TESTING LATE ASSUMPTIONS

ISMAEL MOURIFIÉ[†] AND YUANYUAN WAN[‡]

ABSTRACT. In this paper, we discuss the key conditions for the identification and estimation of the local average treatment effect (LATE, Imbens and Angrist, 1994): the valid instrument assumption (LI) and the monotonicity assumption (LM). We show that the joint assumptions of LI and LM have a testable implication that can be summarized by a sign restriction defined by a set of intersection bounds. We propose an easy-to-implement testing procedure that can be analyzed in the framework of Chernozhukov, Lee, and Rosen (2013) and implemented using the Stata package of Chernozhukov, Kim, Lee, and Rosen (2013). 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, conditionally more compliers.

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 must be independent of all potential outcomes and potential treatments but be dependent on the observed treatment; in the meantime, it must have no effect on the observed outcome beyond its effect on the observed treatment. In the presence of a valid instrument, Holland (1988) shows that the causal effect can be consistently estimated using the Wald estimand when the treatment effect is homogeneous that is, the effect of the treatment is constant across individuals. Imbens and Angrist (1994, IA1994 hereafter) showed, however, that a

valid instrument itself does not ensure the "no sign reversal property" of the Wald estimand when the treatment effect is heterogeneous, meaning that when the effect of the treatment is non-constant across individuals, the Wald estimand can be zero, positive, or negative, even if the causal effect of the treatment on the outcome is strictly positive for every individual. To deal with this issue, IA1994 introduced the treatment monotonicity assumption (LM, also known as the "no defiers" assumption), which assumes that the instrument affects the treatment decision in the same direction for every individual. Whenever both instrument validity (LI) and LM hold, IA1994 showed that the Wald estimand is no longer sign reversal and can be used to consistently estimate the average treatment effect (ATE) for the subpopulation of compliers, namely, the local average treatment effect (LATE).

Although the results of IA1994 have being widely influential in the applied economics literature, there are 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 because LM itself is not testable. 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) proposed a specification test for testing the joint assumptions of LM and LI in the binary treatment case. 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.

In this paper, we revisit the existing discussions on testing joint validity of LM and LI and propose a new and easy-to-implement test. In particular, we show that LI and LM imply a set of conditional moment inequalities; and show that this set of conditional moment inequalities is a sharp characterization of LM and LI, in the sense that, whenever it holds, there always exists another observationally equivalent model in which LI and LM hold. The novelty and a nice feature of these restrictions 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. The test we propose can be analyzed under the intersection bounds framework of Chernozhukov, Lee, and Rosen (2013). A Stata package provided by Chernozhukov, Kim, Lee, and Rosen (2013) is readily available for empirical researchers to use.

2

As an extension to LM, we discuss an alternative testable condition of conditionally more compliers (CMC), which is also discussed in Chaisemartin (2013). CMC is weaker than LM because it does not rules out the existence of defiers, instead, it only requires that the proportion of defiers is no greater than the proportion of the compliers in every subpopulations (strata) defined by the potential outcome variable. Chaisemartin (2013) referred to it as "weak more compliers than defiers" (WMC); we name it as conditionally more compliers to emphasize that it is a restriction imposed on subpopulations. In addition to the discussion in Chaisemartin (2013), we show that CMC and LI have the same testable implication as LM and LI. We further show that when LI holds, the testable implication of LM is equivalent to CMC. In other words, CMC is testable when LI holds. Therefore, our test can be interpreted as test for "more compliers than defiers" whenever LI holds. We also show that when CMC+LI hold, the Wald estimand is still no sign reversal for a class of potential outcomes models which allow for heterogeneous treatment effect.

There are other weaker notions of monotonicity. Small and Tan (2007) discussed the stochastic monotonicity (SM) condition, which requires the proportion of defiers to be smaller than compliers in every subpopulation defined by the vector of two potential outcomes. Compared with SM, CMC only requires this holds in subpopulations defined by a single potential outcome. SM is therefore weaker than LM but stronger than CMC. Chaisemartin (2013) proposed a "compliers-defiers" (CD) assumption and showed that the IV can still be a relevant estimator under CD. CD also allows for the existence of defiers, but it requires that a subgroup of compliers exists that has the same size and same marginal distribution of both potential outcomes as defiers. Unlike CMC, neither SM nor CD is directly testable even when LI holds.

From the testing perspective, our paper complements but differs from the work of Kitagawa (2008, 2013) in the following aspects. Indeed, the latter papers focus on restrictions of the joint distribution of the outcome and treatment variables conditional on the instrument. For this approach, the number of restrictions (inequalities) needs to be checked equals to the number of all subsets of the support of the observable outcome. The computational burden will be very high when the support contains a large number of values. When the set of observable outcomes is infinite, the problem becomes infinite dimensional. One way to solve this issue is to find the smallest core determining class, [1] which is defined as a suitably lowest cardinality class of sets that are sufficient to characterize all

---

[1] See Galichon and Henry (2006, 2011) for detailed discussions.

the restrictions imposed by the joint LI and LM assumptions on the observables. The smallest core determining class has not been discussed in Kitagawa (2008, 2013). The inferential method they use relies on a set of restrictions that is not guaranteed to exhaust all the restrictions implied by the joint LI and LM assumption and hence leads to more conservative inference than could be achieved and also does not guaranteed the test to be consistent against all fixed alternatives (more discussions in Section 4). On the other hand, we propose, first, a new characterization of the testable implication that avoids the research of a core determining class. Second, our test is consistent against any fixed alternatives. Third, our test is also consistent against a class of local alternatives, whereas Kitagawa (2008, 2013) did not establish such results. Finally, as we mentioned earlier, our testing procedure can be easily implemented using the Stata package provided by Chernozhukov, Kim, Lee, and Rosen (2013).

Our last contribution is empirical. 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 whether the first two children are of the same sex as the instrument. Our test does not reject the testable implication of LI+LM for "drat 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 metro area of southern states (as well as for the whole sample). The rejection mainly takes place among the 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. Section 5 propose the CMC condition. We illustrate our test by simulation examples in Section 6. We discuss empirical applications in Section 7.

2. ANALYTICAL FRAMEWORK

We adopt the potential outcome model of Vytlacil (2002). 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

4

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 had the Z 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). *For all $z \in \{0, 1\}$, $Z \perp (Y_1, Y_0, D_0, D_1)$ 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 $D_1^{-1}(1) \subseteq D_0^{-1}(1)$. 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 derive a set of testable implications of the LATE assumptions (LI and LM) and show that these testable implications are "sharp" characterization of the LATE assumptions.

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.

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}

5

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

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

where the third inequality 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

TABLE 2. Observed subgroups and unobserved subpopulations

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

easily see that LM is equivalent to the nonexistence of defiers (i.e. $\pi_{10} = 0$). LM and LI necessarily imply that for an arbitrary $A \in \mathcal{B}_{\mathcal{Y}}$,

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

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 $A \subseteq \mathcal{B}_{\mathcal{Y}}$ such that either Equation (1) or (2) is violated, we must reject the joint assumptions of LM + LI assumptions. Note that Equations (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 Equations (1) and (2) hold but LM+LI is violated.

Equations (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 Equations (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).*

Proof. See Appendix A.1. □

Theorem 1 is essentially equivalent to, but presented in a different way from, Kitagawa (2013, Proposition 2). The sharpness result shows that Equations (1) and (2) are the most informative observable restrictions we have to assess the validity of the joint LI and LM assumptions.

Kitagawa (2013) developed a Kolmogorov-Smirnov type test based on Equations (1) and (2). 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. It is therefore desirable to find a "sufficient sub-class" of $\mathcal{B}_{\mathcal{Y}}$. Following Galichon and Henry (2006, 2011), we can define the (smallest) core determining class as a low (lowest) cardinality class of sets that are sufficient to characterize all the restrictions imposed by Equations (1) and (2). To the best of our knowledge, the issue of finding the smallest core determining class in a generic setup remains open. Kitagawa (2013) proposed, without providing theoretical justification, to consider the half lines and some sub-intervals. It is not clear if this class is a core determining class; if not, it could neglect some subsets of $\mathcal{Y}$, which could provide more power.

To deal with this important issue, we propose to use an alternative representation.

$$\mathbb{P}(Y \in A, D = 1 | Z = 1) = \frac{\mathbb{P}(D = 1, Z = 1 | Y \in A)}{\mathbb{P}(Z = 1)} \mathbb{P}(Y \in A),$$

inequalities (1) and (2) can be written as

$$\mathbb{P}(D = 1, Z = 0 | Y = y)\mathbb{P}(Z = 1) \leq \mathbb{P}(D = 1, Z = 1 | Y = y)\mathbb{P}(Z = 0), \quad (3)$$

and

$$\mathbb{P}(D = 0, Z = 1 | Y = y)\mathbb{P}(Z = 1) \leq \mathbb{P}(D = 0, Z = 0 | Y = y)\mathbb{P}(Z = 0). \quad (4)$$

for all $y \in \mathcal{Y}$, respectively. The equivalence holds because $\mathbb{P}(D = 1, Z = 1 | Y \in A) = \frac{1}{\mathbb{P}(Y \in A)} \int_A \mathbb{P}(D = 1, Z = 1 | Y = y) d\mathbb{P}(Y \leq y)$. We further define the following pair of functions of y

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

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 a core determining class and can simply apply the existing inference methods in Chernozhukov, Lee, and Rosen (2013), 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 Chernozhukov, Lee, and Rosen (2013), 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.2 for the sake of exposition. The testing procedure can be easily implemented in Stata as follows.

Implementation:

(1) Estimate c_0 and c_1 by $\hat{c}_1 = \frac{1}{n} \sum_{i=1}^{n} Z_i$ and $\hat{c}_0 = 1 - \hat{c}_1$, respectively.

8

(2) Let $L_i^1 = \hat{c}_1 D_i(1 - Z_i) - \hat{c}_0 D_i Z_i$ and $L_i^0 = \hat{c}_0(1 - D_i)Z_i - \hat{c}_1(1 - D_i)(1 - Z_i)$.

(3) Implement the CLRtest command with two conditional moment inequalities. Specify L_i^1 and 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 (2013) for the full set of options. $\square$

We make the following assumptions:

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

Assumption 4. *The density of Y given (D, Z) is bounded away from zero and continuously differentiable over its convex and compact support $\mathcal{Y}$.*

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.*

Assumption 3 is standard. In this paper, we focus on the case where Y is continuous and make Assumption 4. If Y is discrete, the conditional inequalities in Equation (5) can be represented by a finite number of unconditional expectations, and the test can be conducted on each of the support points of Y. Assumptions 5 and 6 are conditions on the choice of kernel and bandwidth. Note that Assumption 6 requires under-smoothing because the conditioning variable Y is univariate.

Proposition 1 is an application of Chernozhukov, Lee, and Rosen (2013, Theorem 6), which verifies the consistency and validity of the proposed testing procedure.[2]

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 there exists a set $\mathcal{Y}_a \subseteq \mathcal{Y}$ with positive Lebesgue measure such that $\sup_{y \in \mathcal{Y}_a, k \in \{0,1\}} \theta(y, k) > \mu_n \sqrt{1/nh}$ for any $\mu_n \xrightarrow{p} \infty$, then $\mathbb{P}(\hat{\theta}_\alpha > 0) \to 1$.*

Proof. See Appendix A.2

---

[2]Chernozhukov, Lee, and Rosen (2013, Theorem 6) holds uniformly over drafting sequences of DGPs. Here we only focus on pointwise results.

We have a few comments. First, our test is a type of sup-tests based on conditional moment inequalities specified in Equation (5) and hence does not require researchers to find a core determining class. Our test is consistent against any fixed alternatives and a class of local alternatives. This result is different from Kitagawa (2013), who did not use the core determining class and only established the consistency against a class of fixed alternatives.[3] Consistency against local alternatives is not established either. Second, Kitagawa (2013) did not discuss how to incorporate continuous covariates into his test. For the test we consider here, 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.

In the following section, we will proceed and address two important questions. (i) What types of restrictions do the Equations (1) and (2) impose on the (unobserved) model primitives, for example, the distributions of the subpopulations of compliers, defiers, always-takers, and never-takers? (ii) Does the Wald estimand still a policy relevant estimand whenever Equations (1) and (2) hold and LM is violated?

5. CONDITIONALLY MORE COMPLIERS

In many empirical studies, the LI assumption seems to be reasonable. See Korn, Baumrind, et al. (1998); Abadie, Angrist, and Imbens (2002); Kling, Liebman, and Katz (2007); and Hahn, Todd, and Van der Klaauw (2001); among others, for discussions of models where LI is plausible. On the other hand, several concerns about the validity of LM have been mentioned in the literature, for instance, in Dawid (2000); Heckman and Vytlacil (2005); Small and Tan (2007). In this section, we propose a weaker condition, namely conditionally more compliers (CMC) as an extension of LM. Unlike LM, which imposes on the model to have no defiers, the CMC assumption only restricts the proportion of compliers to be bigger or equal to the proportion of defiers, for all subpopulations (strata) characterized by a single potential outcome Y_d. This condition imposes a restriction on the aggregate behavior of individuals relative to their reaction to the treatment when the instrument Z is externally set to different values 0 or 1, while the restriction imposed by the LM assumption is on

---

[3]For example, Kitagawa (2013) proposed to consider the class of closed intervals $\{[y, y + h]\}$, where $y \in \mathcal{Y}$ and h is the interval length satisfying $h \geq h_{min} > 0$. h_{min} is a fixed positive constant chosen by researchers and does not converge to zero as sample size increases. If the H_0 is violated over an interval whose length is smaller than $h_{min}/2$, then Kitagawa (2013)'s test may not be able to detect such violations based on this interval class. In our test, the bandwidth $h \to 0$, such alternatives will be rejected with probability approaching one.

the behavior of every individual. Interestingly, CMC is equivalent to Equations (1) and (2) when LI holds.

Proposition 2 (Conditionally More Compliers). *Under Assumption 1, Equations (1) and (2) are equivalent to*

(a) $\mathbb{P}(D_0 = 1|Y_1 = y) \leq \mathbb{P}(D_1 = 1|Y_1 = y)$ *for all y,*
(b) $\mathbb{P}(D_1 = 0|Y_0 = y) \leq \mathbb{P}(D_0 = 0|Y_0 = y)$ *for all y,*

or equivalently to

(a') $\pi_{10|Y_1=y_1} \leq \pi_{01|Y_1=y_1}$ *for all y_1,*
(b') $\pi_{10|Y_0=y_0} \leq \pi_{01|Y_0=y_0}$ *for all y_0,*

where $\pi_{10|Y_d=y_d}$ $\left(\pi_{01|Y_d=y_d}\right)$ denotes the proportion of defiers (compliers) within the strata of individuals with potential outcomes $Y_d = y_d$ for $d = 0, 1$.

Proof. See Appendix A.3.

We name the implied restrictions (a-b) or (a',b') in Proposition 2 as Conditionally More Compliers (CMC). Chaisemartin (2013) referred to it as "weak more compliers than defiers" (WMC); we name it as CMC to emphasize that it is a restriction imposed on subpopulations. Note that LM implies CMC since $\pi_{10|Y_d=y} = 0$ for all y.

Corollary 1. *(LM and LI) and (CMC and LI) have the same testable implication which is characterized by Equations (1) and (2).*

The proof of Corollary 1 is straightforward and hence omitted. If we reject the null hypothesis $H_0 : \theta_0 \equiv \sup_{v \in \mathcal{V}} \theta(v) \leq 0$, then what we actually reject is CMC+LI, which implies the rejection of LM+LI; on the other hand, if we do not reject H_0 and if LI holds, then we shall conclude that we can not reject CMC. Note that LM may not hold even we do not reject CMC. Another difference is that CMC is indeed testable when LI holds.

5.1. **Discussion on the plausibility of the CMC assumption.** To fix the idea, let us discuss an instrument that is widely used in labor economics: the same-sex siblings composition. This instrument was first used in 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)

11

proposed to use the sibling-sex composition to construct the IV estimator of the effect of childbearing on the labor supply. Because sex mix is virtually randomly assigned, a dummy for whether the sex of the second child matches the sex of the first child is likely to satisfy the LI assumption. However, the validity of LM is unclear. The suggestive argument proposed by Angrist and Evans (1998) was based on the empirical observation that the proportion of families with either two boys or two girls that have a third birth in the United States was lower to the proportions of families with one boy and one girl have a third child. This empirical observation just shows that $\mathbb{P}(Y \in \mathcal{Y}, D = 1|Z = 0) \leq \mathbb{P}(Y \in \mathcal{Y}, D = 1|Z = 1)$, where Y is the women's labor income, and D is a dummy for having had a third child. However, this latter inequality is not sufficient to ensure CMC, which requires that for every subpopulation of parents with specific outcome there are more parents who have mixed sibling-sex composition preferences than parent who have same sibling-sex composition preferences. Our contribution is to show that for such empirical applications, one can actually test the validity of CMC by visiting all the possible subsets A of $\mathcal{B}_{\mathcal{Y}}$.

Compared with LM, CMC allows for more heterogeneity in the population than the deterministic LM. Indeed, in that context, LM imposes that there is no individual who has strict preference for the same sibling-sex composition, while CMC suggests that such an individual could exist. CMC imposes that for every subpopulation of parents with specific outcome there are more parents who have mixed sibling-sex composition preferences than parent who have same sibling-sex composition preferences.

There are other weaker notion of monotonicity than LM. Small and Tan (2007) introduced the stochastic monotonicity condition (SM) as a relaxation of LM. The SM assumption restricts the model to have more compliers than defiers for every strata defined by (Y_1, Y_0), while CMC imposes this restriction separately on the strata defined by either Y_1 or Y_0. From this perspective, CMC weakens the SM assumption. Chaisemartin (2013) showed that for the purpose of identifying LATE, the LM assumption can be replaced by a weaker one called "compliers-defiers" (CD), which requires that a subgroup of compliers have the same size and the same distribution of potential outcomes as defiers. Both SM and CD provide interesting insight on the identification of LATE. Unfortunately, like LM, SM, and CD are not testable even when LI holds.

5.2. **Is the Wald estimand no sign reversal under CMC and LI?.** An interesting question is what can we say about the Wald (IV) estimand under heterogeneous treatment models if we do

not reject Equations (1) and (2)? We know that under LM+LI assumption, the (IV) estimand $\beta_{IV} \equiv \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]}$ respects the no sign reversal property[4]. Indeed, under LI, we have

$$\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] = \mathbb{E}[Y_1 - Y_0|c]\pi_{01} - \mathbb{E}[Y_1 - Y_0|def]\pi_{10}.$$

Therefore, in the case of heterogeneous treatment (the individual gain of the treatment varies across individual; i.e. $Y_1 - Y_0 \neq cste$), the (IV) estimand can be zero, positive, or negative, even if the causal effect of D on Y is strictly positive for every individual. However, under the LM assumption, we have $\beta_{IV} = \mathbb{E}[Y_1 - Y_0|c]$ and then no possibility to violate the no sign reversal property.

Since CMC does not rule out the existence of defiers, it is not clear that the (IV) estimand still respect the no sign reversal property whenever CMC holds. In the following, we show that there exists a wide range of potential outcome models for which the (IV) estimand still respect the no sign reversal property when CMC holds even in the presence of defiers. Let us impose the following restriction on the potential outcomes.

Assumption 7. $Y_d = G(d, U)$ where $G(d, u)$ is strictly monotone in u for $d \in \{0, 1\}$.

This assumption imposes three main restrictions on the potential outcome model but encompasses most of the structures used in applied economics so far.

First, we restrict the dimension of the unobservable U to be one. This might be restrictive in some cases as explained in Kasy (2013), but it is usually maintain in many of the research papers considering binary endogenous treatment. Indeed, practitioners use additive separable structure for the potential outcome which imposes one-dimensional unobserved heterogeneity. Recently, the literature on the identification of treatment effect using non-separable potential outcome has been widely developing, but most of them are using model with one-dimensional unobserved heterogeneity. For instance see Chernozhukov and Hansen (2005), Chesher (2005), Jun, Pinkse, and Xu (2011) among many others. The strand of this literature that allows multiple-dimension unobserved heterogeneity considers continuous treatment and impose additional restrictions on the treatment, see for instance, Imbens and Newey (2009) and Kasy (2013).

Second, we assume away the possibility to have treatment-specific unobserved heterogeneity i.e. $Y_d = G(d, U_d)$ with $U_0 \neq U_1$. One important special application of such a model is the Roy model.

---

[4]An estimator satisfies the no sign reversal property if whenever the sign of the treatment effects (+,0 or -) is the same for every subject in the population, then the sign of the estimator converges in probability to the sign of the treatment effects.

However, the pure Roy model imposes particular selection process which is not compatible with the model used here. Indeed, the pure Roy model assume that $D = 1\{Y_1 > Y_0\}$ which does not allow the existence of an instrument. Identification of treatment effect in such model need different types of analysis. A first approach is the one used by Heckman and Honore (1990) which imposes additional restrictions in the model in order to achieve identification. A second approach used in Henry and Mourifie (2014) is a partial identification approach which provides sharp bounds on the average treatment effect (ATE). One way to incorporate an instrument in such a model is to consider the generalized Roy model where the treatment selection takes a more general form i.e. $D = 1\{h(Y_1, Y_0, Z, V) > 0\}$ where V is another source of heterogeneity. However, Heckman and Vytlacil (2005) pointed out that the (IV) estimand does not identify interesting policy effect in such a model.

Third, the "strict" monotonicity would be restrictive. The monotonicity by itself is a standard assumption in such models, but imposing to be "strict" rules out the possibility to have discrete outcome.

Now, let us show that under CMC+LI and Assumption 7, the (IV) estimand still respect the no sign reversal property. Let denote $\Delta(U) \equiv Y_1 - Y_0$, under LI we have

$$\beta_{IV} \equiv \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]} = \frac{\mathbb{E}[\Delta(U)(D_1 - D_0)]}{\mathbb{E}[D_1 - D_0]} = \frac{\mathbb{E}[\Delta(U)\mathbb{E}[D_1 - D_0|U]]}{\mathbb{E}[D_1 - D_0]}.$$

Since CMC holds, $\mathbb{E}[D_1 - D_0|Y_d] \geq 0$ almost surely (a.s) in Y_d, then $\mathbb{E}[D_1 - D_0|U] \geq 0$ a.s. in U under Assumption 7. Recall that, β_{IV} is well defined only when $\mathbb{P}(D = 1|Z = 1) \neq \mathbb{P}(D = 1|Z = 0)$, then there exists a set $A_U \subseteq \mathcal{Y}$ with positive measure under the distribution of U such that $\mathbb{E}[D_1 - D_0|U = u] > 0$ for all $u \in A_U$. Then, if $\Delta(U) > 0$ a.s. in U, then $\beta_{IV} > 0$. Therefore, we have the following result.

Proposition 3. *Suppose that Assumption 7, and CMC + LI hold and $\mathbb{P}(D = 1|Z = 1) \neq \mathbb{P}(D = 1|Z = 0)$, then the Wald estimand is no sign reversal.*

6. SIMULATION

We consider three data generating processes (DGPs). In all three designs, $Z \in \{0, 1\}$ with $\mathbb{P}(Z = 1) = 0.5$, $Y = D + U$, and $(U + 1, V) \sim N(0, 0.25, \rho)$. The three DGPs differ in the treatment functions. In DGP1, we set $D = 1\{V \leq 2Z - 0.5\}$; in DGP2, we set $D = 1\{V \leq 0.25\}$;

in DGP3, we set $D = 1\{|V - Z + 0.5| \geq 1\}$. Therefore, DGP1 belongs to "the interior" of H_0, DGP2 is a least favorable null (the knife-edge case), and DGP3 violates the CMC (note that LI holds in all three DGPs). Let $\alpha \in (0, 0.5)$ be a pre-specified significance level; we then expect that the rejection frequencies in those three DGPs shall be close to 0, α and 1, respectively.

TABLE 3. Rejection Frequency (clrtest)

Sig. level	Parametric			Local		
	10%	5%	1%	10%	5%	1%
DGP1						
$n = 200$	0%	0%	0%	0%	0%	0%
$n = 400$	0%	0%	0%	0%	0%	0%
$n = 800$	0%	0%	0%	0%	0%	0%
DGP2						
$n = 200$	10.8%	5.8%	1.5%	13.9%	7.7%	2.6%
$n = 400$	10.2%	5.9%	0.7%	11.5%	5.7%	1.2%
$n = 800$	10.0%	5.8%	1.4%	11.5%	5.9%	0.9%
DGP3						
$n = 200$	89.9%	83.6%	59.6%	55.5%	40.5%	19.3%
$n = 400$	99.4%	88.1%	91.6%	75.5%	63.4%	26.7%
$n = 800$	100%	100%	99.8%	91.4%	83.8%	63.3%

Based on 1000 replications.

Table 3 lists the preliminary simulation results of our test based on the "clrtest" command under difference choices of sample size and DGPs. In addition to the local linear regression, we also investigate the rejection frequency using the "parametric regression" option. For detailed descriptions of the Stata package, see Chernozhukov, Kim, Lee, and Rosen (2013). All results are computed based on 1000 replications. For DGP3 where the CMC fails to hold, the null hypothesis of LM is rejected with high probability even when the sample size is small, for example when $n = 200$. For DGP2, considered as the least favorable null, the rejection rate is close to the target levels. It is not surprising to see the test does not reject DGP1 since it is in the "interior" of the H_0.

7. 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).

15

7.1. **The "same-sex" instrument.** The context of the use of the same sex instrument in Angrist and Evans (1998) has been discussed in the Section 5. To reiterate, $D = 1$ denotes the household had a third child and $Z = 1$ denotes 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 of $620,573$ individuals from the 1990 Census Public Micro Samples (PUMS). The data contain information on age, gender, race, education, labor income and the number of children for women with more than two children and between 21 and 50 years old. The outcome variable of interest is log wage. Summary statistics for the sample are given in Table 4.

TABLE 4. Summary Statistics of Micro Census Data 1990 from Angrist and Evans (1998)

	Total	D=1	Z=1
Observations	620,573	209,782	312,825
Age	33.431 (5.497)	33.540 (5.092)	33.440 (5.500)
Years of Schooling	10.850 (2.468)	10.426 (2.645)	10.850 (2.464)
Race (Non White)	0.180 (0.385)	0.221 (0.415)	0.180 (0.384)
Having the third child (D=1)	0.338 (0.473)	1.000 (0.000)	0.368 (0.482)
First two same sex (Z=1)	0.504 (0.499)	0.549 (0.497)	1.000 (0.000)
Log Wage	5.853 (4.413)	5.003 (4.465)	5.824 (4.419)

Average and standard deviation (in the parentheses)

TABLE 5. Subgroups of PUMS (1990)

	21-28	29-35	36-42	43-50
White, <HS	21,478	26,272	9,013	1,577
White, HS	59,742	138,830	81,181	10,223
White, >HS	10,346	63,158	72,206	14,571
Non-white, <HS	11,676	14,684	6,134	1,202
Non-white, HS	15,964	26,181	12,143	1,640
Non-white, >HS	1,933	8,892	9,244	2,283

We divided the sample of $620,573$ observations into 24 subgroups according to race (white or non-white), education (lower than HS, HS or higher than HS) and age (21-28, 29-35, 36-42, 43-50) and conducted tests on each of these groups.[5] The subgroups sizes are reported in Table 5. Since the CKLR2013 Stata command does not allow us to work with more than 11,000 observations.

---

[5]The reason is that the Stata command only allows one conditioning variable in the moment inequalities.

Therefore, we randomly drew a subsample of 11,000 individuals from subgroups whose size is over 11,000.

TABLE 6. Same sex local method, truncated sample, $[q_{2.5}, q_{97.5}]$ (clrtest)

| | 21-28 | | | 29-35 | | | 35-42 | | | 43-50 | | |
	10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%
White, <HS	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
White, HS	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
White, >HS	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
Non-white, <HS	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
Non-white, HS	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
Non-white, >HS	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR

"R" stands for rejection and "NR" stands for no rejection.
All observations with zero log-income are excluded.

Throughout this section, we use the default choices of bandwidth and kernel functions recommended in Chernozhukov, Lee, and Rosen (2013) and Chernozhukov, Kim, Lee, and Rosen (2013), 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 $H_0 = H_0^{(1)} \cup H_0^{(2)} \cup \cdots \cup H_0^{(24)}$, where H_0 is defined as "Equation (5) holds for every subpopulation" and $H_0^{(s)}$ is defined as "Equation (5) 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 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). One difficulty is that the Stata command does not report p-value for the "clrtest" and hence we can not formally implement the multiple testing procedure. However, we are able to indirectly verify whether the FWER is controlled at targeted levels. Table 6 reports the testing results for each $H_0^{(s)}$. The testing results imply that the smallest p-values is greater than 0.42%. Therefore we can conclude that the multiple testing

procedures rejects no null hypothesis at $0.42\% \times 24 \approx 10\%$ level.[6] Because sex mix is virtually randomly assigned, this result can be interpreted as an evidence of the relative preference for the mix-sibling sex over the same sex within our population of interest.

As a robust check, 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 7), which is consistent with the results obtained from the local linear methods.

TABLE 7. 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.

7.2. **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 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)[8]. We divided the sample in 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 $[Q_{2.5\%}, Q_{97.5\%}]$.

Testing results for individual groups reported in Table 9. Note that the null hypothesis ($H_0^{(5)}$) is rejected at subgroup 5 of non-white person with high school education at the 90% and 95% levels,

---

[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 Chernozhukov, Lee, and Rosen (2013), "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.

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

TABLE 8. 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)

TABLE 9. 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.	320	867	1482	56	129	173
10%	NR	NR	NR	NR	R	NR
5%	NR	NR	NR	NR	R	NR
1%	NR	NR	NR	NR	NR	NR

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

respectively, but not at the 99% level. 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 (2013) obtained the same result without conditioning on subgroups.

7.3. **The "college proximity" instrument.** Card (1993) studied the causal effect of schooling on earning and used 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 education. The outcome variable is the log wage in 1976. The monotonicity under testing is $D_1 \geq D_0$.

The data is from the National Longitudinal Survey of Young Men (NLSYM) began in 1966 with men age 14-24 and continued with follow-up survey until 1981. Some summary statistics are reported

---

[9]The testing procedure with parametric regression method, however, rejects the null hypothesis at all three levels (see Table 7).
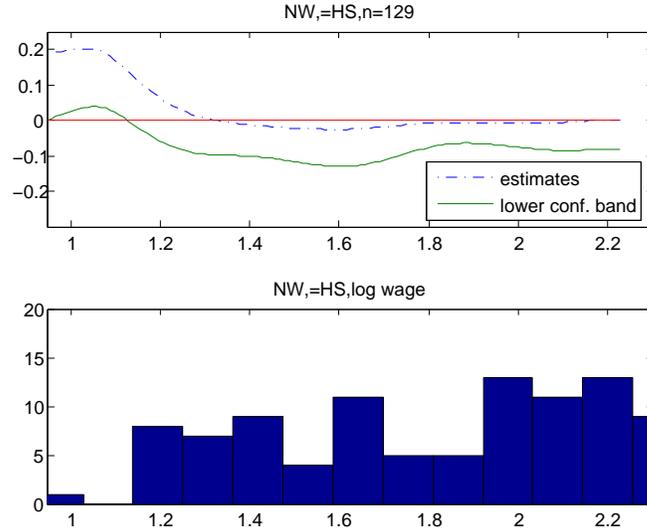
FIGURE 1. $\hat{\theta}(\cdot, 1)$ and $\hat{\theta}(\cdot, 1) - s(\cdot, 1) \times \hat{c}_{\hat{V}, 0.95}$

TABLE 10. 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)

in Table 10.[10] We considered three binary control variables: lived in southern states in 1966, lived in metro area in 1966 and being black. Table 11 reports the corresponding subgroup sizes.

We conduct the test on seven subgroups (exclude the one with only 5 observations). The null hypothesis $H_0^{(4)}$ is rejected for at subgroup 4 of Non-black men who lived in metro area of southern states as well as for the whole sample at 0.5% level. No rejection happens with other subgroups even at 10% level. The results in Table 12 imply that the multiple testing procedure of Holm (1979) would

---

[10]We dropped 608 observations with missing wages.

TABLE 11. 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 12. College proximity, local method (clrtest)

	NB,NS,NM	NB,NS,M	NB,S,NM	NB,S,M	B,NS,M	B,S,NM	B,S,M	All
Subgroup ID	1	2	3	4	5	6	7	
Obs.	429	1191	307	380	138	314	246	3005
5%	NR	NR	NR	R	NR	NR	NR	R
1%	NR	NR	NR	R	NR	NR	NR	R
0.5%	NR	NR	NR	R	NR	NR	NR	R

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

conclude that H_0 is rejected with the FWER be controlled by no more than $0.5\% \times 7 = 3.5\%$. The testing procedure with parametric method gives the same results.

Now it will be interesting to known 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 the subgroup 4 and the whole sample, respectively. It is quite interesting 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 that the density of log wage is reasonably high at this point, and therefore the rejection is unlikely due to boundary issue of the local linear estimation.

To summarize, our result suggests that the Wald estimator in such a case would be sign reversal. Thereby, although the "college proximity" seems to be a good instrument, researchers have to be aware that this instrument would not be a good one to use when the treatment effect is heterogenous.

8. CONCLUSION

In this paper we proposed an easy-to-implement test for testing the (sharp) testable characterization of LI and LM, the key identification conditions for the local average treatment effect. Interestingly, our testing procedure can be easily implemented using the Stata package provided by Chernozhukov, Kim, Lee, and Rosen (2013). Applying our test to the "same sex" instrument, the "draft eligibility"
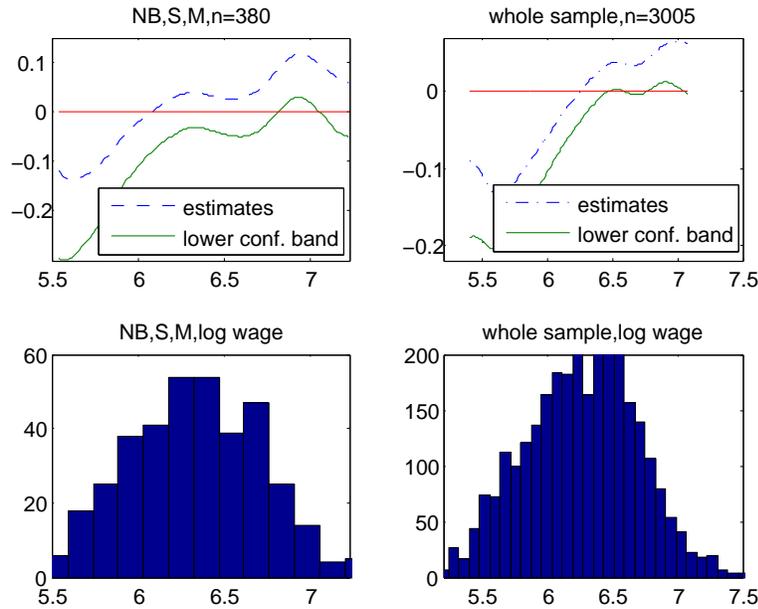
FIGURE 2. $\hat{\theta}(\cdot, 0)$ and $\hat{\theta}(\cdot, 0) - s(\cdot, 0) \times \hat{c}_{\hat{V}, 0.95}$

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., J. ANGRIST, AND G. IMBENS (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70(1), 91–117.

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," .

CHERNOZHUKOV, V., AND C. HANSEN (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245–261.

CHERNOZHUKOV, V., W. KIM, S. LEE, AND A. M. ROSEN (2013): "Implementing Intersection Bounds in Stata," *working paper*.

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

CHESHER, A. (2005): "Nonparametric identification under discrete variation," *Econometrica*, 73(5), 1525–1550.

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.

HAHN, J., P. TODD, AND W. VAN DER KLAAUW (2001): "Identification and Estimation of Treatment Effects with a Regression-discontinuity Design," *Econometrica*, 69(1), 201–209.

HECKMAN, J. J., AND B. E. HONORE (1990): "The empirical content of the Roy model," *Econometrica*, pp. 1121–1149.

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

HENRY, M., AND I. MOURIFIE (2014): "Sharp bounds in the binary Roy model," *Working Paper*.

HOLLAND, P. W. (1988): "Causal Inference, Path Analysis, and Recursive Structural Equations Models," *Sociological methodology*, 18(1), 449–484.

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.

IMBENS, G. W., AND W. K. NEWEY (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77(5), 1481–1512.

JUN, S. J., J. PINKSE, AND H. XU (2011): "Tighter bounds in triangular systems," *Journal of Econometrics*, 161(2), 122–128.

KASY, M. (2013): "Instrumental variables with unrestricted heterogeneity and continuous treatment," Discussion paper, Harvard University OpenScholar.

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

——— (2013): "A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model," Working paper.

KLING, J. R., J. B. LIEBMAN, AND L. F. KATZ (2007): "Experimental Analysis of Neighborhood effects," *Econometrica*, 75(1), 83–119.

KORN, E. L., S. BAUMRIND, ET AL. (1998): "Clinician Preferences and the Estimation of Causal Treatment Differences," *Statistical Science*, 13(3), 209–235.

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

SMALL, D. S., AND Z. TAN (2007): "A Stochastic Monotonicity Assumption for the Instrumental Variables Method," .

VYTLACIL, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331–341.

APPENDIX A. PROOF OF THE MAIN RESULTS

A.1. **Proof of Theorem 1.** (i) is proved in the main text. We show (ii) is true. First, we construct potential treatment $(\tilde{D}_0, \tilde{D}_1)$ which respects the LM assumption. For each z, let

$$\tilde{\pi}_{10|z} = \mathbb{P}(\tilde{D}_0 = 1, \tilde{D}_1 = 0 | Z = z) = 0, \tag{7}$$

$$\tilde{\pi}_{11|z} = \mathbb{P}(\tilde{D}_0 = 1, \tilde{D}_1 = 1 | Z = z) = \mathbb{P}(D = 1 | Z = 0), \tag{8}$$

$$\tilde{\pi}_{00|z} = \mathbb{P}(\tilde{D}_0 = 0, \tilde{D}_1 = 0 | Z = z) = \mathbb{P}(D = 0 | Z = 1), \tag{9}$$

$$\tilde{\pi}_{01|z} = \mathbb{P}(\tilde{D}_0 = 0, \tilde{D}_1 = 1 | Z = z) = 1 - \tilde{\pi}_{11} - \tilde{\pi}_{00}. \tag{10}$$

Since there is no defiers, LM holds. Next, we construct the joint distribution of Y_k and the individual types (compliers, defiers or always takers). For all $y \in \mathcal{Y}$ and $z \in \{0, 1\}$, let

$$\mathbb{P}(\tilde{Y}_1 \leq y, \tilde{c} | Z = z) = \mathbb{P}(Y \leq y, D = 1 | Z = 1) - \mathbb{P}(Y \leq y, D = 1 | Z = 0), \tag{11}$$

$$\mathbb{P}(\tilde{Y}_1 \leq y, \tilde{a} | Z = z) = \mathbb{P}(Y \leq y, D = 1 | Z = 0), \tag{12}$$

$$\mathbb{P}(\tilde{Y}_1 \leq y, \tilde{n} | Z = z) = \mathbb{P}(D = 0 | Z = 1)\mathbb{P}(Y \leq y), \tag{13}$$

$$\mathbb{P}(\tilde{Y}_0 \leq y, \tilde{c} | Z = z) = \mathbb{P}(Y \leq y, D = 0 | Z = 0) - \mathbb{P}(Y \leq y, D = 0 | Z = 1), \tag{14}$$

$$\mathbb{P}(\tilde{Y}_0 \leq y, \tilde{n} | Z = z) = \mathbb{P}(Y \leq y, D = 0 | Z = 1), \tag{15}$$

$$\mathbb{P}(\tilde{Y}_0 \leq y, \tilde{a} | Z = z) = \mathbb{P}(D = 1 | Z = 0)\mathbb{P}(Y \leq y). \tag{16}$$

We now show that Equations (11)–(13) characterize well defined distribution functions for $\tilde{Y}_1$. Note that

$$\mathbb{P}(\tilde{Y}_1 \leq y | Z = z) = \mathbb{P}(\tilde{Y}_1 \leq y, \tilde{c} | Z = z) + \mathbb{P}(\tilde{Y}_1 \leq y, \tilde{a} | Z = z) + \mathbb{P}(\tilde{Y}_1 \leq y, \tilde{n} | Z = z).$$

It is easy to verify that $\mathbb{P}(\tilde{Y}_1 \leq \underline{y} | Z = z) = 0$ and $\mathbb{P}(\tilde{Y}_1 \leq \bar{y} | Z = z) = 1$. It remains to verify that $\mathbb{P}(\tilde{Y}_1 \leq y | z)$ is nondecreasing in y. Since $\mathbb{P}(\tilde{Y}_1 \leq y, \tilde{a} | Z = z)$ and $\mathbb{P}(\tilde{Y}_1 \leq y, \tilde{n} | Z = z)$ is nondecreasing by definition, it is sufficient to show that $\mathbb{P}(\tilde{Y}_1 \leq y, \tilde{c} | Z = z)$ is nondecreasing. Let $y' < y''$, then

$$\mathbb{P}(\tilde{Y}_1 \leq y'', \tilde{c} | Z = z) - \mathbb{P}(\tilde{Y}_1 \leq y', \tilde{c} | Z = z)$$

$$= \mathbb{P}(Y \in (y', y''], D = 1 | Z = 1) - \mathbb{P}(Y \in (y', y''], D = 1 | Z = 0) \geq 0,$$

where the last inequality holds by Equation (1). Similarly, Equations (14), (15) and (16) characterize a well defined distribution function for $\tilde{Y}_0$.

By construction, both LI and LM are satisfied. It therefore remains to show that $\tilde{D}_1, \tilde{D}_0, \tilde{Y}_1, \tilde{Y}_0, Z$ is observationally equivalent to D_1, D_0, Y_1, Y_0, Z. Let consider the two following set of potential outcome models $Y = Y_1 D + Y_0(1 - D)$, $D = D_1 Z + D_0(1 - Z)$ and $\tilde{Y} = \tilde{Y}_1 \tilde{D} + \tilde{Y}_0(1 - \tilde{D})$, $\tilde{D} = \tilde{D}_1 Z + \tilde{D}_0(1 - Z)$. For $z \in \{0, 1\}$, Equations (11)–(16) imply that for all $A \in \mathcal{B}_y$,

$$\mathbb{P}(\tilde{Y} \in A, \tilde{D} = 1 | Z = 1) = \mathbb{P}(\tilde{Y}_1 \in A, \tilde{D}_1 = 1 | Z = 1)$$
$$= \mathbb{P}(\tilde{Y}_1 \in A, \tilde{D}_1 = 1, \tilde{D}_0 = 0 | Z = 1) + \mathbb{P}(\tilde{Y}_1 \in A, \tilde{D}_1 = 1, \tilde{D}_0 = 1 | Z = 1)$$
$$= \mathbb{P}(\tilde{Y}_1 \in A, \tilde{c} | Z = 1) + \mathbb{P}(\tilde{Y}_1 \in A, \tilde{a} | Z = 1) = \mathbb{P}(Y \in A, D = 1 | Z = 1),$$

$$\mathbb{P}(\tilde{Y} \in A, \tilde{D} = 0 | Z = 1) = \mathbb{P}(\tilde{Y}_0 \in A, \tilde{D}_1 = 0 | Z = 1)$$
$$= \mathbb{P}(\tilde{Y}_0 \in A, \tilde{D}_1 = 0, \tilde{D}_0 = 0 | Z = 1) + \mathbb{P}(\tilde{Y}_0 \in A, \tilde{D}_1 = 0, \tilde{D}_0 = 1 | Z = 1)$$
$$= \mathbb{P}(\tilde{Y}_0 \in A, \tilde{n} | Z = 1) = \mathbb{P}(Y \in A, D = 0 | Z = 1).$$

$$\mathbb{P}(\tilde{Y} \in A, \tilde{D} = 1 | Z = 0) = \mathbb{P}(\tilde{Y}_1 \in A, \tilde{D}_0 = 1 | Z = 0)$$
$$= \mathbb{P}(\tilde{Y}_1 \in A, \tilde{D}_1 = 1, \tilde{D}_0 = 1 | Z = 0) + \mathbb{P}(\tilde{Y}_1 \in A, \tilde{D}_1 = 0, \tilde{D}_0 = 1 | Z = 0)$$
$$= \mathbb{P}(\tilde{Y}_1 \in A, \tilde{a} | Z = 0) = \mathbb{P}(Y \in A, D = 1 | Z = 0),$$

$$\mathbb{P}(\tilde{Y} \in A, \tilde{D} = 0 | Z = 0) = \mathbb{P}(\tilde{Y}_0 \in A, \tilde{D}_0 = 0 | Z = 0)$$
$$= \mathbb{P}(\tilde{Y}_0 \in A, \tilde{D}_1 = 1, \tilde{D}_0 = 0 | Z = 0) + \mathbb{P}(\tilde{Y}_1 \in A, \tilde{D}_1 = 0, \tilde{D}_0 = 0 | Z = 0)$$
$$= \mathbb{P}(\tilde{Y}_0 \in A, \tilde{c} | Z = 0) + \mathbb{P}(\tilde{Y}_0 \in A, \tilde{n} | Z = 0) = \mathbb{P}(Y \in A, D = 0 | Z = 0),$$

This completes the proof. □

Lemma 1. *Let $c_0 = \mathbb{P}(Z = 0)$, $m(y) = \mathbb{E}[c_0 D Z | Y = y]$, $\tilde{m}(y)$ 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 $\hat{c}_0$, then

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

Proof. The conclusion follows from the the fact that $\hat{c}_0$ does not depend on y and $\sup_{y \in \mathcal{Y}} |\tilde{m}(y)| < \infty$ with probability one. $\square$

A.2. **Proof of Proposition 1.** First note that by Lemma 1, it is sufficient to treat c_0 and c_1 as if they were known. Let $U(W_i, 1) = c_1 D_i(1 - Z_i) - c_0 D_i Z_i - \theta(Y_i, 1)$, $U(W_i, 0) = c_0(1 - D_i)Z_i - c_1(1 - D_i)(1 - Z_i) - \theta(Y_i, 0)$, $\hat{U}(W_i, 1) = c_1 D_i(1 - Z_i) - c_0 D_i Z_i - \hat{\theta}(Y_i, 1)$ and $\hat{U}(W_i, 0) = c_0(1 - D_i)Z_i - c_1(1 - D_i)(1 - Z_i) - \hat{\theta}(Y_i, 0)$. 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. Furthermore, let $\mathbb{B}(g_v)$ be a $\mathbb{P}$-Brownian bridge with continuous sample path on $\mathcal{Y}$ for each k.

$$\mathbb{G}_n(v) = \frac{\theta(v) - \hat{\theta}(v)}{\sigma(v)}, \quad \sigma^2(v) = \frac{1}{nh} \mathbb{E}[g_v^2(U, Y)], \quad \mathbb{G}_n^*(v) = \frac{\mathbb{B}(g_v)}{\mathbb{E}[g_v^2(U, Y)]}$$

and for i.i.d. $\eta_i \sim N(0, 1)$ independent with data $\mathbf{W}_n$, define

$$\mathbb{G}^o(\hat{g}_v) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i \hat{g}_v(U, Y_i), \quad \mathbb{G}_n^{\star}(v) = \frac{\mathbb{G}^o(v)}{\sqrt{nh} s(v)}, \quad s^2(y, k) = \frac{1}{nh^2} \sum_{i=1}^{n} \frac{\hat{U}^2(W_i, k)}{\hat{f}^2(y)} K^2\left(\frac{y - Y_i}{h}\right).$$

Given these notation, we will verify the conditions NK of CLR2013 based on these notation, which is in turn sufficient to verify conditions (i)-(vi) in CLR2013, Appendix F. Conditions (i) and (ii) hold by Assumption 4. $U(W_i, k)$ is bounded because Y, D and Z are bounded, therefore condition (iii) holds. Since the density of Y is continuous, the density of $U(W_i, k)$ is a finite mixture of continuous densities and hence exists. Again by Assumption 4 it is bounded from above and below. This verifies condition (iv). Condition (v) on kernel function and condition (vi) on bandwidth follows from Assumptions 5 and 6 with $p = 1$. Condition NK(ii) holds by the standard Kernel estimation methods.

Given this, part (1) of Proposition 1 hold by Theorem 6, (a)-(i) of CLR2013; part (2) holds by Theorem 6 (b)-(i,iii) because the contact set $V_0 = V$, therefore condition V and Equation 4.6 of CLR2013 hold with $\rho_n = 1$, $c_n = \infty$.

A.3. **Proof of Proposition 2.** Indeed, for all $A \subseteq \mathcal{B}_y$ we have

$$\mathbb{P}(Y \in A, D = 1 | Z = 1) = \mathbb{P}(Y_1 \in A, D = 1 | Z = 1)$$

$$= \mathbb{P}(D = 1 | Y_1 \in A, Z = 1) \mathbb{P}(Y_1 \in A | Z = 1) = \mathbb{P}(D = 1 | Y_1 \in A, Z = 1) \mathbb{P}(Y_1 \in A)$$

$$= \mathbb{P}(D_1 = 1 | Y_1 \in A) \mathbb{P}(Y_1 \in A).$$

The third and fourth equalities hold under Assumption 1. Therefore,

$$\mathbb{P}(Y \in A, D = 1 | Z = 0) \leq \mathbb{P}(Y \in A, D = 1 | Z = 1) \text{ for all } A \subseteq \mathcal{B}_y \tag{17}$$

is equivalent to

$$\mathbb{P}(D_0 = 1 | Y_1 \in A) \leq \mathbb{P}(D_1 = 1 | Y_1 \in A) \text{ for all } A \subseteq \mathcal{B}_y. \tag{18}$$

which is equivalent to

$$\mathbb{P}(D_0 = 1 | Y_1 = y) \leq \mathbb{P}(D_1 = 1 | Y_1 = y) \text{ for all } y. \tag{19}$$

While the first equivalence is obvious the last one need more explanations. Indeed, $(18) \Rightarrow (19)$ since (19) is a particular case of (18) and the reverse $(18) \Leftarrow (19)$ is also true since $\mathbb{P}(D_1 = 1 | Y_1 \in A) = \frac{1}{\mathbb{P}(Y_1 \in A)} \int_A \mathbb{P}(D_1 = 1 | Y_1 = y) dF_{Y_1}(y)$.