A Discontinuity Test for Identification in Triangular Nonseparable Models

Carolina Caetano, Christoph Rothe, and Nese Yildiz*

Abstract

This paper presents a test for the validity of control variable approaches to identification in triangular nonseparable models. Assumptions commonly imposed to justify such methods include full independence of instruments and disturbances and existence of a reduced form that is strictly monotonic in a scalar disturbance. We show that if the data has a particular structure, namely that the distribution of the endogenous variable has a mass point at the lower (or upper) boundary of its support, validity of the control variable approach implies a continuity condition on an identified function, which can be tested empirically.

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

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

*This Version: July 24, 2014. 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

Empirical specifications with nonseparable unobservables have become increasingly popular in econometrics in recent years (e.g. Matzkin, 2003; Chesher, 2003; Imbens and Newey, 2009; Blundell and Matzkin, 2010). In their most basic form, these models assume that an outcome variable Y is linked to a covariate X and an unobserved quantity U through the relationship

$$Y = m_1(X, U).$$

Compared to classical specifications with additively separable disturbances, these types of models can accommodate very general forms of unobserved heterogeneity. For example, they allow for heterogeneous responses to policy interventions among observationally identical individuals. Both economic theory and empirical evidence strongly suggest that such general forms of unobserved heterogeneity are a common feature of economic data (Heckman, 2001).

The additional flexibility of these models comes of course at a cost. When the covariate X is endogenous, the lack of an additively separable disturbance complicates the identification of many interesting functionals of the structural model. Availability of an instrument, say Z, that is uncorrelated with U but correlated with X does not suffice for identification. Instead, one has to impose additional conditions. One popular approach is based on control variables (Blundell and Powell, 2003; Imbens and Newey, 2009). It entails finding a random variable R that can be written as an identified function of the data, and is such that X and U are stochastically independent conditional on R, that is

$$X \perp U | R.$$

This property ensures that changes in X can be interpreted as causal after conditioning on R, and thus many structural parameters can be identified from the conditional distribution of Y given X and R. Such a control variable arises for example if the model has a triangular structure, that is if the endogenous variable is assumed to be generated in a first stage as

$$X = m_2(Z, V),$$

with Z an instrument and V an unobserved quantity. In a seminal paper, Imbens and Newey (2009) show that if Z is independent of (U, V) and m_2 depends monotonically on the continuously distributed scalar V, then $R := F_{X|Z}(X, Z) = F_V(V)$ is a valid control variable, where $F_{X|Z}(x, z)$ denotes the conditional CDF of X given Z and $F_V(v)$ denotes the unconditional CDF of V. This is because here R is a one-to-one transformation of V, and conditional on V the endogenous variable X only depends on Z.

While this approach to identification is powerful, the postulated triangular specification paired with the restrictions on first stage unobserved heterogeneity implies substantial limitations for the underlying economic model. Indeed, in many empirical settings these conditions can be difficult to justify through theoretical considerations, and thus their validity might be doubtful. Unfortunately, standard approaches to specification testing in models with additively separable disturbances do typically not apply in nonseparable settings. For example, Hausman-type tests for instrument validity, which are widely used in parametric models with additive errors, have no direct analogue in triangular nonseparable models, as there is no notion of overidentification.

In this paper, we show that the conditions necessary for justifying a control variable approach can potentially be refuted empirically using the data and some set of weak maintained assumptions alone if the model exhibits a particular additional structure. Specifically, we study the case where the data generating process is such that the distribution of the endogenous covariate 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. We show that under these conditions validity of the control variable approach implies a continuity condition on a certain function at one particular point, and that this continuity condition can be tested. To the best of our knowledge, our paper is the first to derive testable implications of identifying assumptions in such a context.

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 settings in which our approach can be applied. Yet there are many examples of variables with such a property that appear frequently as potentially endogenous covariates in empirical applications: 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. Our approach should therefore be applicable in a wide range of empirical settings.

The idea behind the derivation of our testable implication is related to that of Caetano (2014), who shows that endogeneity of a covariate X with the above-mentioned properties leads to a discontinuity in the conditional expectation function of Y given X at the mass point. In this paper, the starting point for our analysis is the insight that conditioning on a valid control variable should remove this discontinuity. On the other hand, if the conditional expectation of Y given X and the control variable is discontinuous at the mass point, then the control variable approach must be invalid; and at least one of the assumptions that was made to justify it has to be violated. This basic idea can unfortunately not be implemented directly, because in our setting a valid control variable is only available for those individuals whose realization of the endogenous variable is different from the mass point. We address this issue by integrating out the control variable in such a way that it is no longer necessary to identify it at the mass point. This results in a function that is identifiable and must be continuous if the control variable approach is valid.

The main part of our paper focuses exclusively on establishing a testable implication of the control variable approach. In the appendix, we also propose a test statistic that is based on a direct sample analogue of the function whose continuity we wish to verify. The statistic is similar in structure to estimators used in the regression discontinuity (RD) literature. Its computation is straightforward, involving only standard nonparametric regression techniques and a simple numerical integration step. We derive the asymptotic properties of our test statistic, which is a non-standard problem as it involves running a nonparametric regression 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, 2014) on generated covariates in non- and semiparametric models. We show that the test statistic is asymptotically normal and derive an explicit formula for its asymptotic variance, which can be estimated to obtain critical values.

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 (2014) 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), Chernozhukov, Imbens, and Newey (2007), Torgovitsky (2012) and d'Haultfoeuille and Février (2012) consider IV methods in nonseparable models, but with restrictions on the dimension of the disturbances. Canay, Santos, and Shaikh (2013) show that the completeness condition, which plays a central role for identification in nonparametric IV approaches, is generally not testable.

The remainder of the paper is structured as follows. In Section 2, we introduce the model and the identifying assumptions, and formally state the testing problem. Section 3 contains our main results. Section 4 concludes. In the appendix, we propose a test statistic based on our approach and derive its theoretical properties.

2. Setup

In this section, we first introduce the model and the identifying assumptions whose validity we want to investigate, and then formally state the testing problem we wish to consider.

2.1. Model. We consider a triangular system of equations as the data generating process. To simplify the exposition, the system only contains a single potentially endogenous covariate, but it is allowed to contain an arbitrary number of additional exogenous ones. Extending our arguments to systems with multiple endogenous covariates is conceptually straightforward. The main difference between our setup and standard existing frameworks is that we consider the case of an endogenous variable whose distribution has a mass point at the lower bound of its support. Such a variable could arise in settings where natural or legal restrictions lead to corner solutions in the individuals'

optimization problem that determines the value of the endogenous variable. Examples include weekly hours of work (which have to be non-negative), hourly wages (which have to exceed the minimum wage), or the amount of consumption of a particular good (which has to be non-negative). Specifically, our model is given by

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

$$X = \max\{0, m_2(Z, W, V)\} \tag{2.2}$$

where Y is the outcome of interest, X is a scalar and potentially endogenous covariate, W is a d_W-dimensional vector of additional exogenous covariates, and U and V denote unobserved heterogeneity. We also use the notation that

$$X^* = m_2(Z, W, V).$$

Taking the lower bound of the support of X to be equal to zero is without loss of generality. If the threshold is equal to some other known constant c, the above representation can simply be achieved by subtracting c from both X and $m_2(Z, W, V)$. Similarly, we could allow for a known upper bound on the support of X instead of a lower one. In addition, we assume that the data generating process is such that

$$0 < \Pr(X^* \leq 0|Z, W) < 1 \text{ with probability 1}, \tag{2.3}$$

so that the conditional distribution of X given (Z, W) has an actual mass point at 0, but is not degenerate. We also require that the structural function m_1 satisfies a weak continuity property in its first argument:

$$\lim_{x \downarrow 0} \mathbb{E}(m_1(x, W, U)|W, V) = \mathbb{E}(m_1(0, W, U)|W, V) \text{ with probability 1}. \tag{2.4}$$

A sufficient condition for m_1 having this property is that the mapping $x \mapsto m_1(x, W, U)$ is right-continuous at $x = 0$ with probability 1, which means that variation in X exerts a continuous

causal effect on Y. This seems to be a reasonable assumption for many, although certainly not all, empirical settings.

2.2. Identification Approach. The model that we study in this paper reduces to the one considered by Imbens and Newey (2009) if equation (2.2) would be changed to $X = m_2(Z, W, V)$, and the (then redundant) conditions (2.3)–(2.4) would be dropped. Imbens and Newey (2009) show that in their setting a large class of interesting structural parameters can be identified through control variable arguments. The purpose of this subsection is simply to show that the same is true in our model (2.1)–(2.4) under very similar conditions. To be specific, we focus on identification of the Average Structural Function (ASF), defined by Blundell and Powell (2003) as

$$a(x, w) = \mathbb{E}(m_1(x, w, U)),$$

but the same type of argument applies to a broader class of structural objects. Note that the ASF describes the average outcome of an individual whose covariates (X, W) are exogenously fixed at (x, w). Using the notation that $R = F_{X|ZW}(X; Z, W)$, we impose the following assumptions for identification.

Assumption 1. *The model in* (2.1)–(2.4) *satisfies the following restrictions:*

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

(ii) V *is scalar and continuously distributed with a strictly increasing CDF.*

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

Assumption 2. *The support* $\mathcal{S}(R|X = x, W = w)$ *of* $R = F_{X|ZW}(X; Z, W)$ *conditional on* $X = x$ *and* $W = w$ *is equal to* $(0, 1)$ *for all* $(x, w) \in \mathcal{S}(X, W|X > 0)$.

Assumption 1 is identical to the conditions of Theorem 1 in Imbens and Newey (2009). Part (i) requires full independence between the instruments and the unobserved heterogeneity, whereas parts (ii)–(iii) imply that within the subpopulation that has $X > 0$, individuals with the same realization of the vector (X, Z, W) are also identical in terms of the unobserved heterogeneity V. These

individuals would thus respond identically to exogenous variation in the instruments. Assumption 2, which is again analogous to a condition in Imbens and Newey (2009), is a generalization of the usual rank condition in traditional IV models. It requires the function $(z, w) \mapsto m_2(z, w, V)$ to exhibit a sufficient amount of variation over the support of (Z, W). While this condition is restrictive, it only involves observable quantities and is thus relatively straightforward to verify. Moreover, Imbens and Newey (2009) show that if this assumption fails many structural quantities of interest remain partially identified.

The following proposition shows that Assumption 1 implies the existence of a valid control variable in the subpopulation with $X > 0$, and that taken together Assumptions 1 and 2 ensure identification of the ASF $a(x, w)$ for $x > 0$ through traditional arguments, and for $x = 0$ through the continuity property in equation (2.4).

Proposition 1. *(a) Under Assumption 1, $U \perp (X, W)|R$ in the subpopulation with $X > 0$.*

(b) Under Assumptions 1 and 2,

$$a(x, w) = \begin{cases} \int_0^1 \mathbb{E}(Y|X = x, W = w, R = r)dr & \text{if } x > 0, \\ \lim_{x \downarrow 0} \int_0^1 \mathbb{E}(Y|X = x, W = w, R = r)dr & \text{if } x = 0, \end{cases}$$

and thus the ASF $a(x, w)$ is identified for all $(x, w) \in \mathcal{S}(X, W)$.

Proof. To show part (a), let $R^* = F_{X^*|ZW}(X^*; Z, W)$, and note that it follows from Imbens and Newey (2009) that under Assumption 1 we have that $R^* = F_V(V)$, and that $U \perp (X, W)|V$. The desired result then follows from the fact that $R^* = R$ in the subpopulation with $X > 0$. The proof of part (b) is a minor extension of Theorem 1 in Imbens and Newey (2009). From part (a), it follows that $\mathbb{E}(Y|X = x, W = w, R = r) = \mathbb{E}(m_1(x, w, U)|R^* = r)$ for all $x > 0$ and all r. Since $R^* \sim U[0, 1]$ by construction, we also have that $a(x, w) = \int_0^1 \mathbb{E}(m_1(x, w, U)|R^* = r)dr$. Identification of the ASF for $x > 0$ then follows since Assumption 2 ensures that the integral is well defined. The continuity condition (2.4) then implies identification of the ASF at $x = 0$, because $a(0, w) = \int_0^1 \lim_{x \downarrow 0} \mathbb{E}(m_1(x, w, U)|R^* = r)dr = \lim_{x \downarrow 0} a(x, w)$. □

2.3. Testing Problem. Our interest in this paper is in testing the validity of the assumptions that are necessary to justify a control variable approach described in Section 2.2, namely the model structure described in the equations (2.1)–(2.4), and Assumption 1–2. Since Assumption 2 only involves observable quantities, we focus on cases where this condition is considered to be credible by the analyst, and is thus part of the maintained hypothesis. In fact, for the approach we describe below it suffices to maintain the weaker condition that

$$\mathcal{S}(R|X = x, W = w) = (0, 1) \text{ for all } (x, w) \in \mathcal{S}(X, W|0 < X < \delta) \tag{2.5}$$

for some $\delta > 0$. We also include the basic structure of the data generating process as described in (2.1)–(2.4) in the maintained hypothesis. Note that without further assumptions the equations (2.1)–(2.3) are really just notation, and do not impose any restrictions on the data generating process other than a lower bound on the support of X and strictly positive probability mass at that point. Only the continuity condition (2.4) has to be justified through subject knowledge. This leaves us with Assumption 1 as the remaining restriction whose credibility could be uncertain in an empirical application. The pair of hypotheses we would like to test is thus given by

$$\mathbb{H}_0\text{: Assumption 1 holds} \quad \textit{vs.} \quad \mathbb{H}_1\text{: Assumption 1 is violated} \tag{2.6}$$

under the maintained assumption that conditions (2.1)–(2.5) hold. We say that this problem is testable if there exists an identified functional of the distribution of (Y, X, W, Z) that is equal to zero under $\mathbb{H}_0$ and generally unequal to zero under $\mathbb{H}_1$.

3. Main Results

In this section, we first argue that without the presence of the mass point in the distribution of the endogenous variable the problem in (2.6) would not be testable. We then derive a testable implication that exploits our particular additional structure, and discuss the types of violations of $\mathbb{H}_0$ that can be detected that way.

3.1. Testability without Mass Points. Assumption 1 implies a number substantial limitations for the underlying economic model that generated the data that we observe. For example, it suggests that two individuals in the subpopulation with $X > 0$ that have the same realization of the random vector (X, Z) will react in exactly the same way to exogenous variation in Z. In many empirical applications, the validity of these conditions might be doubtful since they are difficult to justify through theoretical considerations. The testing problem in (2.6) is therefore obviously an important one. However, it is not clear *a priori* that this pair of hypotheses is actually testable. Indeed, the following proposition formally shows that Assumption 1 would have no testable implications in the model considered by Imbens and Newey (2009) beyond continuity of the conditional distribution of X given (Z, W).

Proposition 2. *Suppose that the data are generated according to the model* (2.1)–(2.2) *with* $\inf\{\mathcal{S}(X^*)\} > 0$*, and that* X *is continuously distributed given* (Z, W)*. Then Assumption 1 has no further testable implications.*

Proof. Put $\widetilde{m}_1(x, w, u) = Q_{Y|XW}(u|x, w)$ and $\widetilde{m}_2(z, w, v) = Q_{X|ZW}(v|z, w)$, where $Q_{A|B}(\tau|b)$ denotes the conditional τ-quantile of A given $B = b$. Also let $(\widetilde{U}, \widetilde{V})$ be a random vector drawn independently of the data from the distribution of $(F_{Y|XW}(Y|X, W), F_{X|ZW}(X|Z, W))$. Then the distribution of the random vector $(\widetilde{Y}, \widetilde{X}, W, Z)$ generated by the triangular system

$$\widetilde{Y} = \widetilde{m}_1(\widetilde{X}, W, \widetilde{U}) \text{ and } \widetilde{X} = \widetilde{m}_2(Z, W, \widetilde{V})$$

is identical to the distribution of (Y, X, W, Z). The above system is thus observationally equivalent to (2.1)–(2.2) with $\inf\{\mathcal{S}(X^*)\} > 0$, but its components clearly satisfy Assumption 1 up to differences in notation. □

Remark 1. The non-separability of disturbances is important for obtaining the above result. If the outcome equation was additively separable in the unobservables, as in Newey, Powell, and Vella (1999) for example, the conditional expectation of the outcome given the endogenous covariate and the control variable would be additively separable in the control variable, which would be a testable implication.

3.2. A Testable Implication. Proposition 2 shows that the problem in (2.6) can only testable if we exploit the additional structure that given by our conditions (2.3)–(2.4). Indeed, we now show that under $\mathbb{H}_0$ these two restrictions imply a continuity condition on an identified function, and that this continuity condition is generally violated under $\mathbb{H}_1$. To motivate our approach, suppose for a moment that the latent rank variable $R^* = F_{X^*|ZW}(X^*; Z, W)$ was observable, and write

$$\mu(x, w, r) = \mathbb{E}(Y|X = x, W = w, R^* = r)$$

for the conditional expectation function of Y given (X, W, R^*). Using the structure of the model, it is easily seen that

$$\mu(x, w, r) = \begin{cases} \mathbb{E}(Y|X^* = x, W = w, R^* = r) & \text{if } x > 0, \\ \mathbb{E}(Y|X^* \leq x, W = w, R^* = r) & \text{if } x = 0. \end{cases}$$

Since the conditioning sets in the two conditional expectations on the right-hand side of the previous equation differ, we would generally expect the function $\mu(x, w, r)$ to be discontinuous at $x = 0$ for at least some (and potentially all) values $(w, r) \in \mathcal{S}(W, R^*)$. However, we know from the proof of Proposition 1 that if Assumption 1 holds the latent rank variable R^* is a valid control variable satisfying $U \perp (X, W)|R^*$, and thus

$$\mu(x, w, r) = \mathbb{E}(m_1(x, w, U)|R^* = r)$$

in this case. Since we also have that $R^* = F_V(V)$, it then follows from the continuity condition (2.4) that if Assumption 1 holds, and thus our null hypothesis is true, the function $x \mapsto \mu(x, W, R^*)$ must be right-continuous at $x = 0$ with probability 1. On the other hand, if Assumption 1 is violated, and we are thus under the alternative, we expect the function $x \mapsto \mu(x, W, R^*)$ to be discontinuous at $x = 0$ with positive probability.

This continuity condition is unfortunately not a testable implication of validity of the control variable approach because we can only observe R but not R^*. While these two terms coincide in the subpopulation with $X > 0$, and thus $\mu(x, w, r) = \mathbb{E}(Y|X = x, W = w, R = r)$ for $x > 0$, it is

easy to see that we can only deduce that $0 \leq R^* \leq R$ if $X = 0$. This means that while we are able to learn $\lim_{x \downarrow 0} \mu(x, w, r)$, the data do not point identify $\mu(0, w, r)$, and we can thus not directly check for the presence of a discontinuity. To address this problem, we consider the quantity

$$\Delta(w) = \int \lim_{x \downarrow 0} \mathbb{E}(Y|X = x, W = w, R = r) dF_{R^*|X,W}(r; 0, w) - \mathbb{E}(Y|X = 0, W = w), \tag{3.1}$$

where $F_{R^*|X,W}(r; 0, w)$ denotes the conditional CDF of R^* given $X = 0$ and W. To see why this is useful, note that by construction we have that

$$\Delta(w) = \int \left(\lim_{x \downarrow 0} \mu(x, w, r) - \mu(0, w, r) \right) dF_{R^*|X,W}(r; 0, w),$$

and thus $\Delta(w)$ is equal to the size of the discontinuity of the function

$$(x, w) \mapsto \int \mathbb{E}(Y|X = x, W = w, R^* = r) dF_{R^*|X,W}(r; 0, w)$$

at $x = 0$; which is exactly equal to zero under $\mathbb{H}_0$. Weighting with respect to the conditional distribution of R^* given X and W circumvents the need to identify $\mu(0, w, r)$, but this strategy is of course only useful if $F_{R^*|X,W}(r; 0, w)$ is identified. To see that this is the case, note that it follows from the law of total probability that

$$\begin{aligned} F_{R^*}(r) = &\Pr(R^* \leq r|X > 0, W = w) \Pr(X > 0|W = w) \\ &+ \Pr(R^* \leq r|X = 0, W = w) \Pr(X = 0|W = w). \end{aligned}$$

Since $R^* \sim U[0, 1]$ by construction and $R^* = R$ in the subpopulation with $X > 0$, a simple rearrangement of terms shows that

$$F_{R^*|X,W}(r; 0, w) = \frac{F_{U[0,1]}(r) - \Pr(R \leq r|X > 0, W = w) \Pr(X > 0|W = w)}{\Pr(X = 0|W = w)}, \tag{3.2}$$

and all terms on the right-hand side of the previous equation can be written as a known transformation of the joint distribution of (Y, X, W, Z). In summary, this argument delivers the testable

implication that under the null hypothesis $\Delta(w)$ must be equal to zero over the support of W (up to a set of measure zero), whereas under the alternative it will generally be non-zero. We formally state this finding in the next theorem.

Theorem 1 (Main Testable Implication)**.** *The function* $\Delta(w)$ *given in* (3.1) *is a well defined functional of the distribution of* (Y, X, Z, W)*. Under the null hypothesis,* $\Pr(\Delta(W) = 0) = 1$*, whereas under the alternative* $\Pr(\Delta(W) = 0) \leq 1$*.*

Proof. Equations (3.1) and (3.2) show that the definition of $\Delta(w)$ only involves obervable quantities. Now let $\bar{\Delta}(w, r) = \lim_{x \downarrow 0} \mathbb{E}(Y|X = x, W = w, R^* = r) - \mathbb{E}(Y|X = 0, W = w, R^* = r)$. Then it follows from the above arguments that $P(\bar{\Delta}(W, R^*) = 0) = 1$ under $\mathbb{H}_0$ and $P(\bar{\Delta}(W, R^*) = 0) \leq 1$ under $\mathbb{H}_1$. Since $\Delta(w) = \int \bar{\Delta}(w, r) dF_{R^*|X,W}(r; 0, w)$ this implies the statement of the theorem. □

Remark 2. A natural way to exploit the result in Theorem 1 for creating a feasible testing procedure is to construct a sample analogue $\widehat{\Delta}(w)$ of $\Delta(w)$, and to reject the null hypothesis when the realization of $\widehat{\Delta}(\cdot)$ is "too large" in some suitable norm. We formally describe such an approach in the Appendix.

3.3. Detectable Alternatives. Our approach is generally able to detect violations of each of the three components of Assumption 1, and can thus be a powerful tool for empirical practice. We illustrate this point through numerical examples in the following subsection. However, the discussion preceding Theorem 1 also shows that the condition that $\Delta(W) = 0$ with probability 1 is only necessary but not sufficient for the null hypothesis to hold. To explain the implications of this fact, we first define the function

$$\bar{\Delta}(w, r) = \lim_{x \downarrow 0} \mathbb{E}(Y|X = x, W = w, R^* = r) - \mathbb{E}(Y|X = 0, W = w, R^* = r).$$

Now there are two clases of settings in which $\Delta(W)$ could be equal to zero even though Assumption 1 is violated:

(i) $\bar{\Delta}(w, r) \neq 0$, but $\int \bar{\Delta}(w, r) dF_{R^*|X,W}(r; 0, w) \equiv 0$,

(ii) $\bar{\Delta}(w,r) \equiv 0$ even though $\mathbb{H}_1$ is true.

In case (i), the joint distribution of (Y, X, W, R^*) is such that the function $r \mapsto \bar{\Delta}(w,r)$ takes on both positive and negative values in such a way that it incidentally integrates to zero with respect to $F_{R^*|X,W}(r; 0, w)$ for *every* value of w. We are not aware of any interpretable restrictions on the primitives of the model under which this would be the case, but a numerical example in the following subsection shows that it is possible in principle. We think of this case as a pathological one, and would argue that it is unlikely to be encountered in empirical applications (especially if W has rich support).

Case (ii) is more interesting, as it is possible to give more interpretable conditions under which it might occur. For example, since our approach is based on comparing the subpopulation with $X = 0$ to the one with small positive realizations of X, it is unable to detect violations of $\mathbb{H}_0$ which only affect those individuals with $X > \delta$ for some $\delta > 0$. That is, if the data are generated under a fixed alternative which is such that Assumption 1 only holds within the subpopulation with $X < \delta$ for some $\delta > 0$, then $\bar{\Delta}(w,r) \equiv 0$, and our approach would not be able to detect the violation. In this case, the control variable approach would for example correctly identify the Average Structural Function $\mathbb{E}(m_1(x, w, U))$ for $x < \delta$, but not for $x \geq \delta$.

We would like to stress that we do not consider the fact that there are alternatives that our approach is unable to detect a fundamental flaw of this method. Instead, with Proposition 2 in mind, we see this as an indication of the difficulty to derive *any* testable implication from a condition like Assumption 1.

3.4. Some Numerical Evidence. To further investigate the question of detectable alternatives, we numerically calculate the population value of Δ in a number of different settings. For simplicity, we consider a case in which there are no additional exogenous covariates W. The distribution of the remaining observable quantities (Y, X, Z) is determined by the following class of data generating

Table 1: Population Value of Δ for various Data Generating Processes

Parameter ($\alpha/\beta/\gamma$):	0	.2	.4	.6	.8	1	1.2	1.4	1.6	1.8	2
DPG1 (varying α)	.00	.00	.02	.01	-.08	-.41	-.26	-.27	-.27	-.27	-.26
DPG2 (varying β)	.00	-.03	-.07	-.12	-.17	-.21	-.25	-.28	-.30	-.32	-.34
DPG3 (varying γ)	.00	.21	.40	.57	.73	.85	.95	1.04	1.10	1.16	1.20

processes (DGPs):

$$Y = U_1 + U_2 X$$
$$X = \max\{0, 1 + V + (1 + \alpha V + \beta\bar{\varepsilon}) \cdot Z\},$$

where $\bar{\varepsilon} = \varepsilon^2/\sqrt{2}$, $U_j = (V + \beta\bar{\varepsilon} + \gamma Z + \eta_j)/\sqrt{2 + \beta^2 + \gamma^2}$ for $j = 1, 2$, $(Z, V, \varepsilon, \eta_1, \eta_2) \sim N(0, I)$, and I denotes the identity matrix. The coefficients α, β and γ are real-valued and will be varied for our numerical calculations. Note that for $\alpha = \beta = \gamma = 0$ the DGP corresponds to a linear random coefficient model in the outcome equation and a standard censored linear model in the first stage equation, and thus clearly satisfies $\mathbb{H}_0$. If $\alpha \neq 0$ the monotonicity condition in Assumption 1(iii) fails, for $\gamma \neq 0$ the independence condition Assumption 1(i) is violated, and when $\beta \neq 0$ the first stage no longer only contains a scalar unobservable random variable, and thus Assumption 1(ii) does not hold. We then calculate the value of Δ as defined in (3.1) for the following three scenarios:

- DGP1: $\alpha = 0, .2, .4 \ldots, 2$ and $\beta = \gamma = 0$.
- DGP2: $\beta = 0, .2, .4, \ldots, 2$ and $\alpha = \gamma = 0$.
- DGP3: $\gamma = 0, .2, .4, \ldots, 2$ and $\alpha = \beta = 0$.

The results, given in Table 1, show that our approach is indeed able to pick up all three types of violations of the null hypothesis. We would therefore expect an empirical version of our test to have power against these alternatives. Note that within each of the three DGPs the value of the varying parameter can be interpreted as a measure of distance from the null hypothesis. Since we are only testing a necessary condition for $\mathbb{H}_0$, the value of Δ does not necessarily have to be monotone in the value of that parameter, even though this turns out to be the case for DGP2 and DGP3. On

the other hand, under DGP1 the value of Δ does not vary monotonically with α, and even changes sign over the grid that we consider. In particular, further calculations show that for $\alpha \approx .625$ the joint distribution of (Y, X, Z) is such that the integral on the right-hand-side of equation (3.1) is incidentally equal to $\mathbb{E}(Y|X = 0)$, and thus $\Delta = 0$ in this case. Thus DGP1 with $\alpha \approx .625$ is an example of a violation of $\mathbb{H}_0$ that our approach is unable to detect.

4. Concluding Remarks

In this paper, we have derived a testable implication of the validity of the control variable approach to identification in triangular nonseparable models with endogeneity. To the best of our knowledge, this the first point out a non-trivial testable implication of this particular type of hypothesis. Our approach requires a special 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 setup is certainly restrictive, we argue that the idea is still applicable in a wide range of empirical settings. In the appendix, we show how our approach can be implemented in the context of an empirical application. We propose a test statistic that is easy to compute, show that it is asymptotically normal and derive an explicit formula for its asymptotic variance, which can be estimated to obtain critical values.

A. A Test Statistic and its Theoretical Properties

In this appendix, we describe a feasible approach to test our null hypothesis in an empirical application. For simplicity, we focus on the case in which the distribution of the exogenous covariates W is discrete with support $\{w_1, \ldots, w_K\}$, but extensions to settings with continuous covariates are conceptually straightforward. The main idea behind the construction of our test statistic is to take a sample analogue $\widehat{\Delta}(w)$ of $\Delta(w)$, and to reject the null hypothesis if the vector

$$\widehat{\Delta} := (\widehat{\Delta}(w_1), \ldots, \widehat{\Delta}(w_K))$$

is "too large" in some norm, such as L_2. We construct such a sample analogue as

$$\widehat{\Delta}(w) = \int \widehat{\mu}^+(w, r)\widehat{\Gamma}(dr, w) - \widehat{\mu}(w), \tag{A.1}$$

where $\widehat{\mu}^+(w,v)$ and $\widehat{\mu}(w)$ are nonparametric estimates of $\lim_{x\downarrow 0}\mathbb{E}(Y|X=x,W=w,R=v)$ and $\mathbb{E}(Y|X=0,W=w)$, respectively, and $\widehat{\Gamma}(r,w)$ is a nonparametric estimate of the function $\Gamma(r,w) := F_{R^*|X,W}(r;0,w)$. The construction of these estimates is described below. In Theorem 2, we then show that the vector $\widehat{\Delta}$ is asymptotically normal under standard regularity conditions, and characterize the asymptotic variance.

A.1. Estimation of Δ. We start by describing the construction of the various components that make up the estimate $\widehat{\Delta}$ of $\Delta := (\Delta(w_1),\ldots,\Delta(w_K))$. The data are given by an i.i.d. sample $\{(Y_i,X_i,Z_i,W_i)\}_{i=1}^n$ of size n from the distribution of (Y,X,Z,W). First, we estimate the conditional distribution function $F_{X|Z,W}$ of X given (Z,W) by local linear estimation (Fan and Gijbels, 1996):

$$\widehat{F}_{X|Z,W}(x,z,w) = e_{1,d_z}^\top \operatorname*{argmin}_{(a_1,a_2^\top)} \sum_{i=1}^n \left(\mathbb{I}\{X_i\le x\} - a_1 - a_2^\top(Z_i-z)\right)^2 K_h(Z_i-z)\mathbb{I}\{W_i=w\}.$$

Here $K_g(z) = \prod_{j=1}^{d_z}\mathcal{K}(z_j/g)/g$ is a d_z-dimensional product kernel built from the univariate kernel function $\mathcal{K}$, g is a one-dimensional bandwidth that tends to zero as the sample size n tends to infinity, and $e_{1,d_z} = (1,0,\ldots,0)^\top$ denotes the first unit (d_z+1)-vector. In a second step, we then use this estimated CDF to define estimates $\{\widehat{R}_i\}_{i=1}^n$ of the realizations of the unobserved but identified random variable $R = F_{X|Z,W}(X;Z,W)$ as

$$\widehat{R}_i = \widehat{F}_{X|Z,W}(X_i,Z_i,W_i) \text{ for } i=1,\ldots n. \tag{A.2}$$

Third, we estimate the function $\Gamma(r,w)$ defined in (3.2) by

$$\widehat{\Gamma}(r,w) = \frac{F_{U[0,1]}(r) - \sum_{i=1}^n \mathbb{I}\{\widehat{R}_i\le v, X_i>0, W_i=w\}/\sum_{i=1}^n\mathbb{I}\{W_i=w\}}{\widehat{\pi}(w)},$$

where $F_{U[0,1]}$ is the CDF of the standard uniform distribution, and

$$\widehat{\pi}(w) = \frac{\sum_{i=1}^n \mathbb{I}\{X_i=0,W_i=w\}}{\sum_{i=1}^n\mathbb{I}\{W_i=w\}}$$

is the natural estimate of the conditional probability $P(X=0|W=w)$. Fourth, we define the estimate $\widehat{\mu}^+(r,w)$ of the function $\lim_{x\downarrow 0}\mathbb{E}(Y|X=x,W=w,R=r)$ as

$$\widehat{\mu}^+(r,w) = e_{1,2}^\top \operatorname*{argmin}_{(a_1,a_2^\top)} \sum_{i=1}^n \left(Y_i - a_1 - a_2^\top(X_i,\widehat{R}_i-r)\right)^2 K_h(X_i,\widehat{R}_i-r)\mathbb{I}\{X_i>0,W_i=w\},$$

where $K_h(x,r) = \mathcal{K}(x/h)\mathcal{K}(r/h)/h^2$ is a bivariate 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 $e_{1,2} = (1,0,0)^\top$. Finally, we define the estimate $\widehat{\mu}(w)$ of $\mathbb{E}(Y|X=0,W=w)$ as a sample average of the observed outcomes Y_i among those observations with $(X_i,W_i) = (0,w)$:

$$\widehat{\mu}(w) = \frac{\sum_{i=1}^n Y_i \mathbb{I}\{X_i = 0, W_i = w\}}{\sum_{i=1}^n \mathbb{I}\{X_i = 0, W_i = w\}}.$$

For every $w \in \{w_1, \ldots, w_K\}$, the statistic $\widehat{\Delta}(w)$ is then constructed as described in (A.1). Note that because of the particular structure of the estimate $\widehat{\Gamma}$, the expression given there simplifies to

$$\widehat{\Delta}(w) = \frac{1}{\widehat{\pi}(w)} \left(\int_0^1 \widehat{\mu}^+(r,w)dv - \frac{1}{n}\sum_{i=1}^n \widehat{\mu}^+(\widehat{R}_i, w)\mathbb{I}\{X_i > 0, W_i = w\} \right) - \widehat{\mu}(w).$$

The computation of $\widehat{\Delta}(w)$ is thus straightforward, as it only involves calculating sample averages and a one-dimensional numerical integration problem.

A.2. Asymptotic Theory. Deriving the theoretical properties of $\widehat{\Delta}$ is a non-standard problem because its construction involves a nonparametric regression on the estimated data points $\{\widehat{R}_i\}_{i=1}^n$. We address this issue by using recent results in Mammen, Rothe, and Schienle (2012, 2014) on nonparametric regression with generated covariates. Making use of these 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 bandwidths, and kernel function $\mathcal{K}$.*

(i) For every $w \in \mathcal{S}(W)$, the random vector Z is continuously distributed conditional on $W = w$ with support $S_{Z|w} = \mathcal{S}(Z|W=w) \subset \mathbb{R}^{d_Z}$. The corresponding conditional density function $f_{Z|w}(\cdot)$ is continuously differentiable, bounded, and bounded away from zero on $S_{Z|w}$ for every $w \in \mathcal{S}(W)$.

(ii) The conditional CDF $F_{X|Z,W}(x,z,w)$ of X given (Z,W) is twice continuously differentiable with respect to its second argument on $S_{Z|w}$ for every $w \in \mathcal{S}(W)$.

(iii) For every $w \in \mathcal{S}(W)$, the random vector (X,R) is continuously distributed conditional on $X > 0$ and $W = w$ with support $S_{XR|X>0,w} = \mathcal{S}(X,R|X>0,W=w)$. The corresponding conditional density function $f_{XR|X>0,w}(\cdot)$ is continuously differentiable, bounded, and bounded away from zero on the compact set $S_{\delta,w} = \{(x,v) : (x,v) \in S_{XR|X>0,w} \text{ and } x \leq \delta\}$ with $\delta > 0$ as in (2.5) and every $w \in \mathcal{S}(W)$.

(iv) *The conditional expectation function* $\mathbb{E}(Y|X=x,W=w,R=r)$ *is twice continuously differentiable in (x,r) on* $S_{\delta,w}$ *for every* $w \in \mathcal{S}(W)$.

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

(vi) *The kernel 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 $\widehat{F}_{X|ZW}(x,z,W)$ and $\widehat{\mu}^+(r,w)$ are stable over their respective range of evaluation. Parts (ii) and (iv) are smoothness conditions used to control the magnitude of certain bias terms. Assuming subexponential tails of ε conditional on (X,W,R) in the subpopulation with $X > 0$ in part (v) is necessary to apply certain results from Mammen, Rothe, and Schienle (2012, 2014) 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, 2014).

Assumption 3 allows us to derive the limiting distribution of the random vector $\widehat{\Delta}$. To state the result, we define $f^+_{R|XW}(r,0,w) = \lim_{x\downarrow 0} f_{R|XW}(r,x,w)$ and $f^+_{XW}(0,w) = \lim_{x\downarrow 0} f_{XW}(x,w)$, let $\gamma(r,w) = \partial\Gamma(r,w)/\partial v$, and put

$$\sigma^2_+(w) = \lim_{x\downarrow 0} \mathrm{Var}\left(\varepsilon \cdot \frac{\gamma(R,W)}{f^+_{R|X,W}(R,0,W)} \middle| X = x, W = w \right).$$

where $\varepsilon = Y - \mathbb{E}(Y|X,W,R^*)$ as in Assumption 3(v). Note that under the null hypothesis the function $\sigma^2_+(w)$ can be expressed in the following, somewhat more intuitive form:

$$\sigma^2_+(w) = \lim_{x\downarrow 0} \mathrm{Var}\left(\varepsilon \cdot \frac{f_{V|XW}(V,0,W)}{f^+_{V|X,W}(V,0,W)} \middle| X = x, W = w \right).$$

Also, for $j \in \{0,1,2\}$ we define the constants

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

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

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

and finally let

$$\rho^2(w) = C \cdot \frac{\sigma_+^2(w)}{f_{XW}^+(0,w)}.$$

With this notation, we obtain the following result.

Theorem 2. *Suppose that Assumption 3 holds. Then*

$$\sqrt{nh}\left(\widehat{\Delta} - \Delta\right) \xrightarrow{d} N(0, \mathrm{diag}(\rho^2(w_1), \ldots, \rho^2(w_K))).$$

Proof. See Section A.4 below. □

The theorem shows that in large samples the distribution of the vector $\widehat{\Delta}$ approaches a multivariate normal distribution with mean Δ and diagonal variance matrix $\mathrm{diag}(\rho^2(w_1), \ldots, \rho^2(w_K))/(nh)$. Note that the rate of convergence is the same as that of a standard one-dimensional kernel smoother.

A.3. Test Statistic and Critical Values. Given the result in Theorem 2, a natural test statistic for the testing problem in (2.6) is given by

$$T_n = nh \sum_{k=1}^{K} \left(\frac{\widehat{\Delta}(w_k)}{\widehat{\rho}(w_k)} \right)^2, \tag{A.3}$$

where $\widehat{\rho}^2(w)$ is some consistent estimate of $\rho^2(w)$ for every $w \in \{w_1, \ldots, w_K\}$. Under the conditions of Theorem 2, this test statistic should asymptotically follow a χ^2-distribution with K degrees of freedom under the null hypothesis, and a non-central χ^2-distribution under the alternative. The testing decision is thus to reject $\mathbb{H}_0$ at the nominal level $\alpha \in (0,1)$ if

$$T_n > \chi_K^2(1-\alpha),$$

where $\chi^2_K(\tau)$ denotes the τ-quantile of the χ^2-distribution with K degrees of freedom. The following result formally shows the validity of such an approach.

Theorem 3. *Suppose that Assumption 3 holds, and that* $\widehat{\rho}^2(w) \xrightarrow{p} \rho^2(w)$ *for all* $w \in \mathcal{S}(W)$. *Then the following statements hold.*

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

$$\lim_{n\to\infty} \Pr(T_n > \chi^2_K(1-\alpha)) = \alpha.$$

(ii) Under any fixed alternative that implies $\Delta(w) \neq 0$ *for some* $w \in \mathcal{S}(W)$,

$$\lim_{n\to\infty} \Pr(T_n > \chi^2_K(1-\alpha)) = 1.$$

(iii) Under any local alternative that implies $\Delta(w) = \delta(w)/\sqrt{nh}$ *for all* $w \in \mathcal{S}(W)$,

$$\lim_{n\to\infty} \Pr(T_n > \chi^2_K(1-\alpha)) = 1 - \Xi_{\Gamma,K}(\chi^2_K(1-\alpha))$$

where $\Xi_{\theta,K}$ *is the CDF of the noncentral* χ^2 *distribution with* K *degrees of freedom and noncentrality parameter* $\theta = \sum_{k=1}^K (\delta(w_k)/\rho(w_k))^2$.

Proof. Follows from Theorem 2 using straightforward arguments. □

Since Theorem 3 holds for any consistent estimator $\widehat{\rho}^2(w)$ of $\rho^2(w)$, the only remaining issue for applying our test is to find such an estimator that is feasible to compute in the context of an empirical application. One possible approach would be to develop a direct sample-analogue estimator using boundary-corrected nonparametric estimates of the various components of $\rho^2(w)$. Such an estimator is explicitly described in Section A.5 below. However, the approach can be unattractive in practice because it requires additional smoothing parameters. In some preliminary simulation that we conducted, the performance of the procedure was also quite sensitive with respect to the choice of smoothing parameters. We therefore recommend the use of a nonparametric bootstrap variance estimator. Such a procedure is computationally expensive, but straightforward from a practical point of view. The estimator is obtained as follows. Let $\{(Y_i^*, X_i^*, Z_i^*, W_i^*)\}_{i=1}^n$ be a bootstrap sample drawn with replacement from the observed data $\{(Y_i, X_i, Z_i, W_i)\}_{i=1}^n$, and let $\widehat{\Delta}^*(w)$ be

an estimate of $\Delta(w)$ computed exactly as described above but using the bootstrap sample. Then

$$\widehat{\rho}^2(w) = \mathbb{E}^*((\widehat{\Delta}^*(w) - \widehat{\Delta}(w))^2),$$

where $\mathbb{E}^*$ denotes the expectation with respect to bootstrap sampling.

A.4. Proof of Theorem 2. The result in Theorem 2 follows directly from the three axillary results in Lemma 1–3 below. To simplify the notation, we assume that the additional exogenous covariates W are absent from the model, and can thus be dropped from the notation in the following. That is, we have $\widehat{\Delta}(w) \equiv \widehat{\Delta}$, $\Delta(w) \equiv \Delta$, etc. The first of these three findings gives a bound on the uniform rate of consistency of the estimated function $\widehat{\Gamma}(\cdot)$.

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

$$\sup_{r \in \mathcal{S}(R|X=0)} |\widehat{\Gamma}(r) - \Gamma(r)| = O_P(n^{-1/2}) + O(g^2).$$

Proof. Up to terms that are clearly $O_P(n^{-1/2})$, the estimate $\widehat{\Gamma}(\cdot)$ is equal to a continuous and deterministic transformation of the empirical distribution function of the estimates $\{\widehat{R}_i\}_{i=1}^n$ 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 $\mu^+(r) = \lim_{x \downarrow 0} \mathbb{E}(Y|X = x, R = r)$ that uses the actual realizations of $R_i = F_{X|Z}(X_i, Z_i)$ instead of the corresponding estimated values $\widehat{R}_i$. The corresponding estimator is denoted by $\widetilde{\mu}^+_{Y|X,R}(0, r)$. We also define an infeasible version of our test statistic which uses the population values $\Gamma(\cdot)$ and $\mathbb{E}(Y|X = 0)$ instead of their estimates, and replaces $\widehat{\mu}^+(r)$ with its infeasible version $\widetilde{\mu}^+(r)$:

$$\widetilde{\Delta} = \int \widetilde{\mu}^+(r) d\Gamma(r) - \mathbb{E}(Y|X = 0).$$

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

Lemma 2. *Suppose that the conditions of Theorem 2 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:

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

With this notation, the local linear estimator $\widetilde{\mu}^+(r)$ can be written as

$$\widetilde{\mu}^+(r) = \frac{1}{n}\sum_{i=1}^{n} e_1^\top M_n(r)^{-1} L_i(r) K_h(X_i, R_i - r) Y_i.$$

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

$$\begin{aligned} N_n(r) &= A \lim_{x \downarrow 0} f_{RX}(r, x) + o(1) \\ &= A f^+_{R|X}(r, 0) f^+_X(0) + o(1) \end{aligned}$$

uniformly in r, where the matrix A is given by

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

Note that the structure 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

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

with $\varepsilon_i = Y_i - \mathbb{E}(Y_i | X_i, R_i^*)$, we have that

$$\widetilde{\mu}^+(r) = \mu^+(r) + S_n(r) + O(h^2) + O_P\left(\frac{\log(n)}{nh^2}\right)$$

uniformly over $r \in \mathcal{S}(R^*|X=0)$. Using standard change-of-variables arguments, we find that

$$\int S_n(r)d\Gamma(r) = \frac{1}{n}\sum_{i=1}^{n} e_1^\top N_n(R_i)^{-1} L_i^* K_h(X_i)\Gamma(R_i)\varepsilon_i + O(h^2)$$

with $L_i^* = (1, X_i/h, 0)^\top \cdot \mathbb{I}\{X_i > 0\}$. 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

$$\begin{aligned}
&n^{-1}\mathbb{E}((e_1^\top N_n(R_i)^{-1} L_i^*)^2 K_h(X_i)^2 \Gamma(R_i)^2 \varepsilon_i^2) \\
&\quad = \frac{1}{nhf_X^+(0)^2}\int_0^\infty (e_1^\top A^{-1}(1,x,0)^\top)^2 \mathcal{K}(x)^2 \mathbb{E}\left(\frac{\Gamma(R)^2}{f_{V|X}^+(R,0)^2}\cdot\varepsilon^2 \middle| X = xh\right) f_X(xh)dx + o\left(\frac{1}{nh}\right) \\
&\quad = \frac{1}{nhf_X^+(0)^2}\int_0^\infty (e_1^\top A^{-1}(1,x,0)^\top)^2 \mathcal{K}(x)^2 dx \cdot \lim_{x\downarrow 0}\mathbb{E}\left(\frac{\Gamma(R)^2}{f_{V|X}^+(R,0)^2}\cdot\varepsilon^2 \middle| X = xh\right) \cdot \lim_{x\downarrow 0} f_X(x) + o\left(\frac{1}{nh}\right) \\
&\quad = \frac{1}{nhf_X^+(0)}\cdot C \cdot \sigma_+^2(0) + o\left(\frac{1}{nh}\right).
\end{aligned}$$

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

As the final step of our proof of Theorem 2, 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 $\widehat{\mu}(0) = \mathbb{E}(Y|X=0) + O_P(n^{-1/2})$ and Lemma 1, we find that

$$\widehat{\Delta} = \int \widehat{\mu}^+(r)d\Gamma(r) - \mathbb{E}(Y|X=0) + O_P(n^{-1/2}),$$

since $\widehat{\mu}^+_{Y|X,V}(0,v)$ is easily seen to be a consistent estimate of a bounded function under the conditions of the lemma. Similarly, we have that

$$\widetilde{\Delta} = \int \widetilde{\mu}^+(r)d\Gamma(r) - \mathbb{E}(Y|X=0) + O_P(n^{-1/2}),$$

It only remains to be shown that

$$\int \widehat{\mu}^+(r)d\Gamma(r) = \int \widetilde{\mu}^+(r)d\Gamma(r) + o_P((nh)^{-1/2}).$$

We use recent results on nonparametric regression with generated covariates obtained by Mammen, Rothe, and Schienle (2012, 2014) to show this statement. For convenience, we repeat the following notation, which was already introduced in the proof of Lemma 2:

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

It then follows from an application of Theorem 1 in Mammen, Rothe, and Schienle (2014) that

$$\int \widehat{\mu}^+(r) - \widetilde{\mu}^+(r) - \varphi_n(r; \widehat{F}_{X|Z})d\Gamma(r) = o_P((nh)^{-1/2})$$

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

$$\begin{aligned} \varphi_n(r;\Lambda) = &-(\partial m^+(r)/\partial v)e_1^\top N_n(r)^{-1}\mathbb{E}(L_i(r)K_h(X_i, R_i - v)(\Lambda(X_i, Z_i) - F_{X|Z}(X_i, Z_i))) \\ &+ e_1^\top N_n(r)^{-1}\mathbb{E}(L_i(r)K_h'(X_i, R_i - v)(\Lambda(X_i, Z_i) - F_{X|Z}(X_i, Z_i))\Psi(X_i, Z_i)) \end{aligned}$$

with $\Psi(X_i, Z_i) = \mathbb{E}(Y_i|X_i, Z_i) - \mathbb{E}(Y_i|X_i, R_i)$, and $K_h'(x, r) = \partial K_h(x, r)/\partial r$. We remark that the two expectations in the previous equation are both taken with respect to the distribution of (X_i, Z_i), so that the term $\varphi_n(r; \widehat{F}_{X|Z})$ remains a random variable due to its dependence on the estimate $\widehat{F}_{X|Z}$. Also note that under the null hypothesis the second summand in the formula for φ_n vanishes, because if the model is correctly specified it holds that $\Psi(X_i, Z_i) = 0$. As a consequence, the "index bias" term in Mammen, Rothe, and Schienle (2014) is equal to zero. Next, it follows from the same arguments as in the proof of Theorem 4 in Mammen, Rothe, and Schienle (2014) that

$$\int \varphi_n(r; \widehat{F}_{X|Z})d\Gamma(r) = O_P(n^{-1/2}) + O(h^2) + O(g^2) + O_P\left(\frac{\log n}{ng}\right).$$

This completes our proof. □

A.5. An Alternative Estimate of the Asymptotic Variance. In this section, we describe a plug-in estimator of $\rho^2(w)$ that uses kernel-based nonparametric smoothers to estimate the various density and conditional expectation functions involved in the definition of the asymptotic variance of $\widehat{\Delta}$. Some technical complications arise from the fact that many of these functions need to be evaluated at or close to the limits of their support. This is a problem for standard kernel estimators, which are well-known to be inconsistent at the boundary, and highly biased in its vicinity. Since we only require a consistent estimate of $\rho^2(w)$, and not one that converges with a particular rate, we adopt a simple solution to this problem and introduce a multiplicative correction term into all estimators of density functions. More elaborate procedures could be used to achieve better rates of convergence, but those are not necessary for our main results. The boundary correction terms are of the form

$$s_b(r) = \bar{\mathcal{K}}(\min\{r, 1-r\}/b)^{-1} \text{ with } \bar{\mathcal{K}}(t) = \int_{-\infty}^{t} \mathcal{K}(u)du \tag{A.4}$$

for any $b \in \mathbb{R}$ and $r \in (0,1)$. We then estimate the function $\gamma(r,w) = \partial\Gamma(r,w)/\partial v$ by the sample analogue

$$\widehat{\gamma}(r,w) = (1 - \widehat{g}(r,w))/\widehat{p}(w),$$

where

$$\widehat{g}(r,w) = s_{b_1}(r) \cdot \frac{\sum_{i=1}^n K_{b_1}(\widehat{R}_i - v)\mathbb{I}\{X_i > 0, W_i = w\}}{\sum_{i=1}^n \mathbb{I}\{W_i = w\}}.$$

Here b_1 is a one-dimensional bandwidth that tends to zero as n tends to infinity. By including the boundary correction term $s_{b_1}(r)$ into the definition of $\widehat{g}(r,w)$, we achieve that the estimator $\widehat{\gamma}$ is uniformly consistent under weak regularity conditions. We also define

$$\widehat{f}^{+}_{R|X,W}(r;x,w) = s_{b_2}(r) \cdot \frac{\sum_{i=1}^n K_{b_2}(\widehat{R}_i - v, X_i - x)\mathbb{I}\{X_i > 0, W_i = w\}}{\sum_{i=1}^n K_{b_2}(X_i - x)\mathbb{I}\{X_i > 0, W_i = w\}} \text{ and}$$
$$\widehat{f}^{+}_{XW}(0,w) = \frac{2}{n}\sum_{i=1}^n K_{b_1}(X_i)\mathbb{I}\{X_i > 0, W_i = w\},$$

where b_2 is another one-dimensional bandwidth that tends to zero as n tends to infinity. Again, consistency of these two estimates for the corresponding population counterparts is achieved by including a boundary correction term for the first estimator, and multiplication by two for the second estimator. The estimate the

term $\sigma^2_+(w)$ is given by

$$\widehat{\sigma}^2_+(w) = e_{1,1}^\top \underset{(a_1,a_2)}{\operatorname{argmin}} \sum_{i=1}^n \left(\widehat{\eta}_i - a_1 - a_2(X_i - x)\right)^2 K_{b_1}(X_i - x)\mathbb{I}\{X_i > 0, W_i = w\}.$$

where

$$\widehat{\eta}_i = (Y_i - \widehat{\mu}(X_i, W_i, \widehat{R}_i)) \cdot \frac{\widehat{\gamma}(\widehat{R}_i, W_i)}{\widehat{f}_{R|XW}(\widehat{R}_i, 0, W_i)},$$

and

$$\widehat{\mu}(x, w, v) = \\ e_{1,2}^\top \underset{(a_1,a_2^\top)}{\operatorname{argmin}} \sum_{i=1}^n \left(Y_i - a_1 - a_2^\top (X_i - x, \widehat{R}_i - v)\right)^2 K_h(X_i - x, \widehat{R}_i - v)\mathbb{I}\{X_i > 0, W_i = w\},$$

which is similar in structure to the estimate $\widehat{\mu}^+(w, v)$ defined above. Our final estimator of $\rho^2(w)$ then given by

$$\widehat{\rho}^2(w) = C \cdot \frac{\widehat{\sigma}^2_+(w)}{\widehat{f}^+_{XW}(0, w)}.$$

The constant C depends on the kernel function and can be computed numerically. For example, $C \approx 1.78581$ for the the Gaussian kernel.

Theorem 4. *Suppose that Assumption 3 holds, and that $b_j \to 0$, $nb_j \to \infty$ and $(ng^{d_z}/\log(n) + g^{-4})/b_j^2 \to 0$ as $n \to \infty$ for $j = 1, 2$. Then $\widehat{\rho}^2(w) \xrightarrow{p} \rho^2(w)$ for all $w \in \mathcal{S}(W)$.*

Proof. This result follows from straightforward arguments. Let

$$g(r, w) = \partial \Pr(R \leq r, X > 0 | W = w)/\partial v$$

be the population counterpart of $\widehat{g}(r, w)$, and

$$\widetilde{g}(r, w) = s_{b_1}(r) \cdot \frac{\sum_{i=1}^n K_{b_1}(R_i - v)\mathbb{I}\{X_i > 0, W_i = w\}}{\sum_{i=1}^n \mathbb{I}\{W_i = w\}}.$$

be an infeasible estimator of $g(r, w)$ that uses the the true $R_i = F_{X|ZW}(X, Z, W)$ instead of the estimates

$\widehat{R}_i = \widehat{F}_{X|ZW}(X, Z, W)$. From a simple Taylor expansion, it follows that

$$\sup_{v,w} |\widehat{g}(r,w) - \widetilde{g}(r,w)| = O_P\left(\max_{i=1,\ldots,n} |\widehat{R}_i - R_i|/b_1\right) = o_p(1)$$

since $\max_{i=1,\ldots,n} |\widehat{R}_i - R_i| = O_P((ng^{d_z}/\log(n))^{1/2}) + O(g^{-2})$. Moreover, standard results from kernel density estimation imply that

$$\sup_{v,w} |\widetilde{g}(r,w) - g(r,w)| = o_p(1).$$

Similar arguments can be used to show that $\widehat{f}^{+}_{R|XW}(r;0,w)$ and $\widehat{f}^{+}_{XW}(0,w)$ are uniformly consistent estimates of $\lim_{x\downarrow 0} f_{R|XW}(r;x,w)$ and $\lim_{x\downarrow 0} f_{XW}(x,w)$, respectively. Consistency of $\widehat{\sigma}^2_{+}(0)$ for $\sigma^2_{+}(0)$ then follows from the linearity of the local linear smoothing operator. 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 R. L. Matzkin (2010): "Conditions for the existence of control functions in nonseparable simultaneous equations models," *Working Paper*.

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. (2014): "A Discontinuity Test of Endogeneity," *Working Paper*.

Canay, I. A., A. Santos, and A. M. Shaikh (2013): "On the testability of identification in some nonparametric models with endogeneity," *Econometrica*, 81(6), 2535–2559.

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.

Chesher, A. (2003): "Identification in nonseparable models," *Econometrica*, 71(5), 1405–1441.

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

d'Haultfoeuille, X., and P. Février (2012): "Identification of nonseparable models with endogeneity and discrete instruments," *Working Paper*.

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," *Annals of Statistics*, 33(6), 2904–2929.

Heckman, J. J. (2001): "Micro data, heterogeneity, and the evaluation of public policy," *Journal of Political Economy*, 109(4), 673–748.

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. Cambridge University Press.

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

——— (2014): "Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment," *Review of Economic Studies*, to appear.

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

——— (2014): "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.

Matzkin, R. L. (2003): "Nonparametric estimation of nonadditive random functions," *Econometrica*, 71(5), 1339–1375.

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.

Torgovitsky, A. (2012): "Identification of nonseparable models with general instruments," *Working paper*.