Inference on the Instrumental Quantile Regression Process for Structural and Treatment Effect Models

Victor Chernozhukov[a*], Christian Hansen[b]

[a] Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02142, USA

[b] Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02142, USA

First Version: June 2001 This Version: March 2004

Abstract

We introduce a class of instrumental quantile regression methods for heterogeneous treatment effect models and simultaneous equations models with nonadditive errors and offer computable methods for estimation and inference. These methods can be used to evaluate the impact of endogenous variables or treatments on the entire distribution of outcomes. We describe an estimator of the instrumental variable quantile regression process and the set of inference procedures derived from it. We focus our discussion of inference on tests of distributional equality, constancy of effects, conditional dominance, and exogeneity. We apply the procedures to characterize the returns to schooling in the U.S.

JEL Classification: C10, C11, C13, C15

Keywords: Instrumental Quantile Regression, Structural Estimation, Treatment Effects, Endogeneity, Stochastic Dominance, Hausman Test, Supply-Demand Equations with Random Elasticity, Returns to Education

*Corresponding Author. Tel +1-617-354-6361. Fax +1-617-253-1330. Email: vchern@mit.edu.

The previous working version of this paper was circulated under the title "Instrumental Quantile Regression Methods for Distributional Effects". We benefited from seminars at Cornell, Penn, University of Illinois at Urbana-Champaign (all in 2001), Harvard-MIT in 2002, and the EC2 2001 Conference on Causality and Exogeneity in Econometrics in Louvain-la-Neueve. We would like to thank Peter Boswijk and two anonymous referees for many valuable comments.

1 Introduction

Quantile regression is an important method of modeling heterogeneous effects and accounting for unobserved heterogeneity. The standard quantile regression model may be developed from the basic Skorohod representation. Using this representation, the outcome variable Y, conditional on the exogenous variable of interest $D = d$, takes the form

$$Y = q(D, U_D), \quad U_D|D \sim \text{Uniform}(0, 1) \tag{1.1}$$

where $q(d, \tau)$ is the conditional $\tau-$th quantile of Y given $D = d$ and U_D is the nonseparable error or rank. This model has played a fundamental role in statistics at least since Bhattacharya (1963), Lehmann (1974), and Doksum (1974). The conditional quantile function $\tau \mapsto q(d, \tau)$ captures the impact of D on the outcome at a given quantile, while the rank term U_D represents an index of unobserved heterogeneity. Consequently, the structural quantile effects (SQE) or, equivalently, the quantile treatment effects (QTE), cf. Doksum (1974),

$$\frac{\partial}{\partial d} q(d, \tau) \quad \text{or} \quad q(d, \tau) - q(d', \tau)$$

represent a causal or structural effect of D on observational units, holding the unobserved heterogeneity U_D fixed at $U_D = \tau$. The quantile effects typically vary across τ, implying heterogeneous, non-constant effects. Koenker and Bassett (1978) and Bhattacharya (1963) introduced estimation methods for this model based on the conditional moment restriction

$$P[Y \leq q(D, \tau)|D] = P[U_D \leq \tau|D] = \tau \text{ for each } \tau \in (0, 1), \tag{1.2}$$

and estimation has been further developed by Powell (1986) and Portnoy (1991), among others.

In this paper, we consider estimation and inference for the endogenous generalization of the above model, which is particularly suited to the setting of observational studies where variable D is often endogenous. The model, introduced and analyzed in Chernozhukov and Hansen (2001a), takes the form

$$Y = q(D, U_D), \quad U_D|Z \sim \text{Uniform}(0, 1), \tag{1.3}$$

where Y is the outcome of interest, D is the endogenous or treatment variable of interest, and Z is an instrumental variable that is correlated with D but is independent of rank variable U_D. Chernozhukov and Hansen (2001a) show how (1.3) can be derived from primitive conditions that impose a generalized form of rank invariance and independence of structural (potential) outcomes from the instrument. Under these conditions, the function $q(d, \tau)$ represents the τ-quantile of the outcome in the population under the hypothetical exogenous assignment of variable D. Chernozhukov and Hansen (2001a) also provide the conditions required for nonparametric identification of this function from the instrumental analog of equation (1.2):

$$P[Y \leq q(D, \tau)|Z] = P[U_D \leq \tau|Z] = \tau. \tag{1.4}$$

This paper makes two key contributions. The first contribution is to offer an instrumental variable quantile regression (IV-QR) estimator of the quantile function $\tau \mapsto q(d, \tau)$ for the leading (linear) case

and develop a set of inference tools for examining a number of interesting hypotheses. Our estimator is a quantile analog of two stage least squares.[2] Effectively, as in the canonical two stage least squares model, instrumentation eliminates the endogeneity and selection bias commonly occurring in observational and experimental studies with imperfect compliance. Thus, the IV-QR process allows us to measure the exogenous treatment effect as in a fully controlled experiment, whereas the conventional QR process is inherently biased. The second contribution is the introduction of a class of tests based on the IV-QR process which allow us to examine numerous interesting hypotheses, including **(1)** the hypothesis of **distributional equality**, or whether the treatment or endogenous variable has a significant effect on outcome Y; **(2)** the hypothesis of a constant or **non-varying treatment effect**, a fundamental hypothesis of causal and structural analysis, cf. Heckman (1990) and Doksum (1974); **(3)** the hypothesis of conditional **stochastic dominance**, a fundamental hypothesis as well, cf. Abadie (2002) and McFadden (1989); and **(4)** the hypothesis of **exogeneity**, or whether the treatment variable is exogenous, another essential hypothesis, e.g. Hausman (1978). The critical values are generated by score subsampling, which subsamples the scores or estimated influence functions without recomputing the estimates. This method enables fast, computable implementation.

The use of the approach is illustrated through estimation of the impact of schooling on earnings using the data and instruments of Angrist and Krueger (1991). We analyze the effect of schooling on earnings and find evidence in favor of heterogeneous schooling effects: The effect of an additional year of schooling on earnings varies from almost 30% at low earning quantiles to 10% at high earning quantiles. We also reject the hypothesis of exogeneity and accept the hypothesis of first order stochastic dominance. Other applications of the estimation and inference procedures of this paper can be found in Hausman and Sidak (2002), Januszewski (2002), D'Urso (2002), Frakes and Gruber (2003), and Chernozhukov and Hansen (2003), among others.[3]

This paper accompanies our previous paper, Chernozhukov and Hansen (2001a), that focuses on modeling and identification of QTE in the presence of endogeneity. The present paper introduces and establishes the properties of the instrumental variable quantile regression process and of the inference processes and test statistics derived from it.[4] It also provides practical bootstrap tools to carry out the tests.

The remainder of the paper is organized as follows. In the next section, we briefly discuss the causal model and provide examples demonstrating how economic models may be placed in our modeling framework. Section 3 presents the IV-QR process and develops its sampling theory. Section 4 develops inference procedures for the IV-QR process and presents practical inference for testing distributional hypotheses. Section 5 presents an empirical study, and Section 6 concludes.

[2] We do not use the term "two stage quantile regression" (2SQR) because it is already used to name the procedure proposed by Portnoy and Chen (1996) as an analog of the two stage LAD (2SLAD) of Amemiya (1982) and Powell (1983). This procedure has been widely used to estimate quantile effects under endogeneity. When the QTE vary across quantiles, the 2SQR does not solve (1.4) and thus is inconsistent relative to the treatment parameter of interest. Note that 2SLAD and 2SQR are still excellent strategies for estimating constant treatment effect models.

[3] A brief review of these applications is provided in Section 6.

[4] The IV-QR estimator for a single quantile was first defined and studied in the unpublished working paper, Chernozhukov and Hansen (2001b).

Notation. We use the concepts of stochastic convergence as defined in van der Vaart (1998). We use outer (inner) probabilities, P^* (P_*), to avoid measurability problems; $\to_p$ denotes convergence in probability under P^*; $\to_d$ means convergence in distribution of random vectors; and $\Rightarrow$ means weak convergence in the metric space of bounded functions. The expression "wp $\to 1$" means "with (inner) probability going to 1."

2 The Instrumental Quantile Regression Model

In this section we describe the modeling framework within which we operate. This model is introduced and analyzed in detail in our previous work, cf. Chernozhukov and Hansen (2001a).

2.1 Potential Outcomes and Quantile Effects

Our model is developed within the conventional potential outcome framework, cf. Heckman and Robb (1986) and Imbens and Angrist (1994). Potential real-valued outcomes are indexed against potential values d of the endogenous variable D, and denoted Y_d. For example, Y_d is an individual's outcome when $D = d$. The structural outcomes $\{Y_d\}$ are latent because given the selected treatment D, the observed outcome for each individual or state of the world is $Y \equiv Y_D$. That is, only one component of $\{Y_d\}$ is observed for each observational unit.

Of primary interest to us are the conditional quantiles of potential outcomes, denoted as

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

and the structural quantile effects (SQE) or, equivalently, the quantile treatment effects (QTE) that summarize the difference between the quantiles under different levels of d,

$$q(d, x, \tau) - q(d', x, \tau) \quad \text{or, if defined,} \quad \frac{\partial}{\partial d} q(d, x, \tau).$$

QTE represents a useful way of describing the effect of d on the marginal distribution of outcomes Y_d.[5]

Typically D is selected in relation to $\{Y_d\}$ inducing endogeneity or selection bias, so that the conditional quantile of selected Y given the selected D, is generally not equal to the quantile of potential outcome $q(d, x, \tau)$. This makes the conventional quantile regression inappropriate for the estimation of $q(d, x, \tau)$. Thus, a major obstacle to learning about the quantiles of potential outcomes is sample selectivity or endogeneity. The model presented in the following section states the conditions under which we can recover the quantiles of latent outcomes through a set of conditional moment restrictions.

2.2 The Instrumental Quantile Regression Model

Conditional on $X = x$, the potential outcome Y_d can be related to its quantile functions by the Skorohod representation as

$$Y_d = q(d, x, U_d), \text{ where } U_d \sim U(0, 1), \tag{2.1}$$

[5]The notion of QTE was rigorously introduced by Doksum (1974).

and $q(d, x, \tau)$ is the conditional τ-quantile of potential outcome Y_d. The equation (2.1) is true regardless of the number of disturbances than determine Y_d.[6] The rank variable U_d characterizes heterogeneity of outcomes for individuals or observational units with the same observed characteristics x and treatment d. It also determines their relative ranking in terms of potential outcomes. This allows interpretation of the QTE as *actual effects* on people or units having fixed the level of unobserved heterogeneity U_d at some level τ.

We are now prepared to state the model which is a list of five main conditions (some are representations) that hold jointly. Other, more technical conditions will be added to discuss identification and estimation.

ASSUMPTION 1 ***Main Conditions of the IV-QR Model:*** *Given a common probability space (Ω, F, P), for P-almost every value of X, Z, the following conditions A1-A5 hold jointly:*

A1 Potential Outcomes. *Given $X = x$, for each d, $Y_d = q(d, x, U_d)$, where $U_d \sim U(0, 1)$ and $q(d, x, \tau)$ is strictly increasing in τ.*

A2 Independence. *Given $X = x$, $\{U_d\}$ is independent of Z.*

A3 Selection. *Given $X = x, Z = z$, for unknown function δ and random vector ν, $D \equiv \delta(z, x, \nu)$.*

A4 Rank Invariance or Rank Similarity. *For each d and d', given (ν, X, Z), either*

$$(a) \quad U_d = U_{d'} \quad or \quad (b) \quad U_d \sim U_{d'}.$$

A5 Observed *variables consist of* $Y \equiv q(D, X, U_D)$, $D \equiv \delta(Z, X, \nu)$, X, Z.

The main testable implication of A1-A5, which provides an important link of the parameters of the IV-QR model to a set of conditional moment equations, is given in the following theorem.

THEOREM 1 (*Main Implication*) *Suppose conditions A1-A5 hold, then for any $\tau \in (0, 1)$, a.s.*

$$P[Y \leq q(D, X, \tau)|X, Z] = P[Y < q(D, X, \tau)|X, Z] = \tau, \tag{2.2}$$

and U_D is independent of Z and X.

The proof of Theorem 1 is given in Chernozhukov and Hansen (2001a). Equation (2.2) is a restriction that can be used to estimate the quantile process $\tau \mapsto q(d, x, \tau)$. Identification of the quantile process in the population does not require functional form assumptions, as shown in Chernozhukov and Hansen (2001a). (2.2) is simplest to see under rank invariance, i.e. when $U = U_d$ for each d. Under rank invariance, we have a simple model of the form

$$Y_d = q(d, x, U), \quad U \text{ is independent of } Z, \text{ given } X = x \ .$$

[6] For instance, suppose that the structural outcome (e.g. demand) is $Y_d = m(d, \eta)$, where η is a high-dimensional vector. Then $Y_d = q(d, U)$, where U is uniformly distributed scalar and $q(d, \tau)$ is the quantile function of demand $m(d, \eta)$ for a fixed d. This aggregation of disturbances is remarkable since it works irrespective of non-additivity and the dimension of disturbances η and results in the object $q(d, \tau)$ which is identifiable.

It is then immediate that the event $\{Y \leq q(D, X, \tau)\}$ is equivalent to $\{U \leq \tau\}$ yielding (2.2).

A detailed discussion of A1-A5 is given in Chernozhukov and Hansen (2001a). In the following examples, we briefly illustrate how economic models may be embedded in the IV-QR model. The first example illustrates how a typical schooling model may be considered within the IV-QR framework, and the second example demonstrates that the IV-QR model encompasses a general model of demand with non-separable error.

2.3 Example: A Roy Type Model of Returns to Education

An individual considers several levels of schooling denoted $d \in \mathcal{D} = \{0, 1, ..., \bar{d}\}$. The potential outcome under each schooling level is given by the individual's earnings under the different levels of training $\{Y_d, d = 0, 1, ..., \bar{d}\}$. Suppose that the potential earning outcomes, conditional on $X = x$, are given by

$$Y_d = q(d, x, U), \tag{2.3}$$

where rank $U \sim U(0, 1)$ indexes the unobserved heterogeneity, and $q(d, x, U)$ is increasing in U. $U \sim U(0, 1)$ is a natural normalization in view of the Skorohod representation. Thus the distribution of potential outcome Y_d is characterized by the quantile functions $q(d, x, \tau)$. The rank variable U is assumed to be determined by ability and other unobserved factors that do not vary with d.

The individual selects her schooling level to maximize her expected utility:

$$D = \arg\max_{d \in \mathcal{D}} \; E\left[\; W\{Y_d, d, X\} \Big| X, Z, \nu \right] = \arg\max_{d \in \mathcal{D}} \; E\left[\; W\{q(d, X, U), d, X\} \Big| X, Z, \nu \right], \tag{2.4}$$

where $W\{Y_d, d, X\}$ is the unobserved Bernoulli utility function (for example, $W\{y, d, x\}$ may be increasing in y but decreasing in d, cf. Heckman and Vytlacil (1999)). As a result, the selection is represented as in A3 by $D = \delta(Z, X, \nu)$ for some function δ, where Z and X are observed, and ν is an unobserved information component that is correlated with U and includes other unobserved variables that are relevant to making the education decision. This model is thus a special case of the IV-QR model. In this model, the independence condition A2 only requires that U is independent of Z, conditional on X.

The rank variable U (think of ability, for example) is made invariant to d, which ascribes an important role to conditioning on covariates X. Having a rich set of covariates makes rank invariance a more plausible approximation. The rank similarity condition A4(b) also relaxes rank invariance. This condition allows for noisy, unsystematic variations of rank variable U_d across d, conditional on the information (ν, X, Z) relevant to making the selection decision (2.4). Consider the following simple example, where for $f : \mathbb{R} \to [0, 1]$, $U_d = f(\nu + \eta_d)$, $\{\eta_d\}$ are mutually iid conditional on ν, X, Z. The variable ν represents "mean" rank or ability of a person, while η_d is a noisy adjustment of this rank across treatment states, relative to the group of people that have the same observed characteristics X.[7] This leaves the individual optimization problem (2.4) unaffected, while allowing variation in an individual's rank across different potential outcomes.

[7] Clearly similarity holds in this case : $U_d \overset{d}{=} U_{d'}$ given ν, X, Z.

2.4 Example: Demand with Non-Separable Error

The following is a generalization of the classic supply-demand example. Consider the model

$$\begin{aligned} Y_p &= q\left(p, U\right), \\ \tilde{Y}_p &= \rho\left(p, z, \mathcal{U}\right), \\ P &\in \{p : \rho\left(p, Z, \mathcal{U}\right) = q\left(p, U\right)\}, \end{aligned} \tag{2.5}$$

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. The random variable U is the level of demand and describes the demand curve in different states of the world.[8] Demand is maximal when $U = 1$ and minimal when $U = 0$, holding p fixed. Note that we impose rank invariance in this model by making U invariant to p, which implies A4 (a).

The model (2.5) incorporates traditional additive error models $Y_p = q(p) + \epsilon$, where $\epsilon = Q_\epsilon(U)$. However, the model is more general in that the price can affect the entire distribution of the demand curve, while in traditional models it only affects the location of the distribution of the stochastic demand curve. The τ-quantile of the demand curve $p \mapsto Y_p$ is given by $p \mapsto q(p, \tau)$. Thus with probability τ, the curve $p \mapsto Y_p$ lies below the curve $p \mapsto q(p, \tau)$. Therefore, the various quantiles $q(p, \tau)$ play a key role in describing the distribution and heterogeneity of the stochastic demand curve. The QTE is then characterized by $\partial q(p, \tau)/\partial p$, or, more conveniently, by the elasticity $\partial \ln q(p, \tau)/\partial \ln p$. For example, consider the Cobb-Douglas model $q(p, \tau) = \exp\left(\beta(\tau) + \alpha(\tau) \ln p\right)$ which corresponds to a Cobb-Douglas model for demand with non-separable 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 and depends on the state of the demand U and may vary considerably with U. This variation could arise when the number of buyers varies and aggregation induces a non-constant elasticity across the demand levels. For example, in an application to the Graddy (1995) data from a New York fish market, we find that the elasticity, $\alpha(\tau)$, varies quite substantially from -2 for low quantiles to -0.5 for high quantiles of the demand curve.[9]

The third condition in (2.5), $P \in \{p : \rho\left(p, Z, \mathcal{U}\right) = q\left(p, U\right)\}$, is the equilibrium condition that generates endogeneity – the selection of the market 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 A3,

$$P = \delta(Z, \nu), \text{ where } \nu = (U, \mathcal{U}, \text{“sunspot” variables}),$$

[8] As mentioned previously, the level of demand U may be determined by many disturbances. However, Skorohod representation allows aggregation of the unobserved disturbances into a single variable U. Indeed suppose the demand function is $Y_p = m(p, \eta)$, where η is a high-dimensional vector of unobserved disturbances. Then $Y_p = q(p, U)$, where U is uniformly distributed scalar and $q(p, \tau)$ is the quantile function of demand $m(p, \eta)$ for a fixed p. This aggregation of disturbances also applies to supply, where $\mathcal{U}$ is the level of supply.

[9] These estimation results are available upon request.

where "sunspot" variables are present if there are multiple equilibria. Thus what we observe can be written as simultaneous equations with the form

$$Y \equiv q(P, U), \quad P \equiv \delta(Z, \nu), \quad U \text{ is independent of } Z. \tag{2.6}$$

Identification of the τ-quantile of the demand function, $p \mapsto q(p, \tau)$ is obtained through the use of instrumental variables Z, like weather conditions or factor prices, that shift the supply curve and do not affect the level of the demand curve, U, so that independence assumption A2 is met. Furthermore, the IV-QR model does not require Z to be jointly independent of both U and ν. This is considerably more general than the requirement that both the error U and the unobserved components of ν are independent from the instrument Z. The latter property is violated, for example, when there is measurement error in Z or Z is exogenous relative to the demand equation but endogeneous relative to the supply equation; see Hausman (1977).

3 The Instrumental Variable Quantile Regression

3.1 The Principle

Recall from Koenker and Bassett (1978) that the (conventional) quantile regression estimator is formulated as finding the best predictor of Y given X under the asymmetric least absolute deviation loss $\rho_\tau(u) = (\tau - 1(u < 0))u$. In other words, assuming integrability, the τ-th conditional quantile of Y given X solves the problem

$$Q_{Y|X}(\tau) \in \arg\min_{f \in \mathcal{F}} \; E\left[\rho_\tau(Y - f(X))\right],$$

where $\mathcal{F}$ is the class of measurable functions of X (that can be suitably restricted in applications).

The main implication of Theorem 1,

$$P[Y < q(D, X, \tau)|X, Z] = \tau, \text{ a.s.}, \tag{3.1}$$

is equivalent to the statement that 0 is the τ-th quantile of random variable $Y - q(D, X, \tau)$ conditional on (X, Z):

$$0 = Q_{Y-q(D,X,\tau)}(\tau|X, Z) \quad \text{a.s. for each } \tau. \tag{3.2}$$

Thus, we may pose the problem of finding a function $(d, x) \mapsto q(d, x, \tau)$ solving equation (3.1) as the **instrumental variable** or **inverse quantile regression**. This problem is to find a function $(d, x) \mapsto q(d, x, \tau)$ such that 0 is a solution to the quantile regression of $Y - q(D, X, \tau)$ on (Z, X):

$$0 \in \arg\min_{f \in \mathcal{F}} \; E\rho_\tau\left[(Y - q(D, X, \tau) - f(X, Z))\right], \tag{3.3}$$

where $\mathcal{F}$ is the class of measurable functions of (X, Z) (that will be suitably restricted in applications). The term 'inverse' emphasizes both the evident inverse relation of this problem to the conventional quantile regression of Koenker and Bassett (1978) and a connection to the ill-posed inverse problems of Tikhonov and Arsenin (1977).

3.2 An Instrumental Variable Quantile Regression Process and An Analogy with Two Stage Least Squares

For estimation purposes, we focus on the basic linear-in-parameters model

$$q(d, x, \tau) = d'\alpha(\tau) + x'\beta(\tau), \tag{3.4}$$

where d is an $l \times 1$ vector of treatment variables (possibly interacted with covariates) and x is a $k \times 1$ vector of (transformations of) covariates.

Next we consider a finite-sample analog of the population instrumental variable quantile regression. Define the weighted quantile regression objective function as

$$Q_n(\tau, \alpha, \beta, \gamma) \equiv \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - D_i'\alpha - X_i'\beta - \widehat{\Phi}_i(\tau)'\gamma) \cdot \widehat{V}_i(\tau), \quad \text{where}$$

$\widehat{\Phi}_i(\tau) \equiv \widehat{\Phi}(\tau, X_i, Z_i)$ is an $\dim(\alpha) \times 1$ vector of (transformations of) instruments, and $\widehat{V}_i(\tau) \equiv \widehat{V}(\tau, X_i, Z_i)$ is a positive weight function.

A practical formulation would be to use constant weights, $\widehat{V}_i = 1$, and use the instrument $\widehat{\Phi}_i(\tau)$ formed by the least squares projection of D_i on Z_i and X_i (and possibly their powers). In principle, we could include more elements in vector $\widehat{\Phi}_i(\tau)$ than the dimension of α. However, efficiency can instead be improved by choosing $\widehat{\Phi}_i(\tau)$ and $\widehat{V}_i(\tau)$ appropriately.

The **instrumental variable quantile regression estimator** is defined as follows. Define $\|x\|_A = \sqrt{x'Ax}$, and let $A(\tau)$ be any uniformly positive definite matrix, e.g. $A(\tau) = I$ or $A(\tau) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Phi}_i(\tau)\widehat{\Phi}_i(\tau)'$.[10] Then define

$$\widehat{\alpha}(\tau) = \arg \inf_{\alpha \in \mathcal{A}} \left\| \widehat{\gamma}(\alpha, \tau) \right\|_{A(\tau)}, \qquad \text{where} \tag{3.5}$$

$$(\widehat{\beta}(\alpha, \tau), \widehat{\gamma}(\alpha, \tau)) = \underset{(\beta,\gamma) \in \mathcal{B} \times \mathcal{G}}{\arg \inf} \; Q_n(\tau, \alpha, \beta, \gamma), \tag{3.6}$$

$\mathcal{A}$ and $\mathcal{B}$ are compact parameter sets, and $\mathcal{G}$ is any fixed compact cube centered at 0. The parameter estimates are given by

$$\widehat{\theta}(\tau) \equiv \left(\widehat{\alpha}(\tau), \widehat{\beta}(\tau)\right) \equiv \left(\widehat{\alpha}(\tau), \widehat{\beta}(\widehat{\alpha}(\tau), \tau)\right). \tag{3.7}$$

The estimator (3.7) is a finite-sample instrumental quantile regression. It finds the parameter values for α and β through the inverse step (3.5) such that the value of coefficient $\widehat{\gamma}(\alpha, \tau)$ on the instrument Φ in the quantile regression step (3.6) is driven as close to zero as possible, by analogy with the population problem (3.2). In **practice**, this procedure is simple to implement as follows:

1. For a given probability index τ of interest, define a grid of values $\{\alpha_j, j = 1, ..., J\}$, and run the ordinary τ-quantile regression of $Y_i - D_i'\alpha_j$ on X_i and $\widehat{\Phi}_i(\tau)$ to obtain coefficients $\widehat{\beta}(\alpha_j, \tau)$ and $\widehat{\gamma}(\alpha_j, \tau)$.

[10] Exact form of $A(\tau)$ is not important here due to "exact identification".

2. Choose $\widehat{\alpha}(\tau)$ as the value among $\{\alpha_j, j = 1, ..., J\}$ that makes $\|\widehat{\gamma}(\alpha_j, \tau)\|$ closest to zero. The estimate $\widehat{\beta}(\tau)$ is then given by $\widehat{\beta}(\widehat{\alpha}(\tau), \tau)$.

The **instrumental variable quantile regression process** is then defined as

$$\widehat{\theta}(\cdot) \equiv \left(\widehat{\theta}(\tau), \quad \tau \in \mathcal{T}\right),$$

where $\mathcal{T}$ is a closed subinterval of $(0, 1)$. In practice, we can compute $\widehat{\theta}(\tau_j)$ for a finite collection of probability indices τ, e.g. $\{0.1, ..., 0.9\}$, and interpolate in between.

Perhaps unexpectedly, the proposed estimation method may be viewed as an appropriate quantile regression analog of **two stage least squares**. To explain the analogy, ignore the covariates X for simplicity and consider the least squares analog of (3.5)-(3.6):

$$\widehat{\alpha} = \arg\inf_{\alpha} \left[\widehat{\gamma}(\alpha)' \left(\frac{1}{n}\sum_{i=1}^{n} \Phi_i \Phi_i'\right) \widehat{\gamma}(\alpha)\right], \qquad \text{where } \widehat{\gamma}(\alpha) = \arg\inf_{\gamma} \left[\frac{1}{n}\sum_{i=1}^{n} (Y_i - D_i'\alpha - \Phi_i'\gamma)^2\right], \tag{3.8}$$

and Φ_i denotes a suitable instrument. It is then obvious that (3.8) yields two stage least squares as the solution. With additional covariates X, the proposed estimation method (3.5)-(3.6) may be thought of as a peculiar analog of two stage least squares, modified in a manner that makes computation feasible.

3.3 Computational Properties and Comparison with Other Estimation Approaches

An estimator that is theoretically attractive, but uncomputable, has little value for data analysis. In many cases, the proposed instrumental variable quantile regression estimator is attractive from both a theoretical and a computational point of view. Indeed, there are three principal motivations for this estimator. First, it provides a theoretical link of the IV restrictions (3.1) to the conventional quantile regression. Second, it is computationally convenient, since it efficiently combines convex optimization with low-dimensional searches. The estimates are computed by implementing a series of ordinary quantile regressions (convex optimization problems) implying a need for a grid search only over the α-parameter (typically one-dimensional). Using the interior point-preprocessing methods introduced by Portnoy and Koenker (1997), the convex quantile regression steps are theoretically faster than OLS.[11] Third, the method can be viewed as a computationally attractive method of approximately solving the estimating equations:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} (1(Y_i \leq D_i'\widehat{\alpha} + X_i'\widehat{\beta}) - \tau)(X_i', \widehat{\Phi}_i(\tau)')'\widehat{V}_i(\tau) = o_p(1). \tag{3.9}$$

Thus, the estimator is asymptotically equivalent to a particular GMM estimator and, in principle, it may achieve maximal efficiency by choosing instruments Φ_i and weights V_i appropriately.

As stated, a simple implementation of inverse quantile regression only requires a low-dimensional search over α, where $\dim(\alpha)$ equals one or two in many applications.[12] There are other approaches, used in additive models, that one could use for estimation in the present nonseparable (heterogeneous effects) context:

[11] The computations can be improved further by employing parametric programming as in Koenker and D'Orey (1987). In this approach the quantile regression in (3.5) is initially solved for some α; one then solves for $\widehat{\beta}(\alpha, \tau)$ and $\widehat{\gamma}(\alpha, \tau)$ for nearby α using parametric programming.

[12] Computer programs in Matlab and Ox that implement the estimation and inference are available from the authors.

generalized method of moments based on (3.9), the minimum distance approach,[13] and LIML-type estimation.[14] In contrast to our approach, these other approaches often involve highly non-convex and multi-modal objective functions over many parameters. Implementation of extremum estimators with non-smooth or, more importantly, non-convex objective functions may require grid type searches over a subset of $\mathbb{R}^K$, where $K = \dim(\beta) + \dim(\alpha)$. For example, in the empirical application we consider, $K = 60 + 1$. In contrast, in problems like these the computation of our estimator is quite fast and often trivial. However, it must be noted that the computational advantages of our estimator rapidly diminish as the number of endogenous variables $\dim(\alpha)$ increases. In these cases, one can base a computable estimation procedure on the Markov Chain Monte Carlo approach to GMM estimation developed in Chernozhukov and Hong (2003), which also enables one to approximately solve equations (3.9) and obtain inference statements.

3.4 Theory of Identification, Estimation, and Basic Inference

In order to obtain properties of the IV-QR process, we impose a set of regularity conditions.

ASSUMPTION 2 (Conditions for Identification and Estimation) *In addition to (3.4), suppose*

R1 SAMPLING. *(Y_i, D_i, X_i, Z_i) are iid defined on the probability space (Ω, F, P) and take values in a compact set.*

R2 COMPACTNESS AND CONVEXITY. *For all τ, $(\alpha(\tau), \beta(\tau)) \in$ int $\mathcal{A} \times \mathcal{B}$, $\mathcal{A} \times \mathcal{B}$ is compact and convex.*

R3 FULL RANK AND CONTINUITY. *Y has bounded conditional density , a.s. $\sup_{y \in \mathbb{R}} f_{Y|(X,D,Z)}(y) < K$, and for $\pi \equiv (\alpha, \beta, \gamma)$, $\theta \equiv (\alpha', \beta')$, and*

$$\Pi(\pi, \tau) \equiv E\left[(\tau - 1(Y < D'\alpha + X'\beta + \Phi(\tau)'\gamma))\Psi(\tau)\right],$$
$$\Pi(\theta, \tau) \equiv E\left[(\tau - 1(Y < D'\alpha + X'\beta))\Psi(\tau)\right], \quad \Psi_i(\tau) \equiv V_i(\tau) \cdot [\Phi_i(\tau)', X_i']',$$

Jacobian matrices $\frac{\partial}{\partial(\alpha', \beta')}\Pi(\theta, \tau)$ and $\frac{\partial}{\partial(\beta', \gamma')}\Pi(\pi, \tau)$ are continuous and have full rank, uniformly over $\mathcal{A} \times \mathcal{B} \times \mathcal{G} \times \mathcal{T}$ and the image of $\mathcal{A} \times \mathcal{B}$ under the mapping $(\alpha, \beta) \mapsto \Pi(\theta, \tau)$ is simply-connected.

R4 ESTIMATED INSTRUMENTS AND WEIGHTS. *Wp $\to$ 1, the functions $\widehat{\Phi}(\tau, z, x)$, $\widehat{V}(\tau, z, x) \in \mathcal{F}$ and $\widehat{V}(\tau, z, x) \to_p V(\tau, z, x)$, $\widehat{\Phi}(\tau, z, x) \to_p \Phi(\tau, x, z)$ uniformly in (τ, z, x) over compact sets, where $V(\tau, z, x)$ and $\Phi(\tau, x, z) \in \mathcal{F}$; the functions $f(\tau, z, x) \in \mathcal{F}$ are uniformly smooth functions in (z, x) with the uniform smoothness order $\eta > dim(d, z, x)/2$,[15] and $\|f(\tau', z, x) - f(\tau, z, x)\| < C|\tau - \tau'|^a$, $C > 0, a > 0$, for all (z, x, τ, τ').*

[13] See Hogg (1975), Abadie (1995), Macurdy and Timmins (2000), and Hong and Tamer (2003) for pertinent results that with some work can be adapted to the present case.

[14] Sakata (2001) proposes a LIML-type estimator based on the absolute deviation for the classical location model, which solves: $\max_{\alpha,\beta} \min_{\gamma,\delta} \left[\sum_{i=1}^n |Y_i - D_i'\alpha - X_i'\beta - \Phi_i'\gamma - X_i'\delta| / \sum_{i=1}^n |Y_i - D_i'\alpha - X_i'\beta|\right]$. The computation of this estimator poses a serious challenge. The properties of this LAD-LIML estimator are analogous to those of LIML in the least squares case. We conjecture that the relative theoretical advantages and disadvantages of our estimator vs. Sakata's estimator are similar to the relative properties of 2SLS vs. LIML in the least squares case.

[15] This class of functions C_K^η is defined on page 154 in van der Vaart and Wellner (1996).

Remark 1. Condition R1 imposes iid sampling and compactness on the support of the variables. Compactness is not restrictive in micro-econometric applications, but it can be relaxed. Condition R2 imposes compactness on the parameter space. Such an assumption is needed at least for the parameter $\alpha(\tau)$ since the objective function is not convex in α. The role of R4 is to allow possibly estimated instruments and weights. Smoothness in R4 needs to hold only for the non-discrete sub-component of (d, x, z). Condition R4 allows for a wide variety of nonparametric and parametric estimators of instruments, as shown by Andrews (1994). The smoothness condition in R4 can be replaced by a more general condition of $\mathcal{F}$ having a finite $L_2(P)$-bracketing entropy integral. The condition in R3 implies global identification and the continuity condition in R3 together with R1 suffices for asymptotic normality. Clearly, these conditions may be refined at a cost of more complicated notation and proof.

Remark 2. The parametric identification condition R3 converts an intuitive local identification condition into a global one. Global identification is obtained through the use of a version of Hadamard's theorem. This condition is similar in spirit to the nonparametric identification conditions discussed in Chernozhukov and Hansen (2001a). This condition requires that the instrument Φ impacts the joint distribution of (Y, D) at many relevant points. The condition that the image of the parameter space be simply-connected requires that the image can be continuously homotopized (shrunk) to a point. I.e., it rules out "holes" in the image of the set. This condition may be thought of as ruling out poorly behaved distributions. One sufficient condition for the image to be simply-connected is follows from Mas-Colell (1979a).

LEMMA 1 *A sufficient condition for global identification is as follows: there exists a compact convex set $\mathcal{C}$ such that $\mathcal{A} \times \mathcal{B} \subset \mathcal{C}$, $\mathcal{C}$ has a smooth boundary $\partial\mathcal{C}$, $\det \frac{\partial}{\partial(\alpha', \beta')}\Pi(\theta, \tau) > 0$ over $\mathcal{C}$, and $\frac{\partial}{\partial(\alpha', \beta')}\Pi(\theta, \tau)$ is positive-quasi-definite on $\partial\mathcal{C}$ in the sense defined by Mas-Collel (1979a).*

Theorem 2 describes the identification of the parameters of the IV-QR model.

THEOREM 2 (*Identification by Full Rank Condition*) *Given Assumptions 1 - 2, $(\alpha, \beta')' = (\alpha(\tau)', \beta(\tau)')'$ uniquely solves the system of equations $E\left[\tau - 1(Y < D'\alpha + X'\beta)\Psi(\tau)\right] = 0$ over $\mathcal{A} \times \mathcal{B}$.*

Theorem 3 describes the large sample theory of the IV-QR process in the metric space of bounded functions $\ell^\infty(\mathcal{T})$.

THEOREM 3 (*Estimation*) *Given Assumptions 1-2, for $\epsilon_i(\tau) = Y_i - D_i'\alpha(\tau) + X_i'\beta(\tau)$ and $l_i(\tau, \theta(\tau)) = (\tau - 1(\epsilon_i(\tau) < 0))$,*

$$\sqrt{n}(\widehat{\theta}(\cdot) - \theta(\cdot)) = -J(\cdot)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_i(\cdot, \theta(\cdot))\Psi_i(\cdot) + o_p(1) \Rightarrow b(\cdot), \tag{3.10}$$

where $b(\cdot)$ is a mean zero Gaussian process with covariance function $E\, b(\tau)b(\tau')' = J(\tau)^{-1}S(\tau, \tau')[J(\tau')^{-1}]'$,

$$J(\tau) = E\left[f_{\epsilon(\tau)}(0|X, D, Z)\Psi(\tau)[D', X']\right], \; S(\tau, \tau') = (\min(\tau, \tau') - \tau\tau')E\Psi(\tau)\Psi(\tau')'.$$

Remark 3. (Basic Inference) A basic implication of Theorem 1 is that for any given probability index τ

$$\sqrt{n}\left(\widehat{\theta}(\tau) - \theta(\tau)\right) \to_d \mathcal{N}\Big(0, J(\tau)^{-1}S(\tau,\tau)[J(\tau)^{-1}]'\Big). \tag{3.11}$$

Also, for any finite collection of quantile indices $\{\tau_j, j \in J\}$

$$\Big\{\sqrt{n}\left(\widehat{\theta}(\tau_j) - \theta(\tau_j)\right)\Big\}_{j\in J} \to_d \mathcal{N}\Big(0, \Big\{J(\tau_k)^{-1}S(\tau_k,\tau_l)[J(\tau_l)^{-1}]'\Big\}_{k,l\in J}\Big), \tag{3.12}$$

which gives the joint limit distribution of IV-QR for several quantiles. The result in Theorem 2 is in fact stronger, assuring that the entire empirical instrumental quantile regression process $\widehat{\theta}(\cdot)$ asymptotically behaves continuously, enabling the uniform approximation of $\widehat{\theta}(\cdot)$ by a finite collection of instrumental regression quantiles $\widehat{\theta}(\tau_j), j \in J$ for a suitably fine grid of quantile indices $\{\tau_j, j \in J\}$.

Remark 4. (Standard Errors) The components of the asymptotic variance in (3.11) and (3.12) can be estimated as follows. The matrix $S_\Psi(\tau,\tau')$ can be estimated by its sample counterpart:

$$\widehat{S}_\Psi(\tau,\tau') = (\min(\tau,\tau') - \tau\tau')\frac{1}{n}\sum_{i=1}^{n}\widehat{\Psi}_i(\tau)\widehat{\Psi}_i(\tau')'. \tag{3.13}$$

Following Powell (1986), the estimator of $J_\Psi(\tau)$ takes the form

$$\widehat{J}_\Psi(\tau) = \frac{1}{2nh_n}\sum_{i=1}^{n} I(|\widehat{\epsilon}_i(\tau)| \le h_n)\widehat{\Psi}_i(\tau)[D_i', X_i'], \tag{3.14}$$

where $\widehat{\epsilon}_i(\tau) \equiv Y_i - D_i'\widehat{\alpha}(\tau) - X_i'\widehat{\beta}(\tau)$ and h_n is an appropriately chosen bandwidth, where $h_n \to 0$ and $nh_n^2 \to \infty$.[16] The results (3.11) and (3.12) enable a simple form of inference regarding conditional quantiles for various given probability indices. The next section will address more general inference questions.

Corollary 1 (Distribution-Free Limits) *For $W(\tau) = J(\tau)^{-1}E\Psi(\tau)\Psi(\tau)'[J(\tau)^{-1}]'$, we have $W(\cdot)^{-\frac{1}{2}}\sqrt{n}\left(\widehat{\theta}(\cdot) - \theta(\cdot)\right) \Rightarrow B_p(\cdot)$, where B_p is a standard p-dimensional Brownian bridge B_p ($p = \dim(\alpha) + \dim(\beta)$) with covariance operator $E\, B_p(\tau)B_p(\tau')' \equiv (\min(\tau,\tau') - \tau\tau')I_p$.*

Remark 5. (Weights and Instruments) When we choose the weight and instruments as $V^*(\tau) = f_{\epsilon(\tau)}(0|X,Z)$, $v(\tau) = f_{\epsilon(\tau)}(0|D,X,Z)$, $\Phi^*(\tau) = E[Dv(\tau)|X,Z]\ /\ V^*(\tau)$, and $\Psi^*(\tau) = V^*(\tau)[\Phi^*(\tau)' : X']'$, the variance function becomes $E\ b(\tau)b(\tau)' = \tau(1-\tau)\cdot[E\Psi^*(\tau)\Psi^*(\tau)']^{-1}$. This choice of instruments and weights leads to a pointwise efficient procedure.[17] This can be shown by appealing to the argument of Chamberlain (1987). Regularity condition R4 allows use of a wide variety of nonparametric estimators and parametric approximations of the optimal Φ and V. For particular examples of such procedures, see Amemiya (1977) and Andrews (1994). An example of a simple and practical strategy for empirical work is to construct Φ as an OLS projection of D on Z and X (and possibly their powers) and set $V_i = 1$.

[16] E.g., one may use the Sliverman's rule of thumb. Specific choices of h_n are discussed in Koenker (1994).

[17] The form of optimal weights and instrument for global case remains an interesting open question, since the complete quantile model involves a continuum of moment conditions.

4 General Inference

4.1 Inference Hypotheses and Procedures

It is convenient to embed our hypotheses in the following null hypothesis:

$$R(\tau)\Big(\theta(\tau) - r(\tau)\Big) = 0, \quad \text{for each } \tau \in \mathcal{T}, \tag{4.1}$$

where $R(\tau)$ denotes a known $q \times p$ matrix of rank q, $q \leq p = \dim(\theta(\tau))$, and $r(\tau) \in \mathbb{R}^p$. It is worth noting that this set-up differs from the classical one since $\theta(\cdot)$ and $r(\cdot)$ are functions and, in many cases, both have to be estimated.

The tests will be based on the instrumental variable quantile regression process, $\widehat{\theta}(\cdot)$. We will focus on the basic inference process

$$v_n(\cdot) = R(\cdot)\left(\widehat{\theta}(\cdot) - \widehat{r}(\cdot)\right), \tag{4.2}$$

and statistics of the form $S_n = f(\sqrt{n}v_n(\cdot))$ derived from it. In particular, we will be interested in the Kolmogorov-Smirnov (KS) and Smirnov-Cramer-Von-Misses (CM) statistics, which have

$$S_n = \sqrt{n}\sup_{\tau\in\mathcal{T}} \|v_n(\tau)\|_{\widehat{\Lambda}(\tau)}, \quad S_n = n\int_{\mathcal{T}} \|v_n(\tau)\|^2_{\widehat{\Lambda}(\tau)} d\tau, \tag{4.3}$$

respectively, where the symmetric $\widehat{\Lambda}(\tau) \to_p \Lambda(\tau)$ uniformly in τ, and $\Lambda(\tau)$ is a positive definite symmetric matrix uniformly in τ. The choice of $\Lambda(\tau)$ and $\widehat{\Lambda}(\tau)$ is discussed in Section 4.4. The null hypothesis is rejected if

$$S_n > c(1-\alpha)$$

where the critical value $c(1-\alpha)$ can be obtained using the resampling procedure in Section 4.3.

The following are examples of hypotheses that may be considered in this framework. For simplicity, we set $\dim(\alpha) = 1$ in what follows; extensions to the more general case are straightforward.

Example 1. ***Hypothesis of No Effect.*** A basic hypothesis is that the treatment has no impact on the outcomes: $\alpha(\tau) = 0$ for all τ in $\mathcal{T}$. In this case, $R(\cdot) = R = [1, 0, ...]$ and $r(\cdot) = 0$.

The next example presents a hypothesis of constant treatment effects. The alternative is that the effect varies across quantiles, which is of fundamental importance because it motivates modern structural and causal models developed specifically to cope with varying effects.

Example 2. ***Location-Shift or Constant Effect Hypothesis.*** The hypothesis of a constant effect is that the treatment D affects only the location of outcome Y, but not any other moments. That is, $\exists\alpha : \alpha(\tau) = \alpha,$ for all $\tau \in \mathcal{T}$, which asserts that $\alpha(\tau)$ is constant across all $\tau \in \mathcal{T}$. In this case, $R(\cdot) = R = [1, 0, ...]$ and $r(\cdot) = r = (\alpha, 0, ...)$, implying $Rr = \alpha$. The component r of the null hypothesis can be estimated by any method consistent with the null, e.g. $\widehat{r} = (\widehat{\alpha}(\frac{1}{2}), 0, ...)'$.

Example 3. ***Dominance Hypothesis.*** The test of stochastic dominance, or whether the effect is unambiguously beneficial, involves the dominance null $\alpha(\tau) \geq 0,$ for all $\tau \in \mathcal{T}$, versus the non-dominance alternative

$\alpha(\tau) < 0$, for some $\tau \in \mathcal{T}$. In this case, the *least favorable null* involves $R(\cdot) = R = [1, 0...]$ and $r(\cdot) = 0$, and one may use the one-sided KS or CM statistics,

$$S_n = \sqrt{n} \sup_{\tau \in \mathcal{T}} \max(-\widehat{\alpha}(\tau), 0), \text{ and } S_n = n \int_{\mathcal{T}} \| \max(-\widehat{\alpha}(\tau), 0) \|^2_{\Lambda(\tau)} d\tau$$

to test the hypothesis.

Example 4. ***Exogeneity Hypothesis.*** In the basic linear model, the quantiles of potential or counterfactual outcome Y_d, conditional on X, are given by $d'\alpha(\tau) + x'\beta(\tau)$. Suppose that the treatment D is chosen *independently of outcomes*, that is D is independent of $\{U_d\}$, conditional on X. Then the quantiles of realized outcome Y, conditional on D and X, are given by $D'\alpha(\tau) + X'\beta(\tau)$. Thus, in the absence of endogeneity, $(\alpha(\cdot)', \beta(\cdot)')'$ can be estimated using the conventional quantile regression without instrumenting. The difference between IV-QR estimates, $\widehat{\theta}(\cdot)$, and QR estimate, $\widehat{\vartheta}(\cdot)$, can be used to formulate a Hausman test of the null hypothesis of exogeneity:

$$\alpha(\tau) = \vartheta(\tau)_1 \text{ for each } \tau \text{ in } \mathcal{T}, \text{ where } \vartheta(\tau) \equiv \text{plim } \widehat{\vartheta}(\tau), \tag{4.4}$$

and $\widehat{\vartheta}(\cdot)_1$ is the QR estimate of $\alpha(\cdot)$ obtained without instrumenting. In this case, $R(\cdot) = [1, 0, ...]$ and $r(\cdot) = \vartheta(\cdot)$. The alternative of endogeneity states: $\exists \tau \in \mathcal{T} : \alpha(\tau) \neq \vartheta(\tau)_1$.

4.2 Formal Inference Results

ASSUMPTION 3 (Conditions for Inference)

- ***I.1*** *$R(\cdot)(\theta(\cdot) - r(\cdot)) = g(\cdot)$, where the functions $g(\tau), R(\tau), r(\tau)$ are continuous and either **(a)** $g(\tau) = 0$ for all τ or **(b)** $g(\tau) \neq 0$ for some τ.*
- ***I.2*** *$\sqrt{n}(\widehat{\theta}(\cdot) - \theta(\cdot)) \Rightarrow b(\cdot)$ and $\sqrt{n}(\widehat{r}(\cdot) - r(\cdot)) \Rightarrow d(\cdot)$ jointly in $\ell^\infty(\mathcal{T})$, where $b(\cdot)$ and $d(\cdot)$ are jointly zero mean Gaussian functions that may have different laws under the null and the alternative.*

Remark 6. Conditions I.1(a) and I.1(b) formulate the null and a global alternative. Condition I.2 requires that the estimates of $\theta(\cdot)$ and $r(\cdot)$ are asymptotically Gaussian. In our examples, I.2 holds by the Bahadur type representation of the IV-QR process (3.10) obtained in Theorem 2 and the corresponding representation of QR process. Section 3.4 contains further details. I.2 also permits other asymptotically Gaussian estimators of the parameters of the IV-QR model.

THEOREM 4 (Inference) *For f denoting the two- and one- sided KS or CM statistics*

1. *Under Assumptions 1 and 3:I.1(a), and 3:I.2 $S_n \to_d S \equiv f(v(\cdot))$, where $v(\cdot) = R(\cdot)(b(\cdot) - d(\cdot))$. If $v(\cdot)$ has nondegenerate covariance kernel, then for $\alpha < 1/2$, $P(S_n > c(1-\alpha)) \to \alpha = P(f(v(\cdot)) > c(1-\alpha))$, where $c(1-\alpha)$ is chosen so that $P(f(v(\cdot)) > c(1-\alpha)) = \alpha$.*
2. *Under Assumptions 1 and 3:I.1(b), and 3:I.2, $S_n \to_d \infty$ and $P_n(S_n > c(1-\alpha)) \to 1$.*

Theorem 4 states the limit distribution of the KS and CM statistics under the null and the alternative. In the statement of Theorem 4 we implicitly assume that for the case of one-sided tests in Example 3, the global alternatives to the least favorable null violate the composite null. Theorem 4 alone does not provide us with operational tests, since we do not know the critical value $c(1-\alpha) : P\left(f(v(\cdot)) > c(1-\alpha)\right) = \alpha$. In general, one faces the Durbin problem when estimating $c(1-\alpha)$, since the limit distribution of $f(v(\cdot))$ generally depends on P.[18] This dependence is caused by the estimation of component $r(\cdot)$ in the hypothesis. This leads to the presence of the nonstandard term $d(\cdot)$ in the limit inference process $v(\cdot) \equiv b(\cdot) + d(\cdot)$. In several important cases, such as Examples 1 and 3, the term $d(\cdot) = 0$ because $r(\cdot)$ is known and need not be estimated, which makes it possible to ensure the limit distribution of $f(v(\cdot))$ is independent of P by choosing an appropriate weight matrix $\Lambda(\tau)$ in (4.3).

Corollary 2 (Distribution-Free Inference) *Suppose $d(\cdot) = 0$ in I.2. If $\Lambda(\tau) = [R(\tau)W(\tau)R(\tau)']^{-1}$, with $W(\tau) = J(\tau)^{-1}E[\Psi(\tau)\Psi(\tau)'][J(\tau)^{-1}]'$ and $\widehat{\Lambda}(\tau) = \Lambda(\tau) + o_p(1)$ uniformly in τ, then $f(v_n(\cdot)) \Rightarrow f(B_q(\cdot))$, where B_q is the standard q-dimensional Brownian bridge with covariance function: $EB_q(\tau)B_q(\tau)' = (\min(\tau,\tau') - \tau\tau')I_q$.*

In other important cases, such as Examples 2 and 4, the simple transformation used in Corollary 2 will not provide distribution-free limits, see Durbin (1973). There are several ways to proceed. One method is the Khmaladze martingale transformation, cf. Bai (1997) and Koenker and Xiao (2002). Another method is to use a simple resampling procedure, recentering the inference process around its sample realization, cf. Chernozhukov (2002). The simulation examples in Chernozhukov (2002) suggest that resampling has an accurate size and somewhat better power than Khmaladzation. In the next section, we describe a different resampling method that delivers the same asymptotic quality and is more attractive computationally in the present setting.

4.3 Critical Values by Resampling Scores

The method of resampling we suggest does not require the recomputation of the estimates over the resampling steps, which may be quite laborious since the optimization problem requires many computations of ordinary quantile regressions for many values of α and τ. Instead we resample the linear approximations of the empirical inference processes. In addition, to facilitate a feasible, practical implementation, we employ the m out of n bootstrap (subsampling).

Suppose that we have a linear representation for the inference process:

$$\sqrt{n}(v_n(\cdot) - g(\cdot)) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i(\cdot) + o_p(1), \tag{4.5}$$

where $z_i(\cdot)$ is defined below in Proposition 1. Given a sample of the estimated scores, $\{\widehat{z}_i(\tau), i \leq n, \tau \in \mathcal{T}\}$, consider the following steps. (Estimation of scores is discussed below and other practical details are supplied in Section 4.4.)

[18] See e.g. Durbin (1973), Bai (1997), and Koenker and Xiao (2002) for related discussions

Step 1. Construct B_n randomly chosen subsets of $\{1, ..., n\}$ of size b. Denote such subsets as $I_i, i \leq B_n$.[19] Denote by $v_{j,b,n}(\cdot)$ the inference process computed over the j-th subset of data I_j, i.e. $v_{j,b,n}(\tau) \equiv \frac{1}{b}\sum_{i\in I_j} \widehat{z}_i(\tau)$, and define $S_{j,b,n} \equiv f(\sqrt{b}[v_{j,b,n}(\cdot)])$ as

$$\widehat{S}_{j,b,n} \equiv \sup_{\tau\in\mathcal{T}} \sqrt{b}\|v_{j,b,n}(\tau)\|_{\hat{\Lambda}(\tau)} \text{ or } \widehat{S}_{j,b,n} \equiv b\int_{\mathcal{T}} \|v_{j,b,n}(\tau)\|^2_{\hat{\Lambda}(\tau)} d\tau,$$

for cases when S_n is the Kolmogorov-Smirnov (KS) or Smirnov-Cramer-Von-Misses (CM) statistic, respectively.

Step 2. Define, for $S = f(v(\cdot))$, $\Gamma(x) \equiv P\{S \leq x\}$. Estimate $\Gamma(x)$ by $\widehat{\Gamma}_{b,n}(x) = B_n^{-1}\sum_{j=1}^{B_n} 1\{S_{j,b,n}(\tau) \leq x\}$. The critical value is obtained as the $1-\alpha$-th quantile of $\widehat{\Gamma}_{b,n}(x)$, i.e. $c_{b,n}(1-\alpha) = \inf\{c : \widehat{\Gamma}_{b,n}(c) \geq 1-\alpha\}$. The level α test rejects the null hypothesis when $S_n > c_{b,n}(1-\alpha)$.

In order to obtain the linear expansion (4.5), we maintain the following assumption.

ASSUMPTION 4 (Linear Representations) *In addition to I.1 and I.2*

I.3 *The estimates admit linear representations:* $\sqrt{n}\left(\widehat{\theta}(\cdot) - \theta(\cdot)\right) = -J(\cdot)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n l_i(\cdot,\theta(\cdot))\Psi_i(\cdot) + o_p(1)$ *and* $\sqrt{n}\left(\widehat{r}(\cdot) - r(\cdot)\right) = -H(\cdot)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n d_i(\cdot, r(\cdot))\Upsilon_i(\cdot) + o_p(1)$ *in* $\ell^\infty(\mathcal{T})$*, where* $J(\cdot)$ *and* $H(\cdot)$ *are constant invertible matrices, and* $l_i(\tau,\theta(\tau))\Psi_i(\tau)$ *and* $d_i(\tau, r(\tau))\Upsilon_i(\tau)$ *are i.i.d. mean zero for each* τ*.*

I.4 ***(a)*** *We have estimates* $l_i(\cdot,\widehat{\theta}(\cdot))\widehat{\Psi}_i(\cdot)$ *and* $d_i(\cdot,\widehat{r}(\cdot))\widehat{\Upsilon}_i(\cdot)$ *that take realizations in a Donsker class of functions with a constant envelope and are uniformly consistent in* τ *in the* $L_2(P)$ *norm.*[20] ***(b)*** *Wp* $\to 1$*,* $El_i(\tau,\theta(\tau))f_i(\tau)\big|_{f=\widehat{\Psi}} = 0$ *and* $Ed_i(\tau, r(\tau))f_i(\tau)\big|_{f=\widehat{\Upsilon}} = 0$ *for each* i*.* ***(c)*** $E\|l_i(\tau,\theta) - l_i(\tau,\theta')\| < C\,\|\theta' - \theta\|$*,* $E\|d_i(\tau, r) - d_i(\tau, r')\| < C\|r' - r\|$*, uniformly* $\tau \in \mathcal{T}$ *and in* (θ, θ', r, r') *over compact sets.*

Lemma 4 in Appendix C verifies that I.3 and I.4 are satisfied under Assumption 2 for the particular implementations that we use.

Proposition 1 (Linear Representations) *Under Assumptions 1,3, and 4* $\sqrt{n}(v_n(\cdot) - g(\cdot)) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n z_i(\cdot) + o_p(1)$*, in* $\ell^\infty(\mathcal{T})$*, where* $z_i(\cdot) = R(\cdot)\left[J(\cdot)^{-1}l_i(\cdot,\theta(\cdot))\Psi_i(\cdot) - H(\cdot)^{-1}d_i(\cdot, r(\cdot))\Upsilon_i(\cdot)\right]$*.*

Thus, the estimate of $z_i(\cdot)$ is given by $\widehat{z}_i(\cdot) = R(\cdot)\left[\widehat{J}(\cdot)^{-1}l_i(\cdot,\widehat{\theta}(\cdot))\widehat{\Psi}_i(\cdot) - \widehat{H}(\cdot)^{-1}d_i(\cdot,\widehat{r}(\cdot))\widehat{\Upsilon}_i(\cdot)\right]$, where $\widehat{J}(\cdot)$ and $\widehat{H}(\cdot)$ are any uniformly consistent estimates of $J(\cdot)$ and $H(\cdot)$, cf. Section 4.4.

Remark 7. In Assumption 4, condition I.3 requires that the estimates of $\theta(\cdot)$ and $r(\cdot)$ entering the null hypotheses have asymptotically linear representations of the form defined above and that asymptotic normality applies to these estimates. Note that I.3 is formulated so that other asymptotically Gaussian estimators of the IV-QR model are permitted. In our implementation, this condition is implied by the Bahadur type

[19] The subsampling is done without replacement. However, if $b^2/n \to 0$, subsampling without and subsampling with replacement are equivalent wp $\to 1$.

[20] In the sense that $\widehat{f}(W,\tau)$ is consistent to $f(W,\tau)$ in the $L_2(P)$ norm if $\sup_\tau \mathrm{E}[\|\tilde{f}(w,\tau) - f(w,\tau)\|^2]_{\tilde{f}=\widehat{f}} \to_p 0$.

representation of the IV-QR process (3.10) obtained in Theorem 2 and the corresponding representation of QR process. Conditions I.4(a) and I.4(c) impose sufficient smoothness for developing the theory of the resampling inference. These conditions are also satisfied in all of the examples considered in this paper. Condition I.4(b) is the familiar condition of "orthogonality", cf. Andrews (1994), which implies that the estimation of Ψ_i and Υ_i has no effect on the asymptotic distribution of the linear representation in I.3.

Next, we briefly go through our testing Examples 1-4 and state the scores for each of them.

1. ***Test of No Effect***: Since $r(\cdot) = 0$, $z_i(\tau) = R(\tau)\left[J(\tau)^{-1} l_i(\tau, \theta(\tau))\Psi_i(\tau)\right]$, where $l_i(\tau, \theta(\tau)) = (\tau - 1(Y_i < D_i\alpha(\tau) + X_i'\beta(\tau)))$, $\Psi_i(\tau) = V_i(\tau)[\Phi_i(\tau)', X']'$.
2. ***Test of Constant Effect:*** In this case, $\widehat{r}(\cdot) = \widehat{\theta}(\frac{1}{2})$ is an IV-QR estimate, and for $l_i(\cdot,\cdot)$ defined above $z_i(\tau) = R(\tau)\left[J(\tau)^{-1} l_i(\tau, \theta(\tau))\Psi_i(\tau) - J(\frac{1}{2})^{-1} l_i(\frac{1}{2}, \theta(\frac{1}{2}))\Psi_i(\frac{1}{2}))\right]$.
3. ***Test of Dominance Effect:*** Since $r(\cdot) = 0$, $z_i(\tau) = R(\tau)\left[J(\tau)^{-1} l_i(\tau, \theta(\tau))\Psi_i(\tau)\right]$.
4. ***Test of Exogeneity:*** If $r(\cdot)$ is estimated using conventional quantile regression as defined in Example 4, the score is given by $z_i(\tau) = R(\tau)\left[J(\tau)^{-1} l_i(\tau, \theta(\tau))\Psi_i(\tau) - H(\tau)^{-1} d_i(\tau, \vartheta(\tau))\right]$, where $d_i(\tau, \vartheta(\tau)) = (\tau - 1(Y_i < \tilde{X}_i'\vartheta(\tau))\tilde{X}_i$, $\tilde{X}_i = (D_i', X_i')'$, and $H(\tau) = E f_{Y|\tilde{X}}(\vartheta(\tau)'\tilde{X})\tilde{X}\tilde{X}'$.

Appendix C formally verifies I.3 and I.4 for these examples. Section 4.4 discusses estimation of H and J.

THEOREM 5 (Score Subsampling Inference) *Suppose Assumptions 1, 3 and 4 hold, and that we have* $\widehat{J}(\tau) = J(\tau) + o_p(1)$ *and* $\widehat{H}(\tau) = H(\tau) + o_p(1)$ *uniformly in* τ *over* $\mathcal{T}$. *Then as* $B_n \to \infty, b \to \infty, n \to \infty$: ***(1)*** *Under the null hypothesis, if* Γ *is continuous at* $\Gamma^{-1}(1-\alpha)$: $c_{b,n}(1-\alpha) \to_p \Gamma^{-1}(1-\alpha)$, $P(S_n > c_{b,n}(1-\alpha)) \to \alpha$; ***(2)*** *Under the alternative hypothesis,* $S_n \to_d \infty$, $c_{b,n}(1-\alpha) = O_p(1)$; $P(S_n > c_{b,n}(1-\alpha)) \to 1$; ***(3)*** $\Gamma(x)$ *is absolutely continuous at* $x > 0$ *when the covariance function of* v *is nondegenerate a.e. in* τ.

4.4 Practical Details

This section supplies some necessary implementation details.

Discretization. It is practical to use a grid $\mathcal{T}_n$ in place of $\mathcal{T}$ with the largest cell size $\delta_n \to 0$ as $n \to \infty$.

Corollary 3 *Theorems 1-4 are valid for piecewise constant approximations of the finite-sample processes using* $\mathcal{T}_n$, *given that* $\delta_n \to 0$ *as* $n \to \infty$.

Choice and Estimation of $\Lambda(\tau)$, $J(\tau)$ $H(\tau)$. In order to increase the test's power we could set $\Lambda^*(\tau) = [\Omega^*(\tau)]^{-1} = \mathrm{Var}\,[z_i(\tau)]^{-1}$, which is an Anderson-Darling type weight.[21] In iid samples, there are many methods for estimating $\Lambda^*(\tau)$, *uniformly consistently in* τ. By I.3 and I.4, a uniformly consistent estimate of $\Omega^*(\tau)$ is given by

$$\widehat{\Omega}^*(\tau) = \frac{1}{n}\sum_{i=1}^{n} \widehat{z}_i(\tau)\widehat{z}_i(\tau)', \quad \widehat{z}_i(\tau) = R(\tau)\left[\widehat{J}(\tau)^{-1} l_i(\tau, \widehat{\theta}(\tau))\widehat{\Psi}_i(\tau) - \widehat{H}(\tau)^{-1} d_i(\tau, \widehat{r}(\tau))\widehat{\Upsilon}_i(\tau)\right].$$

[21] This choice is not readily suited to Example 2, since Var $z_i(\frac{1}{2}) = 0$. However, we can cut out $[\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon]$ from the interval $\mathcal{T}$. Alternatively, one may always simply use $\Lambda(\tau) = I$.

A uniformly consistent estimate of $J(\tau)$ is given by the Powell (1986) estimator, $\widehat{J}(\tau) = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}((Y_i - D_i'\widehat{\alpha}(\tau) - X_i'\widehat{\beta}(\tau))\widehat{\Psi}_i(\tau)[D_i', X_i'])$, where $K_h(u) = h^{-1}1[|u| \leq h/2]$ and h_n is chosen as in (3.14). Estimates of $H(\tau)$ are needed in Examples 2 and 4. In Example 2, $\widehat{H}(\tau) = \widehat{J}(\frac{1}{2})$, and in Example 4, a uniformly consistent estimate of $H(\tau)$ is given similarly by $\widehat{H}(\tau) = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(Y_i - \tilde{X}_i'\widehat{\vartheta}(\tau))\tilde{X}_i\tilde{X}_i', \quad \tilde{X}_i = (D_i', X_i')$. We do not discuss the formal properties of these standard estimators to save space.[22]

Choice of the Block Size. In Politis et al. (1999) various rules are suggested for choosing an appropriate subsample size, including the calibration and minimum volatility methods. We use $b = 5n^{2/5}$, though the empirical results are not sensitive to the subsample size.

5 Returns to Schooling in The United States

One of the most widely studied topics in labor economics is the impact of education on earnings. The large volume of research in this area has been motivated by both interest in the causal effect of education on earnings as well as the inherent difficulty in measuring this effect. The difficulty arises due to the possible endogenous relationship between education and earnings. In particular, it seems likely that unobserved individual ability is correlated to both a person's education and wages, thus biasing standard regression estimates of the relation between schooling and earnings. In addition, economists have long believed that the returns to schooling may vary among individuals, further complicating the interpretation of conventional least squares and two stage least squares results.[23]

In order to address the issue of heterogeneity in the returns to schooling and as an illustration of the use of the estimation and inference methods presented in this paper, we use the data and methodology employed in Angrist and Krueger (1991) to estimate the QTE of schooling on earnings. In particular, we estimate linear conditional quantile models of the form

$$Q_{\ln(Y_s)|X}(\tau) = \alpha(\tau)S + X'\beta(\tau),$$

where Y is the weekly wage, S is reported years of schooling and X is a vector of covariates consisting of state and year of birth fixed effects, using quarter of birth as an instrument for education.[24]

The use of quarter of birth as an instrument is motivated by the fact that quarter of birth is correlated to years of schooling through compulsory schooling laws. These laws prohibit students from dropping out of school before reaching a certain age, but in general do not stipulate minimum education levels. However, most school districts do not admit students to the first grade unless they will be six years old by January 1 of the academic year. This means that individuals born earlier in the year reach the minimum drop out age after having attended less school than those born later in the year. Angrist and Krueger (1991), examining

[22] The uniform consistency of $\widehat{\Omega}^*(\tau)$ and $\widehat{J}(\tau)$ can be shown using the uniform laws of large numbers, Theorem 19.3 and Theorem 19.28 in van der Vaart (1998), respectively.

[23] See, for example, Card (1995), Card (1999), and Carneiro et al. (2000) for discussion and Becker and Chiswick (1966), Mincer (1970), Mincer (1974), and Mincer (1995) for early examples.

[24] Specifically, we use the linear projection of S onto the covariates X and three dummies for first through third quarter of birth as the instrumental variable.

data from three decennial censuses, find that people born in the first quarter of the year do indeed have less schooling on average than those born later in the year. Based on this observation, Angrist and Krueger (1991) use quarter of birth as an instrument for years of completed schooling in an attempt to isolate the causal impact of schooling on earnings.[25]

We focus on the specification used in Angrist and Krueger (1991) which includes state of birth effects, year of birth effects, and a constant in the covariate vector.[26] The sample we consider consists of 329,509 males from the 1980 U.S. Census who were born between 1930 and 1939 and have data on weekly wages, years of completed education, state of birth, year of birth, and quarter of birth. The sample was selected using the criteria described in Appendix 1 of Angrist and Krueger (1991).

IV-QR and QR estimates of the schooling coefficient are provided in Figure 1. The shaded region in each panel represents the 95% confidence interval. Both the quantile and instrumental variables quantile regression estimates suggest that the "returns to schooling" vary over the earnings distribution. The second row in Table 1 reports the results from a test of the hypothesis of a constant QTE for the IV-QR estimates which is rejected at the 10% level. The variability of QTE is most apparent in the IV-QR estimates. While the QR estimates do vary statistically,[27] they are all closely clustered around the OLS estimate. The practical lack of variability in the QR estimates is clearly demonstrated in the first panel of Figure 1, which plots both the IV-QR (solid line) and QR (dashed line) estimates. Relative to the IV-QR estimates, the QR estimates appear to be approximately constant.

[Insert Figure 1 about here]

The shapes of the estimated QTE are very interesting. The QR estimates exhibit a distinct u-shape, implying higher returns to schooling for those in the tails of the distribution than for those in the middle. However, if schooling is endogenous to the earnings equation, these estimates do not consistently estimate the true (causal) QTE. IV-QR estimates, on the other hand, are consistent for the QTE under endogeneity and show quite different results than those obtained through standard QR. In particular, the IV-QR results show returns to schooling, as measured by QTE, of approximately 30% per year of additional schooling at low quantiles in the earnings distribution. The returns decrease as the quantile index increases toward the middle of the distribution and then remain approximately constant at levels near the QR and OLS estimates. This implies that the largest gains to additional years of schooling accrue to those at the low end of the earnings distribution. This observation is consistent with the notion that people with high unobserved "ability", as measured by the quantile index τ, will generate high earnings regardless of their education level, while those with lower "ability" gain more from the training provided by formal education.[28] Interpreting the quantile

[25] Angrist and Krueger (1991) also provide evidence that quarter of birth is independent of unobserved taste or ability factors which may affect earnings, which is necessary for quarter of birth to be a valid instrument. For a differing viewpoint, see Bound and Jaeger (1996), who argue that quarter of birth should not be treated as exogenous.

[26] Note that the estimates of the schooling coefficient are not sensitive to the specification of the X vector.

[27] The test for the quantile regression estimates is not reported, but also rejects the null hypothesis of a constant treatment effect.

[28] The term "ability" is used to characterize the unobserved component of earnings, which likely captures elements

Table 1: Process Tests for the Earning Equation. Subsample size = $5n^{2/5}$.

Null Hypothesis	Kolmogorov-Smirnov Statistic	90% Critical Value	95% Critical Value
No Effect. $\alpha(\cdot) = 0$	4.563	2.572	2.935
Constant Effect. $\alpha(\cdot) = \alpha$	2.630	2.442	2.658
Dominance $\alpha(\cdot) \geq 0$	0.000	2.185	2.549
Exogeneity $\alpha(\cdot) = \alpha_{QR}(\cdot)$	2.510	2.465	2.721

index τ as indexing ability, these results are also consistent with a simple model in which individuals acquire education up to the point where the cost equals the rate of return and cost depends negatively on ability.[29] In this case, we would expect the returns to schooling to be decreasing in ability with the lowest ability individuals having the highest returns to education, which is exactly the pattern demonstrated by the IV-QR results.

The first row of Table 1 reports the results from testing that schooling has no causal effect on earnings, while the third row reports the results from the test of stochastic dominance. As would be expected, the tests strongly reject the hypothesis of no effect and fail to reject the null hypothesis of stochastic dominance, confirming our intuition that schooling increases earnings across the distribution. In the final row of Table 1, we test the endogeneity hypothesis. The test rejects the null hypothesis of no endogeneity at the 10% level, providing some evidence on the need to instrument for schooling in the earnings equation. Again, this confirms our intuition that endogeneity contaminates standard estimates of the returns to schooling and underscores the importance of accounting for this endogeneity in estimation.

Overall, the estimation and testing results indicate that the causal effect of schooling on earnings is quite heterogeneous, with the largest returns accruing to those who fall in the lower tail of the earnings distribution. The example also illustrates the variety of interesting distributional hypotheses that can be tested using the methods developed in this paper. The IV-QR results demonstrate considerable heterogeneity in QTEs and provide additional insight into the economic relationships involved which could not be gained by focusing on a single feature of the outcome distribution.

6 Conclusion

In this paper, we described how instrumental variable quantile regression can be used to evaluate the impact of endogenous variables (treatments) on the entire distribution of economic outcomes when the variables are self-selected or selected in relation to potential outcomes. We introduced an instrumental variable quantile regression process and the set of inferences derived from it, focusing on tests of distributional equality, constancy of effects, conditional dominance, and exogeneity. The approach was illustrated through estimation of the returns to schooling. In this example, the hypotheses of constant returns to schooling and exogeneity were rejected. The results suggest that estimates of structural (treatment) effects that focus on

of ability and motivation as well as noise.

[29] See, for example, Card (1999).

a single feature of the outcome distribution may fail to capture the full impact of the treatment and serve to illustrate the variety of distributional hypotheses that can be examined based on the instrumental quantile regression process.

We believe that the results and inference methods presented in this paper will be useful in many economic problems. Indeed, there are now several papers which examine the distributional impacts of economic variables using the model, identification, and estimation results provided in this paper. Hausman and Sidak (2002) consider long-distance price discrimination models with varying coefficients. Januszewski (2002) studies the impact of air traffic delays on airline ticket prices. D'Urso (2002) uses instrumental quantile regression methods to estimate the effect of the internet on home buyer search duration. Chernozhukov and Hansen (2003) estimate the distributional impact of 401(k) participation on assets.

A Proofs

We use the notation for empirical processes following van der Vaart and Wellner (1996). For $W \equiv (Y, D, X, Z)$

$$f \mapsto \mathbb{E}_n f(W) \equiv \frac{1}{n}\sum_{i=1}^{n} f(W_i), \quad f \mapsto \mathbb{G}_n f(W) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \left(f(W_i) - Ef(W_i)\right).$$

If $\widehat{f}$ is an estimated function, $\mathbb{G}_n f(W)$ denotes $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(f(W_i) - Ef(W_i)\right)_{f=\widehat{f}}$.

A.1 Proof of Theorem 1

See Chernozhukov and Hansen (2001a). □.

A.2 Proof of Lemma 1

The result immediately follows from the proof of Theorem 2 in Mas-Colell (1979a). □.

A.3 Proof of Theorem 2

See Step 1 in the Proof of Theorem 3. □.

A.4 Proof of Theorem 3

Define for $\vartheta \equiv (\beta, \gamma)$ and $\varphi_\tau(u) \equiv (1(u < 0) - \tau)$

$$\widehat{f}(W, \alpha, \vartheta, \tau) \equiv \varphi_\tau(Y - D'\alpha - X'\beta - \widehat{\Phi}(\tau)'\gamma)\widehat{\Psi}(\tau),$$
$$f(W, \alpha, \vartheta, \tau) \equiv \varphi_\tau(Y - D'\alpha - X'\beta - \Phi(\tau)'\gamma)\Psi(\tau),$$

$\Psi(\tau) \equiv V(\tau)\cdot(\Phi(\tau)', X')'$, $\Phi(\tau) \equiv \Phi(\tau, X, Z)$, $V(\tau) \equiv V(\tau, X, Z)$, $\widehat{\Psi}(\tau) \equiv \widehat{V}(\tau)\cdot(\widehat{\Phi}(\tau)', X')'$, $\widehat{\Phi}(\tau) \equiv \widehat{\Phi}(\tau, X, Z)$, $\widehat{V}(\tau) \equiv \widehat{V}(\tau, X, Z)$; for $\rho_\tau(u) \equiv (\tau - 1(u < 0))u$

$$\widehat{g}(W, \alpha, \vartheta, \tau) \equiv \rho_\tau(Y - D'\alpha - X'\beta - \widehat{\Phi}(\tau)'\gamma)\widehat{V}(\tau),$$
$$g(W, \alpha, \vartheta, \tau) \equiv \rho_\tau(Y - D'\alpha - X'\beta - \Phi(\tau)'\gamma)V(\tau).$$

Define

$$Q_n(\alpha, \vartheta, \tau) \equiv \mathbb{E}_n\widehat{g}(W, \alpha, \vartheta, \tau), \quad Q(\alpha, \vartheta, \tau) \equiv Eg(W, \alpha, \vartheta, \tau),$$

and

$$\widehat{\vartheta}(\alpha,\tau) \equiv (\widehat{\beta}(\alpha,\tau)', \widehat{\gamma}(\alpha,\tau)') \equiv \quad \arg \inf_{\vartheta \in \mathcal{B}\times\mathcal{G}} Q_n(\alpha,\vartheta,\tau),$$

$$\vartheta(\alpha,\tau) \equiv (\beta(\alpha,\tau)', \gamma(\alpha,\tau)') \equiv \quad \arg \inf_{\vartheta \in \mathcal{B}\times\mathcal{G}} Q\ (\alpha,\vartheta,\tau),$$

$$\widehat{\alpha}(\tau) \ \equiv \arg \inf_{\alpha\in\mathcal{A}} \|\widehat{\gamma}(\alpha,\tau)\|, \quad \alpha^*(\tau) \equiv \arg \inf_{\alpha\in\mathcal{A}} \|\gamma(\alpha,\tau)\|,$$

$$\widehat{\vartheta}(\tau) \equiv (\widehat{\beta}(\tau)', \widehat{\gamma}(\tau)')' \equiv \widehat{\vartheta}(\widehat{\alpha}(\tau),\tau),$$

$$\vartheta(\tau) \equiv (\beta(\tau)', 0')' \equiv \vartheta(\alpha(\tau),\tau),$$

Step 1 (Identification) We show that $(\alpha(\tau)', \beta(\tau)')$ uniquely solves the limit problem for each τ, that is $\alpha^*(\tau) = \alpha(\tau)$ and $\beta(\alpha^*(\tau),\tau) = \beta(\tau)$. Define

$$\Pi(\alpha,\beta,\tau) \equiv E\left[\varphi_\tau(Y - D'\alpha - X'\beta)\Psi(\tau)\right],$$

$$J(\alpha,\beta,\tau) \equiv \frac{\partial}{\partial(\alpha',\beta')} E\left[\varphi_\tau(Y - D'\alpha - X'\beta)\Psi(\tau)\right].$$

By R3 $J(\alpha,\beta,\tau)$ has full rank and is continuous in (α,β), uniformly over $\mathcal{A}\times\mathcal{B}$. Moreover, the image of $\mathcal{A}\times\mathcal{B}$ under the mapping $(\alpha,\beta) \mapsto \Pi(\alpha,\beta,\tau)$ is assumed to be simply connected. As in Chernozhukov and Hansen (2001a), the application of Hadamard's global univalence theorem for general metric spaces, e.g. Theorem 1.8 in Ambrosetti and Prodi (1995), yields[30] that the mapping $\Pi(\cdot,\cdot,\tau)$ is a homomorphism (one-to-one) between $(\mathcal{A}\times\mathcal{B})$ and $\Pi(\mathcal{A},\mathcal{B},\tau)$, the image of $\mathcal{A}\times\mathcal{B}$ under $\Pi(\cdot,\cdot,\tau)$. By Theorem 1, $(\alpha,\beta) = (\alpha(\tau)', \beta(\tau)')'$ solves the equation $\Pi(\alpha,\beta,\tau) = 0$; and it is thus the only solution in $(\mathcal{A}\times\mathcal{B})$. This argument applies for every $\tau \in \mathcal{T}$.

So we have that the true parameters $(\alpha,\beta) = (\alpha(\tau),\beta(\tau))$ uniquely solve the equation

$$E[\varphi_\tau(Y - D'\alpha - X'\beta - \Phi(\tau)'0)\Psi(\tau)] = 0. \tag{A.1}$$

By R3 and in view of the global convexity of $Q(\alpha,\vartheta,\tau)$ in ϑ for each τ and α, $\vartheta(\alpha,\tau)$ is defined by the subgradient condition

$$E[\varphi_\tau(Y - D'\alpha - X'\beta(\alpha,\tau) - \Phi(\tau)'\gamma(\alpha,\tau))\Psi(\tau)]'\nu \geq 0 \text{ for all } \nu : \vartheta(\alpha,\tau) + \nu \in \mathcal{B}\times\mathcal{G}. \tag{A.2}$$

In fact, if $\vartheta(\alpha,\tau)$ is in the interior of $\mathcal{B}\times\mathcal{G}$, it uniquely solves the first order condition version of (A.2):

$$E[\varphi_\tau(Y - D'\alpha - X'\beta(\alpha,\tau) - \Phi(\tau)'\gamma(\alpha,\tau))\Psi(\tau)] = 0. \tag{A.3}$$

We need to find $\alpha^*(\tau)$ by minimizing $\|\gamma(\alpha,\tau)\|$ over α subject to (A.2) holding. By (A.1) $\alpha^*(\tau) = \alpha(\tau)$ makes $\|\gamma(\alpha^*(\tau),\tau)\| = 0$ and satisfies (A.3) and hence (A.2) at the same time. By the preceding paragraph, it is the only such solution. Thus, also by (A.3) $\beta(\alpha^*(\tau),\tau) = \beta(\tau)$.

Step 2 (Consistency) By the bounded density condition in R3, $Q(\alpha,\vartheta,\tau)$ is continuous over $\mathcal{A}\times(\mathcal{B}\times\mathcal{G})\times\mathcal{T}$; and by Lemma 3, $\sup_{(\alpha,\vartheta,\tau)\in\mathcal{A}\times(\mathcal{B}\times\mathcal{G})\times\mathcal{T}} \|Q_n(\alpha,\vartheta,\tau) - Q(\alpha,\vartheta,\tau)\| \to_p 0$. This implies by Lemma 2 the uniform convergence $\sup_{(\alpha,\tau)\in\mathcal{A}\times\mathcal{T}} \|\widehat{\vartheta}(\alpha,\tau) - \vartheta(\alpha,\tau)\| \to_p 0$, (∗), which in turn implies $\sup_{(\alpha,\tau)\in\mathcal{A}\times\mathcal{T}} \|\|\widehat{\gamma}(\alpha,\tau)\|_{A(\tau)} - \|\gamma(\alpha,\tau)\|_{A(\tau)}\| \to_p 0$, which by invoking Lemma 2 again implies $\sup_{\tau\in\mathcal{T}} \|\widehat{\alpha}(\tau) - \alpha(\tau)\| \to_p 0$, which by (∗) implies $\sup_{\tau\in\mathcal{T}} \|\widehat{\beta}(\tau) - \beta(\tau)\| \to_p 0$ and $\sup_{\tau\in\mathcal{T}} \|\widehat{\gamma}(\widehat{\alpha}(\tau),\tau) - 0\| \to_p 0$. (Note that by the implicit function theorem $\vartheta(\alpha,\tau)$ is continuous in τ and α, and $\alpha(\tau)$ is continuous in τ.)

Step 3 (Asymptotics) Consider a collection of closed balls $B_{\delta_n}(\alpha(\tau))$ centered at $\alpha(\tau)$ for each τ, where balls' radius δ_n is independent of τ and $\delta_n \to 0$ slowly enough. Let $\alpha_n(\tau)$ denote any value inside $B_{\delta_n}(\alpha(\tau))$. By the computational properties of the ordinary quantile regression estimator $\widehat{\vartheta}(\alpha_n(\tau),\tau)$, cf. Theorem 3.3 in Koenker and Bassett (1978),

$$O(1/\sqrt{n}) = \sqrt{n}\mathbb{E}_n \widehat{f}(W, \alpha_n(\cdot), \widehat{\vartheta}(\alpha_n(\cdot),\cdot),\cdot). \tag{A.4}$$

[30] Chernozhukov and Hansen (2001a) apply the theorem in the nonparametric context. Other use of this theorem in economic analysis includes Mas-Colell (1979b). Original references are Hadamard (1906) and Caccioppoli (1932).

By Lemma 3, the following expansion of the rhs is valid for any $\sup_{\tau\in\mathcal{T}}\|\alpha_n(\tau)-\alpha(\tau)\|\to_p 0$:[31]

$$\begin{aligned} O(1/\sqrt{n}) &= \sqrt{n}\mathbb{E}_n\widehat{f}(W,\alpha_n(\cdot),\widehat{\vartheta}(\alpha_n(\cdot),\cdot),\cdot) = \mathbb{G}_n\widehat{f}(W,\alpha_n(\cdot),\widehat{\vartheta}(\alpha_n(\cdot),\cdot),\cdot) + \sqrt{n}E\widehat{f}(W,\alpha_n(\cdot),\widehat{\vartheta}_n(\alpha_n(\cdot),\cdot),\cdot) \\ &= \mathbb{G}_n f(W,\alpha(\cdot),\vartheta(\alpha(\cdot),\cdot),\cdot) + o_p(1) + \sqrt{n}E\widehat{f}(W,\alpha_n(\cdot),\widehat{\vartheta}(\alpha_n(\cdot),\cdot),\cdot) \text{ in } \ell^\infty(\mathcal{T}). \end{aligned} \tag{A.5}$$

Expanding the last line further

$$\begin{aligned} O(1/\sqrt{n}) = \mathbb{G}_n f(W,\alpha(\cdot),\vartheta(\cdot),\cdot) + o_p(1) + (J_\vartheta(\cdot)+o_p(1))\sqrt{n}(\widehat{\vartheta}(\alpha_n(\cdot),\cdot)-\vartheta(\cdot)) \\ + (J_\alpha(\cdot)+o_p(1))\sqrt{n}(\alpha_n(\cdot)-\alpha(\cdot)),\ \text{in } \ell^\infty(\mathcal{T}), \end{aligned} \tag{A.6}$$

where

$$\begin{aligned} J_\vartheta(\cdot) &= \frac{\partial}{\partial(\beta',\gamma')}E\left[\varphi_\cdot(Y-D'\alpha(\cdot)-X'\beta-\Phi(\cdot)'\gamma)\Psi(\cdot)\right]_{(\gamma,\beta)=(0,\beta(\cdot))}, \\ J_\alpha(\cdot) &= \frac{\partial}{\partial(\alpha')}E\left[\varphi_\cdot(Y-D'\alpha-X'\beta(\cdot))\Psi(\cdot)\right]_{\alpha=\alpha(\cdot)}. \end{aligned}$$

In other words for any $\sup_{\tau\in\mathcal{T}}\|\alpha_n(\tau)-\alpha(\tau)\|\to_p 0$,

$$\sqrt{n}(\widehat{\vartheta}(\alpha_n(\cdot),\cdot)-\vartheta(\cdot)) = -J_\vartheta^{-1}(\cdot)\mathbb{G}_n f(W,\alpha(\cdot),\vartheta(\cdot),\cdot) - J_\vartheta^{-1}(\cdot)J_\alpha(\cdot)[1+o_p(1)]\sqrt{n}(\alpha_n(\cdot)-\alpha(\cdot)) + o_p(1) \text{ in } \ell^\infty(\mathcal{T}),$$

i.e

$$\sqrt{n}(\widehat{\gamma}(\alpha_n(\cdot),\cdot)-0) = -\bar{J}_\gamma(\cdot)\mathbb{G}_n f(W,\alpha(\cdot),\vartheta(\cdot),\cdot) - \bar{J}_\gamma(\cdot)J_\alpha(\cdot)[1+o_p(1)]\sqrt{n}(\alpha_n(\cdot)-\alpha(\cdot)) + o_p(1) \text{ in } \ell^\infty(\mathcal{T}),$$

where

$$[\bar{J}_\beta(\cdot)' : \bar{J}_\gamma(\cdot)']' \text{ is the conformable partition of } J_\vartheta^{-1}(\cdot).$$

By Step 2 wp $\to 1$

$$\widehat{\alpha}(\tau) = \arg\inf_{\alpha_n(\tau)\in B_n(\alpha(\tau))}\|\widehat{\gamma}(\alpha_n(\tau),\tau)\|_{A(\tau)} \text{ for all } \tau\in\mathcal{T}.$$

By Lemma 3, $\mathbb{G}_n f(W,\alpha(\cdot),\vartheta(\cdot),\cdot) = O_p(1)$, thus

$$\sqrt{n}\|\widehat{\gamma}(\alpha_n(\cdot),\cdot)\|_{A(\cdot)} = \|O_p(1) - \bar{J}_\gamma(\cdot)J_\alpha(\cdot)[1+o_p(1)]\sqrt{n}(\alpha_n(\cdot)-\alpha(\cdot))\|_{A(\cdot)} \text{ in } \ell^\infty(\mathcal{T}).$$

Since $\bar{J}_\gamma(\tau)J_\alpha(\tau)$ and $A(\tau)$ have full rank uniformly in τ, [32] $\sqrt{n}(\widehat{\alpha}(\cdot)-\alpha(\cdot)) = O_p(1)$ in $\ell^\infty(\mathcal{T})$. Hence, using arguments similar to those in the proof of Lemma 2,

$$\sqrt{n}(\widehat{\alpha}(\cdot)-\alpha(\cdot)) = \arg\inf_{\mu\in\mathbb{R}^l}\| - \bar{J}_\gamma(\cdot)\mathbb{G}_n f(W,\alpha(\cdot),\vartheta(\cdot),\cdot) - \bar{J}_\gamma(\cdot)J_\alpha(\cdot)\mu\|_{A(\cdot)} + o_p(1) \text{ in } \ell^\infty(\mathcal{T}),$$

Conclude that in $\ell^\infty(\mathcal{T})$ jointly

$$\sqrt{n}(\widehat{\alpha}(\cdot)-\alpha(\cdot)) = -\Big(J_\alpha(\cdot)'\bar{J}_\gamma(\cdot)'A(\cdot)\bar{J}_\gamma(\cdot)J_\alpha(\cdot)\Big)^{-1}\Big(J_\alpha(\cdot)'\bar{J}_\gamma(\cdot)'A(\cdot)\bar{J}_\gamma(\cdot)\Big)\mathbb{G}_n f(W,\alpha(\cdot),\vartheta(\cdot),\cdot) + o_p(1) = O_p(1),$$

$$\begin{aligned} \sqrt{n}(\widehat{\vartheta}(\widehat{\alpha}(\cdot),\cdot)-\vartheta(\cdot)) = -J_\vartheta^{-1}(\cdot)\Big[I - J_\alpha(\cdot)\Big(J_\alpha(\cdot)'\bar{J}_\gamma(\cdot)'A(\cdot)\bar{J}_\gamma(\cdot)J_\alpha(\cdot)\Big)^{-1}J_\alpha(\cdot)'\bar{J}_\gamma(\cdot)'A(\cdot)\bar{J}_\gamma(\cdot)\Big] \\ \times\mathbb{G}_n f(W,\alpha(\cdot),\vartheta(\cdot),\cdot) + o_p(1) = O_p(1). \end{aligned}$$

Due to invertibility of $J_\alpha(\tau)\bar{J}_\gamma(\tau)$

$$\begin{aligned} \sqrt{n}\left(\widehat{\gamma}(\widehat{\alpha}(\cdot),\cdot)-0\right) &= \Big[-\bar{J}_\gamma(\cdot)\Big[I - J_\alpha(\cdot)\Big(J_\alpha(\cdot)'\bar{J}_\gamma(\cdot)'\Big)^{-1}\bar{J}_\gamma(\cdot)\Big]\mathbb{G}_n f(W,\alpha(\cdot),\vartheta(\cdot),\cdot) + o_p(1) \\ &= 0\times O_p(1) + o_p(1) \text{ in } \ell^\infty(\mathcal{T}). \end{aligned}$$

[31] Note that by convention in empirical process theory $E\widehat{f}(W)$ means $(Ef(W))_{f=\widehat{f}}$.

[32] Indeed, $B(\tau) = \frac{\partial\Pi(\theta,\tau)}{\partial(\beta',\alpha')}^{-1}$ and $G(\tau) = \frac{\partial\Pi(\pi,\tau)}{\partial(\beta',\gamma')}^{-1}$, where the derivatives are evaluated at the true parameter values, exist by the full rank assumption R3. Then using the partitioned inverse formula, note that $\bar{J}_\gamma(\tau)J_\alpha(\tau) = G_{22}(\tau)(B_{22}(\tau))^{-1}$, where $G_{22}(\tau)$ and $B_{22}(\tau)$ are $l\times l$ lower-right blocks of $G(\tau)$ and $B(\tau)$. These blocks are invertible by invertibility of $G(\tau)$ and $B(\tau)$, so $\bar{J}_\gamma(\tau)J_\alpha(\tau)$ is also invertible.

Instead of working out the algebra to see a drastic simplification, using this fact and putting $\left(\alpha_n(\cdot), \widehat{\vartheta}(\alpha_n(\cdot), \cdot)\right) = \left(\widehat{\alpha}(\cdot), \widehat{\vartheta}(\cdot)\right) = \left(\widehat{\alpha}(\cdot), \widehat{\beta}(\cdot), 0 + o_p(1/\sqrt{n})\right)$ back into the expansion (A.6) we have

$$-\mathbb{G}_n f(W, \alpha(\cdot), \vartheta(\cdot), \cdot) = J(\cdot)\sqrt{n} \left(\begin{array}{c} \widehat{\alpha}(\cdot) - \alpha(\cdot) \\ \widehat{\beta}(\cdot) - \beta(\cdot) \end{array} \right) + o_p(1) \text{ in } \ell^\infty(\mathcal{T}) \ .$$

Next by Lemma 3, $\mathbb{G}_n f(W, \alpha(\cdot), \vartheta(\cdot), \cdot) \Rightarrow \mathbb{G}(\cdot)$ in $\ell^\infty(\mathcal{T})$, where $\mathbb{G}(\cdot)$ is the Gaussian process with covariance function $S(\tau, \tau') = (\min(\tau, \tau') - \tau\tau')E\Psi(\tau)\Psi(\tau)'$, which yields the desired conclusion

$$\sqrt{n} \left(\begin{array}{c} \widehat{\alpha}(\cdot) - \alpha(\cdot) \\ \widehat{\beta}(\cdot) - \beta(\cdot) \end{array} \right) \Rightarrow [J(\cdot)]^{-1} \mathbb{G}(\cdot) \text{ in } \ell^\infty(\mathcal{T}). \quad \square$$

A.5 Proof of Theorem 4

Part 1 follows by the Continuous Mapping Theorem. We also need that the distribution function of the limit statistics is continuous on $(0, \infty)$. This follows by Theorem 11.1 in Davydov et al. (1998): The distribution of functionals $f(v(\cdot))$ where f is of the specified sort, is absolutely continuous on $(0, \infty)$ once $v(\cdot)$ has a nondegenerate covariance kernel. Part 2 follows by observing that $f(\sqrt{n}g(\cdot)) \to_p \infty$ implies $f(\sqrt{n}g(\cdot) + G_n(\cdot)) \to_p \infty$, for any tight element $G_n(\cdot) = O_p(1)$ in $\ell^\infty(\mathcal{T})$, once the null is violated (once the composite null is violated for one-sided tests). $\square$

A.6 Proof of Proposition 1

The result is immediate from assumptions. $\square$

A.7 Proof of Theorem 5

We present the proof for the case where all of subsamples of size b constructed by sampling without replacement are used. In practice, a smaller number B_n of randomly chosen subsets can also be used, if $B_n \to \infty$ as $n \to \infty$. The argument extends to the randomly chosen subsets as in Section 2.5 in Politis et al. (1999). To simplify the presentation, we assume that $\Lambda(\tau), J(\tau), H(\tau)$ are known. However, the case with estimated matrices is a straightforward extension which follows, for example, using the arguments in Chernozhukov (2002) in Step II of this proof. Part 1 is shown in Steps I and II, and Parts 2 and 3 are shown in Step III.

Step I. By assumption, wp $\to 1$ realizations of function $\tau \mapsto \widehat{z}(W, \tau)$ belong to a Donsker set of functions denoted as $\{\xi(W, \tau), \tau \in \mathcal{T}, \xi \in \Xi\}$. Consider the empirical process $(\tau, \xi) \mapsto \mathbb{G}_n(\xi(\tau))$, which is Donsker in $\ell^\infty(\mathcal{T} \times \Xi)$ by assumption with the limit law denoted J. Consider also its subsample realizations $(\tau, \xi) \mapsto \mathbb{G}_{j,b,n}(\xi(\tau)) \equiv \frac{1}{\sqrt{b}} \sum_{i \in I_j} \left(\xi(W_i, \tau) - E\xi(W_i, \tau)\right), j = 1, ..., B_n$. Let J_n denote the sampling (outer) law of $(\tau, \xi) \mapsto \mathbb{G}_n(\xi(\tau))$ in $\ell^\infty(\mathcal{T} \times \Xi)$, and let $L_{b,n}$ denote the subsampling law of $(\tau, \xi) \mapsto \mathbb{G}_{j,b,n}(\xi(\tau))$ in $\ell^\infty(\mathcal{T} \times \Xi)$. By Theorem 7.4.1 in Politis et al. (1999)

$$\rho_{BL}\left(L_{b,n}, J_n\right) \to_p 0 \text{ and } \rho_{BL}\left(L_{b,n}, J\right) \to_p 0, \tag{A.7}$$

where ρ_{BL} denotes the Bounded-Lipschitz metric that metrizes weak convergence. See Politis et al. (1999) page 160 for definition.

For f denoting the KS and CM functionals on the empirical processes, let $J_n(\xi)$ denote the (outer) law of $f\left[\mathbb{G}_n(\xi(\cdot))\right]$, let $L_{b,n}(\xi)$ denote the subsampling (outer) law of $f\left[\mathbb{G}_{j,b,n}(\xi(\cdot))\right]$, and let $J(\xi)$ denote the limit law of $f\left[\mathbb{G}_n(\xi(\cdot))\right]$, which exists by the Continuous Mapping Theorem. By (A.7) and definition of ρ_{BL} we have that $\sup_{\xi \in \Xi}\left[\rho_{BL}\left(L_{b,n}(\xi), J_n(\xi)\right)\right] \to_p 0$ and $\sup_{\xi \in \Xi}\left[\rho_{BL}\left(L_{b,n}(\xi), J(\xi)\right)\right] \to_p 0$. This follows because the transformation $f[\mathbb{G}_n(\xi(\cdot))]$ is a uniform Holder-continuous functional of the mapping $(\xi, \tau) \mapsto \mathbb{G}_n(\xi(\tau))$. This immediately gives us that

$$\left[\rho_{BL}\left(L_{b,n}(\xi), J_n(\xi)\right)\right]_{\xi = \hat{z}} \to_p 0 \text{ and hence } \left[\rho_{BL}\left(L_{b,n}(\widehat{z}), J(z)\right)\right] \to_p 0, \tag{A.8}$$

provided we have the asymptotic continuity condition: $\rho_{BL}(J_n(z_n), J(z)) \to 0$ for any sequence $z_n \in \Xi$ such that $\sup_{\tau\in\mathcal{T}}[E\|z_n(\tau) - z(\tau)\|^2] \to 0$. (Recall that by I.4 $\sup_{\tau\in\mathcal{T}}[E\|z_n(\tau) - z(\tau)\|^2]_{z_n=\hat{z}} \to_p 0$). But this asymptotic continuity property is immediate from the stochastic equicontinuity of the process $(\tau, \xi) \mapsto \mathbb{G}_n(\xi(\tau))$ implied by the assumed Donskerness.

Let $H_{b,n}$ denote the subsampling distribution function of $f\left[\mathbb{G}_{j,b,n}(\hat{z}(\cdot))\right]$, and let Γ denote the limit distribution function of $f\left[\mathbb{G}_n(z(\cdot))\right]$. (A.8) is equivalent to pointwise convergence of $H_{b,n}$ to Γ at the continuity points x of $\Gamma(x)$, that is $H_{b,n}(x) \to_p \Gamma(x)$. The set of continuity points of Γ is $(0, \infty)$ as shown in the Proof of Theorem 2.

Step II. Now note that we actually need to show consistency of the subsampling distribution $\Gamma_{b,n}$ of $f\left[v_{j,b,n}(\cdot)\right]$, but difference between $\Gamma_{b,n}$ and $H_{b,n}$ is asymptotically negligible. Indeed,

$$f\left[\mathbb{G}_{j,b,n}(\hat{z}(\cdot))\right] - K_n \leq f\left[v_{j,b,n}(\cdot)\right] \leq f\left[\mathbb{G}_{j,b,n}(\hat{z}(\cdot))\right] + K_n,$$

where e.g. when f is the KS function, for some constants C_j, $j = 1, 2, 3$:

$$K_n = \sup_{\tau\in\mathcal{T}} \left\| \sqrt{b} \cdot [Ez(W, \tau)]_{z=\hat{z}} \right\|_{\Lambda(\tau)} \leq C_1\sqrt{b} \cdot \sup_{\tau\in\mathcal{T}} \left| C_2\|\widehat{\vartheta}(\tau) - \vartheta(\tau)\| + C_3\|\widehat{r}(\tau) - r(\tau)\| \right| = O_p\left(\frac{\sqrt{b}}{\sqrt{n}}\right) = o_p(1),$$

which follows by invoking I.4(b), I.4(c), and I.3. Thus for any $\delta > 0$ wp $\to 1$ $1(E_n) = 1$, where $E_n \equiv \{K_n \leq \delta\}$. Given the event E_n for a small $\epsilon > 0$ there is $\delta > 0$ such that $H_{b,n}(x - \epsilon)1(E_n) \leq \Gamma_{b,n}(x)1(E_n) \leq H_{b,n}(x + \epsilon)1(E_n)$. Hence it follows that wp $\to 1$: $H_{b,n}(x - \epsilon) \leq \Gamma_{b,n}(x) \leq H_{b,n}(x + \epsilon)$. We have by Step I $H_{b,n}(x + c) \to_p \Gamma(x + c)$, for $c = \epsilon$ and $c = -\epsilon$, which implies $\Gamma(x - \epsilon) - \epsilon \leq \Gamma_{b,n}(x) \leq \Gamma(x + \epsilon) + \epsilon$ wp $\to 1$. Since ϵ can be set as small as we like and Γ is continuous at points x of interest, this yields the conclusion $\Gamma_{b,n}(x) \to_p \Gamma(x)$.

Step III. Finally, convergence of quantiles is implied by the convergence of distribution functions at continuity points. Part 2 of Theorem 5 follows by steps that are identical to those in the proof of Part 1 (Steps I and II), except that we have convergence of the subsampling distribution $\Gamma_{b,n}$ to some other distribution $\Gamma' \neq \Gamma$ at the continuity points. Note that by tightness of Γ', $c_{b,n}(1 - \alpha) = O_p(1)$ even if Γ' is not continuous at $\Gamma'^{-1}(1 - \alpha)$. Part 3 of Theorem 5 has already been proven in the proof of Theorem 4. □

B Lemmas

LEMMA 2 (Argmax Process) *Suppose that uniformly in π in a compact set Π and for a compact set K (i) $Z_n(\pi)$ is s.t. $Q_n(Z_n(\pi)|\pi) \geq \sup_{z\in K} Q_n(z|\pi) - \epsilon_n$, $\epsilon_n \searrow 0$; $Z_n(\pi) \in K$ wp $\to 1$, (ii) $Z_\infty(\pi) \equiv \text{argsup}_{z\in K}\, Q_\infty(z|\pi)$ is a uniquely defined continuous process in $\ell^\infty(\Pi)$, (iii) $Q_n(\cdot|\cdot) \to_p Q_\infty(\cdot|\cdot)$ in $\ell^\infty(K \times \Pi)$, where $Q_\infty(\cdot|\cdot)$ is continuous. Then $Z_n(\cdot) = Z_\infty(\cdot) + o_p(1)$ in $\ell^\infty(\Pi)$.*

Proof. The argument is a simple extension of the standard consistency argument, cf. Amemiya (1985). We have $Q_n(z|\pi) = Q_\infty(z|\pi) + o_p(1)$ uniformly in $(z, \pi) \in K \times \Pi$. For any $c > 0$, wp $\to 1$ uniformly in $\epsilon \geq c$ and uniformly in $\pi \in \Pi$ we have that: [i] $Q_n(Z_n(\pi)|\pi) \geq Q_n(Z_\infty(\pi)|\pi) - \epsilon/3$ by definition, [ii] $Q_\infty(Z_n(\pi)|\pi) > Q_n(Z_n(\pi)|\pi) - \epsilon/3$ by the uniform convergence, [iii] $Q_n(Z_\infty(\pi)|\pi) > Q_\infty(Z_\infty(\pi)|\pi) - \epsilon/3$ by the uniform convergence. Hence wp $\to 1$ $Q_\infty(Z_n(\pi)|\pi) > Q_n(Z_n(\pi)|\pi) - \epsilon/3 \geq Q_n(Z_\infty(\pi)|\pi) - 2\epsilon/3 > Q_\infty(Z_\infty(\pi)|\pi) - \epsilon$. Pick any $\delta > 0$. Let $\{B_\delta(\pi), \pi \in \Pi\}$ be a collection of balls with diameter $\delta > 0$, each centered at $Z_\infty(\pi)$. Then $\epsilon \equiv \inf_{\pi\in\Pi}\left[Q_\infty(Z_\infty(\pi)|\pi) - \sup_{z\in K\setminus B_\delta(\pi)} Q_\infty(z)\right] > 0$ a.s. by assumption ii, and for any $\varepsilon > 0$ we can pick $c > 0$ so that for $P(\epsilon \geq c) > 1 - \varepsilon$. It now follows that with probability becoming greater than $1 - \varepsilon$, uniformly in π: $Q_\infty(Z_n(\pi)|\pi) > Q_\infty(Z_\infty(\pi)|\pi) - Q_\infty(Z_\infty(\pi)|\pi) + \sup_{z\in K\setminus B_\delta(\pi)} Q_\infty(z|\pi) = \sup_{z\in K\setminus B_\delta(\pi)} Q_\infty(z|\pi)$. Thus with probability becoming greater than $1 - \varepsilon$, $\sup_{\pi\in\Pi} \|Z_n(\pi) - Z_\infty(\pi)\| \leq \delta$. But ε is arbitrary, so $\sup_{\pi\in\Pi} \|Z_n(\pi) - Z_\infty(\pi)\| \leq \delta$ wp $\to 1$. □

LEMMA 3 (Stochastic Expansions) *Under assumption R1-R4, and using notation defined in the proof of Theorem 3, the following statements are true.*

I. $\sup_{(\alpha,\beta,\gamma,\tau)\in\mathcal{A}\times\mathcal{B}\times\mathcal{G}\times\mathcal{T}} |\mathbb{E}_n[\widehat{g}(W, \alpha, \beta, \gamma, \tau)] - E[g(W, \alpha, \beta, \gamma, \tau)]| = o_p(1)$.

II. $\mathbb{G}_n f(W, \alpha(\cdot), \beta(\cdot), 0, \cdot) \Rightarrow \mathbb{G}(\cdot)$ *in* $\ell^\infty(\mathcal{T})$*, where* $\mathbb{G}$ *is a Gaussian process with covariance function* $S(\tau, \tau')$ *defined in Theorem 2. Furthermore, for any* $\sup_{\tau\in\mathcal{T}} \|(\widehat{\alpha}(\tau), \widehat{\beta}(\tau), \widehat{\gamma}(\tau)) - (\alpha(\tau), \beta(\tau), 0)\| = o_p(1)$*, it is the case that* $\sup_{\tau\in\mathcal{T}} \|\mathbb{G}_n \widehat{f}(W, \widehat{\alpha}(\tau), \widehat{\beta}(\tau), \widehat{\gamma}(\tau), \tau) - \mathbb{G}_n f(W, \alpha(\tau), \beta(\tau), 0, \tau)\| = o_p(1)$.

Proof. We first show II. Denote $\pi = (\alpha, \beta, \gamma)$ and $\Pi = \mathcal{A} \times \mathcal{B} \times \mathcal{G}$ where $\mathcal{G}$ is a closed ball at 0. We first show that the class of functions

$$\mathcal{H} \equiv \Big\{ h = (\Phi, \Psi, \pi, \tau) \mapsto \varphi_\tau(Y - D'\alpha - X'\beta - \Phi(X, Z)'\gamma)\Psi(X, Z), \quad \pi \in \Pi, \Psi \in \mathcal{F}, \Phi \in \mathcal{F} \Big\}$$

is Donsker, where $\mathcal{F}$ is defined in R4. The bracketing number of $\mathcal{F}$ by Cor 2.7.4 in van der Vaart and Wellner (1996) satisfies

$$\log N_{[\cdot]}(\epsilon, \mathcal{F}, L_2(P)) = O(\epsilon^{-\frac{\dim(z,x)}{\eta}}) = O(\epsilon^{-2-\delta'}),$$

for some $\delta' < 0$. Thus $\mathcal{F}$ is Donsker with a constant envelope. By Cor 2.7.4 in van der Vaart and Wellner (1996) the bracketing number of

$$\mathcal{X} \equiv \Big\{ (\Phi, \pi) \mapsto (D'\alpha - X'\beta - \Phi(X, Z)'\gamma), \quad \pi \in \Pi, \Phi \in \mathcal{F} \Big\}$$

satisfies

$$\log N_{[\cdot]}(\epsilon, \mathcal{X}, L_2(P)) = O(\epsilon^{-\frac{\dim(z,d,x)}{\eta}}) = O(\epsilon^{-2-\delta''})$$

for some $\delta'' < 0$. Using the idea contained in Remark 4 in Chernozhukov and Hong (2002), i.e. exploiting the monotonicity and boundedness of indicator function and bounded density condition assumed in R3, it is immediate that the bracketing number of

$$\mathcal{V} \equiv \Big\{ (\Phi, \pi) \mapsto 1(Y < D'\alpha + X'\beta + \Phi(X, Z)'\gamma), \quad \pi \in \Pi, \Phi \in \mathcal{F} \Big\}$$

satisfies

$$\log N_{[\cdot]}(\epsilon, \mathcal{V}, L_2(P)) = O(\epsilon^{-2-\delta''})$$

as well. Since $\mathcal{V}$ also has constant envelope by R1 and R4, it is Donsker. Class $\mathcal{H}$ is formed by taking products and sums of bounded Donsker classes $\mathcal{F}$, $\mathcal{V}$, and $\mathcal{T} \equiv \{\tau \mapsto \tau\}$, i.e. $\mathcal{H} \equiv \mathcal{T} \cdot \mathcal{F} - \mathcal{V} \cdot \mathcal{F}$, which is Lipschitz over $(\mathcal{T} \times \mathcal{F} \times \mathcal{V})$. Hence by Theorem 2.10.6 in van der Vaart and Wellner (1996) $\mathcal{H}$ is Donsker.

Now we show claim II using the established Donskerness. Define the process

$$h = (\Phi, \Psi, \pi, \tau) \mapsto \mathbb{G}_n \varphi_\tau(Y - D'\alpha - X'\beta - \Phi(X, Z)'\gamma)\Psi(X, Z). \tag{B.1}$$

This process is Donsker (asymptotically Gaussian) in $\ell^\infty(\mathcal{H})$. Therefore the process $\tau \mapsto \mathbb{G}_n \varphi_\tau(Y - D'\alpha(\tau) - X'\beta(\tau))\Psi(\tau, X, Z)$ is also Donsker in $\ell^\infty(\mathcal{T})$ by the uniform Holder continuity of $\tau \mapsto (\tau, \alpha(\tau)', \beta(\tau)', \Phi(\tau, X, Z)', \Psi(\tau, X, Z)')'$ in τ with respect to the supremum norm, as imposed by R3 and R4.[33] Thus we have $\mathbb{G}_n \varphi_\cdot(Y - D'\alpha(\cdot) - X'\beta(\cdot))\Psi(\cdot, X, Z) \Rightarrow \mathbb{G}(\cdot)$. $\mathbb{G}(\cdot)$ has covariance function $S(\tau, \tau') = E[\mathbb{G}(\tau)\mathbb{G}(\tau')'] = E[\varphi_\tau(Y - D'\alpha(\tau) - X'\beta(\tau))\Psi(\tau)\varphi_{\tau'}(Y - D'\alpha(\tau') - X'\beta(\tau'))\Psi(\tau')'] = E[(\tau - 1(\epsilon(\tau) \le 0))(\tau' - 1(\epsilon(\tau') \le 0))\Psi(\tau)\Psi(\tau')'] = (\tau\tau' - \tau\tau' - \tau\tau' + \min(\tau, \tau'))E[\Psi(\tau)\Psi(\tau')'] = (\min(\tau, \tau') - \tau\tau')E[\Psi(\tau)\Psi(\tau')']$, where $\epsilon(\tau) \equiv Y - D'\alpha(\tau) - X'\beta(\tau)$. This calculation uses that $P[\epsilon(\tau) \le 0 | Z, X] = \tau$ by Theorem 1.

Since $\widehat{\Psi}(\cdot) \to_p \Psi(\cdot)$ and $\widehat{\Phi}(\cdot) \to_p \Phi(\cdot)$ uniformly over compact sets and $\widehat{\pi}(\tau) - \pi(\tau) \to_p 0$ uniformly in τ, we have by R3 and R4 that $\delta_n \equiv \sup_{\tau\in\mathcal{T}} \rho(h'(\tau), h(\tau))|_{h'(\tau)=\hat{h}(\tau)} \to_p 0$, where ρ is the $L_2(P)$ pseudometric on $\mathcal{H}$:

$$\rho(h, \tilde{h}) \equiv \sqrt{E\Big\|\varphi_\tau(Y - D'\alpha - X'\beta - \Phi(X, Z)'\gamma)\Psi(X, Z) - \varphi_{\tilde{\tau}}(Y - D'\tilde{\alpha} - X'\tilde{\beta} - \tilde{\Phi}(X, Z)'\tilde{\gamma})\tilde{\Psi}(X, Z)\Big\|^2}.$$

Hence as $\delta_n \to 0$

$$\sup_{\tau\in\mathcal{T}} \Big\|\mathbb{G}_n \varphi_\tau(Y - D'\widehat{\alpha}(\tau) - X'\widehat{\beta}(\tau) - \widehat{\Phi}(\tau, X, Z)'\widehat{\gamma}(\tau))\widehat{\Psi}(\tau, X, Z) - \mathbb{G}_n \varphi_\tau(Y - D'\alpha(\tau) - X'\beta(\tau))\Psi(\tau, X, Z)\Big\|$$
$$\le \sup_{\substack{\rho(\tilde{h}, h) \le \delta_n \\ \tilde{h}, h \in \mathcal{H}}} \Big\|\mathbb{G}_n \varphi_{\tilde{\tau}}(Y - D'\tilde{\alpha} - X'\tilde{\beta} - \tilde{\Phi}(X, Z)'\tilde{\gamma})\tilde{\Psi}(X, Z) - \mathbb{G}_n \varphi_\tau(Y - D'\alpha - X'\beta - \Phi(X, Z)'\gamma)\Psi(X, Z)\Big\| = o_p(1),$$

[33] To check the Donskerness, it is easy to verify: (i) the stochastic equicontinuity of the process in τ with respect to the $L_2(P)$ pseudo-metric using the Holder property and stochastic equicontinuity of the process (B.1) in h, and (ii) finite-dimensional asymptotic normality by the Lindeberg-Levy theorem.

by stochastic equicontinuity of $h \mapsto \mathbb{G}_n \varphi_\tau(Y - D'\alpha - X'\beta - \Phi(X,Z)'\gamma)\Psi(X,Z)$.

Having shown claim II, a simple way to show claim I is to note that functions

$$\mathcal{P} = \left\{(\Phi, V, \alpha, \beta, \gamma, \tau) \mapsto \rho_\tau(Y - D'\alpha - X'\beta - \Phi(X,Z)'\gamma)V(X,Z)\right\}$$

are bounded by R1 and uniformly Lipschitz over $(\mathcal{F} \times \mathcal{F} \times \mathcal{A} \times \mathcal{B} \times \mathcal{G} \times \mathcal{T})$ which by Theorem 2.10.6 in van der Vaart and Wellner (1996) implies that $\mathcal{P}$ is Donsker. Donskerness implies a uniform LLN

$$\sup_{h \in \mathcal{H}} \left| \mathbb{E}_n \rho_\tau(Y - D'\alpha - X'\beta - \Phi(X,Z)'\gamma)V(X,Z) - E\rho_\tau(Y - D'\alpha - X'\beta - \Phi(X,Z)'\gamma)V(X,Z) \right| \to_p 0$$

which gives

$$\sup_{(\alpha,\beta,\gamma,\tau) \in (\mathcal{A} \times \mathcal{B} \times \mathcal{G} \times \mathcal{T})} \Big| \mathbb{E}_n \rho_\tau(Y - D'\alpha - X'\beta - \widetilde{\Phi}(\tau, X, Z)'\gamma)\widetilde{V}(\tau, X, Z)$$
$$- E\rho_\tau(Y - D'\alpha - X'\beta - \widetilde{\Phi}(\tau, X, Z)'\gamma)\widetilde{V}(\tau, X, Z)\Big|_{\widetilde{\Phi}(\cdot) = \widehat{\Phi}(\cdot), \widetilde{V}(\cdot) = \widehat{V}(\cdot)} \to_p 0.$$

By uniform consistency of $\widehat{\Phi}(\cdot)$ and $\widehat{V}(\cdot)$ and R4 we also have that

$$\sup_{(\alpha,\beta,\gamma,\tau) \in (\mathcal{A} \times \mathcal{B} \times \mathcal{G} \times \mathcal{T})} \Big| E\rho_\tau(Y - D'\alpha - X'\beta - \widetilde{\Phi}(\tau, X, Z)'\gamma)\widetilde{V}(\tau, X, Z)\Big|_{\widetilde{\Phi}(\cdot) = \widehat{\Phi}(\cdot), \widetilde{V}(\cdot) = \widehat{V}(\cdot)}$$
$$- E\rho_\tau(Y - D'\alpha - X'\beta - \Phi(\tau, X, Z)'\gamma)V(\tau, X, Z)\Big| \to_p 0.$$

The last two displays imply claim I. □.

C Verification of Linear Representations

LEMMA 4 *The conditions I.3 and I.4 hold for the proposed implementation in Examples 1-3 under conditions R.1-R.4. In Example 4 the conditions I.3 and I.4 hold under conditions R.1-R.4 for the IV-QR estimator and the standard regularity conditions for the QR estimator, e.g. those in Angrist et al. (2003).*

Proof. In Example 1, in the test of equality of distributions, I.3 is satisfied for $\widehat{\theta}(\cdot)$ by Theorem 3. Since $r = 0$, $z_i(\tau) = R(\tau)\left[J(\tau)^{-1} l_i(\tau, \theta(\tau))\Psi_i(\tau)\right]$, where

$$l_i(\tau, \theta(\tau)) = \left(\tau - 1(Y_i < D_i\alpha(\tau) + X_i'\beta(\tau))\right), \Psi_i(\tau) = V_i(\tau)[\Phi_i(\tau)', X']'. \tag{C.1}$$

Condition I.4(a) is checked in the proof of Lemma 3 in Appendix B, cf. the class of functions $\mathcal{H}$. Condition I.4.(b) holds by Theorem 1 and since Ψ_i is a function of (X_i, Z_i) only. Condition I.4(c) holds by the bounded density condition R.3. In Example 2, in the test of constant effect, $\widehat{r}(\cdot) = \widehat{\theta}(\frac{1}{2})$ is an IV-QR estimate. Thus for $l_i(\cdot)$ defined in (C.1) $z_i(\tau) = R(\tau)\left[J(\tau)^{-1} l_i(\tau, \theta(\tau))\Psi_i(\tau) - J(\frac{1}{2})^{-1} l_i(\frac{1}{2}, \theta(\frac{1}{2}))\Psi_i(\frac{1}{2}))\right]$, i.e. $d_i(\tau, r(\tau))\Upsilon_i(\tau) = l_i(\frac{1}{2}, \theta(\frac{1}{2}))\Psi_i(\frac{1}{2})$. Thus I.3-I.4 hold by the preceding argument. In Example 3, the test of stochastic dominance, $r = 0$, so the situation is identical to that in Example 1. In Example 4, in the test of exogeneity, the estimate of $\widehat{r}(\tau)$ is given by the ordinary QR estimator of Y on D, X, denoted as $\widehat{\vartheta}(\tau)$. In this case under the regularity conditions specified in Angrist et al. (2003) the estimator $\widehat{\vartheta}(\tau)$ satisfies: $\sqrt{n}\left(\widehat{\vartheta}(\cdot) - \vartheta_n(\cdot)\right) = -H(\cdot)^{-1} n^{-1/2} \sum_{i=1}^n d_i(\cdot, \vartheta(\cdot)) + o_p(1)$, $d_i(\tau, \vartheta(\tau)) = (\tau - 1(Y_i < \tilde{X}_i'\vartheta(\tau))\tilde{X}_i$, $\tilde{X}_i = (D_i', X_i')'$, $H(\tau) = Ef_{Y|\tilde{X}}(\vartheta(\tau)'\tilde{X})\tilde{X}\tilde{X}'$. Thus the score is given by $z_i(\tau) = R(\tau)[J(\tau)^{-1} l_i(\tau, \theta(\tau))\Psi_i(\tau) - H(\tau)^{-1} d_i(\tau, \vartheta(\tau)]$, where $Ed_i(\tau, \vartheta(\tau)) = 0$. The conditions I.3 and I.4 for $l_i(\tau, \theta(\tau))\Psi_i(\tau)$ are checked above. As for $d_i(\tau, \vartheta)$, the proof of Lemma 3 checks I.4.(a) (put $\tilde{X}_i$ in place of Ψ_i and $\gamma = 0$). Note that $Ed_i(\tau, \vartheta(\tau)) = 0$, so I.4.(b) holds, and I.4.(c) holds by the bounded density condition R.3. □

References

Abadie, A., October 1995. Changes in spanish labor income structure during the 1980s: a quantile regression approach, CEMFI Working Paper No. 9521.

Abadie, A., 2002. Bootstrap tests for distributional treatment effects in instrumental variable models. J. Amer. Statist. Assoc. 97 (457), 284–292.

Ambrosetti, A., Prodi, G., 1995. A primer of nonlinear analysis. Vol. 34 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge.

Amemiya, T., 1977. The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. Econometrica 45 (4), 955–968.

Amemiya, T., 1982. Two stage least absolute deviations estimators. Econometrica 50, 689–711.

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press.

Andrews, D., 1994. Empirical process methods in econometrics. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, Vol. 4. North Holland.

Angrist, J., Chernozhukov, V., Fernandez-Vál, I., 2003. Quantile regression under misspecification, with an application to the U.S. wage structure, MIT Working Paper.

Angrist, J., Krueger, A., 1991. Does compulsory school attendance affect schooling and earnings. Quarterly Journal of Economics 106, 979–1014.

Bai, J., 1997. Testing parametric conditional distributions of dynamic models, Working Paper, MIT Department of Economics.

Becker, G., Chiswick, B., 1966. Education and the distribution of earnings. American Economic Review 56, 253–284.

Bhattacharya, P. K., 1963. On an analog of regression analysis. Ann. Math. Statist. 34, 1459–1473.

Bound, J., Jaeger, D. A., 1996. On the validity of season of birth as an instrument in wage equations: A comment on Angrist and Krueger's "Does compulsory attendance affect schooling and earnings?", NBER Working Paper 5835.

Caccioppoli, R., 1932. Sugli elementi uniti delle transformazioni funzionali. Rend. Matem. Padova 3, 1–15.

Card, D., 1995. Earnings, schooling, and ability revisited. In: Polachek, S. (Ed.), Research in Labor Economics, Vol. 14. JAI Press: Greenwich, CT.

Card, D., 1999. The causal effect of education on earnings. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, Vol. 3A. North-Holland, Amsterdam.

Carneiro, P., Heckman, J., Vytlacil, E., 2000. Estimating the returns to education when it varies among individuals, Working Paper, Department of Economics, University of Chicago.

Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. J. Econometrics 34 (3), 305–334.

Chernozhukov, V., 2002. Robust inference on the quantile regression process using subsampling. Working Paper, Department of Economics, MIT.

Chernozhukov, V., Hansen, C., 2001a. An instrumental variable model of quantile treatment effects. Forthcoming in Econometrica.

Chernozhukov, V., Hansen, C., 2001b. An IV model of quantile treatment effects. Working Paper, Department of Economics, MIT.

Chernozhukov, V., Hansen, C., 2003. The impact of 401(K) on savings: An instrumental quantile regression analysis. Forthcoming in The Review of Economics and Statistics.

Chernozhukov, V., Hong, H., 2002. Three-step censored quantile regression and extramarital affairs. J. Amer. Statist. Assoc. 97 (459), 872–882.

Chernozhukov, V., Hong, H., 2003. An MCMC approach to classical estimation. Journal of Econometrics 115, 293–346.

Davydov, Y. A., Lifshits, M. A., Smorodina, N. V., 1998. Local properties of distributions of stochastic functionals. American Mathematical Society, Providence, RI, translated from the 1995 Russian original by V. E. Nazaĭkinskiĭ and M. A. Shishkova.

Doksum, K., 1974. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. Ann. Statist. 2, 267–277.

Durbin, J., 1973. Weak convergence of the sample distribution function when parameters are estimated. Ann. Statist. 1, 279–290.

D'Urso, V. T., 2002. Home buyer search duration and the internet. Working Paper 168, MIT E-Business Center.

Frakes, M., Gruber, J., 2003. The impact of smoking on obesity. Working Paper, MIT Department of Economics.

Graddy, K., 1995. Testing for imperfect competition at the Fulton fish market. Rand Journal of Economics 26(1), 75–92.

Hadamard, J., 1906. Sur les transformations ponctuelles. Bull. Soc. Math. France 34, 71–84.

Hausman, J. A., 1977. Errors in variables in simultaneous equation models. J. Econometrics 5 (3), 389–401.

Hausman, J. A., 1978. Specification tests in econometrics. Econometrica 46 (6), 1251–1271.

Hausman, J. A., Sidak, J. G., 2002. Do the poor and the less-educated pay higher prices for long-distance calls? Working Paper, Department of Economics, MIT.

Heckman, J., 1990. Varieties of selection bias. American Economic Review, Papers and Proceedings 80, 313–338.

Heckman, J., Robb, R., 1986. Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In: Wainer, H. (Ed.), Drawing Inference from Self-Selected Samples. Springer-Verlag, New York, pp. 63–107.

Heckman, J., Vytlacil, E., 1999. Instrumental variables methods for the correlated random coefficients models. J. Human Resources 33 (4), 975–986.

Hogg, R. V., 1975. Estimates of percentile regression lines using salary data. Journal of the American Statistical Association 70 (349), 56–59.

Hong, H., Tamer, E., 2003. Inference in censored models with endogenous regressors. Econometrica 71 (3), 905–932.

Imbens, G. W., Angrist, J. D., 1994. Identification and estimation of local average treatment effects. Econometrica 62 (2), 467–475.

Januszewski, S. I., 2002. The effect of air traffic delays on airline prices. Working Paper, Department of Economics, Unoversity of California in San-Diego.

Koenker, R., 1994. Confidence intervals for regression quantiles. In: Asymptotic statistics (Prague, 1993). Physica, Heidelberg, pp. 349–359.

Koenker, R., Bassett, G. S., 1978. Regression quantiles. Econometrica 46, 33–50.

Koenker, R., D'Orey, V., 1987. Computing regression quantiles. Applied Statistics 36, 383–390.

Koenker, R., Xiao, Z., 2002. Inference on the quantile regression process. Econometrica 70, 1583–1612.

Lehmann, E. L., 1974. Nonparametrics: statistical methods based on ranks. Holden-Day Inc., San Francisco, Calif.

Macurdy, T., Timmins, C., 2000. Application of smoothed quantile estimation. Paper presented at the Konstanz Conference on Economic Applications of Quantile Regression.

Mas-Colell, A., 1979a. Homeomorphisms of compact, convex sets and the Jacobian matrix. SIAM J. Math. Anal. 10 (6), 1105–1109.

Mas-Colell, A., 1979b. Two propositions on the global univalence of systems of cost function. In: General equilibrium, growth, and trade. Academic Press, New York, pp. 323–331.

McFadden, D., 1989. Testing for stochastic dominance. In: Fomny, T., Seo, T. (Eds.), Studies in Economics of Uncertainty in Honor of Josef Hadar.

Mincer, J., 1970. The distribution of labor incomes: A survey with special reference to the human capital approach. Journal of Economic Literature 8, 1–26.

Mincer, J., 1974. Schooling, Experience, and Earnings. Columbia University Press: New York.

Mincer, J., 1995. Investment in human capital and personal income distribution. In: Labor Economics. Volume 2. Employment, Wages, and Education. Elgar: Aldershot U.K., pp. 3–24.

Politis, D. N., Romano, J. P., Wolf, M., 1999. Subsampling. Springer-Verlag, New York.

Portnoy, S., 1991. Asymptotic behavior of regression quantiles in nonstationary, dependent cases. J. Multivariate Anal. 38 (1), 100–113.

Portnoy, S., Chen, L., 1996. Two-stage regression quantiles and two-stage trimmed least squares estimators for structural equation models. Comm. Statist. Theory Methods 25 (5), 1005–1032.

Portnoy, S., Koenker, R., 1997. The Gaussian Hare and the Laplacian Tortoise. Statistical Science 12, 279–300.

Powell, J. L., 1983. The asymptotic normality of two-stage least absolute deviations estimators. Econometrica 51 (5), 1569–1575.

Powell, J. L., 1986. Censored regression quantiles. Journal of Econometrics 32, 143–155.

Sakata, S., 2001. Instrumental variable estimation based on the least absolute deviation estimator. Working Paper, Department of Economics, University of Michigan.

Tikhonov, A. N., Arsenin, V. Y., 1977. Solutions of ill-posed problems. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York. Scripta Series in Mathematics.

van der Vaart, A. W., 1998. Asymptotic statistics. Cambridge University Press, Cambridge.

van der Vaart, A. W., Wellner, J. A., 1996. Weak convergence and empirical processes. Springer-Verlag, New York.

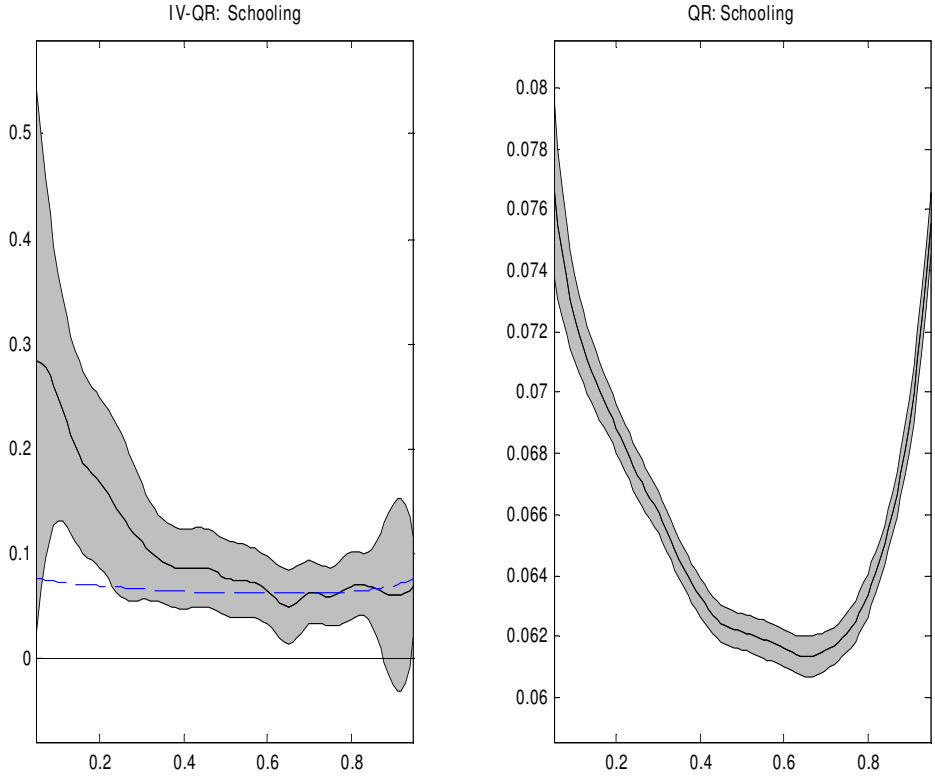

Figure 1: The sample size is 329,509. Coefficient estimates are on the vertical axis, while the quantile index is on the horizontal axis. The shaded region is the 95% confidence band estimated using robust standard errors. The left panel contains estimates of the returns to schooling obtained through instrumental variables quantile regression, and the right panel presents estimates of the effect of years of schooling on earnings obtained through standard quantile regression. For comparison, the dashed line in the first panel plots the schooling coefficient estimated through standard quantile regression. All estimates were computed at .05 unit intervals for $\tau \in [.05, .95]$.