A Discontinuity Test for Identification in Nonlinear Models with Endogeneity

Carolina Caetano, Christoph Rothe, and Nese Yildiz*

PRELIMINARY AND INCOMPLETE

Abstract

In this paper, we consider a triangular system of equations with a potentially endogenous variable whose distribution has a mass point at the lower boundary of its support, but is otherwise continuous. We show that, together with a weak continuity condition on the structural function, this setup yields a testable implication of the set of assumptions that is commonly used in this class of models to achieve identification of various structural quantities through a control variable approach.

JEL Classification: C12, C14, C31, C36, C52

Keywords: *Nonseparable models, triangular systems, control variables, specification testing, identification.*

*This Version: March 5, 2013. Caetano: Department of Economics, University of Rochester, 238 Harkness Hall, P.O. Box: 270156, Rochester, NY 14627, Email: carol.caetano@rochester.edu. Rothe: Department of Economics, Columbia University, 420 W 118th St., New York, NY 10027, Email: cr2690@columbia.edu. Yildiz: Department of Economics, University of Rochester, 238 Harkness Hall, P.O. Box: 270156, Rochester, NY 14627, Email: nese.yildiz@rochester.edu.

1. Introduction

Triangular systems with nonseparable disturbances have recently received considerable attention in econometrics. In these models, nonparametric identification of structural quantities if often achieved via a control variable approach. That is, one imposes assumptions on the primitives of the model that imply that there exists a variable that, when conditioned on, makes covariates and disturbances independent, and that such a control variable can be identified from observable quantities. For example, Imbens and Newey (2009) consider the triangular system

$$Y = m_1(X, U),$$
$$X = m_2(Z, V),$$

where Y is an outcome variable, X is an endogenous covariate, Z is an instrument, and U and V denote unobserved heterogeneity. They show that in this model $V^* = F_{X|Z}(X; Z)$ is a control variable (in the sense that $U \perp X|V^*$), if $Z \perp (U, V)$, V is a scalar, and m_2 is strictly monotonic in V. These conditions are not innocuous, and imply substantial restrictions on the underlying economic model. Having a method to check their validity would thus be of great importance in empirical applications.

In a general triangular model, the conditions that are necessary to justify a control variable approach have no testable implications. In this paper, we argue that the situation is different if the endogenous covariate X has a mass point at the lower (or upper) boundary of its support, is otherwise continuously distributed, and exerts a continuous effect on the outcome variable of interest. Our main contribution is to show that in such a setting it is possible to test the validity of the control variable approach to identification by checking whether a certain conditional expectation function is continuous at one particular point. To the best of our knowledge, our paper is the first to propose a test for this type of hypothesis.

The idea behind our test is related to that in Caetano (2012), who showed that endogeneity of a covariate with the above-mentioned properties can be detected through a discontinuity in the conditional expectation of the outcome variable given the covariate. If the true causal effect of the covariate is continuous, then additional conditioning on a control variable should remove this discontinuity. If the discontinuity remains, how-

ever, one has to conclude that the control variable approach is invalid, and some of the assumptions that were made to justify it have to be violated.

Our paper proposes a test statistic for the null hypothesis of validity of the control variable approach based on the strategy we just laid out. Its computation is straightforward, as it involves only standard nonparametric regression and density estimation techniques, and a simple numerical integration step. The test statistic is asymptotically normal and converges at the one-dimensional nonparametric rate under standard conditions. We give an explicit formula for its asymptotic variance, which can be estimated to obtain critical values. Deriving the theoretical properties of our test statistic is a non-standard problem as it involves running nonparametric regressions on estimated data points (the control variable is unobserved, and has to be estimated from the data). To account for this two-stage structure, we use recent results in Mammen, Rothe, and Schienle (2012, 2013) on generated covariates in nonparametric models.

Requiring the endogenous variable to be bounded from below (or above) and to have a mass point at the lower (or upper) boundary of its support obviously restricts the number of empirical applications in which our testing procedure can be applied. Yet there are many variables with such a property that appear frequently as potentially endogenous covariates in empirical applications. For example, weekly hours of work have to be non-negative, and a sizable fraction of the population does not to work. Hourly wages cannot be lower than the local minimum wage, and a sizable fraction of the population earns exactly the legal minimum. The amount of consumption of some product also has to be non-negative, and a sizable fraction of the population might not consume this product at all. Many other examples are easily constructed. Our test should therefore be useful for a wide range of empirical settings.

Our paper contributes to an extensive literature on identification in nonlinear models with endogeneity. Control variable methods for non- and semi-parametric triangular models are studied by Newey, Powell, and Vella (1999), Blundell and Powell (2003, 2004), Imbens (2007), Imbens and Newey (2009), Rothe (2009) and Kasy (2011), among others. Kasy (2013) considers identification in triangular systems under monotonicity restrictions on the instrument. Instrumental variable (IV) approaches to identification in nonparametric models with additive disturbances are studied in Newey and Powell (2003), Hall

and Horowitz (2005), Blundell, Chen, and Kristensen (2007), or Darolles, Fan, Florens, and Renault (2011). Chernozhukov and Hansen (2005) and Chernozhukov, Imbens, and Newey (2007) consider IV methods in nonseparable models, but with restrictions on the dimension of the disturbances. Canay, Santos, and Shaikh (2012) show that the completeness condition, which is necessary for nonparametric IV approaches, is generally not testable.

The remainder of the paper is structured as follows. In Section 2, we formally introduce our model, review the control variable approach to identification, and explain the testing problem. In Section 3 we describe our testing approach, explain the test statistic, and derive its asymptotic properties. Numerical properties are studied in Section 4, and Section 5 concludes. All proofs and several extensions are collected in the appendix.

2. Model, Identification Strategy, and Testing Problem

2.1. Model. In this paper, we consider a triangular system of simultaneous equations similar to the one studied in e.g. Imbens and Newey (2009). For simplicity, the system only contains a single potentially endogenous variable, and no additional exogenous covariates.[1] The main difference of our setup relative to standard existing ones is that we consider the case of an endogenous variable whose distribution has a mass point at the lower bound of its support. Specifically, our model is given by

$$Y = m_1(X, U), \tag{2.1}$$

$$X = \max\{c, m_2(Z, V)\}, \tag{2.2}$$

where Y is the outcome of interest, X is a scalar and potentially endogenous covariate, U and V denote unobserved heterogeneity, and c is a known constant. We assume that $0 < \Pr(m_2(Z, V) < c) < 1$, so that the distribution of X has an actual mass point at c, but is not degenerate. Such a model could be reasonable in settings where natural or legal restrictions might lead to corner solutions in the individuals' optimization problem that determines the choice of X. Examples for such variables X include weekly hours of work (which have to be non-negative), hourly wages (which have to exceed the minimum

[1] We study an extension that allows for additional covariates in the Appendix.

wage), or the amount of consumption of a particular good (which again has to be non-negative). For the remainder of this paper, we take the threshold c to be zero, which can be done without loss of generality by subtracting c from both X and $m_2(Z, V)$.

2.2. Identification Strategy. There are many structural quantities in models like (2.1)–(2.2) that could potentially be of interest in applications, including features of the distribution of $m_1(x, U)$, such as the Average Structural Function $a(x) = \mathbb{E}(m_1(x, U))$ (cf. Blundell and Powell, 2003). Since (2.1)–(2.2) involves an explicit characterization of the way the endogenous variable X is generated, such structural quantities can be identified through a control variable approach. Under appropriate exogeneity conditions on the instruments Z, the only source for dependence between U and X is their joint dependence on V. Under further conditions on the model, an estimate of a one-two-one transformation of V can be recovered from the data, and endogeneity of X can be controlled by conditioning on that estimate. While our model differs slightly from the one in Imbens and Newey (2009) due to the presence of a endogenous variable with a mass point, the conditions needed to establish existence of a control variable are very similar.

Assumption 1. *The model in* (2.1)–(2.2) *satisfies the following restrictions:*

(i) Z *and* (U, V) *are stochastically independent.*

(ii) V *is scalar and continuously distributed, and without loss of generality its distribution is normalized such that* $V \sim U[0, 1]$*.*

(iii) The function $v \mapsto m_2(Z, v)$ *is strictly increasing with probability 1.*

Assumption 1 is analogous to the conditions of Imbens and Newey (2009), and is central for the validity of control variable arguments. It implies that that $U \perp X|V$, and that V is identified in the subpopulation with $X > 0$. In particular, we have that $V\mathbb{I}\{X > 0\} = V^*$ where $V^* = F_{X|Z}(X; Z)$. Taken together, these two statements imply that

$$U \perp X|(V^*, X > 0), \tag{2.3}$$

which shows that V^* is a control variable in the subpopulation with $X > 0$. The existence of such a control variable implies identification of Average Structural Function, and other structural features of the model (2.1)–(2.2), under a support and a continuity condition.

Assumption 2. *The model in* (2.1)–(2.2) *satisfies the following restrictions:*

(i) The support $supp(V^*|X = x)$ *of* V^* *conditional on* $X = x$ *is equal to* $[0, 1]$ *for all* $x \in supp(X|X > 0)$.

(ii) The function $x \mapsto \mathbb{E}(m_1(x, U)|V)$ *is continuous at* $x = 0$ *with probability 1.*

Assumption 2(i) is a generalization of the usual rank condition in the linear simultaneous equations model. It can be shown to be satisfied if the function $z \mapsto m_2(z, V)$ exhibits a sufficient amount of variation over the support of Z. Assumption 2(ii) is a continuity condition on the structural function m_1, which can often be motivated through subject knowledge. Such a condition would trivially be satisfied if the mapping $x \mapsto m_1(x, U)$ is continuous at $x = 0$ with probability 1. Taken together, Assumption 1–2 imply the following proposition.

Proposition 1. *(i) Suppose that Assumption 1–2*(i) *hold. Then the ASF* $a(x)$ *is identified for all* $x \in supp(X|X > 0)$. *(ii) Suppose that Assumption 1– 2 hold. Then the ASF* $a(x)$ *is identified for all* $x \in supp(X)$.

This result is a minor extension of Imbens and Newey (2009, Theorem 1). To see that it is true, note that it follows from (2.3) that $\mathbb{E}(Y|X = x, V^* = v) = \mathbb{E}(m_1(x, U)|V = v)$ for all $x > 0$ and all v. Identification of the average structural function for $x > 0$ then follows because $a(x) = \int_0^1 \mathbb{E}(m_1(x, U)|V = v)dv$, which is well defined because of Assumption 2(i). Moreover, Assumption 2(ii) then implies identification of the average structural function at $x = 0$, because $a(0) = \int_0^1 \lim_{x \downarrow 0} \mathbb{E}(m_1(x, U)|V = v)dv = \lim_{x \downarrow 0} a(x)$.

2.3. Testing Problem. In a general triangular model, conditions like those in Assumption 1, which are necessary to justify a control variable approach, typically have no testable implications. In this paper, we argue that the situation is different in a model like (2.1)–(2.2), where the endogenous covariate has a mass point at the lower (or upper) boundary of its support. We show that, together with the continuity condition on the

structural function in Assumption 2(ii), this setup yields testable implications that take the form of a continuity condition on a particular conditional expectation function. That is, we show that it is possible to test the null hypothesis

$$H_0: \text{ Assumption 1 holds} \quad \text{vs.} \quad H_1: \text{ Assumption 1 is violated,} \tag{2.4}$$

if Assumption 2(ii) is maintained, and propose a practically feasible testing procedure. This is the main contribution of our paper. Maintaining Assumption 2(ii) is reasonable, as the assumption of continuous causal effects is arguable a more plausible restriction in many applications than exclusion restrictions on the instruments, or monotonicity and dimensionality restrictions on unobservables. Also note that Assumption 2(i) only involves observable quantities, and is thus directly testable.

3. Testing the Validity of a Control Variable Approach

3.1. Testing Approach. To motivate our approach to testing (2.4), consider the conditional expectation of Y given X and V. Due to the structure of the model (2.1)–(2.2), we have that

$$\mathbb{E}(Y|X=x,V=v) = \begin{cases} \mathbb{E}(m_1(x,U)|m_2(Z,v)=x, V=v) & \text{if } x>0, \\ \mathbb{E}(m_1(x,U)|m_2(Z,v)\leq x, V=v) & \text{if } x=0. \end{cases}$$

Since the conditioning sets in the two conditional expectations on the right-hand side of the previous equation differ, we would in general expect the function $x \mapsto \mathbb{E}(Y|X=x,V=v)$ to be discontinuous at $x=0$ for at least some (and potentially all) $v \in [0,1]$. Under the null hypothesis, however, we have that $U \perp X|V$, and thus

$$\mathbb{E}(Y|X=x,V=v) = \mathbb{E}(m_1(x,U)|V=v),$$

which is continuous at $x=0$ for all $v \in [0,1]$ by Assumption 2(ii). This means that if V was identified, we could test our null hypothesis by checking whether

$$\lim_{x \downarrow 0} \mathbb{E}(Y|X=x,V=v) - \mathbb{E}(Y|X=0,V=v) = 0 \text{ for all } v \in (0,1). \tag{3.1}$$

Unfortunately, under the null we are only able to identify the censored variable $V\mathbb{I}\{X>0\}$, but not V itself, and thus this approach is not feasible: while identification of $V\mathbb{I}\{X>0\}$

is sufficient for identifying the first term on the left-hand side of (3.1), it does not suffice for learning $\mathbb{E}(Y|X=0,V=v)$. We therefore consider testing (2.4) by checking a necessary condition for (3.1), namely that

$$\Delta = 0, \tag{3.2}$$

where

$$\begin{aligned}\Delta &:= \int (\lim_{x\downarrow 0} \mathbb{E}(Y|X=x,V=v) - \mathbb{E}(Y|X=0,V=v)) F_{V|X}(dv|0) \\ &= \int \lim_{x\downarrow 0} \mathbb{E}(Y|X=x,V=v) F_{V|X}(dv|0) - \mathbb{E}(Y|X=0).\end{aligned}$$

Such an approach is indeed feasible: the parameter Δ is identified because computation of $\mathbb{E}(Y|X=0)$ and $\lim_{x\downarrow 0}\mathbb{E}(Y|X=x,V=v) = \lim_{x\downarrow 0}\mathbb{E}(Y|X=x,V^*=v)$ only involves the joint distribution of observable and/or identified random variables, and because the conditional CDF $F_{V|X}(\cdot,0)$ of V given $X=0$ is identified. The latter point might seem surprising at first: while we cannot identify V in the subpopulation with $X=0$, we can identify its distribution. To see why that is the case, note that by Assumption 1(ii) the marginal distribution of V is normalized to be the uniform distribution on the interval $[0,1]$, and thus its unconditional CDF F_V is known. It then follows from the Law of Total Probability and the fact that $V\mathbb{I}\{X>0\} = V^*$ that

$$F_{V|X}(v|0) = \frac{1}{\Pr(X=0)} \left(F_V(v) - \Pr(V^* \leq v, X>0)\right), \tag{3.3}$$

and the term on the right-hand side of (3.3) is clearly identified. To test our null hypothesis in applications, we will create a sample analogue $\widehat{\Delta}$ of Δ, and reject the null hypothesis if the realization $\widehat{\Delta}$ is sufficiently large in absolute value.

3.2. Detectable Alternatives. To understand which kind of deviations from the null hypothesis can potentially be detected by our test, recall that Assumption 1 ensures that $U \perp X|V$, and that V is identified in the subpopulation with $X>0$, i.e. $V\mathbb{I}\{X>0\} = V^*$ with $V^* = F_{X|Z}(X;Z)$. Under the alternative, we have that

$$\Delta = \int \lim_{x\downarrow 0} \mathbb{E}(Y|X=x,V^*=v) G(dv) - \mathbb{E}(Y|X=0), \tag{3.4}$$

with $G(v) = \left(H(v) - \Pr(V^* \leq v, X>0)\right) / \Pr(X=0)$ and H the CDF of the standard uniform distribution, and consequently there is in general no reason to expect that $\Delta = 0$.

There are certain types of violations of Assumption 1, however, for which this is the case, and our test would have no power against such alternatives. This includes pathological cases, where the conditional distribution of Y given (X, V^*) is such that the integral in (3.4) is incidentally equal to $\mathbb{E}(Y|X = 0)$, but also cases that have an economically meaningful interpretation. For example, it is easy to see that (3.1) does not fully exploit the restriction that $U \perp X|V$, but would continue to hold if $U \not\perp X|V$ but $U \perp X|(V, 0 \leq X < \delta)$ for some small $\delta > 0$. Our test is thus generally unable to detect violations of the null hypothesis where V is only able to capture the dependence between X and U local to $X = 0$. We would argue, however, that one is unlikely to encounter such alternatives in empirical applications.

3.3. Test Statistic. We now describe the construction of our test statistic for the pair of hypothesis described in (2.4). The data consists of an i.i.d. sample $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ from the distribution of (Y, X, Z). Our test statistic is given by

$$\widehat{\Delta} = \int \widehat{m}^+_{Y|X,V}(0, v) d\widehat{F}_{V|X}(v, 0) - \widehat{m}_{Y|X}(0), \tag{3.5}$$

where $\widehat{m}^+_{Y|X,V}(0, v)$ and $\widehat{m}_{Y|X}(0)$ are nonparametric estimates of $\lim_{x \downarrow 0} \mathbb{E}(Y|X = x, V = v)$ and $\mathbb{E}(Y|X = 0)$, respectively, and $\widehat{F}_{V|X}(\cdot, 0)$ is a nonparametric estimate of the distribution function $F_{V|X}(\cdot, 0)$ of V conditional on $X = 0$. As we show below, this test statistic is easy to compute, as the particular structure of our estimator $\widehat{F}_{V|X}(\cdot, 0)$ simplifies the computation of the integral in (3.5).

We propose to estimate several of the quantities that appear in the definition of the test statistic $\widehat{\Delta}$ by local linear smoothing (Fan and Gijbels, 1996), which is well known to have attractive properties with regard to boundary bias and design adaptivity. For generic random variables $A \in \mathbb{R}$ and $B \in \mathbb{R}^{d_B}$ and a sample $\{(A_i, B_i)\}_{i=1}^n$ we define the local linear estimators of $\mathbb{E}(A|B = b)$ and $\lim_{x \downarrow b} \mathbb{E}(A|B = x)$ by

$$\widehat{\mathbb{E}}^{LL}_h(A|B = b) = e_1^\top \underset{(a_1, a_2^\top)}{\operatorname{argmin}} \sum_{i=1}^n \left(A_i - a_1 - a_2^\top (B_i - b)\right)^2 K_h(B_i - b) \text{ and}$$

$$\widehat{\mathbb{E}}^{LL}_h(A|B = b^+) = e_1^\top \underset{(a_1, a_2^\top)}{\operatorname{argmin}} \sum_{i=1}^n \left(A_i - a_1 - a_2^\top (B_i - b)\right)^2 K_h(B_i - b) \mathbb{I}\left\{B_i > b\right\},$$

respectively, where $K_h(b) = \prod_{j=1}^{d_B} \mathcal{K}(b_j/h)/h$ is a d_z-dimensional product kernel built from the univariate kernel function $\mathcal{K}$, h is a one-dimensional bandwidth that tends to

zero as the sample size n tends to infinity, and $e_1 = (1, 0, \ldots, 0)^\top$ denotes the first unit $(d_B + 1)$-vector.

With this notation, we can now describe the construction of $\widehat{\Delta}$. In a first step, we obtain estimates of the realizations of the unobserved but identified random variable $V^* = F_{X|Z}(X, Z)$. These estimates are given by $\{\widehat{V}_i^*\}_{i=1}^n$, where $\widehat{V}_i^* = \widehat{F}_{X|Z}(X_i, Z_i)$ and

$$\widehat{F}_{X|Z}(x, z) = \widehat{\mathbb{E}}_g^{LL}(\mathbb{I}\{X \leq x\} | Z = z).$$

Second, we exploit relationship (3.3) and put

$$\widehat{F}_{V|X}(v, 0) = \frac{1}{\widehat{\Pr}(X = 0)} \left(H(v) - \widehat{\Pr}(V^* \leq v, X > 0) \right),$$

where $H(v) = v\mathbb{I}\{0 \leq v \leq 1\} + \mathbb{I}\{v > 1\}$ is the CDF of the uniform distribution on the unit interval, $\widehat{\Pr}(X = 0) = n^{-1} \sum_{i=1}^n \mathbb{I}\{X_i = 0\}$ is the proportion of observations with $X_i = 0$, and

$$\widehat{\Pr}(V^* \leq v, X > 0) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\left\{\widehat{V}_i^* \leq v\right\} \mathbb{I}\{X_i > 0\}$$

is the proportion of observations with $\widehat{V}_i^* \leq v$ and $X_i > 0$. Third, we define

$$\widehat{m}_{Y|X,V}^+(0, v) = \widehat{\mathbb{E}}_h^{LL}(Y | X = 0^+, \widehat{V}^* = v).$$

Finally, we define the estimate $\widehat{m}_{Y|X}(0)$ as a sample average of the observed outcomes Y_i among those observations with $X_i = 0$:

$$\widehat{m}_{Y|X}(0) = \frac{1}{n\widehat{\Pr}(X = 0)} \sum_{i=1}^n Y_i \mathbb{I}\{X_i = 0\}.$$

This completes the description of the components of $\widehat{\Delta}$. Note that because of the structure of $\widehat{F}_{V|X}(\cdot, 0)$ our test statistic can be written as

$$\widehat{\Delta} = \frac{1}{\widehat{\Pr}(X = 0)} \left(\int_0^1 \widehat{m}_{Y|X,V}^+(0, v) dv - \frac{1}{n} \sum_{i=1}^n \widehat{m}_{Y|X,V}^+(0, \widehat{V}_i^*) \mathbb{I}\{X_i > 0\} \right) - \widehat{m}_{Y|X}(0),$$

which is easy to compute as it only involves a one-dimensional numerical integration problem.

3.4. Asymptotic Theory. In this subsection, we study the theoretical properties of the test statistic $\widehat{\Delta}$. This is a non-standard problem because it involves a nonparametric regression on the estimated data points $\{\widehat{V}_i^*\}_{i=1}^n$. We address this issue by using recent results in Mammen, Rothe, and Schienle (2012, 2013) on nonparametric regression with generated covariates. Making use of this results requires the following assumption, which is largely similar to conditions that are commonly imposed in the context of local linear estimation.

Assumption 3. *We assume the following properties for the data distribution, the bandwidth, and kernel function $\mathcal{K}$.*

(i) The random vector Z is continuously distributed with support $S_Z = supp(Z) \subset \mathbb{R}^{d_Z}$. The corresponding density function $f_Z(\cdot)$ is continuously differentiable, bounded, and bounded away from zero on S_Z.

(ii) The conditional CDF $F_{X|Z}(x,z) = \mathbb{E}(\mathbb{I}\{X \leq x\}|Z = z)$ is twice continuously differentiable with respect to its second argument on S_Z.

(iii) The random vector (X,V) is continuously distributed conditional on $X > 0$ with support $S_{XV|X>0} = supp(X,V|X>0)$. The corresponding density function $f_{XV|X>0}(\cdot)$ is continuously differentiable, bounded, and bounded away from zero on the compact set $S_\delta = \{(x,v) : (x,v) \in S \text{ and } x \leq \delta\}$ for some sufficiently small $\delta > 0$.

(iv) The conditional expectation function $m_{Y|XV}(x,v) = \mathbb{E}(Y|X = x, V = v)$ is twice continuously differentiable on S_δ.

(v) There exist a constant $c > 0$ and some constant $l > 0$ small enough such that the residuals $\varepsilon = Y - \mathbb{E}(Y|X,V)$ satisfy the inequality $\mathbb{E}(\exp(l|\varepsilon|)|X,V,X>0) \leq c$.

(vi) The function $\mathcal{K}$ is twice continuously differentiable and satisfies the following conditions: $\int \mathcal{K}(u)du = 1$, $\int u\mathcal{K}(u)du = 0$, and $\mathcal{K}(u) = 0$ for values of u not contained in some compact interval, say $[-1,1]$.

(vii) The bandwidths g and h satisfy the following conditions as $n \to \infty$: (a) $nh^5 \to 0$, (b) $nh^3/\log(n) \to \infty$, (c) $nhg^4 \to 0$ and (d) $h^2(ng^{d_z}/\log(n) + g^{-4}) \to \infty$.

As stated above, Assumption 3 collects conditions that are very common in the literature on nonparametric regression. Parts (i) and (iii) ensures that the estimates $\widehat{F}_{X|Z}(x,z)$ and $\widehat{m}^+_{Y|X,V}(0,v)$ are stable over their respective range of evaluation. Parts (ii) and (iv) are smoothness conditions used to control the magnitude of certain bias terms. Note that part (iv) implies together with (2.1)–(2.2) that $\lim_{x\downarrow 0}\mathbb{E}(Y|X=x,V=v) = \mathbb{E}(m_1(0,U)|m_2(Z,v)=0,V=v)$ under both the null and the alternative. Assuming subexponential tails of ε conditional on $(X,V,X>0)$ in part (v) is necessary to apply certain results from Mammen, Rothe, and Schienle (2012, 2013) in our proofs. Part (vi) describes a standard kernel function with compact support. At the expense of technically more involved arguments, this part could be relaxed to also allow for certain kernels with unbounded support. In particular, the Gaussian kernel would be allowed. Finally, part (vii) collects a number of restrictions on the bandwidths that are partly standard, and partly sufficient for certain "high-level" conditions in Mammen, Rothe, and Schienle (2012, 2013).

The following theorem derives the asymptotic distribution of our test statistic under the assumptions stated above. To state the result, we write $f^+_{V|X}(v,0) = \lim_{x\downarrow 0} f_{V|X}(v,x)$ and $f^+_X(0) = \lim_{x\downarrow 0} f_X(x)$, and define

$$\sigma^2_+(0) = \lim_{x\downarrow 0}\mathbb{E}(\eta^2|X=x)$$

where $\eta = \varepsilon f_{V|X}(V,0)/f^+_{V|X}(V,0)$. Also, for $j\in\{0,1,2\}$ we define the constants

$$\kappa_j = \int_0^\infty x^j\mathcal{K}(x)dx \quad \text{and} \quad \lambda_j = \int_0^\infty x^j\mathcal{K}(x)^2dx,$$

which depend on the kernel function $\mathcal{K}$ only, and put

$$C = \frac{\kappa_2^2\lambda_0 - 2\kappa_1\kappa_2\lambda_1 + \kappa_1^2\lambda_2}{(\kappa_2\kappa_0-\kappa_1^2)^2}.$$

With this notation, we have the following result.

Theorem 1. *Under Assumption 2*(ii) *and 3 the following statements hold.*

(i) Under the null hypothesis, i.e. if $\Delta = 0$*,*

$$\sqrt{nh}\cdot\widehat{\Delta} \xrightarrow{d} N\left(0, C\cdot\frac{\sigma^2_+(0)}{f^+_X(0)}\right).$$

(ii) Under any fixed alternative that implies $\Delta \neq 0$,

$$\lim_{n\to\infty} \Pr(|\widehat{\Delta}| > c) = 1$$

for all constants $c > 0$.

Theorem 1 suggests that a test of (2.4) could be based on an appropriately studentized version of $\widehat{\Delta}$, with critical values being obtained from the standard normal distribution. Let $\widehat{\sigma}^2_+(0)$ and $\widehat{f}^+_X(0)$ be consistent estimates of $\sigma^2_+(0)$ and $f^+_X(0)$, and define

$$\widehat{\Delta}_S = \left(\frac{C}{nh} \cdot \frac{\widehat{\sigma}^2_+(0)}{\widehat{f}^+_X(0)} \right)^{-1/2} \widehat{\Delta}.$$

Then the test decision is to reject H_0 at the nominal level α if $|\widehat{\Delta}_S| > \Phi(1-\alpha/2)$, where Φ is the CDF of the standard normal distribution. Theorem 1 shows that this is a test of nominal size α, which is consistent against all fixed alternatives that imply that $\Delta \neq 0$.

Theorem 2. *Under Assumption 2*(ii) *and 3 the following statements hold.*

(i) Under the null hypothesis, i.e. if $\Delta = 0$,

$$\lim_{n\to\infty} \Pr(|\widehat{\Delta}_S| > \Phi(1-\alpha/2)) = \alpha.$$

(ii) Under any fixed alternative that implies $\Delta \neq 0$,

$$\lim_{n\to\infty} \Pr(|\widehat{\Delta}_S| > \Phi(1-\alpha/2)) = 1.$$

for all $\alpha > 0$.

4. Numerical Evidence

Results from a Monte Carlo study and an empirical application. To be completed.

5. Concluding Remarks

In this paper, we have proposed a test for the validity of a control variable approach to identification in triangular models with endogeneity. To the best of our knowledge, this is the first test of this type of hypothesis. Our test requires a particular data structure,

namely that the endogenous covariate has a mass point at the lower (or upper) boundary of its support, and is otherwise continuously distributed. While this is certainly restrictive, we argue that our test is still useful for a wide range of empirical applications. We propose a test statistic that is simple to compute, and derive its asymptotic properties.

A. Mathematical Appendix

A.1. Proof of Theorem 1. The result in Theorem 1 follows directly from the three axillary results in Lemma 1–3 below. The first of these three findings gives a bound on the uniform rate of consistency of the estimated function $\widehat{F}_{V|X}(\cdot, 0)$.

Lemma 1. *Suppose that the conditions of Theorem 1 hold. Then*

$$\sup_{v \in supp(V|X=0)} |\widehat{F}_{V|X}(v,0) - F_{V|X}(v,0)| = O_P(n^{-1/2}) + O(g^2).$$

Proof. The structure of $\widehat{F}_{V|X}(\cdot, 0)$ is very similar to that of an empirical distribution function of the estimates $\{\widehat{V}_i^*\}_{i=1}^n$ in the subset of the sample with $X_i > 0$. The result then follows from arguments analogous to those in Akritas and Van Keilegom (2001). □

To state our next result, we introduce an infeasible estimator of the function $m^+_{Y|X,V}(0, v) = \lim_{x \downarrow 0} \mathbb{E}(Y|X = x, V = v)$ that uses the actual realizations of $V_i^* = F_{X|Z}(X_i, Z_i)$ instead of the corresponding estimated values $\widehat{V}_i^*$. The estimator is given by $\widetilde{m}^+_{Y|X,V}(0, v) = \mathbb{E}_h^{LL}(Y|X = 0^+, V = v)$. We also define an infeasible version of our test statistic which uses the population values $F_{V|X}(\cdot, 0)$ and $\mathbb{E}(Y|X = 0)$ instead of their estimates, and replaces $\widehat{m}^+_{Y|X,V}(0, v)$ with its infeasible version $\widetilde{m}^+_{Y|X,V}(0, v)$:

$$\widetilde{\Delta} = \int \widetilde{m}^+_{Y|X,V}(0,v) dF_{V|X}(v,0) - \mathbb{E}(Y|X=0).$$

The following lemma derives the asymptotic properties of the infeasible test statistic $\widetilde{\Delta}$.

Lemma 2. *Suppose that the conditions of Theorem 1 hold. Then*

$$\sqrt{nh}(\widetilde{\Delta} - \Delta) \xrightarrow{d} N\left(0, C \cdot \frac{\sigma^2_+(0)}{f^+_X(0)}\right).$$

as $n \to \infty$.

Proof. To show this result, we first introduce the additional notation that:

$$L_i(v) = (1, X_j/h, (V_i - v)/h)^\top \cdot \mathbb{I}\{X_i > 0\},$$
$$M_n(v) = \frac{1}{n}\sum_{i=1}^{n} L_i(v)L_i(v)^\top K_h(X_j, V_i - v),$$
$$N_n(v) = \mathbb{E}(L_i(v)L_i(v)^\top K_h(X_i, V_i - v)).$$

With this notation, the local linear estimator $\widetilde{m}^+_{Y|XV}(0, v)$ can be written as

$$\widetilde{m}^+_{Y|XV}(0, v) = \frac{1}{n}\sum_{i=1}^{n} e_1^\top M_n(v)^{-1} L_i(v) K_h(X_i, V_i - v) Y_j.$$

It also follows from straightforward calculations that the term $N_n(v)$ satisfies

$$N_n(v) = AP(X > 0) \lim_{x \downarrow 0} f_{VX}(v, x) + o(1)$$

uniformly in v, where

$$A = \begin{pmatrix} \kappa_0 & \kappa_1 & 0 \\ \kappa_1 & \kappa_2 & 0 \\ 0 & 0 & \kappa_2^* \end{pmatrix} \quad \text{and} \quad \kappa_2^* = \int_{-\infty}^{\infty} x^2 \mathcal{K}(x) dx.$$

Here the form of A follows from the assumption that the kernel function $\mathcal{K}$ is a symmetric density function. We now introduce a particular stochastic expansion for this estimator, which follows from standard results in e.g. Masry (1996). Writing

$$S_n(v) = \frac{1}{n}\sum_{i=1}^{n} e_1^\top N_n(v)^{-1} L_i(v) K_h(X_i, V_i - v)\varepsilon_i$$

with $\varepsilon_j = Y_i - \mathbb{E}(Y_i|X_i, V_i)$, we have that

$$\widetilde{m}^+_{Y|XV}(0, v) = m^+_{Y|XV}(0, v) + S_n(v) + O(h^2) + O_P\left(\frac{\log(n)}{nh^2}\right)$$

uniformly over $v \in supp(V|X = 0)$. Using standard change-of-variables arguments, we find that

$$\int S_n(v) dF_{V|X}(v, 0) = \frac{1}{n}\sum_{i=1}^{n} e_1^\top N_n(V_i)^{-1} L_i^* \mathcal{K}_h(X_i) f_{V|X}(V_i, 0)\varepsilon_i + O(h^2)$$

with $L_j^* = (1, X_j/h, 0)^\top \cdot \mathbb{I}\{X_i > 0\}$. The first term on the right-hand-side of the last equation is a sample average of n independent random variables, and clearly has mean zero. On the

other hand, its variance is equal to

$$
\begin{aligned}
&n^{-1}\mathbb{E}((e_1^\top N_n(V_i)^{-1}L_j^*)^2\mathcal{K}_h(X_i)^2 f_{V|X}(V_i,0)^2\varepsilon_i^2) \\
&\quad= \frac{1}{nh}\int_0^\infty (e_1^\top A^{-1}(1,x,0)^\top)^2\mathcal{K}(x)^2\mathbb{E}\left(\frac{f_{V|X}(V,0)^2}{f^+_{V|X}(v,0)^2}\cdot\varepsilon^2\middle|\, X=xh\right)\frac{f_X(xh)}{f_X^+(0)^2}dx + o\left(\frac{1}{nh}\right) \\
&\quad= \frac{1}{nh}\int_0^\infty (e_1^\top A^{-1}(1,x,0)^\top)^2\mathcal{K}(x)^2\mathbb{E}\left(\eta^2\middle|\, X=xh\right)\frac{f_X(xh)}{f_X^+(0)^2}dx + o\left(\frac{1}{nh}\right) \\
&\quad= \frac{C}{nh}\cdot\frac{\sigma_+^2(0)}{f_X^+(0)} + o\left(\frac{1}{nh}\right)
\end{aligned}
$$

The statement of the lemma then follows from an application of Ljapunov's Central Limit Theorem. □

As the final step of our proof of Theorem 1, the following lemma shows that $\widetilde{\Delta}$ and $\widehat{\Delta}$ have the same first order asymptotic properties.

Lemma 3. *Suppose that the conditions of Theorem 1 hold. Then*

$$\widetilde{\Delta} - \widehat{\Delta} = o_P((nh)^{-1/2})$$

as $n \to \infty$.

Proof. First, using that $\widehat{m}_{Y|X}(0) = \mathbb{E}(Y|X=0) + O_P(n^{-1/2})$ and Lemma 1, we find that

$$\widehat{\Delta} = \int \widehat{m}^+_{Y|X,V}(0,v)dF_{V|X=0}(v) - \mathbb{E}(Y|X=0) + O_P(n^{-1/2}),$$

since $\widehat{m}^+_{Y|X,V}(0,v)$ is easily seen to be a consistent estimate of a bounded function under the conditions of the lemma. It thus remains to be shown that

$$\int \widehat{m}^+_{Y|X,V}(0,v)dF_{V|X}(v,0) = \int \widetilde{m}^+_{Y|X,V}(0,v)dF_{V|X}(v,0) + o_P((nh)^{-1/2}).$$

We use recent results on nonparametric regression with generated covariates obtained by Mammen, Rothe, and Schienle (2012, 2013) to show this statement. For convenience, we repeat the following notation, which was already introduced in the proof of Lemma 2:

$$
\begin{aligned}
L_i(v) &= (1, X_i/h, (V_i - v)/h)^\top \cdot \mathbb{I}\{X_i > 0\}, \\
M_n(v) &= \frac{1}{n}\sum_{i=1}^n L_i(v)L_i(v)^\top K_h(X_i, V_i - v), \\
N_n(v) &= \mathbb{E}(L_i(v)L_i(v)^\top K_h(X_i, V_i - v)).
\end{aligned}
$$

It then follows from an application of Theorem 1 in Mammen, Rothe, and Schienle (2013) that

$$\int \widehat{m}^{+}_{Y|X,V}(0,v) - \widetilde{m}^{+}_{Y|X,V}(0,v) - \varphi_n(v;\widehat{F}_{X|Z})dF_{V|X}(v,0) = o_P((nh)^{-1/2})$$

under the conditions of the lemma, where for any conformable function G

$$\begin{aligned}\varphi_n(v;G) = &-(\partial m^{+}_{Y|XV}(0,v)/\partial v)e_1^{\top}N_n(v)^{-1} \\ &\times \mathbb{E}(L_i(v)K_h(X_i,V_i-v)(G(X_i,Z_i)-F_{X|Z}(X_i,Z_i)))\end{aligned}$$

Note that the formula for φ_n is simpler than the general expression in Mammen, Rothe, and Schienle (2013) because our model implies that $E(Y|X,V) = \mathbb{E}(Y|X,Z)$ under the null hypothesis, and thus their "index bias" term vanishes. Next, it follows from the same arguments as in the proof of Theorem 4 in Mammen, Rothe, and Schienle (2013) that

$$\int \varphi^{A}_n(v;\widehat{F}_{X|Z})dF_{V|X}(v,0) = O_P(n^{-1/2}) + O(h^2) + O(g^2) + O_P\left(\frac{\log n}{ng}\right).$$

This completes our proof. □

References

Akritas, M. G., and I. Van Keilegom (2001): "Non-parametric Estimation of the Residual Distribution," *Scandinavian Journal of Statistics*, 28(3), 549–567.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75(6), 1613–1669.

Blundell, R., and J. Powell (2003): "Endogeneity in nonparametric and semiparametric regression models," in *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, vol. 2, pp. 655–679.

Blundell, R., and J. Powell (2004): "Endogeneity in semiparametric binary response models," *The Review of Economic Studies*, 71(3), 655–679.

Caetano, C. (2012): "A Discontinuity Test of Endogeneity," *Working Paper*.

Canay, I. A., A. Santos, and A. M. Shaikh (2012): "On the Testability of Identification in some Nonparametric Models with Endogeneity," Discussion paper, Working paper.

Chernozhukov, V., and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73(1), 245–261.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): "Instrumental variable estimation of nonseparable models," *Journal of Econometrics*, 139(1), 4–14.

Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011): "Nonparametric instrumental regression," *Econometrica*, 79(5), 1541–1565.

Fan, J., and I. Gijbels (1996): *Local polynomial modelling and its applications*. CRC Press.

Hall, P., and J. L. Horowitz (2005): "Nonparametric methods for inference in the presence of instrumental variables," *The Annals of Statistics*, 33(6), 2904–2929.

Imbens, G., and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77(5), 1481–1512.

Imbens, G. W. (2007): "Nonadditive models with endogenous regressors," in *Advances in Economics and Econometrics*, ed. by R. Blundell, W. Newey, and T. Persson, vol. 43.

Kasy, M. (2011): "Identification in Triangular Systems using Control Functions," *Econometric Theory*, 27, 663–671.

——— (2013): "Identification in General Triangular Systems," *Working Paper*.

Mammen, E., C. Rothe, and M. Schienle (2012): "Nonparametric Regression with Nonparametrically Generated Covariates," *Annals of Statistics*.

——— (2013): "Semiparametric Estimation with Generated Covariates," *Working Paper*.

Masry, E. (1996): "Multivariate local polynomial regression for time series: uniform strong consistency and rates," *Journal of Time Series Analysis*, 17(6), 571–599.

Newey, W., J. Powell, and F. Vella (1999): "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 67(3), 565–603.

Newey, W. K., and J. L. Powell (2003): "Instrumental variable estimation of nonparametric models," *Econometrica*, 71(5), 1565–1578.

Rothe, C. (2009): "Semiparametric estimation of binary response models with endogenous regressors," *Journal of Econometrics*, 153(1), 51–64.