A closed-form estimator for quantile treatment effects with endogeneity

Kaspar Wüthrich*

Revised: January 30, 2019

Abstract

This paper studies the estimation of quantile treatment effects based on the instrumental variable quantile regression (IVQR) model (Chernozhukov and Hansen, 2005). I develop a class of flexible plug-in estimators based on closed-form solutions derived from the IVQR moment conditions. The proposed estimators remain tractable and root-n-consistent, while allowing for rich patterns of effect heterogeneity. Functional central limit theorems and bootstrap validity results for the estimators of the quantile treatment effects and other functionals are provided. Monte Carlo simulations demonstrate favorable finite sample properties of the proposed approach. I apply my method to reanalyze the causal effect of 401(k) plans.

JEL Classification: C21, C26

Keywords: instrumental variables, conditional and unconditional quantile treatment effects, distribution regression, exchangeable bootstrap

*UC San Diego, Department of Economics, 9500 Gilman Dr., La Jolla, CA 92093, email: kwuthrich@ucsd.edu.

1 Introduction

Quantile regression is a powerful tool for characterizing the heterogeneous impact of policy variables on different points of an outcome distribution. However, as with ordinary least squares, endogeneity renders standard quantile regression methods inconsistent for estimating quantile treatment effects (QTE). One approach to overcome this problem is to use instrumental variable (IV) methods.

The goal and main contribution of this paper is to develop a flexible yet tractable approach for estimating conditional and unconditional QTE based on the instrumental variable quantile regression (IVQR) model (Chernozhukov and Hansen, 2005) with binary endogenous variables and binary instruments. The principal feature of the IVQR model is the rank similarity assumption, a restriction on the evolution of individual ranks across treatment states. Rank similarity implies a conditional moment restriction that can be used for constructing estimators of population QTE. However, estimation based on this moment restriction is complicated by the non-smoothness and non-convexity of the corresponding generalized method of moments (GMM) objective function, which occurs even for linear-in-parameters quantile models (Chernozhukov and Hansen, 2013; Chernozhukov et al., 2017).

In this paper, I derive closed-form solutions for the IVQR estimands of the potential outcome cumulative distribution functions (cdfs).[1] These closed-form solutions are compositions of observable conditional cdfs and probabilities. I estimate the conditional cdfs using distribution regression (DR) and the conditional probabilities using parametric binary choice models. I then apply the closed-form solutions to obtain plug-in estimators of the conditional potential outcome cdfs and the conditional QTE. Finally, unconditional QTE are estimated by integrating the estimators of the conditional potential outcome cdfs with respect to the empirical distribution of the covariates.[2]

[1] These identification results generalize Wüthrich (2018), who derives similar closed-from solutions under more restrictive assumptions.

[2] I refer to Firpo (2007) or Frölich and Melly (2013) for a discussion of the differences between conditional and unconditional QTE.

This plug-in estimation strategy naturally bypasses the challenges associated with optimizing the IVQR GMM objective function. As a result, the proposed estimators are easy to implement and remain computationally tractable, while allowing for rich patterns of treatment effect heterogeneity with respect to observables covariates. Moreover, unlike other IVQR estimators, the plug-in approach does not require practitioners to model the counterfactual structural quantile function. Instead, they only need to specify and estimate the observable conditional probabilities and cdfs with well-established and flexible parametric models for which specification tests are readily available.

Under standard regularity conditions, the plug-in estimators of the conditional and unconditional QTE are uniformly consistent and satisfy functional central limit theorems. Moreover, I prove the validity of the exchangeable bootstrap for estimating the limiting laws. These results allow me to construct uniform confidence bands and to test functional hypotheses such as no-effects, positive effects, constant effects, or stochastic dominance. Monte Carlo simulations demonstrate favorable finite sample properties of the proposed estimation and inference procedures. The simulations further suggest that the plug-in approach is robust against different types of misspecification of the underlying parametric restrictions. By contrast, I show that existing estimators based on linear-in-parameters IVQR models may be severely biased whenever the true quantile function is nonlinear.

Although the focus of this paper is on estimating QTE, the plug-in approach to estimation and inference also covers many other smooth functionals of the conditional and unconditional potential outcome cdfs. Examples include average treatment effects (ATE), distributional treatment effects, Lorenz curves, and Gini coefficients.

I illustrate my method by reanalyzing the distributional effect of 401(k) plans on individual savings behavior using data from the 1991 Survey of Income and Program Participation (SIPP)

as in Chernozhukov and Hansen (2004) and Belloni et al. (2017). I find that the effect of 401(k) participation on individual assets is moderate at the lower tail, but increases substantially along the distribution. Interestingly, my results suggest that there is considerably more treatment effect heterogeneity than implied by the estimates based on the linear IVQR model of Chernozhukov and Hansen (2004).

1.1 Related literature

This paper contributes to the literature on estimation and inference based on the IVQR model. Linear-in-parameters conditional quantile models have been analyzed by Chernozhukov and Hansen (2006), Chernozhukov et al. (2007a), Chernozhukov and Hansen (2008), Chernozhukov et al. (2009), Kaplan and Sun (2017), and Chen and Lee (2018). Chernozhukov and Hong (2003) have considered potentially nonlinear parametric quantile functions. Nonparametric approaches have been studied by Chernozhukov et al. (2007b), Horowitz and Lee (2007), Chen and Pouzo (2009), Chen and Pouzo (2012), and Gagliardini and Scaillet (2012). Chernozhukov and Hansen (2013) and Chernozhukov et al. (2017) have provided surveys of the IVQR model including references to empirical applications.

Abadie et al. (2002) have introduced an alternative approach for estimating QTE with endogenous binary treatments based on the local average treatment effects framework (cf. Imbens and Angrist, 1994). They exploit a re-weighting result in Abadie (2003) to identify and estimate conditional QTE for the compliers, referred to as local QTE (LQTE). Frandsen et al. (2012) have studied identification and estimation of LQTE in regression discontinuity designs, Frölich and Melly (2013) and Hsu et al. (2015) have proposed estimators for unconditional LQTE, and Belloni et al. (2017) have suggested estimation and inference methods in high-dimensional settings. Generalizations to non-binary instruments have been analyzed by Carneiro and Lee (2009) and Yu (2014). A recent survey of the LQTE framework is provided by Melly and Wüthrich (2017).

It is important to note that the assumptions of the IVQR model neither imply nor are implied by the assumptions of the LQTE model. The IVQR model imposes rank similarity on the outcome equation, whereas the LQTE model relies on monotonicity of the selection equation. Moreover, the estimands of both models are different. The IVQR model recovers QTE for the whole population, whereas the LQTE model only identifies QTE for the subpopulation which reacts to the instrument—the compliers. Section 3.2.2 discusses the relationship between the IVQR and the LQTE model in more detail; see also Wüthrich (2018).

While the identification and estimation results in this paper cannot be generalized beyond binary treatments, the underlying IVQR model naturally accommodates continuous endogenous variables. Alternative approaches to identification and estimation in nonseparable models with continuous regressors have been proposed by Chesher (2003), Ma and Koenker (2006), Lee (2007), Jun (2009), Imbens and Newey (2009), D'Haultfoeuille and Février (2015), and Torgovitsky (2015) among others.

The proposed estimators are also related to several papers that rely on estimating conditional distributions using distribution or quantile regression models as ingredients for deriving plug-in estimators. Chernozhukov et al. (2013) have analyzed counterfactual distributions, Yu (2014) has proposed a flexible parametric estimation approach for marginal QTE, and Melly and Santangelo (2015) and De Chaisemartin and D'Haultfoeuille (2018) have analyzed nonlinear difference-in-differences models.

1.2 Outline

The remainder of the paper is organized as follows. In Section 2, I present the IVQR model, discuss the key identification conditions, and derive the closed-form solutions. Section 3 describes the plug-in estimators and compares them to existing methods. In Section 4, I give the asymptotic results

and establish bootstrap validity. Section 5 presents Monte Carlo evidence on the finite sample performance of the proposed estimation and inference procedures. In Section 6, I apply my method to estimate the distributional effects of 401(k) plans on accumulated assets. Section 7 concludes. All proofs and some additional results are collected in the appendix.

2 Model and closed-form solutions

2.1 The instrumental variable quantile regression model

This paper focuses on a setup with an absolutely continuous outcome variable Y, a binary treatment D, a binary instrument Z, and a vector of exogenous covariates X. Appendix B discusses how to incorporate non-binary instruments into the analysis. The symbols $\mathcal{D}$, $\mathcal{Z}$, and $\mathcal{X}$ denote the supports of D, Z, and X, $\mathcal{T}$ is a set of quantile indices, and $\mathcal{Y} \subset \mathbb{R}$ is the region of interest for Y. Define $p(z,x) := P(D = 1 \mid Z = z, X = x)$ and let $F_{Y|D,Z,X}$ and $f_{Y|D,Z,X}$ denote the cdf and the density of $Y \mid D = d, Z = z, X = x$. Moreover, I define $\mathcal{YX} := \{(y,x) : y \in \mathcal{Y}, x \in \mathcal{X}\}$ and generate other index sets accordingly; for example $\mathcal{DZ} := \{(d,z) : d \in \mathcal{D}, z \in \mathcal{Z}\}$. The analysis is developed within the potential outcomes framework (e.g., Rubin, 1974). Let Y_1 and Y_0 (indexed by D) denote the two absolutely continuous potential outcomes and denote the corresponding regions of interest as $\mathcal{Y}_1$ and $\mathcal{Y}_0$. Potential outcomes are related to the observed outcome as $Y = DY_1 + (1-D)Y_0$. For simplicity, I assume that the support of $Y_d \mid X = x$ is equal to the marginal support of Y for $(d,x) \in \mathcal{DX}$ and I set $\mathcal{Y}_0 = \mathcal{Y}_1 = \mathcal{Y}$.[3] Neither of the results in this paper rely on this restriction and it could be relaxed at the expense of a more complicated notation.

After conditioning on covariates $X = x$, by the Skorohod representation of random variables,

[3] I use y to denote a generic realization of Y, Y_1, and Y_0. Whenever it is important to distinguish between realizations of Y_1 or Y_0, I use y_1 and y_0.

potential outcomes can be represented as

$$Y_d = Q_{Y_d|X}(U_d \mid x) \quad \text{with} \quad U_d \sim U(0,1),$$

where $Q_{Y_d|X}(\cdot \mid x)$ is the quantile function of $Y_d \mid X = x$. This representation is essential for the IVQR model.

The IVQR model is based on the following set of assumptions (some of which are representations) (Chernozhukov and Hansen, 2005, Assumptions A1–A5):

Assumption 1. *Given a common probability space (Ω, F, P), the following conditions hold jointly with probability one:*

1. *Potential outcomes: Conditional on $X = x$, for each d, $Y_d = Q_{Y_d|X}(U_d \mid x)$, where $Q_{Y_d|X}(\tau \mid x)$ is strictly increasing in τ and $U_d \sim U(0,1)$.*

2. *Independence: Conditional on $X = x$, $\{U_d\}$ are independent of Z.*

3. *Selection: $D := \rho(Z, X, V)$ for some unknown function ρ and random vector V.*

4. *Rank similarity: Conditional on $X = x, Z = z, V = v$, $\{U_d\}$ are identically distributed.*

5. *Observed variables: Observed variables consist of $Y := Q_{Y_D|X}(U_D \mid X)$, D, X, and Z.*

Assumption 1.1 restates the Skorohod representation of random variables and imposes strict monotonicity of the structural quantile function, thus ruling out discrete outcome variables. Assumption 1.2 imposes conditional independence between the potential outcomes and the instrument. Assumption 1.3 states a very general selection equation in which the random vector V captures unobserved factors affecting selection into treatment. Assumption 1.4 is arguably the most important condition of the IVQR model. It requires that individual ranks are constant across potential outcome distributions up to random "slippages" from a common level U. Finally, Assumption

1.5 summarizes the observable variables. The interested reader is referred to Chernozhukov and Hansen (2005, 2013) and Chernozhukov et al. (2017) for in-depth discussions of Assumption 1.

The main statistical implication of Assumption 1 is the following nonlinear moment condition (Chernozhukov and Hansen, 2005, Theorem 1):

$$P\big(Y \leq Q_{Y_D|X}(\tau \mid X) \mid X, Z\big) = \tau \tag{1}$$

Estimation based on (1) is challenging because the sample analogue of the GMM objective function is non-smooth and generally non-convex. The existing approaches to overcome this problem are referenced in the introduction and further discussed in Section 3.2.1.

2.2 Conditions for point identification

The IVQR moment restrictions (1) do not point identify $Q_{Y_D|X}(\tau \mid x)$ without additional assumptions. For vectors (y_0, y_1) and $X = x$ define

$$\Pi(y_0, y_1, x) := (P(Y \leq y_D \mid Z = 0, X = x) - \tau,\ P(Y \leq y_D \mid Z = 1, X = x) - \tau)',$$

where $y_D := Dy_1 + (1 - D)y_0$. Chernozhukov and Hansen (2005) show that the key condition for point identification is full rank of the Jacobian of $\Pi(y_0, y_1, x)$ with respect to (y_0, y_1):

$$\Pi'(y_0, y_1, x) := \begin{pmatrix} f_{Y|D,Z,X}(y_0 \mid 0, 0, x)(1 - p(0, x)) & f_{Y|D,Z,X}(y_1 \mid 1, 0, x)p(0, x) \\ f_{Y|D,Z,X}(y_0 \mid 0, 1, x)(1 - p(1, x)) & f_{Y|D,Z,X}(y_1 \mid 1, 1, x)p(1, x) \end{pmatrix}. \tag{2}$$

Lemma 1 below derives closed-form solutions for the potential outcome cdfs under this full rank condition.

Assumption 2 (Identification)**.** *For all $x \in \mathcal{X}$,*

1. $0 < P(Z = 1 \mid X = x) < 1$

2. *The support of* $(Y_0, Y_1) \mid X = x$ *is a closed rectangle in* $\mathbb{R}^2$.

3. $\Pi'(y_0, y_1, x)$ *is continuous in* (y_0, y_1) *in the support of* $(Y_0, Y_1) \mid X = x$.

4. $\Pi'(y_0, y_1, x)$ *is of full rank for all* (y_0, y_1) *in the support of* $(Y_0, Y_1) \mid X = x$.

Assumption 2.1 is a standard overlap assumption that requires Z to be nontrivially assigned conditional on $X = x$. Assumption 2.2 corresponds to Chernozhukov and Hansen (2005)'s assumption that the parameter space is a closed rectangle. Assumption 2.3 imposes continuity of the Jacobian and corresponds to Condition (i) in Theorem 2 of Chernozhukov and Hansen (2005). As noted by Chernozhukov and Hansen (2005, p.252), conditional on $X = x$, Assumption 2.4 "requires the impact of instrument Z on the joint distribution of (Y, D) to be sufficiently rich". I refer to Chernozhukov and Hansen (2005, 2013) and Chernozhukov et al. (2017) for in-depth discussions of Assumption 2.4 and to Appendix A for simple sufficient conditions. Finally, note that Assumption 2 is in principle directly testable because it restricts the distribution of observables.

2.3 Closed-form solutions

The first contribution of this paper is to derive closed-form solutions for the conditional potential outcome cdfs $F_{Y_1|X}$ and $F_{Y_0|X}$ under Assumption 2.

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

$$\begin{aligned} F_{Y_1|X}(y_1 \mid x) &= p(1,x) F_{Y|D,Z,X}(y_1 \mid 1,1,x) \\ &\quad + (1 - p(1,x)) F_{Y|D,Z,X}\left(Q^c_{Y_0|X}\left(F^c_{Y_1|X}(y_1 \mid x) \mid x\right) \mid 0,1,x\right), \qquad (3) \\ F_{Y_0|X}(y_0 \mid x) &= (1 - p(0,x)) F_{Y|D,Z,X}(y_0 \mid 0,0,x) \\ &\quad + p(0,x) F_{Y|D,Z,X}\left(Q^c_{Y_1|X}\left(F^c_{Y_0|X}(y_0 \mid x) \mid x\right) \mid 1,0,x\right), \qquad (4) \end{aligned}$$

where[4]

$$F^c_{Y_1|X}(y_1 \mid x) = p(1,x)F_{Y|D,Z,X}(y_1 \mid 1,1,x) - p(0,x)F_{Y|D,Z,X}(y_1 \mid 1,0,x), \tag{5}$$

$$F^c_{Y_0|X}(y_0 \mid x) = (1-p(0,x))F_{Y|D,Z,X}(y_0 \mid 0,0,x) - (1-p(1,x))F_{Y|D,Z,X}(y_0 \mid 0,1,x). \tag{6}$$

The proof of Lemma 1 proceeds in two steps. First, I show that, under Assumption 1, Assumption 2 implies strict monotonicity of $F^c_{Y_1|X}(\cdot \mid x)$ and $F^c_{Y_0|X}(\cdot \mid x)$. In the second step, I exploit strict monotonicity to obtain closed-form solutions based on the IVQR moment conditions (1).

Remark 1. *Lemma 1 demonstrates that it is possible to derive closed-form solutions under the same identification conditions as in Chernozhukov and Hansen (2005). It thus extends Wüthrich (2018), which derives similar expressions under more restrictive conditions.*[5]

Based on the closed-form solutions in Lemma 1, the conditional QTE are identified as

$$\delta_X(\tau \mid x) = F^{-1}_{Y_1|X}(\tau \mid x) - F^{-1}_{Y_0|X}(\tau \mid x),$$

where $F^{-1}_{Y_d|X}$ denotes the (left-)inverse of $F_{Y_d|X}$. Conditional QTE are useful for analyzing effect heterogeneity by observable characteristics and across different quantiles. However, because the plug-in approach allows for rich patterns of treatment effect heterogeneity, the conditional QTE are generally high-dimensional objects that are hard to summarize and convey. Therefore, one is often also interested in unconditional QTE, which are informative about the effect of the treatment on the marginal outcome distribution:

$$\delta(\tau) = F^{-1}_{Y_1}(\tau) - F^{-1}_{Y_0}(\tau).$$

[4] The superscript c stands for *compliers* because $F^c_{Y_0|X}$ and $F^c_{Y_1|X}$ correspond to complier cdfs (multiplied by the proportion of compliers) under the assumptions of the LQTE framework; see Appendix A for a further discussion.

[5] An additional more subtle difference between Lemma 1 and the results in Wüthrich (2018) is that the closed-form solutions (3) and (4) are functions of $F^c_{Y_d|X}(y \mid x)$ instead of $F^c_{Y_d|X}(y \mid x)/(p(1,x)-p(0,x))$ as in Wüthrich (2018). This allows for weaker regularity conditions in the theoretical analysis and facilitates the practical implementation of the plug-in estimation approach.

The unconditional potential outcome cdfs are obtained by integrating the conditional potential outcome cdfs with respect to the marginal distribution of the covariates, F_X:

$$\begin{aligned} F_{Y_1}(y_1) &= \int_{\mathcal{X}} F_{Y_1|X}(y_1 \mid x) dF_X(x), \\ F_{Y_0}(y_0) &= \int_{\mathcal{X}} F_{Y_0|X}(y_0 \mid x) dF_X(x). \end{aligned}$$

Note that all the previous estimands are functions of $F_{Y|D,Z,X}$, $p(z,x)$, and F_X only, which naturally suggests a plug-in estimation approach.

3 Estimation

3.1 Plug-in estimators

This section proposes a plug-in estimation approach based on the closed-form solutions derived in Lemma 1. To make estimation practical and realistic and to overcome the curse of dimensionality, I use flexible parametric models for the conditional cdfs and probabilities.

Assumption 3 (Parametric restrictions)**.** *For all* $(y,d,z,x) \in \mathcal{YDZX}$, $p(z,x) = \Lambda(x'\gamma_z)$ *and* $F_{Y|D,Z,X}(y \mid d,z,x) = \Lambda(x'\beta_{dz}(y))$, *where* $\Lambda(\cdot)$ *is either the Probit or Logit link function.*[6]

The conditional probabilities are estimated using binary choice models:

$$\hat{p}(z,x) = \Lambda\big(x'\hat{\gamma}_z\big), \tag{7}$$

where

$$\hat{\gamma}_z = \arg\max_g \sum_{i=1}^{n} 1\{Z_i = z\}\big[1\{D_i = 1\}\ln\big[\Lambda\big(X_i'g\big)\big] + 1\{D_i = 0\}\ln\big[1 - \Lambda\big(X_i'g\big)\big]\big].$$

[6] Assumption 3 imposes that the link functions for $p(z,x)$ and $F_{Y|D,Z,X}$ are the same. The results in this paper do not depend on this restriction and it could be relaxed at the expense of a more complicated notation.

The conditional cdfs are estimated using DR (e.g., Williams and Grizzle, 1972; Foresi and Peracchi, 1995; Chernozhukov et al., 2013):

$$\hat{F}_{Y|D,Z,X}(y \mid d, z, x) = \Lambda\Big(x'\hat{\beta}_{dz}(y)\Big), \tag{8}$$

with

$$\hat{\beta}_{dz}(y) = \arg\max_{b} \sum_{i=1}^{n} 1\{D_i = d, Z_i = z\}\big[1\{Y_i \leq y\} \ln\big[\Lambda\big(X_i'b\big)\big] + 1\{Y_i > y\} \ln\big[1 - \Lambda\big(X_i'b\big)\big]\big].$$

DR is a comprehensive and flexible approach for modeling the entire conditional distribution, which allows for a different impact of covariates at different points of the outcome distribution.

In finite samples, the estimated conditional distributions may not be monotonic. To overcome this problem, I apply the rearrangement procedure proposed by Chernozhukov et al. (2010). Under correct specification (Assumption 3), these rearrangements do not affect the asymptotic properties of the estimators, and I keep them implicit throughout the paper.

Remark 2. *When the specifications* (7) *and* (8) *are fully saturated, the estimated conditional cdfs and probabilities are numerically equivalent to the empirical cdfs and probabilities in each cell. Thus, the parametric restrictions are without loss of generality in this important special case.*[7]

Remark 3. *Instead of DR, the conditional cdfs could be estimated by quantile regression (Koenker and Bassett, 1978). In the linear quantile regression model it is assumed that, for all* $(\tau, d, z, x) \in \mathcal{T}\mathcal{D}\mathcal{Z}\mathcal{X}$*,*

$$Q_{Y|D,Z,X}(\tau \mid d, z, x) = x'\beta_{dz}(\tau).$$

[7] The specifications (8) and (7) are fully saturated when X contains indicators for all points in the support of some original vector of discrete covariates $\tilde{X}$.

The conditional quantiles are estimated as

$$\hat{Q}_{Y|D,Z,X}(\tau \mid d, z, x) = x'\hat{\beta}_{dz}(\tau),$$

where

$$\hat{\beta}_{dz}(\tau) = \arg\min_{b} \sum_{i=1}^{n} 1\{D_i = d, Z_i = z\}\big[\tau - 1\big\{Y_i \leq X_i'b\big\}\big]\big[Y_i - X_i'b\big].$$

The conditional distribution is then estimated as

$$\hat{F}_{Y|D,Z,X}(y \mid d, z, x) = \varepsilon + \int_{\varepsilon}^{1-\varepsilon} 1\Big\{x'\hat{\beta}_{dz}(\tau) \leq y\Big\}d\tau,$$

where $\varepsilon > 0$ is a trimming constant to avoid estimation of tail quantiles. The interested reader is referred to Chernozhukov et al. (2013) and Leorato and Peracchi (2015) for comparisons between distribution and quantile regression. All the theoretical results in this paper can be extended to accommodate quantile regression estimators of the conditional cdfs; see Remark 4.

Based on $\hat{F}_{Y|D,Z,X}$ and $\hat{p}(z,x)$, plug-in estimators for $F_{Y_1|X}$ and $F_{Y_0|X}$ are constructed as

$$\begin{aligned}
\hat{F}_{Y_1|X}(y_1 \mid x) &= \hat{p}(1,x)\hat{F}_{Y|D,Z,X}(y_1 \mid 1,1,x) \\
&\quad +(1-\hat{p}(1,x))\hat{F}_{Y|D,Z,X}\Big(\hat{Q}^c_{Y_0|X}\Big(\hat{F}^c_{Y_1|X}(y_1 \mid x) \mid x\Big) \mid 0,1,x\Big), \qquad (9) \\
\hat{F}_{Y_0|X}(y_0 \mid x) &= (1-\hat{p}(0,x))\hat{F}_{Y|D,Z,X}(y_0 \mid 0,0,x) \\
&\quad +\hat{p}(0,x)\hat{F}_{Y|D,Z,X}\Big(\hat{Q}^c_{Y_1|X}\Big(\hat{F}^c_{Y_0|X}(y_0 \mid x) \mid x\Big) \mid 1,0,x\Big), \qquad (10)
\end{aligned}$$

where

$$\hat{F}^c_{Y_1|X}(y_1 \mid x) = \hat{p}(1,x)\hat{F}_{Y|D,Z,X}(y_1 \mid 1,1,x) - \hat{p}(0,x)\hat{F}_{Y|D,Z,X}(y_1 \mid 1,0,x), \quad (11)$$

$$\hat{F}^c_{Y_0|X}(y_0 \mid x) = (1-\hat{p}(0,x))\hat{F}_{Y|D,Z,X}(y_0 \mid 0,0,x) - (1-\hat{p}(1,x))\hat{F}_{Y|D,Z,X}(y_0 \mid 0,1,x). \quad (12)$$

The plug-in estimators for the conditional and unconditional QTE are

$$\hat{\delta}_X(\tau \mid x) = \hat{F}^{-1}_{Y_1|X}(\tau \mid x) - \hat{F}^{-1}_{Y_0|X}(\tau \mid x) \text{ and } \hat{\delta}(\tau) = \hat{F}^{-1}_{Y_1}(\tau) - \hat{F}^{-1}_{Y_0}(\tau),$$

where the unconditional cdfs, $\hat{F}_{Y_1}$ and $\hat{F}_{Y_0}$, are estimated by integrating the estimators of the conditional cdfs, $\hat{F}_{Y_1|X}$ and $\hat{F}_{Y_0|X}$, with respect to the empirical distribution of the covariates $\hat{F}_X$, where $\hat{F}_X(x) = 1/n \sum_{i=1}^n 1\{X_i \leq x\}$:

$$\hat{F}_{Y_1}(y_1) = \int_{\mathcal{X}} \hat{F}_{Y_1|X}(y_1 \mid x) d\hat{F}_X(x),$$
$$\hat{F}_{Y_0}(y_0) = \int_{\mathcal{X}} \hat{F}_{Y_0|X}(y_0 \mid x) d\hat{F}_X(x).$$

Plug-in estimators for other functionals can be constructed similarly.

3.2 Relationship to the existing literature

3.2.1 Relationship to alternative estimators based on the IVQR model

Fully nonparametric estimation of conditional QTE based on the moment restriction (1) naturally suffers from the curse of dimensionality and is complicated by ill-posed inverse problems. One way to overcome these challenges is to analyze linear-in-parameters models such as

$$Q_{Y_d|X}(\tau \mid x) = d\delta_X(\tau) + x'\beta(\tau). \quad (13)$$

However, even for linear models as simple as (13), estimation is complicated by the non-smoothness and non-convexity of the corresponding GMM objective function. Different approaches have been proposed to circumvent the numerical problems. Chernozhukov and Hong (2003) employ a quasi-Bayesian approach that relies on MCMC sampling and averaging instead of minimizing the objective function, the inverse quantile regression (IQR) estimator of Chernozhukov and Hansen (2006) combines robust grid search methods with convex quantile regressions, Kaplan and Sun (2017) suggest to smooth the moment conditions, and Chen and Lee (2018) reformulate the estimation problem as a mixed integer quadratic programming problem for which well-established solution algorithms exist.

The plug-in estimators and the estimators based on the linear-in-parameters model both rely on parametric assumptions to overcome the curse of dimensionality. However, these two approaches are generally non-nested unless the specifications are fully saturated in which case their probability limits coincide. One important advantage of the plug-in approach is that it naturally bypasses the numerical problems associated with optimizing the non-smooth and non-convex GMM criterion function. As a result, the proposed method is easy to implement, computationally tractable, and tuning-free, while allowing for rich patterns of treatment effect heterogeneity with respect to the observable covariates. Moreover, in contrast to linear-in-parameter models, practitioners do not bear the burden of directly specifying the structural quantile function. Instead, they need to model observable conditional cdfs and probabilities with well-established and flexible parametric models for which specification tests are readily available in the literature.[8] On the other hand, relying on closed-form solutions, the proposed method is inherently limited to binary treatments, whereas linear-in-parameters models can accommodate nonbinary endogenous variables. Furthermore, parsimonious linear-in-parameters models such as (13) have the advantage of being easier to interpret

[8]Note that the parametric models for the conditional cdfs and probabilities imply a particular nonlinear model for the structural quantile function.

in the context of conditional QTE because the estimated coefficients correspond to the quantile effects of interest. Finally, based on linear-in-parameters models, it is rather straightforward to conduct inference which is robust to the presence of weak instruments and identification failures (e.g., Chernozhukov and Hansen, 2008). By contrast, robust inference procedures have not yet been developed for the plug-in approach developed here.

Because the plug-in estimators and the estimation approaches based on linear-in-parameters models rely on parametric assumptions, they may suffer from misspecification bias. In contrast, the nonparametric minimum–distance-type estimators listed in the introduction do not impose any parametric restrictions. However, these methods naturally suffer from the curse of dimensionality and require the choice of tuning parameters. Moreover, regularization is often required to overcome the ill-posedness of the underlying inverse problems.

3.2.2 Relationship to the local quantile treatment effects framework

The LQTE model (Abadie et al., 2002) is a popular alternative framework for identifying and estimating conditional and unconditional QTE with binary treatments. This model identifies QTE for the subpopulation of compliers through a monotonicity assumption on the selection equation.

There are two major conceptual differences between the LQTE model and the IVQR model. First, the underlying assumptions differ and are generally non-nested and non-contradictory. The IVQR model imposes rank similarity on the outcome equation, whereas the LQTE model relies on monotonicity of the selection equation.[9]

Second, the resulting estimands are different. The IVQR model recovers QTE for the whole

[9] As an example of a scenario where the IVQR assumptions appear plausible while the LQTE assumption are questionable, suppose that one is interested in the effect of an information meeting about retirement savings options on individual savings (in the spirit of Duflo and Saez, 2003). To encourage the treatment group to attend, subjects are given a small financial incentive upon attendance. The LQTE monotonicity assumption is questionable here because financial incentives sometimes crowd-out intrinsic motivation as pointed out by De Chaisemartin (2017). By contrast, it appears reasonable that individual ranks in the savings distributions are the same with and without the information meeting (at least in expectation).

population, whereas the LQTE model only identifies QTE for the compliant subpopulation. I refer to Wüthrich (2018) for a detailed theoretical analysis of the relationship between both models.

Common to both frameworks is that, in the absence of parametric assumptions, estimation of conditional QTE suffers from the curse of dimensionality. In this paper, I overcome this problem by parametrizing the conditional cdfs and probabilities. A similar strategy has been used by Yu (2014) in the context of the LQTE framework. Alternatively, one can directly parametrize the structural quantile function as in Chernozhukov and Hansen (2006) and Abadie et al. (2002).

In the context of unconditional QTE, Frölich and Melly (2013) have shown that it is possible to construct fully nonparametric root-n-consistent estimators. They establish pointwise asymptotic normality of their estimator and prove that its variance attains the semiparametric efficiency bound; see also Hsu et al. (2015) and Belloni et al. (2017) for uniform results. To the best of my knowledge, no such results have been derived for estimators based on the IVQR model.

4 Asymptotic theory and inference

4.1 Limiting distribution

Assumption 4 provides regularity conditions under which the plug-in estimators are asymptotically Gaussian.

Assumption 4 (Regularity conditions)**.**

1. $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$ *are i.i.d.*

2. $\mathcal{Y}$ *is a compact interval in* $\mathbb{R}$ *and* $\mathcal{X}$ *is a compact subset of* $\mathbb{R}^{\dim(X)}$.[10] $\mathcal{T} \subset (0,1)$ *is a compact interval such that an* ε*-expansion of* $\{Q_{Y_d|X}(\tau \mid x) : \tau \in \mathcal{T}\}$ *is contained in* $\mathcal{Y}$ *for all* $(d,x) \in \mathcal{DX}$*. For* $(d,z) \in \mathcal{DZ}$*, the conditional density function* $f_{Y|D,Z,X}(y_d \mid d,z,x)$

[10] Note that $\mathcal{YX}$ must be chosen such that the $F_{Y_1|X}(y_1 \mid x)$ and $F_{Y_0|X}(y_0 \mid x)$ in Lemma 1 are well-defined on $\mathcal{YX}$ whenever the components $\{F_{Y|D,Z,X}(y_d \mid d,z,x)\}_{(d,z,x)\in\mathcal{DZX}}$, $\{F^c_{Y_d|X}(y \mid x)\}_{(d,x)\in\mathcal{DX}}$ are well-defined on $\mathcal{YX}$.

exists, is uniformly bounded and uniformly continuous in (y_d, x) *in the support of* (Y_d, X) *and* $f_{Y|D,Z,X}(y_d|d,z,x) > 0$ *on* $\mathcal{Y}\mathcal{X}$.

3. $\mathbb{E}||X||^2 < \infty$ *and, for* $(d,z) \in \mathcal{D}\mathcal{Z}$, *the minimum eigenvalues of*

$$J_{\gamma_z} = \mathbb{E}\left[1\{Z = z\}\frac{\lambda(X'\gamma_z)^2}{\Lambda(X'\gamma_z)[1-\Lambda(X'\gamma_z)]}XX'\right],$$

and

$$J_{\beta_{dz}}(y_d) = \mathbb{E}\left[1\{D = d, Z = z\}\frac{\lambda(X'\beta_{dz}(y_d))^2}{\Lambda(X'\beta_{dz}(y_d))[1-\Lambda(X'\beta_{dz}(y_d))]}XX'\right]$$

are bounded away from zero uniformly over $y_d \in \mathcal{Y}$, *where* λ *is the derivative of* Λ.

4. *For* $d \in \mathcal{D}$, $F^c_{Y_d|X}(y_d \mid x)$ *admits a non-zero derivative* $f^c_{Y_d|X}(y_d \mid x)$ *which is uniformly bounded and uniformly continuous in* (y_d, x) *in the support of* (Y_d, X).

Assumptions 4.1 – 4.3 are standard regularity conditions (e.g., Condition DR in Chernozhukov et al., 2013) that ensure that functional central limit theorems and bootstrap validity results apply for the conditional distributions and conditional probabilities. Assumption 4.4 strengthens Assumption 2, which implies that $f^c_{Y_d|X}(\cdot \mid x)$ is non-zero, to guarantee Hadamard differentiability of the closed-form solutions.

To describe the results, let $\ell^\infty(\mathcal{U})$ denote the set of bounded and measurable functions $h : \mathcal{U} \mapsto \mathbb{R}$ for a generic index set $\mathcal{U}$ and let $\ell^\infty(\mathcal{U})^2 := \ell^\infty(\mathcal{U}) \times \ell^\infty(\mathcal{U})$. The following theorem provides the joint limiting distribution of the estimated conditional potential outcome cdfs.

Theorem 1. *Suppose that Assumptions 1 – 4 hold. Then*

$$\sqrt{n}\begin{pmatrix} \hat{F}_{Y_1|X}(y_1 \mid x) - F_{Y_1|X}(y_1 \mid x) \\ \hat{F}_{Y_0|X}(y_0 \mid x) - F_{Y_0|X}(y_0 \mid x) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{Z}_{F_{Y_1|X}}(y_1 \mid x) \\ \mathbb{Z}_{F_{Y_0|X}}(y_0 \mid x) \end{pmatrix}$$

as stochastic processes in $\ell^\infty(\mathcal{Y}\mathcal{X})^2$*, where* $\mathbb{Z}_{F_{Y_1|X}}$ *and* $\mathbb{Z}_{F_{Y_0|X}}$ *are tight zero-mean Gaussian processes defined in Appendix C.*

Theorem 1 shows that, under the maintained assumptions, the estimators of the potential outcome distributions converge weakly to a two-dimensional Gaussian process. This result is the basis for deriving the asymptotic distribution of the estimators of the conditional QTE (Theorem 2), unconditional QTE (Theorem 3), as well as general smooth functionals (Theorem 4). The main ingredient for proving Theorem 1 is Hadamard differentiablity of the closed-form solutions in Lemma 1. The result then follows from the functional delta method and existing functional central limit theorems for the conditional cdfs and probabilities.

Remark 4. *The Hadamard differentiability of the closed-form solutions, which is established in Step 2 of the proof of Theorem 1, is a generic result and may be of independent interest. Combined with the functional delta method, it can be used to derive the asymptotic properties of* $\hat{F}_{Y_1|X}$ *and* $\hat{F}_{Y_0|X}$ *based on many different first-step estimators, provided that these first-step estimators satisfy a functional central limit theorem. This high-level condition can be verified for many estimators, including linear quantile regression (cf. Chernozhukov et al., 2013).*

The next theorem presents the limiting distribution of the conditional QTE estimators.

Theorem 2. *Suppose that Assumptions 1 – 4 hold. Then*

$$\sqrt{n}\Big(\hat{\delta}_X(\tau \mid x) - \delta_X(\tau \mid x)\Big) \rightsquigarrow \mathbb{Z}_{\delta_X}(\tau \mid x),$$

as a stochastic process in $\ell^\infty(\mathcal{T}\mathcal{X})$*, where* $\mathbb{Z}_{\delta_X}$ *is a mean-zero tight Gaussian process defined as*

$$\begin{aligned}\mathbb{Z}_{\delta_X}(\tau \mid x) \quad := \quad & \mathbb{Z}_{F_{Y_0|X}}\big(Q_{Y_0|X}(\tau \mid x) \mid x\big)/f_{Y_0|X}\big(Q_{Y_0|X}(\tau \mid x) \mid x\big) \\ & -\mathbb{Z}_{F_{Y_1|X}}\big(Q_{Y_1|X}(\tau \mid x) \mid x\big)/f_{Y_1|X}\big(Q_{Y_1|X}(\tau \mid x) \mid x\big).\end{aligned}$$

Functional central limit theorems for the unconditional QTE are derived based on the Hadamard differentiability of the counterfactual operator, $\phi(G, F) = \int G(y, x)dF(x)$, established in Chernozhukov et al. (2013).

Theorem 3. *Suppose that Assumptions 1 – 4 hold. Then*

$$\sqrt{n}\Big(\hat{\delta}(\tau) - \delta(\tau)\Big) \rightsquigarrow \mathbb{Z}_\delta(\tau)$$

as a stochastic process in $\ell^\infty(\mathcal{T})$*, where* $\mathbb{Z}_\delta(\tau) := \mathbb{Z}_{Q_{Y_1}}(\tau) - \mathbb{Z}_{Q_{Y_0}}(\tau)$ *is a mean-zero tight Gaussian process and* $\mathbb{Z}_{Q_{Y_1}}$ *and* $\mathbb{Z}_{Q_{Y_0}}$ *are defined in Appendix C.*

Finally, I provide a general result that characterizes the limiting distribution of a generic Hadamard differentiable functional of $F_{Y_1|X}$ and $F_{Y_0|X}$. Examples of Hadamard differentiable functionals include the ATE, distributional treatment effects, Lorenz curves, and Gini coefficients.

Theorem 4. *Suppose that Assumptions 1 – 4 hold and that the map* $\varphi\big(F_{Y_1|X}, F_{Y_0|X}\big)(w)$ *(indexed by* w*) is Hadamard differentiable with derivatives* $\varphi'_{F_{Y_1|X}}$ *and* $\varphi'_{F_{Y_0|X}}$*. Then*

$$\sqrt{n}\Big(\varphi\Big(\hat{F}_{Y_1|X}, \hat{F}_{Y_0|X}\Big)(w) - \varphi\big(F_{Y_1|X}, F_{Y_0|X}\big)(w)\Big) \rightsquigarrow \Big(\varphi'_{F_{Y_1|X}}\mathbb{Z}_{F_{Y_1|X}}\Big)(w) + \Big(\varphi'_{F_{Y_0|X}}\mathbb{Z}_{F_{Y_0|X}}\Big)(w)$$

as a stochastic process indexed by w*.*

The above characterizations of the limit processes can be used to perform inference using standard analytical methods. However, because the asymptotic variances contain terms that are difficult

to estimate such as conditional densities, I recommend using the resampling methods discussed in the next section.

4.2 Inference

Here I establish the validity of a general resampling procedure called the exchangeable bootstrap (e.g., Van der Vaart and Wellner, 1996; Chernozhukov et al., 2013). To describe the bootstrap procedure, let $(w_1, \ldots, w_n)$ be a vector of nonnegative random weights that are independent of the data and satisfy the following assumption.[11]

Assumption 5. *For each n, let $(w_1, \ldots, w_n)$ be an exchangeable[12], nonnegative random vector, which is independent of the data, such that for some $\varepsilon > 0$,*

$$\sup_n \mathbb{E}[w_1^{2+\varepsilon}] < \infty, \quad \frac{1}{n}\sum_{i=1}^n (w_i - \bar{w})^2 \xrightarrow{p} 1, \quad \bar{w} := \frac{1}{n}\sum_{i=1}^n w_i \xrightarrow{p} 1.$$

The exchangeable bootstrap uses $(w_1, \ldots, w_n)$ as random sampling weights to construct bootstrap versions of the estimators. Specifically, the bootstrap versions of the conditional distribution and the conditional probabilities are given by $\hat{F}^*_{Y|D,Z,X}(y \mid d, z, x) = \Lambda\Big(x'\hat{\beta}^*_{dz}(y)\Big)$ and $\hat{p}^*(z, x) = \Lambda(x'\hat{\gamma}_z{}^*)$, where

$$\hat{\beta}^*_{dz}(y) = \arg\max_b \sum_{i=1}^n w_i 1\{D_i = d, Z_i = z\}\big[1\{Y_i \le y\}\ln\big[\Lambda\big(X_i'b\big)\big] + 1\{Y_i > y\}\ln\big[1 - \Lambda\big(X_i'b\big)\big]\big],$$

and

$$\hat{\gamma}^*_z = \arg\max_g \sum_{i=1}^n w_i 1\{Z_i = z\}\big[1\{D_i = 1\}\ln\big[\Lambda\big(X_i'g\big)\big] + 1\{D_i = 0\}\ln\big[1 - \Lambda\big(X_i'g\big)\big]\big].$$

[11] This assumption corresponds to condition EB in Chernozhukov et al. (2013).

[12] "A sequence of random variables $X_1, X_2, \ldots, X_n$ is exchangeable if for any finite permutation σ of indices $1, 2, \ldots, n$ the joint distribution of the permuted sequence $X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(n)}$ is the same as the joint distribution of the original sequence." (Chernozhukov et al., 2013, p.2230)

Finally, $\hat{F}^*_X(x) = (\sum_{i=1}^n w_i)^{-1} \sum_{i=1}^n w_i 1\{X_i \leq x\}$ is a bootstrap version of the estimator of the marginal covariate distribution.

Remark 5. *As explained in Van der Vaart and Wellner (1996) and Chernozhukov et al. (2013), by appropriately choosing the weights, the exchangeable bootstrap covers many resampling schemes as special cases. For example, the empirical bootstrap corresponds to the case where* $(w_1, \ldots, w_n)$ *is a multinomial vector with parameter* n *and probabilities* $(1/n, \ldots, 1/n)$*. The weighted bootstrap corresponds to the case where* $(w_1, \ldots, w_n)$ *are i.i.d. nonnegative random variables with* $\mathbb{E}[w_1] = Var(w_1) = 1$*. The* m *out of* n *bootstrap is nested by letting* $(w_1, \ldots, w_n)$ *be equal to* $\sqrt{n/m}$ *times multinomial vectors with parameter* m *and probabilities* $(1/n, \ldots, 1/n)$*. Finally, subsampling corresponds to letting* $(w_1, \ldots, w_n)$ *be a row in which the number* $n(n-m)^{-1/2}m^{-1/2}$ *appears* m *times and* 0 *appears* $n - m$ *times ordered at random, independent of the data.*

The next theorem formally establishes validity of the exchangeable bootstrap.

Theorem 5. *Suppose that Assumptions 1 – 5 hold. Then the exchangeable bootstrap consistently estimates the limiting laws for the processes in Theorems 1 – 4.*

The exchangeable bootstrap distributions can be used to perform asymptotically valid inference for the causal effects of interest. Here I focus on uniform inference methods. These methods cover standard pointwise methods as special cases and, in addition, allow for testing richer functional parameters and hypotheses (Chernozhukov et al., 2013). For example, one can construct asymptotic simultaneous $(1-\alpha)$-confidence bands for the whole QTE process by inverting a Kolmogorov–Smirnov-type test:[13]

$$\hat{\delta}^{\pm}(\tau) = \hat{\delta}(\tau) \pm \hat{t}_{1-\alpha} \hat{\Sigma}(\tau)^{1/2} / \sqrt{n}$$

[13] Chernozhukov et al. (2013) and Melly and Santangelo (2015) use similar constructions.

such that

$$\lim_{n\to\infty} P\Big\{\delta(\tau) \in \Big[\hat{\delta}^{-}(\tau), \hat{\delta}^{+}(\tau)\Big] \text{ for all } \tau \in \mathcal{T}\Big\} = 1-\alpha,$$

where $\hat{\Sigma}(\tau)$ is a uniformly consistent estimator of $\Sigma(\tau)$, the asymptotic variance function of $\sqrt{n}\Big(\hat{\delta}(\tau) - \delta(\tau)\Big)$, and $\hat{t}_{1-\alpha}$ is a consistent estimator of the $(1-\alpha)$-quantile of the Kolmogorov–Smirnov maximal t-statistic,

$$t = \sup_{\tau\in\mathcal{T}} \sqrt{n}\hat{\Sigma}(\tau)^{-1/2} \ \big|\hat{\delta}(\tau) - \delta(\tau)\big|.$$

The critical value $\hat{t}_{1-\alpha}$ can be estimated using the exchangeable bootstrap. Uniform confidence bands for other functionals of interest can be obtained similarly.

5 Simulations

This section reports simulation evidence on the finite sample properties of the proposed estimation and inference procedures.

5.1 Setup

To make the simulations realistic, they are designed based on the empirical application in Section 6. The outcome of interest (Y) is net financial assets, the endogenous variable (D) is a binary indicator for 401(k) participation, and the instrument (Z) is a binary indicator for 401(k) eligibility. I consider a simplified setup with two continuous covariates: income (X_1) and age (X_2). To describe the data generating processes (DGPs), it is useful to write the potential outcome quantile functions under

Assumption 3 as

$$Q_{Y_d|X}(\tau \mid x) = G_d\Big(\tau, \big\{\Lambda\big(x'\gamma_z\big)\big\}_{z\in\mathcal{Z}}, \big\{\Lambda\big(x'\beta_{dz}(y)\big)\big\}_{(y,d,z)\in\mathcal{Y}\mathcal{D}\mathcal{Z}}\Big) \quad \text{for} \quad d \in \mathcal{D},$$

where G_d is (implicitly) defined as the inverse of the corresponding closed-form solution derived in Lemma 1.

I analyze five different DGPs that are designed to capture different degrees and types of misspecification. For all DGPs, I draw (X_1, X_2) from their joint empirical distribution and generate $Z \sim \text{Bernoulli}\big(\bar{Z}\big)$, where $\bar{Z}$ is the sample average of Z in the empirical application. The treatment D is obtained as

$$D = Z \cdot 1\{U > 0.9 \cdot V\}, \tag{14}$$

where $U \sim U(0,1)$ and $V \sim N(0,1)$ are mutually independent and independent of (X_1, X_2, Z).[14] The DGP for D in equation (14) is chosen to roughly match the joint empirical distribution of (D, Z). Potential outcomes are either generated as

$$Y_d = G_d\Big(U, \big\{\Lambda\big(X'\gamma_z\big)\big\}_{z\in\mathcal{Z}}, \big\{\Lambda\big(X'\beta_{dz}(y)\big)\big\}_{(y,d,z)\in\mathcal{Y}\mathcal{D}\mathcal{Z}}\Big), \tag{15}$$

or based on the following simple linear-in-parameters quantile model

$$Y_d = \delta_X(U)d + X'\beta(U). \tag{16}$$

Finally, observed outcomes Y are obtained as $Y = DY_1 + (1-D)Y_0$.

[14] To generate the potential outcomes as described below, the $U(0,1)$ distribution of U is approximated by a discrete uniform distribution with 100 points.

The five DGPs differ with respect to the outcome model, the specification of X, as well as the link function $\Lambda(\cdot)$:

1. Model (15) with $X = (X_1, X_2, 1)'$ and a Logit link.
2. Model (15) with $X = (X_1, X_2, 1)'$ and a linear link.
3. Model (15) with X as in the main specification in Section 6 (quadratic splines for income and age) and a Logit link.
4. Model (16) with $X = (X_1, X_2, 1)'$.
5. Model (16) with X as in the main specification in Section 6 (quadratic splines for income and age)

The coefficients in models (15) and (16) are set to the corresponding coefficient estimates based on the dataset from the empirical application. I estimate $\beta_{dz}(y)$ using DR, γ_z using binary choice models, and $(\delta_X(\tau), \beta(\tau)')'$ using IQR with X and $\Lambda(\cdot)$ as specified by the corresponding DGP.

I use DGP1 – DGP5 to assess and compare the following two estimators:

M1: Plug-in estimator with $X = (X_1, X_2, 1)'$ and a Logit link

M2: IQR estimator based on the linear model (13) with $X = (X_1, X_2, 1)'$

To enable a direct comparison of both approaches, I focus on unconditional QTE. Estimation of unconditional QTE based on M2 proceeds in three steps. First, the conditional potential outcome cdfs are computed by inverting the corresponding quantile functions estimated using IQR. Second, I compute unconditional cdfs by integrating the conditional cdfs with respect to the empirical distribution of the covariates. Finally, the unconditional cdfs are inverted to obtain estimates of the unconditional QTE. Note that the second and the third step are identical to the plug-in

estimation approach. Consequently, differences between the estimates of both models are solely due to differences between the estimated conditional potential outcome cdfs.

M1 is correctly specified under DGP1 and misspecified under DGP2 – DGP5, which cover different types of misspecification: a misspecified link function (DGP2), misspecification of the functional form of the linear index (DGP3), and misspecification of the structural quantile function (DGP4 and DGP5). M2 is correctly specified under DGP4 and misspecified under all other DGPs.

In all simulations, I approximate $\mathcal{Y}$ using a grid of 100 evenly-spaced quantiles of the true outcome distribution (removing duplicates) and $\mathcal{T}$ using a grid of 100 quantiles. For the IQR algorithm, I choose a grid of 100 evenly-spaced points between -10,000 and 35,000.

5.2 Bias and RMSE

Figure 1 displays the finite sample bias as a function of the quantile level for the sample sizes $n = 500$ and $n = 1000$ based on Monte Carlo simulations with 500 replications. The results show that, while both estimators perform well under correct specification (M1 under DGP1 and M2 under DGP4), their performance differs substantially when the underlying parametric assumptions are violated. M1 is robust to different types of misspecification, including misspecification of the structural quantile function (DGP4 and DGP5). By contrast, M2 exhibits large biases whenever the true structural quantile function is nonlinear, even if only X is misspecified (DGP5).

[Figure 1 about here.]

Figure 2 reports the estimated root mean squared error (RMSE). Under DGP1 – DGP3, the favorable robustness properties of M1 translate into a lower RMSE at most quantile levels, despite M2 exhibiting a somewhat lower variance. When M2 is correctly specified (DGP4), it exhibits a lower RMSE at most quantiles. This can mainly be attributed to its lower variance since the bias is similar for both models. Finally, under DGP5, M1 dominates M2 in the center of the distribution,

where the bias of M2 is maximal, while M2 exhibits a lower RMSE in the tails.

[Figure 2 about here.]

Two main conclusions can be drawn from these simulation results. First, the flexibility of the plug-in approach translates into good robustness properties against different types of misspecification of the underlying parametric restrictions. This is in sharp contrast to the IQR estimator based on model (13), which can be severely biased when the true structural quantile function is nonlinear. Second, the plug-in estimator exhibits a somewhat larger variance than IQR, suggesting that there is a bias-variance trade-off.

5.3 Coverage properties and power analysis

Here I examine the coverage rates of the uniform confidence bands and the power properties of the underlying Kolmogorov–Smirnov-type test. The critical values $\hat{t}_{1-\alpha}$ are estimated using the empirical bootstrap. I generate the data according to DGP1 described in Section 5.1. The unconditional QTE process implied by this DGP is positive and increasing across quantiles. To assess the power properties of the Kolmogorov–Smirnov-type test, I report rejection frequencies for testing the functional null hypothesis of a zero effect, i.e., $H_0 : \delta(\tau) = 0$ for all $\tau \in \mathcal{T}$.

In Table 1, I show empirical coverage rates and rejection probabilities for the sample sizes $n = 500$ and $n = 1000$. The results are based on 500 Monte Carlo replications. Critical values $\hat{t}_{1-\alpha}$ are estimated using 100 bootstrap replications. The uniform confidence bands exhibit close-to-correct empirical coverage for both sample sizes. In terms of power, the Kolmogorov–Smirnov-type tests exhibit good finite sample properties with rejection rates close to one, even in relatively small samples with $n = 500$ observations.

[Table 1 about here.]

6 Empirical application

In this section, I illustrate the proposed method by estimating the distributional impact of 401(k) plans on accumulated assets as in Chernozhukov and Hansen (2004) and Belloni et al. (2017).

6.1 Empirical setup

As explained by Chernozhukov and Hansen (2004), the 401(k) plans were introduced in the United States in the early 1980s in an effort to increase individual savings. 401(k) plans are provided by employers and allow individuals to deduct contributions from taxable income. The main challenge when estimating the effect of 401(k) plans (D) on accumulated assets is the potential endogeneity of the actual participation status caused by non-random enrollment. To overcome this problem, Abadie (2003), Chernozhukov and Hansen (2004), and Belloni et al. (2017) use 401(k) eligibility as an instrument (Z) for the actual participation status, arguing that eligibility can be taken to be exogenous after conditioning on income and a small set of other observable factors.[15] I adopt this identification strategy, noting that there are also arguments that eligibility is not conditionally exogenous (e.g., Engen et al., 1996).

I use the same dataset as Chernozhukov and Hansen (2004) and Belloni et al. (2017). The data consist of 9,913 observations from a sample of households from the 1991 Survey of Income and Program Participation (SIPP).[16] Descriptive statistics are presented in Tables 1 and 2 in Chernozhukov and Hansen (2004). The outcome variables of interest are two measures of wealth: net financial assets and total wealth. A key component of the argument underlying the exogeneity of 401(k) eligibility is that eligibility may only be taken as exogenous after conditioning on covariates and, in particular, income. I therefore use a flexible specification that includes quadratic splines

[15] This argument is detailed in Poterba et al. (1994, 1995, 1998) and Benjamin (2003).

[16] Note that the original dataset has 9,915 observations but I delete the two observations with a negative income.

in income and age[17], dummies for education categories, a marital status indicator, dummies for family size, two-earner status, defined benefit pension status, individual retirement account participation status, and homeownership. All computations have been carried out using the programming language R (R Core Team, 2018).

Because only individuals who were eligible could actually enroll in 401(k) plans, 401(k) eligibility satisfies one-sided non-compliance, i.e., $p(0,x) = 0$ for all $x \in \mathcal{X}$. One-sided non-compliance has two important implications. First, the key identification condition, Assumption 2.4, automatically holds under one-sided non-compliance (cf. Chernozhukov and Hansen, 2005, Section 2.4). Second, the computational burden is reduced because $p(0,x)$ and $F_{Y|D,Z,X}(y \mid 1,0,x)$ need not be estimated.

6.2 Specification tests

One important practical advantage of the plug-in estimation approach is that the underlying parametric assumptions (Assumption 3) are directly testable. I illustrate this point by assessing the validity of the parametric models for the conditional cdfs using the specification test by Rothe and Wied (2013). This test is based on Kolmogorov–Smirnov- and Cramér–von–Mises-type distances between the joint distribution of (Y,X) implied by the DR models and an unrestricted nonparametric alternative.

[Table 2 about here.]

Table 2 presents the results. Critical values are estimated using the semiparametric boostrap procedure of Rothe and Wied (2013) with 500 replications. The employed DR specifications are not rejected by the data for both outcomes, all three conditional cdfs, and both test statistics. This provides evidence in favor of the flexible DR specifications underlying the plug-in estimates reported in Section 6.3.

[17] Specifically, I allow for income, age, income-squared and age-squared and interact these variables with the seven income respectively five age dummies used in the original specification by Chernozhukov and Hansen (2004).

6.3 Empirical results and comparison to linear-in-parameters model

Figure 3 shows unconditional QTE estimated by the plug-in approach with a Logit link.[18] I approximate $\mathcal{Y}$ using a grid of 200 evenly-spaced quantiles of the unconditional empirical outcome distribution (removing duplicates) and $\mathcal{T}$ using a grid of 200 quantiles. 95% pointwise and uniform confidence bands are constructed based on the empirical bootstrap with 200 replications. For both outcomes, 401(k) participation has a small to moderate impact on accumulated assets at the low quantiles while having a much larger impact at high quantiles.[19] This pattern is more pronounced for net financial assets than for total wealth. Looking at the confidence bands, one can see that the estimates for total wealth are much noisier than those for net financial assets.

[Figure 3 about here.]

Figure 4 compares the plug-in estimates in Figure 3 to QTE estimated based on the linear-in-parameters model of Chernozhukov and Hansen (2004),

$$Q_{Y_d|X}(\tau \mid x) = d\delta_X(\tau) + x'\beta(\tau), \tag{17}$$

where the specification of X is the same as described in Section 6.1. Model (17) is estimated using the IQR algorithm with a grid search over 200 points. Unconditional QTE are obtained as described in Section 5.1.

[Figure 4 about here.]

The results based on the linear-in-parameters model indicate that the effect of 401(k) participation on net financial assets is increasing along the distribution, while being relatively constant for total wealth. There are considerable quantitative differences between the QTE estimates of both

[18] The results are robust with respect to the choice of the link function. Detailed results are available upon request.

[19] This pattern can be primarily attributed to the fact that the outcomes are measured in levels. Results for percentage treatment effects are provided upon request.

models. In particular, for total wealth, the relatively constant pattern of the estimates based on the linear model sharply contrasts the increasing shape of the QTE function estimated by the plug-in approach. Interestingly, the plug-in estimates are similar to the unconditional LQTE estimates based on a flexible nonparametric model reported in Belloni et al. (2017), while the results based on the linear-in-parameters IVQR model are comparable to the estimates based on linear LQTE models; see Chernozhukov and Hansen (2004).

6.4 Computational performance

One practical advantage of the plug-in estimation approach is that it remains computationally tractable, while allowing for rich patterns of effect heterogeneity with respect to covariates. Table 3 reports computation times for estimating the unconditional QTE process for net financial assets.[20] I consider two different specifications for X: a simple specification where the continuous and multi-valued covariates enter linearly ($\dim(X) = 12$) and the quadratic splines specification described in Section 6.1 ($\dim(X) = 48$). For comparison, I also report the corresponding computation times for the IQR estimator based on the linear-in-parameters model (17). Note that the computationally expensive steps are computing the conditional cdfs (plug-in), computing the conditional quantile functions (IQR), as well as inverting distribution and quantile functions at various steps (both methods).

[Table 3 about here.]

Table 3 shows that the plug-in estimator remains tractable with flexible specifications of X and fine grids for $\mathcal{Y}$ and $\mathcal{T}$ in samples of practically relevant size. Computation times for the IQR estimator are at least one order of magnitude higher and IQR becomes computationally prohibitive for flexible models, especially when $\dim(X) = 48$.

[20] Computations were carried out on a standard desktop computer with a 3.2 GHz Intel Core i5 processor and 8GB RAM.

7 Conclusion and directions for future research

This paper proposes a practical and flexible approach for estimating QTE based on the IVQR model. The key idea is to exploit analytic closed-form solutions to construct plug-in estimators. The proposed approach remains computationally tractable and root-n-consistent, while allowing for rich patterns of treatment effect heterogeneity with respect to observable characteristics. I prove functional central limit theorems and establish the validity of the exchangeable bootstrap for estimating the limiting laws. Monte Carlo simulations demonstrate favorable properties of the proposed approach in finite samples. I apply the plug-in estimator to reanalyze the effect of 401(k) plans on individual assets. My findings suggest that the effect of 401(k) plans is positive and increasing along the distribution. Interestingly, these results imply substantially more effect heterogeneity than the estimates based on the linear-in-parameters IVQR model of Chernozhukov and Hansen (2004).

The proposed plug-in estimators rely on parametric models for the conditional cdfs and the conditional probabilities. It would be interesting to extend the estimation approach to accommodate fully nonparametric first-stage estimators. Nonparametric plug-in estimators could also serve as the basis for developing specification tests for linear-in-parameters models. These extensions are beyond the scope of this paper but certainly worth pursuing in future research.

One important limitation of the proposed method is that it is inherently limited to binary treatments since closed-from solutions are only available for this important special case. Moreover, despite the fact that nonbinary instruments can be accommodated as described in Appendix B, the efficiency of the plug-in estimators could be improved by extending Lemma 1 to nonbinary instruments and overidentified GMM objective functions. Such results could be obtained by extending Lemma 1 using similar arguments as in Wüthrich (2018). Deriving more general analytic closed-form solutions and extending the plug-in approach accordingly constitutes a promising

extension.

Acknowledgements

This paper was previously circulated under the title "Semiparametric estimation of quantile treatment effects with endogeneity" and is based on a chapter of my PhD dissertation at the University of Bern. I have benefited from numerous discussions with Blaise Melly. I would like to thank Isaiah Andrews, Daniel Burkhard, Andreas Bachmann, Victor Chernozhukov, Lutz Dümbgen, Iván Fernández-Val, Anna Mikusheva, Andres Santos, Yixiao Sun, three anonymous Referees, an Associate Editor, the Editor, and seminar participants at MIT, the University of Bern, and the California Econometrics Conference 2016 for very helpful comments. Pietro Spini provided excellent research assistance. I would like to thank Chris Hansen for providing me with the data for the empirical application. Financial support by the Swiss National Science Foundation (Doc.Mobility Project P1BEP1_155467) is gratefully acknowledged. The usual disclaimer applies.

References

Abadie, A., 2003. Semiparametric instrumental variable estimation of treatment response models. Journal of Econometrics 11, pp. 231–263.

Abadie, A., Angrist, J., Imbens, G., 2002. Instrumental variable estimates of the effect of subsidized training on the quantile of trainee earnings. Econometrica 70 (1), pp. 91–117.

Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C., 2017. Program evaluation and causal inference with high-dimensional data. Econometrica 85 (1), 233–298.

Benjamin, D., 2003. Does 401(k) eligibility increase saving?: Evidence from propensity score subclassification. Journal of Public Economics 87, pp. 1259–1290.

Carneiro, P., Lee, S., 2009. Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. Journal of Econometrics 149 (2), pp. 191–208.

Chen, L.-Y., Lee, S., 2018. Exact computation of gmm estimators for instrumental variable quantile regression models. Journal of Applied Econometrics 33 (4), 553–567.

Chen, X., Pouzo, D., 2009. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. Journal of Econometrics 152 (1), pp. 46–60.

Chen, X., Pouzo, D., 2012. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. Econometrica 80 (1), pp. 277–321.

Chernozhukov, V., Fernandez-Val, I., Galichon, A., 2010. Quantile and probability curves without crossing. Econometrica 78 (3), pp. 1093–1125.

Chernozhukov, V., Fernandez-Val, I., Melly, B., 2013. Inference on counterfactual distributions. Econometrica 81 (6), pp. 2205–2268.

Chernozhukov, V., Hansen, C., 2004. The effects of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis. The Review of Economics and Statistics 86 (3), pp. 735–751.

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

Chernozhukov, V., Hansen, C., 2006. Instrumental quantile regression inference for structural and treatment effects models. Journal of Econometrics 132, pp. 491–525.

Chernozhukov, V., Hansen, C., 2008. Instrumental variable quantile regression: A robust inference approach. Journal of Econometrics 142 (1), pp. 379–398.

Chernozhukov, V., Hansen, C., 2013. Quantile models with endogeneity. Annual Review of Economics 5 (1), pp. 57–81.

Chernozhukov, V., Hansen, C., Jansson, M., 2007a. Inference approaches for instrumental variable quantile regression. Economics Letters 95 (2), pp. 272–277.

Chernozhukov, V., Hansen, C., Jansson, M., 2009. Finite sample inference for quantile regression models. Journal of Econometrics 152 (2), pp. 93–103.

Chernozhukov, V., Hansen, C., Wüthrich, K., 2017. Instrumental variable quantile regression. In: Chernozhukov, V., He, X., Koenker, R., Peng, L. (Eds.), Handbook of Quantile Regression. CRC Chapman-Hall, pp. 119–143.

Chernozhukov, V., Hong, H., 2003. An MCMC approach to classical estimation. Journal of Econometrics 115 (2), pp. 293–346.

Chernozhukov, V., Imbens, G. W., Newey, W. K., 2007b. Instrumental variable estimation of nonseparable models. Journal of Econometrics 139 (1), pp. 4–14.

Chesher, A., 2003. Identification in nonseparable models. Econometrica 71 (5), pp. 1405–1441.

De Chaisemartin, C., 2017. Tolerating defiance? Local average treatment effects without monotonicity. Quantitative Economics 8 (2), 367–396.

De Chaisemartin, C., D'Haultfoeuille, X., 2018. Fuzzy differences-in-differences. The Review of Economic Studies 85 (2), 999–1028.

D'Haultfoeuille, X., Février, P., 2015. Identification of nonseparable triangular models with discrete instruments. Econometrica 83 (3), pp. 1199–1210.

Duflo, E., Saez, E., 2003. The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. The Quarterly Journal of Economics 118 (3), 815–842.

Engen, E. M., Gale, W. G., Scholz, J. K., 1996. The illusory effects of saving incentives on saving. Journal of Economic Perspectives 10 (4), pp. 113–138.

Firpo, S., 2007. Efficient semiparametric estimation of quantile treatment effects. Econometrica 75 (1), pp. 259–276.

Foresi, S., Peracchi, F., 1995. The conditional distribution of excess returns: An empirical analysis. Journal of the American Statistical Association 90 (430), pp. 451–466.

Frandsen, B. R., Frölich, M., Melly, B., 2012. Quantile treatment effects in the regression discontinuity design. Journal of Econometrics 168 (2), pp. 382–395.

Frölich, M., Melly, B., 2013. Unconditional quantile treatment effects under endogeneity. Journal of Business & Economic Statistics 31 (3), pp. 346–357.

Gagliardini, P., Scaillet, O., 2012. Nonparametric instrumental variable estimation of structural quantile effects. Econometrica 80 (4), pp. 1533–1562.

Horowitz, J. L., Lee, S., 2007. Nonparametric instrumental variables estimation of a quantile regression model. Econometrica 75 (4), pp. 1191–1208.

Hsu, Y.-C., Lai, T.-C., Lieli, R. P., 2015. Estimation and inference for distribution functions and quantile functions in endogenous treatment effect models, iEAS Working Paper, 15-A003.

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

Imbens, G. W., Newey, W. K., 2009. Identification and estimation of triangular simultaneous equations models without additivity. Econometrica 77 (5), pp. 1481–1512.

Jun, S. J., 2009. Local structural quantile effects in a model with a nonseparable control variable. Journal of Econometrics 151 (1), 82 – 97.

Kaplan, D. M., Sun, Y., 2017. Smoothed estimating equations for instrumental variables quantile regression. Econometric Theory 33 (1), 105 – 157.

Koenker, R., Bassett, Gilbert, J., 1978. Regression quantiles. Econometrica 46 (1), pp. 33–50.

Lee, S., 2007. Endogeneity in quantile regression models: A control function approach. Journal of Econometrics 141 (2), pp. 1131–1158.

Leorato, S., Peracchi, F., 2015. Comparing distribution and quantile regression. EIEF Working

Paper 15/11.

Ma, L., Koenker, R., 2006. Quantile regression methods for recursive structural equation models. Journal of Econometrics 134 (2), pp. 471 – 506.

Melly, B., Santangelo, G., 2015. The changes-in-changes model with covariates. Working Paper.

Melly, B., Wüthrich, K., 2017. Local quantile treatment effects. In: Chernozhukov, V., He, X., Koenker, R., Peng, L. (Eds.), Handbook of Quantile Regression. CRC Chapman-Hall, pp. 145–164.

Poterba, J. M., Venti, S. F., Wise, D. A., 1994. 401(k) plans and tax-deferred saving. In: Wise, D. A. (Ed.), Studies in the Economics of Aging. University of Chicago Press.

Poterba, J. M., Venti, S. F., Wise, D. A., 1995. Do 401(k) contributions crowd out other personal saving? Journal of Public Economics 58 (1), pp. 1–32.

Poterba, J. M., Venti, S. F., Wise, D. A., 1998. Personal retirement saving programs and asset accumulation: Reconciling the evidence. In: Wise, D. A. (Ed.), Frontiers in the Economics of Aging. University of Chicago Press.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rothe, C., Wied, D., 2013. Misspecification testing in a class of conditional distributional models. Journal of the American Statistical Association, pp. 314–324.

Rubin, D. B., 1974. Estimating causal effects of treatment in randomized and nonrandomized studies. Journal of Educational Psychology 66 (5), pp. 688–701.

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

Van der Vaart, A., Wellner, J., 1996. Weak Convergence and Empirical Processes: With Application to Statistics. Springer-Verlag.

Williams, O. D., Grizzle, J. E., 1972. Analysis of contingency tables having ordered response categories. Journal of the American Statistical Association 67 (337), 55–63.

Wüthrich, K., 2018. A comparison of two quantile models with endogeneity. Journal of Business and Economic Statistics (forthcoming).

Yu, P., 2014. Marginal quantile treatment effect and counterfactual analysis. Working Paper, HKU.

Appendix to "A closed-form estimator for quantile treatment effects with endogeneity"

Kaspar Wüthrich

University of California, San Diego, email: kwuthrich@ucsd.edu.

A Assumption 2.4 under the LQTE framework

To develop further intuition for the full rank condition in Assumption 2, it is useful to analyze this abstract condition under the LQTE framework (Abadie et al., 2002).

To lighten up the exposition, the conditioning on $X = x$ is suppressed. Let (D_1, D_0) denote the potential treatments (indexed by Z). Based on their compliance behavior, individuals can be classified by four different types $T = t$: never-takers ($D_0 = D_1 = 0$; $T = n$), always-takers ($D_0 = D_1 = 1$; $T = a$), defiers ($D_0 = 1, D_1 = 0$; $T = d$), and compliers ($D_0 = 0, D_1 = 1$; $T = c$). In what follows, I denote the proportions, potential outcome cdfs, and densities for type $T = t$ as π_t, $F_{Y_d|t}$, and $f_{Y_d|t}$. The LQTE framework is based on the following assumptions (e.g., Abadie et al., 2002, Assumption 2.1).

Assumption 6 (LQTE Framework)**.**

1. *Independence:* (Y_1, Y_0, D_1, D_0) *are jointly independent of* Z*.*
2. *Nontrivial assignment:* $P(Z = 1) \in (0, 1)$*.*
3. *First-stage:* $\mathbb{E}[D_1] \neq \mathbb{E}[D_0]$*.*
4. *Monotonicity:* $P(D_1 \geq D_0) = 1$*.*

The key restriction of the LQTE framework is the monotonicity Assumption 6.4, which rules out the existence of defiers. Consequently, never-takers, always-takers, and compliers exhaustively

partition the whole population. I refer to Melly and Wüthrich (2017) for a detailed discussion of Assumption 6.

In the proof of Lemma 1, I show that Assumption 2.4 is equivalent to $f^c_{Y_1}$ and $f^c_{Y_0}$ being non-zero. Under Assumption 6, $f^c_{Y_1}$ and $f^c_{Y_0}$ correspond to potential outcome density functions of the compliers multiplied by the fraction of compliers:

$$f^c_{Y_1}(y_1) = \pi_c f_{Y_1|c}(y_1) \quad \text{and} \quad f^c_{Y_0}(y_0) = \pi_c f_{Y_0|c}(y_0).$$

Because the fraction of compliers, π_c, is strictly positive by Assumption 6.3, Assumption 2.4 is equivalent to

$$f_{Y_1|c}(y_1) > 0 \quad \text{and} \quad f_{Y_0|c}(y_0) > 0.$$

In other words, under the LQTE framework, the IVQR full rank assumption is equivalent to full support of the complier potential outcome distributions. This analysis provides an easy-to-interpret sufficient condition for the abstract IVQR identification assumption. Moreover, it suggests simple setups where the full rank condition fails. For example, if the support of Y_0 (Y_1) for the never-takers (always-takers) is not contained in the support of Y_0 (Y_1) for the compliers.

Remark 6. *We know from Vytlacil (2002) that Assumption 6 is equivalent to a threshold crossing model, $D = 1\{p(Z) \geq V\}$, where $V \sim U(0,1)$ is independent of Z. This equivalence result provides an alternative way of analyzing the full rank condition under the LQTE framework.*[21] *Based on the threshold crossing model for the selection equation, compliers can be described by $V \in (p(0), p(1)]$.*[22] *It follows that $F^c_{Y_1}(y) = P(p(0) < V \leq p(1))P(Y_1 \leq y \mid p(0) < V \leq p(1))$. Because $P(p(0) < V \leq$*

[21] I would like to thank an anonymous referee for suggesting this interpretation.
[22] Without loss of generality it is assumed that $p(0) < p(1)$.

$p(1)) > 0$ *by Assumption 6.3, Assumption 2.4 is equivalent to*

$$f_{Y_1|c}(y) = \int_{p(0)}^{p(1)} f_{Y_1|V}(y \mid v)dv > 0, \tag{18}$$

where $f_{Y_1|V}$ *is the density of* $Y_1 \mid V = v$*. In other words, Assumption 2.4 requires that* $f_{Y_1|V}(y \mid v) > 0$ *for all* v *in some subinterval of* $(p(0), p(1)]$*. That is, full rank is equivalent to full support for at least a subpopulation of the compliers. A similar result can be derived for* $F^c_{Y_0}$*.*

B Nonbinary instruments and simple specification tests

If the original instrument is multivalued or continuous, estimation can proceed based on a dichotomized version of instrument. If there are multiple instruments, the same strategy can be applied based on the propensity score $p(Z, X)$. Under the assumptions put forth in the main text, the choice of the dichotomization does not matter for consistency of the estimators. However, efficiency could be improved by developing plug-in estimators based on analytic solutions with general instruments.

With nonbinary instruments one can construct simple overidentification-type specification tests. To fix ideas, suppose that the researcher has access to two different binary instruments Z_1 and Z_2, which are obtained as transformations of the original instrument(s), and let $\delta_{Z_1,X}(\tau \mid x)$ and $\delta_{Z_2,X}(\tau \mid x)$ denote the associated QTE estimands. Under Assumption 1, the conditional moment restriction (1) implies that

$$\delta_{Z_1,X}(\tau \mid x) = \delta_{Z_2,X}(\tau \mid x) \quad \text{for all} \quad (\tau, x) \in \mathcal{T}\mathcal{X}.$$

The intuition behind this testable restriction is as follows. Under Assumption 1, the IVQR model identifies QTE for the whole population. These treatment effects do not depend on the choice of the

instrument as they are not local effects for an instrument-specific subpopulation.[23] Consequently, one can construct overidentification-type specification tests based on the following testing problem

$$H_0 : \delta_{Z_1,X}(\tau \mid x) = \delta_{Z_2,X}(\tau \mid x) \quad \text{for all} \quad (\tau, x) \in \mathcal{TX}$$

against the unrestricted alternative.[24] Kolmogorov–Smirnov- and Cramér–von Mises-type test statistics can be used to test this hypothesis. Under the conditions presented in the main text, critical values can be obtained using the exchangeable bootstrap.

C Proofs

Notation

To ease the exposition, I introduce some additional notation. Define

$$\begin{aligned}
\gamma &:= (\gamma_0', \gamma_1')', \\
\beta(y_0, y_1) &:= (\beta_{00}(y_0)', \beta_{01}(y_0)', \beta_{10}(y_1)', \beta_{11}(y_1)')', \\
W(y_0, y_1) &:= (W_{\gamma_0}', W_{\gamma_1}', W_{\beta_{00}}(y_0)', W_{\beta_{01}}(y_0)', W_{\beta_{10}}(y_1)', W_{\beta_{11}}(y_1)')'
\end{aligned}$$

where, for all $(d, z) \in \mathcal{DZ}$,

$$\begin{aligned}
W_{\gamma_z} &:= \mathbb{G}(\kappa_{\gamma_z}), \\
W_{\beta_{dz}}(y_d) &:= \mathbb{G}(\kappa_{\beta_{dz}}(y_d)),
\end{aligned}$$

[23] This is in sharp contrast to the LQTE framework; see Section 3.2.2 and Appendix A.

[24] Since the first version of this paper was written, Yu (2017) and Kim and Park (2017) have developed formal overidentification-type testing procedures for rank similarity based on similar insights.

and

$$\begin{aligned}\kappa_{\gamma_z} &:= 1\{Z=z\}\big[\Lambda\big(X'\gamma_z\big)-D\big]H(X'\gamma_z)X,\\ \kappa_{\beta_{dz}}(y_d) &:= 1\{D=d,Z=z\}\big[\Lambda\big(X'\beta_{dz}(y_d)\big)-1(Y\leq y_d)\big]H(X'\beta_{dz}(y_d))X,\end{aligned}$$

where $H(\cdot) := \lambda(\cdot)/\{\Lambda(\cdot)[1-\Lambda(\cdot)]\}$ and $\mathbb{G}$ is a P-Brownian bridge. Finally, define the matrix $J(y_0, y_1)$ as

$$J(y_0,y_1) := \begin{pmatrix} J_{\gamma_0} & 0 & 0 & 0 & 0 & 0\\ 0 & J_{\gamma_1} & 0 & 0 & 0 & 0\\ 0 & 0 & J_{\beta_{00}}(y_0) & 0 & 0 & 0\\ 0 & 0 & 0 & J_{\beta_{01}}(y_0) & 0 & 0\\ 0 & 0 & 0 & 0 & J_{\beta_{10}}(y_1) & 0\\ 0 & 0 & 0 & 0 & 0 & J_{\beta_{11}}(y_1)\end{pmatrix}.$$

Proof of Lemma 1

To simplify the exposition, I suppress the conditioning on $X = x$ and use the notation $p_{dz} := P(D = d \mid Z = z)$. The proof proceeds in two steps. First, I show that Assumption 2 is equivalent to $f^c_{Y_1}$ and $f^c_{Y_0}$ being non-zero, which implies strict monotonicity of $F^c_{Y_1}$ and $F^c_{Y_0}$. In the second step, I exploit strict monotonicity to derive the closed-form solutions.

Step 1: Here I show that full rank of $\Pi'(y_0, y_1)$ is equivalent to $f^c_{Y_1}$ and $f^c_{Y_0}$ being both strictly positive or strictly negative. I start with two preliminary observations. First, note that full rank of $\Pi'(y_0, y_1)$ is equivalent to $\det(\Pi'(y_0, y_1)) \neq 0$, where

$$\det\big(\Pi'(y_0,y_1)\big) = f_{Y|D,Z}(y_0 \mid 0,0)p_{00}f_{Y|D,Z}(y_1 \mid 1,1)p_{11} - f_{Y|D,Z}(y_0 \mid 0,1)p_{01}f_{Y|D,Z}(y_1 \mid 1,0)p_{10}.$$

Second, by definition of $F^c_{Y_1}$ and $F^c_{Y_0}$,

$$f^c_{Y_1}(y_1) = f_{Y|D,Z}(y_1 \mid 1,1)p_{11} - f_{Y|D,Z}(y_1 \mid 1,0)p_{10} \tag{19}$$

and

$$f^c_{Y_0}(y_0) = f_{Y|D,Z}(y_0 \mid 0,0)p_{00} - f_{Y|D,Z}(y_0 \mid 0,1)p_{01}. \tag{20}$$

FULL RANK $\Rightarrow$ MONOTONICITY:[25]

As shown in equation (24) below, the IVQR moment restrictions (1) imply that $F^c_{Y_1}(Q_{Y_1}(\tau)) = F^c_{Y_0}(Q_{Y_0}(\tau))$. By Assumption 1.1, Q_{Y_1} and Q_{Y_0} are strictly increasing and thus admit strictly increasing inverses. Thus, substituting $\tau = F_{Y_0}(y_0)$ yields

$$F^c_{Y_0}(y_0) = F^c_{Y_1}(Q_{Y_1}(F_{Y_0}(y_0))) \tag{21}$$

for $y_0 \in \{Q_{Y_0}(\tau), \tau \in (0,1)\}$. Equation (21) implies that $f^c_{Y_0}(y_0)$ and $f^c_{Y_1}(y_1(y_0))$ are either both positive, zero, or negative for $y_0 \in \{Q_{Y_0}(\tau), \tau \in (0,1)\}$, where $y_1(y_0) := Q_{Y_1}(F_{Y_0}(y_0))$.

Next, I show that $f^c_{Y_0}(y_0) \neq 0$ and $f^c_{Y_1}(y_1(y_0)) \neq 0$ on $\{Q_{Y_0}(\tau), \tau \in (0,1)\}$ by contradiction. Suppose not, then there exists $y_0^* \in \{Q_{Y_0}(\tau), \tau \in (0,1)\}$ such that $f^c_{Y_0}(y_0^*) = 0$ and $f^c_{Y_1}(y_1(y_0^*)) = 0$. Thus, by (19) and (20),

$$f_{Y|D,Z}(y_0^* \mid 0,0)p_{00} = f_{Y|D,Z}(y_0^* \mid 0,1)p_{01} \quad \text{and} \quad f_{Y|D,Z}(y_1(y_0^*) \mid 1,1)p_{11} = f_{Y|D,Z}(y_1(y_0^*) \mid 1,0)p_{10},$$

which implies that $\det(\Pi'(y_0^*, y_1(y_0^*))) = 0$ and thus that $\Pi'(y_0^*, y_1(y_0^*)))$ is not of full rank.

Because $\{Q_{Y_1}(\tau), \tau \in (0,1)\} = Q_{Y_1}(F_{Y_0}(\{Q_{Y_0}(\tau), \tau \in (0,1)\}))$, it follows that $f^c_{Y_d}(y_d) \neq 0$ on $\{Q_{Y_d}(\tau), \tau \in (0,1)\}$ for $d \in \mathcal{D}$. Moreover, by continuity, $f^c_{Y_d}(y_d) > 0$ or < 0 for all $y_d \in \{Q_{Y_d}(\tau), \tau \in$

[25] This step is similar to the proof in the Online Supplement B.3 of Vuong and Xu (2017).

$(0,1)\}$. The desired implication follows because, by equation (21), $f^c_{Y_0}(y_0)$ and $f^c_{Y_1}(y_1(y_0))$ have the same sign for all $y_0 \in \{Q_{Y_0}(\tau), \tau \in (0,1)\}$.

MONOTONICITY $\Rightarrow$ FULL RANK: Consider the case where $f^c_{Y_1}$ and $f^c_{Y_0}$ are both strictly positive. Thus,

$$\begin{aligned} f^c_{Y_1}(y_1) &= f_{Y|D,Z}(y_1 \mid 1,1)p_{11} - f_{Y|D,Z}(y_1 \mid 1,0)p_{10} > 0, \\ f^c_{Y_0}(y_0) &= f_{Y|D,Z}(y_0 \mid 0,0)p_{00} - f_{Y|D,Z}(y_0 \mid 0,1)p_{01} > 0. \end{aligned}$$

Consequently, we have

$$f_{Y|D,Z}(y_0 \mid 0,0)p_{00} f_{Y|D,Z}(y_1 \mid 1,1)p_{11} > f_{Y|D,Z}(y_0 \mid 0,1)p_{01} f_{Y|D,Z}(y_1 \mid 1,0)p_{10},$$

implying that $\det(\Pi'(y_0, y_1)) \neq 0$, which is in turn equivalent to full rank of the Jacobian. The case when $f^c_{Y_1}$ and $f^c_{Y_0}$ are strictly negative is symmetric and thus omitted. This proves the desired implication.

Step 2: I now establish the closed-form solutions. Under Assumption 1, by Theorem 1 in Chernozhukov and Hansen (2005),

$$\begin{aligned} P(Y \leq Q_{Y_D}(\tau) \mid Z = 1) = \tau \\ P(Y \leq Q_{Y_D}(\tau) \mid Z = 0) = \tau. \end{aligned}$$

By the law of iterated expectations and the definition of a conditional cdf,

$$p_{11} F_{Y|D,Z}(Q_{Y_1}(\tau) \mid 1,1) + p_{01} F_{Y|D,Z}(Q_{Y_0}(\tau) \mid 0,1) = \tau \quad (22)$$

$$p_{10} F_{Y|D,Z}(Q_{Y_1}(\tau) \mid 1,0) + p_{00} F_{Y|D,Z}(Q_{Y_0}(\tau) \mid 0,0) = \tau. \quad (23)$$

Equating (22) and (23) and rearranging terms yields

$$p_{11}F_{Y|D,Z}(Q_{Y_1}(\tau) \mid 1,1) - p_{10}F_{Y|D,Z}(Q_{Y_1}(\tau) \mid 1,0) =$$
$$p_{00}F_{Y|D,Z}(Q_{Y_0}(\tau) \mid 0,0) - p_{01}F_{Y|D,Z}(Q_{Y_0}(\tau) \mid 0,1)$$

Thus,

$$F^c_{Y_1}(Q_{Y_1}(\tau)) = F^c_{Y_0}(Q_{Y_0}(\tau)), \tag{24}$$

by definition. By step 1, $F^c_{Y_1}$ and $F^c_{Y_0}$ are strictly monotonic and thus one-to-one with strictly monotonic inverses. Moreover, Q_{Y_1} and Q_{Y_0} are strictly increasing by Assumption 1.1 and thus one-to-one with strictly increasing inverses. Therefore,

$$Q^c_{Y_1}\left(F^c_{Y_0}(y_0)\right) = Q_{Y_1}(F_{Y_0}(y_0)) \tag{25}$$

and

$$Q^c_{Y_0}\left(F^c_{Y_1}(y_1)\right) = Q_{Y_0}(F_{Y_1}(y_1)). \tag{26}$$

Substituting $F_{Y_1}(y_1) = \tau$ in equation (22) and $F_{Y_0}(y_0) = \tau$ in equation (23), we obtain

$$\begin{aligned} F_{Y_1}(y_1) &= p_{11}F_{Y|D,Z}(y_1 \mid 1,1) + p_{01}F_{Y|D,Z}(Q_{Y_0}(F_{Y_1}(y_1)) \mid 0,1), \\ F_{Y_0}(y_0) &= p_{00}F_{Y|D,Z}(y_0 \mid 0,0) + p_{10}F_{Y|D,Z}(Q_{Y_1}(F_{Y_0}(y_0)) \mid 1,0). \end{aligned}$$

The result then follows by plugging-in (25) and (26).

Proof of Theorem 1

The proof has two steps. In the first step, I show that the conditional probabilities and conditional distributions converge jointly to tight mean-zero Gaussian processes. This step builds on the proof strategy detailed in Chernozhukov et al. (2013) and Yu (2014). The second step shows that the closed-form solutions are Hadamard differentiable maps, partly building on work by Melly and Santangelo (2015) and De Chaisemartin and D'Haultfoeuille (2018,b).

Step 1: Under Assumptions 3 and 4.1 – 4.3, it follows from the arguments in Chernozhukov et al. (2013, Appendix E) and Yu (2014, Proofs of Theorems 4 and 8) that

$$\sqrt{n}\begin{pmatrix} \hat{p}(0,x) - p(0,x) \\ \hat{p}(1,x) - p(1,x) \\ \hat{F}_{Y|D,Z,X}(y_0 \mid 0,0,x) - F_{Y|D,Z,X}(y_0 \mid 0,0,x) \\ \hat{F}_{Y|D,Z,X}(y_0 \mid 0,1,x) - F_{Y|D,Z,X}(y_0 \mid 0,1,x) \\ \hat{F}_{Y|D,Z,X}(y_1 \mid 1,0,x) - F_{Y|D,Z,X}(y_1 \mid 1,0,x) \\ \hat{F}_{Y|D,Z,X}(y_1 \mid 1,1,x) - F_{Y|D,Z,X}(y_1 \mid 1,1,x) \end{pmatrix} \rightsquigarrow \varphi_{\gamma,\beta(\cdot)}\big(-J^{-1}(y_0,y_1)W(y_0,y_1)\big) =: \begin{pmatrix} \mathbb{Z}_{p_0}(x) \\ \mathbb{Z}_{p_1}(x) \\ \mathbb{Z}_{F_{00}}(y_0 \mid x) \\ \mathbb{Z}_{F_{01}}(y_0 \mid x) \\ \mathbb{Z}_{F_{10}}(y_1 \mid x) \\ \mathbb{Z}_{F_{11}}(y_1 \mid x) \end{pmatrix},$$

a tight mean-zero Gaussian process, where the map $\varphi_{\gamma,\beta(\cdot)}(\eta,\alpha)(y_0,y_1,x)$ is given by

$$\varphi_{\gamma,\beta(\cdot)}(\eta,\alpha)(y_0,y_1,x) = \begin{pmatrix} \lambda(x'\gamma_0)x'\eta_0 \\ \lambda(x'\gamma_1)x'\eta_1 \\ \lambda(x'\beta_{00}(y_0))x'\alpha_{00}(y_0) \\ \lambda(x'\beta_{01}(y_0))x'\alpha_{01}(y_0) \\ \lambda(x'\beta_{10}(y_1))x'\alpha_{10}(y_1) \\ \lambda(x'\beta_{11}(y_1))x'\alpha_{11}(y_1) \end{pmatrix}.$$

Hence, the details are omitted for brevity.

Step 2: This step establishes Hadamard differentiability of the closed-form solution. To simplify the exposition and to keep track of the exact expressions for the limit processes, I proceed step-by-step, which is justified by the chain rule for Hadamard derivatives (e.g., Van der Vaart and Wellner, 1996, Lemma 3.9.3), and separately for $F_{Y_1|X}$ and $F_{Y_0|X}$. Consider first $\hat{F}^c_{Y_1|X}(y_1 \mid x)$ and $\hat{F}^c_{Y_0|X}(y_0 \mid x)$ defined by equations (11) and (12) in the main text. By the functional delta method,

$$\sqrt{n}\begin{pmatrix} \hat{F}^c_{Y_1|X}(y_1 \mid x) - F^c_{Y_1|X}(y_1 \mid x) \\ \hat{F}^c_{Y_0|X}(y_0 \mid x) - F^c_{Y_0|X}(y_0 \mid x) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{Z}_{F^c_1}(y_1 \mid x) \\ \mathbb{Z}_{F^c_0}(y_0 \mid x) \end{pmatrix} \quad \text{in} \quad \ell^\infty(\mathcal{Y}\mathcal{X})^2,$$

where $\mathbb{Z}_{F^c_1}$ and $\mathbb{Z}_{F^c_0}$ are tight mean-zero Gaussian processes given by

$$\begin{aligned} \mathbb{Z}_{F^c_1}(y_1 \mid x) := \quad & F_{Y|D,Z,X}(y_1 \mid 1,1,x)\mathbb{Z}_{p_1}(x) - F_{Y|D,Z,X}(y_1 \mid 1,0,x)\mathbb{Z}_{p_0}(x) \\ & + p(1,x)\mathbb{Z}_{F_{11}}(y_1 \mid x) - p(0,x)\mathbb{Z}_{F_{10}}(y_1 \mid x) \end{aligned}$$

and

$$\begin{aligned} \mathbb{Z}_{F^c_0}(y_0 \mid x) \quad := \quad & F_{Y|D,Z,X}(y_0 \mid 0,1,x)\mathbb{Z}_{p_1}(x) - F_{Y|D,Z,X}(y_0 \mid 0,0,x)\mathbb{Z}_{p_0}(x) \\ & + (1-p(0,x))\mathbb{Z}_{F_{00}}(y_0 \mid x) - (1-p(1,x))\mathbb{Z}_{F_{01}}(y_0 \mid x). \end{aligned}$$

Under Assumptions 4.2 and 4.4, the inverse map is Hadamard differentiable uniformly with respect to an index (e.g., Chernozhukov et al. (2010)). Therefore, by the functional delta method

$$\sqrt{n}\begin{pmatrix} \hat{Q}^c_{Y_1|X}(\tau \mid x) - Q^c_{Y_1|X}(\tau \mid x) \\ \hat{Q}^c_{Y_0|X}(\tau \mid x) - Q^c_{Y_0|X}(\tau \mid x) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{Z}_{Q^c_1}(\tau \mid x) \\ \mathbb{Z}_{Q^c_0}(\tau \mid x) \end{pmatrix},$$

where $\mathbb{Z}_{Q_1^c}$ and $\mathbb{Z}_{Q_1^c}$ are tight mean-zero Gaussian processes given by

$$\begin{aligned}\mathbb{Z}_{Q_1^c}(\tau \mid x) &:= -\mathbb{Z}_{F_1^c}\Big(Q^c_{Y_1|X}(\tau \mid x) \mid x\Big)/f^c_{Y_1|X}\Big(Q^c_{Y_1|X}(\tau \mid x) \mid x\Big)\\ \mathbb{Z}_{Q_0^c}(\tau \mid x) &:= -\mathbb{Z}_{F_0^c}\Big(Q^c_{Y_0|X}(\tau \mid x) \mid x\Big)/f^c_{Y_0|X}\Big(Q^c_{Y_0|X}(\tau \mid x) \mid x\Big)\end{aligned}$$

By Assumptions 4.2 and 4.4, Lemma 3.9.27 in Van der Vaart and Wellner (1996) (which is valid uniformly with respect to an index under Assumption 4.2 and 4.4) and the functional delta method imply

$$\sqrt{n}\Big(\hat{Q}^c_{Y_0|X}\Big(\hat{F}^c_{Y_1|X}(y_1 \mid x) \mid x\Big) - Q^c_{Y_0|X}\Big(F^c_{Y_1|X}(y_1 \mid x) \mid x\Big)\Big) \rightsquigarrow \mathbb{Z}_{Q_0^c \circ F_1^c}(y_1 \mid x) \text{ in } \ell^\infty(\mathcal{Y}\mathcal{X}),$$

where $\mathbb{Z}_{Q_0^c \circ F_1^c}$ is a mean-zero Gaussian process defined as

$$\mathbb{Z}_{Q_0^c \circ F_1^c}(y \mid x) := \mathbb{Z}_{Q_0^c}\Big(F^c_{Y_1|X}(y_1 \mid x) \mid x\Big) + \frac{\mathbb{Z}_{F_1^c}(y_1 \mid x)}{f^c_{Y_0|X}\Big(Q^c_{Y_0|X}\Big(F^c_{Y_1|X}(y_1 \mid x) \mid x\Big) \mid x\Big)},$$

and

$$\begin{aligned}\sqrt{n}\Big(\hat{F}_{Y|D,Z,X}\Big(\hat{Q}^c_{Y_0|X}\Big(\hat{F}^c_{Y_1|X}(y_1 \mid x) \mid x\Big)|0,1,x\Big) - F_{Y|D,Z,X}\Big(Q^c_{Y_0|X}\Big(F^c_{Y_1|X}(y_1 \mid x) \mid x\Big) \mid 0,1,x\Big)\Big)&\\ \rightsquigarrow \mathbb{Z}_{F_{01} \circ Q_0^c \circ F_1^c}(y_1 \mid x) \text{ in } \ell^\infty(\mathcal{Y}\mathcal{X}),&\end{aligned}$$

where $\mathbb{Z}_{F_{01} \circ Q_0^c \circ F_1^c}$ is a tight mean-zero Gaussian process defined as

$$\begin{aligned}\mathbb{Z}_{F_{01} \circ Q_0^c \circ F_1^c}(y_1 \mid x) &:= \mathbb{Z}_{F_{01}}\Big(Q^c_{Y_0|X}\Big(F^c_{Y_1|X}(y_1 \mid x) \mid x\Big) \mid x\Big)\\ &\quad + f_{Y|D,Z,X}\Big(Q^c_{Y_0|X}\Big(F^c_{Y_1|X}(y_1 \mid x) \mid x\Big) \mid 0,1,x\Big)\mathbb{Z}_{Q_0^c \circ F_1^c}(y_1 \mid x).\end{aligned}$$

Using similar arguments, obtain

$$\sqrt{n}\Big(\hat{F}_{Y|D,Z,X}\Big(\hat{Q}^c_{Y_1|X}\Big(\hat{F}^c_{Y_0|X}(y_0 \mid x) \mid x\Big) \mid 1,0,x\Big) - F_{Y|D,Z,X}\Big(Q^c_{Y_1|X}\Big(F^c_{Y_0|X}(y_0 \mid x) \mid x\Big) \mid 1,0,x\Big)\Big)$$
$$\rightsquigarrow \mathbb{Z}_{F_{10}\circ Q^c_1 \circ F^c_0}(y_0 \mid x) \text{ in } \ell^\infty(\mathcal{YX}),$$

where

$$\begin{aligned}\mathbb{Z}_{F_{10}\circ Q^c_1 \circ F^c_0}(y \mid x) &:= \mathbb{Z}_{F_{10}}\Big(Q^c_{Y_1|X}\Big(F^c_{Y_0|X}(y_0 \mid x) \mid x\Big)|x\Big) \\ &\quad + f_{Y|D,Z,X}\Big(Q^c_{Y_1|X}\Big(F^c_{Y_0|X}(y_0 \mid x) \mid x\Big)|1,0,x\Big)\mathbb{Z}_{Q^c_1\circ F^c_0}(y_0 \mid x),\end{aligned}$$

and

$$\mathbb{Z}_{Q^c_1\circ F^c_0}(y \mid x) := \mathbb{Z}_{Q^c_1}\Big(F^c_{Y_0|X}(y_0 \mid x) \mid x\Big) + \frac{\mathbb{Z}_{F^c_0}(y_0 \mid x)}{f^c_{Y_1|X}\Big(Q^c_{Y_1|X}\Big(F^c_{Y_0|X}(y_0 \mid x) \mid x\Big) \mid x\Big)}.$$

Finally, consider $\hat{F}_{Y_1|X}(y_1 \mid x)$ and $\hat{F}_{Y_0|X}(y_0 \mid x)$ defined in equations (9) and (10). By the functional delta method,

$$\sqrt{n}\begin{pmatrix}\hat{F}_{Y_1|X}(y_1 \mid x) - F_{Y_1|X}(y_1 \mid x) \\ \hat{F}_{Y_0|X}(y_0 \mid x) - F_{Y_0|X}(y_0 \mid x)\end{pmatrix} \rightsquigarrow \begin{pmatrix}\mathbb{Z}_{F_{Y_1|X}}(y_1 \mid x) \\ \mathbb{Z}_{F_{Y_0|X}}(y_0 \mid x)\end{pmatrix} \text{ in } \ell^\infty(\mathcal{YX})^2,$$

where $\mathbb{Z}_{F_{Y_1|X}}$ and $\mathbb{Z}_{F_{Y_0|X}}$ are tight mean-zero Gaussian processes defined as

$$\begin{aligned}\mathbb{Z}_{F_{Y_1|X}}(y_1 \mid x) &:= \Big(F_{Y|D,Z,X}(y_1 \mid 1,1,x) - F_{Y|D,Z,X}\Big(Q^c_{Y_0|X}\Big(F^c_{Y_1|X}(y_1 \mid x) \mid x\Big) \mid 0,1,x\Big)\Big)\mathbb{Z}_{p_1}(x) \\ &\quad + p(1,x)\mathbb{Z}_{F_{11}}(y_1 \mid x) + (1-p(1,x))\mathbb{Z}_{F_{01}\circ Q^c_0 \circ F^c_1}(y_1 \mid x) \\ \mathbb{Z}_{F_{Y_0|X}}(y_0 \mid x) &:= \Big(F_{Y|D,Z,X}\Big(Q^c_{Y_1|X}\Big(F^c_{Y_0|X}(y_0 \mid x) \mid x\Big) \mid 1,0,x\Big) - F_{Y|D,Z,X}(y_0 \mid 0,0,x)\Big)\mathbb{Z}_{p_0}(x) \\ &\quad + (1-p(0,x))\mathbb{Z}_{F_{00}}(y_0 \mid x) + p(0,x)\mathbb{Z}_{F_{10}\circ Q^c_1 \circ F^c_0}(y_0 \mid x).\end{aligned}$$

This completes the proof of the theorem.

Proof of Theorem 2

Under the assumptions of the theorem, the inverse map is Hadamard differentiable uniformly with respect to an index (Chernozhukov et al., 2010). The results therefore follows from Theorem 1 and the functional delta method.

Proof of Theorem 3

Under Assumption 4, Donsker's theorem implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(f(Y_i,X_i)-\int f dP\right) \rightsquigarrow \mathbb{Z}_X(f),$$

as a stochastic process indexed by $f \in \mathcal{F}$, where $\mathcal{F}$ is a universal Donsker class. The limit process $\mathbb{Z}_X$ is a tight P-Brownian bridge (e.g., Chernozhukov et al., 2013). Therefore, Lemma D.1 in Chernozhukov et al. (2013) (Hadamard differentiability of the counterfactual operator) and the functional delta method imply

$$\sqrt{n}\begin{pmatrix}\hat{F}_{Y_1}(y_1)-F_{Y_1}(y_1)\\ \hat{F}_{Y_0}(y_0)-F_{Y_0}(y_0)\end{pmatrix} \rightsquigarrow \begin{pmatrix}\int_{\mathcal{X}}\mathbb{Z}_{F_{Y_1|X}}(y_1 \mid x)dF_X(x)+\mathbb{Z}_X\left(F_{Y_1|X}(y_1 \mid \cdot)\right)\\ \int_{\mathcal{X}}\mathbb{Z}_{F_{Y_0|X}}(y_0 \mid x)dF_X(x)+\mathbb{Z}_X\left(F_{Y_0|X}(y_0 \mid \cdot)\right)\end{pmatrix} =: \begin{pmatrix}\mathbb{Z}_{F_{Y_1}}(y_1)\\ \mathbb{Z}_{F_{Y_0}}(y_0)\end{pmatrix}.$$

Furthermore, under the assumptions of the theorem, the inverse map is Hadamard differentiable (e.g., Van der Vaart and Wellner, 1996, Lemma 3.9.23). Thus, apply the functional delta method to obtain

$$\sqrt{n}\begin{pmatrix}\hat{Q}_{Y_1}(\tau)-Q_{Y_1}(\tau)\\ \hat{Q}_{Y_0}(\tau)-Q_{Y_0}(\tau)\end{pmatrix} \rightsquigarrow \begin{pmatrix}\mathbb{Z}_{Q_{Y_1}}(\tau)\\ \mathbb{Z}_{Q_{Y_0}}(\tau)\end{pmatrix} \quad \text{in} \quad \ell^{\infty}(\mathcal{T})^2,$$

where

$$\mathbb{Z}_{Q_{Y_1}}(\tau) := -\mathbb{Z}_{F_{Y_1}}(Q_{Y_1}(\tau))/f_{Y_1}(Q_{Y_1}(\tau)), \tag{27}$$

$$\mathbb{Z}_{Q_{Y_0}}(\tau) := -\mathbb{Z}_{F_{Y_0}}(Q_{Y_0}(\tau))/f_{Y_0}(Q_{Y_0}(\tau)). \tag{28}$$

The desired result then follows from another application of the functional delta method.

Proof of Theorem 4

Follows directly from the functional delta method.

Proof of Theorem 5

Under Assumptions 3 and 4.1 – 4.3, it follows from Corollary 5.4 in Chernozhukov et al. (2013) and Step 1 in the proof of Theorem 1 that the exchangeable bootstrap is valid for the conditional distribution functions and conditional probabilities. The result then follows from the established Hadamard differentiability of all maps involved and the functional delta method for the bootstrap (e.g., Van der Vaart and Wellner, 1996, Section 3.9).

References

Abadie, A., Angrist, J., Imbens, G., 2002. Instrumental variable estimates of the effect of subsidized training on the quantile of trainee earnings. Econometrica 70 (1), pp. 91–117.

Chernozhukov, V., Fernandez-Val, I., Galichon, A., 2010. Quantile and probability curves without crossing. Econometrica 78 (3), pp. 1093–1125.

Chernozhukov, V., Fernandez-Val, I., Melly, B., 2013. Inference on counterfactual distributions. Econometrica 81 (6), pp. 2205–2268.

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

De Chaisemartin, C., D'Haultfoeuille, X., 2018a. Fuzzy differences-in-differences. The Review of Economic Studies 85 (2), 999–1028.

De Chaisemartin, C., D'Haultfoeuille, X., 2018b. Supplement to: "fuzzy differences-in-differences".

Kim, J. H., Park, B. G., 2017. Testing rank similarity in the heterogeneous treatment effect model. Working Paper, UNC.

Melly, B., Santangelo, G., 2015. The changes-in-changes model with covariates. Working Paper.

Melly, B., Wüthrich, K., 2017. Local quantile treatment effects. In: Chernozhukov, V., He, X., Koenker, R., Peng, L. (Eds.), Handbook of Quantile Regression. CRC Chapman-Hall, pp. 145–164.

Van der Vaart, A., Wellner, J., 1996. Weak Convergence and Empirical Processes: With Application to Statistics. Springer-Verlag.

Vuong, Q., Xu, H., 2017. Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity. Quantitative Economics 8 (2), 589–610.

Vytlacil, E., 2002. Independence, monotonicity, and latent index models: An equivalence result. Econometrica 70 (1), pp. 331–341.

Yu, P., 2014. Marginal quantile treatment effect and counterfactual analysis. Working Paper, HKU.

Yu, P., 2017. Testing conditional rank similarity with and without covariates. Working Paper, HKU.

Tables

Table 1: Finite sample performance of inference procedure

	Coverage		Power	
	$(1-\alpha) = 0.95$	$(1-\alpha) = 0.90$	$\alpha = 0.05$	$\alpha = 0.10$
$n = 500$	0.958	0.918	0.930	0.984
$n = 1000$	0.950	0.904	0.998	1.000

Notes: Simulation based on 500 replications with 100 empirical bootstrap draws each. Coverage: empirical coverage of uniform $(1-\alpha)$-confidence bands. Power: estimated power of the α-level Kolmogorov–Smirnov-type test underlying the uniform confidence bands

Table 2: Rothe and Wied (2013) specification test

	NFA		TW	
	KS	CvM	KS	CvM
$F_{Y\|D,Z,X}(y \mid 1,1,x)$	0.746	0.936	0.298	0.996
$F_{Y\|D,Z,X}(y \mid 0,1,x)$	0.324	0.976	0.986	0.998
$F_{Y\|D,Z,X}(y \mid 0,0,x)$	0.374	0.718	0.612	0.700

Notes: NFA: net financial assets, TW: total wealth, KS: Kolmogorov–Smirnov, CvM: Cramér–von–Mises

Table 3: Comparison computation times

	# of grid points		
	50	100	200
	$\dim(X) = 12$		
Plug-in	12.73	19.63	33.48
IQR	127.83	509.99	2042.63
	$\dim(X) = 48$		
Plug-in	26.04	45.68	85.63
IQR	889.48	3586.72	14375.67

Notes: Computation time in seconds. Sample size $n = 9913$. The number of grid points refers to the grids for $\mathcal{Y}$ and $\mathcal{T}$ for the plug-in estimator and $\mathcal{Y}$, $\mathcal{T}$, and grid search for IQR. Computations were carried out on a standard desktop computer with a 3.2 GHz Intel Core i5 processor and 8GB RAM.

Figures

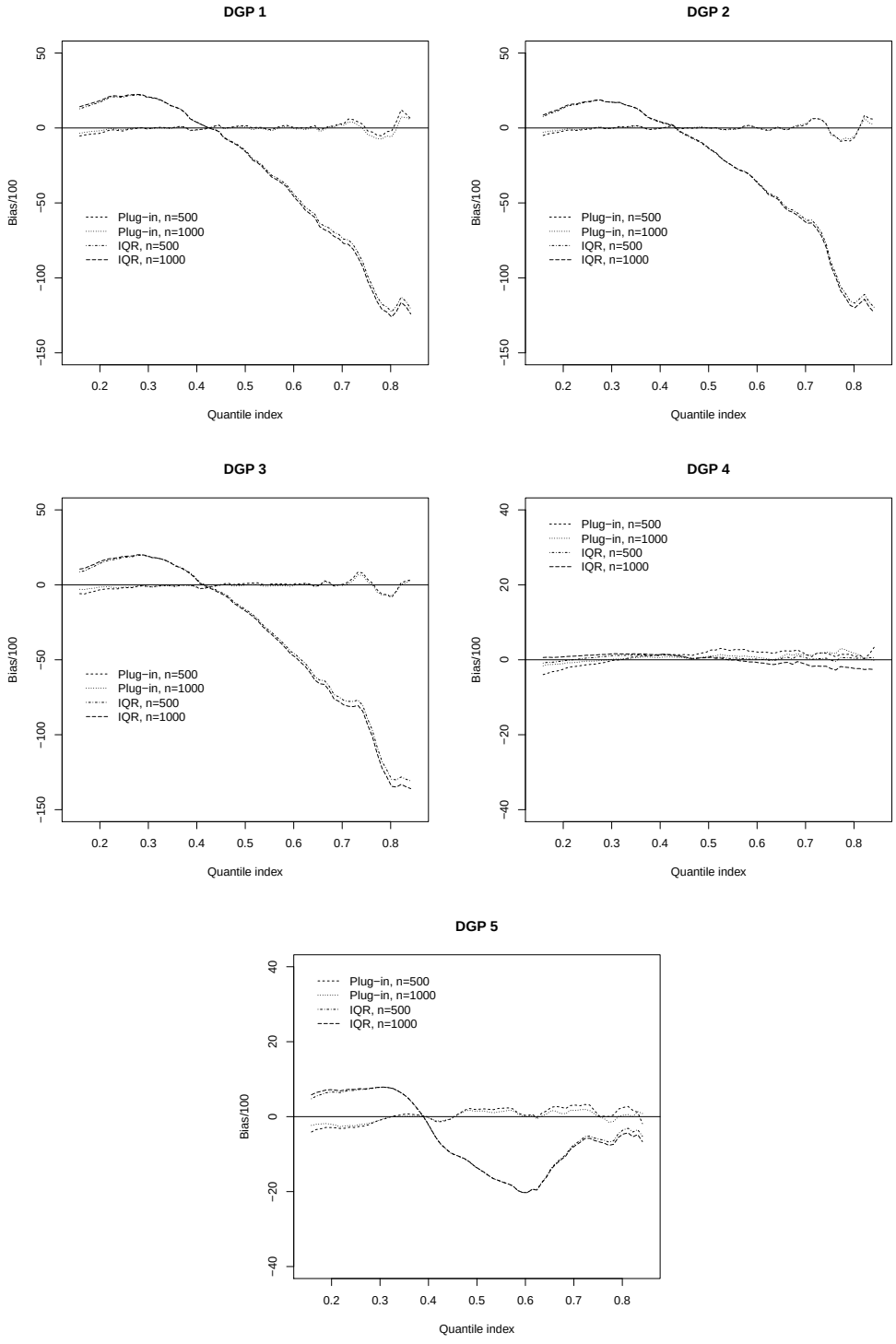

Figure 1: Simulation based on 500 repetitions. DGP1 – DGP5 are described in the main text.

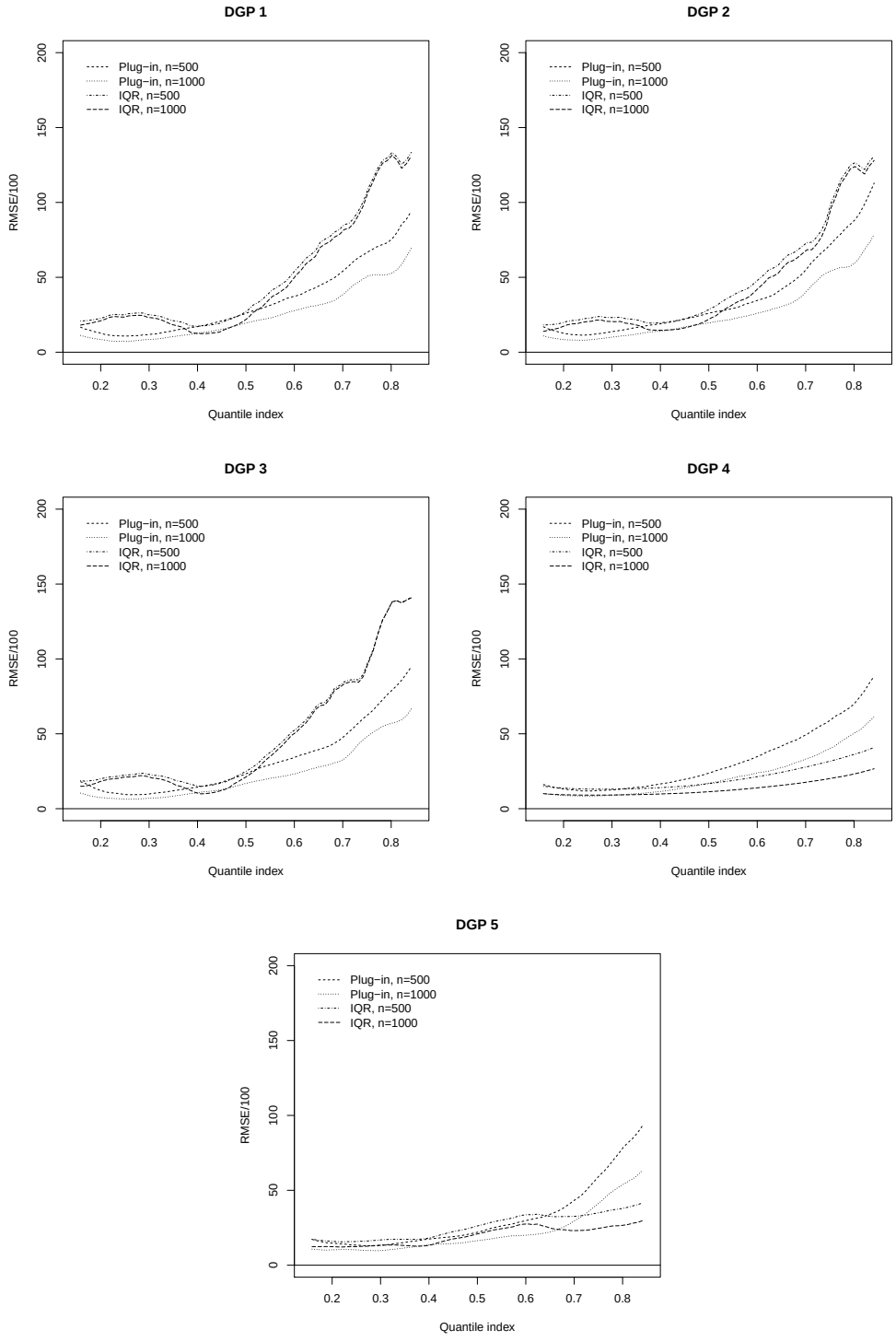

Figure 2: Simulation based on 500 repetitions. DGP1 – DGP5 are described in the main text.

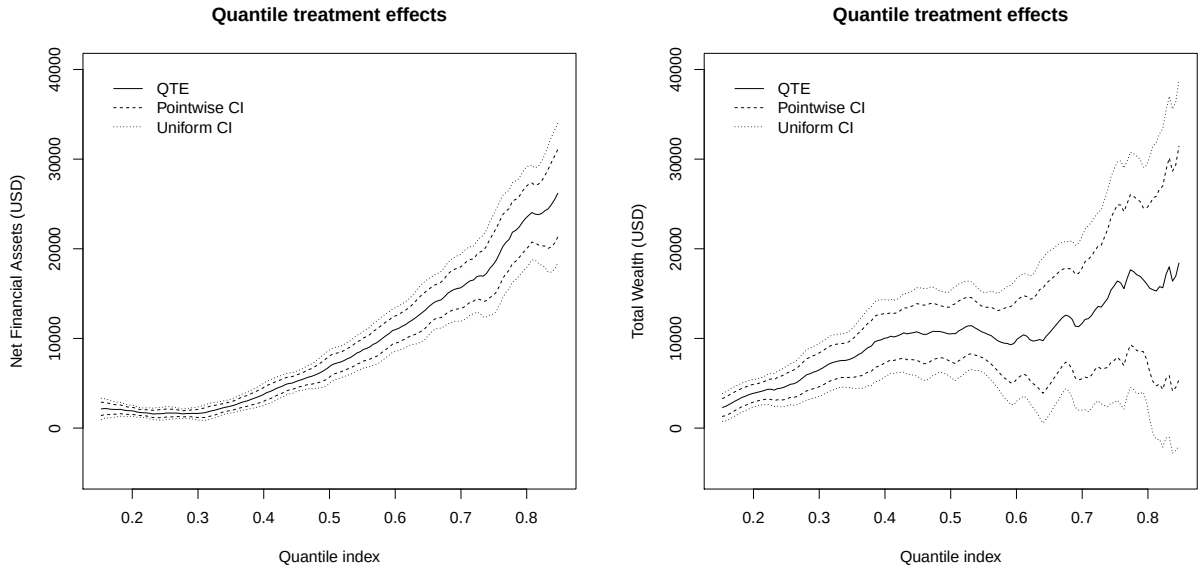

Figure 3: Sample size $n = 9,913$. Uniform and pointwise 95%-confidence intervals are obtained based on the empirical bootstrap with 200 replications.

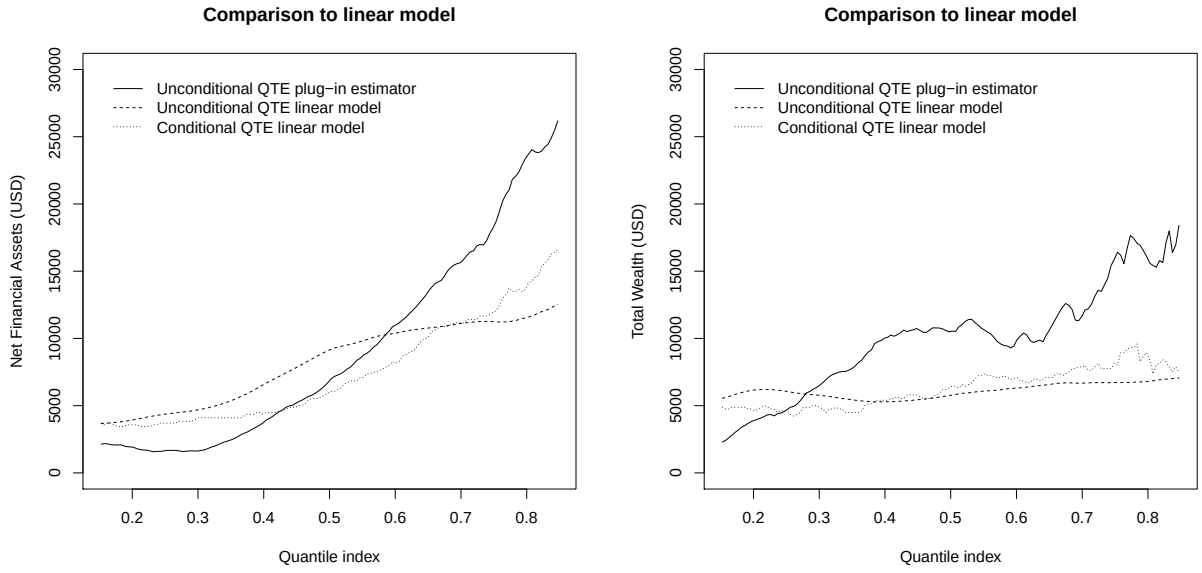

Figure 4: Sample size $n = 9,913$.