Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT*

Stephen G. Donald†  Yu-Chin Hsu‡  Robert P. Lieli§

September 9, 2011

Abstract

We propose inverse probability weighted estimators for 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. 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, instrumental variables, unconfoundedness, inverse probability weighted estimation, nonparametric estimation

*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.
†Department of Economics, University of Texas, Austin.
‡Department of Economics, University of Missouri, Columbia.
§Department of Economics, Central European University, Budapest and the National Bank of Hungary.
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.\footnote{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).}

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\footnote{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.} 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 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 potentially allows for 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 twofold. First, 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 (admittedly small) 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).

More 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 unconfoundedness 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 apply our test to data set on training programs administered under the Job Training Partnership Act (JTPA).\textsuperscript{3} Of interest is the effect

\textsuperscript{3}The same data set is analyzed by Abadie et al. (2002) and Frölich and Melly (2008b) among others.
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 (first order) asymptotic properties. The test for the unconfoundedness assumption is presented in Section 4, and the implications of unconfoundedness are
discussed in more detail. The empirical application is given in Section 5. Section 6 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 $D \equiv D(Z) = D(1)Z + D(0)(1 - Z)$.
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 $V = (Y(0,0), Y(0,1), Y(1,0), Y(1,1), D(1), D(0))^\prime$. 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): $P[D(1) \geq D(0) \mid X_1] = 1$. 

4
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 $P(Z = 1 | X_1)$ 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

$$\tau \equiv E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0]$$

$$\tau_t \equiv E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0, D = 1].$$

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,

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

$$\tau_t = \frac{E[W(1) - W(0) | Z = 1]}{E[D(1) - D(0) | Z = 1]}.$$ (2)

The quantities on the rhs of (1) and (2) are nonparametrically identified from the joint distribution of the observables $(Y, D, Z, X_1)$. Under Assumption 1, the following identification results are implied, for example, by Theorem 3.1 in Abadie (2003):

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

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

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

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

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

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. $E[Y(d) | D, X_2] = E[Y(d) | X_2], d \in \{0, 1\}$.

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 $\text{ATE} = E[Y(1) - Y(0)]$ and $\text{ATT} = E[Y(1) - Y(0) | D = 1].$ 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): $P[D(0) = 0] = 1.$

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

\[
\text{LATT} = E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0, D = 1]
\]

\[= E[Y(1) - Y(0) \mid D(1) = 1, D = 1]
\]

\[= E[Y(1) - Y(0) \mid D = 1] = \text{ATT}.
\]

Thus, under one-sided non-compliance, \( \text{ATT} = \text{LATT} \). 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 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 \( \tau \) and \( \tau_t \) are based on sample analog expressions for (3) through (6):

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

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

where \( \hat{q}(\cdot) \) is a suitable nonparametric estimator of the the propensity score function \( q(x_1) = E(Z \mid X_1 = x_1) \). Following Ichimura and Linton (2005), 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 suitable estimators of \( q(\cdot) \) were used instead.\(^5\)

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

\[
\min_{\beta_0, \ldots, \beta_p} \sum_{i=1}^{n} K\left(\frac{X_{1i} - x_1}{h}\right) \left[Z_i - \hat{\beta}_0 - \hat{\beta}_1(X_{1i} - x_1) - \ldots - \hat{\beta}_p(X_{1i} - x_1)^p\right]^2,
\]

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 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 \( i \)th one.

### 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

\(^5\)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. As shown by Tamer and Khan (2010), this condition is necessary for obtaining \( \sqrt{n} \)-asymptotics.
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:

$m_z(x_1) = E(Y | X_1 = x_1, Z = z)$, $\mu_z(x_1) = E(D | X_1 = x_1, Z = z)$, $s_z^2(x_1) = var(Y | X_1 = x_1, Z = z)$, $\sigma_z^2(x_1) = var(D | X_1 = x_1, Z = z)$, and $c_z(x_1) = cov(Y, D | X_1 = x_1, Z = z)$. 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 $nh_n^{2r}/(\log n)^2 \to \infty$.

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

**Theorem 1 (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
\( V_t = E[\psi_t^2(Y, D, Z, X_1)] \) with the functions \( \psi \) and \( \psi_t \) given by

\[
\psi(y, d, z, x_1) = \frac{1}{\Gamma} \left\{ \frac{z [y - m_1(x_1) - \tau(d - \mu_1(x_1))]}{q(x_1)} - \frac{(1 - z) [y - m_0(x_1) - \tau(d - \mu_0(x_1))]}{1 - q(x_1)} + m_1(x_1) - m_0(x_1) - \tau \mu_1(x_1) - \mu_0(x_1) \right\},
\]

and

\[
\psi_t(y, d, z, x_1) = \frac{q(x_1)}{Q \Gamma_t} \left\{ \frac{z [y - m_1(x_1) - \tau_t(d - \mu_1(x_1))]}{q(x_1)} - \frac{(1 - z) [y - m_0(x_1) - \tau_t(d - \mu_0(x_1))]}{1 - q(x_1)} + \frac{z [m_1(x_1) - m_0(x_1) - \tau_t \mu_1(x_1) - \mu_0(x_1)]}{q(x_1)} \right\},
\]

with \( Q = E(Z) \).

(b) \( V \) is equal to the semiparametric efficiency bound for LATE without the knowledge of \( q(x_1) \)

(c) \( V_t \) is equal to the semiparametric efficiency bound for LATT without the knowledge of \( q(x_1) \).

Theorem 1 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). Theorem 1 follows from the fact that, under the conditions stated, \( \hat{\tau} \) and \( \hat{\tau}_t \) can be expressed as asymptotically linear with influence functions \( \psi \) and \( \psi_t \), 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 1 are derived in Frölich (2007) and Hong and Nekipelov (2008), respectively. To use Theorem 1 for statistical inference, one needs consistent estimators for \( V \) and \( V_t \). Such estimators can be obtained by constructing (uniformly) consistent estimates for \( \psi \) and \( \psi_t \) and then averaging the squared estimates over the sample observations \( \{(Y_i, D_i, Z_i, X_{1i})\}_{i=1}^{n} \).
In estimating $\psi$ and $\psi_t$, one replaces the functions $m_z(x_1)$, $\mu_z(x_1)$ and $q(x_1)$ with nonparametric estimators that are uniformly consistent over $X$. The quantities $\Gamma$, $\Gamma_t$ and $Q$ are also replaced with sample analogs.

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:

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

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

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

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

**Theorem 2** 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

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

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

11
The additional regularity conditions referred to in Theorem 2 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

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

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 2 follows directly from this result. Let $\hat{\psi}_t(\cdot)$ and $\hat{\phi}_t(\cdot)$ be (uniformly) consistent estimators of $\psi_t$ and $\phi_t$ obtained, e.g., by the sample analog principle (see the discussion after Theorem 1). A consistent estimator for $\sigma^2$ can then be constructed as

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

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 \( \bar{a} = \frac{\text{var}(A_0) - \text{cov}(A_0, A_1)}{\text{var}(A_1) - \text{var}(A_0)} \). Then:

(a) \( \text{var}(A_a) \leq \text{var}(A_0) \) for all \( a \in \mathbb{R} \).

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

(c) \( \text{var}(A_a) < \text{var}(A_1) \) 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 \( \tau_t \) 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
\[ \hat{a} = \frac{\sum_{i=1}^{n} \hat{\phi}_t(Y_i, D_i, Z_i, X_{1i})(\hat{\phi}_t(Y_i, D_i, Z_i, X_{1i}) - \hat{\psi}_t(Y_i, D_i, X_{2i}))}{\sum_{i=1}^{n} (\hat{\phi}_t(Y_i, D_i, Z_i, X_{1i}) - \hat{\psi}_t(Y_i, D_i, X_{2i}))^2}. \]

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 \( V \text{ar}(\phi_t) = \text{Cov}(\phi_t, \psi_t) \), 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 3** Suppose that Assumption 1 parts (i), (iii), (iv), (v) and Assumption 3 are satisfied, and let \( V = (Y(0), Y(1)) \). If, in addition, \( X_1 = X_2 = X \),
\[ E(V \mid Z, D, X) = E(V \mid X) \quad \text{and} \quad E(VV' \mid Z, D, X) = E(VV' \mid X), \quad (10) \]
then \( \bar{a} = 0 \).

The proof of Theorem 3 is provided in Appendix C. The conditions of Theorem 3 are stronger than those of Theorem 2. 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 (10). If the null of unconfoundedness is accepted due to (10) 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 (10) to fail hence the conclusion is not automatic.

It must be pointed out that the theoretical results discussed in this subsection are qualified by the fact that one needs to pre-test for unconfoundedness. The construction of a more efficient ATT estimator takes the unconfoundedness assumption as given. However, when the proposed test does not reject unconfoundedness, there is some probability that a type 2 error has been committed. In this case \( \hat{\beta}_t \) is an asymptotically biased estimator of ATT (at least in general), and it is not clear whether taking a linear combination with \( \hat{\tau}_t \) will result in an improvement in MSE relative to \( \hat{\tau}_t \) itself. While a type 2 error happens with smaller probability when the violation of unconfoundedness (and the resulting bias of \( \hat{\beta}_t \)) is severe, pre-testing, in general, causes the MSE gain of \( \hat{\beta}_t(\bar{a}) \) to be less than the efficiency gain implied by Lemma 1 alone. Similarly, with a small (and controllable) probability, the test user will reject unconfoundedness when it actually holds. This again erodes the MSE gain from taking or not taking linear combinations based on the result of the test.
5 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. 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 1, 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 1. 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

---

6The data set consists of 11,204 applicants who were categorized as adult males or females; data on youth are not included.
Table 1: JPTE variables

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

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).

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.\(^7\) One can construct \( \hat{p}(X_i), \hat{m}_1(X_i), \) etc. analogously (these functions will not generally be trivial in X). The asymptotic theory presented in this paper, including the test for unconfoundedness, remains valid if these estimators are used in computing \( \hat{\tau}, \hat{\tau}_i \) and \( \hat{\beta}_i \). The drawback of this approach is that if \( r \)

\(^7\)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 \( X = X_1 = X_2 = \text{SEX} \). 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 2. 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 2 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 2: \( X = \text{SEX} \)

<table>
<thead>
<tr>
<th>Subpop.</th>
<th>Obs. (USD)</th>
<th>( \hat{\tau}_t ) (USD)</th>
<th>std(( \hat{\tau}_t )) (USD)</th>
<th>( \hat{\beta}_t ) (USD)</th>
<th>std(( \hat{\beta}_t )) (USD)</th>
<th>std(( \hat{\tau}_t - \hat{\beta}_t )) (USD)</th>
<th>Test stat. (2-sided)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>11150</td>
<td>1828.1 (506.9)</td>
<td>2979.1 (313.1)</td>
<td>(407.4)</td>
<td>-2.825</td>
<td>0.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>5083</td>
<td>1716.0 (916.4)</td>
<td>4035.6 (557.3)</td>
<td>(740.7)</td>
<td>-3.132</td>
<td>0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>6067</td>
<td>1916.4 (547.8)</td>
<td>2146.7 (346.4)</td>
<td>(436.1)</td>
<td>-0.528</td>
<td>0.597</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: \( \hat{\tau}_t \) is the inverse probability weighted IV estimator of \( \text{LATT} = \text{ATT} \). \( \hat{\beta}_t \) is an estimator of \( \text{ATT} \) 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 \( X = X_1 = X_2 = (\text{SEX, BELOW30, MINORITY, HS}) \). We perform three types of tests: (i) an overall unconfoundedness test w.r.t. \( X \); (ii) conditioning on the possible values of \( X^s = \text{SEX} \), we perform two unconfoundedness tests w.r.t \( X \setminus X^s = (\text{BELOW30, MINORITY, HS}) \) 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 3.

Comparing the LATT estimates for males, females and the whole population across Tables 2 and 3, 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 3. 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 \( \text{BELOW30, MINORITY and HS} \), while female participation appears unconfounded with respect to this set of covariates as well.\(^8\) 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 rejection among

\(^8\)Even if the non-rejection for females in Table 2 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 3: \( X = (\text{SEX, BELOW30, MINORITY, HS}) \)

<table>
<thead>
<tr>
<th>Subpop.</th>
<th>Obs. ( (\text{USD}) )</th>
<th>( \hat{\tau}_t ) ( (\text{USD}) )</th>
<th>std(( \hat{\tau}_t )) ( (\text{USD}) )</th>
<th>( \hat{\beta}_t ) ( (\text{USD}) )</th>
<th>std(( \hat{\beta}_t )) ( (\text{USD}) )</th>
<th>std(( \hat{\tau}_t - \hat{\beta}_t )) ( (\text{USD}) )</th>
<th>Test stat. ( (\text{2-sided}) )</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X^* = \emptyset; X \setminus X^* = (\text{SEX, BELOW30, MINORITY, HS}) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>11150</td>
<td>1810.3</td>
<td>501.9</td>
<td>2804.2</td>
<td>312.6</td>
<td>404.1</td>
<td>-2.460</td>
<td>0.014</td>
</tr>
<tr>
<td>( X^* = \text{SEX}; X \setminus X^* = (\text{BELOW30, MINORITY, HS}) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>5083</td>
<td>1805.9</td>
<td>904.5</td>
<td>3936.0</td>
<td>554.8</td>
<td>732.1</td>
<td>-2.909</td>
<td>0.004</td>
</tr>
<tr>
<td>Females</td>
<td>6067</td>
<td>1813.7</td>
<td>545.3</td>
<td>1912.4</td>
<td>347.2</td>
<td>434.6</td>
<td>-0.227</td>
<td>0.820</td>
</tr>
</tbody>
</table>

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.

males may be partly due to the relatively small number of observations available in some of these groups. The general pattern shown in Table 3 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 3 can again be interpreted as an estimate of ATE/ATT.
(Compared with Table 2, 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 “one-shot” 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.

6 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.

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 is not sufficient to account for men’s participation decision. In contrast, we cannot reject the hypothesis that female participation is essentially random.
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

The derivation of equations (1) and (2)  We can write

\[
E[W(1) - W(0)] = E[(D(1) - D(0))[Y(1) - Y(0)]
\]

\[
= E[Y(1) - Y(0) | D(1) - D(0) = 1]P[D(1) - D(0) = 1]
\]

\[
= \tau \cdot E[D(1) - D(0)],
\]

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,

\[
E[W(1) - W(0) | Z = 1] = E[(D(1) - D(0))[Y(1) - Y(0) | Z = 1]]
\]

\[
= E[Y(1) - Y(0) | D(1) - D(0) = 1, Z = 1]P[D(1) - D(0) = 1 | Z = 1]
\]

\[
= \tau_t \cdot E[D(1) - D(0) | Z = 1],
\]

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

B. The proof of Theorem 1

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

\[
\hat{\Delta} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{Z_iY_i}{q(X_i)} - \frac{(1-Z_i)Y_i}{1-q(X_i)} \right\}, \quad \hat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{Z_iD_i}{q(X_i)} - \frac{(1-Z_i)D_i}{1-q(X_i)} \right\},
\]

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

Lemma 2 Under the conditions of Theorem 1, $\sqrt{n}(\hat{\Delta} - \Delta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \delta(Y_i, D_i, Z_i, X_i) + o_p(1)$ and $\sqrt{n}(\hat{\Gamma} - \Gamma) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \gamma(Y_i, D_i, Z_i, X_i) + o_p(1)$, 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
\[ f(\hat{\Delta}, \hat{\Gamma}) = \frac{\Delta}{\Gamma} \] around the point \((\Delta, \Gamma)\) to write
\[ \sqrt{n}(\hat{\tau} - \tau) = \sqrt{n} \left( \frac{\Delta}{\Gamma} - \frac{\hat{\Delta}}{\hat{\Gamma}} \right) = \frac{1}{\hat{\Gamma}} \sqrt{n}(\hat{\Delta} - \Delta) - \frac{\tau}{\hat{\Gamma}} \sqrt{n}(\hat{\Gamma} - \Gamma) + o_p(1). \] (12)

Substituting the influence function representations given in Lemma 2 into (12) 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 \(\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}(0, V)\).

To derive the asymptotic distribution of \(\hat{\tau}_t\), we write
\[ \tilde{\Delta}_t = \sum_{i=1}^{n} \left\{ \hat{q}(X_i) \left( \frac{Z_i Y_i}{\hat{q}(X_i)} - \frac{(1 - Z_i) Y_i}{1 - \hat{q}(X_i)} \right) \right\} / \sum_{i=1}^{n} \hat{q}(X_i), \]
\[ \tilde{\Gamma}_t = \sum_{i=1}^{n} \left\{ \hat{q}(X_i) \left( \frac{Z_i D_i}{\hat{q}(X_i)} - \frac{(1 - Z_i) D_i}{1 - \hat{q}(X_i)} \right) \right\} / \sum_{i=1}^{n} \hat{q}(X_i), \]
so that \(\hat{\tau}_t = \tilde{\Delta}_t / \tilde{\Gamma}_t\). Then:

**Lemma 3** Under the conditions of Theorem 1,
\[ \sqrt{n}(\tilde{\Delta}_t - \Delta_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} q(X_i) \left\{ \frac{Z_i(1 - m_1(X_i))}{q(X_i)} - \frac{(1 - Z_i)(1 - m_0(X_i))}{1 - q(X_i)} \right\} + o_p(1), \]
\[ \sqrt{n}(\tilde{\Gamma}_t - \Gamma_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} q(X_i) \left\{ \frac{Z_i(1 - m_1(X_i))}{q(X_i)} - \frac{(1 - Z_i)(1 - m_0(X_i))}{1 - q(X_i)} \right\} \]
\[ + \frac{(m_1(X_i) - m_0(X_i) - \Delta_t)Z_i}{q(X_i)} + o_p(1). \]

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

We complete the proof of Theorem 1 by verifying Lemma 2. The proof of Lemma 3 is omitted. \(\blacksquare\)

**The proof of Lemma 2** Our argument is based on Ichimura and Linton (2005) 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 \(x^\lambda = x_1^{\lambda_1} \cdots x_r^{\lambda_r} \in \mathbb{R}\). For a non-negative integer \(\ell\), 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 \(\ell\) 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
\[ \tilde{X}_t^{(i)} = [(X_t^1 - X_t^1)^{\Lambda(0)}, (X_t^1 - X_t^1)^{\Lambda(1)}, \ldots, (X_t^r - X_t^r)^{\Lambda(r)}]' \].
Then the leave-one-out local polynomial regression estimator of \( q(X_i) \) is the first component of the vector \( \hat{\beta} \) that solves \( \min_{\beta} \sum_{t:t\neq i} K \left( \frac{X_t - X_i}{h} \right) (Z_t - \hat{X}^{(i)}_t \beta)^2 \). Letting \( e_1 \) denote the first unit vector having the same dimension as \( X_t^{(i)} \), we can write this estimator as

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

where \( \omega_{it} \) depends only on \( X_1, ..., X_n \) and is given by

\[
\omega_{it} = e_1' \left( \sum_{j:j\neq i} K \left( \frac{X_j - X_i}{h} \right) \hat{X}^{(i)}_j \hat{X}^{(j)t}_j \right) ^{-1} \hat{X}^{(i)}_t K \left( \frac{X_t - X_i}{h} \right).
\] (13)

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 \( H = \text{diag}(1, h u_1, ..., h^r u_r) \), where \( u_j \) is a vector of ones with the same dimensionality as \( (X_t' - X_i') \). Then, noting that \( e_1'H = e_1' \), we can write

\[
\omega_{it} = e_1' \left( H^{-1} \frac{1}{nh^r} \sum_{j:j\neq i} K \left( \frac{X_j - X_i}{h} \right) \hat{X}^{(i)}_j \hat{X}^{(j)t}_j H^{-1} \right) ^{-1} H^{-1} \frac{1}{nh^r} \hat{X}^{(i)}_t K \left( \frac{X_t - X_i}{h} \right).
\]

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

\[
|\omega_{it} - \omega_{ti}| = \frac{1}{nh^r} \left| K \left( \frac{X_t - X_i}{h} \right) \right| \cdot \left| e_1' \tilde{K}^{-1}(X_i) H^{-1} \hat{X}^{(i)}_t - e_1' \tilde{K}^{-1}(X_i) H^{-1} \hat{X}^{(i)}_i \right|
\leq \frac{1}{nh^r} \left| K \left( \frac{X_t - X_i}{h} \right) \right| \cdot \left\{ \left| e_1' \tilde{K}^{-1}(X_i) - \tilde{K}^{-1}(X_i) \right| H^{-1} \hat{X}^{(i)}_t \right| + \left| e_1' \tilde{K}^{-1}(X_i) H^{-1} [\hat{X}^{(i)}_t - \hat{X}^{(i)}_i] \right| \right\}. \tag{14}
\]

The first term in the braces in (14) is bounded as follows. The elements of \( \tilde{K}(X_i) \) are of the form

\[
\frac{1}{nh^r} \sum_{j:j\neq i} (X_j - X_i)^\lambda K \left( \frac{X_j - X_i}{h} \right)
\]

for some \( r \)-vector of nonnegative integers \( \lambda \) 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 (15) converges in probability to \( f(X_i) \int u^\lambda K(u) du \) uniformly in \( X_i \). Hence \( \sup_{i} \|\tilde{K}(X_i) - f(X_i)K\|_\infty = o_p(1) \) for a constant, symmetric matrix \( K \) that only depends on the kernel \( K \). By continuity and \( \mathcal{X} \) compact, it follows that \( \sup_i \|\tilde{K}^{-1}(X_i) - f^{\neg1}(X_i)K^{-1}\|_\infty = o_p(1) \). Since \( f \) is continuously differentiable and is bounded away from zero, \( |f^{\neg1}(x_1) - f^{\neg1}(x_2)| \leq C\|x_1 - x_2\|_\infty \) for all \( x_1, x_2 \in \mathcal{X} \). Combining these observations yields

\[
\|\tilde{K}^{-1}(X_i) - \tilde{K}^{-1}(X_t)\|_\infty
\leq \sup_i \|\tilde{K}^{-1}(X_i) - f^{-1}(X_i)K^{-1}\|_\infty + \|f^{-1}(X_i)K^{-1} - f^{-1}(X_t)K^{-1}\|_\infty + \sup_i \|\tilde{K}^{-1}(X_i) - f^{-1}(X_t)K^{-1}\|_\infty
\leq C\|X_i - X_t\|_\infty + o_p(1) = Ch + o_p(1) \equiv M_n,
\]
where $M_n = o_p(1)$ and is independent of $X_i$ and $X_t$. Hence the first term in the braces in (14) 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 (14) is bounded as follows. Suppose that $r$ is even. Then

$$H^{-1}[\hat{X}_t^{(i)} - \tilde{X}_t^{(i)}] = \left(0, 2 \left(\frac{X_t' - X_t'}{h}\right)^{\Lambda(1)}, z_2, 2 \left(\frac{X_t' - X_t'}{h}\right)^{\Lambda(3)}, \ldots, z_r\right),$$

where the $z_j$, $j$ even, are zero vectors with the same dimensionality as $(X_t' - X_t'^{\Lambda(j)})$. (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 (16) is bounded by 2 since $\|X_t - X_i\|_\infty \leq h$. Therefore it is possible to write

$$\left| e_t'(K^{-1}H^{-1}[\hat{X}_t^{(i)} - \tilde{X}_t^{(i)}]) \right| \leq \left| \frac{1}{f(X_t)} e_t'K^{-1}H^{-1}[\hat{X}_t^{(i)} - \tilde{X}_t^{(i)}] \right| + R_n,$$

where $R_n = o_p(1)$ and does not depend on $X_t$ or $X_i$. By the symmetry properties of the kernel (Assumption 7), the first row of the matrix $K$ has zeros precisely at those positions at which the vector $H^{-1}[\hat{X}_t^{(i)} - \tilde{X}_t^{(i)}]$ is nonzero. A straightforward linear algebra argument (available on request) shows that the first row of the matrix $K^{-1}$ also has zeros at the same positions. Hence, the second term in the braces in (14) is simply bounded by $R_n$. Combining the bounds on the components of (14) shows that

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

where $\tilde{M}_n + R_n = o_p(1)$ and does not depend on $X_t$ or $X_i$. 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 (2005). Let $w = (y, d, z, x)$ and

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

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

$$\Psi(W_i, \Delta, q(X_i)) = Z_iY_i - \frac{(1 - Z_i)Y_i}{1 - q(X_i)},$$

$$\Psi_q(W_i, \Delta, q(X_i)) = -\left( \frac{Z_iY_i}{q^2(X_i)} + \frac{(1 - Z_i)Y_i}{(1 - q(X_i))^2} \right),$$

$$\Psi_{qq}(W_i, \Delta, q(X_i)) = \frac{2Z_iY_i}{q^3(X_i)} - \frac{2(1 - Z_i)Y_i}{(1 - q(X_i))^3},$$

24
and we further define

\[ S_q(X_i) = E[\Psi_q(W_i, \Delta, q(X_i))|X_i] = -\left(\frac{m_1(X_i)}{q(X_i)} + \frac{m_0(X_i)}{1 - q(X_i)}\right), \]

\[ \zeta_i = \Psi_q(W_i, \Delta, q(X_i)) - S_q(X_i), \]

\[ \epsilon_i = Z_i - q(X_i), \]

\[ \beta_n(X_i) = E[\hat{q}(X_i)|X_1, \ldots, X_n] - q(X_i) = \sum_{j:i \neq j} \omega_{ij} q(X_j) - q(X_i), \]

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

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

\[
\sqrt{n}(\hat{\Delta} - \Delta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi(W_i, \Delta, \hat{q}(X_i)) \\
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi(W_i, \Delta, q(X_i)) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_q(W_i, \Delta, q(X_i)) (\hat{q}(X_i) - q(X_i)) \\
+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_q(W_i, \Delta, q^*(X_i)) (\hat{q}(X_i) - q(X_i))^2 \\
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi(W_i, \Delta, q(X_i)) \\
+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_q(X_i) (\hat{q}(X_i) - q(X_i)) \\
+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_i (\hat{q}(X_i) - q(X_i)) \\
+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_q(W_i, \Delta, q^*(X_i)) (\hat{q}(X_i) - q(X_i))^2 \\
\equiv J_0 + J_1 + J_2 + J_3,
\]

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

\[ J_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_q(X_i) (\hat{q}(X_i) - q(X_i)) \\
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_q(X_i) \left( \sum_{j \neq i} \omega_{ij} Z_j - q(X_i) \right) \\
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_q(X_i) \left( \sum_{j \neq i} \omega_{ij} (\epsilon_j + q(X_j)) - q(X_i) \right) \\
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_q(X_i) \epsilon_i + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_q(X_i) \left( \sum_{j \neq i} \omega_{ij} \epsilon_j - \epsilon_i \right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_q(X_i) \left( \sum_{j \neq i} \omega_{ij} q(X_j) - q(X_i) \right) \\
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_q(X_i) \epsilon_i + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \left( \sum_{j \neq i} \omega_{ij} S_q(X_j) - S_q(X_i) \right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_q(X_i) \beta_n(X_i) \\
\equiv J_{11} + J_{12} + J_{13}.\]
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 \( \sup_i \left| \sum_{j:j \neq i} (\omega_{ji} - \omega_{ij}) S_q(X_j) \right| = o_p(1) \). By the bound in (17),

\[
\sup_i \left| \sum_{j:j \neq i} (\omega_{ji} - \omega_{ij}) S_q(X_j) \right| \leq \sup_i \left| \sum_{j:j \neq i} (\omega_{ji} - \omega_{ij}) S_q(X_j) \right| \leq C(M_n + R_n) \sup \frac{1}{n h^r} \left| K \left( \frac{X_j - X_i}{h} \right) \right| = o_p(1) \cdot O_p(1) = o_p(1).
\]

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

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

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

\[
\sup_i \left| \sum_{j:j \neq i} \omega_{ij} S_q(X_j) - S_p(X_i) \right| = o_p(1),
\]

it is true that conditional on the sample path of the \( X_i \), with probability approaching one, \( \sum_{j:j \neq i} \omega_{ij} S_q(X_j) - S_q(X_i) \) is uniformly bounded over \( i \) and converges to zero uniformly over \( i \). Also, the \( \epsilon_i \) 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,

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \left( \sum_{j:j \neq i} \omega_{ij} S_p(X_j) - S_q(X_i) \right) = o_p(1),
\]

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

\[
|J_{13}| = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_q(X_i) \beta_n(X_i) \leq \sup_{x \in \mathcal{X}} \left| \sqrt{n} \beta_n(x) \right| \frac{1}{n} \sum_{i=1}^{n} |S_q(X_i)| = o_p(1) \cdot O_p(1) = o_p(1).
\]

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,

\[
\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_{qq}(W_i, \Delta, q^*(X_i))(\hat{q}(X_i) - q(X_i))^2 \right| \leq \left( \sup_i \sqrt{n}(\hat{q}(X_i) - q(X_i))^2 \right) \frac{1}{n} \sum_{i=1}^{n} \left| \Psi_{qq}(W_i, \Delta, q^*(X_i)) \right| = o_p(1) \cdot O_p(1) = o_p(1).
\]
As a result, we have
\[ \sqrt{n}(\hat{\Delta} - \Delta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi(W_i, \Delta, q(X_i)) + S_q(X_i)e_i + o_p(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \delta(Y_i, D_i, Z_i, X_i) + o_p(1), \]
where the function \( \delta(\cdot) \) is as defined in Lemma 2.

C. Efficiency arguments

The proof of Theorem 3  
First, we provide some simple facts. Under the conditions of Theorem 3, including one-sided non-compliance, the following expressions hold true:

\[ p = P(D = 1) = P(Z = 1, D(1) = 1) = P(D(1) = 1|Z = 1)P(Z = 1) = \Gamma Q, \]
\[ p(x) = P(D = 1|X = x) = P(D(1) = 1|Z = 1, X = x)P(Z = 1|X = x) = \mu_1(x)q(x), \]
\[ \mu_0(x) = E[D|Z = 0, X = x] = E[D(0)|X = x] = 0. \]

Furthermore,
\[ m_1(x) = E[Y|Z = 1, X = x] \]
\[ = E[Y|Z = 1, D = 1, X]P[D = 1|Z = 1, X = x] \]
\[ + E[Y|Z = 1, D = 0, X]P[D = 0|Z = 1, X = x] \]
\[ = E[Y(1)|Z = 1, D = 1, X = x]P[D(1) = 1|Z = 1, X = x] \]
\[ + E[Y(0)|Z = 1, D = 0, X = x]P[D(1) = 0|Z = 1, X = x] \]
\[ = E[Y(1)|X = x]\mu_1(x) + E[Y(0)|X = x](1 - \mu_1(x)) \]
\[ = \rho_1(x) - (1 - \mu_1(x))(\rho_1(x) - \rho_0(x)) \]

and
\[ m_0(x) = E[Y|Z = 0, X = x] \]
\[ = E[Y|Z = 0, D = 0, X = x] \]
\[ = E[Y(0)|Z = 0, D = 0, X = x] = E[Y(0)|X = x] = \rho_0(x). \]

Note that the fourth equality involving the \( m_1(x) \) term holds because the IV assumption and unconfoundedness assumption hold jointly as in (10). This assumption implies both Assumption 1 (IV) and Assumption 2 (unconfoundedness), but the reverse implication is not generally true. 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
\[
\phi_t(Y, D, X) = \frac{p(X)}{p} \left\{ \frac{D(Y - \rho_1(X))}{p(X)} \left( 1 - \frac{(1 - D)(Y - \rho_0(X))}{p(X)} + \frac{D(\rho_1(X) - \rho_0(X) - \beta_1)}{p(X)} \right) \right\}
\]
and rewrite \(\psi_t(Y, D, Z, X)\) as
\[
\psi_t(Y, D, Z, X) = \frac{q(X)}{Q_t} \left\{ \frac{Z[Y - m_1(X) - \tau_t(D - \mu_1(X))]}{q(X)} - \frac{(1 - Z)[Y - m_0(X) - \tau_t(D - \mu_0(X))]}{1 - q(X)} \right.+ \frac{Z[m_1(X) - m_0(X) - \tau_t(\mu_1(X) - \mu_0(X))]}{q(X)}
\]
\[
- \frac{(1 - X)(Y - \rho_0(X))}{1 - q(X)} + \frac{Z(\rho_1(X) - \rho_0(X) - \beta_1)\mu_1(X)}{q(X)}
\]
\[
= \frac{p(X)}{p} \left\{ \frac{Z(Y - \rho_1(X))}{p(X)} + \frac{Z(Y - \mu_1(X))(\rho_1(X) - \rho_0(X))}{p(X)} - \frac{Z\beta_1(D - \mu_1(X))}{p(X)} \right.
\]
\[
- \frac{(1 - Z)(Y - \rho_0(X))}{p(X)} + \frac{Z(\rho_1(X) - \rho_0(X) - \beta_1)\mu_1(X)}{p(X)}
\]
\[
= \frac{p(X)}{p} \left\{ \frac{Z(Y - \mu_1(X))}{p(X)} + \frac{Z(Y - \mu_1(X))(\rho_1(X) - \rho_0(X))}{p(X)} - \frac{Z\beta_1((D - 1) + (1 - \mu_1(X)))}{p(X)} \right.
\]
\[
- \frac{(1 - Z)(Y - \rho_0(X))}{p(X)} + \frac{D(\rho_1(X) - \rho_0(X) - \beta_1)\mu_1(X)}{p(X)}
\]
\[
= \frac{p(X)}{p} \left\{ \frac{Z(Y - \mu_1(X))}{p(X)} - \frac{(1 - Z)(Y - \rho_0(X))}{p(X)} + \frac{Z(\rho_1(X) - \rho_0(X) - \beta_1)}{p(X)} - \frac{Z\beta_1(D - 1)}{p(X)} \right\}
\]
\[
= \frac{p(X)}{p} \{ \psi_1 - \psi_2 + \psi_3 - \psi_4 \}.
\]

Note that
\[
E[\phi_t \psi_1 | X] = \frac{E \left[ Z D(Y - \rho_1(X))^2 | X \right]}{p^2(X)} = \frac{E \left[ D(Y - \rho_1(X))^2 | X \right]}{p^2(X)}
\]
\[
= \frac{E \left[ D(Y - \rho_1(X))^2 | X, D = 1 \right] p(D = 1 | X = x)}{p^2(X)} = \frac{\sigma^2(x)}{p(X)},
\]
where \(\sigma^2(x) = V(Y(1)|X)\). 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
\[
E[\phi_t \psi_3 | X] = \frac{\rho_1(X) - \rho_0(X) - \beta_1}{p^2(X)} E[DZ(Y - \rho_1(X)) | X]
\]
\[
= \frac{\rho_1(X) - \rho_0(X) - \beta_1}{p^2(X)} E[D(Y(1) - \rho_1(X)) | X] = 0,
\]
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,

$$E[\phi_1 \psi_1 | X] = \frac{E[Z(1-D)(Y - \rho_1(X))(Y - \rho_0(X))|X]}{p(X)(1-p(X))}$$

$$= \frac{E[Z(1-D)(Y - \rho_1(X))(Y - \rho_0(X))|X, Z = 1, D = 0]P(Z = 1, D = 0|X)}{p(X)(1-p(X))}$$

$$= \frac{E[(Y(0) - \rho_1(X))(Y(0) - \rho_0(X))|X, Z = 1, D = 0](1 - \mu_1(X))q(X)}{p(X)(1-p(X))}$$

$$= \frac{\sigma_0^2(X)(1 - \mu_1(X))q(X)}{p(X)(1-p(X))} \frac{1 - \mu_1(X)}{1 - p(X)} \mu_1(X).$$

$$E[\phi_1 \psi_2 | X] = \frac{E[(1-Z)(1-D)(Y - \rho_0(X))^2|X]}{(1-p(X))(1-q(X))\mu_1(X)}$$

$$= \frac{E[(1-Z)(1-D)(Y - \rho_0(X))^2|X, Z = 0, D = 0]P(Z = 0, D = 0|X)}{(1-p(X))(1-q(X))\mu_1(X)}$$

$$= \frac{E[(Y(0) - \rho_0(X))^2|X, D = 0]P(Z = 0|X)}{(1-p(X))(1-q(X))\mu_1(X)}$$

$$= \frac{\sigma_0^2(X)(1 - q(X))}{(1-p(X))(1-q(X))\mu_1(X)} = \frac{\sigma_0^2(X)}{1 - p(X)} \frac{1}{\mu_1(X)}.$$

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

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

Consequently,

$$Cov(\phi_i, \psi_i) = E[\phi_i \psi_i]$$

$$= E \left[ \frac{p^2(X)}{p} \left\{ \frac{\sigma_0^2(X)}{p(X)} \frac{1 - \mu_1(X)}{\mu_1(X)} \frac{1}{1 - p(X)} \right\} \right]$$

$$= E[\phi_i^2] = Var(\phi_i).$$

This shows Theorem 3. ■
References


