are mean-independent of D conditional on X_2, i.e. $E[Y(d) \mid D, X_2] = E[Y(d) \mid X_2]$, $d \in \{0, 1\}$.

Assumption 2 is stronger than Assumption 1 in the sense that it rules out selection to treatment based on unobservable factors and permits nonparametric identification of ATE $= E[Y(1) - Y(0)]$ and ATT $= E[Y(1) - Y(0) \mid D = 1]$.[4] As mentioned above, these parameters are often of more interest to decision makers than local treatment effects, but are not generally identified under Assumption 1 alone. A partial exception is when the instrument Z satisfies a strengthening of the monotonicity property called one-sided non-compliance (Frölich and Melly 2008a):

ASSUMPTION 3 (One-sided non-compliance): $P[D(0) = 0] = 1$.

Assumption 3 means that those individuals for whom $Z = 0$ are excluded from the treatment group, while those for whom $Z = 1$ generally have the option to accept or decline treatment (e.g., Z might represent eligibility to receive treatment). Hence, there are no always-takers; non-compliance with the intention-to-treat variable Z is only possible when $Z = 1$. More formally, for such an instrument $D = ZD(1)$, and so $D = 1$ implies $D(1) = 1$ (the treated are a subset of the compliers). Therefore,

$$\begin{aligned} \text{LATT} &= E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0, D = 1] \\ &= E[Y(1) - Y(0) \mid D(1) = 1, D = 1] \\ &= E[Y(1) - Y(0) \mid D = 1] = \text{ATT}. \end{aligned}$$

Thus, under one-sided non-compliance, ATT=LATT. The ATE parameter, on the other hand, remains generally unidentified under Assumptions 1 and 3 alone.

In Section 4 we will show how one can test Assumption 2 when a binary instrument, valid conditional on X_1 and satisfying one-sided non-compliance, is available. Frölich and Lechner (2006) also consider some consequences for identification of the IV assumption and unconfoundedness

[4]For example, ATE is identified by an expression analogous to (5): replace Z with D and $q(X_1)$ with $p(X_2) = P(D = 1 \mid X_2)$. Similarly, ATT is identified by an expression analogous to (6): make the same substitutions as above.

holding simultaneously (without one-sided non-compliance), but they do not discuss estimation by inverse probability weighting, propose a test, or draw out implications for efficiency.

3 The estimators and their asymptotic properties

3.1 Inverse propensity weighted estimators of LATE and LATT

Let $\{(Y_i, D_i, Z_i, X_{1i})\}_{i=1}^n$ denote a random sample of observations on (Y, D, Z, X_1). The proposed inverse probability weighted estimators for τ and τ_t are based on sample analog expressions for (3) through (6):

$$\hat{\tau} = \sum_{i=1}^{n} \left\{ \frac{Z_i Y_i}{\hat{q}(X_{1i})} - \frac{(1-Z_i)Y_i}{1-\hat{q}(X_{1i})} \right\} \Big/ \sum_{i=1}^{n} \left\{ \frac{Z_i D_i}{\hat{q}(X_{1i})} - \frac{(1-Z_i)D_i}{1-\hat{q}(X_{1i})} \right\},$$

$$\hat{\tau}_t = \sum_{i=1}^{n} \hat{q}(X_{1i}) \left\{ \frac{Z_i Y_i}{\hat{q}(X_{1i})} - \frac{(1-Z_i)Y_i}{1-\hat{q}(X_{1i})} \right\} \Big/ \sum_{i=1}^{n} \hat{q}(X_{1i}) \left\{ \frac{Z_i D_i}{\hat{q}(X_{1i})} - \frac{(1-Z_i)D_i}{1-\hat{q}(X_{1i})} \right\},$$

where $\hat{q}(\cdot)$ is a suitable nonparametric estimator of the the propensity score function $q(x_1) = E(Z \mid X_1 = x_1)$. In this paper we use local polynomial regression to estimate $q(\cdot)$. The first order asymptotic results presented in Section 3.2 do not depend critically on this choice—the same conclusions could be obtained under similar conditions if other estimators of $q(\cdot)$ were used instead.[5] In contrast, the second order results presented in Section 3.3 are specific to this estimator.

The local polynomial regression estimator of a conditional mean function solves a weighted least squares problem at each point of evaluation. These regressions are local in the sense that the weights assigned to observations decrease rapidly with the distance from the point of evaluation. For example, if X_1 is a scalar, then for any given point x_1 in the support of X_1, $q(x_1)$ can be estimated by the constant term $\hat{\beta}_0$ in a regression of the form

$$\min_{\hat{\beta}_0,\ldots,\hat{\beta}_p} \sum_{i=1}^{n} K\left(\frac{X_{1i}-x_1}{h}\right) \Big[Z_i - \hat{\beta}_0 - \hat{\beta}_1(X_{1i}-x_1) - \ldots - \hat{\beta}_p(X_{1i}-x_1)^p \Big]^2, \tag{7}$$

where $K(\cdot)$ is a weighting function (kernel) and h is a smoothing parameter (bandwidth). The parameter p is referred to as the order of the estimator. If X_1 is higher dimensional then all powers

[5] For example, HIR employ the series logit estimator. A minor disadvantage of local polynomial regression is that $\hat{q}$ is not necessarily bounded between 0 and 1. Therefore, a trimmed estimate might be preferred in practice. As far as asymptotic theory is concerned, we will handle this issue by assuming that the propensity score is bounded away from 0 and 1. This is mainly for convenience—we could consider an explicit trimming rule instead at the expense of complicating the exposition.

of all components of $X_1 - x_1$ up to order p, as well as all unique cross-products up to order p, are included in (7) as additional regressors.

The estimators $\hat{\tau}$ and $\hat{\tau}_t$ call for estimating the propensity score at each sample observation X_{1i} rather than at fixed points in the support of X_1. In computing $\hat{q}(X_{1i})$, we use the leave-one-out version of the estimator, i.e. the sum in (7) runs over all observations except the ith one. Some of the more delicate second-order asymptotic results are dependent on this construction.

3.2 First order asymptotic results

We now state conditions under which $\hat{\tau}$ and $\hat{\tau}_t$ are $\sqrt{n}$-consistent, asymptotically normal and efficient.

ASSUMPTION 4 (Distribution of X_1): (i) The distribution of the r-dimensional vector X_1 is absolutely continuous with probability density $f(x)$; (ii) the support of X_1, denoted $\mathcal{X}$, is a Cartesian product of compact intervals; (iii) $f(x)$ is twice continuously differentiable, bounded above, and bounded away from 0 on $\mathcal{X}$.

Though standard in the literature, this assumption is fairly restrictive in that it rules out discrete covariates. This is mostly a matter of convenience; working with continuous X_1 makes the use of standard nonparametric regression techniques straightforward. Discrete variables could also be allowed in X_1 at the expense of more cumbersome notation and the modification of some of the technical conditions. Alternatively, one can partition the population by the possible values of the discrete covariates and carry out the analysis in each subpopulation. We will demonstrate how to implement this approach in the empirical section of this paper.

Next we impose restrictions on various conditional moments of Y, D and Z, starting with the propensity score function.

ASSUMPTION 5 (Propensity Score): (i) $q(x_1)$ is continuously differentiable of order $\bar{q} > r$; (ii) $q(x_1)$ is bounded away from zero and one on $\mathcal{X}$.

In addition, we define the following conditional moments: $m_z(x_1) = E(Y \mid X_1 = x_1, Z = z)$, $\mu_z(x_1) = E(D \mid X_1 = x_1, Z = z)$, $s_z^2(x_1) = \text{var}(Y \mid X_1 = x_1, Z = z)$, $\sigma_z^2(x_1) = \text{var}(D \mid X_1 = x_1, Z = z)$, and $c_z(x_1) = \text{cov}(Y_z, D_z \mid X_1 = x_1, Z = z)$. Then:

ASSUMPTION 6 (Conditional Moments of Y and D): $m_z(x_1)$, $\mu_z(x_1)$, $s_z^2(x_1)$, $\sigma_z^2(x_1)$, $c_z(x_1)$ are continuously differentiable over $\mathcal{X}$ for $z = 0, 1$.

The last two assumptions specify the estimator used for the propensity score function.

ASSUMPTION 7 (Kernel Function): (*i*) The kernel function $K(u)$ is supported on $[-1,1]^r$ and is symmetric in each argument; (*ii*) $K(u)$ is continuously differentiable in each argument.

ASSUMPTION 8 (Propensity Score Estimator): The propensity score function is estimated by leave-one-out local polynomial regression of order r with the bandwidth sequence h_n satisfying $nh_n^{2r+2} \to 0$ and $nh_n^{2r}/\log n \to \infty$.

The first-order asymptotic properties of $\hat{\tau}$ and $\hat{\tau}_t$ are stated in the following theorem.

Theorem 2 *(Asymptotic properties of $\hat{\tau}$ and $\hat{\tau}_t$): Suppose that Assumption 1 and Assumptions 4 through 8 are satisfied. Then:*

(a) $\sqrt{n}(\hat{\tau}-\tau) \xrightarrow{d} \mathcal{N}(0,\mathcal{V})$ *and* $\sqrt{n}(\hat{\tau}_t-\tau_t) \xrightarrow{d} \mathcal{N}(0,\mathcal{V}_t)$, *where* $\mathcal{V} = E[\psi^2(Y,D,Z,X_1)]$ *and* $\mathcal{V}_t = E[\psi_t^2(Y,D,Z,X_1)]$ *with the functions* ψ *and* ψ_t *given by*

$$\begin{aligned}&\psi(y,d,z,x_1)\\ &\quad=\frac{1}{\Gamma}\left\{\frac{z\big[y-m_1(x_1)-\tau(d-\mu_1(x_1))\big]}{q(x_1)}-\frac{(1-z)\big[y-m_0(x_1)-\tau(d-\mu_0(x_1))\big]}{1-q(x_1)}\right.\\ &\qquad\left.+m_1(x_1)-m_0(x_1)-\tau[\mu_1(x)-\mu_0(x_1)]\right\},\end{aligned}$$

and

$$\begin{aligned}&\psi_t(y,d,z,x_1)\\ &\quad=\frac{q(x_1)}{Q\Gamma_t}\left\{\frac{z\big[y-m_1(x_1)-\tau_t(d-\mu_1(x))\big]}{q(x_1)}-\frac{(1-z)\big[y-m_0(x_1)-\tau_t(d-\mu_0(x_1))\big]}{1-q(x_1)}\right.\\ &\qquad\left.+\frac{z\big[m_1(x_1)-m_0(x_1)-\tau_t(\mu_1(x_1)-\mu_0(x_1))\big]}{q(x_1)}\right\},\end{aligned}$$

with $Q = E(Z)$.

(b) $\mathcal{V}$ *is equal to the semiparametric efficiency bound for LATE without the knowledge of* $q(x_1)$

(c) $\mathcal{V}_t$ *is equal to the semiparametric efficiency bound for LATT without the knowledge of* $q(x_1)$.

Theorem 2 follows from the fact that, under the conditions stated, $\hat{\tau}$ and $\hat{\tau}_t$ can be expressed as asymptotically linear with influence functions ψ and ψ_t, respectively:

$$\sqrt{n}(\hat{\tau} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i, D_i, Z_i, X_{1i}) + o_p(1), \tag{8}$$

$$\sqrt{n}(\hat{\tau}_t - \tau_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_t(Y_i, D_i, Z_i, X_{1i}) + o_p(1). \tag{9}$$

These representations are developed in Appendix B. The semiparametric efficiency bounds referenced in part (b) and (c) of Theorem 2 are derived in Frölich (2007) and Hong and Nekipelov (2008), respectively. To use Theorem 2 for statistical inference, one needs consistent estimators for $\mathcal{V}$ and $\mathcal{V}_t$. Such estimators can be obtained by constructing (uniformly) consistent estimates for ψ and ψ_t and then averaging the squared estimates over the sample observations $\{(Y_i, D_i, Z_i, X_{1i})\}_{i=1}^n$. In estimating ψ and ψ_t, one replaces the functions $m_z(x_1)$, $\mu_z(x_1)$ and $q(x_1)$ with nonparametric estimators that are uniformly consistent over $\mathcal{X}$. The quantities Γ, Γ_t and Q are also replaced with sample analogs.

Theorem 2 shows that the inverse probability weighted estimators of LATE and LATT presented in this paper are first order asymptotically equivalent to the matching/imputation based estimators developed by Frölich (2007) and Hong and Nekipelov (2008). Our point estimators are easier to implement in that they only require nonparametric estimation of $q(x_1)$, while matching or nonparametric imputation requires estimates for $m_z(x_1)$ and $\mu_z(x_1)$, $z = 0, 1$ as well. However, this advantage disappears if an estimate of the asymptotic variance is also desired.

We will now undertake a more careful comparison of the inverse probability weighted and imputation estimators based on higher order mean square expansions.

3.3 Second order asymptotic results and bandwidth selection

To gain a better approximation to the finite sample properties of the inverse probability weighted estimators versus the imputation based estimators, we derive and compare higher order asymptotic mean squared error (AMSE) expansions. Since this is a tedious and notation-heavy exercise, we focus exclusively on the LATE parameter and the case where X_1 is a scalar ($r = 1$). Based on the results we suggest a data-driven procedure to pick the bandwidth for our estimator in finite samples.

Let $\hat{\Delta}$ and $\hat{\Gamma}$ denote estimators of $\Delta = E[W(1) - W(0)]$ and $\Gamma = E[D(1) - D(0)]$, respectively. In particular, we will consider two types of estimators: (i) a bias-corrected version of the inverse

probability weighted estimator (denoted $\hat{\Delta}_b$ and $\hat{\Gamma}_b$) and (ii) a nonparametric imputation estimator (denoted $\hat{\Delta}_m$ and $\hat{\Gamma}_m$). More specifically, $\hat{\Delta}_b$ and $\hat{\Gamma}_b$ are given by the numerator and denominator of $\hat{\tau}$, respectively, minus bias correction terms suggested by Ichimura and Linton (2002). The exact formulas are stated in Appendix C. The imputation estimators are taken from Frölich (2007) and are given by

$$\hat{\Delta}_m = \frac{1}{n}\sum_{i=1}^{n}[\hat{m}_1(X_{1i}) - \hat{m}_0(X_{1i})] \quad \text{and} \quad \hat{\Gamma}_m = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_1(X_{1i}) - \hat{\mu}_0(X_{1i})],$$

where $\hat{m}_z(\cdot)$ and $\hat{\mu}_z(\cdot)$ are nonparametric estimates of the functions $m_z(\cdot)$ and $\mu_z(\cdot)$, respectively. We assume that all nonparametric regressions required to compute $\hat{\Delta}_b, \hat{\Gamma}_b, \hat{\Delta}_m$ and $\hat{\Gamma}_m$ are estimated by local linear regression; for the inverse probability weighted estimator the propensity score must be estimated by the leave-one-out version of the estimator.

Both estimators of Δ and Γ are asymptotically linear and share the same influence function:

$$\sqrt{n}(\hat{\Delta}_e - \Delta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i + o_p(n^{-\alpha}) \text{ and } \sqrt{n}(\hat{\Gamma}_e - \Gamma) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma_i + o_p(n^{-\alpha}),\ e = b, m \quad (10)$$

where $\alpha \geq 0$, $\delta_i = \delta(Y_i, D_i, Z_i, X_{1i})$, $\gamma_i = \gamma(Y_i, D_i, Z_i, X_{1i})$ with $E(\gamma_i) = E(\delta_i) = 0$ (Lemma 2 in Appendix B provides explicit expressions for δ and γ). For $e = b$ and $\alpha = 0$, these representations are based on Ichimura and Linton (2002); for $e = m$ and $\alpha = 0$, they follow from Heckman et al. (1998) and Frölich (2007). We extend these results by showing that (10) holds even for $\alpha = \frac{1}{10}$. As we will shortly see, our AMSE expansion relies on this faster rate of convergence.

The LATE estimators of interest are of the form $\hat{\Delta}/\hat{\Gamma}$ (as the following discussion pertains to both types of estimators, subscripts are omitted). Taking the second order Taylor expansion of this ratio around the point Δ/Γ yields

$$\frac{\hat{\Delta}}{\hat{\Gamma}} - \frac{\Delta}{\Gamma} = \frac{1}{\Gamma}(\hat{\Delta} - \Delta) - \frac{\Delta}{\Gamma^2}(\hat{\Gamma} - \Gamma) - \frac{1}{\Gamma^2}(\hat{\Delta} - \Delta)(\hat{\Gamma} - \Gamma) + \frac{2\Delta}{\Gamma^3}(\hat{\Gamma} - \Gamma)^2 + o_p(n^{-1}) \quad (11)$$

Grouping the terms on the rhs in a specific way, we take the square of both sides and rescale by n:

$$\begin{aligned}
& n\left(\frac{\hat{\Delta}}{\hat{\Gamma}} - \frac{\Delta}{\Gamma}\right)^2 \\
& = \sqrt{n}\left(\frac{1}{\Gamma}(\hat{\Delta} - \Delta) - \frac{\Delta}{\Gamma^2}(\hat{\Gamma} - \Gamma)\right)^2 && (12)\\
& -\frac{2}{\sqrt{n}}\left(\frac{1}{\Gamma}\sqrt{n}(\hat{\Delta} - \Delta) - \frac{\Delta}{\Gamma^2}\sqrt{n}(\hat{\Gamma} - \Gamma)\right)\left(\frac{1}{\Gamma^2}n(\hat{\Delta} - \Delta)(\hat{\Gamma} - \Gamma) + \frac{2\Delta}{\Gamma^3}n(\hat{\Gamma} - \Gamma)^2\right) && (13)\\
& +O_p(n^{-1}) && (14)
\end{aligned}$$

To approximate the MSE of $\hat{\Delta}/\hat{\Gamma}$, we consider the expectation of the terms (12) and (13). We show that for suitable choices of h, consistent with Assumption 8, these expectations can be further expanded as a sum of terms of order $O(1)$, $O(n^{-3/5})$ and $o(n^{-3/5})$. Our MSE approximation consists of retaining the terms $O(n^{-3/5})$ or larger and ignoring those $o(n^{-3/5})$, which of course includes the expectation of (14).[6]

More specifically, for both types of estimators the expectation of (12) can be approximated as

$$E\left[\sqrt{n}\left(\frac{1}{\Gamma}(\hat{\Delta}-\Delta)-\frac{\Delta}{\Gamma^2}(\hat{\Gamma}-\Gamma)\right)^2\right]=C_0+C_1n^{-1}h^{-1}+C_2nh^4+o(n^{-1}h^{-1}+nh^4), \qquad (15)$$

where C_0, C_1 and C_2 are positive constants. For $\hat{\Delta}=\hat{\Delta}_b$ and $\hat{\Gamma}=\hat{\Gamma}_b$, this expansion follows from Theorem 2 in Ichimura and Linton (2002). On the other hand, for $\hat{\Delta}=\hat{\Delta}_m$ and $\hat{\Gamma}=\hat{\Gamma}_m$, equation (15) is a consequence of Lemma 4.3 in Kalyanaraman (2009).[7] The expectation of the term (13) can be evaluated by substituting in the influence function representations of Δ and Γ given in (10). Straightforward calculations (shown in Appendix C) yield:

$$\begin{aligned}&-\frac{2}{\sqrt{n}}E\left[\left(\frac{1}{\Gamma}\sqrt{n}(\hat{\Delta}-\Delta)-\frac{\Delta}{\Gamma^2}\sqrt{n}(\hat{\Gamma}-\Gamma)\right)\left(\frac{1}{\Gamma^2}n(\hat{\Delta}-\Delta)(\hat{\Gamma}-\Gamma)+\frac{2\Delta}{\Gamma^3}n(\hat{\Gamma}-\Gamma)^2\right)\right]\\&=-\frac{2}{n\Gamma^3}E[(\delta_1-\tau\gamma_1)\delta_1\gamma_1]-\frac{4\tau}{n\Gamma^3}E[(\delta_1-\tau\gamma_1)\gamma_1^2]+o(n^{-\alpha-1/2}). \qquad (16)\end{aligned}$$

Suppose that $h\propto n^{-2/5}$. (As will be argued shortly, this choice is not only consistent with Assumption 8, but is also optimal in the sense of minimizing MSE.) For this rate choice, the second and third terms in (15) are both $O(n^{-3/5})$, while the remainder is $o(n^{-3/5})$. The first two terms in (16) are $O(n^{-1})$, and hence $o(n^{-3/5})$. Thus, the desired expansion is given by the first three terms in (15) provided that $-\alpha-\frac{1}{2}\leq-\frac{3}{5}$, i.e. $\alpha\geq\frac{1}{10}$. Lemmas 4 and 5 in Appendix C show that for both types of estimators it is indeed possible to take $\alpha=\frac{1}{10}$ in (10).

The following theorem summarizes our MSE expansion and provides explicit expressions for the constants C_0, C_1 and C_2 for both estimators. The conditions stated in Section 3.2 are sufficiently strong to deliver the results. To state the theorem, let $\tilde{Y}=Y-\tau D$, and for any previously defined

[6] Of course, $X_n=o_p(1)$ implies $E(X_n)=o(1)$ only under certain regularity conditions such as uniform integrability of the sequence X_n. Following most of the related literature, we do not explicitly verify or impose such regularity conditions.

[7] Application of the cited results is made possible by regarding (12) as the asymptotic MSE of the combined "estimator" $\frac{1}{\Gamma}(\hat{\Delta}-\tau\hat{\Gamma})$ of the parameter $\frac{1}{\Gamma}(\Delta-\tau\Gamma)=0$; see Appendix C for details. This obviates the need to consider explicitly the covariance between $\hat{\Delta}$ and $\hat{\Gamma}$ in evaluating (12).

piece of notation, let the tilde accent denote replacing Y with $\tilde{Y}$ in the definition of that object. E.g., $\tilde{m}_1(x_1) = E[\tilde{Y} \mid X_1 = x_1, Z = 1]$, etc. Further, let $\hat{\tau}_b = \hat{\Delta}_b/\hat{\Gamma}_b$ and $\hat{\tau}_m = \hat{\Delta}_m/\hat{\Gamma}_m$.

Theorem 3 *Suppose that Assumption 1 and Assumptions 4 through 8 are satisfied with $r = 1$. If, in addition, $h \propto n^{-2/5}$, then:*

$$E[n(\hat{\tau}_b - \tau)^2] = F_0 + F_1 n^{-1}h^{-1} + F_2 nh^4 + o(n^{-1}h^{-1} + nh^4), \tag{17}$$

$$E[n(\hat{\tau}_m - \tau)^2] = G_0 + G_1 n^{-1}h^{-1} + G_2 nh^4 + o(n^{-1}h^{-1} + nh^4), \tag{18}$$

where

$$\begin{aligned}
F_0 &= E[\psi^2(Y, D, Z, X_1)], \\
F_1 &= \frac{\|K\|^2}{\Gamma^2} E\left\{ \frac{1}{f(X_1)} \Big[\frac{1 - q(X_1)}{q^2(X_1)} \tilde{s}_1^2(X_1) + \frac{q(X_1)}{(1 - q(X_1))^2} \tilde{s}_0^2(X_1) \Big] \right\} \\
&\quad + \frac{3\|K - K * K\|^2}{\Gamma^2} E\left\{ \frac{1}{f(X_1)} \Big[\frac{1 - q(X_1)}{q(X_1)} \tilde{m}_1(X_1) - \frac{q(X_1)}{1 - q(X_1)} \tilde{m}_0(X_1) \Big]^2 \right\}, \\
F_2 &= \frac{\nu_2^2(K)}{\Gamma^2} \left(E\left[q''(X_1) \left(\frac{\tilde{m}_1(X_1)}{q(X_1)} + \frac{\tilde{m}_0(X_1)}{1 - q(X_1)} \right) \right] \right)^2,
\end{aligned}$$

and

$$\begin{aligned}
G_0 &= E[\psi^2(Y, D, Z, X_1)], \\
G_1 &= \frac{\|K\|^2}{\Gamma^2} E\left[\frac{1}{f(X_1)} \Big(\frac{\tilde{s}_1^2(X_1)}{q(X_1)} + \frac{\tilde{s}_0^2(X_1)}{1 - q(X_1)} \Big) \right], \\
G_2 &= \frac{\nu_2^2(K)}{4\Gamma^2} \big(E[\tilde{m}_1''(X_1) - \tilde{m}_0''(X_1)] \big)^2,
\end{aligned}$$

with

$$\nu_2(K) = \int u^2 K(u) du, \quad \|K\|^2 = \int K^2(u) du, \quad K * K(t) = \int K(u) K(t - u) du.$$

Comparing the coefficients in the two expansions, one can observe that neither estimator *generally* dominates the other in terms of variance (F_1 vs. G_1) or bias (F_2 vs. G_2). There are some special cases in which these quantities are easier to compare, but they are fairly limited in scope. For example, if $q(x)$ is roughly linear, then F_2 will be small, while G_2 can still be large. But if in addition $q(x) \approx 0.5$ for most of the support of X_1, then G_1 will be smaller than F_1, so the joint effect is ambiguous. Alternatively, if the difference between $\tilde{m}_0$ and $\tilde{m}_1$ is approximately linear in x, then G_2 will be small, while F_2 might still be large or small. If $\tilde{m}_0(x) \equiv 0$ (and hence $\tilde{s}_0^2(x) \equiv 0$), then F_1 will be larger than G_1 when $q(x)$ is close to zero on most of $\mathcal{X}$, etc.

More importantly, the expansions given in Theorem 3 can be used to guide finite sample bandwidth selection for both estimators. We look for an optimal bandwidth sequence of the form $h = An^{\alpha}$. Ignoring the remainder terms in (17) and (18), the optimal rate must equalize $n^{-1}h^{-1}$ and nh^4, yielding $h \propto n^{-2/5}$. To pin down the constant of proportionality, one solves

$$\min_{A\geq 0}(F_1 A^{-1} n^{-\frac{3}{5}} + F_2 A^4 n^{-\frac{3}{5}}) \text{ and } \min_{A\geq 0}(G_1 A^{-1} n^{-\frac{3}{5}} + G_2 A^4 n^{-\frac{3}{5}}),$$

which gives $A_b^* = [F_1/(4F_2)]^{1/5}$ and $A_m^* = [G_1/(4G_2)]^{1/5}$. Then $F_1 n^{-1}h^{-1} + F_2 nh^4 < G_1 n^{-1}h^{-1} + G_2 nh^4$, and hence $\hat{\tau}_b$ is predicted to outperform $\hat{\tau}_m$ in terms of finite sample MSE, if and only if

$$(F_1/G_1)^4 < G_2/F_2. \tag{19}$$

Using a pilot bandwidth, one can consistently estimate $\hat{A}_b^*$, $\hat{A}_m^*$ and the ratios in (19) based on the sample analog principle. Then one can compare the two sides of (19) to judge whether $\hat{\tau}_b$ can be expected to outperform $\hat{\tau}_m$ when bandwidth is chosen as $\hat{A}_b^* n^{-2/5}$ and $\hat{A}_m^* n^{-2/5}$, respectively. Thus, we have a practically feasible criterion for choosing between the two estimators.

Two comments about the bandwidth selection procedure are in order. First, the procedure is valid only if $F_2 > 0$ and $G_2 > 0$. As suggested above, these conditions do not always hold. In this case the bias-variance tradeoff stated in (18) disappears, and the procedure calls for infinitely large bandwidth (extreme oversmoothing). Kalyanaraman (2009) proposes a regularized mean square objective to deal with this problem; pursuing this approach is beyond the scope of this paper.

Second, the argument given might seem somewhat circular in that Theorem 3 directly assumes $h \propto n^{-2/5}$. Nevertheless, one can formally expand the MSE of these estimators as

$$C_0 + C_1 n^{-1}h^{-1} + C_2 nh^4 + \text{remainder}$$

for the range of bandwidths given in Assumption 8. For h not proportional to $n^{-2/5}$, it is however not generally possible to argue that the remainder term is of order $o(n^{-1}h^{-1}+nh^4)$; hence, dropping the remainder in choosing the optimal bandwidth is not justified. If one *assumes* that the remainder term can be dropped, it is clear that the optimal tradeoff between the remaining two terms that depend on h occurs when $h \propto n^{-2/5}$. What Theorem 3 shows is that this assumption is self-confirming, i.e. the remainder term is indeed negligible when $h \propto n^{-2/5}$.

4 Testing for unconfoundedness

4.1 The proposed test procedure

If treatment assignment is unconfounded conditional on a subset X_2 of X, then, under regularity conditions, one can consistently estimate ATT ($\equiv \beta_t$) using the estimator proposed by HIR:

$$\hat{\beta}_t = \sum_{i=1}^{n} \hat{p}(X_{2i}) \Big(\frac{D_i Y_i}{\hat{p}(X_{2i})} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_{2i})} \Big) \Big/ \sum_{i=1}^{n} \hat{p}(X_{2i}),$$

where $\hat{p}(x)$ is a suitable nonparametric estimator of $p(x_2) = P(D=1|X_2=x_2)$. (HIR originally proposed the series logit estimator, but local polynomial regression can also be used.) Given a binary instrument that is valid conditional on a subset X_1 of X, one-sided non-compliance implies ATT=LATT, and hence ATT can also be consistently estimated by $\hat{\tau}_t$. On the other hand, if the unconfoundedness assumption does not hold, then $\hat{\tau}_t$ is still consistent, but $\hat{\beta}_t$ is generally not consistent. Hence, we can test the unconfoundedness assumption (or at least a necessary condition of it) by comparing $\hat{\tau}_t$ with $\hat{\beta}_t$. Let $\rho_d(x_2) = E[Y(d)|X_2=x_2] = E[Y|D=d, X_2=x_2]$, $d=0,1$, $p = P(D=1)$, and

$$\begin{aligned} &\phi_t(y,d,x_2) \\ =&\frac{p(x_2)}{p} \left\{ \frac{d(y-\rho_1(x_2))}{p(x_2)} - \frac{(1-d)(y-\rho_0(x_2))}{1-p(x_2)} + \frac{d(\rho_1(x_2)-\rho_0(x_2)-\beta_t)}{p(x_2)} \right\}. \end{aligned}$$

The asymptotic properties of the difference between $\hat{\tau}_t$ and $\hat{\beta}_t$ are summarized in the following theorem:

Theorem 4 *Suppose that Assumptions 1 through 8 are satisfied. If, in addition,* $P[D(1)=1] \neq 1$, $P[Y(0)=0] \neq 1$ *and some additional regularity conditions stated by HIR hold, then*

$$\sqrt{n}(\hat{\tau}_t - \hat{\beta}_t) \xrightarrow{d} N(0, \sigma^2),$$

where $\sigma^2 = E[(\psi_t(Y,D,Z,X_1) - \phi_t(Y,D,X_2))^2]$.

The additional regularity conditions referred to in Theorem 4 restrict the distribution of X_2, impose smoothness of $p(x_2)$, etc. For the test to "work", it is also required that $P[D(1)=1] \neq 1$ and $P[Y(0)=0] \neq 1$. If $P[D(1)=1]=1$, then one-sided non-compliance implies $P[D=Z]=1$. Therefore, instrument validity and unconfoundedness are one and the same. On the other hand, if

$P[Y(0) = 0] = 1$, then $Y = DY(1)$, and so $(1 - Z)Y = 0$ and $(1 - D)Y = 0$. Hence, our LATT estimator reduces to

$$\hat{\tau}_t = \sum_{i=1}^{n} Z_i Y_i \Big/ \sum_{i=1}^{n} Z_i D_i = \sum_{i=1}^{n} D_i Y_i \Big/ \sum_{i=1}^{n} D_i,$$

where the second equality holds since $ZY = ZDY(1) = DY(1) = DY$ and $ZD = D$. It can be shown that $\hat{\tau}_t = \sum_{i=1}^{n} D_i Y_i / \sum_{i=1}^{n} D_i$ is asymptotically equivalent to $\hat{\beta}_t$ in the sense that the difference between the two is $o_p(n^{-1/2})$. That is, whether or not the unconfoundedness assumption holds, $\sqrt{n}(\hat{\tau}_t - \hat{\beta}_t) = o_p(1)$ and the test is not valid.

Under the unconfoundedness assumption, HIR show that the asymptotic linear representation of $\hat{\beta}_t$ is given by

$$\sqrt{n}(\hat{\beta}_t - \beta_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_t(Y_i, D_i, X_i) + o_p(1).$$

Theorem 4 follows directly from this result. Let $\widehat{\psi}_t(\cdot)$ and $\widehat{\phi}_t(\cdot)$ be (uniformly) consistent estimators of ψ_t and ϕ_t obtained, e.g., by the sample analog principle (see the discussion after Thorem 2). A consistent estimator for σ^2 can then be constructed as

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{\phi}_t(Y_i, D_i, X_i) - \widehat{\psi}_t(Y_i, D_i, Z_i, X_i)\big)^2.$$

As a result, one can use a simple z-test with the statistic $\sqrt{n}(\hat{\tau}_t - \hat{\beta}_t)/\hat{\sigma}$ to test the unconfoundedness assumption. Since the difference between LATT and ATT can generally be of either sign, a two-sided test is appropriate.

The proposed test is quite flexible in that it does not place any restrictions on the relationship between X_1 and X_2. The two vectors can overlap, be disjoint, or one might be contained in the other. The particular case in which X_2 is empty corresponds to testing whether treatment assignment is completely random. Finally, we note that if the instrument is not entirely trusted, then the interpretation of the test should be more conservative; namely, it should be regarded as a joint test of unconfoundedness and the IV conditions.

4.2 The implications of unconfoundedness

What are the benefits of (potentially) having the unconfoundedness assumption at one's disposal in addition to IV conditions? An immediate one is that the ATE parameter also becomes identified

and can be consistently estimated, for example, by the inverse probability weighted estimator proposed by HIR or by nonparametric imputation as in Hahn (1998).

A more subtle consequence has to do with the efficiency of $\hat{\beta}_t$ and $\hat{\tau}_t$ as estimators of ATT. If an instrument satisfying one-sided compliance is available, and the unconfoundedness assumption holds at the same time, then both estimators are consistent. Furthermore, the asymptotic variance of $\hat{\tau}_t$ attains the semiparametric efficiency bound that prevails under the IV conditions alone, and the asymptotic variance of $\hat{\beta}_t$ attains the corresponding bound that can be derived from the unconfoundedness assumption alone. The simple conjunction of these two identifying conditions does not generally permit an unambiguous ranking of the efficiency bounds even when $X_1 = X_2$. Nevertheless, by taking appropriate linear combinations of $\hat{\beta}_t$ and $\hat{\tau}_t$, one can obtain estimators that are more efficient than either of the two. This observation is based on the following elementary lemma:

Lemma 1 *Let A_0 and A_1 be two random variables with finite variance. Define $A_a = (1-a)A_0 + aA_1$ for any $a \in \mathbb{R}$. Let $\bar{a} = \frac{\text{var}(A_0)-\text{cov}(A_0,A_1)}{\text{var}(A_1-A_0)}$. Then:*

(*a*) $\text{var}(A_{\bar{a}}) \leq \text{var}(A_a)$ *for all* $a \in \mathbb{R}$.

(*b*) $\text{var}(A_{\bar{a}}) < \text{var}(A_0)$ *when* $\bar{a} \neq 0$, *i.e.* $\text{var}(A_0) \neq \text{cov}(A_0, A_1)$.

(*c*) $\text{var}(A_{\bar{a}}) < \text{var}(A_1)$ *when* $\bar{a} \neq 1$, *i.e.* $\text{var}(A_1) \neq \text{cov}(A_0, A_1)$.

To be more specific, let $\hat{\beta}_t(a) = (1-a)\hat{\beta}_t + a\hat{\tau}_t$ and $\mathcal{V}_t(a) = \text{var}[(1-a)\phi_{ti} + a\psi_{ti}]$, where $\psi_{ti} = \psi_t(Y_i, D_i, Z_i, X_{1i})$ and $\phi_{ti} = \phi_t(Y_i, D_i, X_{2i})$. Then for any $a \in \mathbb{R}$, $\hat{\beta}_t(a)$ is consistent for τ_t and is asymptotically normal with asymptotic variance $\mathcal{V}_t(a)$, i.e. $\sqrt{n}(\hat{\beta}_t(a) - \tau_t) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}_t(a))$. The optimal weight $\bar{a}$ can be obtained as

$$\bar{a} = \frac{\text{var}(\phi_t) - \text{cov}(\phi_t, \psi_t)}{\text{var}(\phi_t) + \text{var}(\psi_t) - 2\text{cov}(\phi_t, \psi_t)},$$

so that $\mathcal{V}_t(\bar{a}) \leq \mathcal{V}_t(a)$ for all $a \in \mathbb{R}$. In other words, $\hat{\beta}_t(\bar{a})$ will be the most efficient estimator among all linear combinations of $\hat{\beta}_t$ and $\hat{\tau}_t$. Although $\bar{a}$ is unknown in general, it can be consistently estimated by

$$\hat{a} = \frac{\sum_{i=1}^n \widehat{\phi}_t(Y_i, D_i, Z_i, X_{1i})\big(\widehat{\phi}_t(Y_i, D_i, Z_i, X_{1i}) - \widehat{\psi}_t(Y_i, D_i, X_{2i})\big)}{\sum_{i=1}^n \big(\widehat{\phi}_t(Y_i, D_i, Z_i, X_{1i}) - \widehat{\psi}_t(Y_i, D_i, X_{2i})\big)^2}.$$

Slutsky's theorem implies that $\sqrt{n}(\hat{\beta}_t(\hat{a}) - \tau_t)$ has the same asymptotic distribution as $\sqrt{n}(\hat{\beta}_t(\bar{a}) - \tau_t)$.

If $Var(\phi_t) = Cov(\phi_t, \psi_t)$, then $\bar{a} = 0$, which implies that $\hat{\beta}_t$ itself is more efficient than $\hat{\tau}_t$ (or any linear combination of the two). We give sufficient conditions for this result.

Theorem 5 *Suppose that Assumption 1 parts (i), (iii), (iv), (v) and Assumption 3 are satisfied, and let* $V = (Y(0), Y(1))$. *If, in addition,* $X_1 = X_2 = X$,

$$E(V \mid Z, D, X) = E(V \mid X) \quad \textit{and} \quad E(VV' \mid Z, D, X) = E(VV' \mid X), \tag{20}$$

then $\bar{a} = 0$.

The proof of Theorem 5 is provided in Appendix D. The conditions of Theorem 5 are stronger than those of Theorem 4. The latter theorem only requires that the IV assumption and unconfoundedness both hold at the same time, which in general does not imply the stronger *joint* mean-independence conditions given in (20). If the null of unconfoundedness is accepted due to (20) actually holding, then $\hat{\beta}_t$ itself is the most efficient estimator of ATT in the class $\{\hat{\beta}_t(a) : a \in \mathbb{R}\}$. Nevertheless, it is in principle possible for unconfoundedness to hold and (20) to fail hence the conclusion is not automatic.

5 Monte Carlo simulations

We present two small-scale Monte Carlo studies to check the finite sample accuracy of our MSE approximations in Section 3.3 and to illustrate the size and power properties of the proposed test. The data generating process is given by

$$Y = D(X + \epsilon), \quad D = Z \cdot 1(X > \epsilon), \quad Z = 1[q(X) > U],$$

where (X, U, ϵ, ν) are mutually independent uniform $[0, 1]$ random variables and $q(X)$, the propensity score, is a given function. The model is chosen so that closed form expressions can be obtained for the functions in terms of which the MSE expansion coefficients are defined. Thus the true value of these coefficients can be computed without nonparametric estimation (only averaging over a very large sample is required). LATE can also be computed analytically (it is equal to one).

In comparing the inverse probability weighted and imputation estimators of LATE we vary the functional parameter $q(X)$. The first specification is given by $q_1(X) = \Lambda(-1 + X)$, where $\Lambda(\cdot)$ is the logistic cdf. In this case $q_1(x)$ is approximately linear over the support of X so that F_2 is very small. In particular, $G_2/F_2 \approx 2390$ and $(F_1/G_1)^4 \approx 7$. By (19), this means that if the

Table 1: MSE simulations

	$q(X) = q_1(X)$						$q(X) = q_2(X)$					
	MSE$[\sqrt{n}(\hat{\tau}_b - \tau)]$			MSE$[\sqrt{n}(\hat{\tau}_m - \tau)]$			MSE$[\sqrt{n}(\hat{\tau}_b - \tau)]$			MSE$[\sqrt{n}(\hat{\tau}_m - \tau)]$		
n	h	Approx.	Sim.	h	Approx.	Sim.	h	Approx.	Sim.	h	Approx.	Sim.
100	0.81	0.71	0.78	0.16	0.78	0.89	0.18	0.99	1.03†	0.16	0.89	1.01†
200	0.62	0.70	0.72	0.12	0.75	0.77	0.14	0.91	1.04	0.12	0.85	0.91
300	0.53	0.69	0.70	0.10	0.74	0.72	0.12	0.88	0.93	0.10	0.83	0.84
500	0.43	0.69	0.68	0.08	0.71	0.69	0.10	0.85	0.85	0.09	0.82	0.82

Note: h denotes the optimal bandwidth based on the higher order MSE approximation. Approx=the corresponding approximate MSE. Sim=simulated (exact) MSE based on 5000 Monte Carlo repetitions. †Denotes cases where *median* squared error was calculated due to a few large outliers. For $q_1(x)$ the asymptotic variance (i.e., the leading term is the MSE expansion) is 0.68; for $q_2(x)$ it is 0.77.

optimal bandwidth given in Section 3.3 is used to compute both estimators, then $\hat{\tau}_b$ is expected to outperform $\hat{\tau}_m$ in terms of MSE.[8] This is confirmed by the simulations shown in the left panel of Table 1. As the sample size increases, the accuracy of the MSE approximation improves and the difference between the two estimators becomes smaller and smaller.

The second specification for $q(X)$ is given by $q_2(X) = 0.5\Lambda(-33X + 85X^2 - 55X^3) + 0.25$. In this case $q_2(x)$ is non-monotonic and has a lot of curvature with $G_2/F_2 \approx 1$ and $(F_1/G_1)^4 \approx 19$. Therefore, $\hat{\tau}_m$ can expected to outperform $\hat{\tau}_b$ in finite samples, as seen in the right panel of Table 1. Again, as the sample size increases, so does the accuracy of the approximation. In both examples considered here the propensity score is safely bounded away from zero and one. If $q(X)$ gets close to these boundaries over a large part of the support of X, we found that the quality of the MSE approximations can be rather poor even in moderately large samples.

To study the properties of our test statistic we use a slightly modified data generating process for which $Y(0) \neq 0$ and we can control the extent to which selection to treatment depends on

[8]We calculate the theoretically optimal bandwidths based on the simulated (i.e., true) values of the asymptotic MSE expansion coefficients. For smaller sample sizes some observations X_i may not have enough neighbors within the radius of the optimal bandwidth to compute the necessary local linear regressions. For these observations three times the theoretically optimal bandwidth is used. Such observations can still lead to outlier LATE estimates.

unobservables:

$$Y = (1 - D)(X + \epsilon), \quad D = Z \cdot 1[a\epsilon + (1 - a)\nu > 0.25 + 0.5X], \quad Z = 1[q(X) > U],$$

where (X, U, ϵ, ν) are mutually independent uniform $[0, 1]$ random variables, $a \in [0, 1]$, and $q(X) = q_1(X)$. Clearly, $D(0) = 0$ so that one-sided non-compliance is satisfied. For $a = 0$ treatment assignment is unconfounded, but for $a > 0$, D and Y are correlated conditional on X (the larger a, the stronger the correlation). Hence, we set $a = 0$ to study size control and $a = 0.25$, 0.5 and 1 to study power. We consider two different sample sizes: $n = 100$ and 300.

We use local linear regression with various bandwidths to compute the test statistic; in particular, for $n = 100$, we set $h = 0.8$, 1.6 and 2.4, and for $n = 300$, we set $h = 0.6$, 1.2 and 1.8. As a reference point, we also implement the test using the series logit estimator (with up to quadratic terms). Rejection rates of the proposed test, computed over 1000 Monte Carlo cycles, are shown in Table 2 for nominal size equal to 5%.

An important finding is that size and power are rather sensitive to bandwidth choice; proper size control, in particular, seems to require a fair amount of smoothing. Since the estimator is implemented with the Epanechnikov kernel, supported on $[-1, 1]$, bandwidths greater than one imply that all sample observations receive non-zero weight in computing $\hat{q}(X_i)$, $\hat{p}(X_i)$, etc. As expected, power increases with the parameter a as well as sample size. Series logit tends to be conservative, especially for $n = 100$.

Table 2: Unconfoundedness test simulations

		$n = 100$				$n = 300$			
		$h = 0.8$	$h = 1.6$	$h = 2.4$	SLE	$h = 0.6$	$h = 1.2$	$h = 1.8$	SLE
$a = 0$	size	0.170	0.066	0.061	0.029	0.252	0.053	0.055	0.043
$a = 0.25$	power	0.285	0.190	0.177	0.054	0.450	0.319	0.325	0.200
$a = 0.5$	power	0.592	0.526	0.524	0.266	0.839	0.921	0.907	0.820
$a = 1$	power	0.906	0.959	0.950	0.835	0.996	1.000	1.000	1.000

Note: The simulated rejection rates are based on 1000 Monte Carlo repetitions; nominal size is 5%. h=bandwidth used in computing local linear regressions. SLE=series logit.

6 Empirical application

We apply our method to estimate the impact of JTPA training programs on subsequent earnings and to test whether participation is unconfounded conditional on a vector of observables. Abadie et al. (2002) and Frölich and Melly (2008a) use the same data set to examine the distributional effect of this program on earnings; the data set is publicly available at the URL

http://econ-www.mit.edu/faculty/angrist/data1/data/abangim02

As explained by Abadie et al. (2002), the JPTA program involved collecting data specifically for purposes of evaluation. At some of the service delivery sites, between Nov. 1987 and Sept. 1989, applicants were randomly selected as eligible to receive a job-related service (classroom training, on-the-job training, job search assistance, probationary employment, other) or were denied services and excluded from the program for 18 months.[9] Clearly, this random offer of services (Z) can be used as an instrument for evaluating the effect of actual program participation (D) on earnings (Y), measured as the sum of earnings in the 30 month period following the eligibility decision. About 36 percent of those who were offered services chose not to participate; conversely, a small fraction of applicants, less than 0.5 percent, ended up participating despite the fact that they were ruled ineligible. Hence, the instrument satisfies one-sided non-compliance almost perfectly; the small number of observations violating this condition were dropped from the sample. The total number of observations is then 11,150; of these, 6,067 are female and 5,083 are male.

A number of covariates describing the socio-economic status of applicants are also available in the data set. These are all dummy variables and are summarized in Table 3, along with the variables discussed above. As Z is completely randomly assigned, and presumably has no direct effect on the outcome, it is a valid instrument regardless of whether one conditions on additional covariates. In all exercises presented in this section we will set $X_1 = X_2 \equiv X$, where X is some subset of the covariates listed in Table 3. A further implication of random assignment is that LATE=LATT for the instrument.

As all available covariates are discrete, one cannot use local polynomial regression to estimate regression functions of X. An r-dimensional vector of dummy variables partitions the population into 2^r subpopulations, each corresponding to a different setting of X. One can then use data from

[9]The data set consists of 11,204 applicants who were categorized as adult males or females; data on youth are not included.

Table 3: JPTE variables

Variable name	Definition	Mean
Z	=1 if offered any job-related service	0.6715
D	=1 if actually participated	0.4309
Y	=30-month earning following eligibility decision (dollars)	15,826
SEX	=1 if male	0.4559
MINORITY	=1 if black or Hispanic	0.3680
HS	=1 if high school graduate (or GED)	0.7263
MS	=marital status; =1 if married	0.3316
W13	=1 if worked less than 13 weeks past year	0.5183
BELOW30	=1 if below 30 years of age	0.4398

Notes: All dummy variable definitions are to be interpreted as "= 0 otherwise". For HS, MS and W13 a few observations are recorded as strictly between zero and one. We treat all non-zero observations as one. Means are computed after enforcing one-sided non-compliance. More detailed age group dummies are available than the BELOW30 variable considered here. Some additional covariates are also available but not listed or used; see Abadie et al. (2002).

each subpopulation to construct the required nonparametric estimates. For example, for a given setting of X, the relative frequency of observations with $Z_i = 1$ in the corresponding subsample is a $\sqrt{n}$-consistent estimate of the value of $q(X)$. Formally, for $s \in \{0,1\}^r$, let $N_s = \sum_{i=1}^n 1(X_i = s)$ and $\hat{q}_s = \frac{1}{N_s} \sum_{i=1}^n 1(Z_i = 1, X_i = s)$. Then $\hat{q}(X_i)$ may be computed as

$$\hat{q}(X_i) = \sum_s \hat{q}_s 1(X_i = s). \tag{21}$$

Of course, as Z is randomly assigned, each $\hat{q}_s$ will converge to the constant $P(Z = 1)$, and one could estimate this quantity simply by the unconditional sample mean $\frac{1}{n} \sum_{i=1}^n Z_i$. Nevertheless, it is possible to show that using the "partitioned" estimator, i.e. exploiting the fact that Z is also valid conditional on X, may lead to efficiency gains in estimating LATE and LATT.[10] One can construct $\hat{p}(X_i)$, $\hat{m}_1(X_i)$, etc. analogously (these functions will not generally be trivial in X). The first order asymptotic theory presented in this paper, including the test for unconfoundedness, remains valid if these estimators are used in computing $\hat{\tau}$, $\hat{\tau}_t$ and $\hat{\beta}_t$. The drawback of this approach is that if r

[10]Giving a formal proof of this claim is beyond the immediate scope of the paper. The result is similar to Theorem 11 of Frölich and Melly (2008b).

is even moderately large, then the sample at hand might contain very few or no observations from certain subpopulations. (Of course, this is a problem for estimating nontrivial functions of X only.) On the other hand, no bandwidth choice is required to implement the estimators.

With purely discrete covariates it is often natural to define subpopulations of special interest using some components of X, say X^s, and estimate LATE/LATT and ATT within those subpopulations. The unconfoundedness of treatment participation can then be tested separately within each subpopulation w.r.t. $X \setminus X^s$ (which might be empty). Conducting such tests subpopulation by subpopulation is not entirely equivalent to conducting a "joint" unconfoundedness test w.r.t. the whole vector X, just as testing whether regression coefficients are individually zero is not equivalent to testing whether they are all zero at the same time. Nevertheless, these individual tests can provide additional insight. We now provide some concrete examples.

In our first exercise we simply set $X = X_1 = X_2 =$ SEX. We estimate LATT and ATT in the entire population as well as among males and females separately. We conduct an overall unconfoundedness test w.r.t. X, and also individual tests of random treatment participation within the two subpopulations. Results are shown in Table 4. The LATT estimator $\hat{\tau}_t$ is interpreted as follows. Take, for example, the value 1916.4 for females. This means that female compliers who actually participated in the program (i.e., were assigned $Z = 1$), are estimated to increase their 30-month earnings by \$1916.4 on average. Since Z is randomly assigned, this number can also be interpreted as an estimate of LATE, i.e. the average effect among all compliers. Further, by one-sided non-compliance, a third interpretation is that 1916.4 is an estimate of the female ATT, i.e. the average effect of the program among all females that chose participation. The corresponding standard error (547.8) shows that the effect is statistically significant. Turning to the unconfoundedness tests, the hypothesis that treatment participation is unconfounded conditional on gender has a p-value of 0.005 and hence is strongly rejected. The individual tests show strong evidence of selection on unobservables within the male subpopulation but not at all among females. This is valuable information that the overall test does not reveal.

As in Table 4 there are no additional covariates besides SEX, the estimator $\hat{\beta}_t$ for males/females is numerically equivalent to taking the difference between the mean earnings of treated males/females and non-treated males/females. Since the hypothesis of random treatment participation cannot be rejected for females, this figure can then be interpreted as a consistent estimate of ATE (as well as ATT, of course). In contrast, $\hat{\beta}_t$ is a biased and inconsistent estimate of male ATE. Furthermore,

Table 4: $X = \text{SEX}$

Subpop.	Obs.	$\hat{\tau}_t$ (USD)	std($\hat{\tau}_t$) (USD)	$\hat{\beta}_t$ (USD)	std($\hat{\beta}_t$) (USD)	std($\hat{\tau}_t - \hat{\beta}_t$) (USD)	Test stat.	p-value (2-sided)
$X^s = \emptyset$; $X \setminus X^s = \text{SEX}$								
All	11150	1828.1	(506.9)	2979.1	(313.1)	(407.4)	-2.825	0.005
$X^s = \text{SEX}$; $X \setminus X^s = \emptyset$								
Males	5083	1716.0	(916.4)	4035.6	(557.3)	(740.7)	-3.132	0.002
Females	6067	1916.4	(547.8)	2146.7	(346.4)	(436.1)	-0.528	0.597

Note: $\hat{\tau}_t$ is the inverse probability weighted IV estimator of LATT=ATT. $\hat{\beta}_t$ is an estimator of ATT under unconfoundedness. Numbers in parenthesis are standard errors.

using the results in Section 4, one can take a weighted average of $\hat{\tau}_t$ and $\hat{\beta}_t$ to obtain a more efficient estimate of female ATE/ATT. The estimated optimal combination puts nearly all weight on $\hat{\beta}_t$, so the actual efficiency gain from doing so is negligible in this example. However, note that without testing for (and accepting) the unconfoundedness assumption, the only valid estimate of female ATT is $\hat{\tau}_t$, which has a much larger standard error than $\hat{\beta}_t$.

In the second exercise we set $X = X_1 = X_2 =$ (SEX, BELOW30, MINORITY, HS). We perform three types of tests: (i) an overall unconfoundedness test w.r.t. X; (ii) conditioning on the possible values of $X^s =$ SEX, we perform two unconfoundedness tests w.r.t $X \setminus X^s =$ (BELOW30, MINORITY, HS) among males and females separately; (iii) conditioning on the possible values of $X^s = X$, we perform tests of random treatment assignment in each of the resulting 16 subpopulations. Results are reported in Table 5.

Comparing the LATT estimates for males, females and the whole population across Tables 4 and 5, we see that the numbers are reasonably close. Differences between the two sets of estimates are due solely to the way the propensity score is estimated. Note the (very) slight drop in standard errors in Table 5. Unconfoundedness conditional on X is rejected in the population as a whole. There is again strong evidence of male participation being based on factors other than BELOW30, MINORITY and HS, while female participation appears unconfounded with respect to this set of covariates as well.[11] Individual tests of random treatment participation within the 16 subpopulations tend not to reject except for one, or maybe two, male groups. The lack of stronger

[11] Even if the non-rejection for females in Table 4 is due to $(Y(0), Y(1))$ and D being fully independent, it does not automatically follow that these variables are independent conditional on X.

Table 5: X = (SEX, BELOW30, MINORITY, HS)

Subpop.	Obs.	$\hat{\tau}_t$ (USD)	std($\hat{\tau}_t$) (USD)	$\hat{\beta}_t$ (USD)	std($\hat{\beta}_t$) (USD)	std($\hat{\tau}_t - \hat{\beta}_t$) (USD)	Test stat.	p-value (2-sided)
		$X^s = \emptyset$; $X \setminus X^s$ = (SEX, BELOW30, MINORITY, HS)						
All	11150	1810.3	(501.9)	2804.2	(312.6)	(404.1)	-2.460	0.014
		X^s = SEX; $X \setminus X^s$ = (BELOW30, MINORITY, HS)						
Males	5083	1805.9	(904.5)	3936.0	(554.8)	(732.1)	-2.909	0.004
Females	6067	1813.7	(545.3)	1912.4	(347.2)	(434.6)	-0.227	0.820
		X^s = (SEX, BELOW30, MINORITY, HS); $X \setminus X^s = \emptyset$						
M, u. 30, minority, hs	614	1068.1	(2324.7)	2326.2	(1414.6)	(1825.6)	-0.689	0.491
M, u. 30, minority, no hs	235	5745.0	(3288.3)	3832.4	(2318.6)	(3069.9)	0.623	0.533
M, u. 30, white, hs	990	805.8	(2162.7)	2976.4	(1308.0)	(1710.2)	-1.269	0.204
M, u. 30, white, no hs	425	7117.8	(3059.3)	6653.8	(1718.6)	(2739.6)	0.169	0.865
M, o. 30, minority, hs	666	-502.9	(2400.2)	2858.8	(1464.6)	(2053.1)	-1.637	0.102
M, o. 30, minority, no hs	270	2874.9	(3831.6)	5091.5	(2401.1)	(3136.8)	-0.707	0.480
M, o. 30, white, hs	1354	1501.7	(1892.4)	4909.7	(1184.5)	(1464.4)	-2.327	0.020
M, o. 30, white, no hs	529	2160.6	(2703.1)	4180.9	(1613.7)	(2207.4)	-0.915	0.360
F, u. 30, minority, hs	830	1887.5	(1365.0)	913.0	(884.6)	(1105.6)	0.881	0.378
F, u. 30, minority, no hs	297	1920.5	(2116.7)	293.6	(1143.1)	(1837.5)	0.885	0.376
F, u. 30, white, hs	1097	2415.7	(1378.0)	2410.4	(849.3)	(1082.6)	0.005	0.996
F, u. 30, white, no hs	416	62.5	(1773.4)	1292.5	(1150.0)	(1375.2)	-0.894	0.371
F, o. 30, minority, hs	858	2022.4	(1466.4)	1331.3	(942.4)	(1218.1)	0.567	0.571
F, o. 30, minority, no hs	333	1065.0	(2019.8)	2668.5	(1189.4)	(1714.1)	-0.935	0.350
F, o. 30, white, hs	1689	1646.7	(1076.3)	2215.3	(711.9)	(834.7)	-0.681	0.496
F, o. 30, white, no hs	547	2242.7	(1731.1)	3445.7	(1002.9)	(1370.5)	-0.878	0.380

Note: $\hat{\tau}_t$ is the inverse probability weighted IV estimator of LATT=ATT. $\hat{\beta}_t$ is an estimator of ATT under unconfoundedness. Numbers in parenthesis are standard errors. M=male; F=female; u. 30=under 30 years of age; o. 30=over 30 years of age; hs=high school diploma or GED.

rejection among males may be partly due to the relatively small number of observations available in some of these groups. The general pattern shown in Table 5 turns out to be quite robust. If one adds W13 or MS to X, uses finer age dummies, etc., individual tests of random treatment participation in the resulting subpopulations tend not to reject perhaps with a couple of exceptions. Unconfoundedness within the male subpopulation is always rejected, but it is never rejected among females. The p-value for unconfoundedness in the whole population is usually well below 5% as well. Thus, for females the value of $\hat{\beta}_t$ reported in Table 5 can again be interpreted as an

estimate of ATE/ATT. (Compared with Table 4, the value of $\hat{\beta}_t$ has changed somewhat with the incorporation of covariates, but its standard error is virtually the same.)

Finally, we caution that the unconfoundedness test developed in this paper is a pairwise test. Thus, if the test is used in a sequential procedure where some of the tests are performed based on the outcome of previous tests, then size distortions will occur. Of course, this caveat applies quite generally in econometrics—consider, for example, specification testing in regression models based on simple t or F-tests.

7 Conclusion

Given a conditionally valid binary instrument, nonparametric estimators of LATE and LATT can be based on imputation or matching, as in Frölich (2007), or weighting by the estimated propensity score, as proposed in this paper. The two approaches are shown to be asymptotically equivalent; in particular, both types of estimators are $\sqrt{n}$-consistent and efficient. Higher order AMSE expansions suggest that neither estimator generally dominates the other in finite samples. This is confirmed by some simple Monte Carlo experiments. The MSE approximation leads to a data-driven bandwidth selection rule and can also be used to pick between the two types of estimators in practice.

When the available binary instrument satisfies one-sided non-compliance, the proposed estimator of LATT is compared with the ATT estimator of HIR to test the assumption that treatment assignment is unconfounded given a vector of observed covariates. To our knowledge, this is the first such test in the literature. We apply our methods to data obtained under the Job Training Partnership Act. A particularly robust finding is that the set of available covariates does not fully explain mens' participation decision. In contrast, we cannot reject the hypothesis that female participation is essentially random.

A possible direction for future research is to develop tests of unconfoundedness for more general (i.e., non-binary) instruments or finding conditions other than one-sided non-compliance that make such tests possible.

Appendix

In order to simplify notation, we set $X_1 = X$ throughout the Appendix. Furthermore, we use $C > 0$ to denote a generic constant whose value might change from equation to equation.

A. Identification results

Establishing equations (1) and (2) We can write

$$\begin{aligned} E[W(1) - W(0)] &= E\{[D(1) - D(0)][Y(1) - Y(0)]\} \\ &= E[Y(1) - Y(0) \mid D(1) - D(0) = 1]P[D(1) - D(0) = 1] \\ &= \tau \cdot E[D(1) - D(0)], \end{aligned}$$

where the second equality follows from the fact that under monotonicity (Assumption $1(v)$) the random variable $D(1) - D(0)$ is either zero or one. Similarly,

$$\begin{aligned} E[W(1) - W(0) \mid Z = 1] &= E\{[D(1) - D(0)][Y(1) - Y(0) \mid Z = 1]\} \\ &= E[Y(1) - Y(0) \mid D(1) - D(0) = 1, Z = 1]P[D(1) - D(0) = 1 \mid Z = 1] \\ &= \tau_t \cdot E[D(1) - D(0) \mid Z = 1], \end{aligned}$$

where the third equality follows from the fact that under monotonicity $D(1) - D(0) = 1$ implies $D = Z$. ∎

The proof of Theorem 1 The moment conditions in Assumption $1(i)$ ensure that all expectations stated in the theorem are well defined. We will only show equation (3); the remaining claims can be verified similarly. Write

$$\begin{aligned} E\left[\frac{ZY}{q(X)}\right] &= E\left[\frac{ZW}{q(X)}\right] = E\left[\frac{ZW(1)}{q(X)}\right] = E\left\{\frac{ZE[W(1) \mid X, Z]}{q(X)}\right\} \\ &= E\left\{\frac{ZE[W(1) \mid X]}{q(X)}\right\} = E\left\{\frac{E(Z \mid X)E[W(1) \mid X]}{q(X)}\right\} = E[W(1)], \end{aligned}$$

where the first equality on the second line follows from Assumption $1(ii)$. An analogous argument shows $E[(1 - Z)Y/(1 - q(X))] = E[W(0)]$; combining the two results yields (3). ∎

B. The proof of Theorem 2

We give a detailed proof for the estimator $\hat{\tau}$ and a brief outline of the proof for $\hat{\tau}_t$. The details of the latter argument are analogous to those of the former and are omitted.

We analyze the numerator and denominator of $\hat{\tau}$ separately. Let

$$\widehat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z_i Y_i}{\hat{q}(X_i)} - \frac{(1 - Z_i)Y_i}{1 - \hat{q}(X_i)}\right\}, \qquad \widehat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z_i D_i}{\hat{q}(X_i)} - \frac{(1 - Z_i)D_i}{1 - \hat{q}(X_i)}\right\}.$$

so that $\hat{\tau} = \widehat{\Delta}/\widehat{\Gamma}$. The asymptotic properties of $\hat{\Delta}$ and $\hat{\Gamma}$ are established in the following lemma.

Lemma 2 *Under the conditions of Theorem 2,* $\sqrt{n}(\widehat{\Delta}-\Delta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta(Y_i,D_i,Z_i,X_i)+o_p(1)$ *and* $\sqrt{n}(\widehat{\Gamma}-\Gamma)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma(Y_i,D_i,Z_i,X_i)+o_p(1)$*, where*

$$\delta(Y_i,D_i,Z_i,X_i)=\frac{Z_iY_i}{q(X_i)}-\frac{(1-Z_i)Y_i}{1-q(X_i)}-\Delta-\left(\frac{m_1(X_i)}{q(X_i)}+\frac{m_0(X_i)}{1-q(X_i)}\right)(Z_i-q(X_i))$$
$$\gamma(Y_i,D_i,Z_i,X_i)=\frac{Z_iD_i}{q(X_i)}-\frac{(1-Z_i)D_i}{1-q(X_i)}-\Gamma-\left(\frac{\mu_1(X_i)}{q(X_i)}+\frac{\mu_0(X_i)}{1-q(X_i)}\right)(Z_i-q(X_i))$$

Taking Lemma 2 as given for now, we can use the first order Taylor expansion of the bivariate function $f(\hat{\Delta},\hat{\Gamma})=\hat{\Delta}/\hat{\Gamma}$ around the point (Δ,Γ) to write

$$\sqrt{n}(\hat{\tau}-\tau)=\sqrt{n}\left(\frac{\widehat{\Delta}}{\widehat{\Gamma}}-\frac{\Delta}{\Gamma}\right)=\frac{1}{\Gamma}\sqrt{n}(\widehat{\Delta}-\Delta)-\frac{\tau}{\Gamma}\sqrt{n}(\widehat{\Gamma}-\Gamma)+o_p(1). \tag{22}$$

Substituting the influence function representations given in Lemma 2 into (22) establishes the representation in (8). It is easy to check that under Assumption 1(*i*), $E[\psi(Y,D,Z,X)]=0$ and $E[\psi^2(Y,D,Z,X)]<\infty$. Applying the Lindeberg-Levy CLT and Slutsky's theorem to (8) shows $\sqrt{n}(\hat{\tau}-\tau)\xrightarrow{d}\mathcal{N}(0,\mathcal{V})$.

To derive the asymptotic distribution of $\hat{\tau}_t$, we write

$$\widehat{\Delta}_t=\sum_{i=1}^{n}\left\{\hat{q}(X_i)\Big(\frac{Z_iY_i}{\hat{q}(X_i)}-\frac{(1-Z_i)Y_i}{1-\hat{q}(X_i)}\Big)\right\}\Big/\sum_{i=1}^{n}\hat{q}(X_i),$$
$$\widehat{\Gamma}_t=\sum_{i=1}^{N}\left\{\hat{q}(X_i)\Big(\frac{Z_iD_i}{\hat{q}(X_i)}-\frac{(1-Z_i)D_i}{1-\hat{q}(X_i)}\Big)\right\}\Big/\sum_{i=1}^{n}\hat{q}(X_i),$$

so that $\hat{\tau}_t=\widehat{\Delta}_t/\widehat{\Gamma}_t$. Then:

Lemma 3 *Under the conditions of Theorem 2,*

$$\sqrt{n}(\widehat{\Delta}_t-\Delta_t)=\frac{1}{\sqrt{n}}\frac{1}{Q}\sum_{i=1}^{n}q(X_i)\left\{\frac{Z_i(Y_i-m_1(X_i))}{q(X_i)}-\frac{(1-Z_i)(Y_i-m_0(X_i))}{1-q(X_i)}\right.$$
$$\left.+\frac{(m_1(X_i)-m_0(X_i)-\Delta_q)Z_i}{q(X_i)}\right\}+o_p(1),$$
$$\sqrt{n}(\widehat{\Gamma}_t-\Gamma_t)=\frac{1}{\sqrt{n}}\frac{1}{Q}\sum_{i=1}^{n}q(X_i)\left\{\frac{Z_i(D_i-\mu_1(X_i))}{q(X_i)}-\frac{(1-Z_i)(D_i-\mu_0(X_i))}{1-q(X_i)}\right.$$
$$\left.+\frac{(\mu_1(X_i)-\mu_0(X_i)-\Gamma_q)Z_i}{q(X_i)}\right\}+o_p(1).$$

Combining Lemma 3 with a Taylor expansion argument as above establishes the influence function representation (9), from which it follows that $\sqrt{n}(\hat{\tau}_t-\tau_t)\xrightarrow{d}\mathcal{N}(0,\mathcal{V}_t)$.

We complete the proof of Theorem 2 by verifying Lemma 2. The proof of Lemma 3 is omitted. ∎

The proof of Lemma 2 Our argument is based on Ichimura and Linton (2002) with the generalization that X is allowed to be an r-dimensional vector rather than a scalar. For a matrix $A=(a_{ij})$ we write $\|A\|_\infty=\sup|a_{ij}|$ and $\|A\|_1=\sum|a_{ij}|$.

STEP 1 (Some properties of $\hat{q}(X_i)$). For $x \in \mathbb{R}^r$ and $\lambda \in \mathbb{N}^r$ define $x^\lambda = x_1^{\lambda_1} \cdot \ldots \cdot x_r^{\lambda_r} \in \mathbb{R}$. For a non-negative integer ℓ, let $x^{\Lambda(\ell)}$ denote the vector $(x^\lambda)_{\lambda_1+\ldots+\lambda_r=\ell}$ (along with some rule to order these elements). Thus, $x^{\Lambda(\ell)}$ contains all polynomial terms of order exactly ℓ that can be constructed from the components of x, and is interpreted as a row vector if x is a row vector and as a column vector if x is a column vector. E.g., $x^{\Lambda(0)} = 1$, $(x_1, \ldots, x_r)^{\Lambda(1)} = (x_1, \ldots, x_r)$, $(x_1, x_2)^{\Lambda(2)} = (x_1^2, x_2^2, x_1 x_2)$, etc. For each observation X_t on the vector of covariates, we define

$$\tilde{X}_t^{(i)} = [(X_t' - X_i')^{\Lambda(0)}, (X_t' - X_i')^{\Lambda(1)}, \ldots, (X_t' - X_i')^{\Lambda(r)}]'.$$

Then the leave-one-out local polynomial regression estimator of $q(X_i)$ is the first component of the vector $\hat{\beta}$ that solves $\min_\beta \sum_{t:t\neq i} K\left(\frac{X_t - X_i}{h}\right)(Z_t - \tilde{X}_t^{(i)\prime}\beta)^2$. Letting e_1 denote the first unit vector having the same dimension as $X_t^{(i)}$, we can write this estimator as

$$\hat{q}(X_i) = e_1' \left(\sum_{t:t\neq i} K\left(\frac{X_t - X_i}{h}\right) \tilde{X}_t^{(i)} \tilde{X}_t^{(i)\prime} \right)^{-1} \sum_{t:t\neq i} K\left(\frac{X_t - X_i}{h}\right) \tilde{X}_t^{(i)} Z_t = \sum_{t:t\neq i} \omega_{it} Z_t,$$

where ω_{it} depends only on $X_1, ..., X_n$ and is given by

$$\omega_{it} = e_1' \left(\sum_{j:j\neq i} K\left(\frac{X_j - X_i}{h}\right) \tilde{X}_j^{(i)} \tilde{X}_j^{(i)\prime} \right)^{-1} \tilde{X}_t^{(i)} K\left(\frac{X_t - X_i}{h}\right). \tag{23}$$

The first property we will need in later arguments is a bound on $|\omega_{it} - \omega_{ti}|$. By Assumption 7, $\omega_{it} = \omega_{ti} = 0$ for $\|X_t - X_i\|_\infty > h$. Now assume $\|X_t - X_i\|_\infty \leq h$. Let $H = diag(1, hu_1, ..., h^r u_r)$, where u_j is a vector of ones with the same dimensionality as $(X_t' - X_i')^{\Lambda(j)}$. Then, noting that $e_1' H = e_1'$, we can write

$$\omega_{it} = e_1' \left(H^{-1} \frac{1}{nh^r} \sum_{j:j\neq i} K\left(\frac{X_j - X_i}{h}\right) \tilde{X}_j^{(i)} \tilde{X}_j^{(i)\prime} H^{-1} \right)^{-1} H^{-1} \frac{1}{nh^r} \tilde{X}_t^{(i)} K\left(\frac{X_t - X_i}{h}\right).$$

Let the matrix inside the inverse operator be denoted as $\widehat{\mathcal{K}}(X_i)$. Then

$$\begin{aligned} |\omega_{it} - \omega_{ti}| &= \frac{1}{nh^r} \left| K\left(\frac{X_i - X_t}{h}\right) \right| \times \left| e_1' \widehat{\mathcal{K}}^{-1}(X_i) H^{-1} \tilde{X}_t^{(i)} - e_1' \widehat{\mathcal{K}}^{-1}(X_t) H^{-1} \tilde{X}_i^{(t)} \right| \\ &\leq \frac{1}{nh^r} \left| K\left(\frac{X_i - X_t}{h}\right) \right| \\ &\times \left\{ \left| e_1' [\widehat{\mathcal{K}}^{-1}(X_i) - \widehat{\mathcal{K}}^{-1}(X_t)] H^{-1} \tilde{X}_t^{(i)} \right| + \left| e_1' \widehat{\mathcal{K}}^{-1}(X_t) H^{-1} [\tilde{X}_t^{(i)} - \tilde{X}_i^{(t)}] \right| \right\}. \end{aligned} \tag{24}$$

The first term in the braces in (24) is bounded as follows. The elements of $\widehat{\mathcal{K}}(X_i)$ are of the form

$$\frac{1}{nh^{r+\sum \lambda_k}} \sum_{j:j\neq i} (X_j - X_i)^\lambda K\left(\frac{X_j - X_i}{h}\right) \tag{25}$$

for some r-vector of nonnegative integers λ with $0 \leq \sum \lambda_k \leq 2r$. Using arguments similar to those in, e.g., Section 3.7 of Fan and Gijbels (1996), one can show that (25) converges in probability to $f(X_i) \int u^\lambda K(u) du$ uniformly in X_i. Hence $\sup_i \|\widehat{\mathcal{K}}(X_i) - f(X_i)\mathcal{K}\|_\infty = o_p(1)$ for a constant, symmetric matrix $\mathcal{K}$ that only

depends on the kernel K. By continuity and $\mathcal{X}$ compact, it follows that $\sup_i \|\widehat{\mathcal{K}}^{-1}(X_i) - \frac{1}{f(X_i)}\mathcal{K}^{-1}\|_\infty = o_p(1)$. Since f is continuously differentiable and is bounded away from zero, $|f^{-1}(x_1) - f^{-1}(x_2)| \leq C\|x_1 - x_2\|_\infty$ for all $x_1, x_2 \in \mathcal{X}$. Combining these observations yields

$$\begin{aligned}
&\|\widehat{\mathcal{K}}^{-1}(X_i) - \widehat{\mathcal{K}}^{-1}(X_t)\|_\infty \\
\leq &\sup_i \|\widehat{\mathcal{K}}^{-1}(X_i) - f^{-1}(X_i)\mathcal{K}^{-1}\|_\infty + \|f^{-1}(X_i)\mathcal{K}^{-1} - f^{-1}(X_t)\mathcal{K}^{-1}\|_\infty + \sup_t \|\widehat{\mathcal{K}}^{-1}(X_t) - f^{-1}(X_t)\mathcal{K}^{-1}\|_\infty \\
\leq &C\|X_i - X_t\|_\infty + o_p(1) = Ch + o_p(1) \equiv M_n,
\end{aligned}$$

where $M_n = o_p(1)$ and is independent of X_i and X_t. Hence the first term in the braces in (24) is bounded by $M_n\|H^{-1}X_t^{(i)}\|_1 \leq \tilde{M}_n$, where $\tilde{M}_n$ incorporates a multiplicative constant that only depends on r and bounds $\|H^{-1}X_t^{(i)}\|_1$ (each component of $H^{-1}X_t^{(i)}$ is bounded by one since $\|X_t - X_i\|_\infty \leq h$).

The second term within the braces in (24) is bounded as follows. Suppose that r is even. Then

$$H^{-1}[\tilde{X}_t^{(i)} - \tilde{X}_i^{(t)}] = \left(0, \quad 2\left(\frac{X_t' - X_i'}{h}\right)^{\Lambda(1)}, \quad z_2 \quad , \quad 2\left(\frac{X_t' - X_i'}{h}\right)^{\Lambda(3)}, \quad \dots \quad , \quad z_r\right)', \tag{26}$$

where the z_j, j even, are zero vectors with the same dimensionality as $(X_t' - X_i')^{\Lambda(j)}$. (The only difference when r is odd is that the last term in this alternating partition is not a zero vector.) Note that each component of (26) is bounded by 2 since $\|X_t - X_i\|_\infty \leq h$. Therefore it is possible to write

$$\left|e_1'\widehat{\mathcal{K}}^{-1}(X_t)H^{-1}[\tilde{X}_t^{(i)} - \tilde{X}_i^{(t)}]\right| \leq \left|\frac{1}{f(X_t)}e_1'\mathcal{K}^{-1}H^{-1}[\tilde{X}_t^{(i)} - \tilde{X}_i^{(t)}]\right| + R_n,$$

where $R_n = o_p(1)$ and does not depend on X_i or X_t. By the symmetry properties of the kernel (Assumption 7), the first row of the matrix $\mathcal{K}$ has zeros precisely at those positions at which the vector $H^{-1}[\tilde{X}_t^{(i)} - \tilde{X}_i^{(t)}]$ is nonzero. A straightforward linear algebra argument (available on request) shows that the first row of the matrix $\mathcal{K}^{-1}$ also has zeros at the same positions. Hence, the second term in the braces in (24) is simply bounded by R_n. Combining the bounds on the components of (24) shows that

$$|\omega_{it} - \omega_{ti}| \leq \frac{\tilde{M}_n + R_n}{nh^r}\left|K\left(\frac{X_t - X_i}{h}\right)\right|, \tag{27}$$

where $\tilde{M}_n + R_n = o_p(1)$ and does not depend on X_i or X_t. Finally, observe that the inequality holds for any value of h, not just for $\|X_t - X_i\|_\infty \leq h$. This is the bound we will need.

The second property of $\hat{q}(X_i)$ that we will make use of is its uniform convergence rate. As shown by Masry (1996), the "include all" version of the estimator satisfies

$$\sup_{x \in \mathcal{X}} |\hat{q}(x) - q(x)| = O_p\left(h^{r+1} + \sqrt{\frac{\log n}{nh^r}}\right).$$

Given the range of h in Assumption 8 it follows that $\sup_i |\hat{q}(X_i) - q(X_i)| = o_p(n^{-1/4})$ for the include all as well as the leave-one-out version.

STEP 2 (Expanding $\hat{\Delta}$). We define notation similar to that in Ichimura and Linton (2002). Let $w = (y, d, z, x)$ and

$$\Psi(w, \Delta, q) \equiv \frac{zy}{q} - \frac{(1-z)y}{1-q} - \Delta.$$

Let Ψ_q and Ψ_{qq} denote the partial derivative of Ψ w.r.t. the argument q, and let $W_i = (Y_i, D_i, Z_i, X_i)$. Then

$$\begin{aligned}
\Psi(W_i, \Delta, q(X_i)) &= \frac{Z_i Y_i}{q(X_i)} - \frac{(1-Z_i)Y_i}{1-q(X_i)}, \\
\Psi_q(W_i, \Delta, q(X_i)) &= -\left(\frac{Z_i Y_i}{q^2(X_i)} + \frac{(1-Z_i)Y_i}{(1-q(X_i))^2} \right), \\
\Psi_{qq}(W_i, \Delta, q(X_i)) &= \frac{2Z_i Y_i}{q^3(X_i)} - \frac{2(1-Z_i)Y_i}{(1-q(X_i))^3},
\end{aligned}$$

and we further define

$$\begin{aligned}
S_q(X_i) &= E[\Psi_q(W_i, \Delta, q(X_i))|X_i] = -\left(\frac{m_1(X_i)}{q(X_i)} + \frac{m_0(X_i)}{1-q(X_i)} \right), \\
\zeta_i &= \Psi_q(W_i, \Delta, q(X_i)) - S_q(X_i), \\
\epsilon_i &= Z_i - q(X_i), \\
\beta_n(X_i) &= E[\hat{q}(X_i)|X_1, ..., X_n] - q(X_i) = \sum_{j:i \neq j} \omega_{ij} q(X_j) - q(X_i),
\end{aligned}$$

where the last quantity is the bias of the estimator conditional on $X_1, \ldots, X_n$.

By a Taylor series expansion around $q(X_i)$,

$$\begin{aligned}
\sqrt{n}(\widehat{\Delta} - \Delta) =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi(W_i, \Delta, \hat{q}(X_i)) \\
=& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi(W_i, \Delta, q(X_i)) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_q(W_i, \Delta, q(X_i))(\hat{q}(X_i) - q(X_i)) \\
&+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_{qq}(W_i, \Delta, q^*(X_i))(\hat{q}(X_i) - q(X_i))^2 \\
=& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi(W_i, \Delta, q(X_i)) \\
&+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_q(X_i)(\hat{q}(X_i) - q(X_i)) \\
&+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_i(\hat{q}(X_i) - q(X_i)) \\
&+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_{qq}(W_i, \Delta, q^*(X_i))(\hat{q}(X_i) - q(X_i))^2 \\
\equiv& J_0 + J_1 + J_2 + J_3,
\end{aligned}$$

where $q^*(X_i)$ is a value between $\hat{q}(X_i)$ and $q(X_i)$ for all i, and the J's are defined line by line. We further

expand the J_1 term as

$$\begin{aligned}
J_1 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_q(X_i)(\hat{q}(X_i) - q(X_i)) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_q(X_i)\Big(\sum_{j:j\neq i}\omega_{ij}Z_j - q(X_i)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_q(X_i)\Big(\sum_{j:j\neq i}\omega_{ij}(\epsilon_j + q(X_j)) - q(X_i)\Big) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_q(X_i)\epsilon_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_q(X_i)\Big(\sum_{j:j\neq i}\omega_{ij}\epsilon_j - \epsilon_i\Big) \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_q(X_i)\Big(\sum_{j:j\neq i}\omega_{ij}q(X_j) - q(X_i)\Big) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_q(X_i)\epsilon_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\Big(\sum_{j:j\neq i}\omega_{ji}S_p(X_j) - S_p(X_i)\Big) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_q(X_i)\beta_n(X_i) \\
&\equiv J_{11} + J_{12} + J_{13}.
\end{aligned}$$

STEP 3 (Evaluating J_0, J_{11}, J_{12}, J_{13}, J_2 and J_3). By the central limit theorem, J_0 and J_{11} are $O_p(1)$ and together they give the influence function representation in Lemma 2. We will show that the rest of the terms are $o_p(1)$.

For J_{12}, we claim that $\omega_{ij} \approx \omega_{ji}$ in that $\sup_i \Big|\sum_{j:i\neq j}(\omega_{ji} - \omega_{ij})S_p(X_j)\Big| = o_p(1)$. By the bound in (27),

$$\begin{aligned}
&\sup_i \Big|\sum_{j:i\neq j}(\omega_{ji} - \omega_{ij})S_p(X_j)\Big| \leq \sup_i \sum_{j:i\neq j}|(\omega_{ji} - \omega_{ij})||S_p(X_j)| \\
\leq& C(\tilde{M}_n + R_n)\sup_i \sum_{j:j\neq i}\frac{1}{nh^r}\Big|K\Big(\frac{X_j - X_i}{h}\Big)\Big| = o_p(1)\cdot O_p(1) = o_p(1).
\end{aligned}$$

The second inequality holds since $S_p(x)$ is bounded on $\mathcal{X}$. Further,

$$\sup_{x\in\mathcal{X}}\left|\sum_{i=1}^{n}\frac{1}{nh^r}\Big|K\Big(\frac{X_j - X_i}{h}\Big)\Big| - f(x)\int|K(u)|du\right| = o_p(1),$$

which implies $\sup_i \sum_{j:j\neq i}\frac{1}{nh^r}\left|K\left(\frac{X_j - X_i}{h}\right)\right| = O_p(1)$. Given

$$\sup_i \Big|\sum_{j:j\neq i}\omega_{ij}S_p(X_j) - S_p(X_i)\Big| = o_p(1),$$

it is true that conditional on the sample path of the X_i, with probability approaching one, $\sum_{j:j\neq i}\omega_{ji}S_p(X_j) - S_p(X_i)$ is uniformly bounded over i and converges to zero uniformly over i. Also, the ϵ_i are mutually independent conditional on the sample path of the X_i. Hence, conditional on the sample path of the X_i with probability approaching one,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\Big(\sum_{j:j\neq i}\omega_{ji}S_p(X_j) - S_p(X_i)\Big) = o_p(1),$$

which is sufficient to show that $J_{12} = o_p(1)$.

For J_{13}, observe that $\beta_n(x)$ is the (conditional) bias of $\hat{q}(x)$, which is of order h_n^{r+1} uniformly (Masry 1996). By the assumptions on h_n, we have $\sup_{x\in\mathcal{X}}|\beta_n(x)| = o_p(n^{-1/2})$. It follows that

$$|J_{13}| = \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_p(X_i)\beta_n(X_i)\Big| \leq \sup_{x\in\mathcal{X}}\big|\sqrt{n}\beta_n(x)\big|\frac{1}{n}\sum_{i=1}^{n}|S_p(X_i)| = o_p(1)\cdot O_p(1) = o_p(1).$$

For J_2, observe that $\sup_{x\in\mathcal{X}}|\hat{q}(x) - q(x)| = o_p(1)$ and argue similarly as in showing $J_{12} = o_p(1)$.

Finally, for J_3. Given that $\hat{q}(x)$ is uniformly bounded in probability on $\mathcal{X}$, and $q^*(X_i)$ is between $\hat{q}(X_i)$ and $q(X_i)$, it follows that $q^*(X_i)$ is uniformly bounded and also bounded away from zero in probability. Also, $\sup_i|\hat{q}(X_i) - q(X_i)| = o_p(n^{-1/4})$, so $\sup_i n^{1/2}(\hat{q}(X_i) - q(X_i))^2 = o_p(1)$. Hence,

$$\begin{aligned}&\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_{qq}(W_i,\Delta,q^*(X_i))(\hat{q}(X_i) - q(X_i))^2\Big| \\ \leq&\Big(\sup_i\frac{1}{\sqrt{n}}(\hat{q}(X_i) - q(X_i))^2\Big)\frac{1}{n}\sum_{i=1}^{n}\Big|\Psi_{qq}(W_i,\Delta,q^*(X_i))\Big| = o_p(1)\cdot O_p(1) = o_p(1).\end{aligned}$$

As a result, we have

$$\sqrt{n}(\widehat{\Delta} - \Delta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi(W_i,\Delta,q(X_i)) + S_q(X_i)\epsilon_i + o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta(Y_i,D_i,Z_i,X_i) + o_p(1),$$

where the function $\delta(\cdot)$ is as defined in Lemma 2. ∎

C. Second order asymptotic results

Inverse probability weighted estimators of Δ and Γ with bias correction Following Ichimura and Linton (2002), we define $\hat{\Delta}_b = \hat{\Delta} - \frac{\hat{b}}{2nh}$, where $\hat{\Delta}$ is the numerator of $\hat{\tau}$, and

$$\hat{b} = \frac{1}{n}\sum_{i=1}^{n}\|K\|^2\hat{S}_{qq}(X_i)\frac{\hat{q}(X_i)[1-\hat{q}(X_i)]}{\hat{f}(X_i)}, \tag{28}$$

with

$$\hat{S}_{qq}(X_i) = \frac{2\hat{m}_1(X_i)}{\hat{q}^2(X_i)} - \frac{2\hat{m}_0(X_i)}{[1-\hat{q}(X_i)]^2}.$$

$\hat{S}_{qq}(X_i)$ is of course an estimate of the function $S_{qq}(X_i) = E[\Psi_{qq}(W_i,\Delta,q(X_i)) \mid X_i]$; cf. Appendix B. The nonparametric estimates $\hat{m}_0$, $\hat{m}_1$ and $\hat{q}$ are based are on local linear regression (use the leave-one-out version for q). A kernel density estimator can used be for $\hat{f}$. The estimator $\hat{\Gamma}_b$ is defined analogously (replace Y_i with D_i throughout). ∎

Justifying equation (16) Substituting the influence function representations (10) into (13) gives

$$
\begin{aligned}
&-\frac{2}{\sqrt{n}}\left(\frac{1}{\Gamma}\sqrt{n}(\hat{\Delta}-\Delta)-\frac{\Delta}{\Gamma^2}\sqrt{n}(\hat{\Gamma}-\Gamma)\right)\left(\frac{1}{\Gamma^2}n(\hat{\Delta}-\Delta)(\hat{\Gamma}-\Gamma)+\frac{2\Delta}{\Gamma^3}n(\hat{\Gamma}-\Gamma)^2\right)\\
&=-\frac{2}{\Gamma\sqrt{n}}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\delta_i-\tau\gamma_i)+o_p(n^{-\alpha})\right]\\
&\times\left[\frac{1}{\Gamma^2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i+o_p(n^{-\alpha})\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma_i+o_p(n^{-\alpha})\right)+\frac{2\Delta}{\Gamma^3}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma_i+o_p(n^{-\alpha})\right)^2\right]\\
&=-\frac{2}{n^2\Gamma^3}\left(\sum_{i=1}^{n}(\delta_i-\tau\gamma_i)\right)\left(\sum_{i=1}^{n}\delta_i\right)\left(\sum_{i=1}^{n}\gamma_i\right)-\frac{4\tau}{n^2\Gamma^3}\left(\sum_{i=1}^{n}(\delta_i-\tau\gamma_i)\right)\left(\sum_{i=1}^{n}\gamma_i\right)^2+o_p(n^{-\alpha-1/2}).
\end{aligned}
$$

Considering the expectation of the first two terms:

$$
-\frac{2}{n^2\Gamma^3}E\left[\left(\sum_{i=1}^{n}(\delta_i-\tau\gamma_i)\right)\left(\sum_{i=1}^{n}\delta_i\right)\left(\sum_{i=1}^{n}\gamma_i\right)\right]=-\frac{2}{n^2\Gamma^3}\sum_{i=1}^{n}E[(\delta_i-\tau\gamma_i)\delta_i\gamma_i]=-\frac{2}{n\Gamma^3}E[(\delta_1-\tau\gamma_1)\delta_1\gamma_1],
$$

where, by random sampling, all cross product terms have zero expectations. Similarly,

$$
-\frac{4\tau}{n^2\Gamma^3}E\left[\left(\sum_{i=1}^{n}(\delta_i-\tau\gamma_i)\right)\left(\sum_{i=1}^{n}\gamma_i\right)^2\right]=-\frac{4\tau}{n\Gamma^3}E[(\delta_1-\tau\gamma_1)\gamma_1^2].
$$

Combining the two results above with the expectation of the remainder term gives equation (16). ∎

The proof of Theorem 3 As argued in the main text, the non-neglected terms in expansions (17) and (18) correspond to the first three terms of equation (15); the remainder term combines the remainder in (15), all terms in (16) and the expectation of (14). As shown by Lemmas 4 and 5 below, one can take $\alpha=\frac{1}{10}$ in (16) and hence the remainder term is of the stated order (specifically, $o_p(n^{-3/5})$). Thus, we can complete the proof by arguing that a mean square expansion of the form (15) is valid for both types of estimators and verifying the expressions for the constants C_0, C_1 and C_2 given in Theorem 3.

Let $\tilde{Y}=Y-\tau D$. To derive equation (15) for the imputation estimator first note that one can write

$$
\hat{\Delta}_m=\frac{1}{n}\sum_{i=1}^{n}[\hat{m}_1(X_i)-\hat{m}_0(X_i)]=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}Y_j\quad\text{and}\quad\hat{\Gamma}_m=\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_1(X_i)-\hat{\mu}_0(X_i)]=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}D_j
$$

where the weights α_{ij} depend only on $X_1,\ldots,X_n$ and $Z_1,\ldots,Z_n$. Therefore,

$$
\begin{aligned}
&\frac{1}{\Gamma}(\hat{\Delta}_m-\Delta)-\frac{\Delta}{\Gamma^2}(\hat{\Gamma}_m-\Gamma)=\frac{1}{\Gamma}(\hat{\Delta}_m-\tau\hat{\Gamma}_m)-0\\
&=\frac{1}{n\Gamma}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}\tilde{Y}_j=\frac{1}{n\Gamma}\sum_{i=1}^{n}[\widehat{\tilde{m}}_1(X_i)-\widehat{\tilde{m}}_0(X_i)],
\end{aligned}
\tag{29}
$$

where $\widehat{\tilde{m}}_1$ and $\widehat{\tilde{m}}_0$ are the local linear regression estimates of the functions $\tilde{m}_1(x)=E[\tilde{Y}\mid Z=1,X=x]$ and $\tilde{m}_0(x)=E[\tilde{Y}\mid Z=0,X=x]$. Regarding (29) as an estimator of $\frac{1}{\Gamma}(\Delta-\tau\Gamma)=0$, the mean square expansion given in equation (15) follows directly from Lemma 4.3 of Kalyanaraman (2009). Verifying that $C_0=G_0$,

$C_1 = G_1$ and $C_2 = G_2$ is straightforward based on the generic expressions provided by Kalyanaraman (2009). (That $C_0 = G_0 = F_0$ also follows from the first order asymptotic results given in our Theorem 2 and the fact that $\hat{\tau}$, $\hat{\tau}_b$ and $\hat{\tau}_m$ are asymptotically equivalent.)

To derive equation (15) for the inverse probability weighted estimator we write

$$\begin{aligned}
&\frac{1}{\Gamma}(\hat{\Delta}_b - \Delta) - \frac{\Delta}{\Gamma^2}(\hat{\Gamma}_b - \Gamma) = \frac{1}{\Gamma}(\hat{\Delta}_b - \tau\hat{\Gamma}_b) - 0 \\
&= \frac{1}{n\Gamma}\sum_{i=1}^{n}\left\{\frac{Z_i\tilde{Y}_i}{\hat{q}(X_i)} - \frac{(1-Z_i)\tilde{Y}_i}{1-\hat{q}(X_i)}\right\} - \frac{1}{n\Gamma}\sum_{i=1}^{n}\|K\|^2\widehat{\tilde{S}}_{qq}(X_i)\frac{\hat{q}(X_i)[1-\hat{q}(X_i)]}{\hat{f}(X_i)},
\end{aligned} \tag{30}$$

where $\widehat{\tilde{S}}_{qq}(X_i) = \frac{2\widehat{m}_1(X_i)}{\hat{q}^2(X_i)} - \frac{2\widehat{m}_0(X_i)}{[1-\hat{q}(X_i)]^2}$. Regarding (30) as an estimator of $\frac{1}{\Gamma}(\Delta - \tau\Gamma) = 0$, the mean square expansion given in equation (15) is a direct application of Theorem 2 in Ichimura and Linton (2002). Verifying that $C_0 = F_0$, $C_1 = F_1$ and $C_2 = F_2$ is a matter of tedious but straightforward calculations. ∎

Lemma 4 *Under the assumptions of Theorem 3,* $\sqrt{n}(\widehat{\Delta}_b - \Delta) = n^{-1/2}\sum_{i=1}^{n}\delta_i + o_p(n^{-1/10})$ *and* $\sqrt{n}(\widehat{\Gamma}_b - \Gamma) = n^{-1/2}\sum_{i=1}^{n}\gamma_i + o_p(n^{-1/10})$.

Proof: Let $\hat{\Delta}$ denote the inverse probability weighted estimator of Δ without bias correction (i.e., $\hat{\Delta}$ is the numerator of $\hat{\tau}$). For $h \propto n^{-2/5}$, the expansion of $\sqrt{n}(\hat{\Delta} - \Delta)$ given in equation (17) of Ichimura and Linton (2002) is almost sufficient to show $\sqrt{n}(\hat{\Delta} - \Delta) = n^{-1/2}\sum_{i=1}^{n}\delta_i + o_p(n^{-1/10})$ except that the last term is $O_p(n^{-1/10})$ instead of $o_p(n^{-1/10})$. This term can be further expanded as:

$$\begin{aligned}
&\sqrt{n}(\hat{\Delta} - \Delta) \\
&= \ldots + \frac{1}{2h\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n}\|K\|^2 S_{qq}(X_i)\frac{q(X_i)[1-q(X_i)]}{f(X_i)} + O_p(h^4\sqrt{n}) \\
&= \ldots + \frac{1}{2h\sqrt{n}}\hat{b} + \frac{1}{2h\sqrt{n}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\|K\|^2 S_{qq}(X_i)\frac{q(X_i)[1-q(X_i)]}{f(X_i)} - b\right) + (b - \hat{b})\right] + O_p(h^4\sqrt{n}),
\end{aligned} \tag{31}$$

where $b = \|K\|^2 E\left[S_{qq}(X_i)\frac{q(X_i)[1-q(X_i)]}{f(X_i)}\right]$ and $\hat{b}$ is as defined in (28). The first term in (31) is $O_p(n^{-1/10})$ and corresponds to the bias correction applied to $\hat{\Delta}$. Thus, it disappears from the analogous expansion of $\hat{\Delta}_b$. The two terms inside the square brackets are both $O_p(n^{-1/2})$; if multiplied by $\sqrt{n}$, the first one satisfies a CLT and, as argued by Ichimura and Linton (2002), $\sqrt{n}(\hat{b} - b)$ admits an influence function representation. Since $\frac{1}{h\sqrt{n}}O_p(n^{-\frac{1}{2}}) = O_p(n^{-\frac{1}{2}-\frac{1}{10}}) = o_p(n^{-\frac{1}{10}})$ and $O_p(h^4\sqrt{n}) = o_p(n^{-\frac{1}{10}})$, the proof is complete. ∎

Lemma 5 *Under the assumptions of Theorem 3,* $\sqrt{n}(\widehat{\Delta}_m - \Delta) = n^{-1/2}\sum_{i=1}^{n}\delta_i + o_p(n^{-1/10})$ *and* $\sqrt{n}(\widehat{\Gamma}_m - \Gamma) = n^{-1/2}\sum_{i=1}^{n}\gamma_i + o_p(n^{-1/10})$.

The proof of Lemma 5 relies on a sequence of auxiliary lemmas. These will be stated and proven first and then we return to the proof of Lemma 5. For a given sample size n, let $n_1 \equiv \sum_{i=1}^{n}(Z_i = 1)$, and suppose that the observations in the sample are arranged so that so that the first n_1 are those with $Z_i = 1$. By

the law of large numbers $n_1/n \to P(Z=1)$ for almost all realizations of $\{Z_i\}_{i=1}^{\infty}$, and hence $n_1 \to \infty$ as $n \to \infty$, at the same rate, almost surely. All arguments in the rest of Appendix C will be conditional on the realization of the instrument sequence $\{Z_i\}_{i=1}^{\infty}$. This conditioning will not be made explicit in the notation (e.g. when taking expectations). All subsequent results will hold for almost all realizations of $\{Z_i\}_{i=1}^{\infty}$, and hence unconditionally as well.

We rewrite the the imputation estimator for Δ as

$$\widehat{\Delta}_m = \frac{1}{n}\sum_{i=1}^{n}[\hat{m}_1(X_i) - \hat{m}_0(X_i)] = \frac{1}{n}\sum_{i=1}^{n_1}[\hat{m}_1(X_i) - \hat{m}_0(X_i)] + \frac{1}{n}\sum_{i=n_1+1}^{n}[\hat{m}_1(X_i) - \hat{m}_0(X_i)],$$

and define

$$\begin{aligned}
&\nu_j = Y_j - m_1(X_j) \\
&M_{1n}(x) = \begin{pmatrix} \kappa_0(x) & \kappa_1(x) \\ \kappa_1(x) & \kappa_2(x) \end{pmatrix}, \quad \kappa_\ell(x) = \int \frac{1}{h}\Big(\frac{t-x}{h}\Big)^\ell K\Big(\frac{t-x}{h}\Big) f_{X|Z=1}(t)dt, \\
&\widehat{M}_{1n}(x) = \begin{pmatrix} \hat{\kappa}_0(x) & \hat{\kappa}_1(x) \\ \hat{\kappa}_1(x) & \hat{\kappa}_2(x) \end{pmatrix}, \quad \hat{\kappa}_\ell(x) = \frac{1}{n_1 h}\sum_{j:Z_j=1}\Big(\frac{X_j-x}{h}\Big)^\ell K\Big(\frac{X_j-x}{h}\Big), \\
&\xi_1^m(Y_j, X_j, x) = (1,0)\cdot M_{1n}^{-1}(x)\cdot \begin{pmatrix} 1 \\ \frac{X_j-x}{h} \end{pmatrix} \cdot K\Big(\frac{X_j-x}{h}\Big)\cdot \nu_j.
\end{aligned}$$

Using the notation defined above, we can express the estimator $\hat{m}_1(x)$ in a way similar to Appendix B:

$$\hat{m}_1(x) = (1,0)\widehat{M}_{1n}^{-1}(x)\frac{1}{n_1 h}\sum_{j=1}^{n_1}\begin{pmatrix} 1 \\ \frac{X_j-x}{h} \end{pmatrix} K\Big(\frac{X_j-x}{h}\Big)Y_j.$$

Therefore,

$$\begin{aligned}
\hat{m}_1(x) - m_1(x) &= (1,0)\widehat{M}_{1n}^{-1}(x)\frac{1}{n_1 h}\sum_{j=1}^{n_1}\begin{pmatrix} 1 \\ \frac{X_j-x}{h} \end{pmatrix} K\Big(\frac{X_j-x}{h}\Big)\nu_j \\
&+ (1,0)\widehat{M}_{1n}^{-1}(x)\frac{1}{n_1 h}\sum_{j=1}^{n_1}\begin{pmatrix} 1 \\ \frac{X_j-x}{h} \end{pmatrix} K\Big(\frac{X_j-x}{h}\Big)m_1(X_j) - m_1(x) \\
&\equiv A(x) + B(x),
\end{aligned}$$

where $A(x)$ and $B(x)$ are defined line by line.

Lemma 6 *Under the assumptions of Theorem 3,*

$$A(x) = \frac{1}{n_1}\sum_{j=1}^{n_1}\xi_1^m(Y_j, X_j, x) + R_1(x),$$

where $n_1^{-1/2}\sum_{j=1}^{n_1} R_1(X_j) = o_p(n^{-1/10})$.

Proof of Lemma 6: The lemma is similar to Lemma 2 of Heckman et al. (1998) but with the stronger conclusion that $n_1^{-1/2}\sum_{j=1}^{n_1} R_1(X_j)$ converges to zero at the rate $o_p(n^{-1/10})$. (Also, we specialize to the case where X is a scalar.) We write

$$\begin{aligned} A(x) &= (1,0)M_{1n}^{-1}(x)\frac{1}{n_1 h}\sum_{j=1}^{n_1}\begin{pmatrix}1\\ \frac{X_j-x}{h}\end{pmatrix}K\Big(\frac{X_j-x}{h}\Big)\nu_j \\ &+ (1,0)(\widehat{M}_{1n}^{-1}(x)-M_{1n}^{-1}(x))\frac{1}{n_1 h}\sum_{j=1}^{n_1}\begin{pmatrix}1\\ \frac{X_j-x}{h}\end{pmatrix}K\Big(\frac{X_j-x}{h}\Big)\nu_j \\ &= \frac{1}{n_1}\sum_{j=1}^{n_1}\xi_1^m(Y_j,X_j,x)+R_1(x), \end{aligned}$$

where $R_1(x)$ is defined by the expression on the second line.

Let $(r_{n0}^1(x), r_{n0}^2(x)) = (1,0)M_{1n}^{-1}(x)$ be the first row of $M_{1n}^{-1}(x)$, and $(\hat{r}_n^1(x), \hat{r}_n^2(x)) = (1,0)\widehat{M}_{1n}^{-1}(x)$ be the first row of $\widehat{M}_{1n}^{-1}(x)$. Given $\epsilon > 0$, we define the ϵ-neighborhoods of the functions $r_{n0}^\ell(x)$, $\ell = 1,2$, as

$$N_n^\ell = \Big\{r(x) : \sup_{x\in\mathcal{X}}|r(x)-r_{n0}^\ell(x)| \le \epsilon\Big\}, \quad \ell = 1,2.$$

We consider two zero mean stochastic processes indexed by the elements of N_n^ℓ, $\ell = 1,2$:

$$\sum_{j=1}^{n_1}\sum_{i=1}^{n_1} g_{r_n}^\ell(\nu_i, X_i, X_j), \quad r_n \in N_n^\ell,$$

where

$$g_{r_n}^\ell(\nu_i, X_i, X_j) = n^{-\frac{3}{2}}h^{-1}r_n(X_j)\Big(\frac{X_i-X_j}{h}\Big)^{\ell-1}K\Big(\frac{X_i-X_j}{h}\Big)\nu_i.$$

Interest in these processes is justified by the fact that

$$\frac{1}{\sqrt{n_1}}\sum_{j=1}^{n_1}R_1(X_j) = \sum_{j=1}^{n_1}\sum_{i=1}^{n_1} g_{\hat{r}_n^1}^1(\nu_i,X_i,X_j) - g_{r_{n0}^1}^1(\nu_i,X_i,X_j) \tag{32}$$

$$+\sum_{j=1}^{n_1}\sum_{i=1}^{n_1} g_{\hat{r}_n^2}^2(\nu_i,X_i,X_j) - g_{r_{n0}^2}^2(\nu_i,X_i,X_j). \tag{33}$$

It follows from Lemmas 5 and 6 of Heckman et al. (1998) that

$$\sup_{x\in\mathcal{X}}|\hat{r}_n^\ell(x) - r_{n0}^\ell(x)| = o_p(\sqrt{h}), \tag{34}$$

so that $\hat{r}_n^\ell \in N_n^\ell$ for n sufficiently large.

Our goal is to show that (32) and (33) are both $o_p(n^{-1/10})$. First set $\ell = 1$. For any fixed sequence $r_n \in N_n^1$, decompose the process of interest as

$$\sum_{j=1}^{n_1}\sum_{i=1}^{n_1} g_{r_n}^1(\nu_i,X_i,X_j) = \sum_{i=1}^{n_1} g_{r_n}^1(\nu_i,X_i,X_i) + \sum_{j\neq i} g_{r_n}^1(\nu_i,X_i,X_j). \tag{35}$$

For the first term on the rhs of (35), we have

$$\sum_{i=1}^{n_1} g^1_{r_n}(\nu_i, X_i, X_i) = \frac{1}{nh}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n_1} r_n(X_i)\nu_i K(0)\right) = O_p(n^{-1}h^{-1}) = O_p(n^{-\frac{3}{5}})$$

where the result follows from the CLT, the fact that $E[r_n(X_i)K(0)\nu_i] = 0$, and $h \propto n^{-\frac{2}{5}}$. Further, it is possible to show that the process $\sum_{i=1}^{n_1} g^1_{r_n}(\nu_i, X_i, X_i)$ is stochastically equicontinuous in r_n in a neighborhood around r^1_{n0}. The argument is very similar to that used by Heckman et al. (1998) following their Lemma 3, and is omitted. Then (34) implies, at least,

$$\sum_{i=1}^{n_1} g^1_{\hat{r}^1_n}(\nu_i, X_i, X_i) - \sum_{i=1}^{n_1} g^1_{r^1_{n0}}(\nu_i, X_i, X_i) = o_p(n^{-3/5}). \tag{36}$$

For the second term in (35), we further define

$$\begin{aligned}
&\vartheta_i \equiv (\nu_i, X_i),\\
&g_{r_n}(\vartheta_i, \vartheta_j) = \frac{g^1_{r_n}(\nu_i, X_i, X_j) + g^1_{r_n}(\nu_j, X_j, X_i)}{2},\\
&\varphi_{r_n}(\vartheta_i) = E\left[g_{r_n}(\vartheta_i, \vartheta_j)\middle|\vartheta_i\right],\\
&\tilde{g}_{r_n}(\vartheta_i, \vartheta_j) = g_{r_n}(\vartheta_i, \vartheta_j) - \varphi_{r_n}(\vartheta_i) - \varphi_{r_n}(\vartheta_j).
\end{aligned}$$

Using these definitions, we can write

$$\sum_{j\neq i} g^1_{r_n}(\nu_i, X_i, X_j) = \sum_{j\neq i} \tilde{g}_{r_n}(\vartheta_i, \vartheta_j) + 2(n_1 - 1)\sum_{i=1}^{n_1} \varphi_{r_n}(\vartheta_i). \tag{37}$$

Consider $U_n \equiv \frac{1}{2n_1(n_1-1)}\sum_{j\neq i} h^{-\frac{1}{2}} n_1^{\frac{3}{2}} \tilde{g}_{r_n}(\vartheta_i, \vartheta_j)$. Since $E[\tilde{g}_{r_n}(\vartheta_i, \vartheta_j)|\vartheta_i] = 0$ and $E[\varphi_{r_n}(\vartheta_i)] = 0$, U_n is a degenerate U-statistic whose kernel depends on n. The kernel further satisfies

$$\begin{aligned}
E\left[\left(h^{-\frac{1}{2}} n_1^{\frac{3}{2}} \tilde{g}_{r_n}(\vartheta_i, \vartheta_j)\right)^2\right] &= \frac{n_1^3}{h} E\left[\left(\frac{1}{2}g^1_{r_n}(\nu_i, X_i, X_j) - \varphi_{r_n}(\vartheta_i) + \frac{1}{2}g^1_{r_n}(\nu_j, X_j, X_i) - \varphi_{r_n}(\vartheta_j)\right)^2\right]\\
&\leq \frac{n_1^3}{h} E\left[\left(\frac{1}{2}g^1_{r_n}(\nu_i, X_i, X_j) + \frac{1}{2}g^1_{r_n}(\nu_j, X_j, X_i))\right)^2\right]\\
&\leq \frac{n_1^3}{h} 2\left(E\left[\left(\frac{1}{2}g^1_{r_n}(\nu_i, X_i, X_j)\right)^2\right] + E\left[\left(\frac{1}{2}g^1_{r_n}(\nu_j, X_j, X_i)\right)^2\right]\right)\\
&= \frac{n_1^3}{h} E[(g^1_{r_n}(\nu_i, X_i, X_j))^2]\\
&\leq \frac{C}{h^2} = Cn_1\frac{1}{n_1 h^2} = O_p(n_1)\cdot o_p(1) = o_p(n_1).
\end{aligned}$$

The first inequality above holds since $E[(Q_1 - E[Q_1|Q_2])^2] \leq E[Q_1^2]$ for any two random variables Q_1 and Q_2 with finite second moments. The second inequality follows from the fact that $(a+b)^2 \leq 2(a^2+b^2)$ for any two real numbers a and b. The equality on the fourth line holds since $E[(g^1_{r_n}(\nu_i, X_i, X_j))^2] = E[(g^1_{r_n}(\nu_j, X_j, X_i))^2]$. Finally, the inequality on the last line holds since $E[(g^1_{r_n}(\nu_i, X_i, X_j))^2] \leq Cn_1^{-3}h^{-1}$.

Thus, we can apply Lemma 3.1 of Powell et al. (1989) to U_n, which yields $\sqrt{n_1}U_n = o_p(1)$. This is enough to show that

$$\sqrt{n_1}\left(\frac{1}{n_1^2}\sum_{j\neq i} n_1^{\frac{3}{2}}\tilde{g}_{r_n}(\vartheta_i,\vartheta_j)\right) = \sum_{j\neq i}\tilde{g}_{r_n}(\vartheta_i,\vartheta_j) = o_p(\sqrt{h}) = o_p(n^{-1/5}).$$

It is possible to show, as above, that the process $\sum_{i\neq j}^{n_1}\tilde{g}_{r_n}(\vartheta_i,\vartheta_j)$ is also stochastically equicontinuous in r_n in a neighborhood around r^1_{n0}. Then (34) implies, at least,

$$\sum_{i\neq j}^{n_1}\tilde{g}_{\hat{r}^1_n}(\vartheta_i,\vartheta_j) - \sum_{i\neq j}^{n_1}\tilde{g}_{r^1_{n0}}(\vartheta_i,\vartheta_j) = o_p(n^{-1/5}). \tag{38}$$

Turning to the second term in (37),

$$\begin{aligned}
2\varphi_{r_n}(\vartheta_i) &= E\left[g^1_{r_n}(\nu_i,X_i,X_j)\Big|\nu_i,X_i\right] \\
&= n^{\frac{-3}{2}}\nu_i\int r_n(u)\frac{1}{h}K\Big(\frac{X_i-u}{h}\Big)f_{X|Z=1}(u)du \\
&= n^{\frac{-3}{2}}\nu_i\int r_n(X_i+uh)K(u)f_{X|Z=1}(X_i+uh)du \\
&\equiv n^{\frac{-3}{2}}\nu_i I_{r_n}(X_i).
\end{aligned} \tag{39}$$

Therefore,

$$2(n_1-1)\sum_{i=1}^{n_1}\varphi_{r_n}(\vartheta_i) = \frac{n_1-1}{n_1}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\nu_i I_{r_n}(X_i) = O_p(1), \tag{40}$$

since $(n_1-1)/n_1 = O_p(1)$, $E[\varphi_{r^1_n}(\vartheta_i)] = 0$ and $E[\varphi_{r^1_n}(\vartheta_i)^2] < \infty$. Again, it is possible to show that the process $2(n_1-1)\sum_{i=1}^{n_1}\varphi_{r_n}(\vartheta_i)$ is stochastically equicontinuous in r_n around a neighborhood of the zero function. Further, it follows from (39) that $a\cdot\varphi_{r_n}(\vartheta_i) + b\cdot\varphi_{r'_n}(\vartheta_i) = \varphi_{ar_n+br'_n}(\vartheta_i)$ for all real numbers a,b and functions r_n, r'_n. Then (40) and the fact that $\sup_{x\in\mathcal{X}}|\hat{r}^1_n(x)/\sqrt{h} - r^1_{n0}(x)/\sqrt{h}| = o_p(1)$ implies

$$\begin{aligned}
&\frac{1}{\sqrt{h}}2(n_1-1)\sum_{i=1}^{n_1}\varphi_{\hat{r}^1_n}(\vartheta_i) - \frac{1}{\sqrt{h}}2(n_1-1)\sum_{i=1}^{n_1}\varphi_{r^1_{n0}}(\vartheta_i) \\
=&2(n_1-1)\sum_{i=1}^{n_1}\varphi_{\hat{r}^1_n/\sqrt{h}-r^1_{n0}/\sqrt{h}}(\vartheta_i) = o_p(1).
\end{aligned}$$

As a result,

$$2(n_1-1)\sum_{i=1}^{n_1}\varphi_{\hat{r}^1_n}(\vartheta_i) - 2(n_1-1)\sum_{i=1}^{n_1}\varphi_{r^1_{n0}}(\vartheta_i) = o_p(\sqrt{h}) = o_p(n^{\frac{-1}{5}}) \tag{41}$$

Combining (36), (38) and (41) shows that (32) is $o_p(n^{-1/10})$ [in fact, $o_p(n^{-1/5})$]. Similar arguments apply to (33), completing the proof. ∎

Lemma 7 *Under the assumptions of Theorem 3,*

$$B(x) = b(x) + R_2(x) \tag{42}$$

with $b(x) = O_p(h^2)$ *and* $n_1^{-1/2}\sum_{j=1}^{n_1}R_2(X_j) = o_p(n^{-1/10})$.

Proof of Lemma 7: We rewrite $B(x)$ as

$$\begin{aligned}
B(x) =&(1,0)\widehat{M}_{1n}^{-1}(x)\begin{pmatrix} E\Big[m_1(X_j)K\Big(\frac{X_j-x}{h}\Big)\Big] \\ E\Big[m_1(X_j)K\Big(\frac{X_j-x}{h}\Big)\Big(\frac{X_j-x}{h}\Big)\Big]\end{pmatrix} - m_1(x) \\
&+(1,0)\widehat{M}_{1n}^{-1}(x)\begin{pmatrix} \frac{1}{n_1 h}\sum_{j=1}^{n_1} m_1(X_j)K\Big(\frac{X_j-x}{h}\Big) - E\Big[m_1(X_j)K\Big(\frac{X_j-x}{h}\Big)\Big] \\ \frac{1}{n_1 h}\sum_{j=1}^{n_1} m_1(X_j)K\Big(\frac{X_j-x}{h}\Big)\Big(\frac{X_j-x}{h}\Big) - E\Big[m_1(X_j)K\Big(\frac{X_j-x}{h}\Big)\Big(\frac{X_j-x}{h}\Big)\Big]\end{pmatrix} \\
\equiv& b(x) + R_2(x),
\end{aligned}$$

where $b(x)$ and $R_2(x)$ are defined line by line. First, using an argument similar to the proof of Lemma 7 of Heckman et al. (1998), $b(x) = O(h^2)$. Second, the assertion involving the remainder $R_2(x)$ can be proven in a way similar to Lemma 6. ∎

Lemma 8 *Under the assumptions of Theorem 3,*

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i:Z_i=1}\hat{m}_1(X_i) - m_1(X_i) &= \frac{1}{\sqrt{n}n_1}\sum_{i:Z_i=1}\sum_{j:Z_j=1}\xi_1^m(Y_j,X_j,X_i) + o_p(n^{\frac{-1}{10}}) \\
\frac{1}{\sqrt{n}}\sum_{i:Z_i=0}\hat{m}_1(X_i) - m_1(X_i) &= \frac{1}{\sqrt{n}n_1}\sum_{i:Z_i=0}\sum_{j:Z_j=1}\xi_1^m(Y_j,X_j,X_i) + o_p(n^{\frac{-1}{10}}).
\end{aligned}$$

Proof of Lemma 8: Recall that $\hat{m}_1(X_i) - m_1(X_i) = A(X_i) + B(X_i)$. To show the first equation, apply Lemmas 6, 7 and the fact that $n^{-1/2}\sum_{i:Z_i=1} b(X_i) = O_p(\sqrt{n}h^2) = o_p(n^{-1/10})$ and $n_1/n = P(Z=1) + O_p(n^{-1/2})$. To show the second equation, one needs the additional facts that $\frac{1}{\sqrt{n}}\sum_{j=n_1+1}^{n} R_1(X_j) = o_p(n^{-\frac{1}{10}})$ and $\frac{1}{\sqrt{n}}\sum_{j=n_1+1}^{n} R_2(X_j) = o_p(n^{-\frac{1}{10}})$. These are established in a way similar to Lemmas 6 and 7. ∎

Lemma 9 *Under the assumptions of Theorem 3,*

$$\frac{1}{\sqrt{n}n_1}\sum_{i:Z_i=1}\sum_{j:Z_j=1}\xi_1^m(Y_j,X_j,X_i) = \frac{1}{\sqrt{n}}\sum_{j:Z_j=1} E[\xi_1^m(Y_j,X_j,X)|Y_j,X_j] + o_p(n^{-\frac{1}{10}}),$$

where the expectation is taken w.r.t. the distribution $f_{X|Z=1}(x)$. Furthermore,

$$\frac{1}{\sqrt{n}(n-n_1)}\sum_{i:Z_i=0}\sum_{j:Z_j=1}\xi_1^m(Y_j,X_j,X_i) = \frac{1}{\sqrt{n}}\sum_{j:Z_j=1} E[\xi_1^m(Y_j,X_j,X)|Y_j,X_j] + o_p(n^{-\frac{1}{10}}),$$

where the expectation is taken w.r.t. the distribution $f_{X|Z=0}(x)$.

Proof of Lemma 9: Note that

$$\begin{aligned}
&n_1^{-\frac{3}{2}}\sum_{i:Z_i=1}\sum_{j:Z_j=1}\xi_1^m(Y_j,X_j,X_i) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} g_{r_{n0}^1}^1(Y_j,X_j,X_i) + g_{r_{n0}^2}^2(Y_j,X_j,X_i), \text{ and} \\
&n_1^{-\frac{1}{2}}\sum_{j:Z_j=1} E[\xi_1^m(Y_j,X_j,X)|Y_j,X_j] = \frac{n_1}{n_1-1}2(n_1-1)\sum_{i=1}^{n_1}\varphi_{r_n^1}(\vartheta_i) + \varphi_{r_n^2}(\vartheta_i) \\
&= 2(n_1-1)\sum_{i=1}^{n_1}\big(\varphi_{r_n^1}(\vartheta_i) + \varphi_{r_n^2}(\vartheta_i)\big) + O_p(n_1^{-1}).
\end{aligned}$$

Hence, using the arguments in the proof of Lemma 6 and the fact that $n_1/n = P(Z=1) + O_p(n^{-1/2})$, we can show

$$\frac{1}{\sqrt{n}n_1}\sum_{i:Z_i=1}\sum_{j:Z_j=1}\xi_1^m(Y_j,X_j,X_i) - \frac{1}{\sqrt{n}}\sum_{j:Z_j=1}E[\xi_1^m(Y_j,X_j,X)|Y_j,X_j] = o_p(n^{-\frac{1}{10}}),$$

which is the first assertion in Lemma 9. The proof of the second assertion is similar. ∎

Lemma 10 *Under the assumptions of Theorem 3, and for $i \neq j$,*

$$E[\xi_1^m(Y_j,X_j,X_i)|Y_j,X_j,Z_j=1,Z_i=1] = (Y_j - m_1(X_j))(1+O_p(h))$$
$$E[\xi_1^m(Y_j,X_j,X_i)|Y_j,X_j,Z_j=1,Z_i=0] = (Y_j - m_1(X_j))\frac{f_{X|Z=0}(X_j)}{f_{X|Z=1}(X_j)}(1+O_p(h)).$$

Proof of Lemma 10: It is straightforward to show that

$$\kappa_0(x) = \int \frac{1}{h}K\Big(\frac{x_j-x}{h}\Big)f_{X|Z=1}(x_j)dx_j = f_{X|Z=1}(x)k_0 + O_p(h^2),$$
$$\kappa_1(x) = \int \frac{1}{h}\Big(\frac{x_j-x}{h}\Big)K\Big(\frac{x_j-x}{h}\Big)f_{X|Z=1}(x_j)dx_j = O_p(h)$$
$$\kappa_2(x) = \int \frac{1}{h}\Big(\frac{x_j-x}{h}\Big)^2K\Big(\frac{x_j-x}{h}\Big)f_{X|Z=1}(x_j)dx_j = f_{X|Z=1}(x)k_2 + O_p(h^2),$$

where $k_\ell = \int u^\ell K(u)du$. Hence,

$$M_{1n}^{-1}(x) = \frac{1}{f_{X|Z=1}(x)}\begin{pmatrix} k_0^{-1}+O(h) & O(h) \\ O(h) & k_2^{-1}+O(h)\end{pmatrix}.$$

Therefore,

$$\begin{aligned}
&E[\xi_1^m(Y_j,X_j,X_i)|Y_j,X_j,Z_j=1,Z_i=1]\\
=&(Y_j-m_1(X_j))\int\frac{1}{h}\frac{1}{f_{X|Z=1}(x_i)}\Big(\frac{1}{k_0}+O(h)+O(h)\Big(\frac{X_j-x_i}{h}\Big)\Big)K\Big(\frac{X_j-x_i}{h}\Big)f_{X|Z=1}(x_i)dx_i\\
=&(Y_j-m_1(X_j))(1+O(h)).
\end{aligned}$$

This establishes the first assertion in Lemma 10. The proof of the second assertion is similar. ∎

Lemma 11 *Under the assumptions of Theorem 3,*

$$\frac{1}{\sqrt{n}}\sum_{j:Z_j=1}E[\xi_1^m(Y_j,X_j,X)|Y_j,X_j] = \frac{1}{\sqrt{n}}\sum_{j:Z_j=1}(Y_j-m_1(X_j)) + o_p(n^{-\frac{1}{10}})$$
$$\frac{1}{\sqrt{n}}\sum_{j:Z_j=0}E[\xi_1^m(Y_j,X_j,X)|Y_j,X_j] = \frac{1}{\sqrt{n}}\sum_{j:Z_j=0}(Y_j-m_1(X_j))\frac{f_{X|Z=0}(X_j)}{f_{X|Z=1}(X_j)} + o_p(n^{-\frac{1}{10}}).$$

Proof of Lemma 11: From Lemma 10, it follows that

$$\frac{1}{\sqrt{n}}\sum_{j:Z_j=1} E[\xi_1^m(Y_j,X_j,X)|Y_j,X_j] - \frac{1}{\sqrt{n}}\sum_{j:Z_j=1}(Y_j - m_1(X_j)) = \frac{1}{\sqrt{n}}\sum_{j:Z_j=1}(Y_j - m_1(X_j))b_j$$

where b_j's are $O_p(h)$ uniformly and b_j only depends on X_j. Hence, it is true that

$$\frac{1}{\sqrt{n}}\sum_{j:Z_j=1}(Y_j - m_1(X_j))b_j = O_p(h) = o_p(n^{-\frac{1}{10}}).$$

The proof of the second assertion is similar. ∎

Lemma 12 *Under the assumptions of Theorem 3,*

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i:Z_i=1}\hat{m}_1(X_i) - m_1(X_i) = \frac{1}{\sqrt{n}}\sum_{j:Z_j=1}(Y_j - m_1(X_j)) + o_p(n^{\frac{-1}{10}}),\\
&\frac{1}{\sqrt{n}}\sum_{i:Z_i=0}\hat{m}_1(X_i) - m_1(X_i) = \frac{1}{\sqrt{n}}\sum_{j:Z_j=1}(Y_j - m_1(X_j))\frac{1-q(X_j)}{q(X_j)} + o_p(n^{\frac{-1}{10}}),\\
&\frac{1}{\sqrt{n}}\sum_{i:Z_i=1}\hat{m}_0(X_i) - m_0(X_i) = \frac{1}{\sqrt{n}}\sum_{j:Z_j=0}(Y_j - m_0(X_j))\frac{q(X_j)}{1-q(X_j)} + o_p(n^{\frac{-1}{10}}),\\
&\frac{1}{\sqrt{n}}\sum_{i:Z_i=0}\hat{m}_0(X_i) - m_0(X_i) = \frac{1}{\sqrt{n}}\sum_{j:Z_j=0}(Y_j - m_0(X_j)) + o_p(n^{\frac{-1}{10}}).
\end{aligned}$$

Proof of Lemma 12: The first assertion follows directly from combining Lemmas 8, 9 and 11. To show the second equation, let $n_0 = n - n_1$ and again combine Lemmas 8, 9 and 11 to obtain

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i:Z_i=0}\hat{m}_1(X_i) - m_1(X_i)\\
=&\frac{1}{\sqrt{n}}\frac{n_0}{n_1}\sum_{j:Z_j=1}(Y_j - m_1(X_j))\frac{f_{X|Z=0}(X_j)}{f_{X|Z=1}(X_j)} + o_p(n^{\frac{-1}{10}})\\
=&\frac{1}{\sqrt{n}}\Big(\frac{P(Z=0)}{P(Z=1)} + O_p(n^{\frac{-1}{2}})\Big)\sum_{j:Z_j=1}(Y_j - m_1(X_j))\frac{P(Z=1)}{P(Z=0)}\frac{1-q(X_j)}{q(X_j)} + o_p(n^{\frac{-1}{10}})\\
=&\frac{1}{\sqrt{n}}\sum_{j:Z_j=1}(Y_j - m_1(X_j))\frac{1-q(X_j)}{q(X_j)} + o_p(n^{\frac{-1}{10}}).
\end{aligned}$$

The second equality holds by Bayes' Theorem and the fact that $n_0/n_1 = P(Z=0)/P(Z=1) + O_p(n^{-1/2})$. The third equality is implied by $\frac{1}{\sqrt{n}}\sum_{j:Z_j=1}(Y_j - m_1(X_j))\frac{1-q(X_j)}{q(X_j)} = O_p(1)$. The last two assertions in Lemma 12 follow from the first two by switching between $Z=0$ and $Z=1$ throughout. ∎

Now we are in a position to provide a simple proof to Lemma 5.

Proof of Lemma 5: We have

$$\begin{aligned}
&\sqrt{n}(\widehat{\Delta}_m - \Delta) \\
=&\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_1(X_i) - \hat{m}_0(X_i) - \Delta\right)\right] \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\hat{m}_1(X_i) - m_1(X_i)\right) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\hat{m}_0(X_i) - m_0(X_i)\right) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(m_1(X_i) - m_0(X_i) - \Delta\right) \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i + o_p(n^{\frac{-1}{10}}).
\end{aligned}$$

The last line holds by splitting the first two sums according to $Z_i = 0$ versus $Z_i = 1$ and applying Lemma 12 to each piece. The proof for the estimator $\widehat{\Gamma}_m$ is similar and we omit it. ∎

D. Efficiency arguments

The proof of Theorem 5 First, we provide some simple facts. Under the conditions of Theorem 5, including one-sided non-compliance, the following expressions hold true:

$$\begin{aligned}
p =&P(D=1) = P(Z=1, D(1)=1) = P(D(1)=1|Z=1)P(Z=1) = \Gamma_t Q, \\
p(x) =&P(D=1|X=x) = P(D(1)=1|Z=1, X=x)P(Z=1|X=x) = \mu_1(x)q(x), \\
\mu_0(x) =&E[D|Z=0, X=x] = E[D(0)|X=x] = 0.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
m_1(x) =&E[Y|Z=1, X=x] \\
=&E[Y|Z=1, D=1, X]P[D=1|Z=1, X=x] \\
&+ E[Y|Z=1, D=0, X]P[D=0|Z=1, X=x] \\
=&E[Y(1)|Z=1, D=1, X=x]P[D(1)=1|Z=1, X=x] \\
&+ E[Y(0)|Z=1, D=0, X=x]P[D(1)=0|Z=1, X=x] \\
=&E[Y(1)|X=x]\mu_1(x) + E[Y(0)|X=x](1-\mu_1(x)) \\
=&\rho_1(x) - (1-\mu_1(x))(\rho_1(x) - \rho_0(x))
\end{aligned}$$

and

$$\begin{aligned}
m_0(x) =&E[Y|Z=0, X=x] \\
=&E[Y|Z=0, D=0, X=x] \\
=&E[Y(0)|Z=0, D=0, X=x] = E[Y(0)|X=x] = \rho_0(x).
\end{aligned}$$

Note that the fourth equality involving the $m_1(x)$ term holds because the IV assumption and unconfoundedness assumption hold jointly as in (20). In general, this equality will not hold when Assumption 1 and

Assumption 2 hold, but do not hold jointly. This is the reason we require a stronger condition in Theorem 5. The second equality regarding the $m_0(x)$ term holds since $D = 0$ when $Z = 0$. As a result, adding $D = 0$ does not change the conditional expectation.

We define

$$
\begin{aligned}
&\phi_t(Y, D, X) \\
=&\frac{p(X)}{p}\left\{\frac{D(Y-\rho_1(X))}{p(X)} - \frac{(1-D)(Y-\rho_0(X))}{(1-p(X))} + \frac{D(\rho_1(X)-\rho_0(X)-\beta_t)}{p(X)}\right\} \\
\equiv&\frac{p(X)}{p}\left\{\phi_1 - \phi_2 + \phi_3\right\},
\end{aligned}
$$

and rewrite $\psi_t(Y, D, Z, X)$ as

$$
\begin{aligned}
&\psi_t(Y, D, Z, X) \\
=&\frac{q(X)}{Q\Gamma_t}\left\{\frac{Z\big[Y - m_1(X) - \tau_t(D-\mu_1(X))\big]}{q(X)} - \frac{(1-Z)\big[Y - m_0(X) - \tau_t(D-\mu_0(X))\big]}{1-q(X)}\right. \\
&\qquad\left. + \frac{Z\big[m_1(X) - m_0(X) - \tau_t(\mu_1(X)-\mu_0(X))\big]}{q(X)}\right\}, \\
=&\frac{p(X)}{p}\frac{1}{\mu_1(X)}\left\{\frac{Z(Y-\rho_1(X))}{q(X)} + \frac{Z(Y-\mu_1(X))(\rho_1(X)-\rho_0(X))}{q(X)} - \frac{Z\gamma_t(D-\mu_1(X))}{q(X)}\right. \\
&\qquad\left. - \frac{(1-X)(Y-\rho_0(X))}{1-q(X)} + \frac{Z(\rho_1(X)-\rho_0(X)-\beta_t)\mu_1(X)}{q(X)}\right\} \\
=&\frac{p(X)}{p}\left\{\frac{Z(Y-\rho_1(X))}{p(X)} + \frac{Z(Y-\mu_1(X))(\rho_1(X)-\rho_0(X))}{p(X)} - \frac{Z\beta_t(D-\mu_1(X))}{p(X)}\right. \\
&\qquad\left. - \frac{(1-Z)(Y-\rho_0(X))}{(1-q(X))\mu_1(X)} + \frac{X(\rho_1(X)-\rho_0(X)-\beta_t)\mu_1(X)}{p(X)}\right\} \\
=&\frac{p(X)}{p}\left\{\frac{Z(Y-\mu_1(X))}{p(X)} + \frac{Z(Y-\mu_1(X))(\rho_1(X)-\rho_0(X))}{p(X)} - \frac{Z\beta_t((D-1)+(1-\mu_1(X)))}{p(X)}\right. \\
&\qquad\left. - \frac{(1-Z)(Y-\rho_0(X))}{(1-q(X))\mu_1(X)} + \frac{D(\rho_1(X)-\rho_0(X)-\beta_t)\mu_1(X)}{p(X)}\right\} \\
=&\frac{p(X)}{p}\left\{\frac{Z(Y-\mu_1(X))}{p(X)} - \frac{(1-Z)(Y-\rho_0(X))}{(1-q(X))\mu_1(X)} + \frac{X(\rho_1(X)-\rho_0(X)-\beta_t)}{p(X)} - \frac{Z\beta_t(D-1))}{p(X)}\right\} \\
\equiv&\frac{p(X)}{p}\left\{\psi_1 - \psi_2 + \psi_3 - \psi_4\right\}.
\end{aligned}
$$

Note that

$$
\begin{aligned}
E[\phi_1\psi_1|X] &= \frac{E\left[ZD(Y-\rho_1(X))^2\big|X\right]}{p^2(X)} = \frac{E\left[D(Y-\rho_1(X))^2\big|X\right]}{p^2(X)} \\
&= \frac{E\left[D(Y-\rho_1(X))^2\big|X, D=1\right]p(D=1|X=x)}{p^2(X)} = \frac{\sigma_1^2(X)}{p(X)},
\end{aligned}
$$

where $\sigma_1^2(X) = V(Y(1)|X)$. Also, $E[\phi_1\psi_2] = 0$ since $(1-Z)D = 0$ with probability one and $E[\phi_1\psi_4|X] = 0$ since $D(1-D) = 0$. Note that

$$\begin{aligned}
E[\phi_1\psi_3|X] &= \frac{\rho_1(X) - \rho_0(X) - \beta_t}{p^2(X)} E[DZ(Y - \rho_1(X))|X] \\
&= \frac{\rho_1(X) - \rho_0(X) - \beta_t}{p^2(X)} E[D(Y(1) - \rho_1(X))|X] = 0,
\end{aligned}$$

where the first equality in second line holds since $ZD = 1$ with probability one and the second equality holds since $E[D(Y(1) - \rho_1(X))|X] = 0$. Furthermore,

$$\begin{aligned}
E[\phi_1\psi_1|X] &= \frac{E[Z(1-D)(Y - \rho_1(X))(Y - \rho_0(X))|X]}{p(X)(1-p(X))} \\
&= \frac{E[Z(1-D)(Y - \rho_1(X))(Y - \rho_0(X))|X, Z=1, D=0]P(Z=1, D=0|X)}{p(X)(1-p(X))} \\
&= \frac{E[(Y(0) - \rho_1(X))(Y(0) - \rho_0(X))|X, Z=1, D=0](1-\mu_1(X))q(X)}{p(X)(1-p(X))} \\
&= \frac{\sigma_0^2(X)(1-\mu_1(X))q(X)}{p(X)(1-p(X))} = \frac{\sigma_0^2(X)}{1-p(X)} \frac{1-\mu_1(X)}{\mu_1(X)}, \\
E[\phi_1\psi_2|X] &= \frac{E[(1-Z)(1-D)(Y - \rho_0(X))^2|X]}{(1-p(X))(1-q(X))\mu_1(X)} \\
&= \frac{E[(1-Z)(1-D)(Y - \rho_0(X))^2|X, Z=0, D=0]P(Z=0, D=0|X)}{(1-p(X))(1-q(X))\mu_1(X)} \\
&= \frac{E[(Y(0) - \rho_0(X))^2|X, D=0]P(Z=0|X)}{(1-p(X))(1-q(X))\mu_1(X)} \\
&= \frac{\sigma_0^2(X)(1-q(X))}{(1-p(X))(1-q(X))\mu_1(X)} = \frac{\sigma_0^2(X)}{1-p(X)} \frac{1}{\mu_1(X)}.
\end{aligned}$$

Also, we have $E[\phi_2\psi_3] = 0$, $E[\phi_2\psi_4] = 0$, $E[\phi_3\psi_1] = 0$, $E[\phi_3\psi_2] = 0$ and $E[\phi_3\psi_4] = 0$. Finally,

$$E[\phi_3\psi_3] = \frac{E[D(\rho_1(X) - \rho_0(X) - \beta_t)^2|X]}{p^2(X)} = \frac{(\rho_1(X) - \rho_0(X) - \beta_t)^2}{p(X)}.$$

Consequently,

$$\begin{aligned}
Cov(\phi_t, \psi_t) =& E[\phi_t\psi_t] \\
=& E\left[\frac{p^2(X)}{p}\left\{\frac{\sigma_1^2(X)}{p(X)} - \frac{\sigma_0^2(X)}{1-p(X)}\frac{1-\mu_1(X)}{\mu_1(X)}\right.\right. \\
&\left.\left. + \frac{\sigma_0^2(X)}{1-p(X)}\frac{1}{\mu_1(X)} + \frac{(\rho_1(X) - \rho_0(X) - \beta_t)^2}{p(X)}\right\}\right] \\
=& E[\phi_t^2] = Var(\phi_t).
\end{aligned}$$

This shows Theorem 5. ∎

References

Abadie, A. (2003). Semiparametric Instrumental Variable Estimation of Treatment Response Models. *Journal of Econometrics 113*, 231–263.

Abadie, A., J. Angrist, and G. Imbens (2002). Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. *Econometrica 70*, 91–117.

Deaton, A. (2009). Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Working paper. Center for Health and Wellbeing, Princeton University.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall.

Frölich, M. (2007). Nonparametric IV Estimation of Local Average Treatment Effects with Covariates. *Journal of Econometrics 139*, 35–75.

Frölich, M. and M. Lechner (2006). Exploiting Regional Treatment Intensity for the Evaluation of Labour Market Policies. IZA Discussion Paper No. 2144.

Frölich, M. and B. Melly (2008a). Identification of Treatment Effects on the Treated with One-Sided Non-Compliance. IZA Discussion Paper No. 3671.

Frölich, M. and B. Melly (2008b). Unconditional Quantile Treatment Effects under Endogeneity. IZA Discussion Paper No. 3288.

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

Heckman, J. (1997). Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources 32*, 441–462.

Heckman, J., H. Ichimura, and P. Todd (1998). Matching as an Econometric Evaluation Estimator. *The Review of Economic Studies 65*, 261–294.

Heckman, J. and S. Urzúa (2009). Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. IZA Discussion Paper, 3980.

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

Hong, H. and D. Nekipelov (2008). Semiparametric Efficiency in Nonlinear LATE Models. Working paper. Department of Economics, University of California, Berkeley.

Ichimura, H. and O. Linton (2002). Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators. Working paper. Department of Economics, University College, London.

Imbens, G. (2009). Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzúa (2009). Working paper. Department of Economics, Harvard University.

Imbens, G. and J. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica 62*, 467–475.

Kalyanaraman, K. (2009). Bandwidth choice for regression functionals with application to average treatment effects. Working paper. Department of Economics, Harvard University.

Masry, E. (1996). Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates. *Journal of Time Series Analysis 17*, 571–599.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric Estimation of Index Coefficients. *Econometrica 57*, 1403–1430.

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist. 6*, 34–58.