TESTING LOCAL AVERAGE TREATMENT EFFECT ASSUMPTIONS

Ismael Mourifié and Yuanyuan Wan*

Abstract—In this paper, we propose an easy-to-implement procedure to test the key conditions for the identification and estimation of the local average treatment effect (LATE; Imbens & Angrist, 1994). We reformulate the testable implications of LATE assumptions as two conditional inequalities, which can be tested in the intersection bounds framework of Chernozhukov, Lee, and Rosen (2013) and easily implemented using the Stata package of Chernozhukov et al. (2015). We apply the proposed tests to the draft eligibility instrument in Angrist (1991), the college proximity instrument in Card (1993), and the same-sex instrument in Angrist and Evans (1998).

I. Introduction

THE instrumental variable (IV) method is one of the most commonly used techniques in applied economics to identify the causal effect of an endogenous treatment on a particular outcome. In the framework of potential outcome models, a valid instrument is often assumed to be independent of all potential outcomes and potential treatments; in the meantime, it must have no effect on the observed outcome beyond its effect on the observed treatment. Imbens and Angrist (1994, IA hereafter) showed that a valid instrument itself does not ensure that the IV estimand identifies the average treatment effect (ATE) when the treatment effect is heterogeneous. To deal with this issue, IA (1994) introduced LM (LATE monotonicity also known as the "no defiers" assumption), which assumes the instrument affects the treatment decision in the same direction for every individual. When both LI (LATE independence) and LM hold, IA (1994) showed that the IV estimand identifies the ATE for the subpopulation of compliers, namely, the LATE.

Although the results of IA (1994) have been widely influential in the applied economics literature, there are still concerns about the validity of the key assumptions. For instance, Dawid (2000) discussed applications where LM is likely to be violated. Such concerns, however, cannot be directly verified since LM itself is not testable, as discussed in IA (1994). Balke and Pearl (1997) and Heckman and Vytlacil (2005) first discussed testable implications of the joint assumptions of LI and LM. Based on these insights, Kitagawa (2008, 2015) showed that this set of testable implications is a sharp characterization of LM and LI, in the sense that it is the most informative set of testable implications for detecting observable violations of the joint LI and LM assumptions and first proposed a test for these implications.

In this paper, we revisit the existing discussions on testing the joint validity of LM and LI and show that this set of testable implications can be tested in an easy-to-implement way. In particular, we reveal that the sharp characterization of LI and LM can be represented by a set of conditional moment inequalities. The novelty, and a nice feature, of this conditional moment inequality representation, is that the outcome variable enters the inequalities as a conditioning variable, and one can easily incorporate additional covariates into the moment inequalities as additional conditioning variables. Interestingly, with this representation, the sharp testable implications of both LI and LM assumptions can be tested using the intersection bounds framework of Chernozhukov, Lee, and Rosen (2013, CLR hereafter). The test can be implemented with the Stata package provided by Chernozhukov et al. (2015), which is readily available for empirical researchers to use.

This testing procedure is different from but complements the (variance-weighed) Kolmogorov-Smirnov test proposed by Kitagawa (2015). First, the two tests have different power properties. Kitagawa's (2015) test has nontrivial power against root-n local alternatives provided that the limit of the alternatives admits a contact set of outcome variable with strictly positive probability mass. We consider a conditional moment inequality reformulation and apply CLR's test, which has nontrivial power against local alternatives subject to a nonparametric rate but does not require such a contact set restriction. As discussed in CLR, both cases are important in applications. Second, the proposed testing procedure requires local linear regression and therefore the choice of a smoothing constant. We follow CLR and use the rule-of-thumb choice given by Fan and Gijbels (1996) in our empirical applications. Kitagawa's (2015) test is based on empirical distribution functions, whose variance-weighted version requires a choice of a trimming constant to ensure the inverse weighting terms to be bounded away from 0. Third, the test can accommodate continuous covariates within the same framework. Indeed, as we further elaborate in section VI, it requires no more than adding covariates as new conditioning variables in the moment inequalities and estimating the conditional expectation of the instrument given covariates. Kitagawa (2015) follows Andrews and Shi's (2013) approach to transform the testable implication to unconditional moment restrictions. Finally, our testing procedure can be easily implemented using the Stata package provided by Chernozhukov et al. (2015). Other papers discuss testing issues under a different setup. Machado, Shaikh, and Vytlacil (2013) proposed tests for LM or outcome monotonicity (in treatment) in a binary treatment,

Received for publication May 20, 2015. Revision accepted for publication March 30, 2016. Editor: Bryan S. Graham.

* University of Toronto.

We are grateful to Victor Aguirregabiria, Clement de Chaisemartin, Christian Gourieroux, Bryan Graham, James Heckman, Marc Henry, Sung Jae Jun, Toru Kitagawa, Alfonso Flores-Lagunes, Matt Masten, Angelo Melino, Joris Pinkse, Adam Rosen, Paul Schrimpf, and two anonymous referees for valuable comments. We have greatly benefited from insightful comments from Edward Vytlacil. We benefited from discussions with participants at the 9th GNYEC, the LATE workshop at Pennsylvania State University, Cowles Foundation Conference 2014, AMES 2014, CESG 2014, SEA 2014, and department seminars at University of Texas at Austin, University of Rochester, National University of Singapore, Duke University, and Northwestern University. We thank Michael Anthony for excellent research assistance. Y.W. thanks SSHRC Insight Development grant 430130275 for support. I.M. thanks Connaught Fund for support. All errors are ours.

The Review of Economics and Statistics, May 2017, 99(2): 305–313

doi:10.1162/REST_a_00622

binary instrument, and binary outcome setup while maintaining the LI assumption. Huber and Mellace (2013) considered a model in which the instrument respects mean independence rather than full independence and proposed a specification test based on a different set of testable implications.

Our paper also contributes to the empirical literature. We apply the proposed test to three well-known instruments used in the literature: the draft eligibility instrument, the college proximity instrument, and the same-sex instrument. Angrist (1991) analyzed the effect of veteran status on civilian earnings using the binary indicator of the draft eligibility as instrument. Card (1993) analyzed the effect of schooling on earning using a binary indicator of whether an individual was born close to a four-year college. Angrist and Evans (1998) studied the causal relationship between fertility and women's labor income using the variable that the first two children are of the same sex as the instrument. Our test does not reject the testable implication of LI + LM for draft eligibility and same-sex instrument. We do, however, find that the implication is rejected for the college proximity instrument on the subgroup of nonblack men who lived in the metropolitan area of southern states. The rejection mainly takes place among individuals with higher labor income.

The rest of the paper is organized as follows. Section II presents the analytical framework. In section III, we revisit the testable implications of the LATE assumptions, followed by section IV, which presents our testing procedure. We discuss empirical applications in section V. The last section extends our analysis to the case with additional covariates.

II. Analytical Framework

We adopt Rubin's (1974) potential outcome model. Let $Y = Y_1D + Y_0(1 - D)$, where Y is the observed outcome taking values from the support $\mathcal{Y}$, $D \in \{0, 1\}$ is the observed treatment indicator, and (Y_1, Y_0) are potential outcomes. Let Z be the instrumental variable. For simplicity, we assume $Z \in \mathcal{Z} = \{0, 1\}$, but our analysis can be extended to allow for multivalued Z. For each $z \in \mathcal{Z}$, let D_z be the potential treatment if Z had been exogenously set to z. With this notation, we can also write the observed treatment $D = D_1Z + D_0(1 - Z)$.

The two well-known identification assumptions for LATE as introduced by IA (1994) are restated as the following:

Assumption 1 (LATE Independence -LI). $Z \perp (Y_1, Y_0, D_0, D_1)$ and $\mathbb{P}(D = 1|Z = 0) \neq \mathbb{P}(D = 1|Z = 1)$.

Assumption 2 (LATE Monotonicity -LM). Either $D_0 \leq D_1$ almost surely or $D_0 \geq D_1$ almost surely.

For each d and z, let $D_z^{-1}(d)$ denote the subset of the individuals in the population who would select treatment d had the instrument been exogenously set to z. LM then implies that we have either $D_0^{-1}(1) \subseteq D_1^{-1}(1)$ or $D_1^{-1}(1) \subseteq D_0^{-1}(1)$. In general, the economic context suggests to empirical researchers the direction of the monotonicity. In this paper, we assume that the hypothetical direction is known to researchers. Without loss of generality (w.l.o.g.), we focus on the direction of $D_0 \leq D_1$ in the rest of the paper.

TABLE 1.—SUBPOPULATIONS

	D_0	D_1	Proportion
a: Always takers	1	1	π_{11}
def: Defiers	1	0	π_{10}
c: Compliers	0	1	π_{01}
n: Never takers	0	0	π_{00}

TABLE 2.—OBSERVED SUBGROUPS AND UNOBSERVED SUBPOPULATIONS

	$Z = 0$	$Z = 1$
$D = 0$	$\pi_{00} + \pi_{01}$	$\pi_{00} + \pi_{10}$
$D = 1$	$\pi_{10} + \pi_{11}$	$\pi_{01} + \pi_{11}$

III. Testable Implications of the LATE Assumptions

In this section, we revisit a set of sharp testable implications of the LATE assumptions (LI and LM). For ease of exposition, we first list in table 1 the standard notion of four subpopulations defined by the potential treatments—always takers, defiers, compliers, and never takers—and we use π_{ij}, $i, j \in \{0, 1\}$ to denote the corresponding probability mass.

Every observed subgroup $\{D = d, Z = z\}$ for $d, z \in \{0, 1\}$ is composed of a mixture of unobserved subpopulations. Indeed,

$$\begin{aligned}\mathbb{P}(D = 0|Z = 0) &= \mathbb{P}(D_1Z + D_0(1 - Z) = 0|Z = 0)\\ &= \mathbb{P}(D_0 = 0|Z = 0) = \mathbb{P}(D_0 = 0, D_1 = 0)\\ &\quad + \mathbb{P}(D_0 = 0, D_1 = 1) = \pi_{00} + \pi_{01},\end{aligned}$$

where the third equality holds under assumption 1. By a similar derivation, we can obtain the other three conditional probabilities, as summarized in table 2.

Notice that by definition, we can easily see that LM is equivalent to the nonexistence of defiers (i.e., $\pi_{10} = 0$). Let $\mathcal{B}_\mathcal{Y}$ be a collection of Borel sets generated from $\mathcal{Y}$; then LM and LI necessarily imply that for an arbitrary $A \in \mathcal{B}_\mathcal{Y}$,

$$\begin{aligned}\mathbb{P}(Y \in A, D = 1|Z = 0) &= \mathbb{P}(Y_1 \in A, D = 1|Z = 0)\\ &= \mathbb{P}(Y_1 \in A, D_0 = 1|Z = 0) = \mathbb{P}(Y_1 \in A, D_0 = 1)\\ &\leq \mathbb{P}(Y_1 \in A, D_1 = 1) = \mathbb{P}(Y \in A, D = 1|Z = 1), \quad (1)\end{aligned}$$

where the third and fourth equalities hold by LI and the first inequality holds by LM. Similarly, we have

$$\mathbb{P}(Y \in A, D = 0|Z = 1) \leq \mathbb{P}(Y \in A, D = 0|Z = 0). \quad (2)$$

Therefore, as long as there exists $A \in \mathcal{B}_\mathcal{Y}$ such that either inequality (1) or (2) is violated, we must reject the joint assumptions of LM + LI assumptions. Note that inequalities (1) and (2) are not sufficient for the joint assumptions to hold in the sense that there could exist a potential outcome model in which both (1) and (2) hold but LM + LI is violated.[1]

Inequalities (1) and (2) need not be the only set of testable implications of LM and LI. Theorem 1 shows, however, that they are the sharp characterization of LI and LM in the sense that whenever inequalities (1) and (2) hold, there always exists another potential outcome model compatible with the data in which LI and LM hold.

Theorem 1 (Sharp characterization of the LATE assumptions). *Let Y, D_1, D_0, Y_1, Y_0, Z define a potential outcome model $Y = Y_1 D + Y_0(1 - D)$. (i) If LM and LI hold, then equations (1) and (2) hold. (ii) If equations (1) and (2) hold, there exists a joint distribution of $(\tilde{D}_1, \tilde{D}_0, \tilde{Y}_1, \tilde{Y}_0, Z)$ such that LM and LI hold and $(\tilde{Y}, \tilde{D}, Z)$ has the same distribution as (Y, D, Z).*

Theorem 1 is essentially equivalent to but presented in a different way from Kitagawa (2015, proposition 1.1) and the proof is therefore omitted. The sharpness result shows that inequalities (1) and (2) are the most informative observable restrictions for assessing the validity of the joint LI and LM assumptions. However, whenever the cardinality of the outcome space is large, the number of inequalities to visit is very high because the number of inequalities to be checked is equal to the number of subsets of the set of observable outcomes. When Y is continuous, there are infinitely many elements in $\mathcal{B}_{\mathcal{Y}}$. In practice, the performance of a test also depends on the subsets we search through, especially when many of the inequalities are redundant. One solution is to follow the idea discussed in Galichon and Henry (2006, 2011) and Chesher, Rosen, and Smolinski (2013) to find a low (or the lowest) cardinality collection of sets that are sufficient to characterize all the restrictions imposed by inequalities (1) and (2). To the best of our knowledge, the issue of finding the smallest collection of sets in a generic setup remains open. To deal with this important issue, we propose to use an alternative representation. Note that for every $A \in \mathcal{B}_{\mathcal{Y}}$, there is $\mathbb{P}(Y \in A, D = 1 | Z = 1)\mathbb{P}(Z = 1) = \mathbb{P}(D = 1, Z = 1, Y \in A)$. Let $\mathbf{1}_{Y\in\cdot}$ be the indictor function. Inequalities (1) and (2) can be written as

$$\mathbb{E}[\mathbf{1}_{Y\in A} D(1 - Z)]\mathbb{P}(Z = 1) \le \mathbb{E}[\mathbf{1}_{Y\in A} DZ]\mathbb{P}(Z = 0) \quad (3)$$

and

$$E[\mathbf{1}_{Y\in A}(1 - D)Z]\mathbb{P}(Z = 0) \le E[\mathbf{1}_{Y\in A}(1 - D)(1 - Z)]\mathbb{P}(Z = 1). \quad (4)$$

Since $A \in \mathcal{B}_{\mathcal{Y}}$, the above inequalities hold with a class of cubes too. We can apply Andrews and Shi (2013, lemma 3) and further write them as $\forall y \in \mathcal{Y}$,

$$\begin{cases} \theta(y, 1) \equiv \mathbb{E}[c_1 D(1 - Z) - c_0 DZ | Y = y] \le 0 \\ \theta(y, 0) \equiv \mathbb{E}[c_0(1 - D)Z - c_1(1 - D)(1 - Z) | Y = y] \\ \qquad \le 0, \end{cases} \quad (5)$$

where $c_k = \mathbb{P}(Z = k)$ for $k = 0, 1$. Let $\mathcal{V} = \mathcal{Y} \times \{0, 1\}$, and then the null hypothesis can be formulated as

$$H_0 : \theta_0 \equiv \sup_{v\in\mathcal{V}} \theta(v) \le 0, \quad H_1 : \theta_0 > 0. \quad (6)$$

The advantage of considering the hypothesis stated in equation (6) is to facilitate implementation.

With our formulation, researchers do not have to find the lowest cardinality collection of sets and can simply apply the existing inference methods in CLR, as explained in the following section.

IV. Testing Procedures

In this section, we formalize a testing procedure for the hypotheses specified in equation (6), that is,

$$H_0 : \theta_0 \equiv \sup_{v\in\mathcal{V}} \theta(v) \le 0, \quad H_1 : \theta_0 > 0,$$

where $v \in \mathcal{Y} \times \{0, 1\}$. We propose to use the intersection bounds framework of CLR, which provides an inference procedure for bounds defined by supremum (or infimum) of a nonparametric function. To be more specific, let $0 < \alpha < \frac{1}{2}$ be the prespecified significance level, and we reject the H_0 if $\hat{\theta}_\alpha > 0$, where

$$\hat{\theta}_\alpha \equiv \sup_{v\in\mathcal{V}} \{\hat{\theta}(v) - s(v)k_\alpha\},$$

and $\hat{\theta}(\cdot)$ is the local linear estimator for $\theta(\cdot)$. $s(\cdot)$ and k_α are estimates for point-wise standard errors and critical value, respectively. For implementation, one does not have to calculate $\hat{\theta}$, s, and k_α explicitly; therefore, we leave their expressions in appendix A.1 for the sake of exposition. The testing procedure can be easily implemented in Stata as follows:

Implementation

1. Estimate c_1 and c_0 by $\hat{c}_1 = \frac{1}{n}\sum_{i=1}^{n} Z_i$ and $\hat{c}_0 = 1 - \hat{c}_1$, respectively.
2. Let $\hat{L}_i^1 = \hat{c}_1 D_i(1 - Z_i) - \hat{c}_0 D_i Z_i$ and $\hat{L}_i^0 = \hat{c}_0(1 - D_i)Z_i - \hat{c}_1(1 - D_i)(1 - Z_i)$.
3. Implement the CLRtest command with two conditional moment inequalities. Specify $\hat{L}_i^1$ and $\hat{L}_i^0$ as the dependent variables for each conditional inequality, respectively. Specify Y_i as the conditioning variable for both inequalities. See Chernozhukov et al. (2015) for the full set of options.

[1] Chaisemartin (2013) refers to this as "weak more compliers than defiers."

We make the following assumptions:

Assumption 3. $\{(D_i, Y_i, Z_i)\}_{i=1}^n$ are i.i.d observations.

Assumption 4. $\mathcal{Y}$ is convex and compact. For each (d, z), the conditional density of Y given $(D, Z) = (d, z)$ is bounded away from 0 and twice continuously differentiable.

We assume the continuity of Y in assumption 4 only for the purpose of exposition. If Y has finite discrete support, the conditional inequalities in equation (10) can be represented by a finite number of unconditional expectations. In this scenario, the test is "parametric" and can still be implemented within the framework.[2] Assumptions 5 and 6 are conditions on the choice of kernel and bandwidth, respectively.

Assumption 5. $K(\cdot)$ has support on $[-1, 1]$, is symmetric and twice differentiable, and satisfies $\int K(u)du = 1$.

Assumption 6. $nh^4 \to \infty$, and $nh^5 \to 0$ at polynomial rates in n.

Proposition 1 is an application of CLR (theorem 6) that verifies the consistency and validity of the proposed testing procedure.

Proposition 1. *Suppose that assumptions 3 to 6 are satisfied; then equation (1) under* H_0, $\mathbb{P}(\hat{\theta}_\alpha > 0) \leq \alpha + o(1)$*; equation (2) if* $\theta(y, k) = 0$ *for all* $y \in \mathcal{Y}$ *and* $k \in \{0, 1\}$, *then* $\mathbb{P}(\hat{\theta}_\alpha > 0) \to \alpha$*; and equation (3) if* $\sup_{y \in \mathcal{Y}_a, k \in \{0,1\}} \theta(y, k) > \mu_n \sqrt{\log n / nh}$ *for any* $\mu_n \to \infty$, *then* $\mathbb{P}(\hat{\theta}_\alpha > 0) \to 1$.

Proof. See section A.1 in the appendix.

Several observations were formed. First, our test is a type of sup-tests based on conditional moment inequalities specified in expression (10) and hence does not require researchers to find the lowest collection of sets. Our test is consistent against any fixed alternatives and local alternatives subject to the nonparametric estimation rate of $\theta(\cdot, \cdot)$. Second, regarding our test, continuous covariates can be easily incorporated as additional conditioning variables. Finally, because of the availability of the STATA package, our test can be easily applied by empirical researchers to assess the validity of the LATE assumptions.

Table 3.—Summary Statistics of a PUMS Sample

Observations	Total 403,011	$D = 1$ 119,221	$Z = 1$ 202,232
Age	33.805 (5.420)	34.026 (4.968)	33.820 (5.423)
Years of schooling	11.119 (2.339)	10.760 (2.493)	11.119 (2.332)
Race (nonwhite)	0.177 (0.381)	0.220 (0.415)	0.179 (0.382)
Having the third child ($D = 1$)	0.296 (0.456)	1.000 (0.000)	0.325 (0.468)
First two same sex ($Z = 1$)	0.502 (0.499)	0.553 (0.497)	1.000 (0.000)
Log wage	9.014 (1.227)	8.803 (1.278)	9.010 (1.229)

Average (standard deviation).

V. Applications

In this section we apply our test to three well-known instruments used in the literature: the same-sex instrument in Angrist and Evans (1998), the draft eligibility instrument in Angrist (1991), and the college proximity instrument in Card (1993).

A. *The Same-Sex Instrument*

Our first application is about the same-sex instrument used by Angrist and Evans (1998), who studied the relationship between fertility and labor income. This study was complicated by the endogeneity of fertility. Angrist and Evans (1998), proposed using the sibling-sex composition to construct the IV estimator of the effect of childbearing on the labor supply. In this application, $D = 1$ denotes that the household had a third child, and $Z = 1$ denotes that the first two children are of the same sex. The direction of monotonicity under testing is $D_1 \geq D_0$.

We consider a sample from the 1990 Census Public Micro Samples (PUMS). The data contain information on age, gender, race, education, labor income, and number of children. We consider women with at least two children, between 21 and 50 years old, and with positive wage. This gives us a sample of of $403,011$ individuals. The outcome variable of interest is log wage. Summary statistics for the sample to which we apply the test are given in table 3.

We divided $403,011$ observations into 24 subgroups according to race (white or nonwhite), education (lower than high school, high school, or higher than high school), and age (21–28, 29–35, 36–42, 43–50) and conducted tests on each of these groups. The subgroups' sizes are reported in table 4. Due to the memory constraint of our computer, we implemented our test on randomly drawn subsamples of 25,000 for subgroups whose sizes are larger than this number.[3]

Throughout this section, we use the default choices of bandwidth and kernel functions recommended in

[2] In the discrete outcome case, we can show that $\{\{y_1\}, \{y_2\}, \cdots, \{y_J\}\}$ is the lowest cardinality collection of sets that are sufficient to characterize all the restrictions imposed on the model. Therefore, without loss of generality, restriction (1) can be written as

$$\theta(y, 1) \equiv \sum_{j=1}^{J} \mathbf{1}[y = y_j]\beta_{1j} \leq 0,$$

where $\beta_{1j} = \mathbb{P}(Z = 1)\mathbb{E}[D(1 - Z)|Y = y_j] - \mathbb{P}(Z = 0)\mathbb{E}[DZ|Y = y_j]$. $\theta(y, 0)$ and β_{0j} can be similarly defined for restriction (2). Both β_{1j} and β_{0j} can be consistently estimated by estimators that converge at root-n rate and have limiting normal distribution with estimable covariance matrix. To implement, one can then follow the discussions in CLR (p. 709).

[3] As a robustness check, we repeated the test over different subsamples of 25,000 for each of these large subgroups and obtained the same conclusion.

TABLE 4.—SUBGROUP SIZES OF THE PUMS SAMPLE

	21–28	29–35	36–42	43–50
White, <high school	9,871	13,986	4,788	751
White, high school	36,386	89,449	55,279	6,749
White, >high school	7,234	43,376	52,793	10,906
Nonwhite, <high school	4,718	7,283	3,195	597
Nonwhite, high school	10,137	18,468	8,771	1,135
Nonwhite, >high school	1,395	6,724	7,223	1,797

TABLE 5.—TESTING RESULTS FOR ALL THREE INSTRUMENTS: LINEAR SPECIFICATION

AE1998 Same Sex			Angrist 1991 Lottery			Card 1993 Proximity		
10%	5%	1%	10%	5%	1%	10%	5%	1%
NR	NR	NR	R	R	R	R	R	R

R: rejection; NR: no rejection. For linear specification, see the parametric option of the CKLR package.

CLR and Chernozhukov et al. (2015), that is, $K(u) = \frac{15}{16}(1-u^2)^2\mathbf{1}\{|u| \leq 1\}$ and $h_{ROT} \times \hat{s} \times n^{\frac{1}{5}} \times n^{-\frac{2}{7}}$, where h_{ROT} is the rule-of-thumb choice given by Fan and Gijbels (1996). To avoid the boundary issue, for each subgroup, we compute the maximum in the test statistics over the interval $[Q_{2.5\%}, Q_{97.5\%}]$, where Q_α is the α-quantile of the subgroup under testing.

Since we conducted tests on 24 subpopulations $s \in \{1, 2, 3, \cdots, 24\}$, we can view $H_0 = H_0^{(1)} \cap H_0^{(2)} \cap \cdots \cap H_0^{(24)}$, where H_0 is defined as "inequality (10) holds for every subpopulation" and $H_0^{(s)}$ is defined as "inequality (10) holds for the subpopulation s." Rejection of any of $H_0^{(s)}$ implies rejection of H_0. Since we are checking a large number of subpopulations, it is desirable to ensure that the familywise error rate (FWER) is controlled at targeted levels. We consequently adapt the multiple testing procedure of Holm (1979), a suitable framework to consider (see also an empirical implementation in Bhattacharya, Shaikh, & Vytlacil, 2012). The testing results show that the smallest p-values among all 24 groups are greater than 10%.[4] Hence we are able to conclude that the multiple testing procedures reject no null hypothesis at the 10% level.[5] Because sex mix is virtually randomly assigned, this result can be interpreted as evidence of the relative preference for the mix-sibling sex over the same sex within our population of interest.

We also conducted the test with the parametric regression method,[6] using the three demographic variables as regressors. The null hypothesis is not rejected at all three significance levels (see table 5), which is consistent with the results obtained from the local linear methods.

[4] The Stata command does not report the p-value for the "clrtest," but one can always set difference significance levels and find the marginal one that gives rejection.

[5] In the sample, there are 35.06% of observations with missing wage. We excluded those observations. We also conducted a pointwise test conditional on the missing wage subsample. The null hypothesis is not rejected either.

[6] In CLR, "parametric regression" means that $\theta(y, k)$ is a known function (up to finite dimensional parameters) of y for each k. In the Stata package, "parametric regression" specifically means $\theta(y, k)$ is linear in y for each k. It is worth noting that the Stata parametric regression option has the advantage of allowing for multiple conditioning variables.

TABLE 6.—SUMMARY STATISTICS OF SIPP DATA FROM ANGRIST (1991)

Observations	Total 3,027	Draft Eligible ($Z = 1$) 1,379	Veteran ($D = 1$) 994
Age	34.063 (2.804)	34.685 (2.607)	35.064 (2.494)
Veteran ($D = 1$)	0.328 (0.470)	0.403 (0.491)	1.000 (0.000)
Draft eligible ($Z = 1$)	0.456 (0.498)	1.000 (0.000)	0.560 (0.497)
Years of schooling	13.522 (2.864)	13.578 (2.834)	13.443 (2.260)
Race (nonwhite)	0.118 (0.322)	0.116 (0.320)	0.080 (0.272)
log (weekly wage)	2.217 (0.532)	2.247 (0.534)	2.248 (0.498)

Average (standard deviation).

TABLE 7.—TESTING RESULTS FOR THE LOTTERY INSTRUMENT

	W,<HS	W,=HS	W,>HS	NW,<HS	NW,=HS	NW,>HS
Subgroup ID	1	2	3	4	5	6
Observations	317	865	1,478	56	129	171
10%	NR	NR	NR	NR	R	NR
5%	NR	NR	NR	NR	R	NR
1%	NR	NR	NR	NR	R	NR

W: white; NW: nonwhite; HS: high school; R: rejection; NR: no rejection.

B. *The Draft Eligibility Instrument*

Our second empirical application is about the draft eligibility instrument in Angrist (1991), who studied the effect of veteran status on civilian earnings. Endogeneity arises since enlisting for military service possibly involves self-selection. To deal with the issue, Angrist (1991) constructed the binary indicator of draft eligibility, which is theoretically randomly assigned based on one's birth date through the draft lotteries. In this application, $D = 1$ denotes the veteran status and $Z = 1$ denotes the individual was drafted. The direction of monotonicity under testing is $D_1 \geq D_0$.

We used a sample of 3,071 individuals from the 1984 Survey of Income and Program Participation (SIPP).[7] (See the summary statistics in table 6.) The sample was divided into six groups according to race (white or nonwhite) and their educations levels (lower than high school, high school, or higher than high school). We then performed our test using the local method for each group. Again, we compute the maximum in the test statistics over the interval $[Q_{2.5\%}, Q_{97.5\%}]$.

Testing results for individual groups are reported in table 7. Note that the null hypothesis ($H_0^{(5)}$) is rejected at subgroup 5 of nonwhite persons with high school education at the 10% and 5% levels, respectively, but not all three levels. However, as shown in figure 1, it is likely due to the boundary issue or small subgroup size. Therefore, we do not consider this as strong evidence against $H_0^{(5)}$. Following the similar arguments as in the same-sex application, we can indirectly verify that we reject no null hypotheses with FWER controlled at 10%.[8] Kitagawa (2015) obtained the same result without conditioning on subgroups.

[7] The data are available from Angrist's website. After removing all entries with missing information, 3,071 individuals remain.

[8] The testing procedure with linear specification on the conditional expectation, however, rejects the null hypothesis at all three levels (see table 5).

FIGURE 1.—$\hat{\theta}(\cdot, 1)$ AND $\hat{\theta}(\cdot, 1) - s(\cdot, 1) \times \hat{c}_{\hat{V},0.95}$

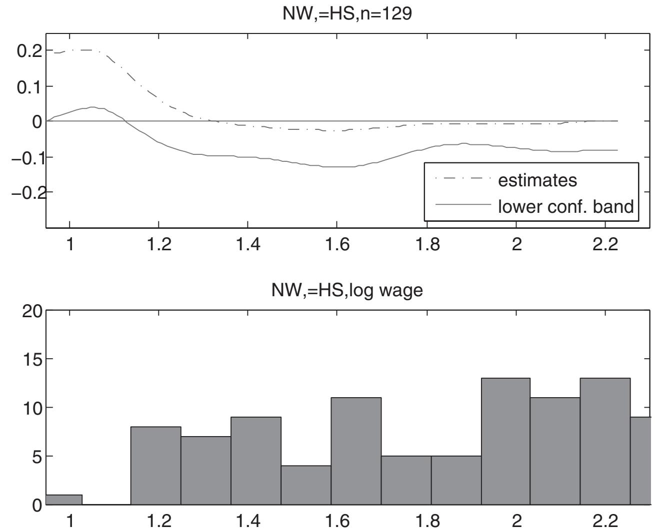


TABLE 8.—SUMMARY STATISTICS OF NLSYM SAMPLE

Observations	Total 3,005	$D = 1$ 2,048	$Z = 1$ 816
Lived in metro area in 1966	0.651 (0.476)	0.693 (0.461)	0.801 (0.399)
Lived in southern states in 1966	0.414 (0.492)	0.313 (0.464)	0.329 (0.470)
Black	0.232 (0.422)	0.099 (0.299)	0.209 (0.407)
Years of schooling in 1976	13.26 (2.675)	16.692 (0.849)	13.532 (2.577)
D (education $\geq$ 16)	0.271 (0.444)	1.000 (0.000)	0.293 (0.455)
Z (college proximity)	0.681 (0.465)	0.736 (0.015)	1.000 (0.000)
Y (log wage in 1976)	6.261 (0.444)	6.428 (0.433)	6.311 (0.440)

Average (standard deviation).

C. The College Proximity Instrument

Card (1993) studied the causal effect of schooling on earnings and employed college proximity as the exogenous source of variation in eduction outcome. In this application, $Z = 1$ denotes there is a four-year college in the local labor market where the individual was born, and $D = 1$ denotes the individual has at least sixteen years of education. The outcome variable is the log wage in 1976. The monotonicity under testing is $D_1 \geq D_0$.

The data from the National Longitudinal Survey of Young Men (NLSYM) began in 1966 with men aged 14 to 24 and continued with a follow-up survey until 1981. Some summary statistics are reported in table 8.[9] We considered three binary control variables: lived in southern states in 1966, lived in a metropolitan area in 1966, and being black. Table 9 reports the corresponding subgroup sizes.

We conduct the test on six subgroups. We exclude the subgroup NS/NM/B because of its small sample size; we also exclude subgroup NS/M/B because of the high frequency of $Z = 1$ (92%). Note that the null hypothesis $H_0^{(4)}$ is rejected in subgroup 4 of nonblack men who lived in the metropolitan area of the southern states, as well as for the whole sample at the 0.5% level. No rejection happens with other subgroups even at the 10% level. The results in table 10 imply that the multiple testing procedure of Holm (1979) would conclude that H_0 is rejected with the FWER controlled by no more than $0.5\% \times 6 = 3\%$. The testing procedure with parametric methods gives the same results.

TABLE 9.—SUBGROUP SIZES OF CARD (1993)

	Nonblack (NB)	Black (B)
Nonsouthern (NS) and Nonmetro (NM)	429	5
Nonsouthern (NS) and Metro (M)	1,191	138
Southern (S) and Nonmetro (NM)	307	314
Southern (S) and Metro (M)	380	246

Southern (south66): lived in southern states in 1966. Metro (smsa66r): lived in urban area in 1966.

Now it will be interesting to know on which subsets of Y the null hypothesis is violated. Figure 2 plots the $\hat{\theta}(\cdot, 0)$ and $\hat{\theta}(\cdot, 0) - s(\cdot, 0) \times \hat{c}_{\hat{V},0.95}$ for subgroup 4 and the whole sample, respectively. Note that θ_0 is in general increasing in Y, and the rejection takes place on higher-income subpopulations (e.g., subpopulations whose observed log wage is around 7). The density of log wage is reasonably high at this point, and therefore the rejection is unlikely due to the boundary issue of the local linear estimation.

To summarize, our result suggests that the Wald estimator in such a case can be inconsistent. Thereby, although the college proximity instrument seems good, researchers must be aware that this instrument would not be a good one to use when the treatment effect is heterogeneous.

VI. Extensions

In this section we discuss three ways of incorporating covariates X into the testing procedure. All three can be implemented with the same test procedure proposed. Let $\mathcal{X}$ be the support of X. We then make the following assumptions:

Assumption 7. $(Y_1, Y_0, D_0, D_1) \perp Z|X = x$ and $\mathbb{P}(D = 1|Z = 0, X = x) \neq \mathbb{P}(D = 1|Z = 1, X = x)$ for all $x \in \mathcal{X}$.

Assumption 7 is common in the literature (see, e.g., Abadie, 2003), which requires the independence assumption to hold conditional on X. Sometimes the independence assumption between potential outcomes and potential treatments may hold for some observed subgroups and not for others. In such a case, researchers would be curious to know for each observed group whether the independence assumption holds. The following assumption could be used to model this case:

Assumption 8. $(Y_1, Y_0, D_0, D_1) \perp Z|X = x^*$ and $\mathbb{P}(D = 1|Z = 0, X = x^*) \neq \mathbb{P}(D = 1|Z = 1, X = x^*)$.

In some contexts, the instrument can be strongly exogenous in the following sense:

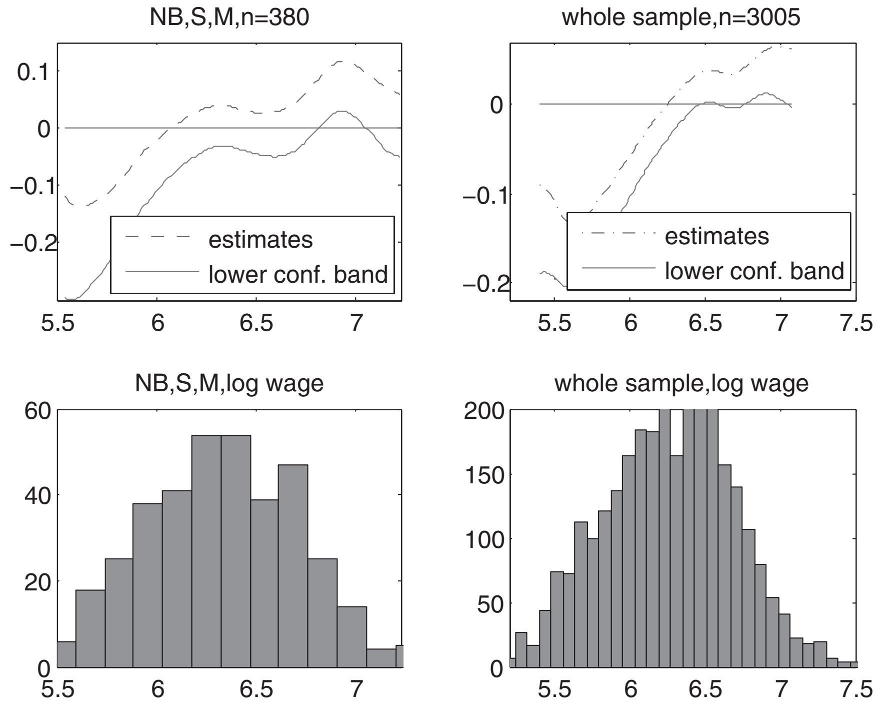
[9] We dropped 608 observations with missing wages.

TABLE 10.—TESTING RESULTS FOR THE COLLEGE PROXIMITY INSTRUMENT

	NB, NS, NM	NB, NS, M	NB, S, NM	NB, S, M	B, S, NM	B, S, M	All
Subgroup ID	1	2	3	4	5	6	
Observations	429	1,191	307	380	314	246	3,005
5%	NR	NR	NR	R	NR	NR	R
1%	NR	NR	NR	R	NR	NR	R
0.5%	NR	NR	NR	R	NR	NR	R

R: rejection; NR: no rejection.

FIGURE 2.—$\hat{\theta}(\cdot, 0)$ AND $\hat{\theta}(\cdot, 0) - s(\cdot, 0) \times \hat{c}_{\hat{V},0.95}$



Assumption 9. $(Y_1, Y_0, D_0, D_1, X) \perp Z$ and $\mathbb{P}(D = 1|Z = 0) \neq \mathbb{P}(D = 1|Z = 1)$.

Our test can be adapted to address all three cases, as summarized by the following corollary:

Corollary 1. *Suppose that assumptions 2 and 7 hold; then for all* $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{cases} \theta^{(1)}(x, y, 1) \equiv \mathbb{E}[c_1(x)D(1 - Z) \\ \quad - c_0(x)DZ|X = x, Y = y] \leq 0 \\ \theta^{(1)}(x, y, 0) \equiv \mathbb{E}[c_0(x)(1 - D)Z \\ \quad - c_1(x)(1 - D)(1 - Z)|X = x, Y = y] \leq 0, \end{cases} \tag{7}$$

where $c_j(x) = \mathbb{P}(Z = j|X = x)$.

If assumptions 2 and 8 hold, then for all $y \in \mathcal{Y}$,

$$\begin{cases} \theta^{(2)}(y, 1) \equiv \mathbb{E}[c_1(x^*)D(1 - Z) \\ \quad - c_0(x^*)DZ|X = x^*, Y = y] \leq 0 \\ \theta^{(2)}(y, 0) \equiv \mathbb{E}[c_0(x^*)1(1 - D)Z \\ \quad - c_1(x^*)(1 - D)(1 - Z)|X = x^*, Y = y] \leq 0. \end{cases} \tag{8}$$

If assumptions 2 and 9 hold, then for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{cases} \theta^{(3)}(x, y, 1) \equiv \mathbb{E}[c_1D(1 - Z) \\ \quad - c_0DZ|X = x, Y = y] \leq 0 \\ \theta^{(3)}(x, y, 0) \equiv \mathbb{E}[c_01(1 - D)Z \\ \quad - c_1(1 - D)(1 - Z)|X = x, Y = y] \leq 0. \end{cases} \tag{9}$$

Proof. See section A.2 in the appendix.

The key difference between equations (7) and (9) is whether the preestimated parameter c_j depends on covariates X. The null hypothesis $H_0^{(k)}$ regarding bounding functions $\theta^{(k)}$ can be defined as

$$H_0^{(k)} : \theta_0^{(k)} \equiv \sup_{(x,y,j) \in \mathcal{X} \times \mathcal{Y} \times \{0,1\}} \theta^{(k)}(x, y, j) \leq 0$$

for $k = 1, 3$, respectively, and

$$H_0^{(2)} : \theta_0^{(2)} \equiv \sup_{(y,j) \in \mathcal{Y} \times \{0,1\}} \theta^{(3)}(y, j) \leq 0.$$

In all three cases, our method is applicable because the estimation rate for $c_j(\cdot)$ or $c_j(x^*)$ is faster than the rate of the bounding functions.

VII. Conclusion

In this paper we provide a reformulation of the testable implications of the key identifying assumptions, LI and LM, of the local average treatment effect, which was first tested by Kitagawa (2008, 2015), with its characterization tracing back to Balke and Pearl (1997) and Heckman and Vytlacil (2005). We show that the testable implications can be written as a set of conditional moment inequality restrictions, which can be tested in the intersection bounds framework of Chernozhukov, Lee, and Rosen (2013) and implemented using the Stata package provided by Chernozhukov et al. (2015). We apply the reformulated testing procedure to the same-sex instrument, the draft eligibility instrument, and the college proximity instrument, respectively. We found that the joint assumption of LI and LM is rejected for the college proximity instrument over some subgroups.

REFERENCES

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

Andrews, D. W. K., and X. Shi, “Inference Based on Conditional Moment Inequalities,” *Econometrica* 81 (2013), 609–666.

Angrist, J. D., “The Draft Lottery and Voluntary Enlistment in the Vietnam Era,” *Journal of the American Statistical Association* 86 (1991), 584–595.

Angrist, J. D., and W. N. Evans, “Children and Their Parents’ Labor Supply: Evidence from Exogenous Variation in Family Size,” *American Economic Review* 88 (1998), 450–477.

Balke, A., and J. Pearl, “Bounds on Treatment Effects from Studies with Imperfect Compliance,” *Journal of the American Statistical Association* 92 (1997), 1171–1176.

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil, “Treatment Effect Bounds: An Application to Swan–Ganz Catheterization,” *Journal of Econometrics* 168 (2012), 223–243.

Card, D., “Using Geographical Variation in College Proximity to Estimate the Returns to Schooling,” NBER working paper 4483 (1993).

Chaisemartin, C. D., “Defying the LATE? Identification of Local Treatment Effects When the Instrument Violates Monotonicity,” Warwick University Department of Economics working paper (2013).

Chernozhukov, V., W. Kim, S. Lee, and A. M. Rosen, “Implementing Intersection Bands in Stata,” *Stata Journal* 15 (2015), 21–44.

Chernozhukov, V., S. Lee, and A. M. Rosen, “Intersection Bounds: Estimation and Inference,” *Econometrica* 81 (2013), 667–737.

Chesher, A., A. M. Rosen, and K. Smolinski, “An Instrumental Variable Model of Multiple Discrete Choice,” *Quantitative Economics* 4 (2013), 157–196.

Dawid, A. P., “Causal Inference without Counterfactuals,” *Journal of the American Statistical Association* 95 (2000), 407–424.

Fan, J., and I. Gijbels, *Local Polynomial Modelling and Its Applications* (Boca Raton, FL: CRC Press, 1996).

Galichon, A., and M. Henry, “Inference in Incomplete Models,” working paper. https://ssrn.com/abstract=886907 (2006).

——— “Set Identification in Models with Multiple Equlibria,” this REVIEW 78 (2011), 1264–1298.

Heckman, J. J., and E. Vytlacil, “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica* 73 (2005), 669–738.

Holm, S., “A Simple Sequentially Rejective Multiple Test Procedure,” *Scandinavian Journal of Statistics* 6 (1979), 65–70.

Huber, M., and G. Mellace, “Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints,” *University of St. Gallen, Department of Economics* discussion paper (2013).

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

Kitagawa, T., “A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model,” Brown University Department of Economics working paper (2008).

——— “A Test for Instrument Validity,” *Econometrica* 83 (2015), 2043–2063.

Machado, C., A. Shaikh, and E. Vytlacil, “Instrumental Variables and the Sign of the Average Treatment Effect,” discussion paper (2013).

Rubin, D. B., “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology* 66 (1974), 688.

APPENDIX

Proofs

Proof of Proposition 1

First, note that $\hat{c}_0$ does not depend on y and $\sup_{y\in\mathcal{Y}} |\tilde{m}(y)| < \infty$ with probability 1, and then it follows that

$$\sup_{y\in\mathcal{Y}} |\hat{m}(y) - \tilde{m}(y)| = O_p\left(\frac{1}{\sqrt{n}}\right),$$

where $c_0 = \mathbb{P}(Z = 0)$, $m(y) = \mathbb{E}[c_0 DZ|Y = y]$, $\tilde{m}(y)$ be the infeasible local linear estimator, which takes c_0 as known, and $\hat{m}(y)$ be the feasible local linear estimator of $m(y)$ in which c_0 is replaced by its frequency count $\hat{c}_0$.

Given the above argument, it is sufficient to treat c_0 and c_1 as if they were known. Recall that

$$\begin{cases} \theta(y,1) \equiv \mathbb{E}[c_1 D(1-Z) - c_0 DZ|Y = y] \le 0 \\ \theta(y,0) \equiv \mathbb{E}[c_0(1-D)Z - c_1(1-D)(1-Z)|Y = y] \le 0 \end{cases}. \tag{10}$$

Let $L_i^1 = c_1 D_i(1-Z_i) - c_0 D_i Z_i$ and $L_i^0 = c_0(1-D_i)Z_i - c_1(1-D_i)(1-Z_i)$. Let $U(W_i,1) = L_i^1 - \theta(Y_i,1)$, $U(W_i,0) = L_i^0 - \theta(Y_i,0)$, $\hat{U}(W_i,1) = L_i^1 - \hat{\theta}(Y_i,1)$ and $\hat{U}(W_i,0) = L_i^0 - \hat{\theta}(Y_i,0)$. Define function $g_v(U,Y)$ as

$$g_{(y,k)}(U,Y) = \frac{U(W,k)}{\sqrt{h}f(y)} K\left(\frac{Y-y}{h}\right).$$

$\hat{g}_v$ is defined similar to g_v with, U and f being replaced by $\hat{U}$ and $\hat{f}$, respectively.

We verify condition NK of CLR and then apply CLR theorem 6. To do so, we first verify that conditions i to vi in CLR appendix F hold in our context, which implies condition NK. We provide these conditions in our notation below:

i. $\theta(y,1)$ and $\theta(y,0)$ are $p+1$ times continuously differentiable with respect to $y \in \mathcal{Y}$, where $\mathcal{Y}$ is convex. *Verify*: $\mathcal{Y}$ being convex is stated in assumption 4. In our context $p = 1$; therefore, we need to verify that $\theta(y,1)$ is twice continuously differentiable. Recall that $\theta(y,1) = \mathbb{E}[L^1|Y = y]$ and L^1 is discrete. Let s be a generic realization of L^1; then $\theta(y,1) = \sum_s s\mathbb{P}(L^1 = s|Y = y)$. So it is sufficient to verify that $\mathbb{P}(L^1 = s|Y = y)$ is twice continuously differentiable with respect to y,

$$\begin{aligned} \mathbb{P}(L^1 = s|Y = y) &= \lim_{\epsilon\to 0} \frac{\mathbb{P}(L^1 = s, y-\epsilon \le Y \le y+\epsilon)}{\mathbb{P}(y-\epsilon \le Y \le y+\epsilon)} \\ &= \lim_{\epsilon\to 0} \frac{\mathbb{P}(y-\epsilon \le Y \le y+\epsilon|L^1 = s)\mathbb{P}(L^1 = s)}{\mathbb{P}(y-\epsilon \le Y \le y+\epsilon)} \\ &= \frac{f(y|L^1 = s)\mathbb{P}(L^1 = s)}{f(y)}, \end{aligned}$$

which is twice continuously differentiable by assumption 4.

ii. The probability density function f of Y_i is bounded above and below from 0 with a continuous derivative on $\mathcal{Y}$. *Verify*: This condition holds by assumption 4.
iii. $U(W_i, 1)$ and $U(W_i, 0)$ are bounded random variables. *Verify*: $U(W_i, k)$ is bounded because Y, D, and Z are bounded.
iv. For each $k \in \{0, 1\}$, the conditional on Y_i density of $U(W_i, k)$ exists and is uniformly bounded from above and below or, more generally, condition R in appendix G (of CLR) holds. *Verify*: The (unconditional) density of $U(W, k)$ exists (with respect to Lebesgue measure). This is because we can write

$$\mathbb{P}(U(W,1) \leq u) = \mathbb{P}(L^1 - \theta(Y,1) \leq u) = \sum_s \mathbb{P}(\theta(Y,1) \geq s - u | L^1 = s)\mathbb{P}(L^1 = s).$$

Since the density of Y given L^1 exists and $\theta(Y, 1)$ is continuously differentiable, we know the conditional density $f_{\theta(1)}$ of $\theta(Y, 1)$ given L^1 exists as long as $\theta(\cdot, 1)$ is a nontrivial measurable function. Taking the derivative with respect to u yields the marginal density of $U(W,1)$;

$$f_{U(W,1)}(u) = \sum_s f_{\theta(1)}(s-u)\mathbb{P}(L^1 = s).$$

Also note that the conditional distribution of $U(W, 1)$ given $Y = y$ is discrete because L^1 is discrete and therefore the condition iv trivially holds for the conditional density of $U(W, k)$ given $Y = y$ (with respect to counting measure). Indeed, our case is analogous to CLR, example B, in that the random variable under the expectation operation is discrete.
v. $K(\cdot)$ has support on $[-1, 1]$ and is twice continuously differentiable: $\int uK(u)du = 0$ and $\int K(u)du = 1$. *Verify*: Condition v is the requirement on the choice of kernel function and is satisfied by many popular kernels (e.g., Epanechnikov kernel). It holds by assumption 5.
vi. $h \to 0$, $nh^{d+|\mathcal{J}|+1} \to \infty$, $nh^{d+2(p+2)} \to 0$, and $\sqrt{n^{-1}h^{-2d}} \to 0$ at polynomial rates in n. *Verify*: Note in our case $|\mathcal{J}| = 2$, $d = 1$ and $p = 1$; therefore, condition vi holds by assumption 6.

CLR show that CLR appendix conditions i to vi imply condition NK(i). Condition NK(ii) holds for the standard nonparametric estimation methods. Then we conclude that parts 1 and 3 of proposition 1 hold by CLR theorem 6, a–i and iii, respectively; part 2 holds by CLR theorem 6 b–i,iii because the contact set $V_0 = \mathcal{V}$; therefore, CLR condition V and equation 4.6 hold with $\rho_n = 1$, $c_n = \infty$.

Proof of Corollary 1

We first verify equation (7). Under assumption 7, the first restriction, equation (1), becomes

$$\mathbb{P}(Y \in A, D = 1|Z = 0, X = x) \leq \mathbb{P}(Y \in A, D = 1|Z = 1, X = x), \quad \forall x \in \mathcal{X},$$

which is equivalent to

$$\mathbb{E}[\mathbf{1}_{Y\in A}\{D(1-Z)c_0(x) - DZc_1(x)\}|X = x] \leq 0, \quad \forall x \in \mathcal{X}.$$

The results hold since the above inequality holds for all $A \in \mathcal{B}_{\mathcal{Y}}$ and, consequently, for the class of cubes. To verify equation 9, simply note that under assumption 9, we have for all $B \in \mathcal{B}_{\mathcal{Y}\times\mathcal{X}}$,

$$\mathbb{P}((Y, X) \in B, D = 1|Z = 0) \leq \mathbb{P}((Y, X) \in B, D = 1|Z = 1).$$

The result follows.