Quantile Models with Endogeneity

V. Chernozhukov[1] and C. Hansen[2]

[1]Department of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142

[2]The University of Chicago Booth School of Business, Chicago, Illinois 60637; email: christian.hansen@chicagobooth.edu

Annu. Rev. Econ. 2013. 5:57–81

First published online as a Review in Advance on May 20, 2013

The *Annual Review of Economics* is online at economics.annualreviews.org

This article's doi: 10.1146/annurev-economics-080511-110952

JEL codes: C21, C26

Keywords

identification, treatment effects, structural models, instrumental variables

Abstract

In this article, we review quantile models with endogeneity. We focus on models that achieve identification through the use of instrumental variables and discuss conditions under which partial and point identification are obtained. We discuss key conditions, which include monotonicity and full-rank-type conditions, in detail. In providing this review, we update the identification results of Chernozhukov & Hansen (2005). We illustrate the modeling assumptions through economically motivated examples. We also briefly review the literature on estimation and inference.

1. INTRODUCTION

Quantile regression is a tool for estimating conditional quantile models that has been used in many empirical studies and has been studied extensively in theoretical econometrics (see Koenker & Bassett 1978, Koenker 2005). One of quantile regression's most appealing features is its ability to estimate quantile-specific effects that describe the impact of covariates not only on the center, but also on the tails of the conditional outcome distribution. Although the central effects, such as the mean effect obtained through conditional mean regression, provide interesting summary statistics of the impact of a covariate, they fail to describe the full distributional impact unless the conditioning variables affect the central and the tail quantiles in the same way. In addition, researchers are interested in the impact of covariates on points other than the center of the conditional distribution in many cases. For example, in a study of the effectiveness of a job-training program, the effect of training on the lower tail of the earnings distribution conditional on worker characteristics may be of more interest than the effect of training on the mean of the distribution.

In observational studies, the variables of interest (e.g., education or prices) are often endogenous. Just as with the conventional linear model, the endogeneity of covariates renders the conventional quantile regression inconsistent for estimating the causal (structural) effects of covariates on the quantiles of economic outcomes. One approach to addressing this problem is to generalize the instrumental variables (IV) framework to allow for the estimation of quantile models. In this article, we review developments in IV approaches to modeling and estimating quantile treatment (structural) effects (QTE) in the presence of endogeneity.

We focus our review on the modeling framework of Chernozhukov & Hansen (2005), which provides conditions for identification of the QTE without functional form assumptions. The principal identifying assumption of the model is the imposition of conditions that restrict how rank variables (structural errors) may vary across treatment states. These conditions allow the use of IV to overcome the endogeneity problem and recover the true QTE. This framework also ties naturally to simultaneous equations models, corresponding to a structural simultaneous equations model with nonadditive errors. Within this framework, estimation and inference procedures for linear quantile models have been developed by Chernozhukov & Hansen (2006, 2008), Chernozhukov et al. (2009), and Jun (2008); nonparametric estimation has been considered by Chernozhukov et al. (2007), Horowitz & Lee (2007), Chen & Pouzo (2009, 2012), and Gagliardini & Scaillet (2012); and inference with discrete outcomes has been explored by Chesher (2005). Moreover, the modeling framework provides a foundation for other estimation methods based on IV median-independence and more general quantile-independence conditions, as in Abadie (1997), Chen et al. (2003), Chernozhukov & Hong (2003), Hong & Tamer (2003), Honore & Hu (2004), and Sakata (2007). It is also important to note that the modeling framework we review can be used to study the nonparametric identification of structural economic models in cases in which quantile effects are not necessarily the chief objects of interest. Berry & Haile (2010) provide an excellent example of this in the context of discrete choice models with endogeneity.

We also briefly review other modeling approaches for quantile effects with endogenous covariates. Abadie et al. (2002) consider a QTE model for the subpopulation of compliers that applies to binary endogenous variables with binary instruments. Imbens & Newey (2009), Chesher (2003), Lee (2007), and Koenker & Ma (2006) use models with triangular structures and show how control functions can be constructed and employed to estimate structural objects of interest. Although these models share some features with the model of Chernozhukov & Hansen (2005), the three approaches are nonnested in general.

Quantile models with endogeneity have been used in many empirical studies in economics (e.g., see Abadie et al. 2002, Chernozhukov & Hansen 2004, Hausman & Sidak 2004, Forbes 2008,

Eren 2009, Kostov 2009, Maynard & Qiu 2009, Wehby et al. 2009, Lamarche 2011, Autor et al. 2012, Somainiy 2012). We do not present a review of empirical applications but note that these papers provide further discussion of how the IV quantile model relates to their specific framework and illustrate some of the rich effects that one can estimate using quantile methods.

2. AN INSTRUMENTAL VARIABLE QUANTILE MODEL

In this section, we present an IV model for QTE, its main econometric implication, and the principal identification result.

2.1. Framework

Our model is developed within the conventional potential (latent) outcome framework (e.g., Heckman & Robb 1986). Potential real-valued outcomes that vary among individuals or observational units are indexed against potential treatment states $d \in \mathcal{D}$ and denoted Y_d. The potential outcomes $\{Y_d\}$ are latent because, given the selected treatment D, the observed outcome for each individual or observational unit is only one component

$$Y := Y_D$$

of the potential outcomes vector $\{Y_d\}$. Throughout the article, capital letters denote random variables, and lowercase letters denote the potential values they may take. We do not explicitly state various technical measurability assumptions as these can be deduced from the context.[1]

The objective of causal or structural analysis is to learn about features of the distributions of potential outcomes Y_d. Of primary interest to us are the τ-th quantiles of potential outcomes under various treatments d, conditional on observed characteristics $X = x$, denoted as

$$q(d, x, \tau).$$

We refer to the function $q(d, x, \tau)$ as the quantile treatment response (QTR) function. We are also interested in the QTE, defined as

$$q(d_1, x, \tau) - q(d_0, x, \tau),$$

that summarize the differences in the impact of treatments on the quantiles of potential outcomes (Doksum 1974, Lehmann 1974).

Typically, the realized treatment D is selected in relation to potential outcomes, inducing endogeneity. This endogeneity makes the conventional quantile regression of observed Y on observed D, which relies on the restriction

$$P[Y \leqslant \theta(D, X, \tau)|X, D] = \tau \text{ a.s.},$$

inappropriate for measuring $q(d, x, \tau)$ and the QTE. Indeed, the function $\theta(d, x, \tau)$ solving these equations will not be equal to $q(d, x, \tau)$ under endogeneity. The model presented next states conditions under which we can identify and estimate the quantiles of latent outcomes through the

[1]For simplicity, we could assume that d takes on a countable set of values $\mathcal{D}$ or make separability assumptions that imply that the stochastic process $\{Y_d, d \in \mathcal{D}\}$ is defined from its definition over a countable subset $\mathcal{D}_0 \subset \mathcal{D}$ (see van der Vaart & Wellner 1996).

use of instruments Z that affect D but are independent of potential outcomes and the nonlinear quantile-type conditional moment restrictions

$$P[Y \leqslant q(D, X, \tau)|X, Z] = \tau \text{ a.s.}$$

2.2. The Instrumental Variable Quantile Treatment Effects Model

Having conditioned on the observed characteristics $X = x$, each latent outcome Y_d can be related to its quantile function $q(d, x, \tau)$ as[2]

$$Y_d = q(d, x, U_d), \text{ where } U_d \sim U(0, 1) \tag{1}$$

is the structural error term. We note that the representation in Equation 1 is essential to what follows.

The structural error U_d is responsible for the heterogeneity of potential outcomes among individuals with the same observed characteristics x. This error term determines the relative ranking of observationally equivalent individuals in the distribution of potential outcomes given the individuals' observed characteristics, and thus we refer to U_d as the rank variable. As U_d drives differences in observationally equivalent individuals, one may think of U_d as representing some unobserved characteristic (e.g., ability or proneness).[3] This interpretation makes quantile analysis an interesting tool for describing and learning the structure of heterogeneous treatment effects and accounting for unobserved heterogeneity (see Doksum 1974, Heckman & Smith 1997, Koenker 2005).

For example, consider a returns-to-training model, where the Y_d's are potential earnings under different training levels d, and $q(d, x, \tau)$ is the conditional earnings function that describes how an individual having training d, characteristics x, and the latent ability τ is rewarded by the labor market. The earnings function may be different for different levels of τ, implying heterogeneous effects of training on earnings of people that have different levels of ability. For example, it may be that the largest returns to training accrue to those in the upper tail of the conditional distribution, that is, to the high-ability workers.[4]

Formally, the instrumental variable quantile treatment effects (IVQTE) model consists of five conditions (some are representations) that hold jointly. Consider a common probability space (Ω, F, P) and the set of potential outcome variables $(Y_d, d \in \mathcal{D})$, the covariate variables X, and the IV Z. The following conditions hold jointly with probability 1:

> **Condition 1 (Potential outcomes):** Conditional on X and for each d, $Y_d = q(d, X, U_d)$, where $\tau \mapsto q(d, X, \tau)$ is nondecreasing on $[0, 1]$ and left-continuous and $U_d \sim U(0, 1)$.
>
> **Condition 2 (Independence):** Conditional on X and for each d, U_d is independent of IV Z.

[2]This follows by the Fisher-Skorohod representation of random variables, which states that given a collection of variables $\{\zeta_d\}$, each variable ζ_d can be represented as $\zeta_d = q(d, U_d)$, for some $U_d \sim U(0, 1)$ (cf. Durrett 1996), where $q(d, \tau)$ denotes the τ-quantile of variable ζ_d.

[3]Doksum (1974) uses the term proneness as in "prone to learn fast" or "prone to grow taller."

[4]It is important to note that the quantile index, τ, in $q(d, x, \tau)$ refers to the quantile of potential outcome Y_d given that exogenous variables are set at $X = x$ and not to the unconditional quantile of Y_d. For example, suppose that one of the control variables in the earnings example is years of schooling. An individual at the 30th percentile of the distribution of Y_d given 20 years of schooling, for example, is not necessarily low income as even a relatively low earner with that level of education may still earn above the median earnings in the overall population.

Condition 3 (Selection): $D := \delta(Z, X, V)$ for some unknown function δ and random vector V.

Condition 4 (Rank similarity): Conditional on (X, Z, V), $\{U_d\}$ are identically distributed.

Condition 5 (Observables): The observed random vector consists of $Y := Y_D$, D, X, and Z.

The following is the main econometric implication of the model.

Theorem 1 (Main statistical implication): Suppose Conditions 1–5 hold. (*a*) Then we have for $U := U_D$, with probability 1,

$$Y = q(D, X, U), \ U \sim U(0,1)|X, Z. \tag{2}$$

(*b*) If Equation 2 holds and $\tau \mapsto q(d, \tau)$ is strictly increasing for each d, then for each $\tau \in (0, 1)$, a.s.

$$P[Y \leqslant q(D, X, \tau)|X, Z] = \tau. \tag{3}$$

(*c*) If Equation 2 holds, then for any closed subset I of $[0, 1]$, a.s.

$$P(U \in I) \leqslant P[Y \in q(D, X, I)|X, Z], \tag{4}$$

where $q(d, x, I)$ is the image of I under the mapping $\tau \mapsto q(d, X, \tau)$.

The first result states that the main consequence of Conditions 1–5 is a simultaneous equations model (Equation 2) with nonseparable error U that is independent of Z and X and is normalized so that $U \sim U(0, 1)$. The second result considers econometric implications when $\tau \mapsto q(D, X, \tau)$ is strictly increasing, which requires that Y is nonatomic conditional on X and Z. In this case, we obtain the conditional moment restriction in Equation 3. This implication follows from the first result and that

$$\{Y \leqslant q(D, X, \tau)\} \text{ is equivalent to } \{U \leqslant \tau\},$$

when $q(D, X, \tau)$ is strictly increasing in τ. The final result deals with the case in which Y may have atoms conditional on X and Z (e.g., when Y is a count or discrete response variable). The first two results were obtained in Chernozhukov & Hansen (2005), and the third result is in the spirit of results given in Chesher et al. (2011), Chesher (2005), and Chesher & Smolinski (2010). The latter results are related to random set/optimal transport methods for identification analysis (see Galichon & Henry 2009, 2011; Ekeland et al. 2010; Beresteanu et al. 2011).

The model and the results of Theorem 1 are useful for two reasons. First, Theorem 1 serves as a means of identifying the QTE in a reasonably general heterogeneous effects model. Second, by demonstrating that the IVQTE model leads to the conditional moment restrictions given in Equations 3 and 4, Theorem 1 provides an economic and causal foundation for estimation based on these restrictions.

2.3. The Identification Regions

The conditions presented above yield the following identification region for the structural quantile function $(d, x, \tau) \mapsto q(d, x, \tau)$. The identification region for the case of strictly increasing $\tau \mapsto q(d, x, \tau)$ can be stated as the set $\mathcal{Q}$ of functions $(d, x, \tau) \mapsto m(d, x, \tau)$ that satisfy the following relations; for all $\tau \in (0, 1]$,

$$P[Y < m(D, X, \tau)|X, Z] = \tau \text{ a.s.} \tag{5}$$

This representation of the identification region $\mathcal{Q}$ is implicit. Nevertheless, statistical inference about $q \in \mathcal{Q}$ can be based on Equation 5 and can be carried out in practice using weak-identification-robust inference as described in Chernozhukov & Hansen (2008), Jun (2008), Chernozhukov et al. (2009), Marmer & Sakata (2012), and Santos (2012). Under conditions that yield point identification, these regions collapse to a singleton, and the aforementioned weak-identification-robust inference procedures retain their validity.

The identification region for the case of weakly increasing $\tau \mapsto q(d, x, \tau)$ can be stated as the set $\mathcal{Q}$ of functions $(d, x, \tau) \mapsto m(d, x, u)$ that satisfy the following relations: For any closed subset I of (0,1],

$$P(U \in I) \leqslant P[Y \in m(D, X, I)|X, Z] \text{ a.s.}, \tag{6}$$

where $m(D, X, I)$ is the image of I under the mapping $\tau \mapsto m(D, X, \tau)$. The inference problem here falls in the class of conditional moment inequalities, and approaches such as those described in Andrews & Shi (2013) and Chernozhukov et al. (2013), for example, can be used. The sets I to be checked could be reduced by determining approximate core-determining subsets (for further discussion, see Galichon & Henry 2009, 2011; Chesher et al. 2011).

2.4. Discussion of the Model

Condition 1 imposes monotonicity on the structural function of interest, which makes its relation to the QTR apparent. Condition 2 states that potential outcomes are independent of Z, given X, which is a conventional independence restriction. Condition 3 is a convenient representation of a treatment selection mechanism, stated for the purposes of discussion. In Condition 3, the unobserved random vector V is responsible for the difference in treatment choices D across observationally identical individuals. Dependence between V and $\{U_d\}$ is the source of endogeneity that makes the conventional exogeneity assumption $U \sim U(0, 1) \mid X, D$ break down. This failure leads to the inconsistency of exogenous quantile methods for estimating the structural quantile function. Within the model outlined above, this breakdown is resolved through the use of IV.

The independence imposed in Conditions 2 and 3 is weaker than the commonly made assumption that both the disturbances $\{U_d\}$ in the outcome equation and the disturbances V in the selection equation are jointly independent of the instrument Z (e.g., Heckman & Robb 1986, Imbens & Angrist 1994). The latter assumption may be violated when the instrument is measured with error, as discussed in Hausman (1977), or the instrument is not assigned exogenously relative to the selection equation, as in Imbens & Angrist (1994, example 2).

Condition 4 restricts the variation in ranks across potential outcomes and is key for identifying the QTR and associated QTE. Its simplest, although strongest, form is rank invariance, when ranks U_d do not vary with potential treatment states d:[5]

[5]Notice that under rank invariance, Condition 3 is a pure representation, not a restriction, because nothing restricts the unobserved information component V.

$$U_d = U \text{ for each } d \in \mathcal{D}. \tag{7}$$

For example, under rank invariance, people who are strong (highly ranked) earners without a training program ($d = 0$) remain strong earners having done the training ($d = 1$). Indeed, the earnings of a person with characteristics x and rank $U = \tau$ in training state 0 are $Y_0 = q(0, x, \tau)$ and in state 1 are $Y_1 = q(1, x, \tau)$.[6] Thus rank invariance implies that a common unobserved factor U, such as innate ability, determines the ranking of a given person across treatment states.

Rank invariance implies that the potential outcomes $\{Y_d\}$ are jointly degenerate, which may be implausible on logical grounds, as pointed out by Heckman & Smith (1997). Moreover, the rank variables U_d may be determined by many unobserved factors. Thus it is desirable to allow the rank U_d to change across d, reflecting some unobserved, asystematic variation. Rank similarity (Condition 4) achieves this property while managing to preserve the useful moment restriction given in Equation 3.

Rank similarity (Condition 4) relaxes exact rank invariance by allowing asystematic deviations, "slippages" in the terminology of Heckman & Smith (1997), in one's rank away from some common level U. Conditional on U, which may enter disturbance V in the selection equation, we have the following condition on the slippages:[7]

$$U_d - U \text{ are identically distributed across } d \in \mathcal{D}. \tag{8}$$

In this formulation, we implicitly assume that one selects the treatment without knowing the exact potential outcomes; in other words, one may know U and even the distribution of slippages but does not know the exact slippages $U_d - U$. This assumption is consistent with many empirical situations in which the exact latent outcomes are not known before receipt of treatment. We also note that conditioning on appropriate covariates X may be important to achieve rank similarity.

In summary, rank similarity is an important restriction of the IVQTE model that allows us to address endogeneity. This restriction is absent in conventional endogenous heterogeneous treatment effect models. However, similarity enables a more general selection mechanism (Condition 3) and weaker independence conditions on instruments than often are assumed in nonseparable IV models. The main force of rank similarity and the other stated assumptions is the implied moment restriction given in Equation 3 of Theorem 1, which is useful for identification and estimation of the QTE.

2.5. Examples

We present some examples that highlight the nature of the model, its strengths, and its limitations.

> **Example 1 (Demand with nonseparable error):** The following is a generalization of the classic supply-demand example. Consider the model
>
> $$\begin{aligned} Y_p &= q(p, U), \\ \tilde{Y}_p &= \rho(p, Z, \mathcal{U}), \\ P &\in \{p : \rho(p, Z, U) = q(p, \mathcal{U})\}, \end{aligned} \tag{9}$$
>
> where functions q and ρ are increasing in the last argument. The function $p \mapsto Y_p$ is the random demand function, and $p \mapsto \tilde{Y}_p$ is the random supply function.

[6] Rank invariance is used in many interesting models without endogeneity (see, e.g., Doksum 1974, Heckman & Smith 1997, Koenker & Geling 2001).

[7] Conditioning is required to be on all components of V in the selection equation in Condition 3.

Additionally, functions q and ρ may depend on covariates X, but this dependence is suppressed.

Random variable U is the level of demand and describes the demand curve at different states of the world. Demand is maximal when $U = 1$ and minimal when $U = 0$, holding p fixed. Note that we imposed rank invariance (Equation 7), as is typical in classic supply-demand models, by making U invariant to p.

The model given in Equation 9 incorporates traditional additive error models for demand, which have $Y_p = q(p) + \varepsilon$, where $\varepsilon = Q_\varepsilon(U)$. The model is much more general in that the price can affect the entire distribution of the demand curve, whereas in traditional models, it affects only the location of the distribution of the demand curve.

The τ-quantile of the demand curve $p \mapsto Y_p$ is given by $p \mapsto q(p, \tau)$. Thus the curve $p \mapsto Y_p$ lies below the curve $p \mapsto q(p, \tau)$ with probability τ. Therefore, the various quantiles of the potential outcomes play an important role in describing the distribution and heterogeneity of the stochastic demand curve. The QTE may be characterized by $\partial q(p, \tau)/\partial p$ or by an elasticity $\partial \ln q(p, \tau)/\partial \ln p$. For example, consider the Cobb-Douglas model $q(p, \tau) = \exp(\beta(\tau) + \alpha(\tau)\ln p)$, which corresponds to a Cobb-Douglas model for demand with nonseparable error $Y_p = \exp(\beta(U) + \alpha(U)\ln p)$. The log transformation gives $\ln Y_p = \beta(U) + \alpha(U)\ln p$, and the QTE for the log-demand equation is given by the elasticity of the original τ-demand curve

$$\alpha(\tau) = \frac{\partial Q_{\ln Y_p}(\tau)}{\partial \ln p} = \frac{\partial \ln q(p, \tau)}{\partial \ln p}.$$

The elasticity $\alpha(U)$ is random, depends on the state of the demand U, and may vary considerably with U. For example, this variation could arise when the number of buyers varies and aggregation induces a nonconstant elasticity across the demand levels. Chernozhukov & Hansen (2008) estimate a simple demand model based on data from a New York fish market that was first collected and used by Graddy (1995). They find point estimates of the demand elasticity, $\alpha(\tau)$, that vary quite substantially from -2 for low quantiles to -0.5 for high quantiles of the demand curve.

The third condition in Equation 3, $P \in \{p : \rho(p, Z, U) = q(p, \mathcal{U})\}$, is the equilibrium condition that generates endogeneity; the selection of the clearing price P by the market depends on the potential demand and supply outcomes. As a result, we have a representation that is consistent with Condition 3, $P = \delta(Z, V)$, where V consists of U and $\mathcal{U}$ and may include sunspot variables if the equilibrium price is not unique. Thus what we observe can be written as

$$Y := q(P, U), \quad P := \delta(Z, V), \quad U \text{ is independent of } Z. \tag{10}$$

Identification of the τ-quantile of the demand function, $p \mapsto q(p, \tau)$, is obtained through the use of IV Z, such as weather conditions or factor prices, that shift the supply curve and do not affect the level of the demand curve, U, so that the independence assumption (Condition 2) is met. Furthermore, the IVQTE model allows arbitrary correlation between Z and V. This property is important as it allows, for

example, Z to be measured with error or to be exogenous relative to the demand equation but endogenous relative to the supply equation.

Example 2 (Savings): Chernozhukov & Hansen (2004) use the framework of the IVQTE model to examine the effects of participating in a 401(k) plan on an individual's accumulated wealth. Because wealth is continuous, wealth, Y_d, in the participation state $d \in \{0, 1\}$ can be represented as

$$Y_d = q(d, X, U_d), \quad U_d \sim U(0, 1),$$

where $\tau \mapsto q(d, X, \tau)$ is the conditional quantile function of Y_d, and U_d is an unobserved random variable. U_d is an unobservable that drives differences in accumulated wealth conditional on X under participation state d. Thus one might think of U_d as the preference for saving and interpret the quantile index τ as indexing rank in the preference for saving distribution. One could also model the individual as selecting the 401(k) participation state to maximize expected utility:

$$D = \arg\max_{d \in \mathcal{D}} E[W\{Y_d, d\}|X, Z, V] = \arg\max_{d \in \mathcal{D}} E[W\{q(d, x, U_d), d\}|X, Z, V], \tag{11}$$

where $W\{Y_d, d\}$ is the random indirect utility derived under participation state d.[8] As a result, the participation decision is represented by

$$D = \delta(Z, X, V),$$

where Z and X are observed, V is an unobserved information component that may be related to ranks U_d and includes other unobserved variables that affect the participation state, and function δ is unknown. This model fits into the IVQTE model, with the independence condition (Condition 2) requiring that U_d is independent of Z, conditional on X.

The simplest form of rank similarity is rank invariance (Equation 7), under which the preference for saving vector U_d may be collapsed to a single random variable $U = U_0 = U_1$. In this case, a single preference for saving is responsible for an individual's ranking across all treatment states. The rank similarity condition (Condition 4) is a more general form of rank invariance. It relaxes the exact invariance of ranks U_d across d by allowing noisy, unsystematic variations of U_d across d, conditional on (V, X, Z). This relaxation allows for variation in rank across the treatment states, requiring only an expectational rank invariance. Similarity implies that given the information in (V, X, Z) employed to make the selection of treatment D, the expectation of any function of rank U_d does not vary across the treatment states. That is, ex ante, conditional on (V, X, Z), the ranks may be considered to be the same across potential treatments, but the realized, ex post, rank may be different across treatment states.

From an econometric perspective, the similarity assumption is nothing but a restriction on the unobserved heterogeneity component, which precludes systematic

[8]It may depend both on observables in X and on realized and unrealized unobservables. Only the dependence on Y_d and d is highlighted.

variation of U_d across the treatment states. To be more concrete, consider the following simple example in which

$$U_d = F_{V+\eta_d}(V + \eta_d),$$

where $F_{V+\eta_d}(\cdot)$ is the distribution function of $V + \eta_d$, and $\{\eta_d\}$ are mutually independent and identically distributed conditional on V, X, and Z. The variable V represents an individual's mean saving preference, whereas η_d is a noisy adjustment.[9] This more general assumption leaves the individual optimization problem (Equation 11) unaffected, while allowing variation in an individual's rank across different potential outcomes.

Although we feel that similarity may be a reasonable assumption in many contexts, imposing similarity is not innocuous. In the context of 401(k) participation, matching practices of employers could jeopardize the validity of the similarity assumption. To be more concrete, let $U_d = F_{V+\eta_d}(V + \eta_d)$ as before but let $\eta_d = dM$ for random variable M, which depends on the match rate and is independent of V, X, and Z. Then conditional on $V = \upsilon$, X, and Z, $U_0 = F_V(\upsilon)$ is degenerate, but $U_1 = F_{V+M}(\upsilon + M)$ is not. Therefore, U_1 is not equal to U_0 in distribution. Similarity may still hold in the presence of the employer match if the rank, U_d, in the asset distribution is insensitive to the match rate. The rank may be insensitive if, for example, individuals follow simple rules of thumb such as target saving when they make their savings decisions. Also, if the variation of match rates is small relative to the variation of individual heterogeneity or if the covariates capture most of the variation in match rates, then similarity may be satisfied approximately.

Example 3 (Discrete choice model with market-level data): Berry & Haile (2010) show that a general model for market-level data realized from a discrete choice problem can fit within the IVQTE model. To keep notation and exposition simple, we consider a much-simplified version of the model from Berry & Haile (2010) in which consumer i's indirect utility from choosing product j is

$$U_{ijt} = u(X_{jt}, P_{jt}, \xi_{jt}, V_{ijt}) = u(\delta_j(X_{jt}, \xi_{jt}), P_{jt}, V_{ijt}),$$

where t indexes markets; X_{jt} are observed exogenous product-market characteristics; P_{jt} is the observed price of product j in market t, which is treated as endogenous; ξ_{jt} are product-market-specific unobservables; and V_{ijt} are individual product-market-specific unobservables that have density $f(\cdot)$. Thus the model imposes that unobserved product-market-specific effects and observed variables X_{jt} may affect utility only through the index $\delta_{jt} = \delta_j(X_{jt}, \xi_{jt})$, where the latter function may differ arbitrarily across products but is the same across all markets. That unobserved product characteristics affect utility only through a scalar index is a substantive restriction but is common in the literature on discrete choice models, in which, for example, one can interpret the index as an aggregate representing product quality.

[9]Clearly, similarity holds in this case, $U_d \stackrel{d}{=} U_{d'}$ given V, X, and Z.

An individual will then choose the product that maximizes individual utility. Letting Y_{it} denote the observed choice of individual i, we have that

$$Y_{it} = \arg\max_{j \leq J} U_{ijt},$$

where we assume that the same J products are available in each market for simplicity.[10] The market share of each product is then given as

$$\begin{aligned} S_{jt} &= \int 1\Big\{u(\delta_{jt}, P_{jt}, \upsilon) = \max_{k \leqslant J} u(\delta_{kt}, P_{kt}, \upsilon)\Big\} f(\upsilon) d\upsilon \\ &:= s_j\Big(\{\delta_{jt}, P_{jt}\}_{j=1}^{J}\Big) = s_j(\delta_t, P_t), \end{aligned}$$

where $\delta_t = (\delta_{1t}, \ldots, \delta_{Jt})'$ and $P_t = (P_{1t}, \ldots, P_{Jt})'$.

To fit this model into the IV quantile regression model, Berry & Haile (2010) make several assumptions to produce a structural relationship that is monotonic in a scalar unobservable. First, they assume that the utility function $u(\delta_{jt}, P_{jt}, V_{ijt})$ is strictly increasing in δ_{jt}. This assumption is standard in the discrete choice literature and coincides with the interpretation of δ_{jt} as product quality, in which higher-quality products are associated with higher utility, all else equal. Monotonicity of the utility function is not sufficient because all that is observed is the market share, which depends on the utility of each potential choice. Thus Berry & Haile (2010) make an additional assumption that they term "connected substitutes." Intuitively, this condition implies that an increase in the quality of every good within some strict subset of the available choices will be associated with the total market share of all goods not in the subset decreasing as long as the quality of no good outside of the subset increases. Berry & Haile (2010) show that the connected substitutes condition is satisfied in usual random utility discrete choice models and that it can hold fairly generally. Using these assumptions, Berry & Haile (2010) use a result from Gandhi (2008) that shows that the system of equations

$$S_{jt} = s_j(\delta_t, P_t)$$

has a unique solution for the vector δ_t as long as all goods present in equilibrium have positive market shares. Thus we may write

$$\delta_{jt} = g_j(S_t, P_t) \tag{12}$$

for some function g_j, where $S_t = (S_{1t}, \ldots, S_{Jt})'$.

From Equation 12, we have that $\delta_j(X_{jt}, \xi_{jt}) = g_j(S_t, P_t)$. To complete the argument, Berry & Haile (2010) assume that the function $\delta_{jt} = \delta_j(X_{jt}, \xi_{jt})$ is strictly increasing in its second argument, ξ_{jt}, which represents unobserved product attributes. This

[10]Obviously, identification of the model requires normalizations. For example, the utility from one of the options is generally normalized to 0. As this model is not the focus of this review, we do not discuss these normalizations, which are discussed in detail in a more general context in Berry & Haile (2010).

condition rules out the case in which ξ_{jt} can represent attributes that would increase utility for some individuals but decrease utility for others and again corresponds to the notion that ξ_{jt} represents unobserved product quality in which an increase unambiguously makes the product more desirable. With the assumed monotonicity in the function δ_j, one obtains

$$\xi_{jt} = \delta_j^{-1}\left(g_j(S_t, P_t); X_{jt}\right) = h_j(S_t, P_t, X_{jt}).$$

It is also clear that $h_j(X_t, P_t, S_t)$ is strictly increasing in S_{jt}, which is proven in Berry & Haile (2010, lemma 5), from which it follows that

$$S_{jt} = q_j(S_{-jt}, P_t, X_{jt}, \xi_{jt}),$$

where S_{-jt} denotes the set of market shares for each product in market t excluding product j, and q_j is an unknown function that is strictly increasing in ξ_{jt}. Then q_j can be taken as the structural function in the IV quantile model after the normalization that ξ_{jt} follows a $U(0, 1)$, assuming that ξ_{jt} has an atomless distribution. The model is then completed by assuming the existence of instruments, Z_t, that are independent of ξ_{jt} conditional on X_{jt} and are related to the endogenous variables through $(S'_{-jt}, P'_t)' = \Delta(Z_t, X_{jt}, V_t)$ for some function Δ and unobservables V_t. Finally, note that the model assumes rank invariance in its construction.

3. THE IDENTIFYING POWER OF INSTRUMENTAL VARIABLE QUANTILE RESTRICTIONS

The purpose of this section is to examine the identifying power of conditional moment restrictions (Equation 3). Specifically, we give various conditions for point identification in this section, summarizing and updating some of the results known in the literature. We remark here that point identification is not required in applications in principle as there exist inference methods that apply without point identification. However, it is useful to know and understand conditions under which moment conditions are informative enough that the identification region shrinks to a single point; in such cases, the inference methods will also produce very informative confidence sets. We present point-identifying conditions first for the binary case, $D \in \{0, 1\}$ and $Z \in \{0, 1\}$, then for the case of D taking a finite number of values, and finally for the continuous case.

3.1. Conditions for Point Identification in the Binary Case

Here we consider the cases in which $D \in \{0, 1\}$ and $Z \in \{0, 1\}$. The following analysis is conditional on $X = x$ and for a given quantile $\tau \in \{0, 1\}$, but we suppress this dependence for ease of notation. Under the conditions of Theorem 1, we know that there is at least one function $q(d) := q(d, x, \tau)$ that solves $P[Y \leqslant q(D)|Z] = \tau$ a.s. The function $q(\cdot)$ can be equivalently represented by a vector of its values $q = (q(0), q(1))'$. Therefore, for vectors of the form $y = (y_0, y_1)'$, we have a vector of moment equations

$$\Pi(y) := (P[Y \leqslant y_D \mid Z = 0] - \tau, P[Y \leqslant y_D \mid Z = 1] - \tau)', \tag{13}$$

where $y_D := (1 - D) \cdot y_0 + D \cdot y_1$. We say that d is identified in some parameter space, $\mathcal{L}$, if $y = q$ is the only solution to $\Pi(y) = 0$ among all $y \in \mathcal{L}$.

We require that the Jacobian $\partial\Pi(y)$ of $\Pi(y)$ with respect to $y = (y_0, y_1)'$ exists and that it takes the form

$$\begin{aligned}\partial\Pi(y) &:= \begin{bmatrix} f_Y(y_0 \mid D=0, Z=0)P[D=0 \mid Z=0] & f_Y(y_1 \mid D=1, Z=0)P[D=1 \mid Z=0] \\ f_Y(y_0 \mid D=0, Z=1)P[D=0 \mid Z=1] & f_Y(y_1 \mid D=1, Z=1)P[D=1 \mid Z=1] \end{bmatrix} \\ &=: \begin{bmatrix} f_{Y,D}(y_0, 0 \mid Z=0) & f_{Y,D}(y_1, 1 \mid Z=0) \\ f_{Y,D}(y_0, 0 \mid Z=1) & f_{Y,D}(y_1, 1 \mid Z=1) \end{bmatrix}. \end{aligned} \tag{14}$$

For local identification, we take $\mathcal{L}$ as an open neighborhood of $q = (q(0), q(1))'$. For global identification, we use some definitions from Mas-Colell (1979) to define $\mathcal{L}$. In what follows, for every proper (nonnull) subspace $L \subset \mathbb{R}^2$, let $\text{proj}_L : \mathbb{R}^l \mapsto L$ denote the perpendicular projection map. A convex, compact polytope is a bounded convex set formed by an intersection of a finite number of closed half-spaces. Such a polytope is of full dimension in $\mathbb{R}^l$ if it has a nonempty interior in $\mathbb{R}^l$. A face of a polytope $\mathcal{L}$ is the intersection of any supporting hyperplane of $\mathcal{L}$ with $\mathcal{L}$ so that faces of a polytope necessarily include the polytope itself. For instance, a rectangle in $\mathbb{R}^2$ has one two-dimensional face given by itself, four one-dimensional faces given by its edges, and four zero-dimensional faces given by its vertices. A subspace spanned by a nonempty face of $\mathcal{L}$ is the translation to the origin of the minimal affine space containing that face.

Theorem 2 (Identification by full rank conditions): Suppose that $\Pi(q) = 0$, the support of D is $\{0, 1\}$, and the support of Z is $\{0, 1\}$. Assume that the conditional density $f_Y(y|D = d, Z = z)$ exists for each $y \in \mathbb{R}$ and $(d, z) \in \{0, 1\} \times \{0, 1\}$. (*a*) (Local identification) Suppose the Jacobian $\partial\Pi$ given by Equation 14 is continuous and has full rank at $y = q$; then the τ-quantiles of potential outcomes, $q = (q(0), q(1))'$, are identified in the region $\mathcal{L}$ given by a sufficiently small open neighborhood of q in $\mathbb{R}^2$. (*b*) (Global identification) Assume that region $\mathcal{L}$ contains q and can be covered by a finite number of compact convex two-dimensional polytopes $\{\mathcal{L}_j\}$, each containing q. Assume that for each j, $\partial\Pi$ is a C^1 Jacobian of $\Pi : \mathcal{L}_j \to \mathbb{R}^2$ and that, possibly after rearranging the rows of $\partial\Pi$, for each $y \in \mathcal{L}_j$ and each subspace $L \subset \mathbb{R}^2$ spanned by a face of $\mathcal{L}_j$ that includes y, the linear map

$$\text{proj}_L \circ \partial\Pi(y) : L \mapsto L$$

has a positive determinant. Then q is identified in $\mathcal{L}$.

The first result is a simple local identification condition of the type considered in Rothenberg (1971) that we provide to fix ideas. The second result is a global identification condition that extends the result in Chernozhukov & Hansen (2005) by allowing nonrectangular sets $\mathcal{L}$. This result is based on the global univalence theorems of Mas-Colell (1979). As explained below, the positive determinant condition requires the impact of instrument Z on the joint distribution of (Y, D) to be sufficiently rich. In particular, the instrument Z should not be independent of the endogenous variable D. We note that the existence of the conditional density $f_Y(y|D = d, Z = z)$ is required only for (d, z) in the support of (D, Z). Outside the support, we can define the conditional density as 0, so the existence condition is not very restrictive. Moreover, the condition is formulated so that $\mathcal{L}$ can take on relatively rich shapes that can carry useful economic restrictions. For instance, in the training context, a useful restriction on

the parameters is that training weakly increases the potential earning quantiles. This restriction can be implemented by taking some natural parameter space and intersecting it with the half-space $H = \{(y_0, y_1) \in \mathbb{R}^2 : y_1 \geqslant y_0\}$. Specifically, a cube $C = \{y \in \mathbb{R}^l : ||y||_\infty \leq K\}$ intersected with the half-space H is an example of a region $\mathcal{L}$ permitted by the global identification result (b).

Comment 1 (Simple sufficient conditions): To illustrate the conditions of the theorem, let us consider the parameter space $\mathcal{L}$ as either $\mathcal{L} = q + C$, i.e., a cube centered at q, or $\mathcal{L} = (q + C) \cap H$, i.e., the intersection of a cube centered at q with the half-space H. Consider the trivial covering of $\mathcal{L}$ by itself, i.e., $\mathcal{L}_j = \mathcal{L}$. Then the positive determinant condition of the theorem is implied by the following simple conditions:

$$\frac{f_{Y,D}(y_1, 1 \mid Z = 1)}{f_{Y,D}(y_0, 0 \mid Z = 1)} > \frac{f_{Y,D}(y_1, 1 \mid Z = 0)}{f_{Y,D}(y_0, 0 \mid Z = 0)} \text{ for all } y = (y_0, y_1) \in \mathcal{L} \tag{15}$$

and

$$f_{Y,D}(y_1, 1 \mid Z = 1) > 0, \quad f_{Y,D}(y_0, 0 \mid Z = 0) > 0, \quad \text{for all } y = (y_0, y_1) \in \mathcal{L}. \tag{16}$$

Alternatively, because we can rearrange the rows of $\partial\Pi$, which corresponds to reordering elements of vector Π, the positive determinant condition of the theorem is implied by the following simple conditions:

$$\frac{f_{Y,D}(y_1, 1 \mid Z = 1)}{f_{Y,D}(y_0, 0 \mid Z = 1)} < \frac{f_{Y,D}(y_1, 1 \mid Z = 0)}{f_{Y,D}(y_0, 0 \mid Z = 0)} \text{ for all } y = (y_0, y_1) \in \mathcal{L} \tag{17}$$

and

$$f_{Y,D}(y_1, 1 \mid Z = 0) > 0, \quad f_{Y,D}(y_0, 0 \mid Z = 1) > 0, \quad \text{for all } y = (y_0, y_1) \in \mathcal{L}. \tag{18}$$

The proof that these are sufficient conditions is given in the Appendix, and below we discuss the economic plausibility of these conditions.

Comment 2 (Plausibility of Equations 15 and 16): The condition given in Equation 16 seems quite mild, so we focus on that given in Equation 15. We can illustrate Equation 15 by considering the problem of evaluating a training program in which the Y's are earnings, the D's $\in \{0, 1\}$ are training states, and the Z's $\in \{0, 1\}$ are offers of training service. The condition given in Equation 15 may be interpreted as a monotone likelihood ratio condition. That is, the instrument Z should have a monotonic impact on the likelihood ratio specified in Equation 15. This monotonicity may be a weak condition in some contexts and a strong condition in others. For instance, if $\mathcal{L}$ is a cube $q + C$, then this condition may be considered relatively strong. However, if we impose monotonicity of the training impact on earning quantiles, so that $q(0) \leq q(1)$, i.e., $q \in \mathcal{L} = (q + C) \cap H$, then the condition given in Equation 15 would be trivially satisfied in many empirical settings. Indeed, it would suffice that the instrument Z, the offer of training services, increases the relative joint likelihood of receiving higher earnings and receiving the training service. In many instances, we also have $P[D = 1|Z = 0] = 0$; e.g., those not offered training services do not receive that training. When $P[D = 1|Z = 0] = 0$, the right-hand side of Equation 15 equals 0, which makes

the identification condition given in Equation 15 satisfied trivially even for the less convenient parameter sets such as $\mathcal{L} = q + C$.

3.2. Identification with Multiple Points of Support

We generalize the result of Theorem 2 to more general discrete treatments with discrete instruments. Consider the case when D has the support $\{1, \ldots, l\}$ and Z has the support $\{1, \ldots, r\}$ $(l \leq r < \infty)$. Note that function $q(\cdot)$ can be represented by a vector $q = (q(1), \ldots, q(l))' \in \mathbb{R}^l$. Under the conditions of Theorem 1, there is at least one function $q(d)$ that solves $P[Y \leq q(D)|Z] = \tau$ a.s. Therefore, for vectors of the form $y = (y_1, \ldots, y_l)'$ and the vector of moment equations

$$\Pi(y) = (P[Y \leqslant y_D \,|\, Z = z] - \tau, z = 1, \ldots, r)', \tag{19}$$

where $y_D := \sum_d 1[D = d] \cdot y_d$, the model is identified if $y = q$ uniquely solves $\Pi(y) = 0$.

We define matrix $\partial\Pi(y)$ as the $r \times l$ matrix with the (d, z) element given by $f_Y(y_d|D = d, Z = z)P[D = d| Z = z]$, where $z = 1, \ldots, r$ and $d = 1, \ldots, l$. We require this to be the Jacobian matrix of the map $y \mapsto \Pi(y)$ and impose full-rank-type conditions on submatrices of this Jacobian. To this end, let m denote any permutation of l distinct integers from $\{1, \ldots, r\}$, called l permutations, and $\mathcal{M}$ be a collection of all such permutations. Let $\Pi_m := (\Pi_j)_{j \in m}$, which maps $\mathbb{R}^l$ to $\mathbb{R}^l$, be a subvector of Π formed by selecting j-th elements of Π according to their order in m.[11] Let $\partial\Pi_m$ denote the corresponding $l \times l$ Jacobian matrix of Π_m. The following theorem generalizes Theorem 2.

> **Theorem 3 (Identification for discrete *D*)**: Suppose $\Pi(q) = 0$, the support of D is $\{1, \ldots, l\}$, and that of Z is $\{1, \ldots, r\}$. Assume that the conditional density $f_Y(y|D = d, Z = z)$ exists for each $y \in \mathbb{R}$, and $(d, z) \in \{1, \ldots, l\} \times \{1, \ldots, r\}$. (*a*) (Local identification) Suppose the Jacobian $\partial\Pi(y)$ defined above is continuous and has rank l at $y = q$. Then the τ-quantiles of potential outcomes, q, are identified in the region $\mathcal{L}$ given by a sufficiently small open neighborhood of q in $\mathbb{R}^l$. (*b*) (Global identification) Assume that region $\mathcal{L}$ contains q and can be covered by a finite number of compact convex l-dimensional polytopes $\{\mathcal{L}_j\}$, each containing q and having the following properties: For each j, there is an l-permutation $m(j) \in \mathcal{M}$, such that $\partial\Pi_{m(j)}$ is the C^1 Jacobian of $\Pi_{m(j)} : \mathcal{L}_j \to \mathbb{R}^l$, and for each $y \in \mathcal{L}_j$ and each subspace $L \subset \mathbb{R}^l$ spanned by a face of $\mathcal{L}_j$ that includes y, the linear map
>
> $$\text{proj}_L \circ \partial\Pi_{m(j)}(y) : L \mapsto L$$
>
> has a positive determinant. Then q is identified in $\mathcal{L}$.

We note that in the theorem existence of the conditional density $f_Y(y|D = d, Z = z)$ is required only for (d, z) in the support of (D, Z). This density can be defined to take on an arbitrary value for (d, z) outside the support. The first result is a simple local identification condition provided to fix ideas. The second result is a global identification condition based on Mas-Colell (1979, global univalence theorem 1). This result complements a similar result given in Chernozhukov & Hansen (2005) based on Mas-Colell (1979, global univalence theorem 2). The positive determinant condition requires the impact of instrument Z on the joint distribution of (Y, D) to be sufficiently rich.

[11] Note that this formulation allows one to reorder the elements of Π, which may be needed to achieve the required positive determinant condition as discussed in the binary case.

Comment 3 (An alternative sufficient condition): Here we recall an alternative sufficient condition from Chernozhukov & Hansen (2005), which is based on Mas-Colell (1979, global univalence theorem 2). Assume that region $\mathcal{L}$ contains q and can be covered by a finite number of compact convex l-dimensional sets $\{\mathcal{L}_j\}$, each containing q and having the following properties: (*a*) For each j, there is a permutation $m(j) \in \mathcal{M}$ such that $\partial\Pi_{m(j)}$ is the C^1 Jacobian of $\Pi_{m(j)} : \mathcal{L}_j \to \mathbb{R}^l$; (*b*) for each $y \in \mathcal{L}_j$,

$$\det\left[\partial\Pi_{m(j)}(y)\right] > 0;$$

(*c*) $\mathcal{L}_j$ possesses a C^1-smooth boundary $\partial\mathcal{L}_j$; and (*d*) for each $y \in \partial\mathcal{L}_j$, $l'[\partial\Pi_{m(j)}(y) + \partial\Pi_{m(j)}(y)']l > 0$ for each $l \in T(y) : l \neq 0$, where $T(y)$ is the subspace tangent to $\mathcal{L}_j$ at point y. Then q is identified in $\mathcal{L}$. This condition seems to require slightly stronger conditions on the boundary than the condition used in Theorem 3. The advantage of the conditions from Chernozhukov & Hansen (2005) is that they more transparently convey the full-rank nature of the conditions imposed.

3.3. Identification with General *D*

Finally we consider conditions for point identification in the case of more general D and Z that may take on a continuum of values. We let d denote elements in the support of D and let z denote elements in the support of Z. Without loss of much generality, we restrict attention to the case in which both Y and D have bounded support. We require the parameter space $\mathcal{L}$ to be a collection of bounded (measurable) functions $m : \mathbb{R}^k \mapsto \mathbb{R}$ containing $q(\cdot)$. We say that $q(\cdot)$ such that $P[Y \leq q(D)|Z] = \tau$ a.s. is identified in $\mathcal{L}$ if for any other $m(\cdot) \in \mathcal{L}$ such that $P[Y \leq m(D)|Z] = \tau$ a.s., $m(D) = q(D)$ a.s. Below, we use $||\cdot||_{p,P}$ to denote the $L_p(P)$ norm.

Theorem 4 (Identification with general *D*): Suppose that $P[Y \leq q(D)|Z] = \tau$ a.s., and both Y and D have bounded support. Consider a parameter space $\mathcal{L}$, which is a collection of bounded (measurable) functions $m : \mathbb{R}^k \mapsto \mathbb{R}$ containing $q(\cdot)$. Assume that for $\varepsilon := Y - q(D)$, the conditional density $f_\varepsilon(e|D, Z)$ exists for each $e \in \mathbb{R}$, a.s. (*a*) (Global identification) Suppose that for each $\Delta(d) := m(d) - q(d)$ with $m(\cdot) \in \mathcal{L}$, $\omega_\Delta(D, Z) := \int_0^1 f_\varepsilon(\delta\Delta(D)|D, Z)d\delta > 0$ a.s. and

$$E[\Delta(D) \cdot \omega_\Delta(D, Z)|Z] = 0 \text{ a.s.} \Rightarrow \Delta(D) = 0 \text{ a.s.} \tag{20}$$

Then $q(\cdot)$ is identified in $\mathcal{L}$. (*b*) (Local identification) Suppose that $\omega_0(D, Z) := f_\varepsilon(0|D, Z) > 0$ a.s., and for each $\Delta(d) := m(d) - q(d)$ with $m(\cdot) \in \mathcal{L}$,

$$E[\Delta(D) \cdot \omega_0(D, Z)|Z] = 0 \text{ a.s.} \Rightarrow \Delta(D) = 0 \text{ a.s.}, \tag{21}$$

and, for some $0 \leq \eta < 1$ and $1 \leq p$,

$$\|E[\Delta(D) \cdot \{\omega_\Delta(D, Z) - \omega_0(D, Z)\}|Z]\|_{p,P} \leqslant \eta\|E[\Delta(D) \cdot \omega_0(D, Z)|Z]\|_{p,P}. \tag{22}$$

Then $q(\cdot)$ is identified in $\mathcal{L}$.

Condition (*a*), mentioned in Chernozhukov & Hansen (2005), states a nonlinear bounded completeness condition for global identification. The required condition given in Equation 20 is not primitive, but it highlights a useful link with the linear bounded completeness condition: $E[\Delta(D)|Z] = 0$ a.s. $\Rightarrow \Delta(D) = 0$ a.s. used by Newey & Powell (2003). The latter condition is needed for identification in the mean IV model $E[Y - q(D)|Z] = 0$ under the assumption of a bounded structural function q. The latter condition is known to be quite weak, as shown by D'Haultfoeuille (2011), and there are many primitive sufficient conditions that imply this condition. Andrews (2011) and Chen et al. (2011) show that linear completeness is generic in several senses under some conditions. Although the condition given in Equation 20 is not primitive, it is not vacuous either as the previous theorems provide primitive conditions for its validity. The local identification condition (*b*), obtained by Chen et al. (2011), provides yet another sufficient condition for result (*a*). Result (*b*) replaces the nonlinear completeness condition given in Equation 20 by the linear completeness condition in Equation 21, which is easier to check. Result (*b*) also requires that the set $\mathcal{L}$ is a sufficiently small neighborhood of q and that functional deviations $m(\cdot) - q(\cdot)$ and the conditional density $f_\varepsilon(\cdot|D, Z)$ are sufficiently smooth. This is explained in detail by Chen et al. (2011), who also provide further primitive smoothness and completeness conditions.

4. OTHER APPROACHES TO QUANTILE MODELS WITH ENDOGENEITY

There are, of course, other sets of modeling assumptions that one could employ to build a quantile model with endogeneity. In this section, we briefly outline two other approaches that have been taken in the literature. The first, due to Abadie et al. (2002), extends the local average treatment effect (LATE) framework of Imbens & Angrist (1994) to QTE. The second, considered in Imbens & Newey (2009) and Lee (2007), uses a triangular structure to obtain identification.

4.1. Local Quantile Treatment Effects with Binary Treatment and Instrument

In fundamental work, Abadie et al. (2002) develop an approach to estimating QTE within the LATE framework of Imbens & Angrist (1994) in the case in which both the instrument and treatment variables are binary. The use of the LATE framework makes this approach appealing as many applied researchers are familiar with LATE and the conditions that allow identification and consistent estimation of this quantity. Importantly, the extension proceeds under exactly the same monotonicity requirement as needed for LATE.

Specifically, Abadie et al. (2002) show that the QTE for a subpopulation is identified if (*a*) the instrument Z is independent of the potential outcome errors, $\{U_d\}$, and the errors in the selection equation, V (independence); (*b*) $P(D_1 \geq D_0|X) = 1$, where D_1 is the treatment state of an individual when $Z = 1$ and D_0 is defined similarly, holds (monotonicity); and (*c*) other standard conditions are met. The subpopulation for whom the QTE is identified is the set of compliers, those individuals with $D_1 > D_0$. In other words, the compliers are the individuals whose treatment is altered by switching the instrument from 0 to 1. Monotonicity is key in this framework. The monotonicity condition rules out defiers, individuals who would receive treatment in the absence of the intervention represented by the instrument but would not receive treatment if placed into the treatment group. The effects for individuals who would always receive treatment or never receive treatment regardless of the value of the instrument are unidentified.

Looking at these conditions, we see that the model of Abadie et al. (2002) replaces the monotonicity assumption (Condition 1), the independence assumption (Condition 2), and the similarity assumption (Condition 4) with a different type of monotonicity and a stronger

independence assumption and identifies a different quantity: the QTE for compliers. The LATE-style approach has not yet been extended beyond cases with a binary treatment and a single binary instrument, whereas the IV quantile model of Chernozhukov & Hansen (2005) applies to any endogenous variables and instruments. Note that neither set of conditions nests the other, and neither framework is more general than the other. Thus the frameworks are best viewed as complements, providing two sets of conditions that can be considered when thinking about a strategy for estimating heterogeneous treatment effects.

Of course, the two sets of conditions may be mutually compatible. One such case is discussed in Chernozhukov & Hansen (2004). In this example, the pattern of results obtained from the two estimators is quite similar, and the difference between the estimates appears small relative to sampling variation. Further exploration of these two approaches and their similarities and differences may be interesting to consider.

4.2. Instrumental Variables Quantile Regression in Triangular Systems

Another compelling framework is based on assuming a triangular structure, as in Imbens & Newey (2009) (for related models and results, see also Chesher 2003, Koenker & Ma 2006, Lee 2007). The triangular model takes the form of a triangular system of equations

$$\begin{aligned} Y &= g(D, \varepsilon), \\ D &= h(Z, \eta), \end{aligned}$$

where Y is the outcome, D is a continuous scalar endogenous variable, ε is a vector of disturbances, Z is a vector of instruments with a continuous component, η is a scalar reduced-form error, and we ignore other covariates for simplicity. Importantly, the triangular system generally rules out simultaneous equations, which typically have the reduced form relating D to Z depending on a vector of disturbances. For example, in a supply and demand system, the reduced form for both price and quantity will generally depend on the unobservables from both the supply equation and the demand equation. Outside of η being a scalar, the key conditions that allow identification of quantile effects in the triangular system are the following: (*a*) The function $\eta \mapsto H(Z, \eta)$ is strictly increasing in η (monotonicity), and (*b*) D and ε are independent conditional on V for some observable or estimable V (independence).

The variable V is thus the control function conditional on which changes in D may be taken as causal. Imbens & Newey (2009) use $V = F_{D|Z}(D, Z) = F_\eta(\eta)$, where $F_\eta(\cdot)$ represents the cumulative distribution function of η as the control function, and show that this variable satisfies the independence condition under the additional condition that (ε, η) is independent of Z. They show that one may use $D = h(Z, \eta)$ to identify V under the assumed monotonicity of $h(Z, \eta)$ in η. Using V obtained in this first step, one may then construct the distribution of $Y|D, V$. Then integrating over the distribution of V and using iterated expectations, one has

$$\begin{aligned} \int F_{Y|D,V}(y, d, v) F_V(dv) &= \int 1(g(d, \varepsilon) \le y) F_\varepsilon(d\varepsilon) \\ &= \Pr(g(d, \varepsilon) \le y) := G(y, d). \end{aligned}$$

It then follows that the τ-th quantile of Y_d is $G^{-1}(\tau, d)$.

As with the framework of Abadie et al. (2002), the triangular model under the conditions given above is neither more nor less general than the model of Chernozhukov & Hansen (2005). The key difference between the approaches is that Chernozhukov & Hansen's (2005) model uses an

essentially unrestricted reduced form but requires monotonicity and a scalar disturbance in the structural equation. The triangular system relies on monotonicity of the reduced form in a scalar disturbance. In addition, the triangular system, as developed in Imbens & Newey (2009), requires a more stringent independence condition in that the instruments need to be independent of both the structural disturbances and the reduced-form disturbance. That the approaches impose structure on different parts of the model makes them complementary, with a researcher's choice between the two dictated by whether it is more natural to impose restrictions on the structural function or the reduced form in a given application.

The triangular model and the model of Chernozhukov & Hansen (2005) can be made compatible by imposing the conditions from the triangular model on the reduced form and the conditions from Chernozhukov & Hansen (2005) on the structural model. Torgovitsky (2012) considers identification and estimation when both sets of conditions are imposed and shows that the requirements on the instruments may be substantially relaxed relative to Chernozhukov & Hansen (2005) or Imbens & Newey (2009) in this case.

5. ESTIMATION AND INFERENCE

In the previous sections, we outline results that are useful for identifying QTE and structural functions that are monotonic in a scalar unobservable. In the following, we briefly review the literature on estimation and inference. We focus on estimation of the model of Chernozhukov & Hansen (2005) presented in Section 2 using the moment conditions derived in Theorem 1. For estimation of the triangular model, readers are referred to Imbens & Newey (2009) for nonparametric estimation and Lee (2007) for a semiparametric approach. Abadie et al. (2002) provide results for estimating the QTE for compliers within the LATE-style framework. Moreover, we only review approaches for estimating parametric quantile functions: $q(D, X, \tau) = g(D, X, \tau, \theta)$ for $\theta \in \Theta \subset \mathbb{R}^m$. Horowitz & Lee (2007) and Gagliardini & Scaillet (2012) present nonparametric estimation and inference results for the IVQTE model using the condition given in Equation 3.

There are two practical issues that make estimation and inference based on the condition in Equation 3 challenging. The first is that the sample analog to the condition in Equation 3 is nonsmooth, and the generalized method of moments (GMM) objective function that would be formed by using Equation 3 as the moment conditions is also generically nonconvex, even for linear quantile models. The second problem is that the model may suffer from weak identification as in the standard linear IV model; Stock et al. (2002) provide a useful introductory survey to weak identification and related inference methods in the linear IV model. In the quantile case, the problem of weak identification is more subtle than in the linear model in that some quantiles may be weakly identified while others may be strongly identified. The relevant object for defining the strength of identification of a given quantile is the covariance between D and Z weighted by the conditional density function of the unobservable at the given quantile (see Chernozhukov & Hansen 2008 for a formal definition of this object and related discussion).

Although the nonsmoothness and nonconvexity of the GMM criterion complicate optimization, they do not render the approach infeasible, especially when the dimension of D and X is not too large. Abadie (1997) considers this approach for estimating an income model and provides further discussion. One could also estimate the model parameters using the Markov chain Monte Carlo approach of Chernozhukov & Hong (2003). This approach bypasses the need for optimization, instead relying on sampling and averaging to estimate model parameters. Note that this approach is not a cure-all because it requires careful tuning in applications. It is also worth noting that standard samplers may perform poorly in even simple linear IV models when

identification is not strong (see Hoogerheide et al. 2007). In an approach related to optimizing the GMM criterion function directly, Sakata (2007) proposes estimating the parameters of an IV quantile model by optimizing a different nonsmooth, nonconvex criterion function.

To partially circumvent the numerical problems in optimizing the full GMM criterion, Chernozhukov & Hansen (2006) suggest a different procedure termed the inverse quantile regression for the linear quantile model $q(D, X, \tau) = D'\alpha(\tau) + X'\beta(\tau)$. The basic intuition for the inverse quantile regression comes from the observation that if one knew the true value of the coefficient on D, $\alpha(\tau)$, the τ-th quantile regression of $Y - D'\alpha(\tau)$ onto X and Z would yield zero coefficients on the instruments Z. This observation allows one to effectively concentrate $\beta(\tau)$ out of the problem and leaves a nonsmooth, nonconvex optimization problem over only the parameters $\alpha(\tau)$. Because D is low dimensional in many applications, one can usually solve this optimization problem using highly robust optimization procedures such as a grid search.

Algorithmically, the inverse quantile regression estimates for a given probability index τ of interest can be obtained as follows using a grid search over $\alpha(\tau)$:

1. Define a suitable set of values $\{\alpha_j, j = 1, \ldots, J\}$ and estimate the coefficients $\beta(\alpha_j, \tau)$ and $\gamma(\alpha_j, \tau)$ from the model $Y - D'\alpha_j = X'\beta(\alpha_j, \tau) + Z'\gamma(\alpha_j, \tau) + \varepsilon$ by running the ordinary τ-quantile regression of $Y - D'\alpha_j$ on X and Z. Call the estimated coefficients $\widehat{\beta}(\alpha_j, \tau)$ and $\widehat{\gamma}(\alpha_j, \tau)$.
2. Save the inverse of the variance-covariance matrix of $\widehat{\gamma}(\alpha_j, \tau)$, which is readily available in any common implementation of the ordinary quantile regression. Denote this variance matrix $\widehat{A}(\alpha_j, \tau)$. Form $W_n(\alpha_j, \tau) = \widehat{\gamma}(\alpha_j, \tau)'\widehat{A}(\alpha_j, \tau)^{-1}\widehat{\gamma}(\alpha_j, \tau)$. Note that $W_n(\alpha_j)$ is the Wald statistic for testing $\gamma(\alpha_j, \tau) = 0$.
3. Choose $\widehat{\alpha}(\tau)$ as a value among $\{\alpha_j, j = 1, \ldots, J\}$ that minimizes $W_n(\alpha, \tau)$. The estimate of $\beta(\tau)$ is then given by $\widehat{\beta}(\widehat{\alpha}(\tau), \tau)$.

Chernozhukov & Hansen (2006, 2008) provide conditions under which the resulting estimator for $\alpha(\tau)$ and $\beta(\tau)$ is consistent and asymptotically normal and provide a consistent variance estimator. Marmer & Sakata (2012) provide a similar multistep algorithm that circumvents the same numeric problems using the objective function of Sakata (2007).

The good behavior of the asymptotic approximations obtained in Chernozhukov & Hansen (2006, 2008) relies on strong identification of the model parameters just as in the linear IV case. Intuitively, strong identification for a quantile of interest requires that a particular density-weighted covariation matrix between D and Z is not local to being rank deficient and that the impact of Z is rich enough to guarantee that the moment equations have a unique solution. The first condition is analogous to the usual full-rank condition in linear IV analysis, and the second condition is required because of the nonlinearity of the problem. Checking these conditions in practice may be difficult, and it is therefore useful to have inference procedures that are robust to violations of these conditions.

Fortunately, there are several inference procedures that remain valid under weak or partial identification. A nice feature of the algorithm defined for estimating $\alpha(\tau)$ above is that it produces a weak-identification-robust inference procedure naturally as a by-product. Chernozhukov & Hansen (2008) show that the Wald statistic, $W_n(\alpha, \tau)$, converges in distribution to $\chi^2_{\dim(Z)}$ under the null that $\alpha = \alpha_0$, where we let α_0 denote the true value of $\alpha(\tau)$ without needing either of the conditions discussed in the preceding paragraph. Thus a valid $(1 - p)\%$ confidence region for $\alpha(\tau)$ may be constructed as the set

$$\{\alpha : W_n(\alpha, \tau) \leq c_{1-p}\}, \tag{23}$$

where c_{1-p} is such that $\Pr(\chi^2_{\dim(Z)} > c_{1-p}) = p$, and the set is approximated numerically by considering the α's in the grid $\{\alpha_j, j = 1, \ldots, J\}$. Chernozhukov & Hansen (2008) show that the

confidence set in Equation 23 is valid when the model parameters are strongly identified and remains valid when the model is weakly identified or even unidentified. Marmer & Sakata (2012) provide a similar procedure. Jun (2008) provides a different approach to performing weak-identification-robust inference in models defined by the condition given in Equation 3. Finally, Chernozhukov et al. (2009) show that one can form statistics for inference about the entire parameter vector θ that are conditionally pivotal in finite samples for models defined by quantile restrictions such as Equation 3. Because the statistics do not depend on unknown nuisance parameters in finite samples, the exact distributions of these statistics can be calculated, and inference can proceed without relying on asymptotic approximations or statements about the strength of identification. The distributions produced in Chernozhukov et al. (2009) are not standard and so must be calculated by simulation.

6. CONCLUSION AND DIRECTIONS FOR FUTURE RESEARCH

In this article, we review approaches for building quantile models in the presence of endogeneity, focusing on conditions that can be used for identification. We also briefly review some of the practical issues that arise in estimation of IV quantile models and approaches to dealing with these issues. The models and estimation strategies outlined and cited above have already seen use in empirical economics, in which they have mostly been used for their ability to uncover interesting distributional effects. We also note above that the identification strategy employed in this article can be used to uncover structural objects, even if quantile effects are not the chief objects of interest, as in Berry & Haile (2010).

Whereas the results reviewed in this article are useful in a variety of contexts, there remain interesting areas for research in quantile models with endogeneity. In some applications, features of the conditional distribution are not the chief objects of interest, and researchers are interested in effects of treatments on unconditional quantiles. Given the set of conditional quantiles, such unconditional effects may be uncovered. In recent work, Froelich & Melly (2008) propose a different approach, related to Abadie et al. (2002), to estimating structural effects of endogenous variables on unconditional quantiles directly. It would also be interesting to think about quantile-like quantities for multivariate outcomes with endogenous covariates. The results reviewed in this article offer one possible approach for quantile modeling with endogeneity, but there remain many interesting directions and other approaches to be explored in further research.

APPENDIX

A.1. Proof of Theorem 1

Conditioning on $X = x$ is suppressed. For P-a.e value z of Z,

$$\begin{aligned} P[U_D \le \tau | Z = z] &\overset{(1)}{=} \int P\big[U_D \leqslant \tau | Z = z, V = v\big] dP\big[V = v | Z = z\big] \\ &\overset{(2)}{=} \int P\big[U_{\delta(z,v)} \leqslant \tau | Z = z, V = v\big] dP\big[V = v | Z = z\big] \\ &\overset{(3)}{=} \int P\big[U_0 \leqslant \tau | Z = z, V = v\big] dP\big[V = v | Z = z\big] \\ &\overset{(4)}{=} P\big[U_0 \leqslant \tau | Z = z\big] \overset{(5)}{=} \tau. \end{aligned} \tag{24}$$

Equality 1 in Equation 24 is by definition. Equality 2 is by the representation in Condition 3. Equality 3 is by the similarity assumption (Condition 4) and representation in Condition 3: Conditional on $(V = v, Z = z)$, $D = \delta(z, v)$ is a constant so that by Condition 4, $U_{\delta(z,v)}$ has the same

distribution as U_0, where 0 denotes any fixed value of D. Equality 4 is by definition, and equality 5 is by the independence assumption (Condition 2). This shows the first result.

The second result follows from the first and the equivalence of the events $\{q(D, U) \leq q(D, \tau)\} = \{U \leq \tau\}$ under $u \mapsto q(d, u)$ strictly increasing for each d on the domain [0, 1]. To show the third result, we note that

$$\{U \in I\} \subseteq (\{u : q(D, u) = q(D, U)\} \cap I \neq \emptyset).$$

Because $Y = q(D, U)$, the latter event is equivalent to the event $\{Y \in q(D, I)\}$, where $q(D, I)$ denotes the image of I under the mapping $u \mapsto q(D, u)$. The third result then follows from the first result.

A.2. Proof of Theorems 2 and 3

The local identification results follow by a standard argument, introduced in Rothenberg (1971), which we omit for brevity. The global identification result is obtained as follows. By assumption, $q \in \mathcal{L}$. Hence we need to check whether $y = d$ is the only solution to $\Pi(y) = 0$ over $\mathcal{L}$. Consider a covering set $\mathcal{L}_j$ and the l permutation $m(j)$ corresponding to it, as defined in the theorem. By assumption, $\Pi_{m(j)}(q) = 0$. By assumption, $q \in \mathcal{L}_j$. The stated rank conditions, compactness, and convexity of the polytope $\mathcal{L}_j$ imply that the mapping $y \to \Pi_{m(j)}(y)$, which maps $\mathcal{L}_j \subset \mathbb{R}^l$ to $\mathbb{R}^l$, is a homeomorphism (one-to-one) between $\mathcal{L}_j$ and $\Pi_{m(j)}(\mathcal{L}_j)$ by the global univalence theorem (Mas-Colell 1979, theorem 1). Thus $y = q$ is the unique solution of $\Pi_{m(j)}(y) = 0$ over $\mathcal{L}_j$. As this argument applies to every j, and $\{\mathcal{L}_j\}$ cover $\mathcal{L}$, it follows that $y = q$ is the unique solution of $\Pi(y) = 0$ over $\mathcal{L}$.

A.3. Proof of Theorem 4

We have that q solves $P[Y \leq q(D)|Z] = \tau$ a.s., and $q \in \mathcal{L}$ by assumption. Hence we need to check whether q is the only solution to $P[Y \leq q(D)|Z] = \tau$ a.s. in $\mathcal{L}$. Suppose there is $m \in \mathcal{L}$ such that $P[Y \leq m(D)|Z] = \tau$ a.s. Define $\Delta(d) := m(d) - q(d)$ and write

$$\begin{aligned} P[Y \leqslant m(D)|Z] - P[Y \leqslant q(D)|Z] &\overset{(1)}{=} E\left[E\left[\int_0^1 f_\varepsilon(\delta\Delta(D)|D, Z)\Delta(D)d\delta|D, Z\right]\Big|Z\right] \\ &\overset{(2)}{=} E\left[\int_0^1 f_\varepsilon(\delta\Delta(D)|D, Z)\Delta(D)d\delta|Z\right] \\ &\overset{(3)}{=} E[\Delta(D) \cdot \omega_\Delta(D, Z)|Z]. \end{aligned} \tag{25}$$

Note that equality 1 in Equation 25 follows by the fundamental theorem of calculus, equality 2 by the law of iterated expectations, and equality 3 by the linearity of the Lebesgue integral. For uniqueness we need that Equation $25 = 0$ a.s. $\Rightarrow \Delta(D) = 0$ a.s., which is assumed. Result (a) follows.

Result (b) is immediate from result (a) by the triangle inequality for $||\cdot||_{p,P}$.

A.4. Proof of the Sufficiency of Equations 15 and 16

Here we show that Equations 15 and 16 are sufficient for identification over the parameter space $\mathcal{L} = (q + C) \cap H$. Let e_1 and e_2 be coordinate vectors in $\mathbb{R}^2$, and let L_k denote the various subspaces spanned by faces of $\mathcal{L}$ containing y. In particular, we have $L_1 := \mathbb{R}^2$ for all y in the two-dimensional face $F_1 := \mathcal{L}$, $L_2 := \text{span}(e_2)$ for all y in the one-dimensional faces F_2 given by the left and right edges of $\mathcal{L}$, $L_3 := \text{span}(e_1)$ for all y in the one-dimensional faces F_3 given by the top and bottom edges of $\mathcal{L}$, and $L_4 = \text{span}(e_1 + e_2)$ for all y in the one-dimensional face F_4 given by the edge produced by the intersection of $\mathcal{L}$ with the 45° line. The subspaces spanned by vertices, which are zero-dimensional faces

of $\mathcal{L}$, are null spaces, so we do not need to consider them. We compute the projections of the Jacobian map onto these subspaces: $\text{proj}_{L_1} \circ \partial\Pi(y)[l] = \partial\Pi(y)l$, $\text{proj}_{L_2} \circ \partial\Pi(y)[l] = f_{Y,D}(y_0, 0|Z = 0)l$, $\text{proj}_{L_3} \circ \partial\Pi(y)[l] = f_{Y,D}(y_1, 1|Z = 1)l$, $\text{proj}_{L_4} \circ \partial\Pi(y)[l] = [(f_{Y,D}(y_1, 1|Z = 1) + f_{Y,D}(y_0, 0|Z = 1) + f_{Y,D}(y_1, 1|Z = 0) + f_{Y,D}(y_0, 0|Z = 0))/2]l$, for $y \in F_k$ and $l \in L_k$ in each of the cases. We then compute the corresponding determinants of the maps

$$\text{proj}_{L_k} \circ \partial\Pi(y) : L_k \to L_k,$$

where determinants are computed with respect to the coordinate systems of L_k, as $\det[\partial\Pi(y)]$ for $k = 1$, $f_{Y,D}(y_0, 0|Z = 0)$ for $k = 2$, $f_{Y,D}(y_1, 1|Z = 1)$ for $k = 3$, and $(f_{Y,D}(y_1, 1|Z = 1) + f_{Y,D}(y_0, 0|Z = 1) + f_{Y,D}(y_1, 1|Z = 0) + f_{Y,D}(y_0, 0|Z = 0))/2$ for $k = 4$. Theorem 2 requires that these determinants are positive for values of $y \in F_k$. This condition is implied by the simpler conditions given in Equations 15 and 16. For the case of $\mathcal{L} = q + C$, verification is analogous, except that we do not need to consider L_4. Thus the positive determinant condition of Theorem 2 is implied by the conditions in Equations 15 and 16 for $\mathcal{L} = q + C$ as well.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank the Reviewing Editor, Isaiah Andrews, Denis Chetverikov, and Ye Luo for excellent comments and much help.

LITERATURE CITED

Abadie A. 1997. Changes in Spanish labor income structure during the 1980s: a quantile regression approach. *Invest. Econ.* 21:253–72

Abadie A, Angrist J, Imbens G. 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* 70:91–117

Andrews DWK. 2011. Examples of L^2-complete and boundedly-complete distributions. *Discuss. Pap. 1801*, Cowles Found., Yale Univ., New Haven, CT

Andrews DWK, Shi X. 2013. Inference based on conditional moment inequalities. *Econometrica* 81:609–66

Autor DH, Houseman SN, Kerr SP. 2012. *The effect of work first job placements on the distribution of earnings: an instrumental variable quantile regression approach*. NBER Work. Pap. 17972

Beresteanu A, Molchanov I, Molinari F. 2011. Sharp identification regions in models with convex moment predictions. *Econometrica* 79:1785–821

Berry ST, Haile PA. 2010. Identification in differentiated products markets using market level data. *Discuss. Pap. 1744*, Cowles Found., Yale Univ., New Haven, CT

Chen X, Chernozhukov V, Lee S, Newey WK. 2011. *Local identification of nonparametric and semiparametric models*. arXiv:1105.3007v1

Chen X, Linton O, Keilegom IV. 2003. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* 71:1591–608

Chen X, Pouzo D. 2009. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *J. Econom.* 152:46–60

Chen X, Pouzo D. 2012. Estimation of nonparametric conditional moment models with possibly nonsmooth moments. *Econometrica* 80:277–322

Chernozhukov V, Hansen C. 2004. The effects of 401(k) participation on the wealth distribution: an instrumental quantile regression analysis. *Rev. Econ. Stat.* 86:735–51
Chernozhukov V, Hansen C. 2005. An IV model of quantile treatment effects. *Econometrica* 73:245–62
Chernozhukov V, Hansen C. 2006. Instrumental quantile regression inference for structural and treatment effect models. *J. Econom.* 132:491–525
Chernozhukov V, Hansen C. 2008. Instrumental variable quantile regression: a robust inference approach. *J. Econom.* 142:379–98
Chernozhukov V, Hansen C, Jansson M. 2009. Finite sample inference for quantile regression models. *J. Econom.* 152:93–103
Chernozhukov V, Hong H. 2003. An MCMC approach to classical estimation. *J. Econom.* 115:293–346
Chernozhukov V, Imbens GW, Newey WK. 2007. Instrumental variable estimation of nonseparable models. *J. Econom.* 139:4–14
Chernozhukov V, Lee S, Rosen A. 2013. Intersection bounds: estimation and inference. *Econometrica* 81:667–739
Chesher A. 2003. Identification in nonseparable models. *Econometrica* 71:1405–41
Chesher A. 2005. Nonparametric identification under discrete variation. *Econometrica* 73:1525–50
Chesher A, Rosen A, Smolinski K. 2011. *An instrumental variable model of multiple discrete choice*. Work. Pap. CWP39/11, CeMMAP, London
Chesher A, Smolinski K. 2010. *Sharp identified sets for discrete variable IV models*. Work. Pap. CWP11/10, CeMMAP, London
D'Haultfoeuille X. 2011. On the completeness condition in nonparametric instrumental regression. *Econom. Theory* 27:460–71
Doksum K. 1974. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Ann. Stat.* 2:267–77
Durrett R. 1996. *Probability: Theory and Examples*. Belmont, CA: Duxbury. 2nd ed.
Ekeland I, Galichon A, Henry M. 2010. Optimal transportation and the falsifiability of incompletely specified economic models. *Econ. Theory* 42:355–74
Eren O. 2009. Does membership pay off for covered workers? A distributional analysis of the free rider problem. *Ind. Labor Relat. Rev.* 62:367–80
Forbes SJ. 2008. The effect of air traffic delays on airline prices. *Int. J. Ind. Organ.* 26:1218–32
Froelich M, Melly B. 2008. Unconditional quantile treatment effects under endogeneity. *Discuss. Pap. 3288*, IZA, Bonn, Ger.
Gagliardini P, Scaillet O. 2012. Nonparametric instrumental variable estimation of structural quantile effects. *Econometrica* 80:1533–62
Galichon A, Henry M. 2009. A test of non-identifying restrictions and confidence regions for partially identified parameters. *J. Econom.* 152:186–96
Galichon A, Henry M. 2011. Set identification in models with multiple equilibria. *Rev. Econ. Stud.* 78:1264–98
Gandhi A. 2008. On the nonparametric foundations of discrete choice demand estimation. *Discuss. Pap.*, Univ. Wisconsin–Madison
Graddy K. 1995. Testing for imperfect competition at the Fulton Fish Market. *Rand J. Econ.* 26:75–92
Hausman JA. 1977. Errors in variables in simultaneous equation models. *J. Econom.* 5:389–401
Hausman JA, Sidak JG. 2004. Why do the poor and the less-educated pay higher prices for long-distance calls? *Contrib. Econ. Anal. Policy* 3:3
Heckman J, Robb R. 1986. Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In *Drawing Inference from Self-Selected Samples*, ed. H Wainer, pp. 63–107. New York: Springer-Verlag
Heckman JJ, Smith J. 1997. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *Rev. Econ. Stud.* 64:487–535
Hong H, Tamer E. 2003. Inference in censored models with endogenous regressors. *Econometrica* 71:905–32
Honore B, Hu L. 2004. On the performance of some robust instrumental variables estimators. *J. Bus. Econ. Stat.* 22:30–39

Hoogerheide LF, Kaashoek JF, van Dijk HK. 2007. On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: an application of flexible sampling methods using neural networks. *J. Econom.* 139:154–80
Horowitz JL, Lee S. 2007. Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica* 75:1191–208
Imbens GW, Angrist JD. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62:467–75
Imbens GW, Newey WK. 2009. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* 77:1481–512
Jun SJ. 2008. Weak identification robust tests in an instrumental quantile model. *J. Econom.* 144:118–38
Koenker R. 2005. *Quantile Regression*. Cambridge, UK: Cambridge Univ. Press
Koenker R, Bassett GS. 1978. Regression quantiles. *Econometrica* 46:33–50
Koenker R, Geling O. 2001. Reappraising medfly longevity: a quantile regression survival analysis. *J. Am. Stat. Assoc.* 96:458–68
Koenker R, Ma L. 2006. Quantile regression methods for recursive structural equation models. *J. Econom.* 134:471–506
Kostov P. 2009. A spatial quantile regression hedonic model of agricultural land prices. *Spat. Econ. Anal.* 4:53–72
Lamarche C. 2011. Measuring the incentives to learn in Columbia using new quantile regression approaches. *J. Dev. Econ.* 96:278–88
Lee S. 2007. Endogeneity in quantile regression models: a control function approach. *J. Econom.* 141:1131–58
Lehmann EL. 1974. *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day
Marmer V, Sakata S. 2012. *Instrumental variables estimation with weak-identification-robust inference based on conditional quantile restriction*. Work. Pap., Univ. South. Calif., Los Angeles
Mas-Colell A. 1979. Homeomorphisms of compact, convex sets and the Jacobian matrix. *SIAM J. Math. Anal.* 10:1105–9
Maynard A, Qiu J. 2009. Public insurance and private savings: Who is affected and by how much? *J. Appl. Econ.* 24:282–308
Newey WK, Powell JL. 2003. Instrumental variable estimation of nonparametric models. *Econometrica* 71:1565–78
Rothenberg TJ. 1971. Identification in parametric models. *Econometrica* 39:577–91
Sakata S. 2007. Instrumental variable estimator based on conditional median restriction. *J. Econom.* 141:350–82
Santos A. 2012. Inference in nonparametric instrumental variables with partial identification. *Econometrica* 80:213–75
Somainiy P. 2012. *Competition and interdependent costs in highway procurement*. Work. Pap., Stanford Univ., Stanford, CA
Stock JH, Wright JH, Yogo M. 2002. A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econ. Stat.* 20:518–29
Torgovitsky A. 2012. *Identification of nonseparable models with general instruments*. Work. Pap., Northwestern Univ., Evanston, IL
van der Vaart AW, Wellner JA. 1996. *Weak Convergence and Empirical Processes*. Springer Ser. Stat. New York: Springer
Wehby GL, Murran JC, Castilla EE, Lopez-Camelo JS, Ohsfeldt RL. 2009. Quantile effects of prenatal care utilization on birth weight in Argentina. *Health Econ.* 18:1307–21

Annual Review of Economics

Volume 5, 2013

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Economics* articles may be found at http://econ.annualreviews.org