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 and Angrist, 1994), namely the valid instrument assumption (LI) and the treatment monotonicity assumption (LM). We reformulate the testable implications of LI and LM 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, Kim, Lee, and Rosen (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).

Keywords: LATE, hypothesis testing, intersection bounds, Testable implications.

This version: Thursday 19th November, 2015

First version: December 7th, 2013

* 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 and Paul Schrimpf and two anonymous referees for valuable comments. We have greatly benefited from insightful comments received 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. Yuanyuan Wan thanks the support from SSHRC Insight Development Grant #430130275. Ismael Mourifié thanks the support from Connaught Fund. All errors are ours.

† Corresponding author. Department of Economics, University of Toronto, 150 St. George Street, Toronto ON M5S 3G7, Canada. ismael.mourifie@utoronto.ca

‡ Department of Economics, University of Toronto. yuanyuan.wan@utoronto.ca.

1. INTRODUCTION

The instrumental variable (IV) method is one of the most-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 but dependent on the observed treatment; in the meantime, it must have no effect on the observed outcome beyond its effect on the observed treatment. Imbens and Angrist (1994, IA1994 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, IA1994 introduced LM (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 and LM hold, IA1994 showed that the IV estimand identifies the ATE for the subpopulation of compliers, namely, the LATE.

Although the results of IA1994 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 IA1994. 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, Kim, Lee, and Rosen (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 (2015)'s test has non-trivial 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 existence of 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 (2015)'s 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 zero. Third, the test can accommodate continuous covariates within the same framework. Indeed, as we further elaborate in Section 6, 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 (2013)'s approach to transform the testable implication to unconditional moment restrictions. Lastly, as we mentioned earlier, our testing procedure can be easily implemented using the Stata package provided by Chernozhukov, Kim, Lee, and Rosen (2015). There are other papers discuss testing issues under different setup. Machado, Shaikh, and Vytlacil (2013) proposed tests for LM and/or outcome monotonicity (in treatment) in a binary treatment, 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 non-black men who lived in the metro 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 2 presents the analytical framework. In Section 3, we revisit the testable implications of the LATE assumptions, followed by Section 4, which presents our testing procedure. We discuss empirical applications in Section 5. The last section extends our analysis to the case with additional covariates.

2. Analytical Framework

We adopt the potential outcome model of Rubin (1974). Let $Y = Y_1 D + Y_0(1-D)$, 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 the sake of simplicity, we assume $Z \in \mathcal{Z} = \{0,1\}$, but our analysis can be extended to allow for multi-valued 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 $D = D_1 Z + D_0(1-Z)$.

The two well-known identification assumptions for LATE as introduced by IA1994 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.

3. 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 the 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.

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}

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

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

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

TABLE 2. Observed subgroups and unobserved subpopulations

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

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 soon 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 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 (1) and (2) hold. (ii) If (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 infinite 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

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

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

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

and

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

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 $\forall y \in \mathcal{Y}$

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

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

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

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.

4. Testing Procedures

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

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

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 pre-specified 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 the purpose of 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, Kim, Lee, and Rosen (2015) for the full set of options. □

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 zero 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 (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] The following Assumptions 5 and 6 are conditions on the choice of kernel and bandwidth, respectively.

[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, the 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 at root-n rate, has limiting normal distribution with estimable covariance matrix. To implement, one can then follow the discussions in CLR (page 709).

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), which verifies the consistency and validity of the proposed testing procedure.

Proposition 1. *Suppose that Assumptions 3 to 6 are satisfied, then (1) under H_0, $\mathbb{P}(\hat{\theta}_\alpha > 0) \leq \alpha + o(1)$; (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 (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 Appendix A.1

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. Lastly, 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.

5. 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).

5.1. 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 the fertility. Angrist and Evans (1998) proposed to use 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 contains information on age, gender, race, education, labor income, and the number of children. We consider women with at least two children, between 21 and 50 years old, and with positive wage.[3] 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.

TABLE 3. Summary Statistics

	Total	D=1	Z=1
Observations	403,011	119,221	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 (Non White)	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 and standard deviation (in the parentheses)

TABLE 4. Subgroups

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

We divided $403,011$ observations into 24 subgroups according to race (white or non-white), 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 size $25,000$ for subgroups whose sizes are larger than this number.[4]

[3]There are 35.06% of observations with missing wage. We also conducted a test on the missing wage subsample. The null hypothesis is not rejected either.

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

Throughout this section, we use the default choices of bandwidth and kernel functions recommended in CLR and Chernozhukov, Kim, Lee, and Rosen (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), which is a suitable framework to consider (see also an empirical implementation in Bhattacharya, Shaikh, and Vytlacil, 2012). The testing results show that the smallest p-values among all 24 groups is greater than 10%.[5] Hence we are able to conclude that the multiple testing procedures rejects no null hypothesis at 10% level.[6] 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 using the "parametric regression" method,[7] 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.

TABLE 5. Application results: Parametric (clrtest)

AE1998 same sex			Angrist1991 lottery			Card1993 proximity		
10%	5%	1%	10%	5%	1%	10%	5%	1%
NR	NR	NR	R	R	R	R	R	R

"R" stands for rejection and "NR" stands for no rejection.

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

[6]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.

[7]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 to allow for multiple conditioning variables.

5.2. **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 enrollment 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 birthdate 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 $D_1 \geq D_0$.

TABLE 6. Summary Statistics of SIPP Data from Angrist (1991)

	Total	Draft Eligible (Z=1)	Veteran (D=1)
Observations	3027	1379	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 (Non White)	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 and standard deviation (in the parentheses)

We used a sample of 3,071 individuals from the 1984 Survey of Income and Program Participation (SIPP)[8]. The sample was divided into 6 different groups according to race (white or non-white) and their educations levels (lower than HS, HS, or higher than HS), where HS stands for high school graduation. We then performed our test using the local method for each group. Again, we compute the maximum in the test statistics over the interval $[Q_{2.5\%}, Q_{97.5\%}]$.

TABLE 7. Lottery local method (clrtest)

	W,<HS	W,=HS	W,>HS	NW,<HS	NW,=HS	NW,>HS
Subgroup ID.	1	2	3	4	5	6
Obs.	317	865	1478	56	129	171
10%	NR	NR	NR	NR	R	NR
5%	NR	NR	NR	NR	R	NR
1%	NR	NR	NR	NR	R	NR

"R" stands for rejection and "NR" stands for no rejection.

[8]The data is available from Angrist's website. 3,071 is the number of individuals after removing all entries with missing information.

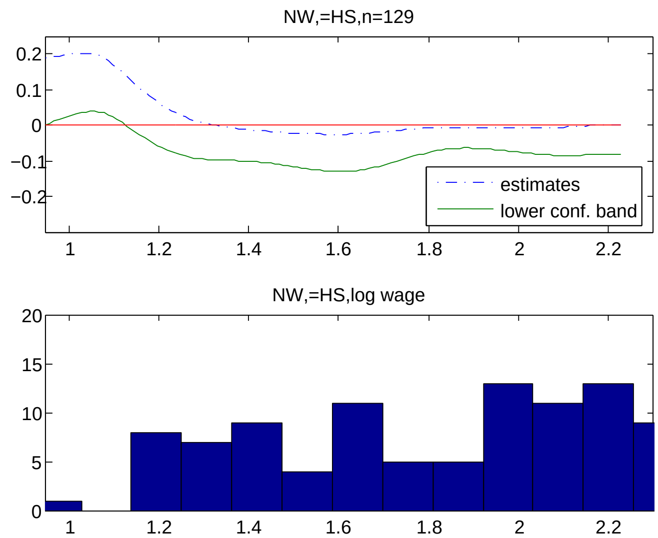

FIGURE 1. $\hat{\theta}(\cdot,1)$ and $\hat{\theta}(\cdot,1) - s(\cdot,1) \times \hat{c}_{\hat{V},0.95}$

Testing results for individual groups reported in Table 7. Note that the null hypothesis ($H_0^{(5)}$) is rejected at subgroup 5 of non-white person 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 and/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 hypothesizes with FWER controlled at 10%.[9] Kitagawa (2015) obtained the same result without conditioning on subgroups.

5.3. **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 4-year college in the local labor market where the individual was born, and $D = 1$ denotes the individual has at least 16 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 14-24 and continued with a follow-up survey until 1981. Some summary statistics are

[9]The testing procedure with parametric regression method, however, rejects the null hypothesis at all three levels (see Table 5).

TABLE 8. Summary Statistics of NLSYM Sample

	Total	$D = 1$	$Z = 1$
Observations	3005	2048	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 (eduction $\geq$ 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 and standard deviation (in the parenthesis)

reported in Table 8.[10] We considered three binary control variables: lived in southern states in 1966, lived in metro area in 1966, and being black. Table 9 reports the corresponding subgroup sizes.

TABLE 9. Subgroup sizes of Card (1993)

	Non-Black (NB)	Black (B)
Non-Southern (NS) & Non-Metro (NM)	429	5
Non-Southern (NS) & Metro (M)	1191	138
Southern (S) & Non-Metro (NM)	307	314
Southern (S) & Metro (M)	380	246

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

TABLE 10. College proximity, local method (clrtest)

	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	
Obs.	429	1191	307	380	314	246	3005
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" stands for rejection and "NR" stands for no rejection.

We conduct the test on six subgroups. We exclude the subgroup NS/NM/B because of small sample size; we also exclude subgroup NS/M/B because the high frequency of $Z = 1$ (92%). Note that the null hypothesis $H_0^{(4)}$ is rejected in subgroup 4 of Non-black men who lived in the metro area of the southern states as well as for the whole sample at 0.5% level. No rejection happens with

[10]We dropped 608 observations with missing wages.

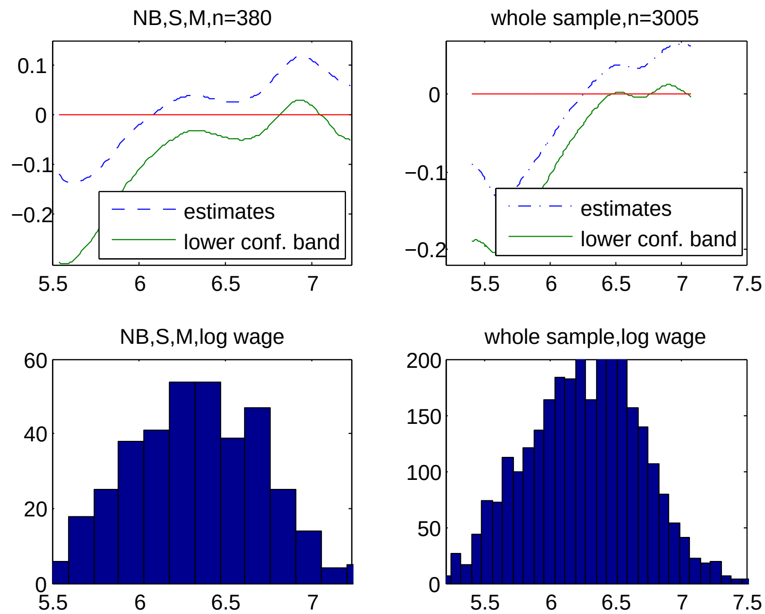

FIGURE 2. $\hat{\theta}(\cdot, 0)$ and $\hat{\theta}(\cdot, 0) - s(\cdot, 0) \times \hat{c}_{\hat{V},0.95}$

other subgroups even at 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.

Now it will be interesting to know on which subsets of $\mathcal{Y}$ 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 the subgroup 4 and the whole sample, respectively. It is quite fascinating to note that θ_0 is in general increasing in Y, and the rejection takes place on higher income subpopulations, e.g. for subpopulations whose observed log wage is around 7. Note 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 could be "sign reversal". Thereby, although the "college proximity" seems to be a good instrument, researchers must be aware that this instrument would not be a good one to use when the treatment effect is heterogenous.

6. EXTENSIONS

In this section we discuss three different ways of incorporating covariates X into the testing procedure. As we will demonstrate below, all three cases 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 holds 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 in knowing for each observed group that 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.

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 adapt to address all three cases, as summarized by the following Corollary.

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

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

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

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

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

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

$$\begin{cases} \theta^{(3)}(x,y,1) \equiv \mathbb{E}[c_1 D(1-Z) - c_0 DZ | X = x, Y = y] \leq 0 \\ \theta^{(3)}(x,y,0) \equiv \mathbb{E}[c_0 1(1-D)Z - c_1(1-D)(1-Z) | X = x, Y = y] \leq 0 \end{cases}. \tag{9}$$

Proof. See Appendix A.2.

The key difference between (7) and (9) is whether the pre-estimated parameter c_j depends on covariates X. The null hypothesis $H_0^{(k)}$ regarding bounding functions $\theta^{(k)}$ 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.

7. 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, Kim, Lee, and Rosen (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 "college proximity" instrument over some subgroups.

References

Abadie, A. (2003): "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, 113(2), 231–263.

ANDREWS, D. W. K., AND X. SHI (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81(2), 609–666.

ANGRIST, J. D. (1991): "The draft lottery and voluntary enlistment in the Vietnam era," *Journal of the American statistical Association*, 86(415), 584–595.

ANGRIST, J. D., AND W. N. EVANS (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 88(3), 450–77.

BALKE, A., AND J. PEARL (1997): "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 92(439), 1171–1176.

BHATTACHARYA, J., A. M. SHAIKH, AND E. VYTLACIL (2012): "Treatment effect bounds: An application to Swan–Ganz catheterization," *Journal of Econometrics*, 168(2), 223–243.

CARD, D. (1993): "Using Geographical Variation in College Proximity to Estimate the Returns to Schooling," Discussion paper.

CHAISEMARTIN, C. D. (2013): "Defying the LATE? Identification of Local Treatment Effects When the Instrument Violates Monotonicity," Working paper.

CHERNOZHUKOV, V., W. KIM, S. LEE, AND A. M. ROSEN (2015): *Stata Journal,*15(1), 21–44.

CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.

CHESHER, A., A. M. ROSEN, AND K. SMOLINSKI (2013): "An instrumental variable model of multiple discrete choice," *Quantitative Economics*, 4(2), 157–196.

DAWID, A. P. (2000): "Causal Inference without Counterfactuals," *Journal of the American Statistical Association*, 95(450), 407–424.

FAN, J., AND I. GIJBELS (1996): *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, vol. 66. CRC Press.

GALICHON, A., AND M. HENRY (2006): "Inference in Incomplete Models," *Working Paper*.

——— (2011): "Set identification in models with multiple equilibria," *The Review of Economic Studies*, 78(4), 1264–1298.

HECKMAN, J. J., AND E. VYTLACIL (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation1," *Econometrica*, 73(3), 669–738.

HOLM, S. (1979): "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70.

HUBER, M., AND G. MELLACE (2013): "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," Discussion paper.

IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

KITAGAWA, T. (2008): "A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model," Working paper.

——— (2015): "A Test for Instrument Validity," *Econometrica*, 83(5), 2043–2063.

MACHADO, C., A. SHAIKH, AND E. VYTLACIL (2013): "Instrumental Variables and the Sign of the Average Treatment Effect," Discussion paper.

RUBIN, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of educational Psychology*, 66(5), 688.

APPENDIX A. PROOFS

A.1. **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 one, 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 similarly as g_v with U and f being replaced by $\hat{U}$ and $\hat{f}$, respectively.

We verify the Condition NK of CLR and then apply CLR-Theorem 6. To do so, we first verify that Conditions (i)-(vi) in CLR Appendix F holds in our context, which implies Condition NK. We recite these conditions in our notation below.

Condition (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 $\mathbb{P}(L^1 = s|Y=y)$ 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.

Condition (ii) The probability density function f of Y_i is bounded above and below from zero with continuous derivative on $\mathcal{Y}$.

Verify: this condition holds by Assumption 4.

Condition (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.

Condition (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

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

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 non-trivial measurable function. Take 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 conditional (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 to be taken expectation is discrete.

Condition (v) $K(\cdot)$ has support on $[-1,1]$, 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.

Condition (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 Condition (i)-(vi) imply Condition NK(i). Condition NK(ii) holds for the standard nonparametric estimation methods. Then we conclude that Part (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 $\rho_n = 1$, $c_n = \infty$.

A.2. **Proof of Corollary 1.** We first verify Equation (7). Under Assumption 7, the first restriction (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 $B \in \mathcal{B}_{\mathcal{Y}\times\mathcal{X}}$, there is

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

The result follows.