INSTRUMENTAL VARIABLES ESTIMATION OF QUANTILE TREATMENT EFFECTS

Alberto Abadie
Joshua D. Angrist
Guido W. Imbens

Technical Working Paper **229**

TECHNICAL WORKING PAPER SERIES

INSTRUMENTAL VARIABLES ESTIMATION OF QUANTILE TREATMENT EFFECTS

Alberto Abadie
Joshua D. Angrist
Guido W. Imbens

Technical Working Paper 229
http://www.nber.org/papers/T0229

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 1998

We thank Gary Chamberlain, Jerry Hausman, Whitney Newey, James Orlin and seminar participants at Berkeley, MIT-Harvard and Penn for helpful comments and discussions. Thanks also go to Moshe Buchinsky and Gary Chamberlain for providing us with their MATLAB code for quantile regression. Abadie acknowledges financial support from the Bank of Spain. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

Instrumental Variables Estimation
of Quantile Treatment Effects
Alberto Abadie, Joshua D. Angrist
and Guido W. Imbens
NBER Technical Working Paper No. 229
March 1998
JEL Nos. C13, C14, C31, J13

ABSTRACT

This paper introduces an instrumental variables estimator for the effect of a binary treatment on the quantiles of potential outcomes. The quantile treatment effects (QTE) estimator accommodates exogenous covariates and reduces to quantile regression as a special case when treatment status is exogenous. Asymptotic distribution theory and computational methods are derived. QTE minimizes a piecewise linear objective function for which a local minimum can be obtained using a modified Barrodale-Roberts algorithm. The QTE estimator is illustrated by estimating the effect of childbearing on the distribution of family income.

Alberto Abadie
Department of Economics, E52-391
Massachusetts Institute of Technology
50 Memorial Drive
Cambridge, MA 02139
aabadie@mit.edu

Joshua D. Angrist
Department of Economics, E52-353
Massachusetts Institute of Technology
50 Memorial Drive
Cambridge, MA 02139
and NBER
angrist@mit.edu

Guido W. Imbens
Department of Economics
University of California, Los Angeles
Los Angeles, CA 90094
and NBER
imbens@econ.ucla.edu

1. INTRODUCTION

Understanding the effect of an event or intervention on distributions of outcomes is of fundamental importance in many areas of empirical economic research. A leading example in labor economics is the impact of union status on the distribution of earnings. One of the earliest studies of the distributional consequences of unionism is Freeman (1980), while more recent studies include Card (1996), and DiNardo, Fortin, and Lemieux (1996), who have asked whether changes in union status can account for a significant fraction of increasing wage inequality in the 1980s. Another application where distribution effects are important is the range of government training programs funded under the Job Training Partnership Act (JTPA). Policy makers hope that subsidized training programs will work to reduce earnings inequality by raising the lower deciles of the earnings distribution and reducing poverty (Lalonde (1995), US Department of Labor (1995)).

Although the importance of distribution effects is widely acknowledged, most evaluation research focuses on mean outcomes, probably because the statistical techniques required to estimate effects on means are easy to use. Many econometric models also implicitly restrict treatment effects to operate in the form of a simple "location shift", in which case the mean effect captures the impact of treatment at all quantiles.[1] Of course, the impact of treatment on a distribution is easy to assess when treatment status is assigned in controlled randomized trials and there is perfect compliance with treatment assignment. Because randomization guarantees that individual outcomes in the treatment group are directly comparable to those of individuals the control group, valid causal inferences can be obtained by simply comparing the distributions of interest in treatment and control groups. The problem of how to draw inferences about distributions in observational studies with non-ignorable or non-random assignment is more difficult, however, and has received less attention.[2]

[1] Traditional simultaneous equations models and the two-stage least absolute deviation estimators introduced by Amemiya (1982) and Powell (1983) fall into this category.

[2] Discussions of average treatment effects include Rubin (1977), Rosenbaum and Rubin (1983), Heckman and Robb (1985), and Imbens and Angrist (1994). Heckman, Smith and Clements (1997), Manski (1994), Imbens and Rubin (1997) and Abadie, (1997a) discuss effects on distributions.

In this paper, we show how to use a source of exogenous variation in treatment status – an instrumental variable – to estimate the effect of treatment on the quantiles of the distribution of the outcome of interest in non-randomized studies (or in randomized studies with imperfect compliance). The treatment effects in this framework are only identified for a subpopulation. We refer to individuals in this subpopulation as *compliers* because in randomized trials with partial compliance, these are people who comply with the treatment protocol.[3] More generally, the subpopulation of compliers consists of individuals whose treatment status can be changed by the instrument. The identification results underlying this *local average treatment effects* (LATE) approach to instrumental variables (IV) models were first established by Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). Imbens and Rubin (1997) extended these results to the identification of the effect of treatment on the distribution of outcomes for compliers, although they did not develop simple estimators, or a scheme for estimating the effect of treatment on quantiles.

We demonstrate our approach to estimating quantile treatment effects (QTE) in an empirical example based on Angrist and Evans (1998), who use the sex composition of the first two siblings as an instrument for the effect of having a third child on labor supply and earnings. This instrument is based on well-documented parental preferences for a mixed sibling sex composition. In particular, parents of two boys or two girls are significantly and substantially more likely to have a third child than parents of a boy-girl pair. Since sex is virtually randomly assigned at birth, it seems unlikely that an indicator for same-sex sibling pairs is associated with parents' labor market outcomes for reasons other than changes in family size. Angrist and Evans' IV estimates show that childbearing reduces labor supply and earnings much more for some groups of women than for others, so it is interesting to consider the effect of childbearing on the distribution of family income. The QTE estimates reported here show that childbearing reduces family income at all quantiles below 0.9, with

[3]See, e.g., Bloom et al. (1997, p. 555), who discuss instrumental variables estimation of average effects for compliers (treatment group members who would not have enrolled if assigned to the control group) in their analysis of the Job Training Partnership Act. An alternative approach develops bounds on average treatment effects for the overall population rather than focusing on compliers. See Manski (1990), Robins (1989) or Balke and Pearl (1997).

effects at low quantiles larger than those estimated using quantile regression.

The paper is organized as follows. Section 2 presents a lemma that provides a foundation for the identification of quantile treatment effects. Section 3 outlines the QTE estimation strategy, which allows for a binary endogenous regressor and reduces to the standard Koenker and Basset (1978) approach when the regressor is exogenous. This section also presents distribution theory based on empirical processes (for a review, see Andrews (1994)). Section 4 discusses the empirical example. The QTE estimator can be computed by minimizing a piecewise linear objective function using a modification of the Barrodale-Roberts algorithm widely used for quantile regression (see, e.g., Buchinsky (1994) and Chamberlain (1991)). Details related to the computation of estimates and the estimation of asymptotic standard errors are discussed in appendices.

2. Conceptual Framework

Throughout the paper, the setup is as follows. The data consist of n observations on a continuously distributed outcome variable, Y, a binary treatment indicator D, and a binary instrument, Z. For example, in a study of the effect of unions, Y is a measure of wages or earnings, D indicates union status, and Z is an instrument for union status, say a dummy indicating individuals who work in firms that were subject to union organizing campaigns (Lalonde, Marschke and Troske (1996)). Another example is the Angrist (1990) study of effects of veteran status, where Y is annual earnings, D indicates veteran status, and Z is an indicator of draft-lottery eligibility. In Angrist and Evans (1998) and the empirical example used here, Y is log family income in families with 2 or more children, D indicates families with more than two children, and Z indicates families where the first two children are of the same sex. We also allow for an $h \times 1$ vector of covariates, X.

As in Rubin (1974, 1977) and earlier work on instrumental variables estimation of causal effects (Imbens and Angrist (1994), Angrist, Imbens, and Rubin (1996)), we define the causal effects of interest in terms of potential outcomes. In particular, we define potential outcomes indexed against Z and D, Y_{ZD}, and potential treatment status indexed against

Z, D_Z. Potential outcomes describe possibly counterfactual states of the world. Thus, D_1 tells us what value D would take if Z were equal to 1, while D_0 tells us what value D would take if Z were equal to 0. Similarly, Y_{zd} tells us what someone's outcome would be if they have $Z = z$ and $D = d$. The objects of causal inference are features of the distribution of potential outcomes, possibly restricted to particular subpopulations.

The observed treatment status is:

$$D = D_0 + (D_1 - D_0) \cdot Z.$$

In other words, if $Z = 1$, then D_1 is observed, while if $Z = 0$, then D_0 is observed. Likewise, the observed outcome variable is:

$$Y = [Y_{00} + (Y_{01} - Y_{00}) \cdot D_0] \cdot (1 - Z) + [Y_{10} + (Y_{11} - Y_{10}) \cdot D_1] \cdot Z. \tag{1}$$

The reason why causal inference is difficult is that although we think of all possible counterfactual outcomes as being defined for everyone, only one potential treatment status and one potential outcome are ever observed for any one person.[4]

2.1. Principal Assumptions

The principal assumptions underlying the potential outcomes framework for IV are stated below:

Assumption 1 *With probability one,*

(i) *(*Independence*)* $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0)$ *is jointly independent of* Z *given* X.

(ii) *(*Exclusion*)* $P(Y_{1D} = Y_{0D}|X) = 1$.

[4]The idea of potential outcomes appears in labor economics in discussions of the effects of union status. See, for example, Lewis' (1986) survey of research on union relative wage effects (p. 2):

> At any given date and set of working conditions, there is for each worker a *pair* of wage figures, one for unionized status and the other for nonunion status. Unfortunately, only one wage figure is observable, namely, that which corresponds to the worker's actual union status at the date. The other wage figure must be estimated, and the estimation task is formidable.

(iii) *(NON-TRIVIAL ASSIGNMENT)* $P(Z = 1|X) \in (0, 1)$.

(iv) *(FIRST-STAGE)* $E[D_1|X] \neq E[D_0|X]$.

(v) *(MONOTONICITY)* $P(D_1 \geq D_0|X) = 1$.

Assumption 1(ii) means we can define $Y_D \equiv Y_{1D} = Y_{0D}$, and this is the notation we use in the remainder of the paper. The random variable Y_1 represents potential outcomes if treated, while Y_0 represents potential outcomes if not. Assumptions 1(i) and 1(ii) are analogous to the conventional instrumental variables assumptions of instrument-error independence and an exclusion restriction. Assumption 1(i) can be thought of as saying that Z is "as good as randomly assigned" given X. Assumption 1(ii) means that the only effect of Z on Y is through D.

Assumption 1(iii) requires that the conditional distribution of the instrument not be degenerate. The relationship between instruments and treatment assignment is restricted in two ways. First, as in simultaneous equations models, we require that there be a relationship between D and Z; this is stated in Assumption 1(iv). Second, Imbens and Angrist (1994) have shown that Assumption 1(v) guarantees identification of a meaningful average treatment effect in any model with heterogeneous potential outcomes that satisfies 1(i)-1(iv). This monotonicity assumption means that the instrument can only affect D in one direction. Monotonicity is plausible in most applications and it is automatically satisfied by linear single-index models for treatment assignment.[5]

The inferential problem in evaluation research requires a comparison of observed and unobserved outcomes. For example, many evaluation studies focus on estimating the difference between the average outcomes of the treated (which is observed) and what this average would have been in the absence of treatment (which is counterfactual). Outside

[5]A linear single-index model specification for participation is

$$D = 1\{\lambda_0 + Z \cdot \lambda_1 - \eta > 0\} = \begin{cases} 1 & \text{if } \lambda_0 + Z \cdot \lambda_1 - \eta > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where λ_0 and λ_1 are parameters and η is an error term that is independent of Z. Then $D_0 = 1\{\lambda_0 > \eta\}$, $D_1 = 1\{\lambda_0 + \lambda_1 > \eta\}$, and either $D_1 \geq D_0$ or $D_0 \geq D_1$ for everyone.

of a randomized trial, the difference in average outcomes by observed treatment status is typically a biased estimate of this effect:

$$E[Y_1|D=1] - E[Y_0|D=0] = \{E[Y_1|D=1] - E[Y_0|D=1]\} + \{E[Y_0|D=1] - E[Y_0|D=0]\}.$$

The first term in brackets is the average effect of the treatment on the treated, which can also be written as $E[Y_1 - Y_0|D=1]$ since expectation is a linear operator; the second is the bias term. For example, comparisons of earnings by union status are biased if the average earnings of nonunion workers do not provide a guide as to what the average earnings of union members would have been if they had not been unionized.

An instrumental variable solves the problem of identifying causal effects for a group of individuals whose treatment status is affected by the instrument. The following result (Imbens and Angrist (1994)) captures this idea formally:

Theorem 1 *Under Assumption 1 (and assuming that the relevant expectations are finite)*

$$\frac{E[Y|Z=1,X] - E[Y|Z=0,X]}{E[D|Z=1,X] - E[D|Z=0,X]} = E[Y_1 - Y_0|X, D_1 > D_0].$$

The parameter identified in Theorem 1 is called Local Average Treatment Effect (LATE). We refer to individuals for whom $D_1 > D_0$ as *compliers* because in a randomized clinical trial with partial compliance, this group would consist of individuals who comply with the treatment protocol whatever their assignment. In other words, the set of compliers is the set of individuals who were affected by the experiment induced by Z. Note that individuals in this set cannot usually be identified (i.e., we cannot name the people who are compliers) because we never observe both D_1 and D_0 for any one person. On the other hand, we can identify certain individuals as non-compliers, as will be shown below.[6]

[6]In the special case when $D_0 = 0$ for everyone, such as in a randomized trial with non-compliance in the treatment group only, all treated units (i.e. units with $D = 1$) are compliers. In such cases, LATE is the effect of treatment on the treated (Imbens and Angrist (1994)).

2.2. Treatment Status is Ignorable For Compliers

The purpose of randomization is to ensure that treatment assignment is independent of potential outcomes, possibly after conditioning on some covariates. Independence of treatment and potential outcomes is sometimes called *ignorable* treatment assignment (Rubin (1978)). Ignorability implies that differences in the distribution of outcomes by treatment status can be attributed to the treatment. Although we have assumed that the instruments are independent of potential outcomes, the actual treatment received is not ignorable. Nevertheless, Theorem 1 shows that instrumental variables methods identify an average causal effect for the group whose treatment status is affected by the instrument, the compliers.

The compliers concept is the heart of the LATE framework and provides a simple explanation for why instrumental variables methods work in this context. To see this, suppose initially that we could know who the compliers are. For these people, Z=D, since it always true that $D_1 > D_0$. This observation plus Assumption 1 leads to the following lemma:

Lemma 1 *Given Assumption 1 and conditional on X, the treatment status, D, is ignorable for compliers:* $(Y_1, Y_0) \amalg D|X, D_1 > D_0$.

This follows from Assumptions 1(i) and 1(ii), because these assumptions imply that $(Y_1, Y_0, D_1, D_0) \amalg Z|X$, so $(Y_1, Y_0) \amalg Z|X, D_1 = 1, D_0 = 0$. When $D_1 = 1$ and $D_0 = 0$, D can be substituted for Z.

A consequence of Lemma 1 is that, for compliers, comparisons of means by treatment status produce LATE even though treatment assignment is not ignorable in the population:

$$E[Y|D = 1, D_1 > D_0, X] - E[Y|D = 0, D_1 > D_0, X] = E[Y_1 - Y_0|X, D_1 > D_0]. \quad (2)$$

Of course, as it stands, Lemma 1 is of no practical use because the subpopulation of compliers is not identified. The reason is that we cannot observe D_1 and D_0 for the same individual. To make Lemma 1 operational, we begin by defining the following function of

D, Z and X:

$$\kappa = \kappa(D, Z, X) = 1 - \frac{D \cdot (1 - Z)}{1 - E[Z|X]} - \frac{Z \cdot (1 - D)}{E[Z|X]}. \tag{3}$$

Note that κ equals one when $D = Z$, otherwise κ is negative. This function is useful because it "identifies compliers" in the following average sense (Abadie (1997b)):

Lemma 2 *Let $\psi(Y, D, X)$ be any measurable real function of (Y, D, X). Then, given Assumption 1,*

$$\frac{E[\kappa \cdot \psi(Y, D, X)]}{P(D_1 > D_0)} = E[\psi(Y, D, X)|D_1 > D_0].$$

To see this, define two groups in the population besides compliers: *always-takers* are individuals who have $D_1 = D_0 = 1$, while *never-takers* have $D_1 = D_0 = 0$. Because of monotonicity, the expectation of ψ given X can be written in terms of expectations for compliers, always-takers, and never-takers as follows:

$$\begin{aligned} E[\psi|X] = {} & E[\psi|X, D_1 > D_0] \cdot P(D_1 > D_0|X) \\ & + E[\psi|X, D_1 = D_0 = 1] \cdot P(D_1 = D_0 = 1|X) \\ & + E[\psi|X, D_1 = D_0 = 0] \cdot P(D_1 = D_0 = 0|X). \end{aligned}$$

Rearranging terms gives,

$$\begin{aligned} E[\psi|X, D_1 > D_0] = {} & \frac{1}{P(D_1 > D_0|X)} \{E[\psi|X] - E[\psi|X, D_1 = D_0 = 1] \cdot P(D_1 = D_0 = 1|X) \\ & - E[\psi|X, D_1 = D_0 = 0] \cdot P(D_1 = D_0 = 0|X)\}. \end{aligned} \tag{4}$$

Now, by monotonicity we know that all individuals with $Z = 0$ and $D = 1$ must be never-takers. Likewise, those with $Z = 0$, $D = 1$ must be always-takers. Moreover, since Z is ignorable given X, we have

$$\begin{aligned} E[\psi|X, D_1 = D_0 = 1] &= E[\psi|X, D = 1, Z = 0] \\ &= \frac{1}{P(D = 1|X, Z = 0)} \cdot E\left[\frac{D \cdot (1 - Z)}{P(Z = 0|X)} \cdot \psi \middle| X\right], \end{aligned}$$

and

$$E[\psi|X, D_1 = D_0 = 0] = E[\psi|X, D = 0, Z = 1]$$
$$= \frac{1}{P(D = 0|X, Z = 1)} \cdot E\left[\frac{(1 - D) \cdot Z}{P(Z = 1|X)} \cdot \psi \middle| X\right].$$

Monotonicity and ignorability of Z given X can also be used to identify the proportions of always-takers and never-takers in the population

$$P(D_1 = D_0 = 1|X) = P(D = 1|X, Z = 0),$$

$$P(D_1 = D_0 = 0|X) = P(D = 0|X, Z = 1).$$

Next plug these results into equation 4, and manipulate to obtain

$$E[\psi|X, D_1 > D_0] = \frac{1}{P(D_1 > D_0|X)} \cdot E\left[\left(1 - \frac{D \cdot (1 - Z)}{P(Z = 0|X)} - \frac{(1 - D) \cdot Z}{P(Z = 1|X)}\right) \cdot \psi \middle| X\right].$$

Applying Bayes' theorem and integrating over X completes the argument.

This derivation shows how monotonicity and ignorability of Z identify expectations for compliers. Monotonicity allows us to divide the population into three subpopulations: compliers, always-takers and never-takers. The average ψ for compliers is then expressed as a function of the average ψ in the population and the correspondent averages for always-takers and never-takers. Finally, ignorability of Z can be used to identify expectations for always-takers and never-takers, so the same expectations are also identified for compliers.

An implication of Lemma 2 is that any statistical characteristic that uniquely solves a moment condition involving (Y, D, X) is identified for compliers. This point is explored in detail in Abadie (1997b).[7] In next section, Lemma 2 is used to identify the causal effect of a treatment on the quantiles of Y_0 and Y_1.

[7]For example, if we define μ and α as

$$(\mu, \alpha) = \text{argmin}_{(m,a)} E[(Y - m - aD)^2|D_1 > D_0],$$

then, $\mu = E[Y_0|D_1 > D_0]$, and $\alpha = E[Y_1 - Y_0|D_1 > D_0]$, so that α is LATE (although μ is not the same intercept that is estimated by conventional IV methods). By Lemma 2, (μ, α) also minimizes $E[\kappa \cdot (Y - m - aD)^2]$.

3. Quantile Treatment Effects

3.1. The QTE Model

Just as conventional IV estimators specialize to ordinary least squares (OLS) in the case where treatment status is an exogenous variable, the QTE estimator is designed so that it collapses to conventional quantile regression (Koenker and Bassett (1978)) when there is no instrumenting. This is accomplished by using a model that restricts the effect of covariates on quantiles to be linear and additive at each quantile.[8]

Assume that the conditional quantiles of the potential outcomes for compliers can be written as

$$\begin{aligned} Q_\theta(Y_0|X, D_1 > D_0) &= X'\beta_\theta, \\ Q_\theta(Y_1|X, D_1 > D_0) &= \alpha_\theta + X'\beta_\theta, \end{aligned} \tag{5}$$

where θ is a quantile index in $(0, 1)$. Recall that $Y = D \cdot Y_1 + (1 - D) \cdot Y_0$. By Lemma 1, D is independent of (Y_0, Y_1) given X and $D_1 > D_0$. The conditional quantile function of Y given D and X for compliers can therefore be written:

$$Q_\theta(Y|X, D, D_1 > D_0) = \alpha_\theta D + X'\beta_\theta.$$

Note that the parameter of primary interest in this model, α_θ, gives the difference in θ-quantiles of Y_1 and Y_0, and not the quantiles of the difference $(Y_1 - Y_0)$. Although, the procedure outlined here can be used to learn whether a training program causes the 10th percentile of the distribution of earnings to move up, we cannot know whether people who were originally at the 10th percentile experienced an increase in earnings. We focus on the marginal distributions of potential outcomes because it is these that would be identified by a randomized trial conducted in the complier population. The parameters revealed by an actual experiment seem like a natural benchmark for identification in observational studies. Also, economists making social welfare comparisons typically use differences in

[8]For expositional purposes, we follow most of the literature on quantile regression and treat the linear model as a literal specification for conditional quantiles. However, the standard errors derived below are robust to misspecification. Chamberlain (1991) discusses quantile regression models where the linear model is viewed as an approximation.

distributions and not the distribution of differences (see, e.g., Atkinson (1970)).[9]

The parameters of the conditional quantile functions in (5) can be expressed as (see Bassett and Koenker (1982)):

$$(\alpha_\theta, \beta_\theta) \equiv \text{argmin}_{(\alpha,\beta)} E[\rho_\theta(Y - \alpha D - X'\beta)|D_1 > D_0],$$

where $\rho_\theta(\lambda)$ is the check function, i.e., $\rho_\theta(\lambda) = (\theta - 1\{\lambda < 0\}) \cdot \lambda$ for any real λ. Because compliers are not identifiable, we cannot use this formulation directly to estimate α_θ and β_θ. However, by Lemma 2, α_θ and β_θ can be defined as

$$(\alpha_\theta, \beta_\theta) \equiv \text{argmin}_{(\alpha,\beta)} E[\kappa \cdot \rho_\theta(Y - \alpha D - X'\beta)].$$

Assume now that we have a random sample $\{y_i, d_i, x_i, z_i\}_{i=1}^n$. Then, following the analogy principle (Manski (1988)) we can estimate the parameters of interest by

$$(\hat{\alpha}_\theta, \hat{\beta}_\theta) \equiv \text{argmin}_{(\alpha,\beta)} \sum_{i=1}^{n} \kappa_i \cdot \rho_\theta(y_i - \alpha d_i - x_i'\beta). \tag{6}$$

Note that when the treatment is assumed to be ignorable and D itself is used as an instrument, then $\kappa_i = 1$ for all $i = 1, ..., n$ and the problem above simplifies to conventional quantile regression.

It remains to discuss minimization of equation (6) in practice. Following Powell's (1994) approach to a similar weighted quantile regression problem, we first estimate $E(Z_i|X_i)$ and then plug this estimate into κ_i. The minimization is then accomplished using a modification of the Barrodale-Roberts (1973) algorithm for quantile regression that exploits the quasi - Linear Programming (LP) nature of this problem. Results were checked using the Nelder-Mead algorithm. Details are given in Appendix II.

3.2. Distribution Theory

This section contains asymptotic results for the QTE estimator. Proofs are given in Appendix I. The next assumption formalizes the presentation of the model outlined in the previous section.

[9]Heckman, Smith and Clements (1997) discuss models where features of the distribution of the difference $(Y_1 - Y_0)$ are identified.

3.2.1. IDENTIFICATION

Assumption 2

There exist unique $\alpha \in \Lambda$ and $\beta \in \Theta$ such that

(i) *The θth quantile of the conditional distribution of Y_0 given X and $D_1 > D_0$ is unique and equal to $X'\beta$.*

(ii) *The θth quantile of the conditional distribution of Y_1 given X and $D_1 > D_0$ is unique and equal to $\alpha + X'\beta$.*

Theorem 2 *(IDENTIFICATION)*

Suppose Assumptions 1 and 2 hold. Then the argmin of

$$E\left[\left(1 - \frac{(1-Z)\cdot D}{Pr(Z=0|X)} - \frac{Z\cdot(1-D)}{Pr(Z=1|X)}\right)\cdot\left(Y - aD - X'b\right)\cdot\left(\theta - 1\{Y - aD - X'b < 0\}\right)\right] \tag{7}$$

over $(a,b) \in (\Lambda \times \Theta)$ is unique and equal to (α, β).

3.2.2. CONSISTENCY

Assumption 3

(i) *Denote $W = (Y, D, X', Z)'$. The random variables $\{W_i\}_{i=1}^n$ are independent and identically distributed.*

(ii) *For a unique $\gamma \in \Gamma$, with Γ being a subset of $\mathcal{R}^L$,*

$$Pr(Z = 1|X) = P(X;\gamma),$$

where for all X and g, $P(X;g)$ is bounded away from zero and one, and is continuous in $g \in \Gamma$.

(iii) *There is a consistent estimator $\hat{\gamma}$ of γ.*

(iv) *$E|Y| < \infty$ and $E\|X\| < \infty$.*

(v) *Λ and Θ are compact.*

(vi) *The function $1\{Y - aD - X'b < 0\}$ is continuous at each (a, b) in $\Lambda \times \Theta$ with probability one.*

Theorem 3 *(CONSISTENCY) Suppose that Assumptions 1-3 hold. Then*

$$(\hat{\alpha}, \hat{\beta}) \equiv argmin_{a \in \Lambda, b \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{d_i \cdot (1 - z_i)}{(1 - P(x_i; \hat{\gamma}))} - \frac{(1 - d_i) \cdot z_i}{P(x_i; \hat{\gamma})} \right) \cdot \rho_\theta(y_i - ad_i - x_i'b),$$

is consistent for (α, β).

3.2.3. ASYMPTOTIC NORMALITY

Assumption 4

(i) *Denote $V = (Z, X')'$. The estimator $\hat{\gamma}$ of γ solves*

$$\frac{1}{n} \sum_{i=1}^{n} q(v_i, g) = 0$$

for g with probability approaching one. The vector of functions $q(\cdot, g)$ has the same dimension as γ and $E\|q(V, \gamma)\|^2 < \infty$.

(ii) *$\alpha \in int(\Lambda)$, $\beta \in int(\Theta)$ and $\gamma \in int(\Gamma)$.*

(iii) *$E\|X\|^2 < \infty$.*

(iv) *There exists a neighborhood $\mathcal{B}$ of (α, β, γ) such that for $(a, b, g) \in \mathcal{B}$*

- **a.** *$P(X; g)$ is continuously differentiable with bounded derivative.*
- **b.** *The vector of functions $q(\cdot, g)$ is differentiable with respect to g.*
- **c.** *The conditional distribution of Y given X, D and Z is absolutely continuous at $aD + X'b$ with respect to the Lebesgue measure. The probability density function $f_{Y|Z,D,X}(aD + X'b)$ is bounded in $\mathcal{B}$ and continuous with probability one at $\alpha D + X'\beta$.*

Theorem 4 *(*ASYMPTOTIC NORMALITY*) Denote* $\delta = (\alpha, \beta)$, $\hat{\delta} = (\hat{\alpha}, \hat{\beta})$ *and*

$$m(W_i, l, g) = \begin{pmatrix} D_i \\ X_i \end{pmatrix} \cdot \kappa_i(g) \cdot (\theta - 1\{Y_i - aD_i - X_i'b < 0\}).$$

where $l = (a, b)$. *Under assumptions 1 to 4,*

$$\sqrt{n}\left(\hat{\delta} - \delta\right) \xrightarrow{d}$$
$$\mathcal{N}\left(0, M_\delta^{-1} E\left[\left\{m(W_i, \delta, \gamma) - M_\gamma Q_\gamma^{-1} q(W_i, \gamma)\right\}\left\{m(W_i, \delta, \gamma) - M_\gamma Q_\gamma^{-1} q(W_i, \gamma)\right\}'\right] M_\delta^{-1\,\prime}\right), \tag{8}$$

where $M_\delta = \partial E\left[m(W, \delta, \gamma)\right]/\partial l'$, $M_\gamma = E\left[\partial m(W, \delta, \gamma)/\partial g'\right]$ *and* $Q_\gamma = E\left[\partial q(V_i, \gamma)/\partial g'\right]$.

4. APPLICATION

In the empirical example, Y is *Log family income* for a sample of women with two or more children, D indicates women with three or more children (*More than two kids*), and Z indicates whether the first two children are both boys or both girls (*Same sex*). The vector of covariates consists of a constant, mother's age, mother's age at first birth, mother's high school graduation status, mother's post high school education status, a dummy for blacks and hispanics, and a dummy for firstborn male children. The relationship of interest is the causal effect of childbearing on family income. If fertility and earnings are jointly determined, as suggested by economic theory (see, e.g., Browning (1992)), OLS or quantile regression estimates of this relationship are not likely to have a causal interpretation. Our empirical example is based on Angrist and Evans (1998), who show that parents whose first two children are both boys or both girls are 6-7 percentage points more likely to go on to have a third child than are parents whose first two children are mixed gender. This relationship suggests that *Same sex* can be used as an instrument for *More than two kids.* The data used here consist of a sample of 346,929 women aged 21-35 in the 1990 Census Public Use Microdata sample (PUMS). For more detailed information about the data see the Angrist and Evans paper.

The basic finding in earlier work using *Same sex* as an instrument for *More than two kids* is that the third child appears to cause a large reduction in average female labor supply

and earnings. On the other hand, while this reduction is especially large for less educated women, it is not observed for more educated women. And female-headed households are naturally most affected by a decline in female earnings. The fact that the impact of childbearing varies with these observed characteristics suggests that childbearing may affect the distribution of family income in ways other than through an additive shift.

Ordinary least squares (OLS) estimates and quantile regression (QR) estimates of the relationship between *Log family income* and *More than two kids* are reported in Table I. Column (1) of the table also shows the mean of each variable used in the analysis. Approximately 36 percent of the sample had 3 or more children. Half of firstborn children are male and half of the first-two sibling pairs are same sex. The OLS estimate of the effect of having a third child in family income is -.092. Quantile regression estimates show an effect at the median of -.066, with smaller effects at higher quantiles and larger effects at lower quantiles. The largest quantile regression estimate is -.098 at the 0.1 quantile. All of these coefficients are estimated very precisely.[10]

The relationship between sibling sex composition and childbearing is captured in the first column of Table II, which reports first-stage coefficient estimates for the dummy endogenous regressor *More than two kids.* Parents with a same sex sibling pair are 6.4 percentage points more likely to go on to have a third child. There is also some evidence of an association between having a firstborn male child and reduced fertility, though this effect (the coefficient on *Boy 1st*) is very small. The conventional two-stage least squares (2SLS) estimate of the effect of *More than two kids* using *Same sex* as an instrument is -.122, with a standard error of .069.

The QTE estimate of the effect of *More than two kids* at the median is -.065 with a standard error of .038.[11] This is smaller (in absolute value) but more precisely estimated than the 2SLS estimate. It is also remarkably similar to the corresponding quantile regression estimate at the median, though the latter is much more precisely estimated. The

[10]Asymptotic standard errors for the QR and QTE estimates were computed using kernel estimates of the conditional density of Y given D, Z and X. See Appendix III for details.

[11]For QTE, the expectations $E[Z_i|X_i = x_i]$ in κ_i were estimated using a linear model. We also experimented with non-parametric cell-by-cell estimators of those expectations obtaining similar results.

quantile regression and QTE estimates at the 0.9 quantile are also close, though the QTE estimate is not significantly different from zero at this quantile. Both of the QTE estimates at quantiles below the median are larger than the corresponding QR estimates, and much larger than either the QR or QTE estimates at the median. The QTE estimate at the 0.1 quantile is -0.18 with a standard error of .097; this is almost 6 times larger than the QTE estimate at the 0.9 quantile and 85% larger than the QR estimate at the 0.10 quantile. The QTE results therefore suggest, even more strongly than the QR estimates, that childbearing reduces the lower tail of the income distribution considerably more than other parts of the income distribution.

5. Summary and Conclusions

This paper introduces an estimator for the effect of a non-ignorable treatment on quantiles. The estimator can be used to determine whether and how an intervention affects the income distribution, or the distribution of any other variable. The QTE estimator is designed to accommodate exogenous covariates and to collapse to conventional quantile regression when the treatment is exogenous. QTE minimizes an objective function that is similar to the check function minimand for conventional quantile regression. The estimates reported here were computed using a modified Barrodale-Roberts (1973) algorithm that exploits the quasi-LP nature of the QTE minimand. As with the Iterated Linear Programming algorithm used by Buchinsky (1994) for censored quantile regression, the computational algorithm used here does not guarantee a global optimum and improving the algorithm is a natural avenue for future research.

The QTE procedure estimates a parametric conditional quantile model for individuals whose treatment status is affected by a binary instrument. Covariate effects and the treatment effect of interest are both estimated for people in this group, whom we call compliers. In many IV applications, compliers are a small proportion of the sample; in the empirical example studied here, this proportion is about 6.4 percent. This leads QTE estimates to be less precise than the corresponding QR estimates. On the other hand, the QTE estimate

of the treatment effect at the median is more precise than the conventional 2SLS estimate. This suggests that the robustness properties of conditional medians (Koenker and Bassett (1978)) may extend to the IV model.

Appendix I: Asymptotic Distribution Theory

Proof of Theorem 2:
Assumption 2 implies that

$$E\left[\Big(Y - h(D,X)\Big) \cdot (\theta - 1\{Y - h(D,X) < 0\})\Big| D_1 > D_0\right]$$

is strictly minimized by choosing $h(D,X)$ to be the θth quantile of the conditional distribution of Y given D and X, and that this quantile is uniquely equal to $\alpha D + X'\beta$. Thus, (α, β) is the unique solution to the problem

$$\min_{(a,b)\in\Lambda\times\Theta} E\left[\Big(Y - aD - X'b\Big) \cdot (\theta - 1\{Y - aD - X'b < 0\})\Big| D_1 > D_0\right]. \tag{9}$$

Then lemma 2 implies the result. $\mathcal{QED}$.

Proof of Theorem 3:
By theorem 2 the function in equation (7) is uniquely minimized at (α, β) over $(\Lambda \times \Theta)$ compact. Denote

$$f(W_i, l, g) = \kappa_i(g) \cdot (\theta - 1\{Y_i - aD_i - X_i'b < 0\}) \cdot (Y_i - aD_i - X_i'b).$$

Then,

$$\begin{aligned} sup_{l\in\Lambda\times\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} f(w_i, l, \hat{\gamma}) - E\left[f(W, l, \gamma)\right] \right\| \\ \leq sup_{l\in\Lambda\times\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} f(w_i, l, \hat{\gamma}) - E\left[f(W, l, \hat{\gamma})\right] \right\| \\ + sup_{l\in\Lambda\times\Theta} \left\| E\left[f(W, l, \hat{\gamma})\right] - E\left[f(W, l, \gamma)\right] \right\|. \end{aligned} \tag{10}$$

By assumption 3(i) the data are iid. Assumptions 3(ii) and 3(vi) imply that $f(w_i, l, g)$ is continuous at each (l, g) in $\Lambda \times \Theta \times \Gamma$ with probability one. By assumption 3(ii), $|\kappa|$ is bounded by some real number $\bar{K}$. Note that $|\theta - 1\{Y - aD - X'b < 0\}|$ is bounded by one. Since the optimization is performed over some compact space $\Lambda \times \Theta$, then there exists a finite real $\bar{l}$ such that $\|l\| \leq \bar{l}$ for all $l \in \Lambda \times \Theta$. Then $\|f(W, l, \gamma)\| \leq \bar{K} \cdot (|Y| + \bar{l} \cdot (1 + \|X\|))$. Assumption 3(iv) implies $E\left[\bar{K} \cdot (|Y| + \bar{l} \cdot (1 + \|X\|))\right] < \infty$. Then, applying Lemma 2.4 in Newey and McFadden (1994), $E[f(W, l, g)]$ is continuous at each (l, g) and

$$sup_{(l,g)\in\Lambda\times\Theta\times\Gamma} \left\| \frac{1}{n}\sum_{i=1}^{n} f(w_i, l, g) - E\left[f(W, l, g)\right] \right\| \xrightarrow{p} 0.$$

Now, the first term of the right hand side of equation (10) is $o_p(1)$. Since $\hat{\gamma} \xrightarrow{p} \gamma$ and by continuity of $E\left[f(W, l, g)\right]$, then $E\left[f(W, l, \hat{\gamma})\right] \xrightarrow{p} E\left[f(W, l, \gamma)\right]$ uniformly in l and the second term of the right hand side of equation (10) is also $o_p(1)$. Theorem 2.1 in Newey and McFadden (1994) shows that these conditions are sufficient for consistency of $(\hat{\alpha}, \hat{\beta})$. $\mathcal{QED}$.

Proof of Theorem 4:
The proof begins with a preliminary lemma:

Lemma 3 *Under assumptions 1 to 4,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(w_i, \hat{\delta}, \hat{\gamma}) = o_p(1).$$

Proof: Note that, given consistency and under assumption 4(ii), with probability approaching one we attain an interior solution for the minimization problem that produces the QTE estimator. Then, an argument similar to the proof of Lemma A.2 in Ruppert and Carroll (1980) shows that each element of $n^{-1/2}\sum_{i=1}^{n} m(w_i, \hat{\delta}, \hat{\gamma})$ is bounded in absolute value by $B_n \equiv n^{-1/2}\sum_{i=1}^{n} \bar{K}\cdot(1 + \|X\|)\cdot 1\{y_i - \hat{\alpha}d_i - x_i'\hat{\beta} = 0\}$ where $\bar{K}$ is an upper bound for $|\kappa|$, that exists by assumption 3(ii). Now assumption 4(iv) implies that $f_{Y|D,X}(aD + X'b)$ is bounded in $\mathcal{B}$ (because $P(Z|D,X) \in [0,1]$), so with probability approaching one the number of observations such that $1\{y_i - \hat{\alpha}d_i - x_i'\hat{\beta} = 0\} = 1$ is not greater than the dimension of (α, β'). Finally, $E\|X\|^2 < \infty$ implies that $E|B_n|^2 \to 0$, so $B_n \xrightarrow{p} 0$, and the lemma follows. $\mathcal{QED}$.

Now, by assumption 4(iv), $m(w_i, l, g)$ is differentiable with respect to g (it is the nondifferentiability with respect to l that is the issue) in a neigborhood $\mathcal{B}$ of (α, β, γ). Since $\hat{\alpha} \xrightarrow{p} \alpha$, $\hat{\beta} \xrightarrow{p} \beta$ and $\hat{\gamma} \xrightarrow{p} \gamma$, then for n large enough the mean value theorem applies,

$$o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(w_i, \hat{\delta}, \hat{\gamma}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(w_i, \hat{\delta}, \gamma) + \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(w_i, \hat{\delta}, \tilde{\gamma})}{\partial g'}\right]\sqrt{n}\,(\hat{\gamma} - \gamma), \qquad (11)$$

for some $\tilde{\gamma}$ in between $\hat{\gamma}$ and γ (where $\tilde{\gamma}$ differs between rows of $\partial m(w_i, \hat{\delta}, \cdot)/\partial g'$). The first equality in equation (11) follows from Lemma 3. Since (i) the data are i.i.d. (assumption 3(i)), (ii) $\partial m(W, l, g)/\partial g'$ is continuous with probability one at (δ, γ) (assumptions 3(ii) and 4(iv)), (iii) $E\left[sup_{(l,g)\in\mathcal{B}}\|\partial m(W, l, g)/\partial g'\|\right] < \infty$ (assumptions 3(ii), 4(iii) and 4(iv)), and (iv) $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ are consistent, then $n^{-1}\sum_{i=1}^{n}\partial m(w_i, \hat{\delta}, \tilde{\gamma})/\partial g' \xrightarrow{p} M_\gamma$ (see Lemma 4.3 in Newey and McFadden (1994)). Define the *empirical process*

$$\nu_n(l, g) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{m(w_i, l, g) - E\left[m(W_i, l, g)\right]\right\}.$$

Empirical processes play an important role in modern large sample theory (see Andrews (1994) for a review). From the last definition,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(w_i, \hat{\delta}, \gamma) = \nu_n(\hat{\delta}, \gamma) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} E\left[m(W_i, \hat{\delta}, \gamma)\right]. \qquad (12)$$

Note that,

$$\begin{aligned} E\left[\begin{pmatrix} D \\ X \end{pmatrix}\cdot\kappa(\gamma)\cdot(\theta - 1\{Y - aD - X'b < 0\})\right] \\ = E\left[\begin{pmatrix} D \\ X \end{pmatrix}\cdot\kappa(\gamma)\cdot\left(\theta - E\left[1\{Y - aD - X'b < 0\}|Z, D, X\right]\right)\right] \\ = E\left[\begin{pmatrix} D \\ X \end{pmatrix}\cdot\kappa(\gamma)\cdot\left(\theta - F_{Y|Z,D,X}\left(aD + X'b\right)\right)\right]. \qquad (13) \end{aligned}$$

Then, assumptions 3(ii), 4(iii) and 4(iv) allow us to apply dominated convergence,

$$\partial E\left[m(W,l,\gamma)\right]/\,\partial l' = -E\left[\begin{pmatrix} D \\ X \end{pmatrix} \cdot \kappa(\gamma) \cdot f_{Y|Z,D,X}\left(aD + X'b\right) \cdot \begin{pmatrix} D \\ X \end{pmatrix}'\right].$$

Now we can apply a Mean Value Theorem as follows:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(w_i,\hat{\delta},\gamma) = \nu_n(\hat{\delta},\gamma) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} E[m(W_i,\delta,\gamma)] + \frac{\partial E[m(W_i,\tilde{\delta},\gamma)]}{\partial l'}\sqrt{n}(\hat{\delta}-\delta). \tag{14}$$

The next lemma shows that the second term of the right hand side of the last equation is zero.

Lemma 4 *Given assumptions 1 to 4,* $E[m(W,\delta,\gamma)] = 0$.

Proof: First, note that under assumptions 1 to 4 $f(\cdot,l,\gamma)$ is $\mathcal{L}^1$-bounded in $\mathcal{B}$ and $\partial f(W,\delta,\gamma)/\partial l = -m(W,\delta,\gamma) \in \mathcal{L}^1$. Now, let us show that the derivative of the limiting objective function (equation (7)) with respect to (a,b) is equal to minus $E[m(W,l,g)]$. Denote

$$\Delta_{\tilde{h}} f(c) \equiv [\kappa(g)\cdot(\theta - 1\{Y-(c+\tilde{h})<0\})\cdot(Y-(c+\tilde{h}))] - [\kappa(g)\cdot(\theta-1\{Y-c<0\})\cdot(Y-c)].$$

It can be easily shown that a Weierstrass domination condition, $|\Delta_{\tilde{h}} f(\alpha D + X'\beta)/\tilde{h}| \leq \bar{K}$ for $\tilde{h} \neq 0$, holds. Then, by assumption 4(iii), $E[(1+\|X\|)\cdot|\Delta_{\tilde{h}} f(\alpha D + X'\beta)/\tilde{h}|] < \infty$ this implies

$$\frac{\partial E\left[f(W,\delta,\gamma)\right]}{\partial l} = -E\left[m(W,\delta,\gamma)\right].$$

Then, Theorem 2 and $(\alpha,\beta) \in int(\Lambda \times \Theta)$ yield $E[m(W,\delta,\gamma)] = 0$. $\mathcal{QED}$.

By assumption 4(iv) $\partial E[m(W_i,l,\gamma)]/\partial l'$ is continuous at δ, this implies that $\partial E[m(W_i,\tilde{\delta},\gamma)]/\partial l' \xrightarrow{p} M_\delta$. Then,

$$\begin{aligned} -\left(M_\delta + o_p(1)\right)\sqrt{n}(\hat{\delta}-\delta) &= \nu_n(\hat{\delta},\gamma) + \left(M_\gamma + o_p(1)\right)\sqrt{n}(\hat{\gamma}-\gamma) + o_p(1) \\ &= \left\{\nu_n(\hat{\delta},\gamma) - \nu_n(\delta,\gamma)\right\} + \nu_n(\delta,\gamma) + M_\gamma\sqrt{n}(\hat{\gamma}-\gamma) + o_p(1). \end{aligned} \tag{15}$$

The first term of the right hand side of equation (15) can be shown to be $o_p(1)$ by using a stochastic equicontinuity result. Each element of the vector $m(W,l,\gamma)$ is an Euclidean class with envelope $F = \bar{K}(1+\|X\|)$. By assumption 4(iii), $E\|X\|^2 < \infty$ so F is square-integrable. Then, assumption 3(vi) implies that each component of $m(W,l,\gamma)$ is $\mathcal{L}^2$ continuous at δ. Under these conditions $\left\{\nu_n(\hat{\delta},\gamma) - \nu_n(\delta,\gamma)\right\}$ is $o_p(1)$ (see Lemma 2.17 in Pakes and Pollard (1989)).

On the other hand, it can be easily shown that, under assumption 4(i),

$$\sqrt{n}(\hat{\gamma}-\gamma) = -\left(E\left[\frac{\partial q(V_i,\gamma)}{\partial g'}\right]\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q(v_i,\gamma) + o_p(1).$$

Then,

$$\sqrt{n}\left(\hat{\delta}-\delta\right) = -M_\delta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{m(w_i,\delta,\gamma) - M_\gamma Q_\gamma^{-1} q(w_i,\gamma)\right\} + o_p(1).$$

Now, under assumptions 3 and 4, $E\|m(W_i,\delta,\gamma) - M_\gamma Q_\gamma^{-1} q(V_i,\gamma)\|^2 < \infty$, then

$$\sqrt{n}\left(\hat{\delta} - \delta\right) \xrightarrow{d}$$
$$\mathcal{N}\left(0, M_\delta^{-1} E\left[\left\{m(W_i,\delta,\gamma) - M_\gamma Q_\gamma^{-1} q(V_i,\gamma)\right\}\left\{m(W_i,\delta,\gamma) - M_\gamma Q_\gamma^{-1} q(V_i,\gamma)\right\}'\right] M_\delta^{-1\,\prime}\right). \quad (16)$$

$\mathcal{QED}$.

Appendix II: Computational Issues

It is well-known that conventional quantile regression has a Linear Programming (LP) representation (see, e.g., Koenker and Bassett (1978)). This LP problem is given by

$$\begin{array}{l} \text{Min}_\tau \ c\tau \\ \text{s.t. } A\tau = y \\ \quad \tau \geq 0 \end{array} \quad (17)$$

where $A = ((d_1,..,d_n)', (x_1,..,x_n)', -(d_1,..,d_n)', -(x_1,..,x_n)', I_n, -I_n)$, $c = (o', o', \theta\cdot\iota', (1-\theta)\cdot\iota')$, $y = (y_1,..,y_n)'$, I_n is the identity matrix of size n, o is a $h \times 1$ vector of zeros and ι is an $n \times 1$ vector of ones. The solution of this problem is interpreted as $\tau = (\hat{\alpha}_\theta^+, \hat{\beta}_\theta^{+\prime}, \hat{\alpha}^-, \hat{\beta}^{-\prime}, \hat{u}_\theta^{+\prime}, \hat{u}_\theta^{-\prime})'$, where $u_\theta = y - \alpha_\theta(d_1,..,d_n)' + (x_1,..,x_n)'\beta_\theta$, e^+ denotes the positive part of the real number e and e^- denotes its negative part. This problem can be solved efficiently using the modification of the Barrodale and Roberts (1973) algorithm developed by Koenker and D'Orey (1987). This algorithm is a specialization of the Simplex method that exploits the particular structure of the quantile regression problem to pass through several adjacent vertices in each Simplex iteration.

A similar representation of QTE sets $c = (o', o', \theta\cdot\mathcal{K}'\ (1-\theta)\cdot\mathcal{K}')$, where $\mathcal{K}$ is the $n \times 1$ vector of κ_i's. However, QTE is not an LP problem because when κ_i is negative we have to include the constraints $u_i^+ \cdot u_i^- = 0$ to make $u_i^+ = max\{u_i, 0\}$ and $u_i^- = max\{-u_i, 0\}$ hold. If we do not include those constraints, the problem is unbounded and the solution method breaks down since we can reduce the objective function as much as we want by increasing both the positive and the negative parts of a residual associated with a negative κ_i. Suppose, for example, that we have a basic solution with $u_i = u_i^+ = \overline{u}_i > 0$ and $u_i^- = 0$. Then, if we make $u_i^+ = \overline{u}_i + \Delta$ and $u_i^- = \Delta$, for every $\Delta > 0$ we still have $u_i = u_i^+ - u_i^- = \overline{u}_i$, so the new solution is feasible. However, the objective function is reduced by $|\kappa_i|\Delta$. As this is true for every $\Delta > 0$, the problem is unbounded.

One way to incorporate the non-linear constraints $u_i^+ \cdot u_i^- = 0$ is to express the minimization as a Mixed Integer Linear Programming (MILP) problem. To do that, we include two additional restrictions and one additional parameter, s_i, for each observation with a negative κ_i:

$$u_i^+ \leq M s_i$$

and

$$u_i^- \leq M(1 - s_i).$$

where $s_i \in \{0, 1\}$ and M is a (non-binding) high number. This formulation imposes $u_i^+ \cdot u_i^- = 0$ for observations with negative κ_i. In principle, a global optimum could be attained by using *branch and bound* algorithms for MILP problems or for LP problems with Special Ordered Sets (SOS).

A special ordered set of type one (SOS1) is a set of nonnegative variables such that at most one of them may be nonzero (see, e.g., Hummeltenberg (1984)). Clearly, the set formed by both the positive and the negative part of a number is an SOS1. However, algorithms for MILP or SOS are very slow for large problems like ours.

Another possible strategy to solve this problem is to combine the Simplex method with a *restricted-basis entry* rule (See, for example, Wagner (1975) pag. 565. A restricted-basis entry rule does not allow the negative part of a residual to enter the basis, that is to take a value greater than zero, if the positive part of that residual is already in the basis, and vice versa.). Because our problem is not convex, this strategy does not guarantee a global optimum. However, restricted-basis entry methods find an optimum among permitted adjacent extreme points; this is a *local star optimum* in the terminology of Charnes and Cooper (1957). Because a global optimum is always a local star optimum, we can search for a global optimum by starting this procedure from different initial values. In practice, we found that an efficient way to implement restricted-basis entry is by using a modification of the Barrodale-Roberts algorithm. By construction, the Barrodale-Roberts algorithm does not allow both the positive and the negative part of residuals and parameters to be in the basis at the same time. Moreover, in addition to being fast, the modified Barrodale-Roberts algorithm has access to more vertices at each iteration than the conventional Simplex method. This feature allows the algorithm to improve over the local star optima found by the Simplex method with restricted-basis entry.

The main modification we made to the Barrodale-Roberts algorithm is related to the way that algorithm passes through several vertices in each iteration. The Barrodale-Roberts algorithm changes the sign of the pivotal row while the marginal cost of the vector entering the basis is positive after that change. In this way, the algorithm reduces the objective function. When changing the sign of the pivotal row makes the marginal cost negative, the algorithm performs a Simplex transformation. For our problem, whether the objective function and the marginal cost of the vector entering the basis increase or decrease with a change in the sign of the pivotal row depends on the sign of the κ_i associated to that row. Taking that into account we choose the vector to enter the basis as that one which accomplish the larger reduction in the objective function. Our simulation experiments indicated that the modified Barrodale-Roberts algorithm is very likely to find the global optimum for the QTE problem, so we chose this procedure for estimation.

For the empirical application, the modified Barrodale-Roberts algorithm was implemented using conventional quantile regression estimates as initial values. Then, the same algorithm was restarted from initial values randomly chosen in wide regions centered at the "best-so-far" points (in particular, we constructed regions with side lengths equal to twice the absolute values of the current estimates.) This step was repeated for each quantile until no improvement was attained in the last twenty trials. Overall, only small changes in some of the coefficients were observed in this step. Finally, a Nelder-Mead algorithm was started from the solution at this point. This final step did not decrease the value of the objective function for any quantile.

Appendix III: Asymptotic Variance Estimation

For the empirical application we used a linear specification, $E[Z|X] = X'\gamma$, for the first step. Since γ is estimated by OLS, this yields:

$$q(V, g) = X \cdot (Z - X'g),$$

$$Q_\gamma = -E[XX'],$$

and

$$M_\gamma = E\left[\begin{pmatrix} D \\ X \end{pmatrix} \cdot \left(-\frac{D \cdot (1-Z)}{(1-X'\gamma)^2} + \frac{(1-D)\cdot Z}{(X'\gamma)^2}\right) \cdot (\theta - 1\{Y - \alpha D - X'\beta < 0\}) \cdot X'\right].$$

We know from Appendix I that

$$M_\delta = -E\left[\begin{pmatrix} D \\ X \end{pmatrix} \cdot \kappa(\gamma) \cdot f_{Y|Z,D,X}\left(\alpha D + X'\beta\right) \cdot \begin{pmatrix} D \\ X \end{pmatrix}'\right].$$

The matrices Q_γ, M_γ, and M_δ were estimated by evaluating the sample counterparts of the last three equations at $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$. Note that $f_{Y|Z,D,X}\left(\alpha D + X'\beta\right) = f_{U|Z,D,X}\left(0\right)$, where $U = Y - \alpha D - X'\beta$. To estimate the density function $f_{U|Z,D,X}\left(0\right)$ for each of the considered quantiles, the data were divided in cells defined by different values of the categorical covariates (same sex, high school graduate, etc.). Then a normal kernel was used to smooth over the age variables and the QTE residual within each of the cells. Let U_θ be the QTE residual for a quantile index equal to θ. Also let A and A_1 be the age of the mother and the age of the mother at first birth. For each of the cells, we estimated $f_{(U_\theta,A,A_1)}(0,a,a_1)$ and $f_{(A,A_1)}(a,a_1)$ for each realized value (a,a_1) of (A,A_1) in the cell. The conditional densities in M_δ were then estimated as

$$\hat{f}_{U_\theta|A=a,A_1=a_1}(0) = \frac{\hat{f}_{(U_\theta,A,A_1)}(0,a,a_1)}{\hat{f}_{(A,A_1)}(a,a_1)}.$$

When there is no instrumenting, the asymptotic variance of QTE reduces to the well-known formula for conventional quantile regression (see e.g. Buchinsky (1994)). The conditional density terms that appear in the asymptotic variance of conventional quantile regression were estimated in the same way as for QTE.

References

Abadie, A., 1997a, "Bootstrap Tests for the Effects of a Treatment on the Distributions of Potential Outcomes," MIT Department of Economics, mimeo.

Abadie, A., 1997b, "Identification of Treatment Effects in Models with Covariates," MIT Department of Economics, mimeo.

Amemiya, T., 1982, "Two Stage Least Absolute Deviations Estimators," Econometrica 50, 689-711.

Andrews, D.W.K., 1994, "Empirical Process Methods in Econometrics," Chapter 37, Handbook of Econometrics, IV. Elsevier Science Publishers.

Angrist, J.D., 1990, "Lifetime Earnings and the Vietnam-Era Draft Lottery: Evidence from Social Security Administrative Records," American Economic Review 80, 313-336.

Angrist, J.D. and W. N. Evans, 1998, "Children and Their Parents' Labor Suppy: Evidence from Exogenous Variation in Family Size," American Economic Review (forthcoming).

Angrist, J.D., G.W. Imbens, and D.B. Rubin, 1996, "Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91, 444-472.

Atkinson, A.B., 1970, "On the Measurement of Inequality," Journal of Economic Theory, 2, 244-263.

Balke A., and J. Pearl, 1997, "Bounds on Treatment Effects From Studies With Imperfect Compliance," Journal of the American Statistical Association, 92, 1171-1176.

Barrodale, I., and F.D.K. Roberts, 1973, "An Improved Algorithm for Discrete l_1 linear approximation," SIAM Journal on Numerical Analysis 10, 839-848.

Bassett, G., and R. Koenker, 1982, "An Empirical Quantile Function for Linear Models with iid Errors," Journal of the American Statistical Association, 77, 407-415.

Bloom, H.S., L.L. Orr, S.H. Bell, G. Cave, F. Doolittle, W. Lin and J.M. Bos, 1997, "The Benefits and Costs of JTPA Title II-A Programs," The Journal of Human Resources 32, 549-576.

Browning, M., 1992, "Children and Household Economic Behavior," Journal of Economic Literature 30, 1434-1475.

Buchinsky, M., 1994, "Changes in the US Wage Structure 1963-87: Application of Quantile Regression," Econometrica 62, 405-458.

Card, D., 1996, "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," Econometrica 64, 957-980.

Chamberlain, G., 1991, "Quantile Regression, Censoring, and the Structure of Wages," Chapter 5 in C.A. Sims, ed., Advances in Econometrics Sixth World Congress, Volume I, Econometric Society Monograph No. 23, Cambridge, Cambridge University Press.

Charnes A., and W.W. Cooper, 1957, "Nonlinear Power of Adjacent Extreme Point Methods in Linear Programming," Econometrica 25, 132-153.

DiNardo, J., Nicole M. Fortin, and T. Lemieux, 1996, "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," Econometrica 64, 1001-1045.

Freeman, R., 1980, "Unionism and the Dispersion of Wages," Industrial and Labor Relations Review 34, 3-23.

Heckman, J. and R. Robb, 1985, "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer, eds., Longitudinal Analysis of Labor Market Data, New York: Cambridge University Press.

Heckman J., J. Smith and N. Clements, 1997, "Making the Most Out of Social Experiments: Accounting for Heterogeneity in Programme Impacts," Review of Economic Studies 64, 487-535.

Hummeltenberg, W., 1984, "Implementations of Special Ordered Sets in MP Software," European Journal of Operational Research 17, 1-15.

Imbens, G.W., and J.D. Angrist, 1994, "Identification and Estimation of Local Average Treatment Effects," Econometrica 62, 467-476.

Imbens, G.W., and D.B. Rubin, 1997, "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," Review of Economic Studies 64, 555-574.

Koenker, R., and G. Bassett, 1978, "Regression Quantiles," Econometrica 46, 33-50.

Koenker, R., and V. D'Orey, 1987, "Computing Regression Quantiles," Journal of the Royal Statistical Society, Applied Statistics, 36, 383-393.

Lalonde, R.J., 1995, "The Promise of Public-Sector Sponsored Training Programs," Journal of Economic Perspectives (Spring), 149-168.

Lalonde, R.J., G. Marschke and K. Troske, 1996, "Using Longitudinal Data on Establishments to Analize the Effects of Union Organizing Campaigns in the United States," Annales d'Économie et de Statistique 41/42, 155-185.

Lewis, H.G., 1986, Union Relative Wage Effects: A Survey, Chicago: University of Chicago Press.

Manski, C.F., 1988, Analog Estimation Methods in Econometrics, New York: Chapman and Hall.

Manski, C.F., 1990, "Nonparametric Bounds on Treatment Effects," American Economic Review, Papers and Proceedings, 80, 319-323.

Manski, C.F., 1994, "The Selection Problem," Chapter 3 in C.A. Sims, ed., Advances in Econometrics Sixth World Congress, Volume II, Econometric Society Monograph No. 23, Cambridge, Cambridge University Press.

Newey, W.K, and D. McFadden, 1994, "Large Sample Estimation and Hypothesis Testing," Chapter 36, Handbook of Econometrics, IV. Elsevier Science Publishers.

Pakes, A. and D. Pollard, 1989, "Simulation and the Asymptotics of Optimization Estimators," Econometrica, 57, 1027-1057.

Powell, J.L., 1983, "The Asymptotic Normality of Two-Stage Least Absolute Deviations Estimators," Econometrica 51, 1569-1575.

Powell, J.L., 1994, "Estimation of Semiparametric Models," Chapter 41, Handbook of Econometrics, IV. Elsevier Science Publishers.

Robins, J.M., 1989, "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in Health Service Research Methodology: A Focus on AIDS, eds. L. Sechrest, H. Freeman, and A. Mulley, Washington, DC: U.S. Public Health Service, pp. 113-159.

Rosenbaum, P.R., and D.B. Rubin, 1983, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika 70, 41-55.

Rubin, D.B., 1974, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology, 66, 688-701.

Rubin, D.B., 1977, "Assignment to Treatment Group on the Basis of a Covariate," Journal of Educational Statistics 2, 1-26.

Rubin, D.B, 1978, "Bayesian Inference for Causal Effects: The Role of Randomization," Annals of Statistics, 6, 34-58.

Ruppert, D., and R.J. Carroll, 1980, "Trimmed Least Squares Estimation in the Linear Model," Journal of the American Statistical Association 75, 828-838.

US Department of Labor, 1995, Office of the Chief Economist, What's Working (and What's Not), A Summary of Research on the Economic Impacts of Employment and Training Programs, Washington, DC: US Government Printing Office, January.

Wagner, H., 1975, Principles of Operations Research, New Jersey: Prentice-Hall.

TABLE I
CONVENTIONAL QUANTILE REGRESSION AND OLS ESTIMATES

	Mean	OLS	Quantile				
	(1)	(2)	0.10	0.25	0.50	0.75	0.90
Log family income	10.3						
	(1.39)						
More than two kids	.363	-.092	-.098	-.077	-.066	-.046	-.027
	(.481)	(.005)	(.008)	(.004)	(.003)	(.003)	(.003)
Constant		7.88	6.14	7.54	8.56	9.19	9.53
		(.024)	(.035)	(.020)	(.016)	(.015)	(.016)
Mother's age	30.5	.042	.049	.039	.034	.030	.027
	(3.44)	(.0008)	(.001)	(.0007)	(.0005)	(.0005)	(.0005)
Mother's age	21.9	.035	.056	.042	.030	.026	.029
at first birth	(3.48)	(.0008)	(.001)	(.0007)	(.0005)	(.0005)	(.0005)
High school	.625	.493	.746	.550	.367	.241	.189
graduate	(.484)	(.008)	(.013)	(.008)	(.006)	(.005)	(.005)
More than	.209	.798	1.12	.813	.588	.462	.443
high school	(.407)	(.009)	(.014)	(.009)	(.006)	(.006)	(.006)
Minority	.178	-.623	-.993	-.721	-.434	-.224	-.141
	(.383)	(.008)	(.013)	(.008)	(.006)	(.005)	(.005)
Boy 1st	.513	-.001	.004	.0008	-.0004	-.002	-.005
	(.500)	(.004)	(.007)	(.004)	(.003)	(.003)	(.003)
Same sex	.505						
	(.500)						

Note: The sample includes 346,929 observations on the family income of black or white women aged 21-35 and with two or more children in the 1990 Census PUMS. Other sample restrictions are as in Angrist and Evans (1998). *Minority* indicates black or hispanic; *Boy 1st* indicates firstborn male children. Column (1) shows sample means and column (2) shows OLS estimates from a regression of log family income on the listed covariates. The remaining columns report quantile regression estimates for the same specification. The numbers reported in parentheses are standard deviations of the variables for column (1) and standard errors of the estimates for the remaining columns. For OLS, robust standard errors are reported.

TABLE II
QUANTILE REGRESSION FOR COMPLIERS AND 2SLS ESTIMATES

	First Stage	2SLS	Quantile				
	(1)	(2)	0.10	0.25	0.50	0.75	0.90
More than two kids		-.122	-.180	-.089	-.065	-.070	-.031
		(.069)	(.097)	(.053)	(.038)	(.036)	(.041)
Constant	.439	7.89	8.00	8.33	8.89	9.39	9.44
	(.008)	(.040)	(.597)	(.337)	(.247)	(.228)	(.255)
Mother's age	.024	.043	.016	.027	.036	.030	.036
	(.0003)	(.002)	(.021)	(.012)	(.008)	(.007)	(.008)
Mother's age	-.037	.034	.034	.032	.020	.020	.019
at first birth	(.0003)	(.003)	(.022)	(.014)	(.009)	(.008)	(.010)
High school	-.071	.491	.671	.476	.255	.183	.186
graduate	(.002)	(.010)	(.207)	(.130)	(.084)	(.069)	(.065)
More than	-.039	.797	.941	.733	.430	.383	.382
high school	(.003)	(.010)	(.245)	(.146)	(.100)	(.086)	(.085)
Minority	.061	-.621	-1.39	-.390	-.217	-.158	-.170
	(.002)	(.009)	(.523)	(.169)	(.112)	(.096)	(.092)
Boy 1st	-.007	-.002	-.032	-.054	-.043	-.0004	.055
	(.002)	(.004)	(.102)	(.055)	(.039)	(.036)	(.042)
Same sex	.064						
	(.002)						

Note: Sample and variable definitions are the same as in Table I. Column (1) reports estimates from a "first-stage" regression of *More than two kids* on the listed covariates and the *Same sex* instrument. Column (2) shows 2SLS estimates from a regression of *Log family income* on the listed covariates with *More than two kids* treated as endogenous and *Same sex* used as excluded instrument. The remaining columns report QTE estimates for the same specification. Standard errors are reported in parentheses. For 2SLS, robust standard errors are reported.