
MASSACHUSETTS
INSTITUTE of TECHNOLOGY
LIBRARIES

Massachusetts Institute of Technology
Department of Economics
Working Paper Series

# AN IV MODEL OF QUANTILE TREATMENT EFFECTS

Victor Chernozhukov, MIT
Christian Hansen, MIT

Working Paper 02-06
December 2001

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

Massachusetts Institute of Technology
Department of Economics
Working Paper Series

# AN IV MODEL OF QUANTILE TREATMENT EFFECTS

Victor Chernozhukov, MIT
Christian Hansen, MIT

Working Paper 02-06
December 2001

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

# An IV Model of Quantile Treatment Effects[1]

Victor Chernozhukov and Christian Hansen
Department of Economics
MIT

MIT Department of Economics Working Paper
Revised: December 2001

Abstract

This paper develops a model of quantile treatment effects with treatment endogeneity. The model primarily exploits similarity assumption as a main restriction that handles endogeneity. From this model we derive a Wald IV estimating equation, and show that the model does not require functional form assumptions for identification.

We then characterize the quantile treatment function as solving an "inverse" quantile regression problem and suggest its finite-sample analog as a practical estimator. This estimator, unlike generalized method-of-moments, can be easily computed by solving a series of conventional quantile regressions, and does not require grid searches over high-dimensional parameter sets. A properly weighted version of this estimator is also efficient. The model and estimator apply to either continuous or discrete variables. We apply this estimator to characterize the median and other quantile treatment effects in a market demand model and a job training program.

Key Words: Quantile Regression, Inverse Quantile Regression, Instrumental Quantile Regression, Treatment Effects, Empirical Likelihood, Training, Demand Models.

JEL Codes: C13, C14, C30, C51, D4, J24, J31.

 We thank the seminar participants at Cornell, University of Pennsylvania, University of Illinois at Urbana-Champaign, Winter Econometric Society 2000, EC2 Conference on Causality and Exogeneity in Econometrics for useful discussions of the research reported in this paper. We would like to express our appreciation to Guido Imbens, Petra Todd, James Heckman, Joel Horowitz, Roger Koenker, Stephen Portnoy, Edward Vytlacil, and especially Joshua Angrist, Jerry Hausman and Whitney Newey for constructive discussions. Conversations with Alberto Abadie at the Konztanz conference, May 2000, Takeshi Amemiya, Han Hong, as well as debates with Tom MaCurdy, while the first author was a student at Stanford, inspired the model and the estimation developed in this paper. We are especially grateful to them.

[ For convenience of refereeing only. To be removed.]

# Contents

# 1 Introduction

The ability of quantile regression models, Koenker and Bassett (1978), to characterize the impact of variables on the distribution of outcomes makes them appealing for examining many economic applications, see e.g. Buchinsky (1998) and Koenker and Hallock (2001). The distributional impacts of social programs, such as welfare, unemployment insurance, and training programs are of large interest to economists. Unfortunately, in all of these cases, treatment is self-selected or endogenous, making conventional quantile regression inappropriate. This paper makes two contributions.

***First***, this paper proposes a model of quantile treatment effects with endogeneity. At the heart of the model is an assumption of similarity (containing rank invariance as a special case) that allows us to address endogeneity. This differs from the monotonicity assumptions of Heckman's nonparametric selection model and Imbens and Angrist's LATE model.[2] We show that this model's *main implication* is a Wald IV estimating equation:

$$P(Y \leq q(D,\tau)|Z) = \tau, \tag{1}$$

where $q(d,\tau)$ is the $\tau$-quantile of the potential or counterfactual outcome when the treatment is exogenously set to the value $d$, $D$ is the actual endogenous treatment, $Y$ is the actual outcome, and $Z$ is an instrument. Thus, the model provides a causal justification and interpretation of the Wald IV estimating equation (1).[3] We also show that the model does not require functional form assumptions for identification.

***Second***, we characterize the function $q$ as solving an inverse quantile regression problem and suggest its finite-sample analog as a practical estimator. This estimator, unlike generalized method-of-moments and other similar estimators, can be easily computed by solving a series of conventional quantile regressions (convex optimization problems), and does not require grid searches over high-dimensional parameter sets. A properly weighted version of this estimator is also efficient. We apply this estimator to characterize the median and other quantile treatment effects in a market demand model and a job training program.

An important aspect of the proposed model is treatment effect heterogeneity, given which conventional linear IV inconsistently estimate the average treatment effect (example 5.1). Thus, even if one is interested in ATE, one has to estimate the QTE's first and integrate them over the quantile index to obtain a consistent estimate of ATE. Alternatively, one may estimate only the median treatment effects to characterize the central effects, using the proposed approach.[4]

[2]See Vytlacil (2001) on the distributional equivalence of these two models.

[3]There is very important prior work that estimates functions $q$ under restrictions like (1). This work starts as early as Hogg (1975); Koenker (1998) characterizes Hogg's estimator as a Wald's IV approach to quantile regression. See Abadie (1995), Christoffersen, Hahn, and Inoue (1999), MaCurdy and Timmins (1998) for GMM-like approaches to estimation and testing. Also see Hong and Tamer (2001) for a fundamental treatment of censoring case. The problem is that a function $q$, satisfying the estimating equation, has not had any known causal meaning within standard IV models with non-constant treatment effects, such as those of Heckman or Imbens and Angrist. Thus, our model provides a causal interpretation and support of (1) and of these previous important estimators.

[4]In expected utility framework, some form of average is typically of interest, but since we (econome-

Further details of the model and the estimator are as follows. The model is developed in the standard potential outcomes framework, and the QTE is defined as the difference in the quantiles of potential outcomes under potential treatments. At the heart of the model is *similarity*, a generalization of rank invariance assumption, which is reasonable in many applications and also facilitates interesting interpretations of QTE, as in Lehmann(1974), Doksum (1974), and Koenker and Geling (2001). This assumption is different from the monotonicity assumptions of the prevalent IV models (the selection – LATE models). Similarity also requires less stringent independence conditions (allowing, for example, measurement error in the instrument). As a result the model differs from the selection-LATE models. However, the two models do contain a large common subclass.

It should also be noted that the model and estimators looked at in this paper both substantively complement and differ from the fundamental model of Amemiya (1982) and the QTE model of Abadie, Angrist, and Imbens (2001) developed within the LATE framework. Amemiya's approach and its extension by Chen and Portnoy (1996), known as two-stage quantile regression (2SQR), allow continuous treatment variables. However, we show that 2SQR is not consistent when the quantile treatment effect differs across quantiles (Appendix B). The inconsistency is noted, since this estimator has often been used expressly to estimate heterogeneous quantile treatment effects. On the other hand, Abadie et al's (2001) approach applies only to binary treatments, and its extension to more general treatments is not known. Allowing general treatments is clearly important.

The approach in this paper expressly allows for QTE that vary across quantiles and applies to arbitrary– continuous, discrete, or binary – treatment variables. Thus it can be used to study education effects, demand systems, and any other non-binary treatments.

On the estimation side, the inverse quantile regression is an easily computable and transparent estimator, unlike GMM that requires grid searches over high-dimensional parameter sets. The estimator is obtained as a link between Koenker-Bassett quantile regression and the Wald IV restrictions. In addition to deriving theoretical properties of inverse quantile regression, we provide user-friendly computer programs that implement the estimator, standard errors, and produce graphical output.

The remainder of the paper is organized as follows. Section 2 presents the model. Section 3 provides two economic models as examples: aggregate demand analysis and the returns to education. Section 4 presents identification results and the inverse quantile regression. Estimation methods are described in Section 5, and Section 6 contains two empirical applications, corresponding to models in section 3.

A word on notation. Following Koenker, we use $F_Y(\cdot|x)$ and $Q_Y(\tau|x)$ to denote the conditional distribution function and the $\tau$-quantile of $Y$ given $X = x$; capitals such as $Y$ denote random variables and $y$ denote the values they take.

tricians) typically do not know (are agnostic about) which particular average, the entire distributional impact needs to be evaluated. In addition, Manski's(1988) ingenious work provides ordinal utility models of decision making under uncertainty, where agents maximize a $\tau$-quantile of utility distribution. In such framework only quantiles of potential outcomes would be of interest to a policy-maker.

# 2 A Model of Quantile Treatment Effects

The section begins with an important preliminary discussion that naturally leads to the QTE model of this paper.

## 2.1 Potential Outcomes and the QTE

We develop our model within the conventional Neymann-Fisher-Rubin potential outcome framework.[5] *Potential* real-valued *outcomes* are indexed against treatment $D$ ($D \in \mathcal{D}$, a subset of $\mathbb{R}^l$), and denoted $Y_d$, while potential treatment status is indexed against the instrument $Z$, and denoted $D_z$. For example, $Y_d$ is an individual's outcome when $D = d$ and $D_z$ is an individual's treatment status when $Z = z$.

The potential or counterfactual outcomes $\{Y_d, \quad d \in \mathcal{D}\}$, such as wages or demand, vary across individuals or states of the world. Given the actual treatment $D$, the observed outcome is

$$Y \equiv Y_D.$$

That is, only the $D$-th component of $\{Y_d, d \in \mathcal{D}\}$ is observed. Typically $D$ is selected in relation to potential outcomes, *inducing endogeneity or sample selectivity.*

The objective of causal analysis is to learn about the features of marginal distributions of potential outcomes $Y_d$. For example, $\mu_{d,d'} = EY_d - EY_{d'}$ is the average treatment effect (ATE). The quantile treatment effect (QTE) is the difference in quantiles of potential outcomes under different potential treatments:[6]

$$Q_{Y_d}(\tau) - Q_{Y_{d'}}(\tau).$$

A main obstacle to learning about the QTE is the sample selectivity or endogeneity.

Early formulations of QTE by Lehmann (1974) and Doksum (1974) axiomatically interpret QTE as a a measure of interaction of the latent ability $\tau$ ( "prone to die at an early age, "prone to learn fast," etc.) and the treatment. The subjects differ in this latent characteristic and their response to the treatment is described by QTE. An assumption that allows such interpretation is rank invariance. Rank invariance was also used by Heckman and Smith (1997) and Koenker and Bilias (2001) in quantile models without endogeneity.[7]

Our model uses similarity as a main restriction that allows to address endogeneity. Similarity facilitates analogous interpretation of the quantile treatment effects in our framework and incorporates rank invariance as a special case.

## 2.2 The Instrumental Quantile Treatment Model.

The first part of the model is a potential outcomes model. The other part relates the treatment choice to the potential outcomes, accounting for endogeneity.

[5] See e.g. Heckman and Robb (1986) and Imbens and Angrist (1994).

[6] Generally, QTE are more informative than ATE, since they summarize the distributional impact, whereas ATE summarize the impact on the first moment of the distribution. In fact, $\mu_{d,d'} = \int_0^1 \left(Q_{Y_d}(\tau) - Q_{Y_{d'}}(\tau)\right) d\tau$.

[7] Heckman and Smith (1997) use rank invariance to identify $Q_{Y_d - Y_{d'}}(\tau) = Q_{Y_d}(\tau) - Q_{Y_{d'}}(\tau)$.

**Assumption 1 (IQT Model)** For almost every value of $(X, Z) = (x, z)$,

**A1** POTENTIAL OUTCOMES. Given $X = x$, for some $U_d \overset{d}{\sim} U(0,1)$,

$$Y_d = q(d, x, U_d),$$

such that $q(d, x, \tau)$ is the $\tau$-th quantile of $Y_d$ for any $0 < \tau < 1$.

**A2** SELECTION. For unknown function $\delta$ and random process $V$, given $X = x, Z = z$,

$$D_z \equiv \delta(z, x, V).$$

**A3** INDEPENDENCE. Given $X$, $\{U_d\}$ is independent of $Z$.

**A4** SIMILARITY. For each $d$ and $d'$, given $(V, X, Z)$

$$U_d \text{ is equal in distribution to } U_{d'}.$$

**A5** OBSERVED variables $W$ consist of ( for $U \equiv U_D$)

$$\begin{cases} Y \equiv q(D, X, U), \\ D \equiv \delta(Z, X, V), \\ X, Z. \end{cases}$$

**Remark 2.1** Of interest also is a much more restrictive special case of A3 and A4

**A3*** FULL INDEPENDENCE. $\{U_d, V\}$, or equivalently $\{Y_d, D_z\}$, are jointly independent of $Z$, given $X$.

**A4*** RANK INVARIANCE. $U_d \equiv U_{d'} \equiv U$ for each $d$.

In A1 the conditional $\tau$- quantile of $Y_d$ is $q(d, x, \tau)$, given $X = x$. Our main interest is the **Conditional Quantile Treatment Effect**

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

the difference in quantiles of potential outcomes distributions conditional on $x$.

In A2, the unobserved random vector $V$ is responsible for the difference in treatment choices $D_z$ across observationally identical individuals. $\delta(\cdot)$ is the (measurable) *selection function.* We do not impose any other assumptions on this function. This is important to accomodate realistic economic examples.

A3 states that potential outcomes are *independent* of $Z$, given $X$. A3 is more general than A3*, the assumption of selection-LATE models, that requires both $\{Y_d\}$ and the potential treatments $\{D_z\}$ to be independent of the instrument $Z$. A3* is a strong assumption that can be easily violated when the instrument is measured with error or there are omitted variables related to $Z$ (a part of the error $V$) in the selection equation. Imbens and Angrist (1994) provide additional examples violating A3*.

Given the same observed characteristic $x$ and treatment $d$, the subjects still differ in terms of potential outcomes. Their *relative ranking* is determined by the *rank* or *ability* vector $\{U_d\}$. This vector can be collapsed to a single variable under assumption A4* – the rank invariance or common error assumption.

A4, similarity, states that given the information $(V, Z, X)$ the *expectation* of (any function of) $U_d$ *does not vary* across the treatment states $d$. In other words, *ex-ante* the ranks are "similar," while *ex-post* the ranks may *differ*. Thus similarity allows substantial ex-post *slippage in the ranks*, the importance of allowing which was shown by Heckman and Smith (1997). See Example 3.2 for an example.

A4 also facilitates interpretation of the QTE as a measure of the interaction between the latent ex-ante ability $\tau$ and the treatment, following Doksum (1974) and Koenker and Bilias (2001). Additionally, A4 is a *key identification device*, leading to the Wald restrictions in Section 4.

Similarity is the main restriction of the IQT model. It is absent in the conventional LATE/selection models. However, A4 enables a more general selection function in A2 that requires neither the monotonicity assumption or stronger independence assumptions of the LATE models. Thus, LATE models mainly exploit monotonicity and stronger independence assumption to address endogeneity, while the present approach uses the similarity assumption. The value of one versus the other has to be judged in each particular application.

## 2.3 A Comparison with a LATE Model with Common Error.

Although the IQT model differs from selection-LATE model, the two do contain a large common subclass. Indeed, consider the following model:

**V1** $Y_d = g(\nu_d(X), U), \quad d \in \{0, 1\}$.

**V2** $D_z = 1(\vartheta(z, X) > V)$ for some real-valued function $\vartheta$.

**V3** $\{Y_d, V\}$ (or $\{Y_d, D_z\}$) are independent of $Z$, given $X$.

**V4** $U$ does not vary across potential treatments $d$.

Assume that $g$ is monotone in $U$, so that error $U$ can be normalized to be uniform. This model is a special case of the IQT model, with assumption V4 corresponding to exact rank invariance or common error assumption A4* (see Doksum (1974), Robins and Tsiatis (1991), Heckman and Smith (1997), and Vytlacil (2000) for various justifications of rank invariance), and V3 being a stronger version of the independence assumption A3, in fact corresponding to A3*. Vytlacil (2000) shows such a model incorporates a wide variety of familiar nonlinear simultaneous equations models. In turn, the IQT model incorporates the model V1-V4 as an important special case.

# 3 Economic Examples

The following examples highlight the nature of the IQT model. The discussion is quite thorough because it underlies the empirical applications in Section 6.

**Example 3.1 (Demand with Non-Separable Error)** The following is a generalization of the classic supply-demand example. Consider the "random coefficient" model

$$\begin{cases} \textbf{i.} & Y_p = q(p, U), \\ \textbf{ii.} & \bar{Y}_p = \rho(p, z, \mathcal{U}), \\ \textbf{iii.} & P \in \{p : q(p, Z, U) = \rho(p, \mathcal{U})\}. \end{cases} \tag{2}$$

The map $p \mapsto Y_p$ is the random demand function, that is, it is the *potential demand* when the price is set (externally) to the value $p$. Likewise, $p \mapsto \bar{Y}_p$ is the random supply function, that is the *potential supply* when the price is set (externally) to $p$. Additionally, $Y_p$ and $\bar{Y}_p$, $q(\cdot)$, and $\rho(\cdot)$ depend on the covariates $X$, but this dependence is suppressed. Random variable $U$ is the level of the demand in the sense that $(p, U) \leq q(p, U')$ when $U \leq U'$. Demand is maximal when $U = 1$ and minimal when $U = 0$, holding $p$ fixed. Likewise, $\mathcal{U}$ is the level of supply. The $\tau$-quantile of the demand curve $p \mapsto Y_p$ is given by

$$p \mapsto Q_{Y_p}(\tau) \equiv q(p, \tau).$$

Thus with probability $\tau$, the curve $p \mapsto Y_p$ lies below the curve $p \mapsto Q_{Y_p}(\tau)$.

The quantile treatment effect is characterized by an elasticity $\partial \ln q(p, \tau)/\partial \ln p$. The elasticity depends on the state of the demand $\tau$ (low or high) and may vary with $\tau$. For example, this variation could arise when the number of buyers varies and aggregation induces non-constant elasticity across the demand levels as a process of summation of individual demand curves, holding the price fixed.

This model incorporates many traditional models with separable error

$$Y_p = q(p) + \mathcal{E}, \text{ where } \mathcal{E} = F_{\mathcal{E}}^{-1}(U). \tag{3}$$

The model **i.** is much more general in that the price can affect the entire distribution of the demand curve, while in (2) it only affects the location of the distribution of the demand curve.

Condition **iii.** is the equilibrium condition that generates *endogeneity* – the *selection* of the actual price by the market depends on the potential demand and supply outcomes **i.** and **ii.** As a result $P = \delta(Z, V)$, where $V$ consists of $U$, $\mathcal{U}$, and other variables (including "sunspot" variables, if the equilibrium price is not unique). Thus what we observe can be written as simultaneous equations of a general form, with observables[8]

$$\begin{cases} Y \equiv q(P, U), \\ P \equiv \delta(Z, V). \end{cases} \tag{4}$$

[8]To appreciate the generality, note that model incorporates, for example, the simultaneous equations model of Imbens and Newey(2001), who assume that $V$ is univariate, $\delta$ is monotone in $V$, both $V$ and $U$ are independent of $Z$, if in addition we assume $U$ is uniform. Imbens and Newey (2001) developed some ingenious identification results using these stronger assumptions.

Because of endogeneity, $Q_{Y|P}(\tau) \neq q(P,\tau)$, therefore the conventional quantile regression will be inappropriate to estimate the $\tau$-th quantile demand curve. Additionally, we show in Appendix A that 2SQR is generally not suitable for estimation purposes.

We show that the instrumental variables $Z$, like weather conditions, that shift the supply curve and do not affect the level of the demand curve $U$ allows identification of the $\tau$-quantile of the demand function, $p \mapsto q(p,\tau)$. Furthermore, the IQT model allows arbitrary correlation between $Z$ and $V$. This allows, for example, measurement error in $Z$ (e.g. in weather conditions). The standard IV approaches (Heckman et al (2001), Imbens and Angrist (1994)) do not accommodate such a possibility.

**Example 3.2 (Education/Training Returns)** Let "earnings" in the "education" states $d \in \{0,1\}$ be determined by a "random coefficients" model

$$Y_1 = q_1(X,U_1), \quad Y_0 = q_0(X,U_0).$$

An individual's training or education decision is given by

$$D = 1\left(\varphi(Z,X,V) \geq 0\right)$$

where unobserved vector $V$ potentially depends on (but is not necessarily determined by) the ability vector $(U_1,U_0)$ and arbitrarily on functions $q_1$ and $q_0$, $X$ and $Z$. The first kind of dependence is endogeneity.

In the standard Roy model, no restrictions are placed on the individual specific variations in earnings, and the individual observes these before making the schooling choice. For identification, we impose *similarity*: conditional on $(Z,X,V)$, $U_1$ equals in distribution to $U_0$. This is more restrictive than the general case, but perhaps not as restrictive as it may appear. This restriction allows arbitrary correlation between $Y_0$ and $Y_1$ and allows the general treatment impacts through the $q_1(\cdot)$ and $q_0(\cdot)$ functions. A main difference between this model and the Roy model is the implicit ex ante nature of the decision process. Instead of knowing the exact outcomes in any state of the world, the subject anticipates the same distribution of ability across treatment states and makes the decision accordingly.[9]

Indeed, consider a simple example that satisfies similarity A4:

$$U_0 = \eta + \nu_0, \quad U_1 = \eta + \nu_1,$$

where $\eta$ is a function of error vector $V$ in the selection equation, and $\nu_0$ and and $\nu_1$ are the slippage terms such that $\nu_0 \overset{d}{\sim} \nu_1$ given $(X,Z,V)$. Rank invariance A4* is a degenerate case when $\nu_1 = \nu_0 = 0$.

Finally note that the similarity only need hold conditional on $Z$, $X$, and $V$. This seems to be a reasonable framework. For example, people generally decide on whether to attend college or not before they observe their rank/ability among college educated and non-college educated individuals with observationally identical characteristics. Thus, it seems a plausible approximation that they would anticipate the same

[9]More precisely, we assume that he has enough information only to anticipate the same distribution of ability across states. The assumption does not require the subject to have correct beliefs.

distribution of their rank/ability across the treatment states **relative to similar** individuals (with the same covariates $X$ and $Z$).

Another difference with conventional IV model is that the IQT model expressly allows for dependence to exist between the instrument $Z$ and $V$ whereas the standard approaches expressly disallow this, as mentioned in the previous example. E.g., consider the following simple schooling decision rule

$$D = 1\{\varphi(Z) + V \geq 0\}.$$

In the schooling or training context, if $Z$ is a family background, it may be measured with a sizable error, so independence between $Z$ and $V$ need not hold. $V$ could also capture omitted variables which are correlated to $Z$ and impact the schooling decision but not the outcome. Note that measurement error or omitted variables also violate the monotonicity assumption often used in the IV literature. See Imbens and Angrist (1994) for other examples of violation.

To summarize, three aspects of the proposed model are highlighted by the above examples. *First*, the IQT model allows *arbitrarily general* quantile treatment effects. The similarity assumption in no way restricts their shape. *Second*, under similarity, we can interpret the QTE as measuring the interaction between the latent ex-ante ability and the treatment, following Doksum (1974) and Koenker and Bilias (2001). The similarity seems reasonable in many settings. *Third*, the similarity allows the selection in A2-A3 to be more general than that in the popular IV approaches, although this should be taken as a subsidiary point.

# 4 Wald IV and Inverse Quantile Regression

Here we establish a link between the IQT model and the Wald-type IV restrictions, relate those to Koenker and Basset's (1978) quantile regression, and show that the model is identified without functional form assumptions.

## 4.1 Main Identification Restriction

The following theorem provides provides an important link of the parameters of the IQT model to the Wald-type IV estimating equations.

**Theorem 1** *Suppose A1-A5 hold, and given $X, Z$*
*i. if $Y$ is continuously distributed ($q(D, X, \tau)$ is strictly increasing in $\tau$ a.s.) then a.s.*

$$\begin{aligned} P[Y \leq q(D, X, \tau)|X, Z] &= \tau, \\ P[Y < q(D, X, \tau)|X, Z] &= \tau, \end{aligned} \tag{5}$$

*ii. otherwise ($q(D, X, \tau)$ is non-decreasing in $\tau$, a.s.), a.s.*

$$\begin{aligned} P[Y \leq q(D, X, \tau)|X, Z] &\geq \tau, \\ P[Y < q(D, X, \tau)|X, Z] &\leq \tau, \end{aligned} \tag{6}$$

*with the last inequality being strict if* $q(D, X, \tau') = q(D, X, \tau)$ *for some* $\tau' > \tau$ *with probability* $P > 0$ *given* $X$ *and* $Z$.

By linking the IQT model to Wald's IV quantile restrictions, Theorem 1 provides an empirical and causal content to these restrictions. In this regard, the IQT model serves the same purpose as the LATE model developed by Imbens and Angrist (1994) to provide the link between the Wald's IV approach and the (local) average treatment effects. However, our results employ the similarity assumption in place of the monotonicity assumptions to obtain this link.

As noted, Theorem 1 allows for an arbitrary variable $Y$, for arbitrary treatment variable $D$, and arbitrary instrument $Z$. Thus equations (5) and (6) lead to natural ways to estimate any model with endogeneity as long as the corresponding quantiles of the potential outcome distribution $q(d, x, \tau)$ may be specified. We focus on the continuous $Y$, but the discrete case is clearly relevant – see e.g. Manski (1985), Horowitz (1992), Powell (1986), and Hong and Tamer (2001).

Before proceeding further, it is very important to note that although the IQT model allows the use of a "conditioning on $Z$" strategy to estimate the quantile treatment effects, it is not possible to use the same "conditioning on $Z$" strategy to estimate other treatment effects of interest. For example, in order to estimate the average treatment effect within the IQT model, we first need to estimate the quantile treatment effects and then integrate them over quantile index $\tau$. Conventional linear IV will not work here. This feature is analogous to that in the selection-LATE models (Heckman 1990).

**Example 4.1 (Average Treatment Effects: Failure of 2SLS)** Within A1-A5, suppose $\mu(d)$ is finite in the equation

$$Y_d = \mu(d) + \epsilon_d, \quad E\epsilon_d = 0,$$

where $\mu(d)$ is the mean treatment function. It would be natural to expect that $E[Y - \mu(D)|Z] = 0$, but this is **false** since generally

$$E[q(D, U) - \mu(D)|Z] \neq 0,$$

because $U$ is not independent of $D$ conditional on $Z$ in general, so that

$$\begin{aligned} E[Y \equiv q(D, U)|Z] &= \int\int_{[0,1]} q(d, u) dP[D = d, U = u|Z] \\ &\neq \int\int_{[0,1]} q(d, u) dP[D = d|Z] \cdot dP[U = u|Z] \equiv E[\mu(D)|Z]. \end{aligned}$$

The equality holds if there is no endogeneity or the treatment effect is constant.

## 4.2 The Inverse Quantile Regression

The main identification restriction of Theorem 1 can be posed as an optimization problem, which we call the *inverse quantile regression* for its "inverse" relation to the

(conventional) quantile regression of Koenker and Bassett (1978). This links the IQT model, the Wald IV restrictions, and quantile regression together.

In order to obtain the link, we note that Theorem 1 states that 0 is the $\tau$-th quantile of random variable $Y - q(X, D, \tau)$ conditional on $(X, Z)$. Therefore, the problem of finding a function $q(x, d, \tau)$ satisfying equations (5) or (6) is the problem of the *inverse quantile regression*:

> Find a function $q(x, d, \tau)$ such that **0** is the solution to the quantile regression problem, in which we regress $Y - q(X, D, \tau)$ on any function of $(Z, X)$.

Theorem 2 formally states this result.

**Theorem 2** *For P-a.e. value $(x, z)$ of $(X, Z)$, the following are **equivalent** statements, for each measurable $q'$*

1. *$q'$ satisfies equation (5) or (6) (in place of $q$).*
2. *$Q_{\epsilon|X,Z}(\tau) = 0$, where $\epsilon \equiv Y - q'(D, X, \tau)$.*
3. *assuming integrability, $q'$ satisfies*

$$0 \equiv \underset{v \in \mathbb{R}}{\text{argminl}}\ E\left[\rho_\tau\left(Y - q'(x, d) - v\right) | x, z\right],$$

*where $\rho_\tau(u) \equiv \tau u^+ + (1 - \tau)u^-$.*

4. *$q'$ is an $\text{argmin}_\varphi \left\|v(x, z)\right\|$, where the minimum is computed over all candidate (measurable) functions $\varphi$, and, assuming integrability,*

$$v(x, z) \equiv \underset{v \in \mathbb{R}}{\text{argminl}}\ E\left[\rho_\tau\left(Y - \varphi(x, d) - v\right) | x, z\right].$$

**Remark 4.1** Integrability conditions can be removed by subtracting $\rho_{\tau'}(Y - q'(x, d) - \bar{v})$, where $\bar{v}$ is a fixed number, inside the expectation. The "argminl" above means "$\lim_{\tau' \uparrow \tau}$ argmin," and is a pure technicality, insuring uniqueness of solution. It is only needed there for non-continuous $Y$ and at most countably many values of $\tau \in (0, 1)$.

Theorem 2 applies to continuous, discrete, or mixed outcomes, so estimation based on Theorem 2 can be applied to such data. Theorem 2 is both interpretive and constructive. First, any consistent estimator asymptotically solves the inverse quantile regression problem. Second, Theorem 2 (part 4) suggests a way to construct practical estimators (in addition to obvious method of moments or minimum distance methods based on equations (5) and (6)).

## 4.3 Conditions for (Global) Identification

Here we show that we do not need functional form assumptions to identify QTE as long as we have a reasonable instrument. We focus on the case of binary $D$, while the

appendix contains generalizations. The following analysis is all conditional on $X = x$, but we suppress this for ease of notation. Define $\mathcal{L}(x)$ as convex hull of the set of functions $\varphi$ mapping $d$ from $\{0,1\}$ to $(y : f_Y(y|d) > 0)$ such that $P(Y \leq \varphi(D,\tau)|Z)$ belongs to $[\tau - \delta, \tau + \delta]$ a.s. for $\delta > 0$.

Define the following function

$$\Pi_{\mathbf{z}}(\varphi, x) = \big(P[Y \leq \varphi(D)|z_1], P[Y \leq \varphi(D)|z_2]\big),$$

where $\mathbf{z} = (z_j, j = 1,2)$. Assuming relevant smoothness define

$$\begin{aligned} J_{\mathbf{z}}(\varphi, x) \equiv \frac{d}{d\varphi}\Pi_{\mathbf{z}}(\varphi) &\equiv \begin{bmatrix} f_Y(\varphi(0)|D=0,z_1)P[D=0|z_1] & f_Y(\varphi(1)|D=1,z_1)P[D=1|z_1] \\ f_Y(\varphi(0)|D=0,z_2)P[D=0|z_2] & f_Y(\varphi(1)|D=1,z_2)P[D=1|z_2] \end{bmatrix} \\ &\equiv \begin{bmatrix} f_{Y,D}(\varphi(0),0|z_1) & f_{Y,D}(\varphi(1),1|z_1) \\ f_{Y,D}(\varphi(0),0|z_2) & f_{Y,D}(\varphi(1),1|z_2) \end{bmatrix}. \end{aligned}$$

We will say that rank $J_{\mathbf{z}}(\varphi, x)$ is full w. pr. $> 0$ if with positive probability $\mathbf{Z} = (Z_1, Z_2)$ is such that rank $J_{\mathbf{Z}}(\varphi, x) = 2$, where $Z_1$ and $Z_2$ are independent replica of $Z$, given $X = x$.

**Theorem 3** *Suppose A1-A5 hold, and that $f_Y(y|d,z,x) > 0$ and finite over the range of $d \mapsto q(d,x,\tau)$. Then $d \mapsto q(d,x,\tau)$ is a unique solution of*

$$P(Y \leq q(d,x,\tau)|x,z) = \tau \text{ for } P\text{-a.e. } z, \text{ given } X = x, \tag{7}$$

*among $\mathcal{L}(x)$ if for any $\varphi \in \mathcal{L}(x)$ $J_{\mathbf{z}}(\varphi, x)$ is finite and has full rank w. pr. $> 0$.*

These conditions are akin to the identification of average treatment effects in Abadie (2001) or Das (2001). The difference is in the weighting by a density. The condition is easy to verify in many applications. For example, suppose $Z = 0$ or 1 as in the JTPA example discussed in Section 6. Then $\det J_{\mathbf{z}} \neq 0$ is equivalent to a *nonconstant likelihood ratio property*:

$$\frac{f_{Y,D}(\varphi(0),0|Z=1)}{f_{Y,D}(\varphi(1),1|Z=1)} \neq \frac{f_{Y,D}(\varphi(0),0|Z=0)}{f_{Y,D}(\varphi(1),1|Z=0)},$$

for any $\varphi \in \mathcal{L}(x)$. The instrument $Z$ should impact the joint distribution of $Y$ and $D$ at all relevant points. In the JTPA data $P[D = 1|Z = 0] = 0$, which means $f_{Y,D}(y,1|Z=0) = 0$ for any $y$, so the condition is always true as long as the left-hand-side is finite. In other cases, the condition is simply plausible.

# 5 Estimation

In this paper, it is natural to focus on estimating the basic linear model, which covers a wide area of applications. In this model a conditional $\tau$-quantile of the potential outcome is given by (or approximated by)

$$Q_{Y_d|X}(\tau) = d'\alpha_\tau + X'\beta_\tau, \tag{8}$$

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. This model is a specialization of A1, and is a foundation of quantile regression research (see e.g. Koenker and Hallock (2000) and Buchinsky (1998) for reviews).

Using Theorem 2 we offer the inverse quantile regression estimator as a finite-sample analog of the inverse quantile regression in the population. In the appendix, for completeness and comparisons, we also provide the results for the generalized empirical likelihood estimators. The presented estimator is perhaps the only practical estimator that can be applied to reasonably general cases. Other strategies such as method of moments or empirical likelihood are typically infeasible,[10] as explained below.

To state the idea clearly, first suppose we have no covariates or simply treat covariates as the part of vector $d$ above. In this case, a simple analog of the population inverse quantile regression is as follows:

> Find $\widehat{\alpha}$ by minimizing a norm of $\widehat{\gamma}[\alpha]$ over $\alpha$ subject to $\widehat{\gamma}[\alpha]$ solving the quantile regression of $Y - D'\alpha$ on $Z$: $\widehat{\gamma}[\alpha] = \operatorname{argmin}_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - D_i'\alpha - Z_i'\gamma)$.

Now suppose we have covariates $X_t$. Then the procedure can be modified as follows:

> Find $\widehat{\alpha}$ by minimizing a norm of $\widehat{\gamma}[\alpha]$ over $\alpha$ subject to $(\widehat{\gamma}[\alpha], \widehat{\beta}[\alpha])$ solving the quantile regression of $Y - D'\alpha$ on $Z$ and $X$:
> $(\widehat{\gamma}[\alpha], \widehat{\beta}[\alpha]) = \operatorname{argmin}_{\gamma,\beta} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - D_i'\alpha - X_i'\beta - Z_i'\gamma)$.

The estimate of $\beta$ can be obtained as a usual quantile regression of $Y - d'\widehat{\alpha}$ on $X$.

In order to improve efficiency, we allow the observations to be weighted differently and allow for estimated instruments. Define the weighted quantile regression objective function:

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

$\Phi_i \equiv \Phi(X_i, Z_i)$, where $\Phi$ is a smooth $r \times 1$ vector function of instruments,
$\widehat{\Phi}_i \equiv \widehat{\Phi}(X_i, Z_i)$, where $\widehat{\Phi}$ is a smooth consistent estimate of $\Phi$, satisfying R5,
$V_i \equiv V(X_i, Z_i) > 0$, where $V$ is a smooth weight function,
$\widehat{V}_i \equiv \widehat{V}(X_i, Z_i) > 0$, where $\widehat{V}$ is a smooth consistent estimate of $V$, satisfying R5.

Note that one may *simply* set $\Phi_i = Z_i$ or $V_i = 1$, which will give us the simpler versions above. Efficient estimation is described in Corollary 1. We can use a wide variety of nonparametric estimators and parametric approximations of $V$ and $\Phi$, satisfying a standard smoothness condition, stated as a technical assumption R5 in appendix G. We also assume $(\alpha_\tau, \beta_\tau)$ belongs to a compact set $\mathcal{A} \times \mathcal{B}$. Other technical conditions

[10] Note, however, that EL has many good properties and purportedly performs well in finite samples. A possible feasible approach is as follows. In the first stage, IQR estimates of the parameters could be obtained. Then, in the second stage, the estimates could be recomputed using EL limiting the domain to a neighborhood around the estimates obtained in the first stage.

are stated as assumptions R1-R5 in the appendix. Most of them are standard in the quantile regression literature.

Now let's formally define the estimation procedure as follows:

$$\widehat{\alpha} = \arg \inf_{\alpha \in \mathcal{A}} \gamma[\alpha]' \widehat{A} \gamma[\alpha], \text{ such that} \tag{9}$$

$$(\widehat{\beta}[\alpha], \widehat{\gamma}[\alpha]) = \underset{(\beta,\gamma) \in \mathcal{B} \times \mathcal{G}}{\arg\inf} Q_n(\alpha, \beta, \gamma). \tag{10}$$

where $\mathcal{G} = [-\delta, \delta]^r$ for $\delta > 0$ and $\widehat{A} \xrightarrow{p} A$ is a positive definite matrix. A final estimate of $\beta_\tau$ is obtained as

$$\widehat{\beta} = \arg \inf_{\beta} Q_n(\widehat{\alpha}, \beta, 0). \tag{11}$$

Equations (9) -(10) are a finite sample *inverse or instrumental quantile regression* (IQR). (10) is the quantile regression step, and (9) is the "inverse" step.

This formulation allows one to effectively reduce the dimensionality of a potentially difficult optimization problem to the dimension of $\alpha$. In GMM, the objective function is highly multi-modal and has zero derivative almost everywhere, implying the need to perform a grid search over a subset of $\mathbb{R}^K$ where $K = \dim(x) + \dim(\alpha)$ (e.g. in Example 2 of the next section $\dim(x) + \dim(\alpha) = 16$). Such an estimator is infeasible, except perhaps when $\dim(x) = 2$ or 3. In contrast, a simple implementation of inverse quantile regression would require only a grid search over a subset of $\mathbb{R}^{\dim(\alpha)}$. The regression quantile steps are solved as fast as OLS by interior point methods combined with preprocessing, see Portnoy and Koenker (1997). The computations may be improved further by employing parametric programming. In this approach the quantile regression in (10) is initially solved for some $\alpha_0$, then one solves for $\widehat{\beta}[\alpha]$ and $\widehat{\gamma}[\alpha]$ for nearby $\alpha$ using a standard sensitivity analysis.

We now turn to the theoretical properties of the estimator. In the appendix we also study the properties of the generalized empirical likelihood estimators.

**Theorem 4** *Under assumptions R1-R6 listed in the appendix*

$$\sqrt{n}(\widehat{\alpha} - \alpha_\tau) \xrightarrow{d} K\, \mathcal{N}(0, S),$$
$$\sqrt{n}(\widehat{\beta} - \beta_\tau) \xrightarrow{d} L\, \mathcal{N}(0, S),$$

*where convergence is joint, and $\mathcal{N}(0, S)$ is normal vector with mean 0 and variance $S = \tau(1-\tau)E\Psi\Psi'$, where $K \equiv (J_\alpha' H J_\alpha)^{-1} J_\alpha' H$, $H \equiv \bar{J}_\gamma' A \bar{J}_\gamma$, $L \equiv J_\beta^{-1}[I_k : 0]M$, $M \equiv I - J_\alpha K$, $\Psi \equiv V \cdot [X' : \Phi']'$, $J_\alpha \equiv E[f_\epsilon(0|X, D, Z)\Psi D']$ and $J_\beta = E[f_\epsilon(0|X, Z)XX']$, where $\epsilon = Y - D\alpha_\tau - X'\beta_\tau$. Finally, $J_\theta \equiv E[1/V \cdot f_\epsilon(0|X, Z)\Psi\Psi']$, where $[\bar{J}_\beta' : \bar{J}_\gamma']'$ is the partition of $J_\theta^{-1}$, such that $\bar{J}_\beta$ is a $k \times (k+r)$ matrix and $\bar{J}_\gamma$ is a $r \times (k+r)$ matrix.*

**Corollary 1** *Generally, when the number of instruments $\Phi$ equals that of enodogenous regressors $D$, the joint asymptotic variance of $\widehat{\alpha}$ and $\widehat{\beta}$ has a simple form*

$$J^{-1} S J^{-1}, \tag{12}$$

*for* $J = E[f_\epsilon(0|X, D, Z)\Psi[D' : X']]$. *Choice of* $A$ *is irrelevant. Further, if* $\Phi = \Phi^* \equiv E[D \cdot v|Z, X]/V^*$, *where* $v \equiv f_\epsilon[0|D, Z, X]$, *and* $V = V^* \equiv f_\epsilon(0|X, Z)$, *the asymptotic variance of* $\widehat{\alpha}$ *and* $\widehat{\beta}$, *simplifies to*

$$\tau(1-\tau)E[\Psi^*\Psi^{*\prime}]^{-1}, \text{ where } \Psi^* = V^* \cdot [X', \Phi^{*\prime}]'. \tag{13}$$

**Corollary 2** *If the number of instruments* $\Phi$ *is larger than the number of enodogenous regressors* $D$, *choice of weighting matrix* $A$ *matters. An optimal choice of* $A$ *is given by* $A \equiv [J_\theta^{-1}]_{22} = (\bar{J}_\gamma J_\theta \bar{J}_\gamma)^{-1}$, *an* $r \times r$ *matrix. In this case the joint asymptotic variance of* $\widehat{\alpha}$ *and* $\widehat{\beta}$ *equals* $NSN'$, *where* $N = (J' J_\theta^{-1} J)^{-1} J' J_\theta^{-1}$. *If in addition,* $V = V^*$, *the joint variance equals* $(JS^{-1}J)^{-1}$.

(13) is the efficiency bound for the GMM estimators under conditional moment restrictions as in Theorem 1. This is the efficiency bound in the sense of Amemiya (1977), Chamberlain (1987), or Newey (1990). See also Newey and Powell (1990).

Corollary 1 suggests a reasonable approach to estimation and inference.

First of all, in section 6, we used a simplest and most transparent strategy, projecting $D$ on $Z$ with OLS to form the instrument $\Phi$, and setting $V_i = 1$. We used methods described in Koenker (1994) to obtain the estimates of standard errors based on the simple formula (12). Powell (1986)'s methods also apply without modification.

Generally, we can use many established methods to either approximate or implement exactly the optimal procedure. We can estimate $f_\epsilon(0|\cdot)$ by the kernel methods described in Andrews (1994) or quantile regression differencing as in Koenker (1994), and $E[Dv|Z, X]$ can be estimated using series estimation (e.g. OLS of $Dv$ on $Z, X$ and their powers), as in Newey (1997) and Andrews and Whang (1990). Assumption R5 allows for a wide variety of nonparametric and parametric estimation procedures – Andrews (1994) discusses a number of them.

In practice, it is often reasonable to use parametric approximations, cf. Amemiya (1975). For example, we may use conditional normality for $f_\epsilon(\cdot|\cdot)$ to get an approximation of the standard errors and optimal weights in the quantile regressions above. On the other hand, $E[Dv|Z, X]$ can be approximated by polynomial functions in $Z, X$ and estimated by OLS. As long as approximation of the optimal procedure is accurate, the standard errors, based on (13) or on a more robust formula (12), will also be accurate.

When there is a compelling reason to use instruments $\Phi$ of dimension larger than that of $D$, Corollary 2 describes the choice of the weighting matrix $A$ that simplifies the asymptotic variance.

The documented computer programs in programming languages R (free software available from www.r-project.org) and Matlab that implement the estimation and inference are available from the authors. The programs implement both the optimal and sub-optimal instrument cases.

# 6 Empirical Applications

This section presents the empirical illustration to the economic models presented in section 3. The first example is a market demand model, and the second example is an evaluation of a job training program.

## 6.1 Demand for Fish

In this section, we present estimates of demand elasticities which may potentially vary with the level of demand, $\tau$. The data contain observations on price and quantity of fresh whiting sold in the Fulton fish market in New York over the five month period from December 2, 1991 to May 8, 1992. These data were used previously in Graddy (1995) to test for imperfect competition in the market and later in Angrist, Graddy, and Imbens (2000) to illustrate use of the conventional IV estimator as a weighted average of heterogeneous demands. The price and quantity data are aggregated by day, with the price measured as the average daily price for the dealer and the quantity as the total amount of fish sold that day. The data also contain information on the day of the week of each observation and variables indicating weather conditions at sea, which are used as instruments to identify the demand equation. The total sample consists of 111 observations for the days in which the market was open.

The demand function we estimate takes a standard Cobb-Douglas form:

$$Q_{\ln(Y_p)|X}(\tau) = \alpha_\tau \ln p + X'\beta_\tau,$$

where $Y_p$ is demand when price is $p$. The elasticity $\alpha_\tau$ varies across the quantiles $\tau$ of demand level. Following discussion in section 3, this is a demand model with non-separable error and random elasticity.

The top two panels of Figure 1 provide the estimates of elasticities obtained by IQR of $\ln(Y)$ on $\ln(P)$ using wind speed as the instrumental variable, while the lower panels depict standard quantile regression (QR) estimates. The shaded region around the point estimates represents the 80 percent confidence interval. While the reported estimates are for a model without covariates, the estimated elasticities are not sensitive to the inclusion of dummy variables for the days of the week or other covariates.

The price effect on quantities sold, as estimated by QR, appears to be approximately constant across the entire range of quantiles. The magnitudes of the effects are also quite small, in all cases much less than unity. IQR estimates, on the other hand, range from -2 to -.5, with the median elasticity of -1, indicating variation of elasticities with the level of demand. Except at high quantiles, the IQR elasticities are uniformly greater in magnitude than the price effects predicted by QR. This is clearly shown in the demand curves plotted in Figure 2. Note that the interpretation of IQR and QR estimates is very different. IQR estimates a (causal) demand model, while QR estimates the conditional quantiles of the equilibrium quantity as a function of the equilibrium price.

The IQR estimates of the demand elasticities $\alpha_\tau$ illustrate heterogeneity across the demand levels. The results indicate that demand elasticity is quite high in magnitude

at low quantiles, but is decreasing in the quantile index. While there are many possible explanations for this demand behavior, it does cast doubt on the hypothesis that the aggregate demand in this market is a sum of the demand curves of numerous identical price-taking agents who randomly arrive at the market. The estimates may also suggest that a single statistic may be insufficient to truly capture the demand function variety.

## 6.2 Evaluation of a JTPA Program

The impact of job training programs on the earnings of participants, especially those with low income, is of great interest to economists, but evaluating the causal effect of training programs on earnings is difficult due to the self-selection of treatment status. However, data available from a randomized training experiment conducted under the Job Training Partnership Act (JTPA) provides a mechanism for addressing this issue. In the experiment, people were randomly assigned the offer of JTPA training services, but because people were able to refuse to participate, the actual treatment receipt was self-selected. Of those offered treatment, only 60 percent participated in the training. There was also a small number of individuals from the control group who received training. The random assignment of the training offer provides a plausible instrument for a person's actual training status. Adadie et. al. (2000) and Heckman and Smith (1997) provide detailed information regarding data collection procedures and institutional details of the JTPA. We limit the analysis to the adult males.

To capture the effects of training on earnings, we estimate a linear model:

$$Q_{Y_d|X}(\tau) = d\alpha_\tau + X'\beta_\tau,$$

where $d$ indicates training status and is instrumented for by assignment to the control group, the potential outcomes $Y_d$ are earnings, and $X$ is a vector of covariates. The data consist of 5,102 observations with data on earnings, training and assignment status, and other individual characteristics. Earnings are measured as total earnings over the 30 month period following the assignment into the treatment or control group. We also include dummies for black and Hispanic persons, a dummy indicating high-school graduates and GED holders, five age-group dummies, a marital status dummy, a dummy indicating whether the applicant worked 12 or more weeks in the 12 months prior to the assignment, a dummy signifying that earnings data are from a second follow-up survey, and dummies for the recommended service strategy.[11]

Results for standard quantile regression are illustrated in Figure 4 and IQR estimates in Figure 3. The shaded region represents the 90 percent confidence interval for the point estimates. The first panel in each figure shows the estimated impact of the participation in the training program across various quantiles. A quick comparison of the two sets of results shows that the standard quantile regression estimates of the *statistical* impacts of training are well above the treatment effect. The quantile regression estimates are uniformly larger than the IQR estimates, and in many cases the difference is quite substantial. This difference is perhaps most important in the

[11] The recommended service strategy was broken into three categories: classroom training, on-the-job training and/or job search assistance, and other forms of training.

low to middle quantiles where the conventional quantile regression estimates indicate a relatively large statistical impact of training on the earnings of participants.

The differences in the standard QR and the IQR estimates, as well as the distributional impacts of the program, are made even more apparent when one considers the impact of training in percentage terms.[12] Quantile regression estimates indicate large percentage impacts, especially in the lower quantiles. The IQR estimates, on the other hand, indicate that the percentage *causal* impact of the training program is relatively constant and low, between 5 and 10 percent, along the whole distribution. This is interesting since the supposed intent of job training programs is to raise the incomes of low income individuals. However, we observe that the impacts were actually the greatest for the upper quantiles.

Coefficient estimates for several of the covariates are also included in the figures. None of the results are particularly surprising. Being Hispanic has no significant impact on potential earnings at any point in the distribution, while at medium and high quantiles, blacks earn significantly less than whites. We also see that education, as measured by high school graduation or having a GED, has a positive impact along almost the entire distribution, with the impact growing monotonically in the quantile index. This pattern is also observed for the marriage effect, which tapers off in the highest quantiles. The effect of having worked little in the previous year runs in almost exactly the opposite direction, impacting earnings negatively at all quantiles and decreasing earnings substantially in the upper tail of the distribution.

We next compare our results with those in Abadie et al.(2001). Since identification in two models comes through different assumptions *and* the estimated treatment effects are for different populations (the Abadie et al's model is for the sub-population of LATE-compliers), the estimation results need not agree. However, the JTPA is an example where *both* sets of assumptions appear to hold. Independence and monotonicity are almost certainly satisfied, and it seems reasonable that, relative to others with similar characteristics, similarity assumption is also fulfilled.[13] Under these conditions, the models *overlap* and the results should indeed be comparable *if* the subpopulation of LATE-compliers is *representative* of the entire population. This appears to be the case in the present example.

Lastly, consider the results of Heckman and Smith (1997). The model of Heckman and Smith (1997) did not incorporate endogeneity (it had a different point). Thus their results correspond to our QR results (fig 4), and differ from the IQR results (fig 3).

## 6.3 Numerical Performance

The objective functions for selected quantiles from Examples 1 and 2 are graphed in Figure 5. The upper three panels in the figure illustrate the objective functions from the fish example, while the lower panels correspond to the JTPA example. The

[12]The percentage impact of training is calculated for both whites, Percentage Impact I, and black, Percentage Impact II. Percentages are calculated for married high-school graduates aged 30 to 35.

[13]Note that conditioning on covariates weakens the required similarity condition requiring that similarity only hold for people with the same covariate values.

objective functions are very well-behaved, especially in the JTPA example. Each of the objective functions from the fish example does have many local minima, which is attributable to the small sample size. However, in all cases, the functions have an obvious unique global minimum.

# 7 Conclusion and Future Research

This paper offered two contributions. First, it proposed a model of quantile treatment effects which allows for treatment endogeneity. The model exploits the similarity as a main identification restriction. The resulting model differs from both Heckman's non-parametric selection model and Imbens and Angrist's LATE model. From this model we derive a Wald IV estimating equation. We show that the model does not require functional form assumptions for identification. Second, we characterized the quantile treatment function as solving an inverse quantile regression problem and suggested its finite-sample analog as a practical estimator. This estimator, unlike generalized method-of-moments, can be easily computed by solving a series of conventional quantile regressions, and does not require grid searches over high-dimensional parameter sets. A properly weighted version of it is also efficient. We applied this estimator to characterize quantile treatment effects in a market demand model and evaluation of a job training program.

An important feature of the proposed model is that even though one may not be interested in quantile treatment effects, one may still have to estimate them. Indeed, the average treatment effects can not be estimated by conventional IV methods, as shown in example 5.1.[14] Instead, quantile treatment effects have to be estimated first and then integrated over the quantile index. Alternatively, one may estimate only the median treatment effects, using the proposed model and estimator.

In companion works, we consider a number of directions. In a joint work with Whitney Newey and Guido Imbens, we explore fully non-parametric estimation, which poses an interesting problem. Other research directions are also considered. For example, an important research question is how to estimate policy-relevant treatment effects in an expected utility framework, given particular social loss functions, known program costs, and effects on choice probabilities (cf. Heckman et al (2001)).

[14] Note that the local average treatment effect may still be identified without estimating the QTE.

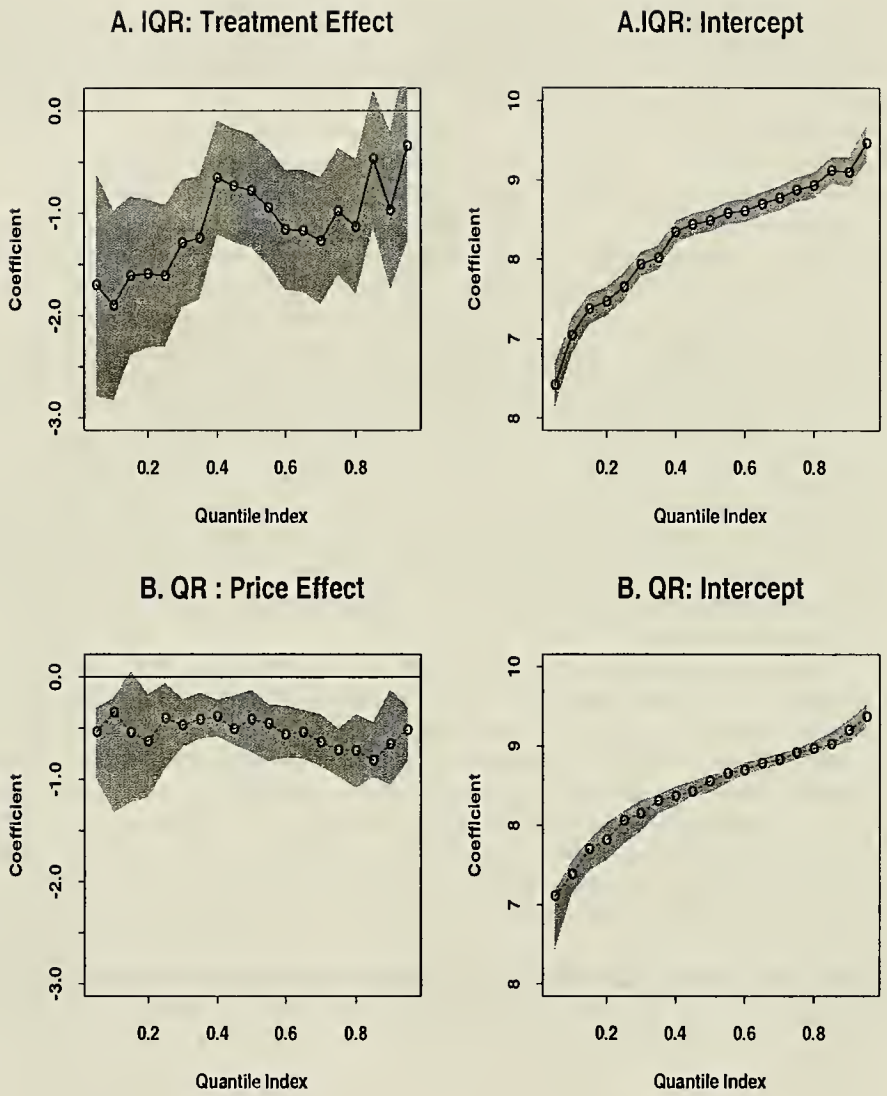


Figure 1: Inverse Quantile Regression and Quantile Regression Results for fish data. The quantile treatment effect, estimated by IQR, is the elasticity of the $\tau$-th quantile demand curve. It tends to be much higher than the "price effect" on the $\tau$-quantiles of quantities sold, estimated by QR.

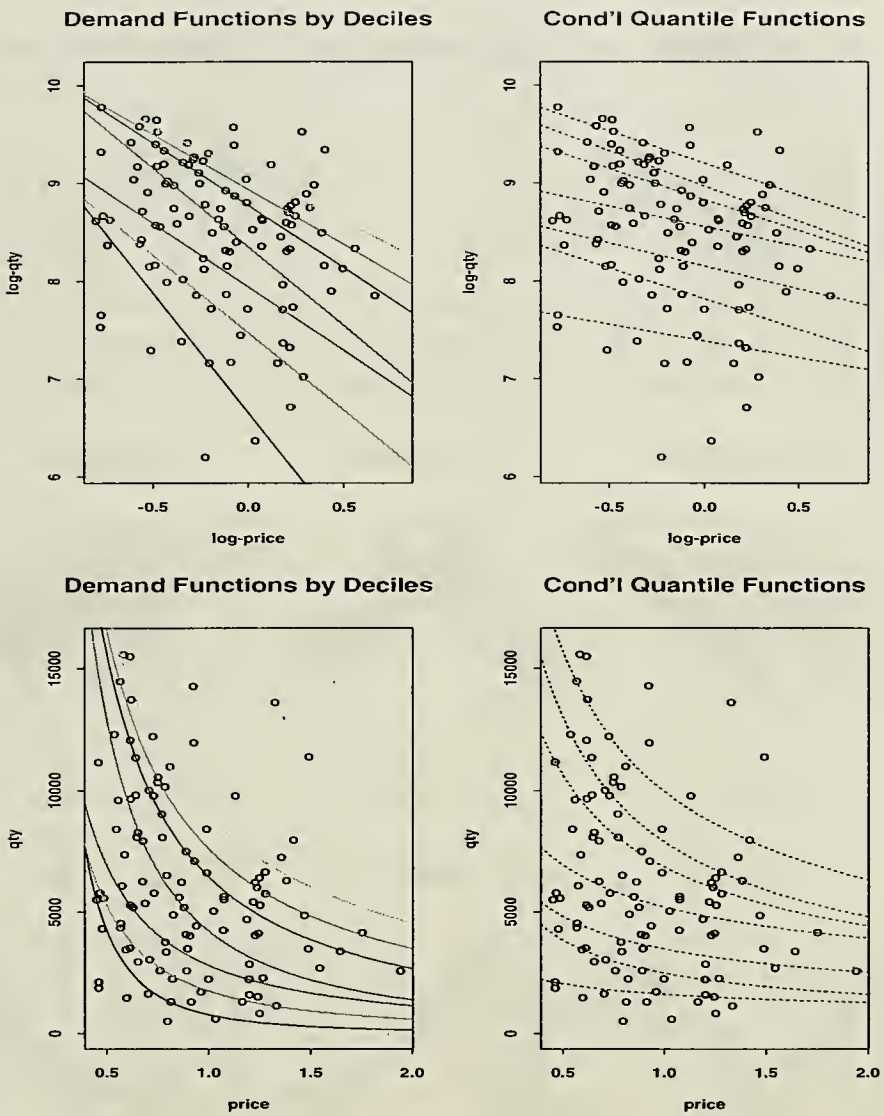


Figure 2: *Left Column*: The estimated by IQR demand curves, indexed by the quantile index (.1, .2, .3, .5, .7, .8. and .9.). The top display is in log-price-log-quantity space. The bottom display is in the original space. *Right Column*: The estimated conditional quantile curves of fish quantity sold as a function of price. The top display is in log-price-log-quantity space. The bottom display is in the original space.

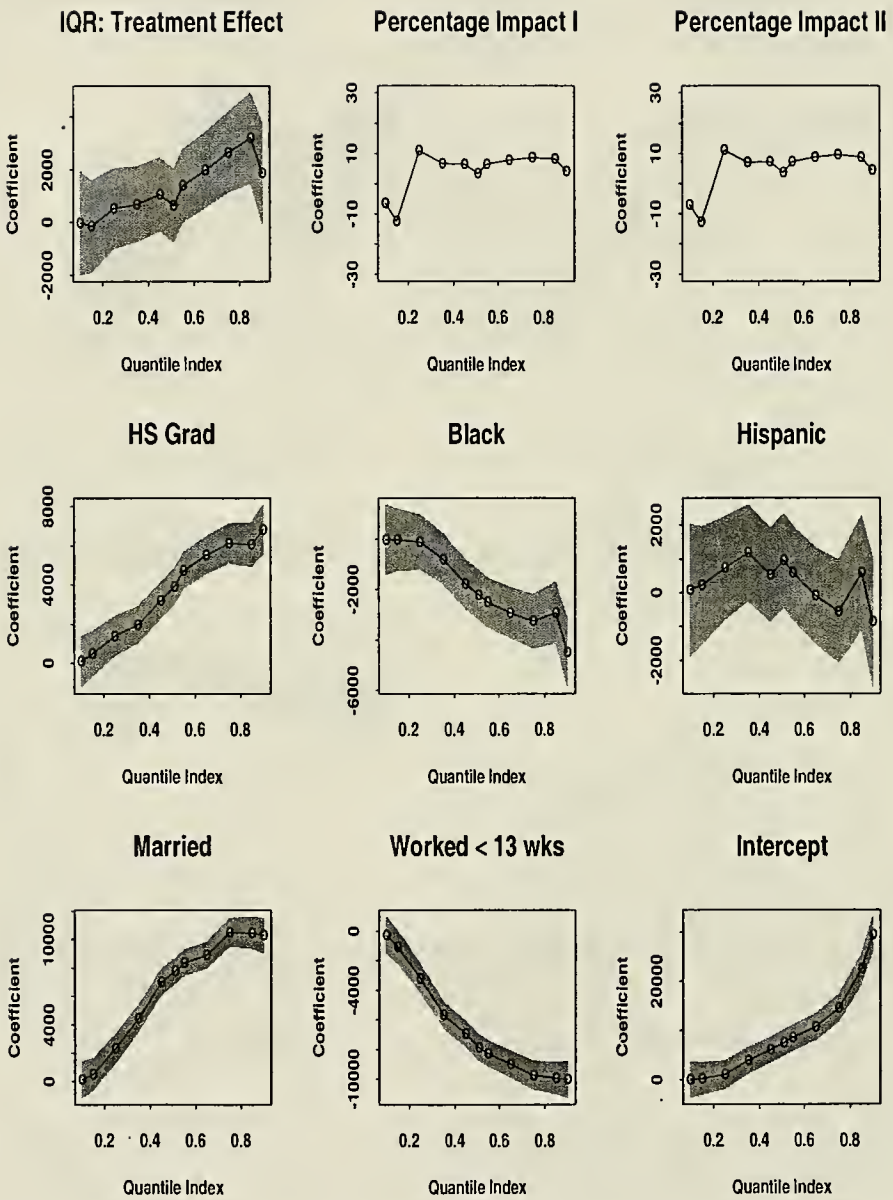


Figure 3: Inverse Quantile Regression results on JTPA data.

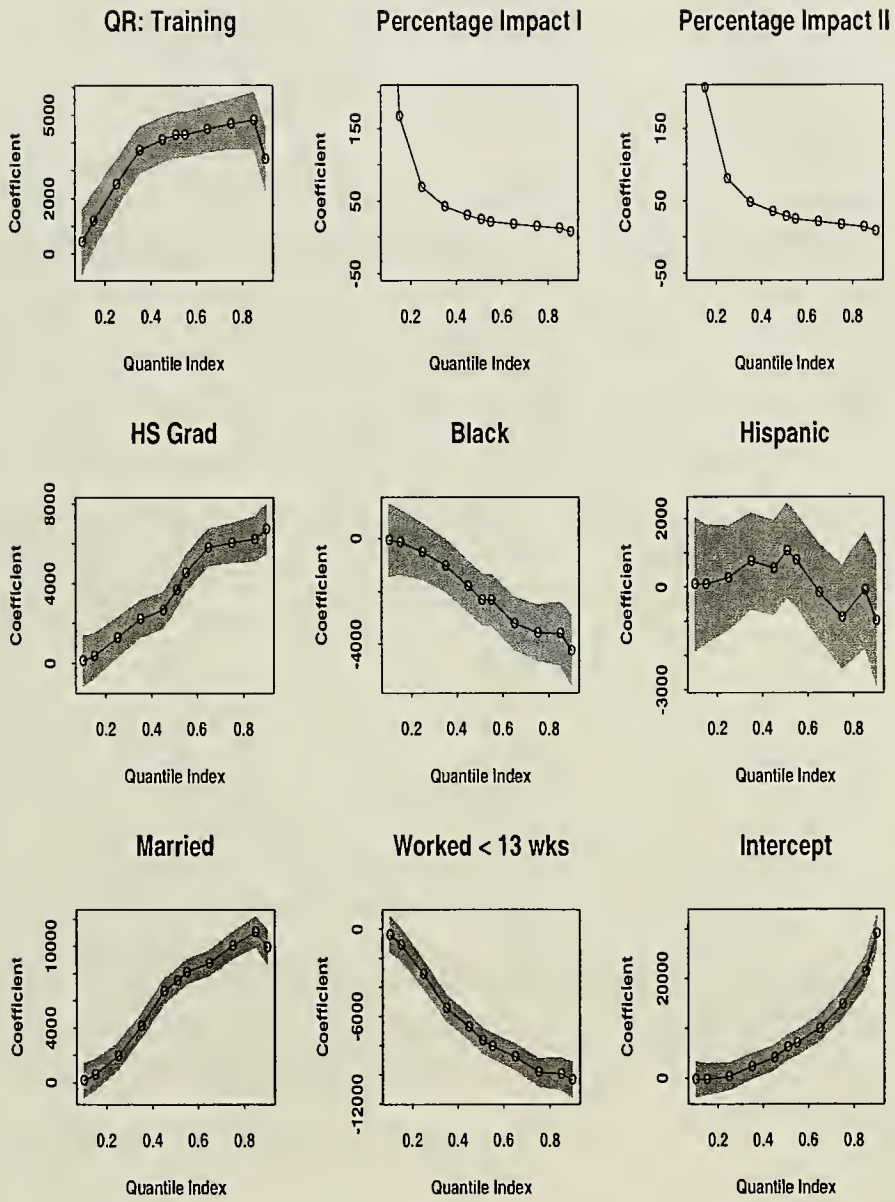


Figure 4: Quantile Regression results on JTPA data.

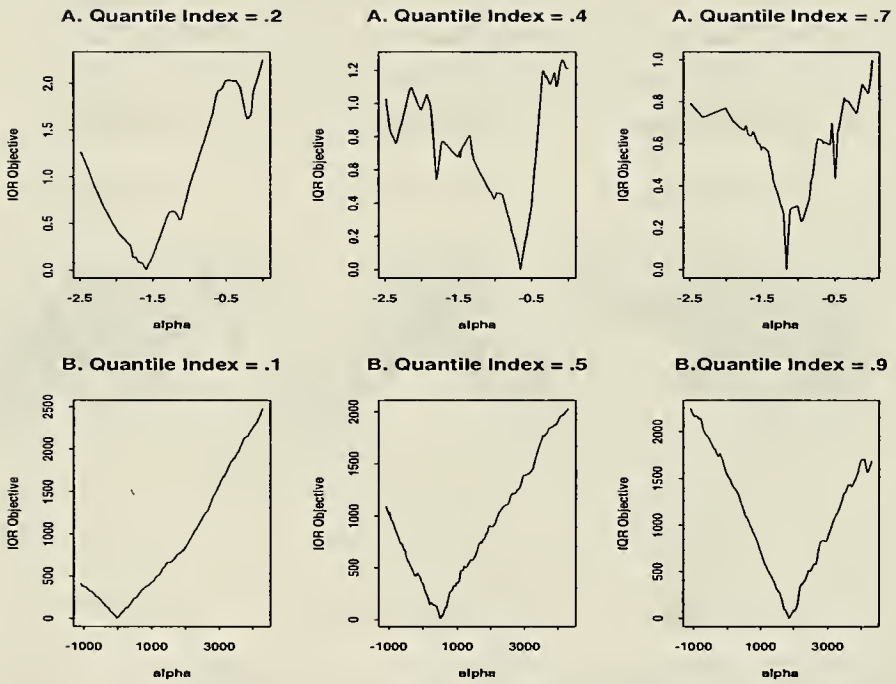


Figure 5: A. IQR objective functions for the fish example. B. IQR objective functions for the JTPA example.

# A Definitions and Lemmas

We use the following empirical processes in the sequel, for $W \equiv (Y, D, X, Z)$

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

For example, if $\widehat{f}$ is estimated function, $\mathbb{G}_n f(W)$ means: $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(W_i) - Ef(W_i))_{f=\widehat{f}}$. Outer and inner probabilities, $P^*$ and $P_*$ are defined as in van der Vaart (1998). In this paper $\xrightarrow{p}$ means convergence in (outer) probability, and $\xrightarrow{d}$ means convergence in distribution. We will say that process $\{l \mapsto v_n(l), l \in \mathcal{L}\}$ is *stochastically equi-continuous* (s.e.) in $\ell^\infty(\mathcal{L})$ if for each $\epsilon > 0$ and $\eta > 0$, there is $\delta > 0$ :

$$\limsup_{n\to\infty} P^*\Big(\sup_{\rho(l,l')<\delta} |v_n(l) - v_n(l')| > \eta\Big) < \epsilon$$

for some pseudo-metric $\rho$ on $\mathcal{L}$, such that $(\mathcal{L}, \rho)$ is totally bounded pseudo-metric space.

The following results are from Knight (1999). They allow general discontinuities and $\bar{\mathbb{R}}$ - valued objective functions. Related literature is Rockafellar and Wets (1998).

**Lemma A.1 (Geyer's Lemma)** *Suppose $\{Q_n\}$ is a sequence of lower-semi-continuous convex $\bar{\mathbb{R}}$-valued random functions, defined on $\mathbb{R}^d$, and let $\mathcal{D}$ be a countable dense subset of $\mathbb{R}^d$. If $Q_n$ converges to $Q_\infty$ in $\bar{\mathbb{R}}$ on $\mathcal{D}$, in finite dimensional sense, where $Q_\infty$ is lsc convex and finite on an open non-empty set a.s., then*

$$\operatorname*{arginf}_{z\in\mathbb{R}^d} Q_n(z) \xrightarrow{d} \operatorname*{arginf}_{z\in\mathbb{R}^d} Q_\infty(z),$$

*provided the latter is uniquely defined a.s. in $\mathbb{R}^d$.*

**Lemma A.2 (Approximate Argmins)** *Suppose*

*i. $Z_n$ is s.t. $Q_n(Z_n) \le \inf_{z\in\mathbb{R}^d} Q_n(z) + \epsilon_n$, $\epsilon_n \searrow 0$; $Z_n = O_p(1)$.*

*ii. $Z_\infty \equiv \operatorname{argmin}_{z\in\mathbb{R}^d} Q_\infty(z)$ is uniquely defined in $\mathbb{R}^d$ a.s.*

*iii. $Q_n(\cdot) \Rightarrow Q_\infty(\cdot)$ in $\ell^\infty(K)$ over any compacts $K$, where $Q_\infty$ is continuous. Then $Z_n \xrightarrow{d} Z_\infty$.*

# B Two-Stage Quantile Regression: Inconsistency when QTE varies with $\tau$

The model proposed by Amemiya consists of two equations

$$\begin{aligned} &(i) \quad Y = D'\delta + X'\beta + U, \\ &(ii) \quad D = Z'\gamma + V, \end{aligned} \tag{14}$$

where $D$ is an endogenous vector, i.e. $D$ depends on the real-valued $U$, $X$ is a vector of exogenous or predetermined variables, $U$ and $V$ are independent of $X$ and $Z$, and $U$ and $V$ are jointly symmetric and absolutely continuous.

Parameters $(\beta, \delta)$ can be estimated by two-stage LAD, Amemiya (1982), by projecting $D$ on $Z$ to get $\widehat{\gamma}$, and using the median regression of $Y$ on $(Z'\widehat{\gamma}, X)$. Another valid second stage is the quantile regression, cf. Chen and Portnoy (1996), known as two-stage quantile regression (2SQR).

Model (14) imposes the *constant* QTE. The treatment variable, if assigned externally, shifts the location of the outcome variable, but does not affect the scale or shape of its distribution. This severely limits the treatment variety.

The assumptions of constant QTE is *crucial for validity* of 2SQR. If QTE effects are non-constant, 2SQR does not consistently estimate them. Unfortunately a rather extensive empirical literature has used 2SQR to estimate the *non-constant* QTE.

To explain the inconsistency, it suffices to consider an example with no endogeneity. Suppose for some increasing one-to-one map $\delta(\cdot)$:

$$\begin{aligned} &Y = D\delta(U), U \stackrel{d}{=} U(0,1), \\ &D = Z'\gamma + V, \\ &Z, U, V \text{ are mutually independent.} \end{aligned} \tag{15}$$

Assume that $Y, V$ have densities conditional on $Z$ and that $D > 0$. To pin down $\gamma$ we may assume $E[V] = 0$. It is sufficient to show that $\delta(\tau)$ is generally not the optimum in the population 2SQR problem. That is, there is no $\alpha$ such that

$$\begin{aligned} &\text{i. } E\left(1(Y \leq \alpha + \delta(\tau)Z'\gamma) - \tau\right) = 0, \\ &\text{ii. } E\left(1(Y \leq \alpha + \delta(\tau)Z'\gamma) - \tau\right) Z'\gamma = 0. \end{aligned} \tag{16}$$

By definition $\{Y \leq \alpha + \delta(\tau)Z'\gamma\} \equiv \{V\delta(U) + Z'\gamma(\delta(U) - \delta(\tau)) \leq \alpha\}$. Equation **i.** implies:

$$\alpha = Q_M(\tau), \text{ where } M \equiv V\delta(U) + Z'\gamma \cdot (\delta(U) - \delta(\tau)),$$

thus it remains to check whether

$$E\left(1(M \leq Q_M(\tau)) - \tau\right) Z'\gamma \stackrel{?}{=} 0. \tag{17}$$

Generally (17) is false. Equation (17) holds when $M$ is $\tau$-quantile independent of $Z$:

$$Q_{M|Z}(\tau) = Q_M(\tau) \;\; P \text{ a.e.} \Leftrightarrow P[M \leq Q_M(\tau)|Z] = \tau \;\; P \text{ a.e.}$$

This necessarily happens when $\delta(\tau) = \delta$, the constant treatment effect case, or, for example, when $\tau = 1/2$ and $M$ is symmetric given $Z$, as in Amemiya (1982).

Simple examples suffice to confirm that (17) indeed fails. The first example involves no endogeneity:

- $V \stackrel{d}{=} 5 + N(0,1)$, truncated to be positive,
- $Z'\gamma \stackrel{d}{=} 5 + N(0,1)$, truncated to be positive,
- $\delta(U) \stackrel{d}{=} N(0,1)/100$, $D = Z'\gamma + V$, $Y = D \cdot \delta(U)$.

The following computation uses monte-carlo integration using 500,000 simulations.

- $E\left(1(M \leq Q_M(\tau)) - \tau\right) Z'\gamma = 0.34$, for $\tau = .7$ with s.e. of .003

The second example involves endogeneity:

- $V \stackrel{d}{=} 5 + N(0,1)$, truncated to be positive,

- $Z'\gamma \stackrel{d}{=} 5 + N(0,1)$, truncated to be positive,
- $\delta(U) = V/100$, $D = Z'\gamma + V$, $Y = D \cdot \delta(U)$.

The following computation uses monte-carlo integration using 500,000 simulations.

- $E\left(1(M \leq Q_M(\tau)) - \tau\right) Z'\gamma = 0.38$, for $\tau = .7$ with s.e. .003

# C Comparison with Abadie et al's Model

In Abadie, Angrist, and Imbens (2001), henceforth AAI, the treatment variable $D$ and the instrument $Z$ are both *binary*. The binary nature of $D$ is critical, and extensions to the general case are not known. The general, non-binary case is clearly important. This approach, however, is well suited to many experimental studies.

The potential outcomes $Y_d$ are indexed by the treatment status $d \in \{0,1\}$, and the potential treatments $D_z$ are indexed by the instrument status $z \in \{0,1\}$. The realized outcome is $Y \equiv Y_D$, while the realized treatment is $D \equiv D_Z$. AAI impose the independence condition

$$(Y_0, Y_1, D_1, D_0) \text{ are independent of } Z, \tag{18}$$

and the monotonicity assumption:

$$D_1 \geq D_0 \text{ a.s.} \tag{19}$$

For example, let

$$D_Z = 1(\varphi(Z) > V), \quad \text{where } Z \text{ is independent of } V, \tag{20}$$

i.e. $D_0 = 1(\varphi(0) + V)$ and $D_1 = 1(\varphi(1) + V)$. $V$ may depend on the potential outcomes $Y_0$ and $Y_1$. Model (20) along with (18) satisfies the independence and monotonicity assumption. (Vytlacil (2001) also shows the converse is true as well, in the sense of distribution equivalence.)

Exploiting that $D$ and $Z$ are binary, independence, and monotonicity, AAI show that in the *subpopulation of compliers*, where $D_1 > D_0$, the realized treatment $D$ is independent of potential outcomes:

$$(Y_1, Y_0) \text{ are independent of } D \,|X, D_1 > D_0.$$

The compliers are manipulated by the instrument and, therefore, randomly receive a treatment status. That is, the treatment status is given to them independently of their potential responses $Y_0$ and $Y_1$, conditional on observed covariates $X$. That is, endogeneity is *removed* in this subpopulation.

Let $Q_{Y|X,C}(\tau)$ denote the $\tau$-quantile for the population of compliers conditional on $(X, D_1 > D_0)$. The quantile treatment effect $\delta(\tau)$ is a difference in the conditional $\tau$-quantiles of $Y_1$ and $Y_0$ for compliers:

$$Q_{Y|X,C}(\tau) = \delta(\tau)D + X'\beta(\tau).$$

AAI suggest an *ingenious* weighting scheme that "finds compliers" (compliers are unobserved) and interpret their estimator as a re-weighted Koenker and Bassett's quantile regression.

The main differences with our approach are the following.

*First*, our model's QTE is defined relative to the population, while AAI's QTE is defined relative to compliers. The compliers may substantially differ from the entire population. For example, in Angrist and Krueger (1992), the compliers are those whose education level is affected by their birthdate. Thus, the 90% QTE in AAI's model may differ substantially from

the 90% QTE in our model. The QTE's of two models may coincide if compliers in AAI's model are representative of the population *and* other assumptions overlap as well.

*Second*, AAI's model applies to binary cases only, while the present approach applies to general cases. *Third*, the estimation procedure are fundamentally different. *Fourth*, we use similarity or rank invariance to identify QTE while AAI use the monotonicity and stronger independence conditions.

# D Proof of Theorem 1

**Part (a).** Conditioning on $X = x$ is suppressed. For $P$-a.e. value $z$ of $Z$

$$\begin{aligned} P\left[Y \leq q\left[D,\tau\right]|Z=z\right] & \overset{(1)}{=} P\left[q\left[D,U_D\right] \leq q\left[D,\tau\right]|Z=z\right] \\ & \overset{(2)}{=} P\left[U_D \leq \tau | Z=z\right], \\ & \overset{(3)}{=} \int P\left[U_D \leq \tau | Z=z, V=v\right] dP\left[V=v|Z=z\right] \\ & \overset{(4)}{=} \int P\left[U_{\delta(z,v)} \leq \tau | Z=z, V=v\right] dP\left[V=v|Z=z\right] \\ & \overset{(5)}{=} \int P\left[U_o \leq \tau | Z=z, V=v\right] dP\left[V=v|Z=z\right] \\ & \overset{(6)}{=} P\left[U_o \leq \tau | Z=z\right] \\ & \overset{(7)}{=} \tau. \end{aligned} \tag{21}$$

Equality (1) is by A1 and A5. Equality (3) is by definition. Equality (4) is by A2. Equality (5) is by the similarity assumption A4: for each $d$, conditional on $(V = v, X = x, Z = z)$

$$U_{\delta(z,v)} \text{ equals in distribution to } U_o.$$

Equality (6) is by definition and equality (7) is by A3. Note that equality (2) is immediate when $\tau \mapsto q(d,\tau)$ is continuous, since we assumed that $\tau \mapsto q(d,\tau)$ is strictly increasing. To show (2) holds more generally, simply note that for $\tau \in (0,1)$ the event $\{U_D \leq \tau\}$ implies the event $\{q\left[D,U_D\right] \leq q\left[D,\tau\right]\}$ by $\tau \mapsto q\left[d,\tau\right]$ non-decreasing on $(0,1)$ for each $d$. On the other hand, the event $\{q\left[D,U_D\right] \leq q\left[D,\tau\right]\}$ implies the event $\{U_D \leq \tau\}$, since $\tau \mapsto q\left[d,\tau\right]$ is strictly-increasing and left-continuous[15] in $(0,1)$ for each $d$.

Finally, since $\tau \mapsto q[d,\tau]$ is strictly increasing, left-continuous, we have

$$P\left[q[D,U_D] = q[D,\tau]|Z=z\right] = 0,$$

so that $P$-a.e.

$$P\left[Y \leq q\left[D,\tau\right]|Z\right] = P\left[Y < q\left[D,\tau\right]|Z\right]. \quad \blacksquare$$

[15] $\tau \mapsto q\left[d,\tau\right]$ is said to be left-continuous if $\lim_{\tau' \uparrow \tau} q\left[d,\tau'\right] = q\left[d,\tau\right]$.

**Part (b).** Conditioning on $X = x$ is suppressed. For $P$-a.e. value $z$ of $Z$

$$\begin{aligned} P[Y \leq q[D,\tau] \,|\, Z = z] & \\ & \overset{(1)}{=} P\left[q[D,U_D] \leq q[D,\tau] \,|\, Z = z\right] \\ & \overset{(2)}{\geq} P\left[U_D \leq \tau | Z = z\right], \\ & \overset{(3)}{\geq} P\left[U_o \leq \tau | Z = z\right] = \tau, \end{aligned} \tag{22}$$

where the equalities (1) and (3) are by the same arguments as in the proof of part (a), and equality (2) follows because the event $\{U_D \leq \tau\}$ is a subset of the event $\{q[D,U_D] \leq q[D,\tau]\}$ by $\tau \mapsto q[d,\tau]$ non-decreasing for each $d$. On the other hand,

$$\begin{aligned} P[Y < q[D,\tau] \,|\, Z = z] & \\ & \overset{(4)}{=} P\left[q[D,U_D] < q[D,\tau] \,|\, Z = z\right] \\ & \overset{(5)}{=} (or <) P\left[U_D < \tau | Z = z\right] \\ & \overset{(6)}{=} P\left[U_o < \tau | Z = z\right] = \tau, \end{aligned} \tag{23}$$

where the equalities (4) and (6) are by the same arguments as in the proof of part (a). (5) holds as an equality if $q[D,\tau]$ is strictly increasing at $\tau$, P-a.e., conditional on $Z = z$, since the event $\{q[D,U_D] < q[D,\tau]\}$ equals the event $\{U_D < \tau\}$ P-a.e. If on the other hand, if $q[D,\tau]$ is flat at $\tau$, P-a.e., conditional on $Z = z$, with prob $> 0$, conditional on $Z = z$, then $\{q[D,U_D] < q[D,\tau]\}$ is a strict subset of the event $\{U_D < \tau\}$ by $\tau \mapsto q[d,\tau]$ non-decreasing and left-continuous for each $d$. ■

# E Proof of Theorem 2

First show that statement (1) $\Leftrightarrow$ statement (2). Let $\epsilon \equiv Y - q'(d,X,\tau)$. This follows immediately by definition $Q_{\epsilon|X,Z}(\tau) \equiv \inf\{m : P[\epsilon \leq m | X,Z] \geq \tau\}$.

We next show that statement (2) $\Leftrightarrow$ statement (3). $0 = Q_\epsilon[\tau|x,z]$ is the conditional quantile. Therefore, it is a best predictor under asymmetric absolute loss, cf. Manski (1985), p. 55. We need to show a stronger fact — (2) $\Leftrightarrow$ (3) — extending his argument. Write for any $v < 0$

$$\begin{aligned} & E[\rho_\tau(\epsilon - v)|x,z] - E\left[\rho_\tau(\epsilon)\,|x,z\right] \\ & \qquad = (1-\tau)\int_{(-\infty,v]} v \; dF_\epsilon[e|x,z] + \int_{(v,0)} [e - \tau v] \; dF_\epsilon[e|x,z] \\ & \qquad + \tau \int_{[0,\infty)} (-v) \; dF_\epsilon[e|x,z] \end{aligned} \tag{24}$$

$$\begin{aligned} & = (1-\tau)\; v\; P[\epsilon \leq v|x,z] + (1-\tau)\; v\; P[\epsilon \in (v,0)|x,z] - \tau v P[\epsilon \geq 0|x,z] \\ & + \int_{(v,0)} (e - v) \; dF_\epsilon[e|x,z] \\ & = v\Big((1-\tau)P[\epsilon < 0|x,z] - P[\epsilon \geq 0|x,z]\tau\Big) + \int_{(v,0)} (e - v) \; dF_\epsilon[e|x,z] > 0. \end{aligned} \tag{25}$$

(25) $> 0$, since $v < 0$ and (i) $P[\epsilon < 0|x,z] \leq \tau$ and (ii) $\int_v^0 (e-v)\ dF_\epsilon[e|x,z] \geq 0$, and one of these inequalities must be strict. Indeed, if $P[\epsilon = 0|x,z] > 0$, the inequality (i) is strict. If on the other hand, $P[\epsilon = 0|x,z] = 0$, then the inequality (ii) must be strict. Indeed, in this case $\int_v^0 (e-v) dF_\epsilon[e|x,z] = 0$ occurs only if $F_\epsilon[e|x,z]$ is flat (assigns no mass) on $(v,0)$, which given that there is no mass at 0, contradicts to $0 = Q_{\epsilon|x,z}(\tau)$.

Next, for any $v > 0$

$$
\begin{aligned}
&E\left[\rho_\tau(\epsilon - v)|x,z\right] - E\left[\rho_\tau(\epsilon)|x,z\right] \\
&\quad = (1-\tau)\int_{(-\infty,0]} v\ dF_\epsilon[e|x,z] + \int_{(0,v)} [(1-\tau)v - e]\ dF_\epsilon[e|x,z] \\
&\quad + \tau \int_{[v,\infty)} (-v)\ dF_\epsilon[e|x,z] \\
&\quad = (1-\tau)\ v\ P[\epsilon \leq 0|x,z] - \tau\ v\ P[\epsilon \in (0,v)|x,z] - \tau v\ P[\epsilon \geq v|x,z] \\
&\quad + \int_{(0,v)} [v-e]\ dF_\epsilon[e|x,z] \\
&\quad = v\Big((1-\tau)\ P[\epsilon \leq 0|x,z] - \tau\ P[\epsilon > 0|x,z]\Big) + \int_{(0,v)} [v-e]\ dF_\epsilon[e|x,z] \geq 0.
\end{aligned}
\tag{26}
$$

(26) $\geq 0$ since $v > 0$ and $P[\epsilon > 0|x,z] \leq 1-\tau$ and $\int_0^v [v-e]\ dF_\epsilon[e|x,z] \geq 0$. (26) $= 0$ if (i) $P[\epsilon > 0|x,z] = 1-\tau$, thus $P[\epsilon \leq 0|x,z] = \tau$ and (ii) the second term is zero. (ii) happens iff $t \mapsto F_\epsilon(t|d,z)$ is flat at $(0,v)$. When (26) $= 0$, 0 is not the unique predictor under $\rho_\tau$ loss. Since $\tau \mapsto Q_{\epsilon|x,z}(\tau)$ is left-continuous, for any sequence $\tau'_m \uparrow \tau$ we have $q_m = Q_{\epsilon|x,z}(\tau'_m) \uparrow 0$ and for any $v > 0$, denoting $\epsilon_m \equiv \epsilon - q_m$

$$
\begin{aligned}
&E\left[\rho_{\tau'_m}(\epsilon_m - v)|x,z\right] - E\left[\rho_{\tau'_m}(\epsilon_m)|x,z\right] \\
&= v\Big((1-\tau'_m)\ P[\epsilon_m \leq q_m|x,z] - \tau'_m\ P[\epsilon_m > q_m|x,z]\Big) + \int_{(0,v)} [v-e] dF_{\epsilon_m}[e|x,z] > 0,
\end{aligned}
\tag{27}
$$

since both of the terms are non-negative by the earlier arguments, and $\int_{(0,v)} |v-e] dF_{\epsilon_m}[e|x,z] \equiv \int_{(q_m, v+q_m)} [v - e - q_m] dF_\epsilon[e|x,z] > 0$, since $0 = Q_{\epsilon|x,z}(\tau) \in (q_m, v+q_m)$ for sufficiently large $m$. In other words, the last statement implies that $F_\epsilon[\cdot|x,z]$ has to assign positive mass to $(q_m, v+q_m)$ for large $m$. In addition, by arguments like in (25) for any $v < 0$

$$
E\left[\rho_{\tau'_m}(\epsilon_m - v)|x,z\right] - E\left[\rho_{\tau'_m}(\epsilon_m)|x,z\right] > 0.
$$

Thus, $Q_\epsilon[\tau'_m|x,z]$ are unique best predictors for for large $m$, and $\lim_{\tau'_m \uparrow \tau} Q_{\epsilon_m}[\tau'_m|x,z] = 0$.

Thus, we demonstrated the equivalence of statement (2) and statement (3). 0 is the unique (modified by the limit operation) best predictor under asymmetric absolute loss. Note that the limit operation is only needed for at most countably many $\tau$ in $(0,1)$.

Finally, equivalence (3) $\Leftrightarrow$ (4) is obvious. ■

# F Proof of Theorem 3

The proof is a special case of Theorem 5 in section 1 ■

# G Assumptions R.1-R.6

The following assumptions are maintained for the inverse quantile regression.

**R1** $W_i = (Y_i, D_i, X_i, Z_i)$ are iid and $(D_i, X_i, Z_i)$ take values in a compact set.

**R2** $(\alpha_\tau, \beta_\tau) \in$ interior $\mathcal{V}$, where $\mathcal{V} \equiv \mathcal{A} \times \mathcal{B}$ is compact, convex, and $(\alpha_\tau, \beta_\tau)$ is unique $(\alpha, \beta)$ : $E\varphi_\tau(Y_i - D_i'\alpha - X_i'\beta)\Psi_i = 0$, where $\Psi_i \equiv V_i \cdot [X_i' : \Phi_i']'$ and $\varphi_\tau(u) \equiv (1(u < 0) - \tau)$.

**R3** $Y$ has bounded conditional density given $X$, $D$, $Z$, uniformly over support of $(X, D, Z)$.

**R4** $J(\pi) \equiv \frac{\partial}{\partial(\alpha', \beta', \gamma')} E[\varphi_\tau(Y - D'\alpha - X'\beta - \Phi'\gamma)\Psi]$ has full column rank and is continuous at each $(\alpha, \beta, \gamma)$ in $\mathcal{A} \times \mathcal{B} \times \mathcal{G}$, where $\mathcal{G}$ is an open ball in $\mathbb{R}^{\dim(\gamma)}$ at zero.

**R5** Functions $(z, x) \mapsto \widehat{\Psi}(z, x)$ and $(z, x) \mapsto \widehat{V}(z, x)$ belong to a set $\mathcal{F}$ wp $\to 1$; $\mathcal{F}$ is a set of boundedly differentiable functions $C_M^\eta$, with smoothness order $\eta > \dim(z, x)/2$.[16] $\widehat{\Phi}(\cdot) \xrightarrow{p} \Phi(\cdot), \widehat{V}(\cdot) \xrightarrow{p} V(\cdot) \in \mathcal{F}$, uniformly over compact sets. $V(\cdot) > 0$.

**Remark G.1** All assumptions, but R5, are analogous to the standard assumptions for quantile regression. They may be refined at a cost of more complicated notation and proof.

**Remark G.2** Smoothness in R5 needs to hold only for the non-discrete sub-component of $(x, z)$. As discussed in the text condition R5 allows for a wide variety of nonparametric and parametric estimators, as shown by Andrews (1994). Ideally, we would like to approximate the optimal instruments and the optimal weight as closely as possible using non-parametric or parametric methods. There is a wide variety of estimators that satisfy assumption R5, such as smooth parametric approximation to $V(X, Z)$ and $\Phi(X, Z)$ or, alternatively, various smooth kernel estimators and smooth series estimators. See Andrews (1994), (1995), Newey (1997), (1990), and Newey and Powell (1990) for a catalogue of estimators that satisfy condition R5.

# H Proof of Theorem 4

**1.** In the proof $W$ denotes $(Y, D, X, Z)$. Define for $\theta \equiv (\beta, \gamma)$ and $\theta_0 \equiv (\beta_\tau, 0)$ and $\varphi_\tau(u) \equiv (1(u < 0) - \tau)$

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

where $\Psi \equiv V \cdot (X', \Phi')', \Phi \equiv \Phi(X, Z)$, $\widehat{\Psi} \equiv \widehat{V} \cdot (X', \widehat{\Phi})', \widehat{\Phi} \equiv \widehat{\Phi}(X, Z)$;

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

where $\rho_\tau(u) \equiv (\tau - 1(u < 0))u$. Let

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

and for $\Theta \equiv \mathcal{B} \times \mathcal{G}$

$$\widehat{\theta}(\alpha) \equiv (\widehat{\beta}(\alpha), \widehat{\gamma}(\alpha)) \equiv \arg \inf_{\theta \in \Theta} Q_n(\alpha, \theta),$$
$$\theta(\alpha) \equiv (\beta(\alpha), \gamma(\alpha)) \equiv \arg \inf_{\theta \in \Theta} Q(\alpha, \theta),$$
$$\widehat{\alpha} \equiv \arg \inf_{\alpha \in \mathcal{A}} \|\widehat{\gamma}(\alpha)\|, \quad \alpha^* \equiv \arg \inf_{\alpha \in \mathcal{A}} \|\gamma(\alpha)\|.$$

[16] See page 154 in van der Vaart and Wellner (1996).

By Theorem 1 (i), the true parameters $(\alpha_\tau, \beta_\tau)$ solve the equation

$$E\varphi_\tau(Y_i - D_i'\alpha_\tau - X_i'\beta_\tau - \Phi_i'0)\Psi_i = 0.$$

On the other hand, by R3 $\theta(\alpha)$ satisfies the equation:

$$E\varphi_\tau(Y_i - D_i'\alpha - X_i'\beta(\alpha) - \Phi_i'\gamma(\alpha))\Psi_i = 0.$$

We need to find $\alpha^*$ such that this equation holds and the norm of $\gamma(\alpha)$ is as small as possible. $\alpha^* = \alpha_\tau$ makes the norm of $\gamma(\alpha^*) = 0$ equal zero. Thus $\alpha^* = \alpha_\tau$ is a solution; by R2 it is unique. Additionally, by R2 $\beta(\alpha^*) = \beta_\tau$.

**2.** For each $\alpha$ and $\theta$, by a LLN ( lemma K.1 )

$$\mathbb{E}_n[\widehat{g}(W, \alpha, \theta) - \widehat{g}(W, \alpha_\tau, \bar{\theta})] \xrightarrow{p} E[g(W, \alpha, \theta) - g(W, \alpha_\tau, \bar{\theta})],$$

for some fixed $\bar{\theta}$ (the subtraction of the terms is to make the summands bounded functions of $W$). The lhs is a finite convex function in $\theta$ and $\alpha$, at least wp $\to 1$. Therefore the convergence is uniform over compact sets. Hence by lemma A.1

$$\widehat{\theta}(\alpha_\tau) \xrightarrow{p} \theta_\tau \quad \text{and} \quad \widehat{\theta}(\widehat{\alpha}) \xrightarrow{p} \theta_\tau, \quad \text{provided } \widehat{\alpha} \xrightarrow{p} \alpha_\tau.$$

**3.** $\widehat{\alpha} \xrightarrow{p} \alpha_\tau$. ( shown below).

**4.** By the computational properties of quantile regression estimator $\widehat{\theta}(\alpha_n)$, for any $\alpha_n$ in a small ball at $\alpha_\tau$

$$O(K/\sqrt{n}) = \sqrt{n}\mathbb{E}_n\widehat{f}(W, \alpha_n, \widehat{\theta}(\alpha_n)). \tag{28}$$

By lemma K.1, the following expansion of r.h.s. is valid for any $\alpha_n \xrightarrow{p} \alpha_\tau$:[17]

$$\begin{aligned}\sqrt{n}\mathbb{E}_n\widehat{f}(W, \alpha_n, \widehat{\theta}(\alpha_n)) &\equiv \mathbb{G}_n\widehat{f}(W, \alpha_n, \widehat{\theta}(\alpha_n)) + \sqrt{n}E\widehat{f}(W, \alpha_n, \widehat{\theta}_n(\alpha_n)) \\ &= \mathbb{G}_n f(W, \alpha_\tau, \theta_\tau) + o_p(1) + \sqrt{n}E\widehat{f}(W, \alpha_n, \widehat{\theta}(\alpha_n))\end{aligned} \tag{29}$$

Expanding the last line further

$$\begin{aligned}&= \mathbb{G}_n f(W, \alpha_\tau, \theta_\tau) + o_p(1) \\ &+ (J_\theta + o_p(1))\sqrt{n}(\widehat{\theta}(\alpha_n) - \theta_\tau) \\ &+ (J_\alpha + o_p(1))\sqrt{n}(\alpha_n - \alpha_\tau).\end{aligned} \tag{30}$$

In other words for any $\alpha_n \xrightarrow{p} \alpha_\tau$

$$\sqrt{n}(\widehat{\theta}(\alpha_n) - \theta_\tau) = -J_\theta^{-1}\mathbb{G}_n f(W, \alpha_\tau, \theta_\tau) - J_\theta^{-1}J_\alpha[1 + o_p(1)]\sqrt{n}(\alpha_n - \alpha_\tau) + o_p(1), \text{ i.e}$$

$$\sqrt{n}(\widehat{\gamma}(\alpha_n) - 0) = -\bar{J}_\gamma\mathbb{G}_n f(W, \alpha_\tau, \theta_\tau) - \bar{J}_\gamma J_\alpha[1 + o_p(1)]\sqrt{n}(\alpha_n - \alpha_\tau) + o_p(1).$$

Over a shrinking ball at $\alpha_\tau$, denoted $B_n(\alpha_\tau)$, wp $\to 1$, for $\|x\|_A \equiv x'Ax$

$$\widehat{\alpha} = \arg \inf_{\alpha_n \in B_n(\alpha_\tau)} \|\widehat{\gamma}(\alpha_n)\|_{\bar{A}}.$$

Observe that

$$\sqrt{n}\|\widehat{\gamma}(\alpha_n)\|_{\bar{A}} = \|O_p(1) - \bar{J}_\gamma J_\alpha[1 + o_p(1)]\sqrt{n}(\alpha_n - \alpha_\tau)\|_{A + o_p(1)},$$

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

Since $\bar{J}_\gamma J_\alpha$ and $A$ have full rank, $\sqrt{n}(\widehat{\alpha} - \alpha_\tau) = O_p(1)$. Hence by lemma A.2

$$\sqrt{n}(\widehat{\alpha} - \alpha_\tau) \stackrel{LD}{=} \arg\inf_{\mu} \| - \bar{J}_\gamma \mathbb{G}_n f(W, \alpha_\tau, \theta_\tau) - \bar{J}_\gamma J_\alpha \mu\|_A.$$

where $\stackrel{LD}{=}$ means that the limit distributions of the lhs and rhs agree. Conclude that:

$$\sqrt{n}(\widehat{\alpha} - \alpha_\tau) \stackrel{LD}{=} -(J_\alpha' \bar{J}_\gamma' A \bar{J}_\gamma J_\alpha)^{-1} J_\alpha' \bar{J}_\gamma' A \bar{J}_\gamma \mathbb{G}_n f(W, \alpha_\tau, \theta_\tau) \text{ and}$$

$$\sqrt{n}(\widehat{\theta} - \theta_\tau) \stackrel{LD}{=} -J_\theta^{-1}[I - J_\alpha (J_\alpha' \bar{J}_\gamma' A \bar{J}_\gamma J_\alpha)^{-1} J_\alpha' \bar{J}_\gamma' A \bar{J}_\gamma] \mathbb{G}_n f(W, \alpha_\tau, \theta_\tau)$$

with $\mathbb{G}_n f(W, \alpha_\tau, \theta_\tau) \xrightarrow{d} N(0, S)$ by CLT.

Finally, the consistency and asymptotic representation for $\widehat{\beta}$ follows analogously to that of $\widehat{\beta}(\widehat{\alpha})$, except that in the definition of $\widehat{\beta}$ instead of $\widehat{\gamma}$ we use its plim 0. Therefore, analogously to (40)- (42):

$$\sqrt{n}(\widehat{\beta} - \beta_\tau) \stackrel{LD}{=} -J_\beta^{-1}[I_k : 0][I - J_\alpha (J_\alpha' \bar{J}_\gamma' A \bar{J}_\gamma J_\alpha)^{-1} J_\alpha' \bar{J}_\gamma' A \bar{J}_\gamma] \mathbb{G}_n f(W, \alpha_\tau, \theta_\tau). \blacksquare$$

**Proof of step 3.** The argument is just slightly more complicated than usual consistency arguments, cf. Amemiya (1985) or Newey and McFadden (1994). For some $\bar{\theta}$, $\widehat{\theta}(\alpha)$ maximizes

$$\bar{Q}_n(\alpha, \theta) \equiv -\mathbb{E}_n[\widehat{g}(W, \alpha, \theta) - \widehat{g}(W, \alpha_\tau, \bar{\theta})] \xrightarrow{p} \bar{Q}_\infty(\alpha, \theta) \equiv -E[g(W, \alpha, \theta) - g(W, \alpha_\tau, \bar{\theta})] \quad (31)$$

where the convergence is uniform in $(\theta, \alpha)$ over compact sets, using step **2**. For any $\epsilon > 0$, wp $\to 1$, uniformly in $\alpha \in \mathcal{A}$: [i] $\bar{Q}_n(\alpha, \widehat{\theta}(\alpha)) \geq \bar{Q}_n(\alpha, \theta(\alpha))$ by definition, [ii] $\bar{Q}_\infty(\alpha, \widehat{\theta}(\alpha)) > \bar{Q}_n(\alpha, \widehat{\theta}(\alpha)) - \epsilon/2$ by (31), [iii] $\bar{Q}_n(\alpha, \theta(\alpha)) > \bar{Q}_\infty(\alpha, \theta(\alpha)) - \epsilon/2$ by (31). Hence wp $\to 1$

$$\bar{Q}_\infty(\alpha, \widehat{\theta}(\alpha)) > \bar{Q}_n(\alpha, \widehat{\theta}(\alpha)) - \epsilon/2 \geq \bar{Q}_n(\alpha, \theta(\alpha)) - \epsilon/2 > \bar{Q}_\infty(\alpha, \theta(\alpha)) - \epsilon.$$

Let $\{B(\alpha), \alpha \in \mathcal{A}\}$ be a collection of balls with diameter $\delta$, each centered at $\theta(\alpha)$. Then $\epsilon \equiv \inf_{\alpha \in \mathcal{A}} \left[\bar{Q}_\infty(\theta(\alpha)) - \sup_{\theta \in \Theta \setminus B(\alpha)} \bar{Q}_\infty(\theta)\right] > 0$, by assumption R4 and concavity in $\theta$ for each $\alpha$. It now follows wp $\to 1$, uniformly in $\alpha$

$$\bar{Q}_\infty(\widehat{\theta}(\alpha)) > \bar{Q}_\infty(\theta(\alpha)) - \bar{Q}_\infty(\theta(\alpha)) + \sup_{\theta \in \Theta \setminus B(\alpha)} \bar{Q}_\infty(\theta(\alpha)) = \sup_{\theta \in \Theta \setminus B(\alpha)} \bar{Q}_\infty(\theta).$$

Thus wp $\to 1$, $\sup_{\alpha \in \mathcal{A}} \|\widehat{\theta}(\alpha) - \theta(\alpha)\| \leq \delta$, for any $\delta > 0$. This implies that $\sup_{\alpha \in \mathcal{A}} \left| \|\widehat{\gamma}(\alpha)\|_A - \|\gamma(\alpha)\|_A \right| \xrightarrow{p} 0$, which by Lemma A.2 implies $\widehat{\alpha} \xrightarrow{p} \alpha^*$. $\blacksquare$

# I Identification Results: Generalizations

The following statements and functions are all conditional on the event $X = x$. For notation sake, we suppress this conditioning. Suppose support of $D$ is a finite set of discrete values in $\mathbb{R}^l$. We can label the points of the support as $\{1, ..., J\}$. Define $\mathcal{L}(x)$ as the convex hull of functions $\varphi$ mapping $d$ from $\{1, ...J\}$ to $(y : f_Y(y|d) > 0)$ such that $P(Y \leq \varphi(D, \tau)|Z)$ belongs to $[\tau - \delta, \tau + \delta]$ a.s. for small $\delta > 0$. Define the following function

$$\mathbf{z} \mapsto \Pi_{\mathbf{z}}(\varphi, x) = [P[Y \leq \varphi(D)|z_j], 1 \leq j \leq J],$$

where $\mathbf{z} = (z_j, 1 \leq j \leq J)$. Define, assuming relevant smoothness $J_{\mathbf{z}}(\varphi, x) \equiv \frac{d}{d\varphi} \Pi_{\mathbf{z}}(\varphi)$

$$\equiv \begin{bmatrix} f_Y(\varphi(1)|D = 1, z_1) P[D = 1|z_1] & \dots & f_Y(\varphi(J)|D = J, z_1) P[D = J|z_1] \\ \vdots & \vdots & \vdots \\ f_Y(\varphi(1)|D = 1, z_J) P[D = 1|z_J] & \dots & f_Y(\varphi(J)|D = J, z_J) P[D = J|z_J] \end{bmatrix}.$$

The statement that rank $J_{\mathbf{Z}}(\varphi, x)$ is full w. pr. $> 0$ means that with positive probability rank $J_{\mathbf{Z}}(\varphi, x) = J$, where $\mathbf{Z} = \{Z_j\}$ are independent replica of $Z$, given $X = x$.

**Theorem 5 (Discrete $D$)** *Suppose A1-A5 hold, and that $f_Y(y|d, z, x) > 0$ and finite over the range of $d \mapsto q(d, x, \tau)$. Then $d \mapsto q(d, x, \tau)$ is a unique solution of*

$$P(Y \leq q(d, x, \tau)|x, z) = \tau \text{ for } P\text{-a.e. } z, \text{ given } X = x, \tag{32}$$

*among $\mathcal{L}(x)$ if for any $\varphi \in \mathcal{L}(x)$ $J_{\mathbf{Z}}(\varphi, x)$ is finite and has full rank w. pr. $> 0$.*

**Proof.** Condition on the event $X = x$. We know that $q(d, x, \tau)$ solves (32) from Theorem 1, hence it belongs to $\mathcal{L}(x)$. Suppose there exits $q^* \in \mathcal{L}(x)$ that also solves (32) such that $d \mapsto q^*(d)$ and $d \mapsto q(d, x, \tau)$ disagree on $\{1, ..., J\}$. Then

$$\Pi_{\mathbf{Z}}(q^*, x) = \mathbf{1}\tau \text{ and } \Pi_{\mathbf{Z}}(q, x) = \mathbf{1}\tau, \quad P - \text{a.s.},$$

for a conformable vector of $1's$, $\mathbf{1}$. Then for any vector $\lambda \in \mathbb{R}^J \setminus \{0\}$ Taylor expansion gives

$$\lambda' (\Pi_{\mathbf{Z}}(q^*, x) - \Pi_{\mathbf{Z}}(q, x)) = \lambda' J_{\mathbf{Z}}(q^*_\lambda, x) = 0, \quad P - \text{a.s.}$$

where $q^*_\lambda \in \mathcal{L}(x)$, which is impossible by the full rank assumption. ■

Finally, we consider continuous $D$.[18] As Newey and Powell (2001) shown, in the model $E(Y - \mu(D)|Z) = 0$, the condition for identification of $\mu$ is the Lehmann-Scheffe completeness condition:

**L1** $E[\Delta(D)|z] = 0$ $P$ a.e. $\Rightarrow \Delta(D) = 0$ $\mathcal{P}$ a.e.,

where $\mathcal{P}$ is a collection of $F_D[\cdot|z]$ as $z$ varies over support of $Z$ given $X = x$. Lehmann (1954) provided a sufficient "happy family" condition:

**L2** $P[D = d|z]$ is a full rank exponential family $h(d) \cdot \exp(\eta(z)'T(d) + \lambda(z))$.

The full rank condition requires $\eta(z)$ to vary over an open rectangle in $\mathbb{R}^{\dim(T(d))}$ and $T(d)$ not to satisfy a linear constraint. **L2** allows for a broad variety of non-parametric distributions.

The statements are conditional on the event $X = x$ but we suppress this conditioning. Define $\mathcal{L}(x)$ as a convex hull of functions $m$ that map a $d$ from the set $\mathcal{D}(x)$, the support of $D$, to $(y : f(y|d) > 0)$ such that $P[Y \leq m(D)|Z] \in [\tau - \delta, \tau + \delta]$ a.s., for small $\delta > 0$ given $X = x$. Solution $q$ is said to be unique if any other solution $m = q$ $\mathcal{P} - a.e.$, where $\mathcal{P}$ is defined above.

**Theorem 6** *Suppose A1-A5 hold, and that $f_Y(y|d, z, x) > 0$ and finite over the range of $d \mapsto q(d, x, \tau)$, uniformly in $z$. Then $d \mapsto q(d, x, \tau)$ is a unique solution of (32) among $\mathcal{L}(x)$ if*

*i. for any $\Delta(d) = m(d) - q(d, x, \tau)$ such that $m \in \mathcal{L}(x)$ and $\epsilon \equiv Y - q(d, x, \tau)$ and independent standard uniform $\zeta$, $E[f_\epsilon(\zeta\Delta(D)|D, z)\Delta(D)|z] = 0$ $P$-a.e. $\Rightarrow \Delta(D) = 0$ $\mathcal{P}$-a.e.*

*ii. sufficient condition for i. is $f(t, d|z) \equiv c \cdot t^{-1} [P_\epsilon[t|d, z] - P_\epsilon[0|d, z]] \cdot f_D[d|z] \propto h(d, t) \cdot \exp(\eta(z)'T(d, t))$ is an exponential family of full rank.*

In the last expression, $f(0, d|z) \equiv \lim_{t \to 0} f(t, d|z)$. Condition **ii.** is a plausible non-parametric condition, with rhs motivated as a Taylor approximation of the log of lhs.

[18] To be removed and is given here for completeness. The full treatment is to be given in the joint work with Whitney Newey and Guido Imbens

**Proof.** By hypothesis there is $m \in \mathcal{L}(x)$ such that $P[Y \leq m(D)|z] = \tau \quad P$ a.e. Then the difference $0 = P[Y \leq m(D)|z] - P[Y \leq q(D)|z]$ equals by Taylor expansion

$$E\Delta(D) \int_0^1 f_\epsilon(\delta\Delta(D)|D,z)d\delta \equiv Ef_\epsilon(\zeta\Delta(D)|D,z)\,[\Delta(D)] = 0, \tag{33}$$

where $\zeta$ is a uniform variable on $[0,1]$, independent of $Z, D$ and $\Delta \equiv m - q$. (33) is proportional to $E^*[\Delta(D)|z]$ where $E^*$ is the expectation distorted by $f_\epsilon(\zeta\Delta(D)|D,z)$. For uniqueness we need that (33)=0 implies $\Delta(D) = 0$ $\mathcal{P}$ -a.e. This proves **i.**. To prove **ii.** is sufficient, by L2 condition **ii.** implies (also using $f(t(d),d|z) = 0 \Leftrightarrow f_D(d|z) = 0$)

$$E_{f(t(d),d|z)}\Delta(d) = 0 \implies \Delta(d) = 0 \quad \mathcal{P} - \text{a.e.} \tag{34}$$

for any measurable function $t(d)$, since $f(t(d),d|z)$ remains a full $\dim(d)$-rank exponential family. Thus if $t(d) \equiv \Delta(d)$, (34) still holds. Now note $\int_0^1 f_{Y-q(D)}(\zeta\varphi(D)|D,z)d\zeta$ equals $(P[Y - q(D) \leq \Delta(D)|D,z] - P[Y - q(D) \leq 0|D,z])/\Delta(D)$, so $E_{f(\Delta(d),d|z)}\Delta(d) \equiv$ lhs of (33). ■

# J Additional Results: Empirical Likelihood

Here we treat the generalized empirical likelihood. Define for $\vartheta \equiv (\alpha,\beta), \varphi_\tau(u) = \tau - 1(u < 0)$

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

$\Psi \equiv \Psi(X,Z)$ is a smooth function of an instrument,
$\widehat{\Psi} \equiv \widehat{\Psi}(X,Z)$ is a smooth consistent estimate of such function, satisfying **L5**.

$$\text{Define also} \qquad Q_n(\vartheta,\gamma) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathrm{s}[\widehat{f}(W,\vartheta)'\gamma],$$

$$\widehat{\gamma} \equiv \arg\inf_\gamma Q_n(\widehat{\vartheta},\gamma), \quad \widehat{\vartheta} \equiv \arg\sup_{\vartheta\in\mathcal{V}}[\inf_\gamma Q_n(\vartheta,\gamma)]. \tag{35}$$

Function $\mathrm{s}(\cdot)$ equals a strictly convex, finite, and four times differentiable function $\mathrm{s}_0$ on an open interval of $\mathbb{R}$ containing 0, and equals $+\infty$ outside it. Normalize $[\partial^j \mathrm{s}(v)/\partial v^j](0) = 1$ for $j = 1,2$. Functions $\mathrm{s}_0(v) = -\ln(1-v), \exp(v), (1+v)^2/2$ lead to the well-known empirical likelihood, exponential tilting, and continuous up-dating GMM estimator. See Imbens (1997), Newey and Smith (2001), and Kitamura and Stutzer (1997).

Just as GMM, GEL is infeasible in our settings with two or more covariates. It may be useful in low-dimensional settings or as a refinement of the IQR estimator. The latter can be used to bring the estimates to a right neighborhood, and the GEL can be recomputed over such a neighborhood. For this purpose, GEL are known to have good finite sample properties. The pivotal objective function of GEL can be used for construction of confidence intervals.

**Assumptions L.1-L.6** The following assumptions are maintained

**L1** $W_i = (Y_i, D_i, X_i, Z_i)$ is *i.i.d.* and $(D_i, X_i, Z_i)$ take values in a compact set.

**L2** $\vartheta_0 \equiv (\alpha_\tau, \beta_\tau) \in$ interior $\mathcal{A} \times \mathcal{B}$, a compact convex set.

**L3** $\vartheta_0$ is unique $\vartheta$ s.t. $E\varphi_\tau(Y - D'\alpha - X'\beta)\Psi = 0$, in $\mathcal{V} \equiv \mathcal{A} \times \mathcal{B}$.

**L4** $S(\vartheta) = E\varphi_\tau(Y - D'\alpha - X'\beta)^2\Psi\Psi'$ is positive definite for each $\vartheta \in \mathcal{V}$. $(S \equiv S(\vartheta_0))$

L5 $(z,x) \mapsto \widehat{\Psi}(z,x) \in \mathcal{F}$, wp $\to 1$, $\mathcal{F}$ is a set of boundedly differentiable functions $C_M^\eta$, with smoothness order $\eta > \dim(z,x)/2$.[19] $\widehat{\Psi}(\cdot) \xrightarrow{p} \Psi(\cdot) \in \mathcal{F}$, uniformly over compacts.

L6 $J(\vartheta) \equiv \frac{\partial}{\partial v} E\left[\varphi_\tau(Y - D'\alpha - X'\beta)\Psi\right] = E\left[f_{Y-D'\alpha-X'\beta}(0|X,D,Z)\Psi[D' : X']\right]$ is defined, finite, has full column rank and continuous at each $\alpha, \beta$ in $\mathcal{A} \times \mathcal{B}$.

**Remark J.1** All assumptions, but L5, are fairly standard. L5 allows for a wide variety of nonparametric and parametric estimators. See Remark G.2.

**Theorem 7 (GEL)** *In the linear model (8) and assumptions L1-L6 listed above*

$$\begin{matrix}\sqrt{n}(\widehat{\vartheta} - \vartheta_0) \\ \sqrt{n}\widehat{\gamma}\end{matrix} \xrightarrow{d} N\left[\begin{matrix} 0, & (J'S^{-1}J)^{-1} & 0 \\ 0, & 0 & S^{-1}[S - J(J'S^{-1}J)^{-1}J']S^{-1}\end{matrix}\right]$$

*where* $S = \tau(1-\tau)E\Psi\Psi'$ *and* $J = E\left[f_\epsilon(0|X,D,Z)\Psi[D' : X']\right]$, $\epsilon \equiv Y - D'\alpha(\tau) - X'\beta(\tau)$.

**Corollary 3** *Further, if we set* $\Psi^* = V^* \cdot [X', \Phi^{*\prime}]'$, *where* $\Phi^* \equiv E[D \cdot v|Z,X]/V^*$, $v \equiv f_\epsilon[0|D,Z,X]$, *and* $V^* \equiv f_\epsilon(0|X,Z)$, *the asymptotic variance of* $\widehat{\alpha}$ *and* $\widehat{\beta}$, *simplifies to the efficiency bound*

$$\tau(1-\tau)E[\Psi^*\Psi^{*\prime}]^{-1}.$$

**Proof.** The proof extends the arguments of Kitamura and Stutzer (1997) and Christoffersen, Hahn, and Inoue (1999). **1.** Define

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

where by $\Psi$ and $\widehat{\Psi}$ we denote vectors $\Psi(X,Z)$ and $\widehat{\Psi}(X,Z)$. Define also

$$Q_n(\vartheta,\gamma) \equiv \mathbb{E}_n \mathrm{s}[\widehat{f}(W,\vartheta)'\gamma], \quad Q(\vartheta,\gamma) \equiv E\ \mathrm{s}[f(W,\vartheta)'\gamma], \text{ and}$$
$$\widehat{\gamma}(\vartheta) \equiv \arg\inf_\gamma Q_n(\vartheta,\gamma), \quad \widehat{\vartheta} \equiv \arg\sup_{\vartheta\in\mathcal{V}} \inf_\gamma Q_n(\vartheta,\gamma),$$

$\gamma(\vartheta) \equiv \arg\inf_\gamma Q(\vartheta,\gamma)$, $\vartheta^* \equiv \arg\sup_{\vartheta\in\mathcal{V}} \inf_\gamma Q(\vartheta,\gamma)$. By arguments of Kitamura and Stutzer (1997) or Newey and Smith, $\vartheta^* = \vartheta_0$ and $\gamma(\vartheta^*) = 0$.

**2.** By Lemma K.1, in $\bar{\mathbb{R}}$, for any $\vartheta_n \xrightarrow{p} \vartheta_0$ $\mathbb{E}_n \mathrm{s}[\widehat{f}(W,\vartheta_n)'\gamma] \xrightarrow{p} E\mathrm{s}\left[f(W,\vartheta_0)'\gamma\right]$ for each $\gamma$ in a dense countable subset of $\mathbb{R}^{\dim(\gamma)}$. Hence by convexity lemma A.1, since $E\mathrm{s}\left[f(W,\vartheta_0)'\gamma\right]$ is finite over an open set by L1

$$\widehat{\gamma}(\vartheta_0) \xrightarrow{p} 0, \quad \widehat{\gamma}(\widehat{\vartheta}) \xrightarrow{p} 0, \text{ provided } \widehat{\vartheta} \xrightarrow{p} \vartheta_0.$$

**3.** By lemma K.1 and consistency proof of Kitamura and Stutzer (1997) or e.g. Newey and Smith (2001) and references therein that do not require smoothness of $f$: $\widehat{\vartheta} \xrightarrow{p} \vartheta_0$.

**4.** Step 4, proved below, shows $\sqrt{n}\mathbb{E}_n\widehat{f}(W,\widehat{\vartheta}) = O_p(1)$.

**5.** In view of steps 2-3, by Lemma K.1 and properties of s, the following expansion of the first order conditions is valid, wp$\to$ 1 for $(\gamma_n,\vartheta_n) = (\widehat{\gamma},\widehat{\vartheta})$ or $(\gamma_n,\vartheta_n) = (\widehat{\gamma}(\vartheta_0),\vartheta_0)$,

$$\begin{aligned} 0 &= \sqrt{n}\mathbb{E}_n\widehat{f}(W,\vartheta_n)\mathrm{s}[f(W,\vartheta_n)\gamma_n] \\ &= \left[\sqrt{n}\mathbb{E}_n\widehat{f}(W,\vartheta_n)\right] + \mathbb{E}_n\widehat{f}(W,\vartheta_n)\widehat{f}(W,\vartheta_n)'\sqrt{n}\gamma_n + O_p(\sqrt{n}\|\gamma_n\|^2) \qquad (36) \\ &= \left[\sqrt{n}\mathbb{E}_n f(W,\vartheta_0) + (J + o_p(1))\sqrt{n}(\vartheta_n - \vartheta_0)\right] + (S + o_p(1))'\sqrt{n}\gamma_n + O_p(\sqrt{n}\|\gamma_n\|^2). \end{aligned}$$

[19] See page 154 in van der Vaart and Wellner (1996).

By step 4, (36), and L4, $\sqrt{n}\gamma_n = O_p(1)$, i.e.

$$\sqrt{n}\widehat{\gamma} = O_p(1) \text{ and } \sqrt{n}\widehat{\gamma}(\vartheta_0) = O_p(1). \tag{37}$$

and by (36), (37), and L6,

$$\sqrt{n}(\widehat{\vartheta} - \vartheta_0) = O_p(1). \tag{38}$$

6. Step 6, proved below, shows

$$\sqrt{n}J'\widehat{\gamma} = o_p(1). \tag{39}$$

7. From (36) $\sqrt{n}\widehat{\gamma} = -S^{-1}\sqrt{n}\mathbb{E}_n f(W,\vartheta_0) - S^{-1}J\sqrt{n}(\widehat{\vartheta} - \vartheta_0) + o_p(1)$, which when put into (39) gives $J'S^{-1}\sqrt{n}\mathbb{E}_n f(W,\vartheta_0) + J'S^{-1}J\sqrt{n}(\widehat{\vartheta} - \vartheta_0) + o_p(1) = 0$, which yields

$$\begin{aligned}
&\sqrt{n}(\widehat{\vartheta} - \vartheta_0) = -(J'S^{-1}J)^{-1}J'S^{-1}\sqrt{n}\mathbb{E}_n f(W,\vartheta_0) + o_p(1) \xrightarrow{d} N(0,(J'S^{-1}J)^{-1}),\\
&\sqrt{n}\widehat{\gamma} = -[S^{-1} - S^{-1}J(J'S^{-1}J)^{-1}J'S^{-1}]\sqrt{n}\mathbb{E}_n f(W,\vartheta_0) + o_p(1)\\
&\quad \xrightarrow{d} N(0, S^{-1}[S - J(J'S^{-1}J)^{-1}J']S^{-1}),
\end{aligned}$$

and also jointly, with asymptotic covariance between $\sqrt{n}(\widehat{\vartheta} - \vartheta_0)$ and $\sqrt{n}\widehat{\gamma}$ equal 0. ■

**Proof of step 4.** By definition, for any $g_n = O_p(1/\sqrt{n})$,

$$\begin{aligned}
-n(\mathbb{E}_n \mathrm{s}[\widehat{f}(W,\widehat{\vartheta})'g_n] - \mathrm{s}(0)) &\le -n(\mathbb{E}_n \mathrm{s}[\widehat{f}(W,\widehat{\vartheta})'\widehat{\gamma}] - \mathrm{s}(0))\\
&\le -n(\mathbb{E}_n \mathrm{s}[\widehat{f}(W,\vartheta_0)'\widehat{\gamma}(\vartheta_0)] - \mathrm{s}(0)).
\end{aligned} \tag{40}$$

By Lemma K.1, wp $\to 1$ the following expansions are valid (by steps like in (36))

$$\begin{aligned}
\text{rhs of (40)} &= \sqrt{n}\mathbb{E}_n[f(W,\vartheta_0)]\widehat{\gamma}(\vartheta_0)\sqrt{n} + \frac{1}{2}\sqrt{n}\widehat{\gamma}(\vartheta_0)'S\widehat{\gamma}(\vartheta_0)\sqrt{n}\\
&\quad + O_p(n\|\widehat{\gamma}(\vartheta_0)\|^3) = O_p(1),
\end{aligned} \tag{41}$$

(since $\sqrt{n}\widehat{\gamma}(\vartheta_0) = O_p(1)$ and $\sqrt{\mathbb{E}_n}[\widehat{f}(W,\vartheta_0)] = \sqrt{\mathbb{E}_n}[f(W,\vartheta_0)]$ by Lemma K.1 ), and

$$\text{lhs of (40)} = \sqrt{n}\mathbb{E}_n[\widehat{f}(W,\widehat{\vartheta})]g_n\sqrt{n} + \frac{1}{2}\sqrt{n}g_n'Sg_n\sqrt{n} + O_p(n\|g_n\|^3). \tag{42}$$

By (40)- (42), because $g_n = O_p(1/\sqrt{n})$ is arbitrary, we have $\sqrt{n}\mathbb{E}_n\widehat{f}(W,\widehat{\vartheta}) = O_p(1)$. ■

**Proof of step 6.** For any $\vartheta_n = O_p(1/\sqrt{n})$, by definition

$$-n(\mathbb{E}_n \mathrm{s}[\widehat{f}(W,\widehat{\vartheta})'\widehat{\gamma}] - \mathrm{s}[0]) \le -n(\mathbb{E}_n \mathrm{s}[\widehat{f}(W,\vartheta_n)'\widehat{\gamma}] - \mathrm{s}[0]). \tag{43}$$

By Lemma K.1 and step 5, the following expansions (by steps like in (36)) are valid

$$\begin{aligned}
\text{lhs of (43)} &= -\sqrt{n}\mathbb{E}_n\widehat{f}(W,\widehat{\vartheta})'\widehat{\gamma}\sqrt{n} - \frac{1}{2}\sqrt{n}\widehat{\gamma}'S\widehat{\gamma}\sqrt{n} + o_p(1),\\
\text{rhs of (43)} &= -\sqrt{n}\mathbb{E}_n\widehat{f}(W,\vartheta_n)'\widehat{\gamma}\sqrt{n} - \frac{1}{2}\sqrt{n}\widehat{\gamma}'S\widehat{\gamma}\sqrt{n} + o_p(1),
\end{aligned} \tag{44}$$

and by Lemma K.1

$$\begin{aligned}
\sqrt{n}\mathbb{E}_n\widehat{f}(W,\vartheta_n) &= \sqrt{n}\mathbb{E}_n f(W,\vartheta_0) + J(\vartheta_n - \vartheta)\sqrt{n} + o_p(1),\\
\sqrt{n}\mathbb{E}_n\widehat{f}(W,\widehat{\vartheta}) &= \sqrt{n}\mathbb{E}_n f(W,\vartheta_0) + J'(\widehat{\vartheta} - \vartheta)\sqrt{n} + o_p(1).
\end{aligned} \tag{45}$$

Putting (43) - (45) together, we have

$$\sqrt{n}(\vartheta_n - \widehat{\vartheta})'J'\widehat{\gamma}\sqrt{n} \le o_p(1). \tag{46}$$

Because (46) holds for any $\vartheta_n = O_p(1/\sqrt{n})$, (46) implies $J'\widehat{\gamma}\sqrt{n} = o_p(1)$. ■.

# K A Lemma

This lemma uses empirical process arguments to obtain some of stochastic relationships.

**Lemma K.1 (Expansions)** *Under assumptions L1-L6, as* $\vartheta_n \xrightarrow{p} \vartheta_0$, *for any real-valued function* $m$ *that is Liphitz over the range of* $\widehat{f}(W)$

*i.* $\mathbb{G}_n \widehat{f}(W, \vartheta_n) = \mathbb{G}_n f(W, \vartheta_0) + o_p(1)$,

*ii.* $\mathbb{E}_n m[\widehat{f}(W, \vartheta_n)] \xrightarrow{p} Em[f(W, \vartheta_0)]$ *( in particular, for* $m(x) = xx'$ *etc.),*

*iii.* $E_n s[\widehat{f}(W, \vartheta_n)'\gamma] \xrightarrow{p} Es[f(W, \vartheta_0)'\gamma]$ *for each* $\gamma$ *in a countable dense subset of* $\mathbb{R}^{\dim(\gamma)}$.

*Under assumption R1-R6,*

*iv. For each* $(\alpha, \theta)$, $\mathbb{E}_n[\widehat{g}(W, \alpha, \theta) - \widehat{g}(W, \alpha_\tau, \bar{\theta})] \xrightarrow{p} E[g(W, \alpha_\tau, \theta) - g(W, \alpha_\tau, \bar{\theta})]$.

*v.* $\mathbb{G}_n \widehat{f}(W, \alpha_n, \widehat{\theta}(\alpha_n)) = \mathbb{G}_n f(W, \alpha_\tau, \theta_\tau) + o_p(1)$, *for any* $\alpha_n \xrightarrow{p} \alpha_\tau$.

**Proof.** Denote $\pi = (\alpha, \beta, \gamma)$ and $\Pi = \mathcal{A} \times \mathcal{B} \times \mathcal{G}$ where $\mathcal{G}$ is a ball at 0. The class of functions

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

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

$$\log N_{[\cdot]}(\epsilon, \mathcal{F}, L_2(P)) = O\left(\frac{1}{\epsilon}^{\frac{\dim(z,x)}{\eta}}\right) = O\left(\frac{1}{\epsilon}^{2+\delta'}\right),$$

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

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

is $O\left(\log N_{[\cdot]}(\epsilon, \mathcal{F}, L_2(P))\right)$, because it has the same smoothness properties. Exploiting the monotonicity and boundedness of indicator function and assumptions R4 or L6, the bracketing number of

$$\mathcal{V} \equiv \left\{ (\Phi, \pi) \mapsto \varphi_\tau(Y - D'\alpha - X'\beta - \Phi(X, Z)'\gamma), \quad \pi \in \Pi, \Phi \in \mathcal{F} \right\}$$

is $O\left(\log N_{[\cdot]}(\epsilon, \mathcal{F}, L_2(P))\right)$ as well. Therefore $\mathcal{V}$ is Donsker. Class $\mathcal{H}$ is formed as a product of these two uniformly bounded (by L1 or R1 and L5 or R5) classes:

$$\mathcal{H} = \mathcal{F}\mathcal{V},$$

so the product is Lipshitz over $(\mathcal{F} \times \mathcal{V})$, and by Theorem 2.10.6 in van der Vaart and Wellner (1996) $\mathcal{H}$ is Donsker.

Now we show **i.** using the established Donskerness. Define the process

$$h = (\Psi, \vartheta) \mapsto \mathbb{G}_n \varphi_\tau(Y - D'\alpha - X'\beta)\Psi(X, Z).$$

Since $\widehat{\Psi} \xrightarrow{p} \Psi_0$ uniformly over compacts and $\vartheta_n \xrightarrow{p} \vartheta_0$, we have $\rho(\widehat{h}, h) \xrightarrow{p} 0$, where $\rho$ is defined by the $L_2(P)$ seminorm $\rho(h) \equiv E\|\varphi_\tau(Y - D'\alpha - X'\beta)\Psi(X, Z)\|$, so that

$$\mathbb{G}_n \varphi_\tau(Y - D'\alpha_n - X'\beta_n)\widehat{\Psi}(X, Z) - \mathbb{G}_n \varphi_\tau(Y - D'\alpha - X'\beta)\Psi(X, Z) = o_p(1)$$

By the above analysis $\mathbb{G}_n m[\widehat{f}(W,\vartheta_n)]$ is Donsker (asymptotically Gaussian) as well, using Theorem 2.10.6 in van der Vaart and Wellner (1996) ($m$ is Lipshitz over bounded subsets to which $\widehat{f}(W,\vartheta_n)$ belongs wp $\to$ 1 by assumption.) From this **ii.** is immediate.

The proof of **v.** and **iv.** follows exactly as **i.** and **ii.**, respectively, using that $\mathcal{H}$ is Donsker.

To show **iii.** note that $\gamma$ is either such that $Es[f(W,\vartheta_0)'\gamma] = +\infty$ or $Es[f(W,\vartheta_0)'\gamma] < \infty$. By convexity and lower-semicontinuity of $s$, the latter set, say $F$, is convex, open, and its boundary is nowhere dense in $\mathbb{R}^{\dim(\gamma)}$. Thus for $\gamma \in F$, $Es[f(W,\vartheta)'\gamma]|_{v=v_n, f=\widehat{f}} < \infty$, wp $\to$ 1. Conditional on this event, step ii. gives $E_n s[\widehat{f}(W,\vartheta_n)'\gamma] \xrightarrow{p} Es[f(W,\vartheta_0)'\gamma] < \infty$. Similarly take $\gamma$ in $\bar{F}^c$, where $\bar{F}$ denotes the closure of $F$. The analogous argument delivers, $E_n s[\widehat{f}(W,\vartheta_n)'\gamma] \xrightarrow{p} Es[f(W,\vartheta_0)'\gamma] = \infty$. So iii. follows by taking all the rationals not in the boundary of $F$. ■

# References


ABADIE, A. (1995): "Changes in Spanish Labor Income Structure During the 1980s: A Quantile Regression Approach," CEMPFI Working Paper No. 9521.

——— (2001): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," Harvard, mimeo.

ABADIE, A., J. ANGRIST, AND G. IMBENS (2001): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, p. forthcoming.

AMEMIYA, T. (1975): "The nonlinear limited-information maximum-likelihood estimator and the modified nonlinear two-stage least-squares estimator," *J. Econometrics*, 3(4), 375–386.

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

——— (1982): "Two Stage Least Absolute Deviations Estimators," *Econometrica*, 50, 689–711.

——— (1985): *Advanced Econometrics*. Harvard University Press.

ANDREWS, D. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden. North Holland.

ANDREWS, D. W. K. (1995): "Nonparametric kernel estimation for semiparametric models," *Econometric Theory*, 11(3), 560–596.

ANDREWS, D. W. K., AND Y.-J. WHANG (1990): "Additive interactive regression models: circumvention of the curse of dimensionality," *Econometric Theory*, 6(4), 466–479.

ANGRIST, J., AND A. KRUEGER (1992): "Age at School Entry," *JASA*, 57, 11–25.

BUCHINSKY, M. (1994): "Changes in U.S. wage structure 1963-87: An application of quantile regression," *Econometrica*, 62, 405–458.

——— (1998): "Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research," *Journal of Human Resources*, 33(1), 88–126.

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

CHEN, L., AND S. PORTNOY (1996): "Two-stage regression quantiles and two-stage trimmed least squares estimators for structural equation models," *Comm. Statist. Theory Methods*, 25(5), 1005–1032.

CHRISTOFFERSEN, P. F., J. HAHN, AND A. INOUE (1999): "Testing, Comparing, and Combining Value-at-Risk Measures," SSRN working paper.

DAS, M. (2001): "Instrumental Variable Estimation of Nonparametric Models with Discrete Endogenous Regressors," mimeo.

DOKSUM, K. (1974): "Empirical probability plots and statistical inference for nonlinear models in the two-sample case," *Ann. Statist.*, 2, 267–277.

HECKMAN, J. (1990): "Varieties of Selection Bias," *American Economic Review, Papers and Proceedings*, 80, 313–338.

HECKMAN, J., AND R. ROBB (1986): "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in *Drawing Inference from Self-Selected Samples*, ed. by H. Wainer, pp. 63–107. Springer-Verlag, New York.

HECKMAN, J. J., AND J. SMITH (1997): "Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts," *Rev. Econom. Stud.*, 64(4), 487–535, With the assistance of Nancy Clements, Evaluation of training and other social programmes (Madrid, 1993).

HOGG, R. V. (1975): "Estimates of Percentile Regression Lines Using Salary Data," *Journal of the American Statistical Association*, 70(349), 56-59.

HONG, H., AND E. TAMER (2001): "Estimation of Censored Regression with Endogeneity," Preprint.

HOROWITZ, J. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Models," *Econometrica*, 60(3).

IMBENS, G. W. (1997): "One-step estimators for over-identified generalized method of moments models," *Rev. Econom. Stud.*, 64(3), 359–383.

IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

IMBENS, G. W., AND W. K. NEWEY (2001): "Estimation of Nonparametric Triangular Simultaneous Equations," mimeo.

KITAMURA, Y., AND M. STUTZER (1997): "An information-theoretic alternative to generalized method of moments estimation," *Econometrica*, 65(4), 861–874.

KNIGHT, K. (1999): "Epi-convergence and Stochastic Equisemicontinuity," Preprint.

KOENKER, R. (1994): "Confidence intervals for regression quantiles," in *Asymptotic statistics (Prague, 1993)*, pp. 349–359. Physica, Heidelberg.

KOENKER, R. (1998): "Treating the treated, varieties of causal analysis," a note.

KOENKER, R., AND G. S. BASSETT (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

KOENKER, R., AND Y. BILIAS (2001): "Quantile Regression for Duration Data: A Reappraisal of the Pennsylvania Reemployment Bonus Experiments," *Empirical Economics*, to appear, also at `http://www.econ.uiuc.edu/~roger/research/home.html`.

KOENKER, R., AND K. HALLOCK (2000): "Quantile Regression: An Introduction," *Journal of Economic Perspectives*, forthcoming.

LEHMANN, E. L. (1974): *Nonparametrics: statistical methods based on ranks*. Holden-Day Inc., San Francisco, Calif., With the special assistance of H. J. M. d'Abrera, Holden-Day Series in Probability and Statistics.

MACURDY, T., AND C. TIMMINS (1998): "Application of Smoothed Quantile Estimation," paper presented at Quantile Regression Conference in Konstanz, mimeo.

MANSKI, C. F. (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313–333.

——— (1988): "Ordinal utility models of decision making under uncertainty," *Theory and Decision*, 25(1), 79–104.

NEWEY, W. K. (1990): "Efficient instrumental variables estimation of nonlinear models," *Econometrica*, 58(4), 809–837.

——— (1997): "Convergence rates and asymptotic normality for series estimators," *J. Econometrics*, 79(1), 147–168.

NEWEY, W. K., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in Engle, R.F. and D. McFadden (eds), Handbook of Econometrics, Vol. IV.

NEWEY, W. K., AND J. L. POWELL (1990): "Efficient estimation of linear and type I censored regression models under conditional quantile restrictions," *Econometric Theory*, 6(3), 295–317.

——— (2001): "Nonparametric Instrumental Variable Estimation," Preprint.

NEWEY, W. K., AND R. SMITH (2001): "Higher Order Properties of Generalized Empirical Likelihood," working paper, MIT.

PORTNOY, S., AND R. KOENKER (1997): "The Gaussian Hare and the Laplacian Tortoise," *Statistical Science*, 12, 279–300.

POWELL, J. L. (1986): "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143–155.

ROBINS, J. M., AND A. A. TSIATIS (1991): "Correcting for noncompliance in randomized trials using rank preserving structural failure time models," *Comm. Statist. Theory Methods*, 20(8), 2609–2631.

ROCKAFELLAR, R. T., AND R. B. WETS (1998): *Variational Analysis*. Springer-Verlag, Berlin.

VAN DER VAART, A. W. (1998): *Asymptotic statistics*. Cambridge University Press, Cambridge.

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak convergence and empirical processes*. Springer-Verlag, New York.

VYTLACIL, E. (2000): "Semiparametric Identification of the Average Treatment Effect in Nonseparable Models," mimeo.

——— (2001): "Selection Model and LATE model:Equivalence Result," *Econometrica*, forthcoming.