UNIVERSITÄT
BERN

Faculty of Business, Economics
and Social Sciences

Department of Economics

Semiparametric estimation of quantile treatment effects with endogeneity

Kaspar Wüthrich

15-09

July 2015

DISCUSSION PAPERS

Schanzeneckstrasse 1
Postfach 8573
CH-3001 Bern, Switzerland
http://www.vwi.unibe.ch

Semiparametric estimation of quantile treatment effects with endogeneity*

Kaspar Wüthrich†

July 17, 2015

Abstract

This paper studies estimation of conditional and unconditional quantile treatment effects based on the instrumental variable quantile regression (IVQR) model (Chernozhukov and Hansen, 2004, 2005, 2006). I introduce a class of semiparametric plug-in estimators based on closed form solutions derived from the IVQR moment conditions. These estimators do not rely on separability of the structural quantile function, while retaining computational tractability and root-n-consistency. Functional central limit theorems and bootstrap validity results for the estimators of the quantile treatment effects and other functionals are provided. I apply my method to reanalyze the effect of 401(k) plans on individual savings behavior.

JEL Classification: C14, C21, C26

Keywords: instrumental variables, quantile treatment effects, distribution regression, functional central limit theorem, Hadamard differentiability, exchangeable bootstrap

---

*Acknowledgments: I would like to thank Isaiah Andrews, Daniel Burkhard, Andreas Bachmann, Victor Chernozhukov, Lutz Dümbgen, Iván Fernández-Val, Blaise Melly, Anna Mikusheva, and seminar participants at MIT and the University of Bern for very helpful comments. Financial support by the Swiss National Science Foundation (Doc.Mobility Project P1BEP1_155467) is gratefully acknowledged. All errors are my own. The usual disclaimer applies.

†University of Bern, Department of Economics, Schanzeneckstrasse 1, CH-3001 Bern, Switzerland, phone: +41 79 830 7302, email: kaspar.wuethrich@vwi.unibe.ch.

1 Introduction

This paper studies estimation of quantile treatment effects (QTE) with endogenous policy variables. As with linear models, endogeneity renders standard quantile regression methods inconsistent for estimating QTE. A common approach to deal with this problem is to use instrumental variable (IV) methods.

The goal and main contribution of this paper is to develop a regression-based semiparametric estimation approach based on the instrumental variable quantile regression (IVQR) model (Chernozhukov and Hansen, 2004, 2005, 2006). The principal feature of the IVQR model is the rank similarity assumption, a condition that restricts the evolution of individual ranks across treatment states. Rank similarity implies a conditional moment restriction that can be used to construct estimators for QTE.

However, estimation is complicated by the nonsmoothness and nonconvexity of the resulting generalized method of moments (GMM) objective function that occurs even for linear-in-parameters quantile models (Chernozhukov and Hansen, 2013). Different approaches have been proposed to overcome this problem: estimation procedures for linear-in-parameters models and nonparametric minimum-distance-type estimators. While linear-in-parameters models typically impose strong additivity assumptions on the structural quantile function and thereby substantially restrict treatment effect heterogeneity, nonparametric approaches suffer from the curse of dimensionality and require choosing tuning parameters.

The semiparametric estimation approach proposed in this paper does neither impose separability restrictions, require the choice of tuning parameters, nor suffer from the curse of dimensionality. Instead, it relies on flexible parametric models for the observed conditional distributions and conditional probabilities. The key idea is to construct analytic plug-in estimators based on closed form solutions for the IVQR estimands of the potential outcome cumulative distribution functions (cdfs), which are available whenever the policy variable is binary.[1] These closed form solutions are compositions of observable conditional distributions and conditional probabilities. I estimate the conditional distributions using distribution regression (DR) and the conditional probabilities using binary choice models.[2] I then apply the closed form solutions to obtain plug-in estimators of the conditional potential outcome distributions and the conditional QTE. Because analytic closed form solutions are only available for binary treatments, this plug-in estimation approach is inherently limited to binary treatments.

Because the semiparametric plug-in estimators do not rely on additive separability of the structural quantile function, the conditional QTE are generically nonlinear functions of the covariates. These high-dimensional objects are typically hard to summarize and convey. Conse-

[1]These analytic closed from solutions have been derived in a companion paper (Wüthrich, 2014).

[2]DR was first proposed by Foresi and Peracchi (1995). Uniform inference results have been derived by Chernozhukov et al. (2013).

quently, one is often more interested in unconditional QTE, which are obtained by integrating the estimators of the conditional potential outcome distributions with respect to the empirical distribution of the covariates.[3]

This paper shows that under standard regularity conditions, the semiparametric estimators of the QTE and other related functionals are uniformly consistent and satisfy functional central limit theorems. Moreover, I prove validity of the exchangeable bootstrap for estimating the limiting laws. These results allow me to construct uniform confidence bands and to test functional hypothesis such as no-effects, positive effects, constant effects, or stochastic dominance. I also suggest simple overidentification specification tests for the IVQR model based on the Kolmogorov-Smirnov distance between the QTE estimates obtained from using different instruments.

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

The method is illustrated by reanalyzing the distributional effect of 401(k) plans on individual savings behavior using the data from the 1991 Survey of Income and Program Participation (SIPP) studied in Chernozhukov and Hansen (2004) and Belloni et al. (2014). My estimates suggest that 401(k) participation has a moderate effect on individual assets at lower quantiles while having a substantive impact at high quantiles. A comparison of these results with the estimates from a linear-in-parameters model shows substantive differences between both approaches, highlighting the importance of analyzing nonseparable models.

1.1 Related literature

This paper contributes to the extensive literature on estimation in the IVQR model. Estimation and inference in linear conditional quantile models have been analyzed by Chernozhukov and Hong (2003), Chernozhukov and Hansen (2006), Chernozhukov et al. (2007a), Chernozhukov and Hansen (2008), Chernozhukov et al. (2009), and Kaplan and Sun (2014). Nonparametric estimation has been studied by Chernozhukov et al. (2007b), Horowitz and Lee (2007), Chen and Pouzo (2009), Chen and Pouzo (2012), Gagliardini and Scaillet (2012), Su and Hosino (2013), Kaplan and Sun (2014), and Belloni et al. (2014). Chernozhukov and Hansen (2013) provide a recent survey on the IVQR model including further references.

The semiparametric estimation strategy is also related to several estimation approaches that rely on estimating conditional distributions using distribution or quantile regression as ingredients for deriving plug-in estimators. Chernozhukov et al. (2013) have analyzed counterfactual

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

distributions, Yu (2014) has proposed a semiparametric estimation approach for marginal QTE, and Melly and Santangelo (2015) have analyzed nonlinear difference-in-differences models.

More broadly, this paper contributes to the literature on identification and estimation of QTE and structural quantile functions with endogeneity. Abadie et al. (2002), Frandsen et al. (2012), and Frölich and Melly (2013) have studied semi- and nonparametric estimation of local QTE with binary treatments. Triangular models with continuous treatments have been analyzed by Chesher (2003), Ma and Koenker (2006), Lee (2007), Imbens and Newey (2009), and Torgovitsky (2012) among others.

1.2 Outline

The remainder of the paper is organized as follows. In Section 2, I present the IVQR model and the analytic closed form solutions. Section 3 describes the estimators and compares them to existing methods. In Section 4, I give the asymptotic results and their proofs. Section 5 extends the analysis to nonbinary instruments and suggests a simple overidentification specification test for the IVQR model. In Section 6, I apply my method to estimate the distributional effects of 401(k) plans on accumulated assets. Section 7 concludes. The appendix contains additional proofs and two simple numerical examples that illustrate the difference between linear separable and nonseparable quantile models.

2 The IVQR model

I consider a setup with a continuous outcome variable Y, a binary treatment D, a binary instrument Z, and a vector of covariates X. While the estimation approach can be extended to nonbinary instruments as discussed in Section 5, it is not possible to extend it to nonbinary treatments because there are no analytic closed form solutions in this case. Let the symbols $\mathcal{Y}$, $\mathcal{D}$, $\mathcal{Z}$, and $\mathcal{X}$ denote the supports of these random variables and let $\mathcal{T} \subset (0,1)$ be a compact interval of quantile indices. Moreover, 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 potential outcomes. Having conditioned on covariates $X = x$, by the Skorohod representation of random variables, potential outcomes can be represented as

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

where $Q_{Y_d|X}(\tau|x)$ is the τ-quantile of Y_d given $X = x$. 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, 2013):

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, $Y_d = Q_{Y_d|X}(U_d|x)$, where $Q_{Y_d|X}(\tau|x)$ 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 $\rho(\cdot)$ 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 $Y := Q_{Y_D|X}(U_D|X)$, D, X, and Z.*

Assumption 1.1 restates the Skorohod representation of random variables and imposes strict monotonicity on the structural quantile function, thus ruling out discrete outcome variables. Assumption 1.2 imposes independence between the potential outcomes and the instrument. Assumption 1.3 states a general selection equation in which the unobservable random vector V leads to different treatment choices between observationally identical individuals. Assumption 1.4 represents the arguably most important condition of the IVQR model. It requires that individual ranks are constant across potential outcome distributions up to random slippages away from a common level U. Finally, Assumption 1.5 summarizes the observable variables. The interested reader is referred to Chernozhukov and Hansen (2005, 2013) 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):

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

Estimation based on (1) is challenging because the sample analogue of the GMM objective function is nonsmooth and generically nonconvex. References to different approaches to overcome these challenges are given in the introduction.

Here I propose a computationally tractable estimation approach that exploits closed form solutions of the potential outcome distributions $F_{Y_1|X}(y|x)$ and $F_{Y_0|X}(y|x)$ to construct semi-parametric plug-in estimators. These analytic closed from solutions have been derived in a companion paper (Wüthrich, 2014).

Lemma 1. *Suppose that Assumption 1 holds and that*

$$F^c_{Y_1|X}(y|x) := \frac{p(1,x)F_{Y|D,Z,X}(y|1,1,x) - p(0,x)F_{Y|D,Z,X}(y|1,0,x)}{p(1,x)-p(0,x)},$$
$$F^c_{Y_0|X}(y|x) := \frac{(1-p(0,x))F_{Y|D,Z,X}(y|0,0,x) - (1-p(1,x))F_{Y|D,Z,X}(y|0,1,x)}{p(1,x)-p(0,x)}, \tag{2}$$

are strictly increasing and continuously differentiable cdfs, then

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

where $p(z,x) := P(D=1|Z=z, X=x)$.

Proof. See appendix B.2. □

Lemma 1 and its proof are closely related to Lemma 1 in the companion paper (Wüthrich, 2014). Based on the closed form solutions in Lemma 1, the conditional QTE are identified as

$$\delta(\tau|x) = F^{\leftarrow}_{Y_1|X}(\tau|x) - F^{\leftarrow}_{Y_0|X}(\tau|x),$$

where $F^{\leftarrow}_{Y_d|X}(y|x)$ denotes the left-inverse of $F_{Y_d|X}(y|x)$. Conditional QTE are useful for analyzing effect heterogeneity by observable characteristics and across different quantiles. Because conditional QTE are generically high-dimensional objects, one is often more interested in unconditional QTE, which are informative about the effect of the treatment on the marginal outcome distribution:

$$\delta(\tau) = F^{\leftarrow}_{Y_1}(\tau) - F^{\leftarrow}_{Y_0}(\tau),$$

where unconditional potential outcome distributions are obtained by integrating the conditional potential outcome distributions with respect to the marginal distribution of the covariates, $F_X(x)$:

$$F_{Y_1}(y) = \int_{\mathcal{X}} F_{Y_1|X}(y|x)dF_X(x),$$
$$F_{Y_0}(y) = \int_{\mathcal{X}} F_{Y_0|X}(y|x)dF_X(x).$$

Note that all the previous estimands are functions of $F_{Y|D,Z,X}(y|d,z,x)$, $p(z,x)$, and $F_X(x)$ only. This suggests a plug-in estimation approach as detailed in Section 3.

3 Estimators

The conditional distributions are estimated using DR. For all $(d,z) \in \mathcal{DZ}$, let

$$\hat{F}_{Y|D,Z,X}(y|d,z,x) = \Lambda\left(x'\hat{\beta}_{d,z}(y)\right) \quad \text{for all} \quad y \in \mathcal{Y},$$

with

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

where $\Lambda(\cdot)$ denotes the logit or probit link. In finite samples, the estimated conditional distributions do not need to be monotone. To overcome this problem, I suggest applying the rearrangement procedure proposed by Chernozhukov et al. (2010). Because these rearrangements do not affect the asymptotic properties of the estimators, I keep them implicit throughout the paper. The conditional probabilities are estimated using binary choice models:

$$\hat{p}(z,x) = \Lambda\left(x'\hat{\gamma}_z\right),$$

where

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

Note that the specifications of the conditional cdfs and conditional probabilities are very flexible in the sense that for a given $\Lambda(\cdot)$, $F_{Y|D,Z,X}(y|d,z,x)$ and $p(z,x)$ can be approximated arbitrarily well by using a rich enough dictionary of transformations of the original covariates (Chernozhukov et al., 2013).

Remark 1. *Instead of DR, the conditional distributions can be estimated using quantile regression (QR). In the linear QR model it is assumed that for all* $(d,z) \in \mathcal{DZ}$*,*

$$\hat{Q}_{Y|D,Z,X}(y|d,z,x) = x'\hat{\beta}_{d,z}(\tau) \quad \textit{for all} \quad \tau \in \mathcal{T},$$

where

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

The conditional distribution is then estimated as

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

where $\varepsilon > 0$ is a trimming constant to avoid estimation of tail quantiles. Based on the results in Chernozhukov et al. (2013), it is straightforward to extend the asymptotic theory to QR. The interested reader is referred to Koenker et al. (2013), Chernozhukov et al. (2013), or Rothe and Wied (2013) for further discussions as well as specification tests of distribution and quantile regression.

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

$$\begin{aligned}\hat{F}_{Y_1|X}(y|x) =& \hat{p}(1,x)\hat{F}_{Y|D,Z,X}(y|1,1,x)\\ &+ (1-\hat{p}(1,x))\,\hat{F}_{Y|D,Z,X}\left(\hat{Q}^c_{Y_0|X}\left(\hat{F}^c_{Y_1|X}(y|x)|x\right)|0,1,x\right),\\ \hat{F}_{Y_0|X}(y|x) =& (1-\hat{p}(0,x))\,\hat{F}_{Y|D,Z,X}(y|0,0,x)\\ &+ \hat{p}(0,x)\hat{F}_{Y|D,Z,X}\left(\hat{Q}^c_{Y_1|X}\left(\hat{F}^c_{Y_0|X}(y|x)|x\right)|1,0,x\right),\end{aligned}$$

where

$$\begin{aligned}\hat{F}^c_{Y_1|X}(y|x) =& \frac{\hat{p}(1,x)\hat{F}_{Y|D,Z,X}(y|1,1,x) - \hat{p}(0,x)\hat{F}_{Y|D,Z,X}(y|1,0,x)}{\hat{p}(1,x)-\hat{p}(0,x)},\\ \hat{F}^c_{Y_0|X}(y|x) =& \frac{(1-\hat{p}(0,x))\hat{F}_{Y|D,Z,X}(y|0,0,x) - (1-\hat{p}(1,x))\hat{F}_{Y|D,Z,X}(y|0,1,x)}{\hat{p}(1,x)-\hat{p}(0,x)}.\end{aligned}$$

For the conditional and unconditional QTE, I also obtain estimators via the plug-in rule

$$\hat{\delta}(\tau|x) = \hat{F}^{\leftarrow}_{Y_1|X}(y|x) - \hat{F}^{\leftarrow}_{Y_0|X}(y|x) \quad \text{and} \quad \hat{\delta}(\tau) = \hat{F}^{\leftarrow}_{Y_1}(y) - \hat{F}^{\leftarrow}_{Y_0}(y),$$

where the unconditional distributions, $\hat{F}_{Y_1}(y)$ and $\hat{F}_{Y_0}(y)$, are estimated by integrating the estimators of the conditional distributions, $\hat{F}_{Y_1|X}(y|x)$ and $\hat{F}_{Y_0|X}(y|x)$, with respect to the empirical distribution of the covariates $\hat{F}_X(x)$.

$$\begin{aligned}\hat{F}_{Y_1}(y) &= \int_{\mathcal{X}} \hat{F}_{Y_1|X}(y|x)d\hat{F}_X(x),\\ \hat{F}_{Y_0}(y) &= \int_{\mathcal{X}} \hat{F}_{Y_0|X}(y|x)d\hat{F}_X(x),\end{aligned}$$

where

$$\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} 1\left\{X_i \leq x\right\}.$$

Similarly, one can construct plug-in estimators for other functionals of interest.

3.1 Comparison to other estimation approaches

The alternative estimation approaches can be classified broadly into two categories: approaches based on the linear-in-parameters models such as

$$Q_{Y_D|X}(\tau|x) = D\delta(\tau) + X'\beta(\tau), \tag{4}$$

and nonparametric minimum-distance type estimators.

The most popular approach for estimating IVQR models is the inverse quantile regression algorithm (IQR) developed by Chernozhukov and Hansen (2006). This approach exploits the linear-in-parameters structure to overcome the problems associated with maximizing the function by combining robust grid search methods with standard quantile regression techniques. The key feature and main limitation of IQR is that the dimensionality of the grid search equals the dimensionality of the endogenous variables. Thus, IQR is computationally tractable only if the dimensionality of the endogenous variables is small, typically one or two. However, this feature crucially limits the appeal of IQR for analyzing models with treatment effect heterogeneity across observable covariates. To see this, note that models such as (4) impose that the conditional QTE $\delta(\tau)$ is constant across $X = x$. Imposing more flexible models including interactions of D and (transformations of) X increases the dimensionality of the endogenous variables and consequently the dimensionality of the grid search, rendering IQR computationally prohibitive. In contrast, the semiparametric estimation approach remains computationally tractable without imposing separability restrictions on the structural quantile function. Instead, it relies on flexible parametric models for the conditional cdfs and conditional probabilities. Moreover, if we use fully saturated specifications for the conditional distributions and the conditional probabilities, the parametric restrictions are without loss of generality and the semiparametric estimation approach presents a computationally convenient procedure to nonparametric estimation of QTE. In contrast, the IQR algorithm is computationally prohibitive in this case, even in situations where X is rather low dimensional. Moreover, to the best of my knowledge, there are no results about the interpretation of linear IVQR models with misspecification.[4] Thus, it is unclear what linear IVQR models estimates when the functional form is misspecified. There are at least two alternatives to IQR, namely quasi-Bayesian estimators (Chernozhukov and Hong, 2003) and estimators based on smoothed estimation equations (Kaplan and Sun, 2014). While both approaches remain computationally tractable with multiple endogenous variables, the former requires a careful tuning in applications and the latter relies on the choice of a smoothing-bandwidth (Chernozhukov and Hansen, 2013).

The nonparametric estimation approaches cited in the introduction do not rely on separability assumptions nor impose parametric models for the conditional cdfs and probabilities.

[4] This is in contrast to the exogenous case, see Angrist et al. (2006).

However, these methods suffer from the curse of dimensionality[5], necessitate the choice tuning parameters, and are typically computationally demanding. In addition, nonparametric estimation is complicated by nonlinear ill-posed inverse problems. In contrast, the semiparametric estimation approach constitutes a computationally feasible and easily implementable approach that does not rely on tuning parameters nor suffer from the curse of dimensionality.

Finally, it is noteworthy that many of the alternative estimation approaches such as the IQR algorithm accommodate multivalued, continuous, or even multiple treatments. This is in sharp contrast to the semiparametric estimation approach that is limited to binary scalar treatments.

4 Asymptotic theory and inference

4.1 Limiting distribution

Assumption 2 gives conditions under which the plug-in estimators are uniformly consistent and asymptotically Gaussian.

Assumption 2 (Regularity conditions)**.**

1. *$\{Y_i, D_i, Z_i, X_i\}$ are i.i.d.*

2. *$p(z,x) = \Lambda(x'\gamma_z)$ and $F_{Y|D,Z,X}(y|d,z,x) = \Lambda(x'\beta_{dz}(y))$ for all $(y,d,z,x) \in \mathcal{YDZX}$, where $\Lambda(\cdot)$ is either the probit oder logit link function.*

3. *The region of interest $\mathcal{Y}$ is a compact interval in $\mathbb{R}$ and the conditional density $f_{Y|D,Z,X}(y|d,z,x)$ exists, is uniformly bounded, and uniformly continuous. Moreover, $\mathcal{X}$ is a compact subset of $\mathbb{R}^{\dim(X)}$.*

4. *$\mathbb{E}||X||^2 < \infty$ and the minimum eigenvalues of*

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

and

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

are bounded away from zero uniformly over $y \in \mathcal{Y}$, where $\lambda(\cdot)$ is the derivative of $\Lambda(\cdot)$.

5. *$P(Z=1|X=x)$ and $|p(1,x) - p(0,x)|$ are bounded away from zero and one for all $x \in \mathcal{X}$, and $F^c_{Y_d|X}(y|x)$ admits a positive, uniformly bounded, and uniformly continuous density on an interval containing an ϵ-enlargement of the set $\{Q^c_{Y_d|X}(\tau|x) : (\tau,x) \in \mathcal{TX}\}$.*

[5] This particularly applies to conditional QTE. For unconditional QTE, one could probably obtain $\sqrt{n}$-consistent estimators as in Frölich and Melly (2013).

Assumptions 2.1 – 2.4 are standard regularity conditions (e.g., Chernozhukov et al., 2013) that ensure that functional central limit theorems and bootstrap validity results apply for the conditional distributions and conditional probabilities. Assumption 2.5 implies point identification based on the moment condition (1) and Hadamard differentiability of the closed form solutions (3).[6]

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

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

$$\sqrt{n}\begin{pmatrix}\hat{F}_{Y_1|X}(y|x) - F_{Y_1|X}(y|x) \\ \hat{F}_{Y_0|X}(y|x) - F_{Y_0|X}(y|x)\end{pmatrix} \rightsquigarrow \begin{pmatrix}\mathbb{Z}_{F_1}(y|x) \\ \mathbb{Z}_{F_0}(y|x)\end{pmatrix}$$

as stochastic processes indexed by $(y, x) \in \mathcal{YX}$ in the metric space $\ell^\infty(\mathcal{YX})^2$, where $\mathbb{Z}_{F_1}(y|x)$ and $\mathbb{Z}_{F_0}(y|x)$ are tight zero-mean Gaussian processes defined in Appendix B.

Proof. See Appendix B. □

The main difficulty in proving Lemma 2 is to show that the closed form solutions (3) are Hadamard differentiable (uniformly with respect to an index). The result then follows from the functional delta method and existing functional central limit theorems for the conditional cdfs and the conditional probabilities.

The next theorem presents the limiting distribution for the conditional quantile functions and the conditional QTE.

Theorem 1. *Suppose that Assumptions 1 and 2 hold and that $F_{Y_d|X}(y|x)$ admits a positive continuous density on an interval containing an ϵ-enlargement of the set $\{Q_{Y_d|X}(\tau|x) : (\tau, x) \in \mathcal{TX}\}$ for $d \in \mathcal{D}$. Then*

$$\sqrt{n}\left(\hat{\delta}(\tau|x) - \delta(\tau|x)\right) \rightsquigarrow \mathbb{Z}_\delta(\tau|x),$$

as a stochastic process indexed by $(\tau, x) \in \mathcal{TX}$ in the metric space $\ell^\infty(\mathcal{TX})$, where $\mathbb{Z}_\delta(\tau|x)$ is a mean-zero tight Gaussian process defined as

$$\begin{aligned}\mathbb{Z}_\delta(\tau|x) :=& \mathbb{Z}_{F_{Y_0|X}}\left(Q_{Y_0|X}(\tau|x)|x\right) / f_{Y_0|X}\left(Q_{Y_0|X}(\tau|x)|x\right) \\ &- \mathbb{Z}_{F_{Y_1|X}}\left(Q_{Y_1|X}(\tau|x)|x\right) / f_{Y_1|X}\left(Q_{Y_1|X}(\tau|x)|x\right)\end{aligned}$$

[6] This condition implies continuity and full rank of the Jacobian of the moment condition (1), which in turn implies point identification of the IVQR estimands (Chernozhukov and Hansen, 2005).

Proof. The proof follows from Lemma 2, Hadamard differentiability of the inverse map uniformly with respect to an index (Chernozhukov et al., 2010), and the functional delta method. □

Functional central limit theorems for the unconditional quantile functions and QTE can be 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 2. *Suppose that Assumptions 1 and 2 hold and that $F_{Y_d}(y)$ admits a positive continuous density on an interval containing an ϵ-enlargement of the set $\{Q_{Y_d}(\tau) : \tau \in \mathcal{T}\}$ for $d \in \mathcal{D}$. Then*

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

as a stochastic process indexed by $\tau \in \mathcal{T}$ in the metric space $\ell^\infty(\mathcal{T})$ 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 $\mathbb{Z}_{Q_{Y_1}}(\tau)$ and $\mathbb{Z}_{Q_{Y_0}}(\tau)$ are defined in (5) and (6).

Proof. By the Donsker theorem, the empirical distribution of the covariates

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

as a stochastic process indexed by $f \in \mathcal{F}$, where $\mathcal{F}$ is a universal Donsker class. The limit process $\mathbb{Z}_x(f(y, x))$ is a tight P-Brownian bridge (Chernozhukov et al., 2013). This convergence is jointly with the conditional potential outcome cdf process in Lemma 2.

By Lemma 2, Hadamard differentiability of the counterfactual operator (Chernozhukov et al., 2013), and the functional delta method, obtain

$$\sqrt{n}\begin{pmatrix} \hat{F}_{Y_1}(y) - F_{Y_1}(y) \\ \hat{F}_{Y_0}(y) - F_{Y_0}(y) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \int_{\mathcal{X}} \mathbb{Z}_{F_1}(y|x) dF_X(x) + \mathbb{Z}_x\left(F_{Y_1|X}(y|\cdot)\right) \\ \int_{\mathcal{X}} \mathbb{Z}_{F_0}(y|x) dF_X(x) + \mathbb{Z}_x\left(F_{Y_0|X}(y|\cdot)\right) \end{pmatrix} := \begin{pmatrix} \mathbb{Z}_{F_{Y_1}}(y) \\ \mathbb{Z}_{F_{Y_0}}(y) \end{pmatrix}.$$

By Hadamard differentiability of the inverse map (e.g., Van der Vaart and Wellner, 1996, Lemma 3.9.20) and the functional delta method,

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

where

$$\mathbb{Z}_{Q_{Y_1}}(\tau) := -\mathbb{Z}_{F_{Y_1}}\left(Q_{Y_1}(\tau)\right) / f_{Y_1}\left(Q_{Y_1}(\tau)\right), \tag{5}$$

$$\mathbb{Z}_{Q_{Y_0}}(\tau) := -\mathbb{Z}_{F_{Y_0}}\left(Q_{Y_0}(\tau)\right) / f_{Y_0}\left(Q_{Y_0}(\tau)\right). \tag{6}$$

The result then follows from the functional delta method. $\square$

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

Theorem 3. *Suppose that Assumptions 1 and 2 hold and that the map $\varphi\left(F_{Y_1|X}, F_{Y_0|X}\right)(w)$ (indexed by w) is Hadamard differentiable with derivative maps $\varphi_{F_{Y_1|X}}(\cdot)$ and $\varphi_{F_{Y_0|X}}(\cdot)$. Then*

$$\sqrt{n}\left(\varphi\left(\hat{F}_{Y_1|X}, \hat{F}_{Y_0|X}\right)(w) - \varphi\left(F_{Y_1|X}, F_{Y_0|X}\right)(w)\right) \rightsquigarrow \varphi_{F_{Y_1|X}}\left(\mathbb{Z}_{F_1}\right)(w) + \varphi_{F_{Y_0|X}}\left(\mathbb{Z}_{F_0}\right)(w)$$

as a stochastic process indexed by $w \in \mathcal{W}$ in $\ell^\infty(\mathcal{W})$.

Proof. Follows directly from the functional delta method. $\square$

Remark 2. *The functional central limit theorems imply that the standard pointwise estimators (e.g., the QTE estimator at a single quantile) converge to normal random variables. Moreover, any finite collection of pointwise estimators (e.g., the QTE estimators at two different quantiles) converges jointly to multivariate normal random variables. The variance-covariance matrices are given by the above expressions.*

The above characterizations of the limit processes can be used to perform inference using standard analytical methods. Because all asymptotic variances contain terms that are difficult to estimate (e.g., conditional densities), I recommend using the bootstrap, whose validity is established in the next section.

4.2 Inference

Here I prove 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 $(w_1, ..., w_n)$ be a vector of nonnegative random weights that are independent of the data and satisfy the following assumption.[7]

Assumption 3. *For each n, let $(w_1, ..., w_n)$ be an exchangeable[8], nonnegative random vector, which is independent of the data, such that for some $\varepsilon > 0$,*

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

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

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

The exchangeable bootstrap uses $(w_1, ..., w_n)$ 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 $\hat{F}^*_{Y|D,Z,X}(y|d,z,x) = \Lambda\left(x'\hat{\beta}^*_{d,z}(y)\right)$ and $\hat{p}^*(z,x) = \Lambda\left(x'\hat{\gamma}_z{}^*\right)$ respectively, where

$$\hat{\beta}^*_{d,z}(y) = \arg\max_b \sum_{i=1}^n w_i 1\{D_i = d, Z_i = z\}\left[1\{D_i = 1\}\ln\left[\Lambda\left(X_i'b\right)\right] + 1\{D_i = 0\}\ln\left[1 - \Lambda\left(X_i'b\right)\right]\right],$$

and

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

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

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 $(w_1, ..., w_n)$ is a multinomial vector with parameter n and probabilities $(1/n, ..., 1/n)$. The weighted bootstrap corresponds to the case where $(w_1, ..., w_n)$ 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 $(w_1, ..., w_n)$ be equal to $\sqrt{n/m}$ times multinomial vectors with parameter m and probabilities $(1/n, ..., 1/n)$. Finally, subsampling corresponds to letting $(w_1, ..., w_n)$ 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 a random.

The next theorem formally establishes validity of the exchangeable bootstrap.

Theorem 4. *Suppose that Assumptions 1 – 3 hold. Then the exchangeable bootstrap consistently estimates the limit laws for the processes in Lemma 2 and Theorems 1 – 3.*

Proof. By Lemma 2 and Corollary 5.4 in Chernozhukov et al. (2013), the exchangeable bootstrap is valid for the conditional distribution functions and conditional probabilities. The result then follows from 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). □

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 hypothesis (Chernozhukov et al., 2013). For example, one can construct asymptotic simultaneous $(1 - \alpha)$%-confidence bands for the whole quantile treatment effect process $\hat{\delta}(\tau)$:[9]

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

such that

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

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

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

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 Nonbinary instruments and a simple specification test

Here I briefly discuss how to incorporate nonbinary instruments and present a simple specification test for the IVQR model.

If the instrument is multivalued or continuous, it can be dichotomized such that Assumption 2 holds. Estimation can then proceed based on the dichotomized 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 previous sections, the choice of the dichotomization does not matter for consistency of the estimators. However, efficiency could be improved upon by developing plug-in estimators based on closed from solutions for general instruments (Wüthrich, 2014, Section 4 and Lemma 2) or overidentified GMM objective functions. Such extensions are beyond the scope of this paper and left for future research.

With nonbinary instruments the IVQR model and, in particular, the rank similarity assumption are testable. Suppose that the researcher has access to two different binary instruments Z_1 and Z_2 that are obtained as transformations of the original instrument respectively instruments and let $\delta_{Z_1}(\tau|x)$ and $\delta_{Z_2}(\tau|x)$ denote the associated QTE estimands. Under Assumption 1, the conditional moment equation (1) implies that

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

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

The intuition behind this testable restriction is relatively simple. Under Assumption 1, the IVQR model yields QTE for the whole population. These treatment effects do not depend on the choice of the instrument because they are not local effects for an instrument-specific subpopulation.[10] Hence, one can use an overidentification-type test to empirically assess the validity of the rank similarity assumption (conditional on having a valid instrument). I consider the following formal testing problem:

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

against

$$H_1 : \delta_{Z_1}(\tau|x) \neq \delta_{Z_2}(\tau|x) \quad \text{for some} \quad (\tau, x) \in \mathcal{TX}$$

Given this representation, specification tests are constructed based on a Kolmogorov-Smirnov-type measure of distance,[11]

$$T_n = \sqrt{n} \sup_{(\tau,x)\in\mathcal{TX}} |\hat{\delta}_{Z_1}(\tau|x) - \hat{\delta}_{Z_2}(\tau|x)|.$$

The test then rejects H_0 when $T_n > \hat{T}_{1-\alpha}$, where $\hat{T}_{1-\alpha}$ is a consistent estimator of the $(1-\alpha)$-quantile of the limiting distribution of T_n. Under the conditions set forth in the previous sections, the critical value $\hat{T}(1-\alpha)$ can be obtained using the exchangeable bootstrap.

6 Empirical application

In this section, I illustrate my method by estimating the distributional impact of 401(k) plans on accumulated assets as in Chernozhukov and Hansen (2004) and Belloni et al. (2014). The goal here is to complement their findings with unconditional QTE based on the IVQR model and to provide an empirical comparison to estimates from a linear IVQR models as in Chernozhukov and Hansen (2004).

6.1 Semiparametric quantile treatment effect estimates

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 problem in estimating the effect of 401(k) plans on accumulated assets is the potential endogeneity of the actual participation status caused by non-random enrollment. To overcome

[10] This is in sharp contrast to the local average treatment effects framework (Imbens and Angrist, 1994; Abadie et al., 2002).

[11] Alternatively, one could consider Cramer-Von-Mises-type test statistics.

this problem, Abadie (2003), Chernozhukov and Hansen (2004), and Belloni et al. (2014) use 401(k) eligibility as an instrument for the actual participation status, arguing that eligibility can be taken to be exogenous after conditioning on a small set of covariates including income.[12] I adopt their 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. (2014). The data consist of 9,915 observations from a sample of households from the 1991 Survey of Income and Program Participation (SIPP). 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. Covariates include dummies for income categories, dummies for age categories, dummies for education categories, a marital status indicator, family size, two-earner status, defined benefit pension status, individual retirement account participation status, homeownership, and a constant. This is identical to the main specification in Chernozhukov and Hansen (2004).

Because only individuals who were eligible could enroll in 401(k) plans, 401(k) eligibility satisfies one-sided non-compliance. Formally, one-sided non-compliance implies that $p(0, x) = 0$ for all $x \in \mathcal{X}$. In practice, this feature reduces the computational burden because $p(0, x)$ and $F_{Y|D,Z,X}(y|1, 0, x)$ need not be estimated.

Figure 1 contains the main empirical results, estimated using a logit link functions and 99 grid points for DR. I construct pointwise and uniform confidence bands using 250 empirical bootstrap replications and a fine grid of quantile indices $\{0.15, 0.16, ..., 0.85\}$. 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. 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.

In Figure 2, I assess the robustness of the results with respect to the choice of the link function. The results indicate that there are almost no visible differences between the probit and the logit link for either outcome variable and only minor differences between these nonlinear link functions and the linear link function at the high quantiles of total wealth. However, these differences are negligible compared to the sampling variation. Overall, the comparison indicates that my results are robust with respect to the choice of the link function.

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

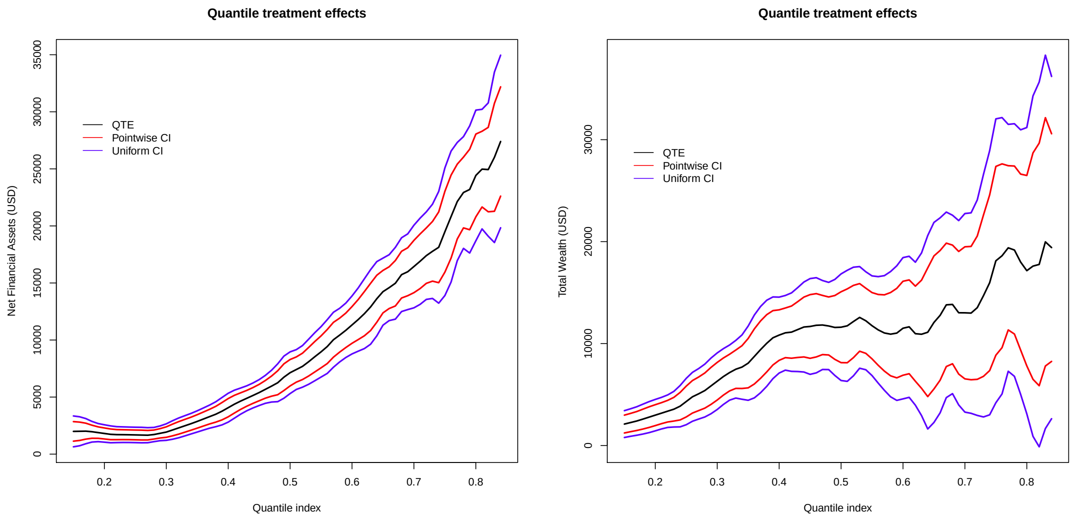

Figure 1: The number of observations is 9,915. The figure reports unconditional QTE, pointwise 95%-CI, and uniform 95%-CI. The CI are obtained from 250 bootstrap replications.

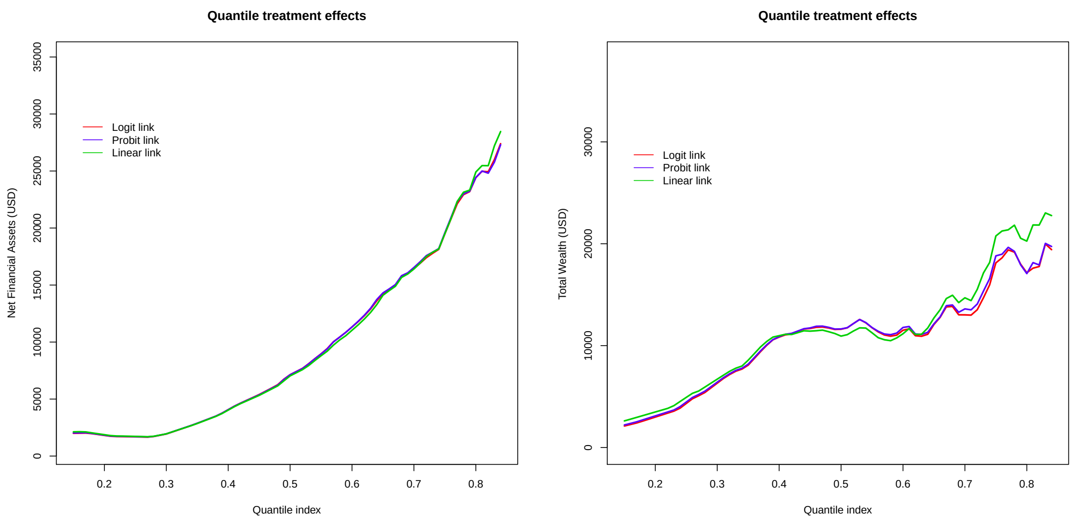

Figure 2: The number of observations is 9,915. The figures compare the QTE estimates based on the logit link to the QTE estimates based on the probit link and a identity link.

6.2 Comparison to the linear-in-parameters model

Figure 3 compares the semiparametric plug-in estimates to the estimates obtained from linear-in-parameters models as in Chernozhukov and Hansen (2004):

$$Q_{Y_D|X}(\tau|x) = D\delta(\tau) + X'\beta(\tau),$$

which are estimated using IQR. The comparison suggests that there are substantial differences regarding the magnitude and shape of the QTE estimates. In particular, for total wealth, the relatively constant pattern for the conditional estimates based on the linear model sharply contrasts with the increasing shape of the unconditional QTE estimates based on the nonseparable model. Interestingly, my estimates are similar to the unconditional QTE estimates for the compliers based on a flexible nonparametric model reported in Belloni et al. (2014), while the results based on the linear-in-parameters IVQR model are comparable to the estimates based on linear models for the compliers (Chernozhukov and Hansen, 2004). The differences between

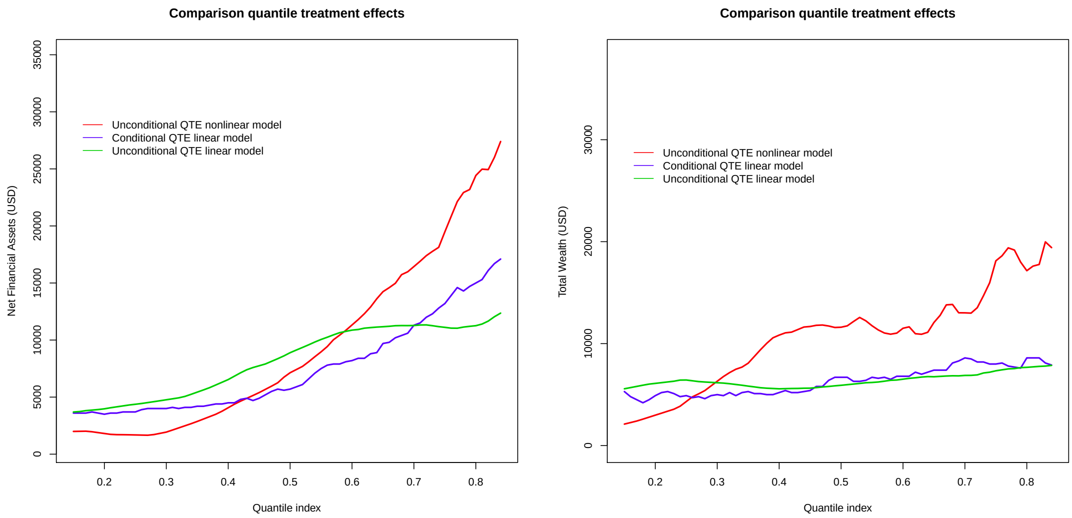

Figure 3: The number of observations is 9,915. The figure displays unconditional QTE estimates based on the semiparametric plug-in approach (red), conditional QTE estimates obtained from a linear model (blue), and unconditional QTE obtained from a linear model (green).

the estimates of both models can be either due to misspecification of the linear model or due to differences between conditional and unconditional QTE. To further assess this issue, I also plot unconditional QTE based on the linear model. These estimates are constructed in three steps. In the first step, the conditional potential outcome cdfs are estimated by inverting the conditional potential outcome quantile functions,

$$\hat{Q}_{Y_1|X}(\tau|x) = \hat{\delta}(\tau) + x'\hat{\beta}(\tau),$$

and

$$\hat{Q}_{Y_0|X}(\tau|x) = x'\hat{\beta}(\tau).$$

Second, I obtain unconditional distributions integrating the conditional distributions with respect to the estimated empirical distribution of the covariates. Finally, I invert the unconditional cdfs to estimate the unconditional QTE. Note that steps two and three are identical to the semiparametric estimation approach. Thus, differences between the estimates are solely due to differences between the conditional potential outcome distributions. The differences between the unconditional QTE estimates suggest that imposing effect homogeneity across covariates can substantially bias the estimates and highlight the importance of more flexible models. Appendix A contains two simple examples to further illustrate this point.

7 Conclusion and directions for future research

This paper proposes semiparametric plug-in estimators for conditional and unconditional QTE based on the IVQR model. Exploiting analytic closed form solutions, the estimation procedure does not rely on separability of the structural quantile function, while retaining computational tractability and $\sqrt{n}$-consistency. I prove functional central limit theorems and establish validity of the exchangeable bootstrap for estimating the limiting laws. The semiparametric estimation approach is applied 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. A comparison to a separable linear-in-parameters model highlights the potential bias arising from separability restrictions and the importance of estimating more flexible models.

The semiparametric plug-in approach relies on parametric first stage estimates of conditional cdfs and conditional probabilities. Although flexible, these parametric models might not be appropriate and one might prefer fully nonparametric estimation approaches. It would thus be interesting to extend the estimation approach to accommodate nonparametric first stage estimators. Such an extension is outside of the the scope of this paper but is certainly worth pursuing in future research.

One important limitation is that the estimation approach is inherently limited to binary treatments because closed from solutions are only available for this important special case. Moreover, despite the fact that nonbinary instruments can be accommodated as outline in Section 5, efficiency of the estimators could be improved upon by using the more general closed form solutions (Wüthrich, 2014, Lemma 2). Deriving more general analytic closed form solution and extending the plug-in approach accordingly thus constitutes a promising extension.

Finally, the specification test for the IVQR model presented in Section 5 could be extended by considering more general test statistics that efficiently make use of all the available overidentifying restrictions.

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.

Angrist, J., Chernozhukov, V., Fernández-Val, I., 2006. Quantile regression under misspecification, with an application to the u.s. wage structure. Econometrica 74 (2), pp. 539–563.

Belloni, A., Chernozhukov, V., Fernández-Vál, I., Hansen, C., 2014. Program evaluation with high-dimensional data. cemmap Working Papers, 33/13.

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

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., 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., D'Haultfoeuille, X., 2014. Supplement to: "fuzzy changes-in-changes". Unpublished Manuscript.

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.

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.

Kaplan, D., Sun, Y., 2014. Smoothed estimating equations for instrumental variables quantile regression. Unpublished Manuscript.

Koenker, R., Leorato, S., Peracchi, F., 2013. Distributional vs. quantile regression. CEIS Tor Vergata, Research Paper Series, Vol. 11, Issue 15, No. 300.

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

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. Unpublished Manuscript.

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.

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.

Su, L., Hosino, T., 2013. Sieve instrumental variable quantile regression estimation of functional coefficient models. Unpublished Manuscript.

Torgovitsky, A., 2012. Identification of nonseparable models with general instruments. Mimeo.

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

Wüthrich, K., 2014. A comparison of two quantile models with endogeneity. Discussion Paper 1408, University of Bern, Department of Economics.

Yu, P., 2014. Marginal quantile treatment effect and counterfactual analysis. Unpublished Manuscript.

A Linear vs. nonlinear models: two simple examples

Section 6.2 shows substantial differences between the unconditional QTE estimates of the linear model and the nonseparable semiparametric estimation approach. To further investigate the differences between linear and nonlinear IVQR models, I analyze two different data generating processes (DGP):

1. $U \sim N(0,1)$, $V \sim N(0,1)$, $Z \sim Bernoulli(0.5)$, $X \sim N(0,1)$, $D = 1\{Z + 0.25 \cdot (X + U) > V\}$, $Y_0 = X \cdot U$, $Y_1 = X + U$, $Y = Y_1 \cdot D + Y_0 \cdot (1 - D)$.

2. Same as DGP 1 but with $Y_1 = X/U$.

DGP 1 and DGP 2 both satisfy Assumption 1 and feature heterogeneity across covariates, the only difference being the specifiation of Y_1. Figure 4 compares unconditional QTE estimated based on a linear model and the IQR algorithm with those based on the nonseparable model estimated using the semiparametric approach proposed in this paper.[13] Panel A shows a similar

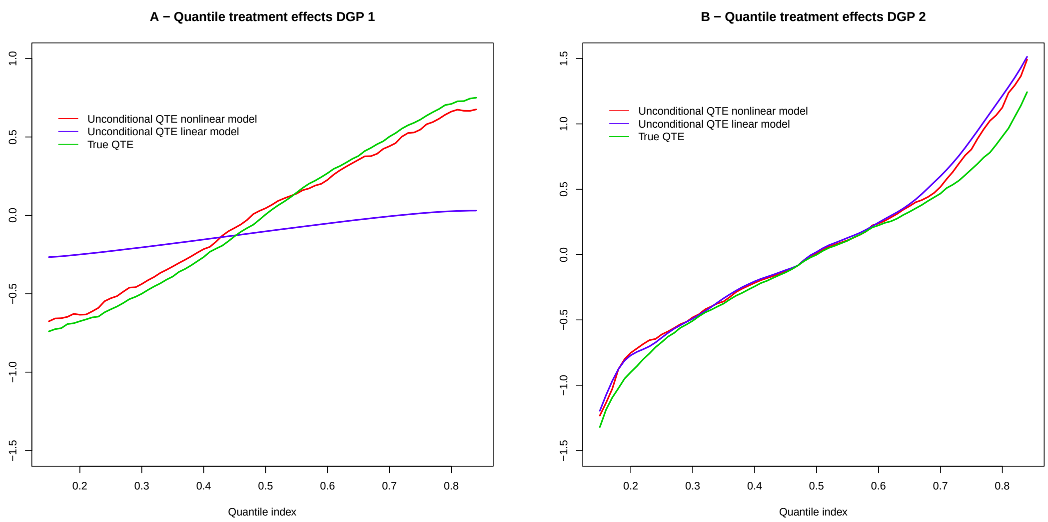

Figure 4: Based on $N = 10'000$. The figure displays unconditional QTE estimates based on the semiparametric plug-in approach, unconditional QTE estimates obtained from a linear model, and the true unconditional QTE.

pattern as in the empirical application: the linear model underestimates the heterogeneity across quantiles compared to the nonseparable model. Moreover, the semiparametric estimates are close to the true QTE. In contrast, Panel B shows a scenario where both estimates are similar and coincide with the true value. These examples illustrate that separability of the structural quantile function can seriously bias QTE estimates for some DGPs while not being restrictive for others.

[13] The IQR algorithm is based on a grid search over $\{-5, -4.99, ..., 5\}$ and the semiparametric estimation approach is based on a logit link function and 199 DR to approximate the conditional distributions.

B Additional Proofs

B.1 Additional notation

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

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

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

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

and

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

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

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

Finally, denote by $C(\mathcal{U})$ the set of continuous functions from $\mathcal{U}$ to $\mathbb{R}$

B.2 Proof of Lemma 1

The proof of Lemma 1 is closely related to the proof of Lemma 1 in Wüthrich (2014).

Step 1: Under Assumption 1, we have that

$$\begin{aligned}
P\left(Y \leq Q_{Y_D|X}(\tau|x)|X = x, Z = 1\right) &= \tau \\
P\left(Y \leq Q_{Y_D|X}(\tau|x)|X = x, Z = 0\right) &= \tau.
\end{aligned}$$

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

$$F_{Y|D,Z,X}\left(Q_{Y_1|X}(\tau|x)|1,1,x\right)p(1,x)+F_{Y|D,Z,X}\left(Q_{Y_0|X}(\tau|x)|0,1,x\right)(1-p(1,x))=\tau \quad (7)$$

$$F_{Y|D,Z,X}\left(Q_{Y_1|X}(\tau|x)|1,0,x\right)p(0,x)+F_{Y|D,Z,X}\left(Q_{Y_0|X}(\tau|x)|0,0,x\right)(1-p(0,x))=\tau. \quad (8)$$

Step 2: Equating (7) and (8) and rearranging terms yields

$$\begin{aligned}&F_{Y|D,Z,X}\left(Q_{Y_1|X}(\tau|x)|1,1,x\right)p(1,x)-F_{Y|D,Z,X}\left(Q_{Y_1|X}(\tau|x)|1,0,x\right)p(0,x)=\\&F_{Y|D,Z,X}\left(Q_{Y_0|X}(\tau|x)|0,0,x\right)(1-p(0,x))-F_{Y|D,Z,X}\left(Q_{Y_0|X}(\tau|x)|0,1,x\right)(1-p(1,x))\end{aligned}$$

Dividing by $p(1,x)-p(0,x)$, we obtain

$$F^c_{Y_1|X}\left(Q_{Y_1|X}(\tau|x)|x\right)=F^c_{Y_0|X}\left(Q_{Y_0|X}(\tau|x)|x\right),$$

by definition. By assumption, $F^c_{Y_1|X}(y|x)$, $F^c_{Y_0|X}(y|x)$, $Q_{Y_1|X}(\tau|x)$, and $Q_{Y_0|X}(\tau|x)$ are strictly increasing. Thus,

$$Q^c_{Y_1|X}\left(F^c_{Y_0|X}(y|x)|x\right)=Q_{Y_1|X}\left(F_{Y_0|X}(y|x)|x\right) \quad (9)$$

and

$$Q^c_{Y_0|X}\left(F^c_{Y_1|X}(y|x)|x\right)=Q_{Y_0|X}\left(F_{Y_1|X}(y|x)|x\right). \quad (10)$$

Step 3: By assumption, $Q_{Y_1|X}(\tau|x)$ and $Q_{Y_0|X}(\tau|x)$ are strictly increasing in τ. Hence, substituting $F_{Y_1|X}(y|x)=\tau$ in equation (7) and $F_{Y_0|X}(y|x)=\tau$ in equation (8), we obtain

$$\begin{aligned}F_{Y_1|X}(y|x)=&p(1,x)F_{Y|D,Z,X}(y|1,1,x)\\&+(1-p(1,x))F_{Y|D,Z,X}\left(Q_{Y_0|X}\left(F_{Y_1|X}(y|x)|x\right)|0,1,x\right),\\F_{Y_0|X}(y|x)=&(1-p(0,x))F_{Y|D,Z,X}(y|0,0,x)\\&+p(0,x)F_{Y|D,Z,X}\left(Q_{Y_1|X}\left(F_{Y_0|X}(y|x)|x\right)|1,0,x\right).\end{aligned}$$

The result then follows by plugging-in (9) and (10) from step 2.

B.3 Proof of Lemma 2

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 earlier work by Melly and Santangelo (2015) and De Chaisemartin and D'Haultfoeuille (2014).

Step 1: Using the arguments in Chernozhukov et al. (2013, Appendix E) and Yu (2014, Proofs of Theorems 4 and 8) it is straightforward to show that

$$\sqrt{n}\left(\hat{\theta}(y)-\theta(y)\right) \rightsquigarrow \mathbb{Z}_\theta(y)$$

as a stochastic process indexed by $y \in \mathcal{Y}$ in the metric space $\ell^\infty(\mathcal{Y})^{\dim(\theta)}$, where $\mathbb{Z}_\theta(y) := -J^{-1}(y)W(y)$. Hence the details are omitted for brevity. Note that, $\mathbb{Z}_\theta(y) = (\mathbb{Z}_\gamma', \mathbb{Z}_\beta'(y))'$, where $\mathbb{Z}_\beta(y)$ is a stochastic process indexed by $y \in \mathcal{Y}$ and $\mathbb{Z}_\gamma$ is a multivariate normal random variable. Next, consider the map $(g, b) \mapsto \varphi(g, b)$, where

$$\varphi(g,b)(x,y) = \begin{pmatrix} \Lambda\left(x'g_0\right) \\ \Lambda\left(x'g_1\right) \\ \Lambda\left(x'b_{00}(y)\right) \\ \Lambda\left(x'b_{01}(y)\right) \\ \Lambda\left(x'b_{10}(y)\right) \\ \Lambda\left(x'b_{11}(y)\right) \end{pmatrix}$$

is Hadamard differentiable at $(g, b(\cdot)) = (\gamma, \beta(\cdot))$ tangentially to $\mathbb{R}^{\dim(\gamma)} \times C(\mathcal{Y})^{\dim(\beta)}$ with derivative map given by $(\eta, \alpha) \mapsto \varphi_{\gamma,\beta(\cdot)}(\eta, \alpha)$, where

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

Therefore, by the functional delta method

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

as a stochastic process indexed by $(y, x) \in \mathcal{YX}$ in the metric space $\ell^\infty(\mathcal{YX})^6$.

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 (Van der Vaart and Wellner, 1996, Lemma 3.9.3). Because,

$$\hat{F}^c_{Y_1|X}(y|x) = \frac{\hat{p}(1,x)\hat{F}_{Y|D,Z,X}(y|1,1,x) - \hat{p}(0,x)\hat{F}_{Y|D,Z,X}(y|1,0,x)}{\hat{p}(1,x)-\hat{p}(0,x)},$$
$$\hat{F}^c_{Y_0|X}(y|x) = \frac{(1-\hat{p}(0,x))\hat{F}_{Y|D,Z,X}(y|0,0,x) - (1-\hat{p}(1,x))\hat{F}_{Y|D,Z,X}(y|0,1,x)}{\hat{p}(1,x)-\hat{p}(0,x)},$$

by the functional delta method

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

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

$$\mathbb{Z}_{F^c_1}(y|x) := \frac{F_{Y|D,Z,X}(y|1,1,x) - F^c_{Y_1|X}(y|x)}{p(1,x)-p(0,x)}\mathbb{Z}_{p_1}(x) + \frac{F^c_{Y_1|X}(y|x) - F_{Y|D,Z,X}(y|1,0,x)}{p(1,x)-p(0,x)}\mathbb{Z}_{p_0}(x)$$
$$+ \frac{p(1,x)}{p(1,x)-p(0,x)}\mathbb{Z}_{F_{11}}(y|x) - \frac{p(0,x)}{p(1,x)-p(0,x)}\mathbb{Z}_{F_{10}}(y|x),$$

and

$$\mathbb{Z}_{F^c_0}(y|x) := \frac{F_{Y|D,Z,X}(y|0,1,x) - F^c_{Y_0|X}(y|x)}{p(1,x)-p(0,x)}\mathbb{Z}_{p_1}(x) + \frac{F^c_{Y_0|X}(y|x) - F_{Y|D,Z,X}(y|0,0,x)}{p(1,x)-p(0,x)}\mathbb{Z}_{p_0}(x)$$
$$- \frac{1-p(1,x)}{p(1,x)-p(0,x)}\mathbb{Z}_{F_{01}}(y|x) + \frac{1-p(0,x)}{p(1,x)-p(0,x)}\mathbb{Z}_{F_{00}}(y|x).$$

Next, by Hadamard differentiability of the inverse map (uniformly with respect to an index) (Chernozhukov et al., 2010) and the functional delta method,

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

where $\mathbb{Z}_{Q^c_1}(y|x)$ and $\mathbb{Z}_{Q^c_1}(y|x)$ are tight mean-zero Gaussian processes given by

$$\mathbb{Z}_{Q^c_1}(y|x) := -\mathbb{Z}_{F^c_1}\left(Q^c_{Y_1|X}(y|x)|x\right)/f^c_{Y_1|X}\left(Q^c_{Y_1|X}(y|x)|x\right)$$
$$\mathbb{Z}_{Q^c_0}(y|x) := -\mathbb{Z}_{F^c_0}\left(Q^c_{Y_0|X}(y|x)|x\right)/f^c_{Y_0|X}\left(Q^c_{Y_0|X}(y|x)|x\right)$$

By Lemma 3.9.27 in Van der Vaart and Wellner (1996), which is valid uniformly with respect to an index under Assumption 2 and the functional delta method, obtain

$$\sqrt{n}\left(\hat{Q}^c_{Y_0|X}\left(\hat{F}^c_{Y_1|X}(y|x)|x\right) - Q^c_{Y_0|X}\left(F^c_{Y_1|X}(y|x)|x\right)\right) \rightsquigarrow \mathbb{Z}_{Q^c_0 \circ F^c_1}(y|x) \;\; \text{in} \;\; \ell^\infty(\mathcal{YX}),$$

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

$$\mathbb{Z}_{Q_0^c \circ F_1^c}(y|x) := \mathbb{Z}_{Q_0^c}\left(F_{Y_1|X}^c(y|x)|x\right) + \frac{\mathbb{Z}_{F_1^c}(y|x)}{f_{Y_0|X}^c\left(Q_{Y_0|X}^c\left(F_{Y_1|X}^c(y|x)|x\right)|x\right)},$$

and

$$\sqrt{n}\left(\hat{F}_{Y|D,Z,X}\left(\hat{Q}_{Y_0|X}^c\left(\hat{F}_{Y_1|X}^c(y|x)|x\right)|0,1,x\right) - F_{Y|D,Z,X}\left(Q_{Y_0|X}^c\left(F_{Y_1|X}^c(y|x)|x\right)|0,1,x\right)\right) \rightsquigarrow \mathbb{Z}_{F_{01} \circ Q_0^c \circ F_1^c}(y|x) \text{ in } \ell^\infty(\mathcal{YX}),$$

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

$$\begin{aligned}\mathbb{Z}_{F_{01} \circ Q_0^c \circ F_1^c}(y|x) :=& \mathbb{Z}_{F_{01}}\left(Q_{Y_0|X}^c\left(F_{Y_1|X}^c(y|x)|x\right)|x\right) \\ &+ f_{Y|D,Z,X}\left(Q_{Y_0|X}^c\left(F_{Y_1|X}^c(y|x)|x\right)|0,1,x\right)\mathbb{Z}_{Q_0^c \circ F_1^c}(y|x).\end{aligned}$$

Similarly obtain,

$$\sqrt{n}\left(\hat{F}_{Y|D,Z,X}\left(\hat{Q}_{Y_1|X}^c\left(\hat{F}_{Y_0|X}^c(y|x)|x\right)|1,0,x\right) - F_{Y|D,Z,X}\left(Q_{Y_1|X}^c\left(F_{Y_0|X}^c(y|x)|x\right)|1,0,x\right)\right) \rightsquigarrow \mathbb{Z}_{F_{10} \circ Q_1^c \circ F_0^c}(y|x) \text{ in } \ell^\infty(\mathcal{YX}),$$

where

$$\begin{aligned}\mathbb{Z}_{F_{10} \circ Q_1^c \circ F_0^c}(y|x) :=& \mathbb{Z}_{F_{10}}\left(Q_{Y_1|X}^c\left(F_{Y_0|X}^c(y|x)|x\right)|x\right) \\ &+ f_{Y|D,Z,X}\left(Q_{Y_1|X}^c\left(F_{Y_0|X}^c(y|x)|x\right)|1,0,x\right)\mathbb{Z}_{Q_1^c \circ F_0^c}(y|x),\end{aligned}$$

and

$$\mathbb{Z}_{Q_1^c \circ F_0^c}(y|x) := \mathbb{Z}_{Q_1^c}\left(F_{Y_0|X}^c(y|x)|x\right) + \frac{\mathbb{Z}_{F_0^c}(y|x)}{f_{Y_1|X}^c\left(Q_{Y_1|X}^c\left(F_{Y_0|X}^c(y|x)\right)|x\right)}.$$

Finally, because

$$\begin{aligned}\hat{F}_{Y_1|X}(y|x) =& \hat{p}(1,x)\hat{F}_{Y|D,Z,X}(y|1,1,x) \\ &+ (1-\hat{p}(1,x))\,\hat{F}_{Y|D,Z,X}\left(\hat{Q}_{Y_0|X}^c\left(\hat{F}_{Y_1|X}^c(y|x)|x\right)|0,1,x\right), \\ \hat{F}_{Y_0|X}(y|x) =& (1-\hat{p}(0,x))\,\hat{F}_{Y|D,Z,X}(y|0,0,x) \\ &+ \hat{p}(0,x)\hat{F}_{Y|D,Z,X}\left(\hat{Q}_{Y_1|X}^c\left(\hat{F}_{Y_0|X}^c(y|x)|x\right)|1,0,x\right),\end{aligned}$$

by the functional delta method obtain

$$\sqrt{n}\begin{pmatrix}\hat{F}_{Y_1|X}(y|x) - F_{Y_1|X}(y|x)\\ \hat{F}_{Y_0|X}(y|x) - F_{Y_0|X}(y|x)\end{pmatrix} \rightsquigarrow \begin{pmatrix}\mathbb{Z}_{F_1}(y|x)\\ \mathbb{Z}_{F_0}(y|x)\end{pmatrix} \quad \text{in} \quad \ell^\infty(\mathcal{Y}\mathcal{X})^2,$$

where $\mathbb{Z}_{F_1}(y|x)$ and $\mathbb{Z}_{F_0}(y|x)$ are tight mean-zero Gaussian processes defined as

$$\begin{aligned}
\mathbb{Z}_{F_1}(y|x) :=& \left(F_{Y|D,Z,X}(y|1,1,x) - F_{Y|D,Z,X}\left(Q^c_{Y_0|X}\left(F^c_{Y_1|X}(y|x)|x\right)|0,1,x\right)\right)\mathbb{Z}_{p_1}(x)\\
&+ p(1,x)\mathbb{Z}_{F_{11}}(y|x) + (1-p(1,x))\,\mathbb{Z}_{F_{01}\circ Q^c_0\circ F^c_1}(y|x)\\
\mathbb{Z}_{F_0}(y|x) :=& \left(F_{Y|D,Z,X}\left(Q^c_{Y_1|X}\left(F^c_{Y_0|X}(y|x)|x\right)|1,0,x\right) - F_{Y|D,Z,X}(y|0,0,x)\right)\mathbb{Z}_{p_0}(x)\\
&+ (1-p(0,x))\,\mathbb{Z}_{F_{00}}(y|x) + p(0,x)\mathbb{Z}_{F_{10}\circ Q^c_1\circ F^c_0}(y|x),
\end{aligned}$$

This completes the proof of the lemma.