A Discontinuity Test for Identification in Triangular Nonseparable Models

Carolina Caetano, Christoph Rothe, and Neşe 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 model, triangular systems, control variable, instrument validity, nonparametric identification.*

[*] This Version: January 26, 2016. 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. We would like to thank Gregorio Caetano, three anonymous referees, and audiences at several seminar and conference presentations for their helpful comments.

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 (e.g. 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 structural parameters. Availability of an instrument, say Z, that is uncorrelated with U but correlated with X does generally not suffice for identification of meaningful quantities. Instead, one has to impose additional conditions. One popular approach is based on control variables (e.g. Blundell and Powell, 2003; Imbens and Newey, 2009). This method 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.

A control variable arises for example if the model has a triangular structure. This means that 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 $R \equiv F_{X|Z}(X, Z) = F_V(V)$ is a valid control variable, where $F_{X|Z}$ denotes the conditional CDF of X given Z and F_V denotes the unconditional CDF of V. This is because in such a model R is a one-to-one transformation of V, and conditional on V all variation in the endogenous variable X comes from variation in Z.

While this approach to identification is powerful, the postulated triangular specification paired with the restrictions on first stage unobserved heterogeneity impose substantial limitations for the underlying economic model. For example, the condition that X depends monotonically on V implies that individuals with common values of X and Z will react in exactly the same way to exogenous variation in Z. In many empirical contexts such a behavioral restriction can be difficult to justify through theoretical considerations alone, and thus its validity might be doubtful. Other conditions, like the full independence of Z and (U, V) might be equally questionable in practice. Unfortunately, one cannot simply adapt established methods for specification testing in models with additively separable disturbances to settings with nonseparable unobserved heterogeneity. For example, tests for instrument validity based on overidentifying restrictions, such as those of Sargan (1958) and Hansen (1982), have no direct analogue in triangular nonseparable models: if the identifying assumptions mentioned above hold with some vector of instruments Z, there is no guarantee that that they will also hold (*mutatis mutandis*) with some subvector of Z.

In this paper, we propose a test of the conditions necessary for justifying a control variable approach in nonseparable models with a particular additional structure. Specifically, we study the case where the data generating process is such that the distribution of the endogenous covariate X has a mass point at some known value, is otherwise continuously distributed, and exerts a continuous effect on the outcome variable of interest. In most applications, the location of the mass point coincides with the lower or upper boundary of the support of X. We show that in this case 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. Validity of the control function approach can thus be potentially refuted using data and a weak set of maintained assumptions alone. To the best of our knowledge, our paper is the first to consider this type of testable implications of identifying assumptions for this type of nonseparable models.

The idea behind the derivation of our testable implication is related to that of Caetano (2015), 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. If not 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, which yields a continuity condition on an identified function.

We also propose a test statistic that is based on a direct sample analogue of the function whose continuity we wish to verify. Its computation only involves standard nonparametric regression techniques and a simple numerical integration step. Deriving its asymptotic properties is a non-standard problem, as it involves nonparametric regression with estimated data points. We use recent results in Mammen et al. (2012, 2015) on generated covariates in non- and semiparametric models to account for this two-stage structure. Through a Monte Carlo study, we show that this leads to a test with good size and power properties in finite samples.

1.1. Potential Applications. Focusing on settings where the endogenous variable has a mass point at the lower or upper boundary of its support is clearly restrictive, but such scenarios can still be found in a wide array of applications. If the endogenous variable is a choice which must be non-negative, such as the quantity consumed of a particular good, our methods can usually be applied. One such case is the problem of estimating the effects of smoking during pregnancy on the baby's birthweight. Smoking amounts cannot be negative, and a sizable proportion of the pregnant women do not smoke. The literature on this topic is extensive (see e.g. Kramer, 1987, 1998; Vogler and Kozlowski, 2002; Almond et al., 2005; Tominey, 2007; Fertig, 2010). Papers using instrumental variable approaches to address the problem of the endogeneity of smoking include Evans and Ringel (1999) and Lien and Evans (2005), who use tax variations across locations and time, and Wehby et al. (2011), who use genetic markers that predict smoking behavior as instruments for smoking. This literature generally focuses on linear models, but our approach could be used to investigate the validity of a triangular nonseparable specification in such a context.

Another interesting class of examples are studies of the effects of labor supply, expressed as "hours of work," on different outcomes. See Blundell and MaCurdy (1999) for a survey on the extensive literature on this topic. Approaches based on instrumental variables include, for example, Heckman and MaCurdy (1980)'s study of the effect of female labor supply on female wages, which uses the wages of the husband or other non-female generated sources of income as instrument for female hours worked. Another example is Connelly and Kimmel (2009), who study of the effect of labor supply on time spent in childcare, using age and education squared, as well as spouse's age and education as instruments for hours of work. A third example can be found in Blau and Grossberg (1992), who study the effect of maternal labor supply on the child's cognitive development, using supply side determinants of maternal labor supply, such as predicted wage, as instruments for the maternal hours worked. Hours of work must of course be non-negative, and a large part of the population supply exactly zero hours, and thus our methods can be used to test the validity of a control function approach in these setups.

In other settings the presence of mass points is due to certain legal restrictions. One example is the variable "schooling." Due to minimum attendance laws (and additionally minimum working age laws), students are forced to remain in school until a certain age threshold is reached. The well known literature on the returns to schooling (see Card (1999)) is concerned with endogeneity of "schooling." Several instrumental variables have been used, such as for example Card (1995)'s proximity to college. We can therefore test the validity of a control function based approach in this setting.

Another variable which is also constrained by law is wage, which must be equal to or larger than a pre-established quantity for all legal employment. Examples of problems and instruments include Stewart and Swaffield (1997), which studies the effect of wages on the desired hours of work, and uses as instrument for wages a combination of variables including years of education and the male age-specific regional unemployment rate. Iwata and Tamada (2014) examines the effect of wages on the commuting time of married women, also using a combination of variables as instrument for wages, including the average market wage for women and the age of the respondent. Lydon and Chevalier (2002) investigates the effect of wages on job satisfaction, and uses as instrument for the person's wages their partner's characteristics, such as wage and age. Our test can be applied to test the validity of a control function approach in such problems.

1.2. Related Literature. 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 et al. (1999), Blundell and Powell (2003, 2004), Imbens (2007), Imbens and Newey (2009), Rothe (2009) and Kasy (2011), among others. Instrumental variable (IV) approaches to identification in nonparametric models with additive disturbances are studied in Newey and Powell (2003), Hall and Horowitz (2005), Blundell et al. (2007), or Darolles et al. (2011). Chernozhukov and Hansen (2005), Chernozhukov et al. (2007), Torgovitsky (2015) and D'Haultfoeuille and Février (2011) consider IV methods in nonseparable models, but with restrictions on the dimension of the disturbances. Canay et al. (2013) show that the completeness condition, which plays a central role for identification in nonparametric IV approaches, is generally not testable.

1.3. Plan of the Paper. The remainder of the paper is structured as follows. In Section 2, we introduce the model and the identification approach. In Section 3, we derive our testable implication, and discuss its strength and limitations. In Section 4, we propose a test statistic and study its asymptotic properties. Section 5 contains the results of a Monte Carlo study. Section 6 concludes. Proofs and some extension are collected in the appendix.

2. Model and Identification

2.1. Model. We consider a triangular system of equations as the data generating process similar to the one investigated by Imbens and Newey (2009), focusing on systems that only contain a single potentially endogenous covariate, and no additional exogenous ones.[1] Our setup differs from others

[1] Systems with multiple endogenous covariates would be rather cumbersome to study, although it would be possible in principle. We do not explore this possibility further due to the high prevalence of systems with only one endogenous

considered in the literature in that we assume that the distribution of the endogenous covariate has a mass point at some known value, which, again for simplicity, we take to be the lower boundary of its support. Specifically, our model is given by

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

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

where Y is the outcome of interest, X is a scalar and potentially endogenous covariate, and U and V denote unobserved heterogeneity.[2] We also assume that the data generating process is such that

$$0 < P(X^* \leq 0|Z) < 1 \text{ with probability 1, where } X^* = m_2(Z, V). \tag{2.3}$$

This means that the conditional distribution of X given Z has an actual mass point at 0, but is not degenerate.

The interpretation of negative realizations of the semi-latent variable X^* depends on the respective empirical setting. Suppose for instance that X denotes average daily cigarette consumption. Then individuals with a large negative realization of X^* can be thought of as "very health conscious types" that would need to receive substantial compensation in utility-equivalent units in order to smoke even a single cigarette per day, whereas individuals with X^* close to zero are those that are almost indifferent between smoking and not smoking. When X denotes hourly wages in excess of the legal minimum, a negative value of X^* can be thought of as the normalized market wage an individual would earn in the absence of minimum wage laws. Similar arguments can be made in other empirical contexts.

Finally, we assume that the average of the structural function m_1 satisfies a weak continuity

variable in applications. If there are additional covariates present in an application, the following analysis can be understood as being conditional on these covariates. We return to the issue of how to incorporate covariates into our analysis below.

[2] 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 $m_2(Z, V)$. Similarly, we could allow for a known upper bound on the support of X instead of a lower one. It would also be possible to consider settings with a mass point in the interior of the support, by postulating models like (2.1)–(2.2) separately for the data points located above and below the mass point, respectively.

property:

$$\lim_{x\downarrow 0}\mathbb{E}(m_1(x,U))=\mathbb{E}(m_1(0,U)). \tag{2.4}$$

Continuity of the average structural function seems to be a reasonable assumption for many, although certainly not all, empirical settings.

Introducing some notation that we will use repeatedly in the following, we put $R = F_{X|Z}(X;Z)$ and $R^* = F_{X^*|Z}(X^*;Z)$. These two quantities denote an individual's rank in the conditional distribution of X and its latent counterpart X^* given Z. Furthermore, for generic random variables A and B, we will write $\mathcal{S}(A)$ to denote the support of A, and $\mathcal{S}(A,B|A>0)\equiv\{(a,b):(a,b)\in\mathcal{S}(A,B)\text{ and }a>0\}$ for the intersection of the support of (A,B) with $\mathbb{R}_+\times\mathbb{R}$.

2.2. Identification. 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 $X = m_2(Z,V)$, 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

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

but the same arguments apply to a broader class of structural objects. Note that the ASF describes the average outcome of an individual whose covariate X is exogenously fixed at x. We consider the following identifying assumptions, which are analogous to those in Imbens and Newey (2009).

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

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

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

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

Assumption 2. *$\mathcal{S}(R|X = x) = [0, 1]$ for all $x \in \mathcal{S}(X|X > 0)$.*

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) are also identical in terms of the unobserved heterogeneity V. Assumption 2 is a generalization of the usual rank condition in traditional IV models. It requires the function $z \mapsto m_2(z, V)$ to exhibit a sufficient amount of variation over the support of Z. This condition is strong and arguably not satisfied in many settings. On the other hand, it only involves observable quantities, and thus its validity can be investigated empirically by studying the range of a suitable estimate of R. Moreover, Imbens and Newey (2009) show that if this assumption fails many structural quantities of interest remain partially identified. Taken together, these assumptions imply that the ASF is identified:

Proposition 1. *Suppose that Assumptions 1 and 2 hold. Then*

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

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

3. A Discontinuity Test for Identification

3.1. Testing Problem. Our interest in this paper is in testing the validity of the assumptions that are necessary to justify a control variable approach as described in Section 2.2. Specifically, the pair of hypotheses that we would like to test is given by

$$\mathbb{H}_0\text{: Assumption 1 holds} \quad \textit{vs.} \quad \mathbb{H}_1\text{: Assumption 1 is violated;} \tag{3.1}$$

under the maintained assumption that conditions (2.1)–(2.4) hold, and that

$$\mathcal{S}(R|X = x) = [0, 1] \text{ for all } x \in \mathcal{S}(X|0 < X < \delta) \tag{3.2}$$

for some $\delta > 0$. While still restrictive, condition (3.2) is weaker than Assumption 2, and thus maintaining it seems reasonable in our setting. 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. The continuity condition (2.4) can usually be well-justified through subject knowledge. This leaves Assumption 1 as the remaining restriction of the model structure whose credibility could be uncertain in an empirical application.

3.2. A Testable Implication. To motivate our approach, suppose for a moment that the latent rank variable $R^* = F_{X^*|Z}(X^*; Z)$ was observable, and define

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

Note that this definition involves observable quantities only as R^* is assumed to be observable for the moment, and is thus meaningful under both the null hypothesis and the alternative. Using the structure of the model, it is then easily seen that

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

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

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

in this case. Since $R^* = F_V(V)$, it then follows from the continuity condition (2.4) that under the null hypothesis that Assumption 1 holds, the function $x \mapsto \mu(x, R^*)$ must be right-continuous at $x = 0$ with probability 1. On the other hand, under the alternative that Assumption 1 is violated, there is no reason to expect that $U \perp X | R^*$, and thus $x \mapsto \mu(x, R^*)$ should generally be discontinuous at $x = 0$ with positive probability.[3]

In our model, we only observe R but not R^*. While these two terms coincide in the subpopu-

[3] As we discuss below, there might still be certain alternatives under which no discontinuities appear, but we argue that those mostly correspond to pathological cases one is unlike to encounter in practice.

lation with $X > 0$, and thus $\mu(x, r) = \mathbb{E}(Y|X = x, R = r)$ 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 $\lim_{x\downarrow 0} \mu(x, r)$, $\mu(0, r)$ is not point identified, and thus we cannot check directly for the presence of a discontinuity. To address this problem, we consider the quantity

$$\Delta = \int \lim_{x\downarrow 0} \mathbb{E}(Y|X = x, R = r)d\Gamma(r) - \mathbb{E}(Y|X = 0), \tag{3.3}$$

where

$$\Gamma(r) = \frac{F_{U[0,1]}(r) - P(R \leq r|X > 0)P(X > 0)}{P(X = 0)}, \tag{3.4}$$

and $F_{U[0,1]}$ denotes the CDF of a uniform distribution on the unit interval. The support condition (3.2) ensures that the integral in (3.3) is well defined, but it is stronger than necessary for that purpose. The following theorem shows that under the null hypothesis Δ must be equal to zero.

Theorem 1 (Main Implication). *The term Δ defined in* (3.3) *is a well-defined functional of the distribution of (Y, X, Z); and a necessary condition for $\mathbb{H}_0$ to be true is that $\Delta = 0$.*

Note that $\Delta = 0$ is not sufficient for the null hypothesis. This means that finding that $\Delta \neq 0$ is evidence of a violation of $\mathbb{H}_0$, but finding that $\Delta = 0$ is compatible with both the null and some elements of the alternative. We elaborate more on this point below. To see why the theorem is true, first note that all terms on the right-hand side of equation (3.3) can be written as a known transformation of the joint distribution of (Y, X, Z), and thus Δ is well-defined under both the null hypothesis and the alternative. Next, note that under $\mathbb{H}_0$ the function $\Gamma(r) = F_{R^*|X}(r; 0)$, the conditional CDF of R^* given $X = 0$. This is because $R^* \sim U[0, 1]$ by construction, $R^* = R$ in the subpopulation with $X > 0$, and

$$F_{R^*}(r) = P(R^* \leq r|X > 0)P(X > 0) + P(R^* \leq r|X = 0)P(X = 0)$$

by the law of total probability. It follows that under $\mathbb{H}_0$ we have that

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

is equal to the size of the discontinuity of the function $x \mapsto \int \mathbb{E}(Y|X = x, R^* = r)dF_{R^*|X}(r; 0)$ at

$x = 0$; which is exactly equal to zero under $\mathbb{H}_0$. Our approach to use weighting with respect to the conditional distribution of R^* given $X = 0$ circumvents the need to identify $\mu(0, r)$. Note that we identify this conditional distribution even though realizations of R^* are never observed if $X = 0$.

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 Monte Carlo experiments below. However, Theorem 1 also implies that there might be alternatives that our approach is unable to detect. This fact should not be understood as a fundamental flaw of our method. Instead, it should be seen as an indication of the difficulty to derive testable implications from a condition like Assumption 1 in such a general setting. To provide some further insights into this issue, let us consider the function

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

With this notation, we can categorize the settings in which $\Delta = 0$ even though Assumption 1 is violated into two cases:

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

(ii) $\bar{\Delta}(r) = 0$ for all r even though $\mathbb{H}_1$ is true.

In case (i), the joint distribution of (Y, X, R^*) is such that the function $r \mapsto \bar{\Delta}(r)$ takes on both positive and negative values in such a way that it incidentally integrates to zero with respect to $F_{R^*|X}(r; 0)$. 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 below shows that it is conceivable in principle. We think of this case as a pathological one, and would argue that it is unlikely to be encountered in empirical applications.

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 $\bar{\Delta}(r) \equiv 0$, 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 $\mathbb{E}(m_1(x, U))$ for $x < \delta$, but not for $x \geq \delta$.

4. A Test Statistic and its Theoretical Properties

4.1. Estimation of Δ. The main idea behind the construction of our test statistic is to take a sample analogue $\widehat{\Delta}$ of Δ, and to reject the null hypothesis if this difference is "too large" in some appropriate sense. We propose to use

$$\widehat{\Delta} = \int \widehat{\mu}^{+}(r) d\widehat{\Gamma}(r) - \widehat{\mu}, \tag{4.1}$$

with $\widehat{\mu}^{+}(v)$ and $\widehat{\mu}$ suitable estimates of $\lim_{x \downarrow 0} \mathbb{E}(Y|X = x, R = v)$ and $\mathbb{E}(Y|X = 0)$, respectively, and $\widehat{\Gamma}(r)$ a suitable estimate of the function $\Gamma(r)$. We assume that the data are an i.i.d. sample $\{(Y_i, X_i, Z_i)\}_{i=1}^{n}$ of size n from the distribution of (Y, X, Z).

In order to compute the components that make up the definition of $\widehat{\Delta}$, we first estimate the conditional distribution function $F_{X|Z}$ of X given Z by local linear estimation (Fan and Gijbels, 1996):

$$\widehat{F}_{X|Z}(x, z) = e_{1,d_z}^{\top} \operatorname*{argmin}_{(a_1, a_2^{\top})} \sum_{i=1}^{n} \Big(\mathbb{I}\{X_i \leq x\} - a_1 - a_2^{\top}(Z_i - z) \Big)^2 K_g(Z_i - z). \tag{4.2}$$

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 $R = F_{X|Z}(X; Z)$ as

$$\widehat{R}_i = \widehat{F}_{X|Z}(X_i, Z_i) \text{ for } i = 1, \ldots n. \tag{4.3}$$

Third, we estimate the function $\Gamma(r)$ defined in (3.4) by

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

where $F_{U[0,1]}$ is the CDF of the standard uniform distribution. Fourth, we define the estimate $\widehat{\mu}^+(r)$ of the function $\lim_{x\downarrow 0}\mathbb{E}(Y|X=x,R=r)$ as

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

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 $\widehat{\mu}$ of $\mathbb{E}(Y|X=0)$ as a sample average of the observed outcomes Y_i among those observations with $X_i = 0$:

$$\widehat{\mu} = \frac{\sum_{i=1}^n Y_i \mathbb{I}\{X_i = 0\}}{\sum_{i=1}^n \mathbb{I}\{X_i = 0\}}.$$

The statistic $\widehat{\Delta}$ is then constructed as described in (4.1). Note that because of the particular structure of the estimate $\widehat{\Gamma}$, the expression given there simplifies to

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

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

4.2. Asymptotic Properties of $\widehat{\Delta}$. 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 et al. (2012, 2015) on nonparametric regression with generated covariates. Making use of these results requires the following assumption.

Assumption 3. *We assume that the data are an i.i.d. sample $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ of size n from the distribution of (Y, X, Z), and that the following conditions hold.*

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

(ii) The conditional CDF $F_{X|Z}(x, z)$ of X given Z is twice continuously differentiable with respect to its second argument on S_Z.

(iii) The random vector (X, R) is continuously distributed conditional on $X > 0$ with support $S_{XR|X>0} = \mathcal{S}(X, R|X > 0)$. The corresponding conditional density function $f_{XR|X>0}(\cdot)$ is continuously differentiable, bounded, and bounded away from zero on the compact set $S_\delta = \{(x, v) : (x, v) \in S_{XR|X>0}$ and $x \leq \delta\}$ with $\delta > 0$ as in (3.2).

(iv) The conditional expectation function $\mathbb{E}(Y|X = x, R = r)$ is twice continuously differentiable in x on S_δ.

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

(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) $h^2(ng^{d_z}/\log(n) + g^{-4}) \to \infty$.

Assumption 3 collects conditions that are largely common in the literature on nonparametric regression. Parts (i) and (iii) ensure that the estimates $\widehat{F}_{X|ZW}(x, z)$ and $\widehat{\mu}^+(r)$ 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, R) in the subpopulation with $X > 0$ in part (v) is necessary to apply certain results from Mammen et al. (2012, 2015) 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 et al. (2012, 2015). We derive the limiting distribution of $\widehat{\Delta}$ under Assumption 3. To state the result, we define

$$f^+_{R|X}(r,0) = \lim_{x\downarrow 0} f_{R|X}(r,x) \text{ and } f^+_X(0) = \lim_{x\downarrow 0} f_X(x),$$

let $\gamma(r) = \partial\Gamma(r)/\partial r$, and put

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

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

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

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, 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, the following result shows that distribution of $\widehat{\Delta}$ is asymptotically normal, with the rate of convergence being the same as that of a standard one-dimensional kernel smoother.

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

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

4.3. Test Statistic and Critical Values. Given the result in Theorem 2, a natural test statistic for the testing problem in (3.1) is given by a simple t-statistic of the form

$$T_n = \frac{\sqrt{nh} \cdot \widehat{\Delta}}{\widehat{\rho}},$$

where $\widehat{\rho}^2$ is some consistent estimate of the asymptotic variance $\rho^2 = C \cdot \sigma^2_+/f^+_X(0)$ of $\widehat{\Delta}$. We discuss several possibilities for obtaining such a variance estimate below. Under the conditions of Theorem 2, this test statistic follows a standard normal distribution under the null hypothesis in large samples, and diverges under any fixed alternative which is such that $\Delta \neq 0$. The testing decision is thus to reject $\mathbb{H}_0$ at the nominal level $\alpha \in (0,1)$ if $|T_n| > c_\alpha$, where $c_\alpha = \Phi^{-1}(1-\alpha/2)$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution. The following result formally shows the validity of such an approach.

Theorem 3. *If Assumption 3 holds, and $\widehat{\rho}^2 \xrightarrow{p} \rho^2$, we have the following results:*

(i) Under the null hypothesis, i.e. if $\Delta = 0$, $\lim_{n\to\infty} P(|T_n| > c_\alpha) = \alpha$.

(ii) Under any fixed alternative that implies $\Delta \neq 0$, $\lim_{n\to\infty} P(|T_n| > c_\alpha) = 1$.

(iii) Under any local alternative that implies $\Delta = \delta/\sqrt{nh}$, $\lim_{n\to\infty} P(|T_n| > c_\alpha) = 1 - \Xi_\theta(c_\alpha)$, where Ξ_θ is the CDF of the absolute value of a normal random variable with mean $\theta = \delta/\rho$ and variance 1.

Since Theorem 3 holds for any consistent estimator of ρ^2, the only remaining issue is to find one such 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 ρ^2. However, such a procedure, which we explicitly describe in Section A.4, can be unattractive in practice because it requires choosing several additional smoothing parameters. In some preliminary simulations that we conducted, the performance of the resulting test was indeed quite sensitive in this regard. We therefore recommend the use of a nonparametric bootstrap variance estimator. Such a procedure is computationally more expensive, but straightforward from a practical point of view. The estimator is obtained as follows. Let $\{(Y_i^*, X_i^*, Z_i^*)\}_{i=1}^n$ be a bootstrap sample drawn with replacement from the observed data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, and let $\widehat{\Delta}^*$ be an estimate of Δ computed exactly as described above but using the bootstrap sample. Then $\widehat{\rho}^2 = \mathbb{E}^*((\widehat{\Delta}^* - \widehat{\Delta})^2)$, where $\mathbb{E}^*$ denotes the expectation with respect to bootstrap sampling. While it is difficult to give interpretable low-level conditions for the consistency of such an estimator, bootstrap variance estimation is widely used in practice due to its generally good behavior in simulations.

4.4. Covariates. In many empirical applications, researchers might want to incorporate additional exogenous covariates into their model. In principle, it would be straightforward to introduce such covariates, say W, into our setting in a fully nonparametric fashion, and to consider a model of the form

$$\begin{aligned} Y &= m_1(X, W, U) \\ X &= \max\{0, m_2(Z, W, V)\}. \end{aligned}$$

Under an appropriate exogeneity condition of W, our identification analysis from above can simply be applied to this model conditional on the value of W. In most realistic applications, which typically use several covariates, a fully nonparametric *implementation* of our test is unlikely to have good finite sample properties due to the curse of dimensionality. For this reason, we expect that applied researchers will generally want to impose some semiparametric restrictions when working with additional covariates. There is of course a vast number of types of restrictions that can be used reduce the effective dimensionality of the problem, and their appropriateness depends critically on the details of the respective application. As it is therefore impossible to present an exhaustive discussion, we focus here on the following "partially linear" specification.

$$Y = m_1(X, U) + W'\theta_1, \tag{4.4}$$

$$X = \max\{0, m_2(Z, V) + W'\theta_2\} \tag{4.5}$$

This specification has the advantage that is seems appropriate for a wide range of applications, and that within this framework including covariates into our existing testing procedure is rather simple from a technical point of view.

To describe how this can be done, suppose that Assumption 1(i) is replaced with the condition that (Z, W) and (U, V) are stochastically independent, define $\tilde{X}_i = X_i - W'\theta_2$ and $\tilde{Y}_i = Y_i - W'\theta_1$, let $F_{\tilde{X}|Z}$ denote the CDF of $\tilde{X}$ given Z, and redefine $R_i = F_{\tilde{X}|Z}(\tilde{X}_i, Z_i)$. $F_{\tilde{X}|Z}(t|z) = F_V(m_2^{-1}(z,t))[1 - F_{W'\theta_2|Z}(-t|z)]$ is strictly monotone in t, where $m_2^{-1}(z,t)$ denotes the inverse of m_2 with respect to its second argument. Moreover, $\tilde{X} = m_2(Z, V)$if $X > 0$, and thus distribution of U conditional on $X = x$ and $R = r$ is the same as the the distribution of U conditional on $R = r$ when $x > 0$,.

In our specification the conditional expectation $\mathbb{E}(X|Z,W) = \mathbb{E}(X|Z,W'\theta_2)$ has an index structure, and thus a $\sqrt{n}$-consistent estimator $\widehat{\theta}_2$ of θ_2 can be obtained under standard regularity conditions by for example using the method of Ichimura and Lee (1991). We can therefore define an estimate $\widehat{X}_i = X_i - W_i'\widehat{\theta}_2$ of $\tilde{X}_i$, and an estimate $\widehat{R}_i = \widehat{F}_{\widehat{X}|Z}(\widehat{X}_i, Z_i)$ of R_i, where $\widehat{F}_{\widehat{X}|Z}$ is computed as described in (4.2), but with $\widehat{X}_i$ replacing X_i.

It is also easy to see that $\mathbb{E}(Y|X=x, W=w, R=r, X>0) = \mathbb{E}(m_1(x,U)|R=r, X>0) + w'\theta_1$ is the sum of a smooth function in (x,r) and a linear function in w. Hence a $\sqrt{n}$-consistent estimator $\widehat{\theta}_1$ of θ_1 can obtained under standard regularity conditions by for example using the method of Robinson (1988) in the subsample with $X>0$; that is, by running a linear regression of $Y - \widehat{\mathbb{E}}(Y|X,\widehat{R})$ on $W - \widehat{\mathbb{E}}(W|X,\widehat{R})$ in the subsample with $X>0$, where $\widehat{\mathbb{E}}(\cdot)$ denotes a nonparametrically estimated conditional expectation function. We can then define an estimate $\widehat{Y}_i = Y_i - W_i'\widehat{\theta}_1$ of $\tilde{Y}_i$, and compute our test statistic as before, but with $\widehat{Y}_i$ replacing Y_i. Since $(\widehat{\theta}_1, \widehat{\theta}_2)$ converges with a parametric rate, the first order asymptotic properties of this test statistic will be the same as that of an infeasible one which uses the actual realizations of $\tilde{X}_i$ and $\tilde{Y}_i$ instead of X_i and Y_i.

5. Some Monte Carlo Evidence

To gain further insights into the properties of our testing procedure, we conduct as a series of Monte Carlo experiments. In these simulations, the distribution of the observable quantities (Y,X,Z) is determined by the following class of data generating processes (DGPs):

$$Y = U_1 + U_2 X,$$
$$X = \max\{0, 1 + V_1 + A \cdot Z\}.$$

Here we define $U_j = (V_1 + \beta V_2^2/\sqrt{2} + \gamma Z + \eta_j)/\sqrt{1+\beta^2+\gamma^2}$ for $j = 1,2$, and $A = 2 \cdot (1 + \alpha V_1 + \beta V_2^2/\sqrt{2})/\sqrt{1+\alpha^2+\beta^2}$ with $(Z, V_1, V_2, \eta_1, \eta_2)$ a vector of independent, standard normally distributed random variables. The coefficients α, β and γ are real-valued and are varied for our Monte Carlo experiments. 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. For our Monte Carlo study, we specifically consider 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 $\alpha = \beta = 0$.

We then apply our test as described in Section 4.3 to samples of size $n = 500$ from each of these DGPs. The number of replications is set to 10,000. We consider a bootstrap-based estimator of the asymptotic variance using $B = 500$ bootstrap samples. The bandwidths are set to $h = .5 \cdot SD \cdot n^{-1/4}$ and $g = 1.5 \cdot SD \cdot n^{-1/4}$, where SD is the empirical standard deviation of the respective covariate.[4]

Table 1: Simulation Results.

Parameter	DGP1 (varying α)		DGP2 (varying β)		DGP3 (varying γ)	
($\alpha/\beta/\gamma$)	Δ	P(rej. $\mathbb{H}_0$)	Δ	P(rej. $\mathbb{H}_0$)	Δ	P(rej. $\mathbb{H}_0$)
0.0	0.000	0.045	0.000	0.045	0.000	0.045
0.2	0.005	0.054	-0.056	0.052	0.128	0.082
0.4	0.044	0.067	-0.128	0.103	0.250	0.202
0.6	-0.021	0.083	-0.191	0.179	0.359	0.398
0.8	-0.195	0.180	-0.259	0.274	0.437	0.632
1.0	-0.420	0.341	-0.301	0.399	0.494	0.803
1.2	-0.735	0.710	-0.345	0.470	0.540	0.920
1.4	-0.507	0.649	-0.371	0.556	0.573	0.972
1.6	-0.411	0.422	-0.396	0.622	0.601	0.992
1.8	-0.382	0.269	-0.416	0.669	0.622	0.994
2.0	-0.367	0.209	-0.442	0.704	0.634	1.000

In Table 1, we report the empirical rejection probabilities of our test under the various DGPs at the usual nominal significance level of 5%. We also numerically compute the population value of Δ for each DGP to gain some additional insights into the behavior of our procedure. The table shows that when the null hypothesis is true (that is, if $\alpha = \beta = \gamma = 0$), the rejection probability of

[4] We also experimented with various multiples of these bandwidths and found the results to be reasonable robust with respect to that choice.

our test is close to the nominal level. It also shows that our test has power against all three types of violations of the null hypothesis.

Table 1 also illustrates one of the consequences of the fact that that we are only testing a necessary condition for the null hypothesis to be true. Note that within each of the three types of DGPs the value of the varying parameter can be interpreted as a measure of distance from the null hypothesis. However, this does not mean that the value of Δ has to be monotone in the value of that parameter. While this turns out to be the case for DGP2 and DGP3, Table 1 shows that under DGP1 the value of Δ varies non-monotonically with α. Therefore some alternatives that are "far" away from the null hypothesis are more difficult to reject for our test than some others that are "closer".

6. Conclusions

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 is the the first to 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. 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. Appendix

A.1. Proof of Proposition 1. The result is a minor extension of Theorem 1 in Imbens and Newey (2009). Let $R^* = F_{X^*|Z}(X^*; Z)$, and note that it follows from Imbens and Newey (2009) that under Assumption 1 we have that $R^* = F_V(V)$, and that $U \perp X|R^*$. Since $R^* \sim U[0,1]$ by construction, we also have that $a(x) = \int_0^1 \mathbb{E}(m_1(x,U)|R^* = r)dr$. 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 $a(0) = \lim_{x\downarrow 0} a(x)$. $\square$

A.2. Proof of Theorem 1. Equations (3.3) and (3.4) show that the definition of Δ only involves observable quantities. Now let $\bar{\Delta}(r) = \lim_{x \downarrow 0} \mathbb{E}(Y|X = x, R^* = r) - \mathbb{E}(Y|X = 0, R^* = r)$. Then it follows from the above arguments that $P(\bar{\Delta}(R^*) = 0) = 1$ under $\mathbb{H}_0$ and $P(\bar{\Delta}(R^*) = 0) \leq 1$ under $\mathbb{H}_1$. Since $\Delta = \int \bar{\Delta}(r) dF_{R^*|X}(r; 0)$ this implies the statement of the theorem. □

A.3. Proof of Theorem 2. The result in Theorem 2 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 $\widehat{\Gamma}(\cdot)$.

Lemma 1. *Suppose that the conditions of Theorem 2 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 define an infeasible estimator of $\mu^+(r) = \lim_{x \downarrow 0} \mathbb{E}(Y|X = x, R = r)$ 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:

$$L_i(r) = (1, X_j/h, (R_i - v)/h)^\top \cdot \mathbb{I}\{X_i > 0\},$$

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

$$N_n(r) = \mathbb{E}(L_i(r)L_i(r)^\top K_h(X_i, R_i - r)).$$

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

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

uniformly in r, where the matrix A is given by

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

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

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

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 2 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 et al. (2012, 2015) to show this statement. For convenience, we repeat the following notation, which was already introduced in the proof of Lemma 2:

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

It then follows from an application of Theorem 1 in Mammen et al. (2015) 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 et al. (2015) is equal to zero. Next, it follows from the same arguments as in the proof of Theorem 4 in Mammen et al. (2015) 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.4. An Alternative Estimate of the Asymptotic Variance. In this section, we describe a plug-in estimator of ρ^2 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 ρ^2, 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

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

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

$$\widehat{\gamma}(r) = (1 - \widehat{g}(r))/(\sum_{i=1}^{n} \mathbb{I}\{X_i = 0\}/n),$$

where

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

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 $\widehat{g}(r)$, we achieve that the estimator $\widehat{\gamma}$ is uniformly consistent under weak regularity conditions. We also define

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

$$\widehat{f}^{+}_{X}(0) = \frac{2}{n} \sum_{i=1}^{n} K_{b_1}(X_i)\mathbb{I}\{X_i > 0\},$$

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 σ^2_+ is given by

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

where

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

and

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

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

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

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 $\widehat{\rho}^2 \xrightarrow{p} \rho^2$.*

Proof. Let

$$g(r) = \partial P(R \leq r, X > 0)/\partial v$$

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

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

be an infeasible estimator of $g(r)$ that uses the true $R_i = F_{X|Z}(X, Z)$ instead of the estimates $\widehat{R}_i = \widehat{F}_{X|Z}(X, Z)$. From a simple Taylor expansion, it follows that

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

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

$$\sup_v |\widetilde{g}(v) - g(v)| = o_p(1).$$

Similar arguments can be used to show that $\widehat{f}^+_{R|X}(r; 0)$ and $\widehat{f}^+_X(0)$ are uniformly consistent estimates of $\lim_{x\downarrow 0} f_{R|X}(r; x)$ and $\lim_{x\downarrow 0} f_X(x)$, respectively. Consistency of $\widehat{\sigma}^2_+$ for σ^2_+ then follows from the linearity of the local linear smoothing operator. □

References

Akritas, M. G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scandinavian Journal of Statistics*, 28(3):549–567.

Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. *Quarterly Journal of Economics*, 120(3):1031–1083.

Blau, F. and Grossberg, A. (1992). Maternal labor supply and children's cognitive development. *Review of Economics & Statistics*, 74(3):474–481.

Blundell, R., Chen, X., and Kristensen, D. (2007). Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669.

Blundell, R. and MaCurdy, T. (1999). Labor supply: A review of alternative approaches. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 3 of *Handbook of Labor Economics*, chapter 27, pages 1559–1695. Elsevier.

Blundell, R. and Matzkin, R. (2010). Conditions for the existence of control functions in nonseparable simultaneous equations models. *Working Paper*.

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

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

Caetano, C. (2015). A test of exogeneity without instrumental variables in models with bunching. *Econometrica*, 83(4):1581–1600.

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

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In Christofides, L. N., Grant, E. K., and Swidinsky, R., editors, *Aspects of labour market behaviour: essays in honour of John Vanderkamp*, pages 201–222. University of Toronto Press, Toronto, Canada.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics*, 3:1801–1863.

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

Chernozhukov, V., Imbens, G. W., and Newey, W. K. (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.

Connelly, R. and Kimmel, J. (2009). Spousal influences on parents' non-market time choices. *Review of Economics of the Household*, 7(4):361–394.

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

D'Haultfoeuille, X. and Février, P. (2011). Identification of nonseparable models with endogeneity and discrete instruments. *Working Paper*.

Evans, W. N. and Ringel, J. S. (1999). Can higher cigarette taxes improve birth outcomes? *Journal of Public Economics*, 72(1):135–154.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall.

Fertig, A. R. (2010). Selection and the effect of prenatal smoking. *Health Economics*, 19(2):209–226.

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

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

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

Heckman, J. J. and MaCurdy, T. E. (1980). A life cycle model of female labour supply. *The Review of Economic Studies*, 47(1):47–74.

Imbens, G. and Newey, W. (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 Blundell, R., Newey, W., and Persson, T., editors, *Advances in Economics and Econometrics*. Cambridge University Press.

Iwata, S. and Tamada, K. (2014). The backward-bending commute times of married women with household responsibility. *Transportation*, 41(2):251–278.

Kasy, M. (2011). Identification in triangular systems using control functions. *Econometric Theory*, 27:663–671.

Kramer, M. S. (1987). Intrauterine growth and gestational duration determinants. *Pediatrics*, 80(4):502–511.

Kramer, M. S. (1998). Socioeconomic determinants of intrauterine growth retardation. *European Journal of Clinical Nutrition*, 52(1):29–32.

Lien, D. S. and Evans, W. N. (2005). Estimating the impact of large cigarette tax hikes the case of maternal smoking and infant birth weight. *Journal of Human Resources*, 40(2):373–392.

Lydon, R. and Chevalier, A. (2002). Estimates of the effect of wages on job satisfaction. Technical Report 20081, London School of Economics and Political Science, LSE Library.

Mammen, E., Rothe, C., and Schienle, M. (2012). Nonparametric Regression with Nonparametrically Generated Covariates. *Annals of Statistics*, 40(2):1132–1170.

Mammen, E., Rothe, C., and Schienle, M. (2015). Semiparametric estimation with generated covariates. *Econometric Theory*.

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

Newey, W. K. and Powell, J. L. (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.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415.

Stewart, M. B. and Swaffield, J. K. (1997). Constraints on the desired hours of work of british men. *The Economic Journal*, 107(441):520–535.

Tominey, E. (2007). Maternal Smoking During Pregnancy and Early Child Outcomes. CEP Discussion Papers dp0828, Centre for Economic Performance, LSE.

Torgovitsky, A. (2015). Identification of nonseparable models using instruments with small support. *Econometrica*, 83(3):1185–1197.

Vogler, G. P. and Kozlowski, L. T. (2002). Differential influence of maternal smoking on infant birth weight: gene-environment interaction and targeted intervention. *JAMA*, 287(2):241–242.

Wehby, G. L., Fletcher, J. M., Lehrer, S. F., Moreno, L. M., Murray, J. C., Wilcox, A., and Lie, R. T. (2011). A genetic instrumental variables analysis of the effects of prenatal smoking on birth weight: evidence from two samples. *Biodemography and Social Biology*, 57(1):3–32.