A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model

Toru Kitagawa[*][†]

Department of Economics, Brown University

This draft: September, 2008

Abstract

This paper develops a test for the conditions of instrument validity in the heterogeneous treatment effect model (Imbens and Angrist (1994)). We show that under the conditions of instrument validity, the point-identified complier's outcome densities must be nonnegative. This provides the testable implication for instrument validity. To infer this testable implication from data, we develop a specification test based on a Kolmogorov-Sminov type statistic to assess the nonnegativity of the densities. We provide a bootstrap algorithm to implement the proposed test and show its asymptotic validity.

Keywords: Treatment Effects, Instrumental Variable, Specification Test, Bootstrap.
JEL Classification: C12, C15, C21.

[*] Email: Toru_Kitagawa@brown.edu
Webpage: http://www.econ.brown.edu/Students/Toru_Kitagawa/index.htm

[†] I am deeply grateful to Frank Kleibergen and Guido Imbens for helpful comments. Financial support from the Merit Dissertation Fellowship is gretefully acknowledged.

1 Introduction

In this paper, we develop a test procedure for the instrumental validity in the heterogeneous treatment effect model. When we suspect that one's participation to a treatment depends on his potential outcomes, a common strategy to extract identifying information for counterfactual causal effects is to employ an instrumental variable Z. As is demonstrated in Imbens and Angrist (1994), Angrist, Imbens and Rubin (1996), and Heckman and Vytracil (1999, 2001), when the instrument satisfies the two key conditions, we can point-identify the average causal effects for those whose participation decision is strictly randomized by the instrument. (the local average treatment effect) These key conditions consist of i) *random treatment assignment (RTA):* an instrument is assigned independently from individual heterogeneities which affect one's outcome and participation response, and ii) *monotonic participation response to instrument (MPR):* one's participation response to the instrument is uniform in a certain sense over the entire population.[1]

When we analyze (quasi-)experimental data with possible incompliance, we often use the initial treatment assignment as an instrument. In this case, the instrumental validity is reasonably satisifed as far as the the initial treatment assignment is completely randomized and incompliance is allowed only for those who are initially assigned to the treatment group (see, for example, Abadie, Angrist, and Imbens (2002) and Kling, Liebman, and Katz (2007)). But, if the incompliance is also allowed for those initially assigned to the control group, we face a risk of violating MPR. Examples of this contain the well-known draft lottery of Angrist (1991) and the applications of the fuzzy regression discontinuity design (Campbell (1969), Hahn, Todd, and Van der Klaauw (2001)) where eligibility for a treatment based on one's attribute is used as an instrument. When we conduct an analysis using observational data, the exogeneity of instrument becomes less credible and therefore, not only MPR, but also RTA becomes a threat for the instrument validity. Although validating an instrument is the core of identifying the causal effects, there have been no procedures proposed to empirically test the aforementioned instrumental validity. Because of this, the instrumental validity is simply assumed or justified by indirect evidence outside of data.

The first contribution of this paper is to clarify the testability of the instrument validity in the heterogenous treatment effect model with a binary treatment and a binary instrument. The refutability result of this paper is closely related to the pointi-identification result of the complier's outcome distributions by Imbens and Rubin (1997). They show that under RTA and MPR, the distribution of complier's treated outcome and that of complier's control outcome are point-identified. But, from the data, the point estimator of the complier's outcome densities can take negative values on some subsets in the outcome support. We focus on this phenomenon as a clue to refute the instrumental validity. That is, if we

[1] MPR considered in this paper stands for the restriction termed as "monotonicity" in Imbens and Angrist (1994). The reason that we call it MPR is to distinguish the monotonicity between one's participation response and instrument from the monotonicity between one's outcome response and instrument considered in Manski and Pepper (2000).

obtain negative estimates for complier's treated outcome or control outcome density on some regions in the outcome support, we interpret it as a counter-evidence for the joint restriction of RTA and MPR since the probability density function cannot be negative. We derive the condition for the data generating process to yield nonnegative complier's potential outcome densities. We demonstrate that the refuting rule based on that condition is most powerful for screening out the violation of the instrumental validity in the heterogenous treatment effect model.

The second contribution of this paper is to develop a specification test for the instrumental validity based on the aforementioned refutability result. We propose a Kolmogorov-Sminov type test statistic to measure how serious the nonnegativity of the compliers outcome density is violated in data. The asymptotic distribution of the proposed test statistic is not analytically tractable, and therefore the critical values are difficult to obtain. In order to overcome this problem, we develop a bootstrap algorithm to obtain asymptotically valid critical values. As Romano (1988) demonstrated, the bootstrap is widely applicable and easy to implement to obtain the critical values of the general Kolmogorov-Sminov type goodness-of-fit statistic. This is also the case for our test procedure.

The rest of the paper is organized as follows. In Section 2, we demonstrate the refutability of the instrumental validity in the heterogeneous treatment effect model. In Section 3, we construct a statistic to test the testable implication obtained in Section 2 and provide an algorithm of the bootstrap procedure. Monte Carlo simulations and two empirical applications are provided in Section 4. Proofs are provided in the appendices.

2 Model

Let Y_1 represent the potential outcome with a treatment, and Y_0 represent the potential outcome without the treatment. They are scalar variables and their support is denoted by $\mathcal{Y}$. The observed outcome is denoted by Y_{obs}. Let D_{obs} indicate the observed participation response such that $D_{obs} = 1$ when one participates to the treatment while $D_{obs} = 0$ if one does not. Thus, the observed outcome is written as $Y_{obs} = Y_1 D_{obs} + Y_0(1 - D_{obs})$. We denote a binary instrument by Z. As in Angrist and Imbens (1994), we introduce D_1 as the potential participation decision that one would take if $Z = 1$. Similarly, we define D_0 for $Z = 0$. Associated with the potential selection indicators, we define the individual type T that indicates individual participation response to the instrument Z.

$T = c$: *complier* if $D_1 = 1, D_0 = 0$
$T = n$: *never-taker* if $D_1 = 0, D_0 = 0$
$T = a$: *always-taker* if $D_1 = 1, D_0 = 1$
$T = d$: *defier* if $D_1 = 0, D_0 = 1$.

The following three assumptions guarantee point-identification of the local average treatment effects for compliers as well as the marginal distributions of the counterfactual outcomes for compliers (see Imbens and Angrist (1994) and Imbens and Rubin (1997)).

Assumption

1. *Random Treatment Assignment (RTA):* Z is jointly independenct of (Y_1, Y_0, D_1, D_0).
2. *Monotonic Participation Response to Instrument (MPR): Without loss of generality, assume* $\Pr(D_{obs} = 1|Z = 1) \geq \Pr(D_{obs} = 1|Z = 0)$. The potential participation indicators satisfy $D_1 \geq D_0$ *with probability one.*

Note that the above assumptions are defined in terms of the potential variables. RTA is stronger than the conventional instrumental exclusion restriction since it only restricts Z to being independent of the potential outcomes. MPR states that the ordering of the potential particiation indicators are identical over the entire population and there are no defiers in the population $\Pr(T = d) = 0$. Since we never observe all the potential variables of the same individual, we cannot directly examine these assumptions from data, and therefore necessary and sufficient testable implications for these assumptions are not available. Hence, we examine the refutability by looking for a testable implication as a necessary condition for the instrumental validity.

To illustrate our analytical framework, we introduce the following notations. Let P and Q be the conditional probability distributions of $(Y_{obs}, D_{obs}) \in \mathcal{Y} \times \{1, 0\}$ given $Z = 1$ and $Z = 0$ respectively. We interpret the data generating process to have the two-sample structure in terms of the assigned value of Z. For a subset $A \subset \mathcal{Y}$ and $d = 1, 0$, $P(Y_{obs} \in A, D_{obs} = d)$ and $Q(Y_{obs} \in A, D_{obs} = d)$ represent $\Pr(Y_{obs} \in A, D_{obs} = d|Z = 1)$ and $\Pr(Y_{obs} \in A, D_{obs} = d|Z = 0)$ respectively. Note that P and Q are the joint distributions of the *observable* variables (Y_{obs}, D_{obs}), and therefore we can consistently estimate P and Q by data.

We now state the refutability result of the instrumental validity. Provided that the population has a strictly positive fraction of compliers, the conclusion of the next proposition is equivalent to the nonnegativity of the complier's outcome densities pinned down under the instrumental validity (Imbens and Rubin (1997)). A proof is given in Appendix A.

Proposition 1 *If a population distribution of (Y_1, Y_0, D_1, D_0, Z) satisfies RTA and MPR, then, the data generating process P and Q satisfies the following inequalities for arbitrary Borel sets B in $\mathcal{Y}$,*

$$\begin{aligned} &P(Y_{obs} \in B, D_{obs} = 1) \geq Q(Y_{obs} \in B, D_{obs} = 1), \\ &P(Y_{obs} \in B, D_{obs} = 0) \leq Q(Y_{obs} \in B, D_{obs} = 0). \end{aligned} \tag{1}$$

Conversely, if the data generating process P and Q satisfies these inequalities for all Borel sets B, then there exists a joint probability law of (Y_1, Y_0, D_1, D_0, Z) that is compatible with the data generating process P and Q, RTA, and MPR.

Let $p(y, D_{obs} = d)$ and $q(y, D_{obs} = d)$ be the probability density function of P and Q on $\mathcal{Y} \times \{d\}$ with respect to a dominating measure μ. . In terms of the density functions, the above two inequalities are equivalent to

$$\begin{aligned} p(y, D_{obs} = 1) &\geq q(y, D_{obs} = 1) \quad \mu\text{-a.e.}, \\ p(y, D_{obs} = 0) &\leq q(y, D_{obs} = 0) \quad \mu\text{-a.e.} \end{aligned}$$

These inequalities imply that when the instrument is valid, we must observe the configuration of the densities as in Figure 1. The left-hand side figure corresponds to Y_1's distribution and the right figure corresponds to Y_0's distribution. The dotted line in each figure represents the probability density of the potential outcomes, i.e., $f_{Y_1}(y)$ is the marginal density of the treated outcome and $f_{Y_0}(y)$ is the marginal density of the control outcome. The solid lines represent $p(y, D_{obs} = d)$ and $q(y, D_{obs} = d)$, which are point-identifiable by data. Note that their integrals are equal to the probability of $D_{obs} = d$ conditional on Z. Therefore, the scale of $p(y, D_{obs} = d)$ and $q(y, D_{obs} = d)$ is smaller than $f_{Y_1}(\cdot)$ and $f_{Y_0}(\cdot)$. Furthermore, $p(y, D_{obs} = d)$ and $q(y, D_{obs} = d)$ both lie below the potential outcome density $f_{Y_d}(\cdot)$. This is because RTA implies

$$\begin{aligned} f_{Y_d}(y) &= f_{Y_d|Z}(y|Z = 1) \\ &= f_{Y_d, D_{obs}|Z}(y, D_{obs} = d|Z = 1) + f_{Y_d, D_{obs}|Z}(y, D_{obs} = 1 - d|Z = 1) \\ &= p(y, D_{obs} = d) + f_{Y_d, D_{obs}|Z}(y, D_{obs} = (1 - d)|Z = 1) \end{aligned}$$

and

$$f_{Y_d}(y) = q(y, D_{obs} = d) + f_{Y_d, D_{obs}|Z}(y, D_{obs} = (1 - d)|Z = 0).$$

The second term in the right hand side of the above equations correspond to the density function for the missing treated or control outcomes, so they must be nonnegative.

When RTA and MPR hold in the population, Proposition 1 implies that the two identifiable density functions $p(y, D = d)$ and $q(y, D = d)$ must be nested as shown in Figure 1. For the treated outcome densities, $p(y, D = 1)$ must lie above $q(y, D = 1)$ and for the control outcome densities, $q(y, D = 0)$ must lie above $p(y, D = 0)$. Under RTA and MSR, we can point-identify the complier's outcome densities by the areas between these two densities rescaled by their area (see the proof of Proposition 1 in Appendix A). Thus, the inequalities of Proposition 1 constitute necessary conditions for the instrument validity.

The converse statement of Proposition 1 clarifies that if the data generating process admits the inequalities (1), then we can construct a population distribution of (Y_1, Y_0, D_1, D_0, Z) which does not contradict the data generating process and the instrument validity. This implies that no other refuting rules can screen out violations of the instrument validity more than the refuting rule based on the inequalities (1) does. In this sense, the refuting rule of Proposition 1 is most powerful in screening out the violations of the instrument validity.

Note that Proposition 1 does not give an if and only if statement for the instrumental validity. That is, an invalid instrument does not necessarily imply a violation of the

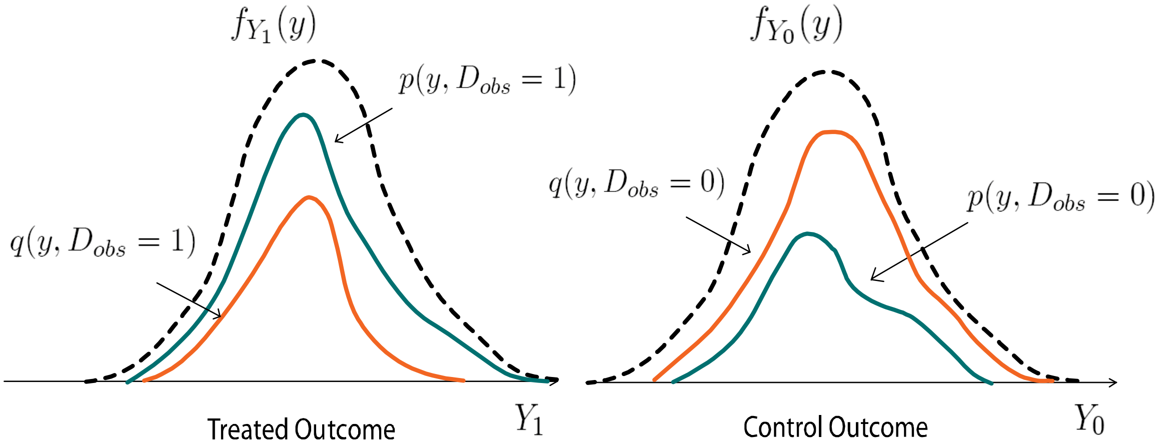

Figure 1: *When we observe that the observable densities* $p(y, D_{obs} = 1)$ and $q(y, D_{obs} = d)$ *are nested as in this figure, the instrumental validity is not refuted.*

inequalities. In this sense, testing the inequalities does not guarantee to screen out all the possible violations of the instrumental validity.

If we observe the configuration of the densities like Figure 2, we can refute at least one of the instrumental validity conditions since some of the inequalities (1) are violated on some subsets of the outcome support. These subsets are labeled as V_1 and V_2 in Figure 2. Although observing the configuration of the densities like Figure 2 does not tell us which conditions are violated in the population, it allows us to conclude that the chosen instrument is not valid to point-identify the local average treatment effects and, hence, the classical IV-estimator breaks down.

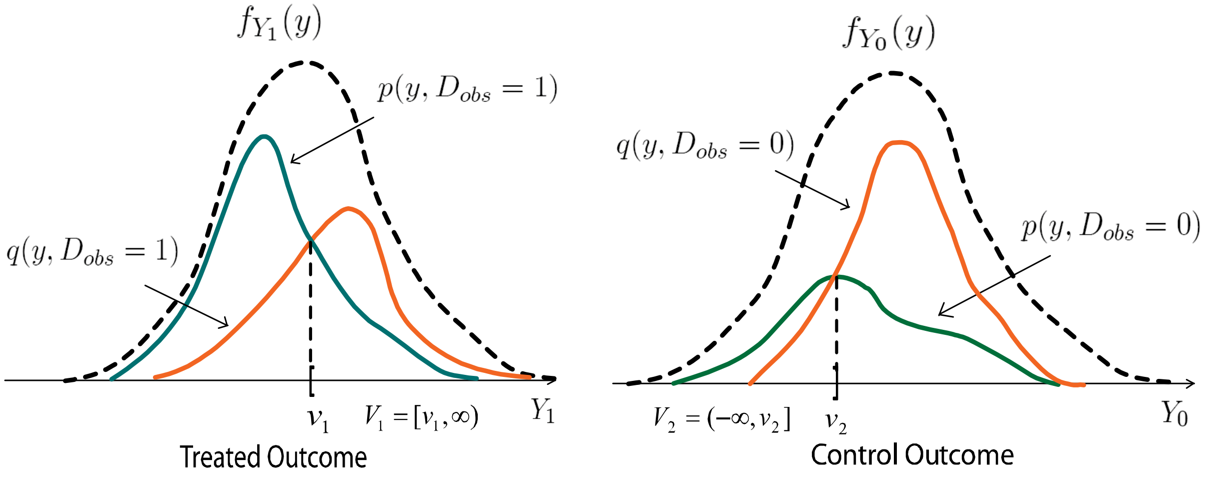

Figure 2: *When we observe the above configuration of the densities, we can refute the instrumental validity since at the subset $V_1 = [v_1, \infty)$, the first inequality in Proposition 1 is violated. The right-hand side picture shows that the second inequality in Proposition 1 is violated at $V_2 = (-\infty, v_2]$.*

3 Test Procedure

P and Q are point-identified by the sampling process, and therefore we can examine the validity of the inequalities (1) by inferring whether estimators for P and Q satisfy them or not.

Let sample consist of N i.i.d observations of (Y_{obs}, D_{obs}, Z). We divide the sample into two subsamples in terms of the value of Z. Let m be the sample size with $Z_i = 1$ and n the sample size with $Z_i = 0$. Let $(Y^1_{obs,i}, D^1_{obs,i})$, $i = 1, \ldots, m$ be the observations with $Z = 1$ and $(Y^0_{obs,j}, D^0_{obs,j})$, $j = 1, \ldots, n$ be those with $Z = 0$. We assume $m/N \to \lambda$ as $N \to \infty$ almost surely where $\lambda \in (\epsilon, 1 - \epsilon)$ for some $\epsilon > 0$. We estimate P and Q by the empirical distributions,

$$\begin{aligned} P_m(V, d) &\equiv \frac{1}{m} \sum_{i=1}^{m} I\{Y^1_{obs,i} \in V \text{ and } D^1_{obs,i} = d\}, \\ Q_n(V, d) &\equiv \frac{1}{n} \sum_{j=1}^{n} I\{Y^0_{obs,j} \in V \text{ and } D^0_{obs,j} = d\}. \end{aligned}$$

We measure the degree of violation of the inequalities (1) by the next statistic.

$$T_N = \left(\frac{mn}{N}\right)^{1/2} \max \left\{ \begin{array}{c} \sup_{V \in \mathbb{V}} \{Q_n(V, 1) - P_m(V, 1)\}, \\ \sup_{V \in \mathbb{V}} \{P_m(V, 0) - Q_n(V, 0)\} \end{array} \right\}, \tag{2}$$

where $\mathbb{V}$ is a collection of subsets in $\mathcal{Y}$.

This test statistic is designed to measure the degree of the violations of the inequalities (1) using the empirical distributions. If the sample counterpart of the inequality (1) is violated for a subset V, then, the first supremum in the max operator of the test statistic is positive. Similarly, when the sample counterpart of the inequality (**??**) is violated for some subset V, then the second term becomes positive. The proposed test statistic returns the maximal deviations of the above inequalities where the maximum is searched over a class of subsets $\mathbb{V}$.

The test statistic can be seen as a variant of the Kolmogorov-Sminov type nonparametric distance test statistic (Romano (1988)). This test statistic is not pivotal due to the discreteness of D_{obs} and the asymptotic distribution can depend on P and Q. Choice of $\mathbb{V}$ will not affect the size of test while it can affect power of the test.

Although Proposition 1 suggests us to take $\mathbb{V}$ as the Borel σ-algebra of $\mathcal{Y}$, we cannot take it to be as rich as the Borel σ-algebra unless Y is discrete. In order for the above test statistic to have an asymptotic distribution, a specified $\mathbb{V}$ has to guarantee the uniform convergence property of the empirical processes of P_m and Q_n. A class of subsets which meets this requirement is the Vapnik-Červonenkis class (VC-class). For example, a collection of left unbounded intervals $\{(-\infty, y]; y \in \mathbb{R}\}$ and a collection of the finite number of disjoint intervals are the examples of the VC-classes. (See e.g., Dudley (1999) and van der Vaart and Wellner (1996) for the general construction of the VC-classes).

We will employ two specific VC-classes in our Monte Carlo studies and empirical applications given in the next section. They are the *half unbounded interval class* $\mathbb{V}_{half}$ and the *histogram class* $\mathbb{V}_{hist}$. The half unbounded interval class is simply a collection of right unbounded intervals and left unbounded intervals,

$$\mathbb{V}_{half} = \{(-\infty, y]; y \in \mathbb{R}\} \cup \{[y, \infty); y \in \mathbb{R}\}. \tag{3}$$

The histogram class is the power set of the histogram bins whose breakpoints can float over $\mathbb{R}$. Algebraically, this can be expressed as follows. Let $h > 0$ be a fixed positive number representing the binwidth and L be the number of bins. Pick an initial breakpoint $y_0 \in \mathbb{R}$ and consider equally distanced L points $-\infty < y_0 < y_1 < \cdots < y_{L-1} < \infty$ where $y_l = y_0 + lh$, $l = 1, \ldots, (L-1)$. Denote the $(L+1)$ disjoint intervals formed by these L points by $H_0(y_0, h) = (-\infty, y_0]$, $H_l(y_0, h) = [y_{l-1}, y_l]$, $l = 1, \ldots, (L-1)$, and $H_L(y_0, h) = [y_{L-1}, \infty)$. Let $I_j(L)$, $j = 1, \ldots, 2^{L+1}$ represent all the possible subsets of the indices $\{0, 1, \ldots, L\}$. Given $\mathcal{Y}_0$ a set of the smallest breakpoint y_0, the histogram class with binwidth h and the number of bins L is defined as

$$\mathbb{V}_{hist}(h, L, \mathcal{Y}_0) = \left\{ \bigcup_{l \in I_j(L)} H_l(y_0, h) : y_0 \in \mathcal{Y}_0,\ j = 1, \ldots, 2^{L+1} \right\}. \tag{4}$$

In contrast to a rather complicated expression, the histogram class is flexible and simple to implement.

For the test statistic (2), $P = Q$ is the least favorable null hypothesis among the composite null hypotheses defined by the inequalities (1). Therefore, we will find the critical value with a nominal level α by estimating the $(1-\alpha)$-th quantile of the asymptotic distribution of T_N under the least favorable null $P = Q$. If the estimated critical values are consistent to the $(1-\alpha)$-th quantile of the asymptotic distribution of T_N under the least favorable null, the resulting testing procedure has correct size.

As discussed in Romano (1988), the resampling method is an attractive approach to estimate asymptotically valid critical values for the Kolmogorov-Sminov type test statistic since its asymptotic distribution generally does not have an analytically tractable distribution function. Bootstrap resolves this issue by estimating the null distribution of the statistic by the empirical distribution of the resampled test statistics. Given that the the composite null has the least favorable null, bootstrap samples are drawn from $\hat{P}$ and $\hat{Q}$, which is consistent to the least favorable null hypothesis, i.e., $\hat{P} = \hat{Q}$. In the two sample hypothesis testing problem with the null hypothesis given by the equality of the two distributions, one choice of the resampling distribution is the pooled empirical distribution H_N, the empirical distribution of the pooled data $(Y^1_{obs,1}, D^1_{obs,1}), \ldots, (Y^1_{obs,m}, D^1_{obs,m}), (Y^0_{obs,1}, D^0_{obs,1}), \ldots, (Y^0_{obs,n}, D^0_{obs,n})$. Abadie (2002) proposes the bootstrap procedure to test hypotheses on distributional features between the complier's treated and control outcomes. Although the null hypothesis and test statistic are different, our bootstrap procedure shown below is analogous to Abadie (2002).

Bootstrap procedure:

1. S*ample* $(Y^*_{obs,i}, D^*_{obs,i})$, $i = 1, \ldots, m$ *randomly with replacement from the pooled empirical distribution* H_N *and construct the bootstrap empirical distribution* P^*_m. *Similarly, sample* $(Y^*_{obs,j}, D^*_{obs,j})$, $j = 1, \ldots, n$ *randomly with replacement from the pooled empirical distribution* H_N *and construct the bootstrap empirical distribution* Q^*_n.

2. *Compute the test statistic* T^*_N *defined in (2) by plugging in the bootstrapped empirical distributions* P^*_m *and* Q^*_n.

3. *Iterate Step 1 and Step 2 and get the empirical distribution of* T^*_N. *For a chosen nominal level* $\alpha \in (0, 1/2)$, *we obtain the bootstrapped critical value* $\hat{c}_{boot}(1-\alpha)$ *from its empirical* $(1-\alpha)$*-th quantile .*

4. *Reject the null hypothesis if* $T_N > \hat{c}_{boot}(1-\alpha)$.

Note that the bootstrap sample is drawn from the pooled empirical distribution because our interest is in estimating the null distribution of T_N under the least favorable null hypothesis, $P = Q$. This enables us to control the supremum of the asymptotic false rejection probabilities at the chosen nominal level α,

$$\sup_{(P,Q)\in H_0} \lim_{N\to\infty} \Pr(T_N > \hat{c}_{boot}(1-\alpha)) = \alpha. \tag{5}$$

This is the conventional definition of the pointwise consistency of test.

The asymptotic validity of the proposed bootstrap is stated in the next proposition. A proof is given in Appendix B.

Proposition 2 *Let $\mathbb{V}$ be a VC-class and $\alpha \in (0, 1/2)$. i) For the null hypothesis of P and Q given by the inequalties (1), the proposed bootstrap test procedure provides pointwise correct asymptotic size (5). ii) If, for a fixed alternative, there exist some $V \in \mathbb{V}$ which violates (1), then the proposed bootstrap testing procedure is consistent, i.e., the rejection probability converges to one as $N \to \infty$.*

4 Monte Carlo Studies and Empirical Applications

4.1 Small sample performance

To examine the finite sample performance of our bootstrap test, we perform a Monte Carlo simulation. We specify the sampling process as the least favoralble null $P = Q$, and therefore the test asymptotically achieves nominal size.

$$\begin{aligned} p(y, D &= 1) = q(y, D = 1) = 0.5 \times \mathcal{N}(1, 1), \\ p(y, D &= 0) = q(y, D = 0) = 0.5 \times \mathcal{N}(0, 1). \end{aligned}$$

We consider two specifications of $\mathbb{V}$. One is the half unbounded interval class $\mathbb{V}_{half}$ and the other is the histogram class $\mathbb{V}_{hist}$ defined in Section 3. The histgram class provides a finer collection of subsets than the half unbounded interval class. This implies that the histogram class has more refutability power in the sense that it can asymptotically reject more alternatives than the half unbounded interval class. In the finite sample situation, however, there will be a trade-off between asymptotic refutability power and finite sample test power. In order to see the effect of a choice of the binwidth of $\mathbb{V}_{hist}$ to test size and power, we consider two different choices of binwidth, 0.8 and 0.4. The number of bins are 12 and 24 respectively. The set of initial breakpoints are $\mathcal{Y}_0 = [-4.4, -3.6)$ for the former histogram class and $\mathcal{Y}_0 = [-4.4, -4.0)$ for the latter.

For each specification of the sample size (m, n), we simulate the test procedure 2000 times with 500 bootstrap iterations. Table 1 shows that for every specification of $\mathbb{V}$, the test has good size performance even for relatively small sample size, $(m, n) = (50, 50)$. The unbalanced sample case, $(m, n) = (50, 250)$, shows a slight size distortion, while size of the test is overall satisfactory. In addition, we can see that size of the test is not affected by the choice of $\mathbb{V}$.

Table 1: Test Size in Small Samples

Monte Carlo iterations 2000, Bootstrap iterations 500.

	Specification of $\mathbb{V}$								
	Nominal test size								
	$\mathbb{V}_{half}$			$\mathbb{V}_{hist}$ binwidth 0.8			$\mathbb{V}_{hist}$ binwidth 0.4		
sample size (m,n)	.10	.05	.01	.10	.05	.01	.10	.05	.01
(50,50)	.085	.042	.008	.098	.049	.009	.106	.053	.010
(50,250)	.124	.073	.022	.098	.046	.008	.118	.058	.014
(100,100)	.108	.054	.015	.113	.052	.015	.104	.054	.001
(500,500)	.092	.046	.011	.104	.057	.017	.112	.062	.014
s.e.	.007	.005	.002	.007	.005	.002	.007	.005	.002

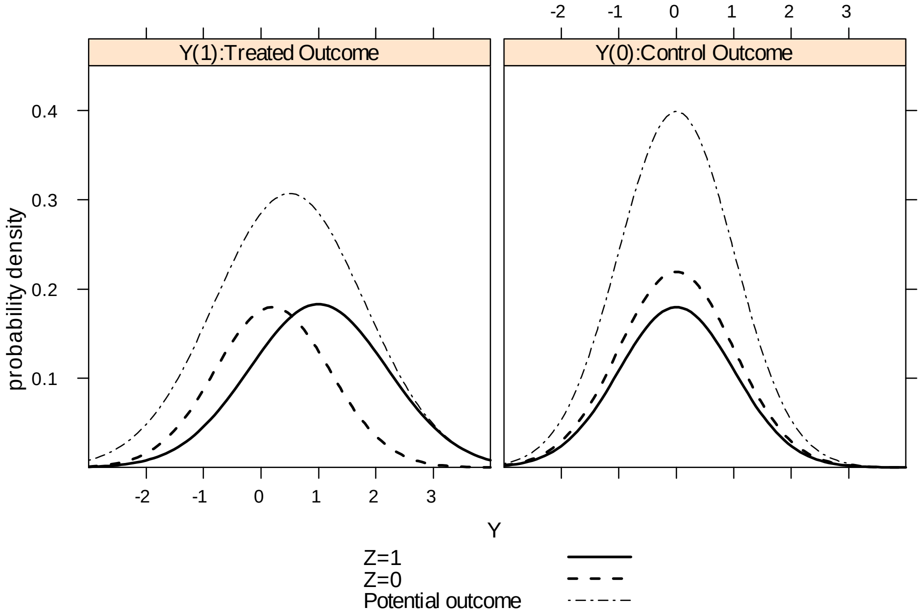

Figure 3: **Simulation of Test Power: Specification of Densities.** *The instrumental validity is refuted since for the treated outcomes the two observable densities intersect. Note that in each panel the density drawn to cover the other two represents the probability density of the potential outcomes.*

Table 2: Power against the Fixed Alternative

Monte Carlo iterations 2000, Bootstrap Iterations 500

	specification of $\mathbb{V}$					
	significance level					
	$\mathbb{V}_{hist}$ with binwidth 0.8			$\mathbb{V}_{hist}$ with binwidth 0.4		
	.10	.05	.01	.10	.05	.01
sample size	rejection probability			rejection probability		
(50,50)	.067	.033	.007	.062	.028	.006
(100,100)	.118	.068	.017	.071	.037	.009
(250,250)	.343	.227	.090	.234	.141	.045
(500,500)	.710	.595	.356	.521	.396	.189

In order to see finite sample power of our test procedure, we simulate the empirical rejection rate of the bootstrap test against a fixed alternative. The data generating process is specifed as

$$\begin{aligned} p(y,D &= 1) = 0.55 \times N(1,1.44), \qquad q(y,D=1) = 0.45 \times N(0.2,1) \\ p(y,D &= 0) = 0.45 \times N(0,1), \qquad\quad q(y,D=0) = 0.55 \times N(0,1). \end{aligned}$$

Figure 3 presents the densities of the specified data generating process. From this figure, we can observe that the instrumental validity is refuted by the configuration of the treated outcome densities since $p(y,D=1)$ intersects with $q(y,D=1)$. Table 2 presents the simulated rejection probabilities. We specify $\mathbb{V}$ as the histogram classes with the binwidth 0.8 or 0.4, the number of bins 12 or 24, and the set of initial breakpoints $\mathcal{Y}_0 = [-6.2,-5.4)$ or $\mathcal{Y}_0 = [-6.2,-5.8)$. For the specified alternative, we find that the simulated power is very poor in the small sample case. It is even lower than nominal size when $(m,n) = (50,50)$. The test procedure gains power for relatively large sample size $(m,n) = (500,500)$. We can also observe that $\mathbb{V}_{hist}$ with the shorter binwidth is less powerful than that with the wider binwidth. This can be explained that as the binwidth gets finer, the distribution of the test statistic under the least favorable null $P = Q$ has more variance and it raises the bootstrap critical values. This makes our test procedure less powerful. This suggests that given the finite sample there is a trade-off between the richness of $\mathbb{V}$, or equivalently, asymptotic refuting power and the finite sample power. Regardless of its practical importance in choosing $\mathbb{V}$, we make the choice of $\mathbb{V}$ out of scope of this paper and leave that as a part of future research.

4.2 Empirical Applications

We illustrate a use of the test procedure with using the following two data sets. The first one is the draft lottery data during Vietnam era used in Angrist (1991). The second one is from Card (1993) on returns to schooling using geographical proximity to college as an instrument.

4.2.1 Draft Lottery Data

The draft lottery data consist of a sample of 10,101 white men, born in 1950-1953. The data source is March Current Population Surveys of 1979 and 1981-1985. The outcome variable is measured in terms of the logarithm of weekly earnings imputed by the annual labor earnings divided by weeks worked. The treatment is whether one has a Vietnam veteran status or not. Since the enrollment for the military service possibly involves self-selection based on one's future earning, the veteran status is not considered to be randomly assigned. In order to solve this endogeneity issue, Angrist (1991) constructs the binary indicator of the draft eligibility, which is randomly assigned based on one's birthdate through the draft lotteries. A justification of the instrumental validity here is that the instrument is generated being independent of any individual characteristics. Hence, it is reasonable to argur that the instrument satisfies RTA. On the other hand, the validity of MPR is less credible since the existence of defiers are not eliminated by the sampling design, i.e., in the sample there are observations who participate to the military service even though they are not initially drafted.

The proposed testing procedure gives a solution to validate these assumptions from data. Figure 4 plots the kernel density estimates for the observed outcome distribution multiplied by the selection probability. We observe that the configuration of the densities in Figure 4 is similar to Figure 1. Therefore, we do not expect that the instrumental validity is refuted by the testing procedure. As Table 3 shows, p-value of the bootstrap test is almost one, and we do not refute the instrumental validity from the data.

4.2.2 Returns to Education: Proximity to College Data

The Card data is based on National Longitudinal Survey of Young Men (NLSYM) began in 1966 with age 14-24 men and continued with follow-up surveys through 1981. Based on the respondents' county of residence at 1966, the Card data provides the presence of a 4-year college in the local labor market. Observations of years of education and wage level are based on the follow-ups' educational attainment and wage level responded in the interview in 1976.

The idea of using the proximity to college as an instrument is stated as follows. Presence of a nearby college reduces a cost of college education by allowing students to live at home, while one's inherited ability is presumably independent of his birthplace. Compliers in this context can be considered to be those who grew up in relatively low-income families and who

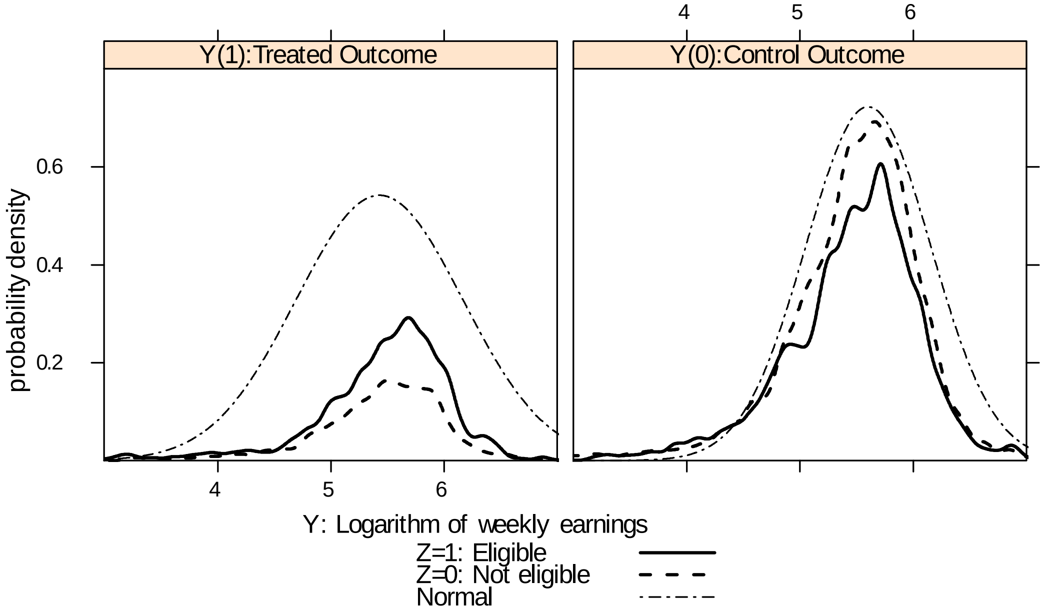

Figure 4: **Kernel Density Estimates for the Draft Lottery Data.** The *Gaussian kernel with bandwidth 0.06 is used. In each panel, we draw a normal density to illustrate the scale of the estimated densities.*

Table 3: Test Results of the Empirical Applications

Bootstrap iterations 500

	Draft lottery data		Proximity to college data			
	Full sample		Full sample		Restricted sample	
sample size (m,n)	(2780,7321)		(2053,957)		(1047,144)	
$\Pr(D=1\|Z=1)$, $\Pr(D=1\|Z=0)$	0.31, 0.19		0.29, 0.22		0.35,0.24	
$\mathbb{V}_{hist}$ binwidth	0.8	0.4	1.0	0.5	1.0	0.5
Bootstrap test, p-value	0.988	1.00	0.00	0.00	0.997	0.997

were not able to go to college without living with their parents. We make the educational level as a binary treatment which indicates one's education years to be greater or equal to 16 years. Roughly speaking, the treatment is considered as a four year college degree.

We specify the measure of outcome to be the logarithm of weekly earnings. In the first specification, we do not control any demographic covariates. This simplification raises a concern for the violation of RTA. For instance, one's region of residence, or whether they were born in the standard metropolitan area or rural area may affect one's wage levels and the proximity to colleges if the urban areas are more likely to have colleges and has higher wage level compared with the rural areas. This kind of confounder may contaminate the validity of RTA. In fact, Card (1993) emphasizes an importance of controlling for regions, residence in the urban area, race, job experience, and parent's education in order to make use of the college proximity as an instrument.

Figure 5 presents the kernel density estimates for observed oucome densities. In contrast to Figure 4, the kernel density estimates in Figure 5 intersect especially for those of the control outcomes. That is, the configuration of the densities are similar to Figure 2, and this indicates the violation of the instrument validity. Our test procedure yields zero p-value and this provides an empirical evidence that, without any covariates, college proximity is not a valid instrument.

We next look at how the test result changes once we control for some covariates. Controlling discrete covariates can be done by simply making the whole analysis conditional on the specified value of the covariates. We consider restricting the sample to be white workers (black dummy is zero), not living in south states in 1966 (south66 dummy is zero), and living in a metropolitan area in 1966 (SMSA66 dummy is one). That is, we are controlling for race, whether or not one grew up in southern states, and whether or not one grew up in urban area. The size of the restricted sample is 1191 ($m = 1047$, $n = 144$). Figure 6 indicates that the kernel density estimates do not reveal a clear evidence for a violation of the instrumental validity. This observation is also supported by the high p-value of the proposed test. Thus, we conclude that the instrumental validity is not refuted once we control for these covariates.

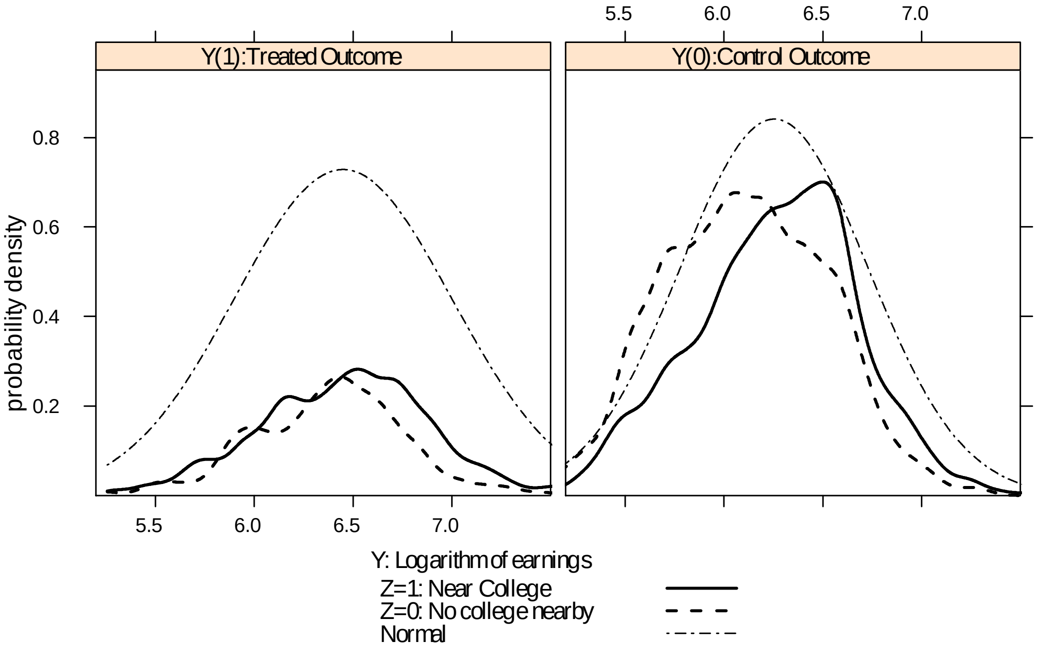

Figure 5: **Kernel Density Estimates for the Proximity to College Data (No covariates controlled)**. *The Gaussian kernel with bandwidth 0.07 is used. In each panel, we draw a normal density to illustrate the scale of the estimated densities.*

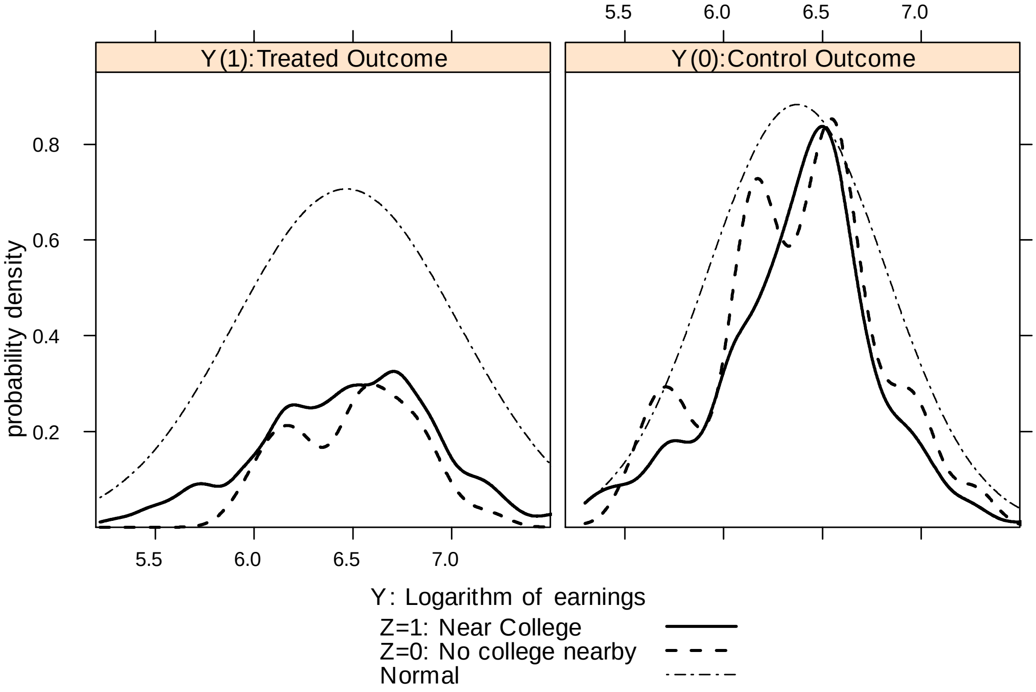

Figure 6: **Kernel Density Estimates for the Proximity to College Data (white workers, not living in south states, and living in a metropolitan area)**. *The Gaussian kernel with bandwidth 0.1 is used. In each panel, we draw a normal density to illustrate the scale of the estimated densities.*

5 Concluding Remarks

In this paper, we develop the bootstrap test procedure to empirically check the conditions of the instrumental validity of Imbens and Angrist (1994). Our testing strategy focuses on the nonnegativity of the complier's outcome densities that are point-identified when the instrument is valid. The nonnegativity of the complier's outcome density is equivalently expressed as the inequalities between the joint probability distributions of Y_{obs} and D_{obs} conditional on Z. We demonstrate that the inequalities provide the testable implication that has most refuting power. Our test statistic is designed to measure the discrepancy of these inequalities, and it has a form of the supremum statistic on the difference between the two empirical distributions over a specified VC-class of subsets. We develop the bootstrap algorithm to derive the critical values since the asymptotic distribution of the proposed statistic is not analytically tractable.

There are some issues left for future work. First of all, we do not formally investigate how to choose a VC-class $\mathbb{V}$ and how it affects the test performance in the finite sample case. We propose the two different choices of $\mathbb{V}$ in our simulation studies, the half unbounded interval class and the histrogram class. We observe that test size is not affected by the choice of $\mathbb{V}$ while test power is sensitive to the specification of $\mathbb{V}$.

Second, this paper exclusively considers the binary instrument case. When an instrument is multi-valued, but as long as its support is discrete, it is possible to test the instrument validity for every pair of two instrumental values. However, it is not clear what is a suitable test statistic when we want to test the instrument validity jointly over multiple instrument values. We leave a further discussion of the multi-valued instrument case for future work.

Appendix A: Proof of Proposition 1

Denote the population distribution of the types by $\pi_t \equiv \Pr(T = t)$, $t \in \{c, n, a, d\}$. Under RTA, $P(B, 1)$, for any Borel set $B \subset \mathcal{Y}$, is expressed as the following.

$$\begin{aligned}
P(B,1) &= \Pr(Y_{obs} \in B | D_{obs} = 1, Z = 1) \Pr(D_{obs} = 1 | Z = 1) \\
&= \left[\sum_{t \in \{c,n,a,d\}} \Pr(Y_1 \in V | D_1 = 1, Z = 1, T = t) \Pr(T = t | D_1 = 1, Z = 1) \right] \\
&\qquad \times \Pr(D_1 = 1 | Z = 1) \\
&= \Big[\sum_{t \in \{c,n,a,d\}} \Pr(Y_1 \in B | D_1 = 1, T = t) \Pr(T = t | T \in \{c, a\}) \Big] \Pr(T \in \{c, a\}) \\
&= \Big[\Pr(Y_1 \in B | T = a) \frac{\pi_a}{\pi_a + \pi_c} + \Pr(Y_1 \in V | T = c) \frac{\pi_c}{\pi_a + \pi_c} \Big] \\
&\qquad \times (\pi_a + \pi_c) \\
&= \Pr(Y_1 \in V | T = a) \pi_a + \Pr(Y_1 \in V | T = c) \pi_c. \qquad (6)
\end{aligned}$$

The second line follows by the law of total probability and the fact that the conditioning event $\{D_{obs} = 1, Z = 1\}$ is identical to $\{D_1 = 1, Z = 1\}$. To obtain the third line, we apply RTA to $\Pr(T = t | D_1 = 1, Z = 1)$, $\Pr(D_1 = 1 | Z = 1)$, and $\Pr(Y_1 \in B | D_1 = 1, Z = 1, T = t)$. Note that the type indicator T gives a finer partition of the sample space than D_1, so we obtain $\Pr(Y_1 \in B | D_1 = 1, T = t) = \Pr(Y_1 \in B | T = t)$ and $\Pr(T = t | D_1 = 1, Z = 1) = \Pr(T = t | T \in \{c, a\})$.

The similar operation to $Q(B, 1)$ yields

$$Q(B,1) = \Pr(Y_1 \in B | T = a) \pi_a + \Pr(Y_1 \in B | T = d) \pi_d. \qquad (7)$$

Under MPR, there do not exist defiers in the population, i.e., $\pi_d = 0$. If we take the difference between (6) and (7), we obtain

$$P(B,1) - Q(B,1) = \Pr(Y_1 \in B | T = c) \pi_c \geq 0.$$

This proves the first inequality of the proposition. The second inequality of the proposition is obtained in an analogous way and we omit its derivation for brevity.

For a proof of converse statement, let a data generating process P and Q satisfying the inequalities (1) be given. Let $p(y, d)$ and $q(y, d)$ be the densities (with respect to a dominating measure μ) of P and Q on $\mathcal{Y} \times \{d\}$. It suffices to show that we can construct a joint distribution of (Y_1, Y_0, T, Z) that is compatible with P and Q and satisfies RTA and MPR. Since the marginal distribution of Z is not important for the analysis, we focus on constructing the conditional distribution of (Y_1, Y_0, T) given Z. Let us consider the nonnegative functions $h_{Y_d,t}(y)$, $d = 1, 0$, $t \in \{c, n, a, d\}$,

$$\begin{aligned}
h_{Y_1,c}(y) &= p(y,1) - q(y,1), \\
h_{Y_1,n}(y) &= \gamma_{Y_1}(y), \\
h_{Y_1,a}(y) &= q(y,1), \\
h_{Y_1,d}(y) &= 0, \\
h_{Y_0,c}(y) &= q(y,0) - p(y,0), \\
h_{Y_0,n}(y) &= p(y,0), \\
h_{Y_0,a}(y) &= \gamma_{Y_0}(y), \\
h_{Y_0,d}(y) &= 0.
\end{aligned}$$

where $\gamma_{Y_1}(y)$ and $\gamma_{Y_0}(y)$ are arbitrary nonnegative functions satisfying $\int_{\mathcal{Y}} \gamma_{Y_1}(y) d\mu = P(\mathcal{Y}, 0)$ and $\int_{\mathcal{Y}} \gamma_{Y_0}(y) d\mu = Q(\mathcal{Y}, 1)$. We construct a conditional probability law of (Y_1, Y_0, T) given Z as, for an arbitrary Borel sets B_1 and B_0 in $\mathcal{Y}$,

$$\begin{aligned}
& \Pr(Y_1 \in B_1, Y_0 \in B_0, T = c | Z = 1) = \Pr(Y_1 \in B_1, Y_0 \in B_0, T = c | Z = 0) \\
\equiv \; & \begin{cases} \frac{\int_{B_1} h_{Y_1,c}(y) d\mu}{\int_{\mathcal{Y}} h_{Y_1,c}(y) d\mu} \times \frac{\int_{B_0} h_{Y_0,c}(y) d\mu}{\int_{\mathcal{Y}} h_{Y_0,c}(y) d\mu} \times [P(\mathcal{Y},1) - Q(\mathcal{Y},1)] & \text{if } [P(\mathcal{Y},1) - Q(\mathcal{Y},1)] > 0 \\ 0 & \text{if } [P(\mathcal{Y},1) - Q(\mathcal{Y},1)] = 0 \end{cases} \\
& \Pr(Y_1 \in B_1, Y_0 \in B_0, T = n | Z = 1) = \Pr(Y_1 \in B_1, Y_0 \in B_0, T = n | Z = 0) \\
\equiv \; & \begin{cases} \frac{\int_{B_1} h_{Y_1,n}(y) d\mu}{\int_{\mathcal{Y}} h_{Y_1,n}(y) d\mu} \times \frac{\int_{B_0} h_{Y_0,n}(y) d\mu}{\int_{\mathcal{Y}} h_{Y_0,n}(y) d\mu} \times P(\mathcal{Y},0) & \text{if } P(\mathcal{Y},0) > 0 \\ 0 & \text{if } P(\mathcal{Y},0) = 0 \end{cases} \\
& \Pr(Y_1 \in B_1, Y_0 \in B_0, T = a | Z = 1) = \Pr(Y_1 \in B_1, Y_0 \in B_0, T = a | Z = 0) \\
\equiv \; & \begin{cases} \frac{\int_{B_1} h_{Y_1,a}(y) d\mu}{\int_{\mathcal{Y}} h_{Y_1,a}(y) d\mu} \times \frac{\int_{B_0} h_{Y_0,a}(y) d\mu}{\int_{\mathcal{Y}} h_{Y_0,a}(y) d\mu} \times Q(\mathcal{Y},1) & \text{if } Q(\mathcal{Y},1) > 0 \\ 0 & \text{if } Q(\mathcal{Y},1) = 0 \end{cases} \\
& \Pr(Y_1 \in B_1, Y_0 \in B_0, T = d | Z = 1) = \Pr(Y_1 \in B_1, Y_0 \in B_0, T = d | Z = 0) \\
\equiv \; & 0
\end{aligned}$$

Note that this is a valid probability measure since it is nonnegative and satisfies

$$\sum_{t \in \{c,n,a,d\}} \Pr(Y_1 \in \mathcal{Y}, Y_0 \in \mathcal{Y}, T = t | Z = z) = 1, \quad z = 1, 0.$$

Furthermore, the proposed probability distribution satisfies RTA and MPR by construction and it is consistent with the given data generating process, i.e.,

$$\begin{aligned}
\Pr(Y_{obs} \in B, D_{obs} = 1|Z = 1) &= \Pr(Y_1 \in B, T = c|Z = 1) + \Pr(Y_1 \in B, T = a|Z = 1) \\
&= \int_B [h_{Y_1,c}(y) + h_{Y_1,a}(y)]d\mu = P(B, 1), \\
\Pr(Y_{obs} \in B, D_{obs} = 0|Z = 1) &= \Pr(Y_0 \in B, T = n|Z = 1) + \Pr(Y_0 \in B, T = d|Z = 1) \\
&= P(B, 0) \\
\Pr(Y_{obs} \in B, D_{obs} = 1|Z = 0) &= \Pr(Y_1 \in B, T = a|Z = 0) + \Pr(Y_1 \in B, T = d|Z = 0) \\
&= Q(B, 1) \\
\Pr(Y_{obs} \in B, D_{obs} = 0|Z = 0) &= \Pr(Y_0 \in B, T = n|Z = 0) + \Pr(Y_0 \in B, T = c|Z = 0) \\
&= Q(B, 0)
\end{aligned}$$

This completes the proof. ■

Appendix B: Proof of Proposition 2

Throughout the proof, it is assumed that the probability law of a binary instrument Z is i.i.d Bernoulli with parameter $\lambda \in (\epsilon, 1-\epsilon)$ for some $\epsilon > 0$.

i)
Step 1: Derive the asymptotic distribution of the test statistic T_N under the null $P = Q$.

Define P_m and Q_n as the empirical probability measure of (Y, D) conditional on $Z = 1$ and $Z = 0$ respectively,

$$P_m = \frac{1}{m}\sum_{i=1}^{m} \delta_{(Y^1_{obs,i}, D^1_{obs,i})}, \qquad Q_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{(Y^0_{obs,j}, D^0_{obs,j})},$$

where $\delta_{(y,d)}$ represents a unit mass measure on $(Y_{obs}, D_{obs}) = (y, d)$.

Given $\mathbb{V}$ a VC-class of subsets in $\mathbb{R}$, we define the class of indicator functions on $\mathbb{R} \times \{1, 0\}$, $\mathcal{F}_1$ and $\mathcal{F}_0$,

$$\mathcal{F}_1 = \{1\{(V, 1)\}; V \in \mathbb{V}\}, \qquad \mathcal{F}_0 = \{1\{(V, 0)\}; V \in \mathbb{V}\}$$

where the first coordinate of the indicator function corresponds to a subset $V \subset \mathbb{R}$ and the second coordinate corresponds to the participation indicator D_{obs}. Following to the notation in van der Vaart and Wellner (1996), for a function $f : \mathbb{R} \times \{1, 0\} \to \mathbb{R}$, Pf stands for the expectation of f with respect to P, $Pf = \int f dP$. Note that $\mathcal{F}_1$ and $\mathcal{F}_0$ are VC-class of functions on $\mathbb{R} \times \{1, 0\}$ since the collection of subsets $\mathbb{V}$ are assumed to be a VC-class.

Consider stochastic processes $G_{1,N} : \mathcal{F} \to \mathbb{R}$ where $\mathcal{F}$ is a class of functions on $\mathbb{R} \times \{1, 0\}$,

$$\begin{aligned} G_{1,N}(\cdot) &= \left(\frac{mn}{N}\right)^{1/2} (Q_n - P_m) \\ &= \left(\frac{m}{N}\right)^{1/2} \sqrt{n}(Q_n - Q) - \left(\frac{n}{N}\right)^{1/2} \sqrt{m}(P_m - P) \\ &\quad + \left(\frac{mn}{N}\right)^{1/2} (Q - P). \end{aligned} \tag{8}$$

Given the above Donsker class of functions $\mathcal{F}_1$, we apply the Donsker theorem (theorem 3.5.1 in van der Vaart and Wellner (1996)) to get the weak convergence of $\sqrt{n}(Q_n - Q)(\cdot)$ and $\sqrt{m}(P_m - P)(\cdot)$ to the brownian bridges on $\mathcal{F}_1$,

$$\begin{aligned} &\sqrt{n}(Q_n - Q) \rightsquigarrow G_Q \quad \text{in} \quad l^\infty(\mathcal{F}_1) \\ &\sqrt{m}(P_m - P) \rightsquigarrow G_P \quad \text{in} \quad l^\infty(\mathcal{F}_1) \end{aligned}$$

where "$\rightsquigarrow$" notates weak convergence, G_P represents the P-brownian bridge, G_Q represents the Q-brownian bridge, and $l^\infty(\mathcal{F})$ denotes the space of l^∞ functions which map from $\mathcal{F}$ into $\mathbb{R}$. Under the null $P = Q$, since $m/N \to \lambda$ almost surely, $G_{1,N}$ converges weakly to a sum of two independent P-brownian bridges G_P and G'_P.

$$G_{1,N} \rightsquigarrow \lambda^{1/2} G_P - (1 - \lambda)^{1/2} G'_P.$$

Note that the probability law of the process $\lambda^{1/2} G_P - (1 - \lambda)^{1/2} G'_P$ is identical to the P-brownian bridge G_P. Hence, we have $G_{1,N} \rightsquigarrow G_P$ in $l^\infty(\mathcal{F}_1)$. Analogously, for stochastic processes $G_{0,N} : \mathcal{F} \to \mathbb{R}$

$$\begin{aligned} G_{0,N} &= \left(\frac{mn}{N}\right)^{1/2} (P_m - Q_n) \\ &= \left(\frac{n}{N}\right)^{1/2} \sqrt{m}(P_m - P) - \left(\frac{m}{N}\right)^{1/2} \sqrt{n}(Q_n - Q) \\ &\quad + \left(\frac{mn}{N}\right)^{1/2} (P - Q). \end{aligned} \tag{9}$$

we obtain $G_{0,N} \rightsquigarrow G_P$ in $l^\infty(\mathcal{F}_0)$.

Notice that the test statistic is written as

$$T_N = \max \left\{ \sup_{f \in \mathcal{F}_1} G_{1,N} f, \sup_{f \in \mathcal{F}_0} G_{0,N} f \right\}.$$

Let $\mathcal{F}^*= \mathcal{F}_1 \cup \mathcal{F}_0$. Note that $\mathcal{F}^*$ is also a Donsker class. For $X \in l^\infty(\mathcal{F}^*)$ with $l^\infty(\mathcal{F}^*)$ equipped with the sup metric, the functional $\sup_{f \in \mathcal{F}_1} Xf$ is continuous with respect to X, since for $X_1, X_2 \in l^\infty(\mathcal{F}^*)$, $|\sup_{f \in \mathcal{F}_1} (X_1 - X_2) f| \leq \sup_{f \in \mathcal{F}^*} |(X_1 - X_2) f| \acute{\leq} \|X_1 - X_2\|$ holds. Since the max operator is clearly continuous, the continuous mapping theorem for stochastic processes (see, e.g., Pollard (1984)) implies

$$T_N \rightsquigarrow T = \max \left\{ \sup_{f \in \mathcal{F}_1} G_P f, \sup_{f \in \mathcal{F}_0} G_P f \right\} = \sup_{f \in \mathcal{F}^*} G_P f. \tag{10}$$

This is the limiting probability law of T_N under the null $P = Q$.

Step 2: Prove the asymptotic consistency of the distribution of the bootstrap statistic.

Let us define the bootstrap empirical measure

$$P_m^* = \frac{1}{m}\sum_{i=1}^{m} \delta_{(Y_{obs,i}^*, D_{obs,i}^*)}, \qquad Q_n^* = \frac{1}{n}\sum_{j=1}^{n} \delta_{(Y_{obs,j}^*, D_{obs,j}^*)}.$$

where $(Y_{obs,i}^*, D_{obs,i}^*)$, $i = 1, \ldots, m$, and $(Y_{obs,j}^*, D_{obs,j}^*)$, $j = 1, \ldots, n$, are drawn randomly from the pooled empirical measure

$$H_N = \frac{m}{N} P_m + \frac{n}{N} Q_n.$$

The bootstrap test statistic is expressed as

$$T_N^* = \max\left\{ \sup_{f \in \mathcal{F}_1} G_{1,N}^* f, \sup_{f \in \mathcal{F}_0} G_{0,N}^* f \right\}$$

where $G_{1,N}^* = \left(\frac{mn}{N}\right)^{1/2} (Q_n^* - P_m^*)$ and $G_{0,N}^* = \left(\frac{mn}{N}\right)^{1/2} (P_m^* - Q_n^*)$. The bootstrap consistency is proved if the distribution of T_N^* converges weakly to the one obtained in (10) under the null $P = Q$ for almost every sampling sequences of $\{(Y_{obs,i}^1, D_{obs,i}^1)\}$ and $\{(Y_{obs,j}^0, D_{obs,j}^0)\}$.

Let $H = \lambda P + (1 - \lambda)Q$. By theorem 3.7.7 in van der Vaart and Wellner (1996), $\sqrt{m}(P_m^* - H_N) \rightsquigarrow G_H$ and $\sqrt{n}(Q_n^* - H_N) \rightsquigarrow G_H$ hold with probability one in terms of the randomness of the sequences, $\{(Y_{obs,i}^1, D_{obs,i}^1)\}$ and $\{(Y_{obs,j}^0, D_{obs,j}^0)\}$.

Thus, by the similar argument to Step 1, $G_{1,N}^*$ and $G_{0,N}^*$ weakly converge to the H-brownian bridge, i.e.,

$$\begin{aligned} G_{1,N}^* &= \left(\frac{mn}{N}\right)^{1/2} (Q_n^* - P_m^*) \\ &= \left(\frac{m}{N}\right)^{1/2} \sqrt{n}(Q_n^* - H_N) - \left(\frac{n}{N}\right)^{1/2} \sqrt{m}(P_m^* - H_N) \\ &\rightsquigarrow \lambda^{1/2} G_H - (1 - \lambda)^{1/2} G'_H = G_H \end{aligned}$$

and $G_{0,N}^* \rightsquigarrow G_H$ for almost every sequence of $\{(Y_{obs,i}^1, D_{obs,i}^1)\}$ and $\{(Y_{obs,j}^0, D_{obs,j}^0)\}$. Therefore, by the continuous mapping theorem,

$$T_N^* \rightsquigarrow \sup_{f \in \mathcal{F}^*} G_H f. \tag{11}$$

Note that, under the null, $H = P$ holds, and therefore the obtained H-brownian bridge is in fact P-brownian bridge. Hence, $T_N^* \rightsquigarrow T$ holds. This implies that the asymptotic distribution of T_N^* coincides with that of T_N under the null for almost every sequence of $\{(Y_{obs,i}^1, D_{obw,i}^1)\}$ and $\{(Y_{obs,j}^0, D_{obs,j}^0)\}$.

Step 3: Prove the asymptotic consistency of the rejection probability based on the bootstrap critical value $\hat{c}_{boot}(1 - \alpha)$.

Let $J_N(\cdot, H_N)$ be the cdf of the bootstrap statistic T_N^* (conditional on H_N). The bootstrap estimates of the critical value is the $(1-\alpha)$-th quantile of $J_N(\cdot, H_N)$, that is,

$$\hat{c}_{boot}(1-\alpha) \equiv \inf\left\{c : \mathrm{Prob}_{H_N}(T_N^* > c) \leq \alpha\right\}.$$

Let $J(\cdot, H)$ be the cdf of T under the null $P = Q(= H)$ and denote its $(1-\alpha)$-th quantile by $c(1-\alpha)$. Since $J_N(\cdot, H_N)$ converges weakly to $J(\cdot, H)$, $\hat{c}_{boot}(1-\alpha)$ converges to the $c(1-\alpha)$ if $J(\cdot, H)$ is continuous and strictly increasing at its $(1-\alpha)$-th quantile (see, e.g., Lemma 1.2.1. in Politis, Romano, and Wolf (1999)).

The absolute continuity of $J(\cdot, H)$ follows by the absolute continuity theorem for the convex functional of the Gaussian processes (Theorem 11.1 of Davydov, Lifshits, and Smorodina (1998)). Note that the test statistic is a convex functional of $l^\infty(\mathcal{F}^*)$, and for some $f \in \mathcal{F}^*$ with nondegenerate $G_P f$, it holds $\Pr(T \leq 0) \leq \Pr(Gf \leq 0) = 1/2$. Therefore, $J(t, H)$ is absolutely continuous for every $t > 0$. Then, the absolute continuity theorem guarantees that, for $\alpha \in (0, 1/2)$, $J(t, H)$ is absolutely continuous at $c(1-\alpha)$. Thus, $\hat{c}_{boot}(1-\alpha) \to c(1-\alpha)$ almost surely in terms of the randomness of H_N.

Finally, by the Slutsky's Theorem, it follows

$$Prob_{P=Q=H}(T_N > \hat{c}_{boot}(1-\alpha)) \to 1 - J(c(1-\alpha), H) = \alpha.$$

ii)

To examine power of the test against a fixed alternative, consider P and Q such that $(Q-P)f > 0$ for some $f \in \mathcal{F}_1$. Then, the last term in (8) diverges to positive infinitiy at these f. Since the Brownian bridge processes as the limiting process of $\sqrt{m}(P_m - P)$ and $\sqrt{n}(Q_n - Q)$ are bounded with probability one, $\sup_{f \in \mathcal{F}_1} G_{1,N} f \to \infty$ with probability one. This implies $T_N \to \infty$ with probability one.

On the other hand, the bootstrap critical value are bounded almost surely (with respect to the original sampling sequence) because T_N^* weakly converges to $\sup_{f \in \mathcal{F}^*} G_H f$ with $H = \lambda P + (1-\lambda)Q$. Therefore,

$$Prob_{P=Q=H}(T_N > \hat{c}_{boot}(1-\alpha)) \to 1$$

as $N \to \infty$.

■

References

[1] Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97, 284-292.

[2] Abadie, A., J. D. Angrist, and G. W. Imbens. (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91-117.

[3] Angrist, J. D. (1991): "The Draft Lottery and Voluntary Enlistment in the Vietnam Era," *Journal of the American Statistical Association,* 86, 584-595

[4] Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-455.

[5] Campbell, D. T. (1969): "Reforms as Experiments," *American Psychologist*, 24 (4), 409-429.

[6] Card, D. (1993): "Using Geographical Variation in College Proximity to Estimate the Returns to Schooling", Natinoal Bureau of Economic Research Working Paper No. 4, 483.

[7] Davydov, Y.A., Lifshits, M.A., and Smorodina, N.V. (1998): *Local Propoerties of Distributions of Stochastic Functionals.* Providence: American Mathematical Society.

[8] Dudley, R. M. (1999): *Uniform Central Limit Theorem.* Cambridge University Press.

[9] Hahn, J., Todd, P. E., and Van der Klaauw, W. (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica* 69 (1), 201-209.

[10] Heckman, J. J. and E. Vytlacil (1999): "Local Instrumental Variables and Latent Variables Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96, 4730-4734.

[11] Heckman, J. J. and E. Vytlacil (2001): "Local Instrumental Variables," in C. Hsiao, K. Morimune, and J. Powell editors, *Nonlinear Statistical Model: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya,* 1-46. Cambridge University Press, Cambridge UK.

[12] Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-475.

[13] Imbens, G. W. and D. B. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variable Models," *Review of Economic Studies*, 64, 555-574.

[14] Kling, J. R., J. B. Liebman, and L. F. Katz (2007): "Experimental Analysis of Neighborhood Effects," *Econometrica,* 75, 83-119.

[15] Romano, J. P. (1988): "A Bootstrap Revival of Some Nonparametric Distance Tests." *Journal of American Statistical Association*, 83, 698-708.

[16] Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling.* New York: Springer.

[17] van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics,* New York: Springer.